# UNIVERSIDAD NACIONAL DE CHIMBORAZO


**FACULTAD DE INGENIERÍA**

**CARRERA DE SISTEMAS Y COMPUTACIÓN**

Proyecto de Investigación previo a la obtención del título de Ingeniero en Sistemas y Computación

**TRABAJO DE TITULACIÓN**

IMPLEMENTACIÓN DE UN MODELO DE PREDICCIÓN BASADO EN REDES NEURONALES ARTIFICIALES PARA LA CLASIFICACIÓN DE INFORMACIÓN ACADÉMICA Y DE INVESTIGACIÓN DE LA UNACH

**AUTOR:**

Ilda Jackeline Quizhpilema Lazo

**TUTOR:**

Ing. Xavier Bustamante., MsC.

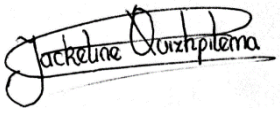
**Riobamba - Ecuador**

**2020**

## AUDITORIA DE LA INVESTIGACIÓN

La responsabilidad del contenido de este proyecto de graduación, corresponde exclusivamente a Ilda Jackeline Quizhpilema Lazo, autora del proyecto de investigación, bajo la dirección del MsC. Wayner Xavier Bustamante Granda y al patrimonio intelectual de la Universidad Nacional de Chimborazo.

MsC. Wayner Xavier Bustamente
1103951677
**Director del Proyecto**



Ilda Jackeline Quizhpilema
0605219294
**Autora**

**PÁGINA DE ACEPTACIÓN**

Los miembros del tribunal de Graduación del proyecto de investigación de título: "IMPLEMENTACIÓN DE UN MODELO DE PREDICCIÓN BASADO EN REDES NEURONALES ARTIFICIALES PARA LA CLASIFICACIÓN DE INFORMACIÓN ACADÉMICA Y DE INVESTIGACIÓN DE LA UNACH", presentado por la Srta. Ilda Jackeline Quizhpilema Lazo, dirigido por el MsC. Wayner Xavier Bustamante Granda. Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación escrito, con fines de graduación en el cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la Facultad de Ingeniería de la UNACH.

Para constancia de lo expuesto firman:

MsC. Xavier Bustamante …….
**Tutor** Firma

PhD. Lida Barba …….
**Miembro del tribunal** Firma

MsC. Ana Congacha …….
**Miembro del tribunal** Firma

**DEDICATORIA**

*Dedico este proyecto de investigación a mis padres Víctor & Nelly, mi padre es el ejemplo perfecto de esfuerzo, constancia, dedicación, mi madre es mi amor más grande, mi fortaleza e inspiración. Todo lo que pueda conseguir en la vida es gracias a las oportunidades que ellos me dieron, no solamente materiales sino espirituales. Gracias por estar a mi lado, creer en mí y apoyarme incondicionalmente en este proceso de formación académica.*

*A mis abuelitos, Salvador y Carlota, por acogerme en su hogar, ser alegría a mi vida y darme todo su cariño.*

*A mis hermanos Beatriz, Belén, Elizabeth, Israel, Joel, Elid, Josué, y Abigail; por brindarme su cariño, amor y ser mí soporte en momentos difíciles.*

***Ilda Jackeline Quizhpilema Lazo***

*"El éxito es la suma de pequeños esfuerzos que se hacen día tras día"* ***Robert Collier****.*

## AGRADECIMIENTO

*A Dios por guiar mis pasos y concederme culminar con éxito mi sueño más anhelado.*

*A la Universidad Nacional de Chimborazo por brindarme la oportunidad, a los maestros de la Escuela de Sistemas y Computación por compartir sus conocimientos, concejos, experiencias y pasión a la enseñanza.*

*Al tutor el MsC. Xavier Bustamante, a los Miembros del tribunal PhD. Lida Barba, MsC. Ana Congacha. quienes fueron un pilar fundamental para la culminación del presente trabajo investigación.*

*A mis compañeros por formar parte de mi vida académica.*

***Ilda Jackeline Quizhpilema Lazo***

*"Porque Jehová da la sabiduría, Y de su boca viene el conocimiento y la inteligencia".* ***Proverbios. 2:6***

**ÍNDICE GENERAL**

## ÍNDICE DE TABLAS

## ÍNDICE DE ILUSTRACIONES

**RESUMEN**

En el presente trabajo de investigación se aplicó técnicas de minería de datos de clasificación, como son las Redes Neuronales Artificiales (ANN), con aprendizaje supervisado, de tipo Perceptrón Multicapas (MLP) con el algoritmo de Retropropagación (BP), para obtener un modelo que sea capaz de clasificar factores que afectan el rendimiento académico de los estudiantes, la evaluación integral a los docentes, la publicación de productos de investigación, y docentes que publican en alto impacto, revistas regionales, capítulo de libro, a partir de las bases de datos del Sistema Informático de Control Académico (SICOA) y el Sistema de Publicaciones de la Dirección de Investigación (SPDI) de la Universidad Nacional de Chimborazo (UNACH), generando conocimiento útil en la toma de decisiones de los directivos, y, por ende, a la comunidad en la que se encuentra inmersa.

La confiabilidad del modelo es de 94,72%, para el desarrollo de esta investigación se utilizó el Proceso Estándar Industrial Híbrido para la Minería de Datos (CRISP-DM) como metodología, se hizo Extracción, Transformación y Carga (ETL) con la herramienta *Talend Data Quality*, se construyó el modelo con las herramientas *RapidMiner* 9.5 y R Studio 1.2, se usó el método de validación cruzada (CV) y finalmente se obtuvo factores, como por ejemplo: el rendimiento académico de los estudiantes es excelente cuando él género es femenino, soltero(a), no es foráneo, no trabaja, no tiene hijos, tiene hermanos, practica actividad deportiva y cultural, y más, que se detallan en el Capítulo IV (Resultados y Discusión).

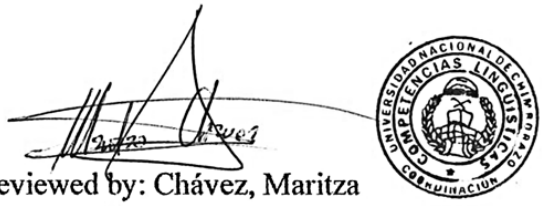
**Palabras claves:** Minería de datos, Redes Neuronales Artificiales, Clasificación, Retropropagación, CRISP-DM, *RapidMiner*.

## ABSTRACT

In the present research work was applicable techniques of classifying data mining, such as Artificial Neural Networks (ANN), with supervised learning, Multilayer Perceptron (MLP) with the Backpropagation algorithm (BP), to obtain a model that is able to classify factors that affect the academic achievement of the students, the comprehensive evaluation to the teachers, the publication of fact-finding products, and teachers who publish in high impact, regional magazines, chapter of book, from the data bases of the Information-Technology System of Academic Control (SICOA) and Publications' System of the fact-finding Management (SPDI) of the National University of Chimborazo (UNACH), generating useful knowledge in the decision making of the executives, and, as a consequence, to the community in which it finds absorbed. The reliability of the model belongs to 94.72 %, developmental of this investigation the Standard Industrial Hybrid Process for the Data Mining (CRISP DM) like methodology used, it became Extraction, Transformation and Loads (ETL) with the tool Talend Data Quality, RapidMiner made the model with the tools 9,5 and R Studio 1,2, the method of crossed validation used (VF) and finally factor obtained, like for instance: the student's academic performance is excellent when gender is Feminine, single, is not foreign, does not work, has no children, has siblings, practices sports and cultural activities, and more, which are detailed in Chapter IV (Results and Discussion).



**Keywords:** Data mining, Artificial Neural Networks, Classification, backpropagation, CRISP-DM, RapidMiner.

Reviewed by: Chávez, Maritza

Language Center Teacher

## INTRODUCCIÓN

Con el avance de la tecnología las Instituciones de Educación Superior (IES), captan, gestionan, almacenan y publican día a día una masiva cantidad de datos relacionados con los procesos docente, investigación, vinculación entre otros (Rodríguez, 2015). Una buena parte de estos datos tiene tal densidad que no puede analizarse con técnicas de análisis tradicionales, el reto consiste en analizar científicamente en los campos de la estadística, inteligencia artificial y matemática computacional (Aluja, 2001). La Minería de Datos (DM) se conoce como descubrimiento de conocimiento, aprendizaje automático y análisis predictivo, puede manejar grandes volúmenes con múltiples atributos e implementar algoritmos complejos que automaticen el proceso de búsqueda de una solución óptima para un problema de datos específico. Cada tarea de minería de datos utiliza algoritmos específicos como árboles de decisión, redes neuronales, k-vecinos más cercanos, agrupación de k-medias, entre otros (Kotu & Deshpande, 2014).

Las Redes Neuronales Artificiales (ANN) simulan el funcionamiento de una red de neuronas biológicas presentes en el cerebro humano, permitiendo aprender a partir de experiencias (Maithili, Kumari, & Rajamanickam, 2012). Existen distintos tipos de redes neuronales, dependiendo del tipo de aprendizaje que requiera. El tipo de red más utilizado en clasificación y predicción es el Perceptrón Multicapas que usa el aprendizaje *backpropagation* (Pitarque, Ruiz, & Roy, 2000), para minimizar la función del error entre la salida deseada y la obtenida del modelo neuronal a partir de un conjunto de observaciones ya clasificadas (Viñuela & Leon, 2004). En el contexto del uso de una ANN para predecir patrones futuros de comportamiento en el ámbito educativo, Pinninghoff, *et al*. (2007), utilizaron una red neuronal multicapas para predecir el éxito o

fracaso de estudiantes utilizando los datos de Programa Internacional para la Evaluación de Estudiantes (PISA), obteniendo una precisión de más del 75% en la clasificación (Pinninghoff, Salcedo, & Contreras, 2007). González (2015), muestra el funcionamiento del Perceptrón Multicapa como modelo clasificador de las conductas adictivas, y la eficacia del modelo seleccionado fue del 74% (González M. V., 2015). Orellana *et al.* (2018), usa el perceptrón multicapa al momento de clasificar una organización y presenta un porcentaje de precisión de 83.19% (Orellana Parapi, 2018). Estos trabajos descritos clasifican conductas y predicen el éxito o fracaso de los estudiantes. En el mismo contexto, el enfoque del presente trabajo fue identificar patrones de comportamiento en el rendimiento académico del estudiante, la calificación obtenida en la evaluación al docente y la clasificación de docentes en si generan o no producción científica, con información almacenada en la base de datos del Sistema Informático de Control Académico y del Sistema de Publicaciones de la Dirección de Investigación de la Universidad Nacional de Chimborazo.

La metodología CRISP-DM, dividida en seis fases: compresión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación es la más usada en problemas de minería de datos, por tal razón fue aplicada para guiar la investigación.

El trabajo de titulación se estructura de: Resumen, Introducción, el capítulo I contiene el planteamiento del problema y la definición de los objetivos de la investigación, en el capítulo II aborda el marco teórico, en el capítulo III define el proceso metodológico empleado en el desarrollo del proyecto, en el capítulo IV se realiza el análisis de resultados, las respectivas conclusiones y recomendaciones.

# CAPÍTULO I.

## 1. PLANTEAMIENTO DEL PROBLEMA

### Problema y Justificación

En las últimas décadas, la educación evidencia cambios de gran trascendencia en el entorno universitario, a nivel demográfico, económico, tecnológico y competitivo. Indudablemente, esto envuelve a los sistemas de dirección, organización y gestión (Llinás-Audet, Girotto, & Solé, 2011). Las IES deben responder a las demandas de su entorno, para asegurar un lugar pertinente en la desafiante sociedad (Duro & Gilart, 2016).

Uno de los retos que tienen que enfrentar las IES para ofrecer una mayor calidad educativa, es mejorar el rendimiento académico de los estudiantes y el nivel de preparación de docentes (González E. G., 2017). La producción científica es una forma tangible y objetiva de medir la experiencia científica y la competencia en investigación (Mejia, Valladares-Garrido, & Valladares-Garrido, 2018) , y es un indicador valioso en el proceso de evaluación y acreditación de las universidades.

Según el Art. 3 del reglamento de evaluación integral al desempeño del personal académico de la Universidad Nacional De Chimborazo su fin es contribuir a los indicadores de evaluación institucionales, de carreras y programas; enfocándose principalmente en la disminución de la retención estudiantil y al incremento de la ciencia terminal; proporcionar a las autoridades resultados cuantitativos y cualitativos, que constituyan un sustento valido para la toma de decisiones (Gerrero, 2019).

Las bases de datos del SICOA y del SPDI contienen información académica e investigativa valiosa, que no ha sido analizada mediante las redes neuronales artificiales que gracias a su excelente comportamiento en predicción y clasificación. A criterio del investigador, no se han analizado factores clave que inciden en el proceso de publicación, entre ellos rendimiento académico de los estudiantes de todas las facultades de la UNACH, la valoración que obtienen los docentes en los procesos de evaluación integral a su desempeño, las características presentadas en los docentes que publican y los que no publican, y el tipo de publicación como; alto impacto, revistas regionales y capitulo de libro.

**Objetivos:**

### 1.1.1. Objetivo General

Implementar un modelo de predicción basado en Redes Neuronales Artificiales para la clasificación de información de los sistemas de control académico y de investigación de la Universidad Nacional de Chimborazo que permita mejorar a la toma de decisiones de los directivos de la institución.

### 1.1.2. Objetivos Específicos

- Identificar parámetros representativos en los campos de las bases de datos del SICOA y del SPDI de la UNACH, para ser usados como entradas a un sistema clasificador implementado con Redes Neuronales Artificiales.
- Aplicar las redes neuronales artificiales *Backpropagation* (BP) utilizando los parámetros representativos obtenidos de los datos académicos y de investigación de la UNACH.
- Analizar los resultados de la aplicación del algoritmo *Backpropagation* de redes neuronales artificiales y verificar la confiabilidad del modelo.

## CAPÍTULO II.

## 2. MARCO TEÓRICO

### Minería de Datos

Es un conjunto de técnicas y tecnologías que permiten navegar dentro de grandes cantidades de información o bases de datos en un mínimo de tiempo de manera automática, permite extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos (Moine, Haedo, & Gordillo, 2011), siendo su principal objetivo encontrar información oculta o implícita, que no es posible obtener mediante métodos estadísticos convencionales (Benalcázar Tamayo, 2017).

### Learning Analytics

Está relacionado con el aprendizaje personalizado y adaptativo, con incidencia en todas las disciplinas educativas, siendo una disciplina que interactúa con otras de gran relevancia en esta década como la Minería de datos educativos (EDM), la inteligencia empresarial (BI), el análisis de redes sociales (SNA) y lo relativo a Machine Learning (ML) (Gutiérrez-Priego, 2015).

Consiste en la interpretación de un amplio rango de datos producidos y recogidos acerca de los estudiantes para orientar su progresión académica, predecir actuaciones futuras e identificar elementos problemáticos. El objetivo de la recolección, registro, análisis y presentación de estos datos es posibilitar que los profesores puedan adaptar de manera rápida y eficaz las estrategias educativas al nivel de necesidad y capacidad de cada alumno. Aun, en sus primeras etapas de desarrollo, las analíticas de aprendizaje responden a la necesidad de llevar a cabo el seguimiento y control de la actividad en el

campus para la toma de decisiones estratégicas. Por otro lado, pretenden aprovechar la gran cantidad de datos producidos por los estudiantes en actividades académicas (Durall, Gros, Maina, Johnson, & Adams, 2012).

**Minería de datos Educativa**

Es una disciplina emergente que busca desarrollar nuevos métodos para explorar la información que se genera dentro de los ambientes educativos con el fin de entender la forma en que los estudiantes aprenden para poder tomar las decisiones adecuadas que garanticen el éxito en el proceso educativo (Huapaya, Lizarralde, Arona, & Massa, 2012).

Por otro lado, Amaury *et al*. (2013), definieron unas categorías para la aplicación de EDM, como son Análisis y visualización de datos, soporte a la instrucción, comportamiento de los estudiantes, predicción del rendimiento académico como también de comportamientos indeseables, caracterización de estudiantes, análisis de redes sociales, desarrollo de mapas conceptuales, currículo y organización de las actividades escolares (Quinteros, Funes, & Ahumada, 2016).

**Tipos de Minería de Datos**

Los problemas de minería de datos se pueden clasificar en términos generales en modelos de aprendizaje supervisados o no supervisados. Supervisados o dirigidos predicen el valor de las variables de salida en función de un conjunto de variables de entrada. Para hacer esto, se desarrolla un modelo de un conjunto de datos de entrenamiento donde los valores de entrada y salida son previamente conocido. El modelo generaliza la relación entre las variables de entrada y salida y lo usa para predecir el conjunto de datos donde solo las variables de entrada son conocidos. La variable de salida que se predice también se denomina clase etiqueta o variable objetivo. La minería de datos supervisada necesita un

número suficiente de registros etiquetados para aprender el modelo de los datos. Sin supervisión o sin dirección la minería de datos descubre patrones ocultos en datos sin etiquetar. En datos no supervisados, no hay variables de salida para predecir. El objetivo de esta clase de técnicas de minería de datos es encontrar patrones en los datos basados en la relación entre los puntos de datos en sí. Una aplicación puede emplear tanto supervisados y aprendices sin supervisión (Kotu & Deshpande, 2014). Las tareas de minería de datos también se pueden agrupar en clasificación, regresión, análisis de asociación, detección de anomalías, series de tiempo, tareas de minería de texto y *feature selection* como muestra la Ilustración 1.

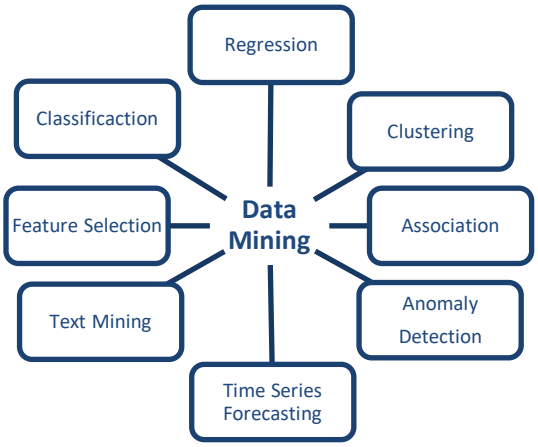



*Ilustración 1. Tareas de Minería de Datos.*
*Adaptado de: (Kotu & Deshpande, 2014)*

**Herramientas de Minería de Datos.**

Las herramientas de minería de datos o las herramientas de programación estadística, como R, *RapidMiner*, SAS *Enterprise Miner*, IBM SPSS, etc., pueden implementar algoritmos con facilidad. Estas herramientas de minería de datos ofrecen una biblioteca de algoritmos como funciones, que pueden interactuar a través del código de programación o la configuración a través de interfaces gráficas de usuario.

### 2.1.1. Inspiración

La ANN están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Estos elementos están organizados de una forma parecida a la que presenta el cerebro humano (Olabe, 1998, pág. 2) Ver Ilustración 2.

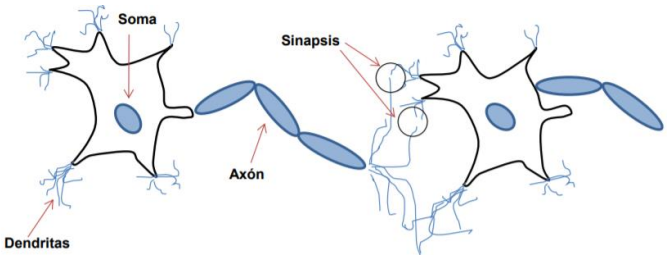



*Ilustración 2. Estructura de una neurona biológica.*
*Recuperado de: (Manjarrez, 2014)*

Según Olabe (1998) las partes de una neurona biológica son:

- **Célula nerviosa(soma)**: Donde las señales son combinadas (**Cuerpo celular**)
- **Dendrita**: Combina el input desde muchas otras células nerviosas (**entradas**)
- **Sinapsis**: Interfaz entre dos neuronas (**Conexiones**)
- **Axón**: Transporta la señal de activación a células nerviosas en diferentes partes (**salida**).

El cerebro humano es el sistema de cálculo más complejo que conoce el hombre. El ordenador y el ser humano realiza diferentes tareas de forma eficiente; así la operación de reconocer el rostro de una persona resulta una tarea relativamente sencilla para el hombre y difícil para el ordenador, mientras que la contabilidad de una empresa es tarea

costosa para un experto contable y una sencilla rutina para un ordenador básico (Olabe, 1998, pág. 1).

### 2.1.2. Descripción

Son capaces de detectar y aprender patrones y características de los datos, además, una vez adiestradas las redes puede hacer previsiones, clasificaciones y segmentación (Aguilar & Estrada, 2012). Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y del pasado, y aplicando tal conocimiento a la resolución de problemas nuevos. Este aprendizaje se obtiene como resultado del adiestramiento ("*training*") y éste permite la sencillez, la potente adaptación y evolución, ante una realidad cambiante muy dinámica.

### 2.1.3. Estructura de una neurona artificial

La distribución de neuronas dentro de la red se realiza formando niveles o capas de un número determinado de neuronas cada una, se pueden distinguir tres tipos de capas:

- De entrada: es la capa que recibe directamente la información proveniente de las fuentes externas a la red
- Ocultas: son internas a la red y no tienen contacto directo con el entorno exterior. Pueden estar interconectadas de distintas maneras, lo que determina junto con su número las distintas tipologías de redes.
- De salida: transfieren información de la red hacia el exterior. (Larranaga, Inza, & Moujahid, 1997, pág. 5).

### 2.1.4. Funciones de activación

La función de activación se elige de acuerdo con la tarea realizada por la neurona. La función de activación que se suele usar en el Perceptrón Multicapa es la función sigmoidal que muestra la ecuación 1. Por lo tanto, los rangos de valores de los atributos deben escalarse a -1 y +1 (González M. V., 2015).

$$y = \frac{1}{1 + e^{-x}} \qquad (1)$$

El tipo de aprendizaje es supervisado, es decir, que es el usuario quien determina la salida deseada. Para que la red pueda aprender y adquiera esa capacidad de generalizar se diferencian dos etapas: de entrenamiento y de funcionamiento (González M. V., 2015).

### 2.1.5. Modelo estándar de neurona artificial

Según los principios descritos en Rumelhart & McClelland (1986), la neurona artificial estándar consiste en:

$$y_j = \varphi\left(\sum_{i=1}^{m} w_{ji} x_i\right) \qquad (2)$$

- Un conjunto de entradas $xi$, pesos sinápticos $wji$, con i = 1, . . ., m.
- Una regla de propagación definida a partir del conjunto de entradas y los pesos sinápticos ($x1$, . . ., $xm$, $w11$, $w12$, …, $w1m$), siendo la más común la aditiva (sumatoria de $wjixi$).
- Una función de activación φ, la cual representa simultáneamente la salida de la neurona y su estado de activación.

### 2.1.6. Topologías de las redes neuronales artificiales

Las topologías comunes en una red neuronal son feed-forward (FNN), la información va en una única dirección desde la entrada a la salida y recurrente (ANN), la información se realimenta dentro de la red neuronal, sin descartar el uso de topologías mixtas como se ve en la Ilustración 3. El peso de la conexión de las neuronas es el principal componente directamente determinado por la topología, también la velocidad de los cálculos es afectada por la topología de la red. En términos de intercambio de datos (Larranaga, Inza, & Moujahid, 1997).

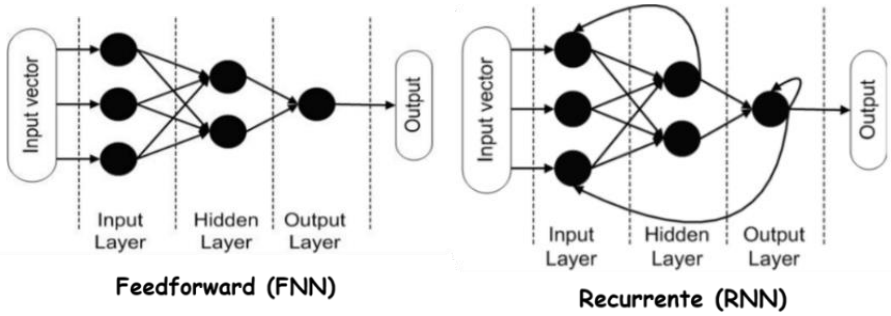



*Ilustración 3. Arquitecturas de redes neuronales.*
*Adaptado de (Larranaga, Inza, & Moujahid, 1997)*

**Perceptrón Multicapa**

El (MLP *Multi-Layer Perceptron*) surge en la década de 1980, como una solución para superar el problema detectado en el Perceptrón Simple, en cuanto a su imposibilidad de aprender clases de funciones no linealmente separables (Larranaga, Inza, & Moujahid, 1997). Esta red es la más usada actualmente, su importancia se debe principalmente a su potencia y generalidad, también es capaz de actuar como aproximador universal de funciones lo que hace de él uno de los modelos más útiles en la práctica. La operación de este sistema se relaciona con la regresión no lineal. El MLP constituye el modelo de ANN más utilizado para la resolución de problemas de ingeniería, varias investigaciones han

demostrado su condición de aproximador universal de funciones (Palmer, Montaño, & Jiménez, 2001).

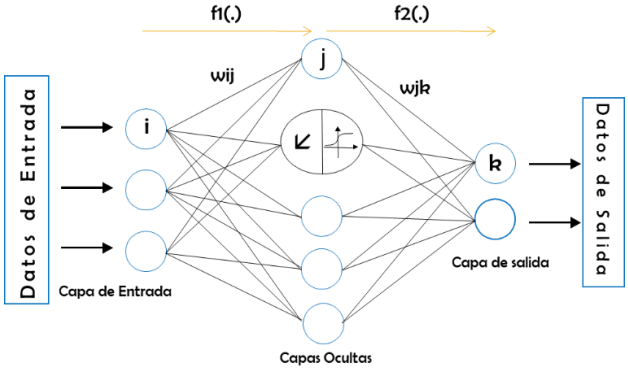



*Ilustración 4. Perceptrón Multicapa*
Adaptado de: (Mitchell, 1997)

Una red neuronal es una función

$$Y = F(X, W) \qquad (3)$$

donde Y es el vector formado por las salidas de la red, X es el vector de entrada a la red, y W es el conjunto de todos los parámetros de la red (pesos y umbrales), y F una función continua no lineal, como muestra la ilustración 5.

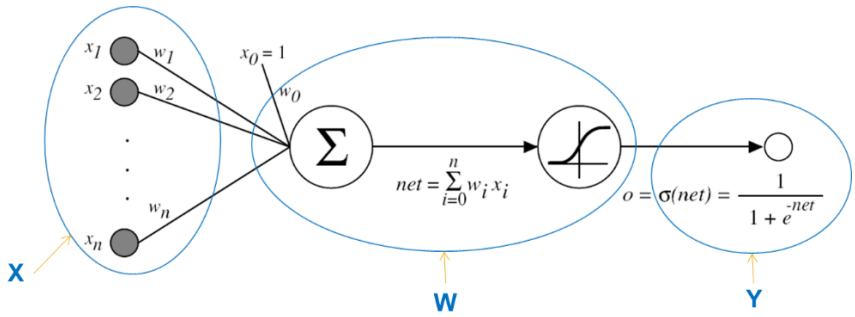



*Ilustración 5. Función de la red neuronal*
Adaptado de: (Mitchell, 1997)

**Algoritmo Backpropagation**

Las redes Backpropagation tienen un método de entrenamiento supervisado. A la red se le presenta parejas de patrones, un patrón de entrada emparejado con un patrón de salida

deseada. Por cada presentación los pesos son ajustados de forma que disminuya el error entre la salida deseada y la respuesta de la red. El algoritmo de aprendizaje backpropagation conlleva una fase de propagación hacia adelante y otra fase de propagación hacia atrás. Ambas fases se realizan por cada patrón presentado en la sesión de entrenamiento (Olabe, 2017).

### 2.1.7. Funcionamiento del Algorimo *Backpropagation* en una ANN

El proceso seguido por una ANN es bastante intuitivo y se parece mucho a la transmisión de señal en las neuronas biológicas. El modelo utiliza cada registro de entrenamiento para estimar el error (*backpropagation*) de la salida predicha en comparación con la salida real. Luego, el modelo usa el error para ajustar los pesos para minimizar el error para el próximo registro de entrenamiento y este paso se repite hasta que el error se encuentre dentro del rango aceptable. A continuación, los pasos del funcionamiento de una ANN.

1. Determinar la topología y la función de activación
2. Iniciación y primer registro de entrenamiento.
3. Error de cálculo (*Backpropagation*)
4. Ajuste de peso

**Paso 1: determinar la topología y la función de activación**

Para este ejemplo, supongamos un conjunto de datos con tres atributos de entrada numéricos ($X_1$ $X_2$ $X_3$ :) y una salida numérica (Y). Para modelar la relación, usando una topología con dos capas y una función de activación de agregación simple.

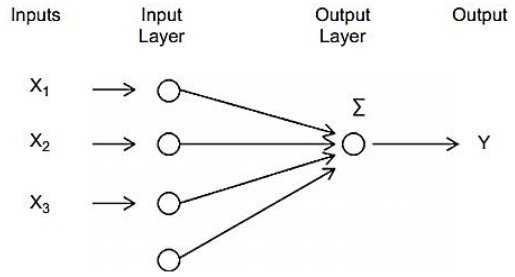



*Ilustración 6. Topología de dos capas*

**Paso 2: iniciación**

Supongamos que los pesos iniciales para los cuatro enlaces son 1, 2, 3 y 4. Tomemos un modelo de ejemplo y un registro de prueba con todas las entradas como 1 y la salida conocida como 15. Entonces, $X_1 = X_2 = X_3 = 1$ y salida Y = 15. La ilustración 7 muestra el inicio del primer registro de entrenamiento.

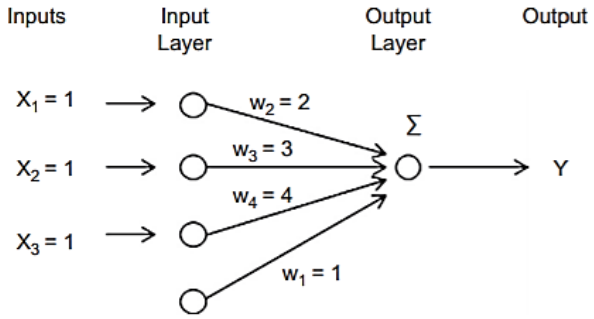



*Ilustración 7. Iniciación y primer registro de entrenamiento.*

**Paso 3: error de cálculo**

Podemos calcular la salida pronosticada del registro de la ilustración 7. Este es un simple proceso de avance de pasar por los atributos de entrada y calcular la salida pronosticada. La salida predicha $\bar{y}$ según el modelo actual es 1+ 1 *2+ 1* 3 + 1* 4= 10. La diferencia entre la salida real del registro de entrenamiento y la salida pronosticada es el error.

$$e = y - \bar{y} \qquad (4)$$

El error para este registro de entrenamiento de ejemplo es 15 -10 =5

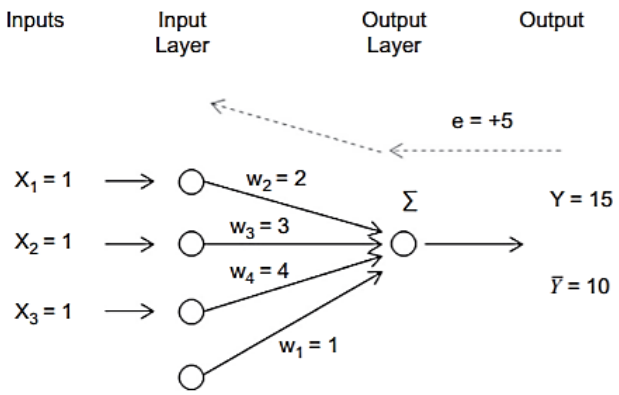



*Ilustración 8. Error de red neuronal propagación hacia atrás.*

**Paso 4: ajuste de peso**

El ajuste de peso es la parte más importante del aprendizaje en una red neuronal artificial. El error calculado en el paso anterior se devuelve desde el nodo de salida a todos los demás nodos en la dirección inversa. Los pesos de los enlaces se ajustan a partir de su valor anterior por una fracción del error. La fracción $\lambda$ aplicada al error se llama tasa de aprendizaje. $\lambda$ toma valores de 0 a 1. Un valor cercano a 1 produce un cambio drástico en el modelo para cada registro de entrenamiento y un valor cercano a 0 produce cambios más pequeños y menos corrección. El nuevo peso del enlace (w) es la suma del peso anterior (w') y el producto de la tasa de aprendizaje y proporción del error ($\lambda * e$).

$$w = w' + \lambda * e \qquad (5)$$

La elección de $\lambda$ puede ser complicado en la implementación de un ANN. Algunos procesos modelo comienzan con $\lambda$ un valor cercano a 1 y reducen el valor de $\lambda$ un tiempo de entrenamiento en cada ciclo. Con este enfoque, los registros atípicos posteriores en el ciclo de capacitación no degradarán la relevancia del modelo. La ilustración 8 muestra la propagación del error en la topología.

El peso actual del primer enlace es $w_2$ =2. Supongamos que la tasa de aprendizaje es 0.5. El nuevo peso será $w_2$=2 +0.5 *5/3= 2.83. El error se divide entre 3 porque el error se devuelve a tres enlaces desde el nodo de salida. Del mismo modo, el peso se ajustará para

todos los enlaces. En el próximo ciclo, se calculará un nuevo error para el próximo registro de entrenamiento. Este ciclo continúa hasta que todos los registros de entrenamiento se procesen mediante ejecuciones iterativas. El mismo ejemplo de entrenamiento puede repetirse hasta que la tasa de error sea inferior a un umbral. Hemos revisado un caso muy simple de una red neuronal artificial. En realidad, habrá múltiples capas ocultas y múltiples enlaces de salida, uno para cada valor de clase nominal. Debido al cálculo numérico, un modelo ANN funciona bien con entradas y salidas numéricas.

Si la entrada contiene un atributo nominal, un paso de preprocesamiento debe ser incluido para convertir el atributo nominal en múltiples atributos numéricos, uno para cada valor de atributo, este proceso es similar a la introducción de variable ficticia, en la Predicción de series de tiempo Este preprocesamiento específico aumenta el número de enlaces de entrada para la red neuronal en el caso de atributos nominales y, por lo tanto, aumenta los recursos informáticos necesarios. Por lo tanto, un ANN es más adecuado para atributos con un tipo de datos numéricos (Kotu & Deshpande, 2014).

**Neuralnet**

En *RapidMiner,* los operadores del modelo ANN están disponibles en la carpeta Clasificación. Hay tres tipos de modelos disponibles: uno simple como es el perceptrón con una capa de entrada y una de salida, un modelo ANN flexible llamado Neural Net con todos los parámetros para la construcción completa del modelo y el algoritmo avanzado de AutoML (Perceptrón multicapa automático) combina conceptos de algoritmos genéticos y tocásticos. Aprovecha un conjunto de ANN con diferentes parámetros, como capas ocultas y tasas de aprendizaje. También se optimiza al reemplazar los modelos de peor desempeño por mejores y mantiene una solución óptima.

Este operador aprende un modelo por medio de una red neuronal alimentada hacia adelante entrenada por un algoritmo de propagación hacia atrás (perceptrón multicapa). Este operador no puede manejar atributos polinomiales.

**Matriz de confusión (o tabla de verdad)**

El rendimiento de la clasificación se describe mejor con una herramienta bien denominada llamada matriz de confusión. Comprender la matriz de confusión requiere familiarizarse con varias definiciones. Pero antes de introducir las definiciones, debemos mirar una matriz de confusión básica para una clasificación binaria o binomial donde puede haber dos clases (por ejemplo, Y o N). La precisión de la clasificación de un ejemplo específico se puede ver de una de las cuatro formas posibles:

- La clase predicha es Y, y la clase real también es Y = es: "Verdadero Positivo" o TP
- La clase predicha es Y, y la clase real es: N = es "Falso Positivo "o FP
- La clase predicha es N, y la clase real es: Y = es "Falso Negativo "o FN
- La clase predicha es N, y la clase real también es: N = es "Verdadero negativo" o TN

*Tabla 1. Matriz de Confusión*

| | | Actual Class (Observation) | |
|---|---|---|---|
| | | Y | N |
| Predicted Class (expectation) | Y | TP (True positive) Resultado Correcto | FP (false positive) Unexpected result |
| | N | FN (False negativo) Missing result | TN (true negative) Correct absence of result |

Recuperado de: (Kotu & Deshpande, 2014)

Una forma rápida de examinar esta matriz o una "tabla de verdad", como también se la llama, es escanear la diagonal desde la parte superior izquierda a la parte inferior derecha. Un rendimiento de clasificación ideal solo tendría entradas a lo largo de esta diagonal principal y los elementos fuera de la diagonal serían cero.

Estos cuatro casos se utilizarán ahora para introducir varios términos de uso común para comprender y explicar el rendimiento de la clasificación. Como se mencionó anteriormente, un clasificador perfecto no tendrá entradas para FP y FN (es decir, el número de FP = número de FN = 0).

*Tabla 2. Medidas de Evaluación*

| Term | Definition | Calculation |
|---|---|---|
| Sensitivity | Ability to select what needs to be selected | TP/(TP+FN) |
| Specificity | Ability to reject what needs to be rejected | TN/(TN+FP) |
| Precision | Proportion of cases found that were relevant | TP/(TP+FP) |
| Recall | Proportion of all relevant cases that were found | TP/(TP+FN) |
| Accuracy | Aggregate measure of classifier performance | (TP+TN)/(TP+TN+FP+FN) |

Recuperado de: (Kotu & Deshpande, 2014)

## Validación Cruzada de "n-folds"

En la ilustración 9 se muestra el *training & test* que usa la validación cruzada, y que consiste en:

- Se subdividen los datos en *n* subconjuntos disjuntos.
- Se considera la evaluación de *n*-1 subconjuntos para el entrenamiento del modelo y 1 subconjunto para la prueba. Este proceso se repite hasta que los n subconjuntos fueron evaluados como prueba.
- Una estimación del error es el promedio de los errores considerados para las *n* evaluaciones de la prueba.
- El caso más usado es una validación cruzada de 10-folds.
- K Validación Cruzada de "*n-folds*"

Se evalúa *K* veces una validación cruzada de *n-folds*:

Loop *K* veces {

Se mezclan los datos y se dividen en *n* subconjuntos

Loop *n veces {*

*(n-1)* folds son usados para entrenamiento

1 fold es usado para prueba
}

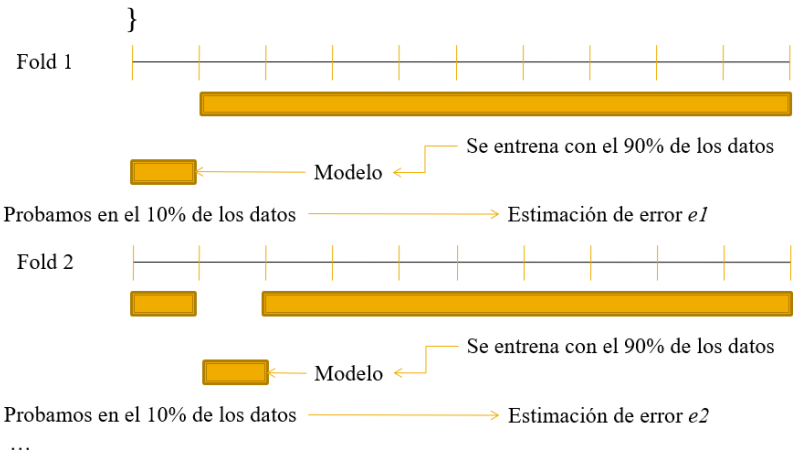
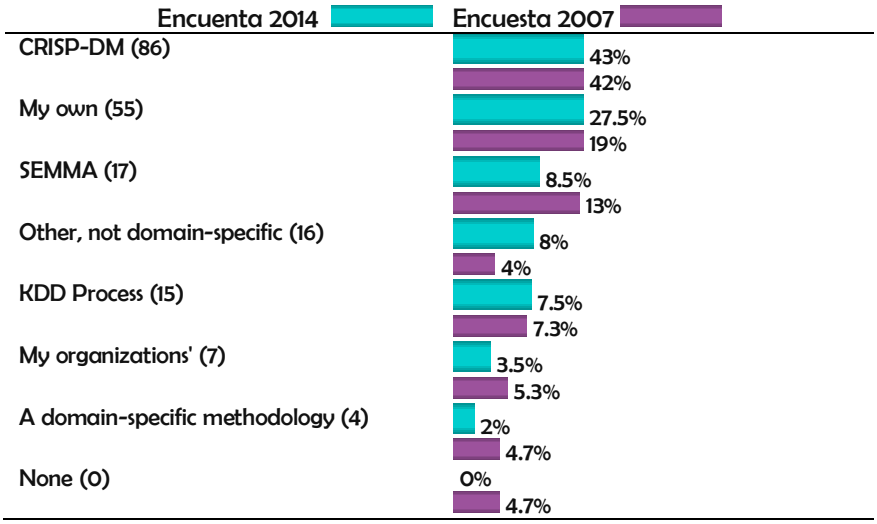



*Ilustración 9. Ejemplo validación Cruzada de "10-folds"*
Recuperado de: (Reveco & Vergara, 2018)

**CRISP-DM**

(*Cross Industry Standard Process for Data Mining*), es una de las metodologías más utilizadas actualmente para proyectos de minería de datos (Piatetsky, 2014), en gran cantidad de organizaciones y desde sus inicios fue respaldada por diversas empresas tanto públicas como privadas, debido a su calidad y efectividad siendo así una de las favoritas (Colina, 2017). Cómo se puede observar la tabla 3 publicada en (kdnuggets, 2007), de 200 votos en total CRISP-DM ha experimentado un ligero ascenso del 1% en los últimos años.

*Tabla 3. Metodologías para proyectos de análisis, minería de datos.*




Recuperado de: (kdnuggets, 2014)

CRISP-DM está estructurado en seis fases o etapas que funcionan de manera cíclica e iterativa, cada una cuenta con tareas generales y específicas que permitan cumplir con los objetivos del proyecto (Chapman, y otros, 2000). En la ilustración 10, se puede observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

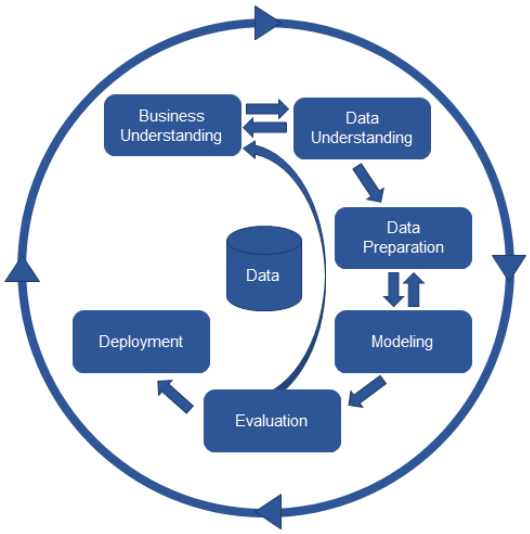



*Ilustración 10. Fases de la Metodología CRISP-DM*
Recuperado de: (kdnuggets, 2017)

**Herramienta *RapidMiner***

*RapidMiner* es una plataforma de minería de datos de código abierto desarrollada y mantenida por *RapidMiner* Inc. El software se conocía anteriormente como YALE (Yet Another Learning Environment). Entorno para aprendizaje automático y para procesos de minería de datos (Bermúdez & Acevedo, 2010). Permite realizar todos los procesos que intervienen en un proyecto: la adquisición de datos, la transformación de los datos, la selección de datos, la selección de atributos, la transformación de los atributos, el aprendizaje/modelización y la validación. Además, permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Lo que hace posible aumentar la productividad a través de modelos que

solucionan los problemas de predicción, clasificación y segmentación de la información (Jaramillo & Arias, 2015).

**Talend Data Quality**

Según Garther (2019), en su informe Gartner Magic Quadrant for Data Integration Tools, esta herramienta es líder en toda la funcionalidad, el rendimiento y la escalabilidad para crear datos precisos y fiables. Talend ofrece Talend Open Studio, Talend Data Fabric, Talend Data Management Platform, Talend Big Data Platform, Talend Data Services Platform, Talend Integration Cloud y Talend Stitch Data Loader. La base de clientes de pago del proveedor para esta cartera de productos es de más de 3,200 organizaciones.


*Ilustración 11. Cuadrante mágico para herramientas de integración de datos Recuperado de: (Zaidi, Heudecker, & Thoo, 2019).*

## CAPÍTULO III.

## 3. METODOLOGÍA

En investigación aplicada es muy común encontrar situaciones en las que debemos estimar o predecir el comportamiento de una variable criterio en función de una o varias variables predictoras. Cuando el criterio es una variable cuantitativa se suele hablar de problemas de predicción o estimación, mientras que cuando es una variable cualitativa/categorial se habla entonces de problemas de clasificación por lo que esta investigación aplicó el enfoque cualitativo que usó variables categóricas para entrada del modelo como son estado civil, nivel de instruido, genero, entre otros, y para la salida excelente, bueno e insuficiente, entre otros.

Los estudios en los cuales se aplica minería de datos se caracterizan por el descubrimiento del conocimiento sin el planteamiento de una hipótesis previamente conocida o preconcebida, donde el investigador comienza examinando los hechos en sí y en el proceso desarrolla una teoría coherente para representar lo que observa.

### Tipo y Diseño de Investigación

Se realizó el tipo de investigación bibliográfica para la revisión de literatura basada en técnicas que se emplearon para obtener información de tesis, artículos científicos, revista, libros, entre otros.

Para el análisis y compresión de datos se efectuó la investigación descriptiva porque se encarga de puntualizar las características de la población que se está estudiando, su objetivo es descubrir la naturaleza de un segmento demográfico, describe el tema de investigación el por qué ocurre, busca especificar las propiedades, las características y los

perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis y la investigación exploratoria para conocer las variables, una comunidad, un contexto, un evento, una situación, es aplicada a problemas de investigación nuevo o poco conocidos, esta se aplica al analizar los datos cualitativos (Hernández, Fernández, & Pilar, 2014), se describieron las tablas y gráficos y se exploraron los valores válidos, no válidos y nulos.

En la preparación de los datos, las investigaciones cualitativas se basan más en una lógica y proceso inductivo (explorar y describir, y luego generar perspectivas teóricas) que al observar las características van de lo particular a lo general. Los algoritmos de aprendizaje por inducción nos permiten obtener resultados de un proceso de aprendizaje supervisado, alimentándose con una colección de datos de entrenamiento. No todos los atributos son relevantes para la clasificación, por lo tanto, la elección de los atributos relevantes interviene el método analítico para comprender mejor comportamiento de cada variable, también se usó la investigación exploratoria porque brinda información y facilita la compresión de los datos al realizar la calidad de datos y obteniendo así los valores categóricos necesarios para el análisis.

En la implementación y evaluación del modelo se aplicó las redes neuronales a través del operador neural net de la herramienta *RapidMiner* versión 9.1.

**Unidad de análisis**

La unidad de análisis para esta investigación está basada en las bases de datos del SICOA y del SPDI.

**Población de estudio**

Son los estudiantes y docentes de las bases de datos SICOA percibido entre el periodo septiembre 2012 - marzo 2013 y octubre 2018 - marzo 2019 y del SPDI desde diciembre de 2013 hasta abril de 2018.

**Tamaño de muestra**

La base de datos del SICOA contiene 16008 estudiantes, 4097 docentes y del SPDI 12050 publicaciones.

**Técnicas de recolección de Datos**

La información fue proporcionada en dos archivos de Excel una base en cada archivo, por el departamento de la Unidad Técnica de Control. Académico (UTECA) de la UNACH.

**Técnicas de Análisis e interpretación de la información**

Se implementó las ANN de clasificación, con un tipo de aprendizaje supervisado, es decir, que es el usuario quien determina la salida deseada, específicamente el Perceptrón Multicapa con el algoritmo de *backpropagation* en la herramienta *RapidMiner*.

**Aplicación de la metodología CRISP-DM**

Divide el proceso en seis fases principales como es Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. Cada fase es descompuesta en varias tareas generales de segundo nivel como se detallan a continuación:

### 3.1.1. Fase 1: Comprensión del negocio o problema

Es la más importante del proyecto donde se determinó los objetivos de DM, se evalúo la situación, se realizó la planificación del proyecto para el rendimiento académico de los estudiantes, la evaluación final de los docentes y la publicación de docentes que se detallan a continuación.

**a) Determinar el objetivo del negocio:** Generar conocimiento que contribuya a la toma de decisiones de los directivos de la institución.

**b) Evaluación de la situación**

En esta tarea se analizó los recursos Hardware, Software, como se muestra en las tablas 4 y 5.

*Tabla 4. Recursos Hardware*

| Marca | Modelo | Procesador | RAM | Disco Duro | Sistema Operativo |
|---|---|---|---|---|---|
| DELL | Inspiron 15-7569 | Intel Core i7-6500U 2.50 GHz 2.60 GHz | 12 GB | 512GB | Windows 10 Home |

*Tabla 5. Recursos Software*

| Software | Utilidad |
|---|---|
| Talend Data Quality Online | Calidad de datos |
| Rapid Miner Studio 9.5 | Implementar la ANN |
| Paquete de Microsoft Office 365 | Digitalización de resultados |

**c) Determinación de los objetivos de DM**

Los objetivos que se determinaron para el análisis son:

- Establecer qué características ostenta un estudiante para tener un promedio Excelente, Bueno o Regular.
- Establecer qué características ostenta un docente en el resultado de evaluación Excelente, Bueno o Insuficiente.

- Establecer qué características ostenta un docente que realiza publicaciones.
- Identificar los factores que influyen que un docente realice publicaciones de alto impacto, revistas regional y capitulo de libro.

**d) Producción de un plan del proyecto**

El plan de proyecto que contiene seis fases comprendidas desde el 26 de noviembre de 2019 y el 20 de febrero de 2020, con una duración de 85 días como se detallan en el Anexo 1.

### 3.1.2. Fase 2: Comprensión de los datos

Esta fase incluye la recopilación inicial de datos, descripción de los datos, exploración de los datos y verificación de la calidad de datos que se detallan a continuación.

**a) Recolección de datos iniciales**

La base de datos del SICOA de la UNACH contiene:

- Información demográfica del estudiante
- Información académica del estudiante
- Información demográfica del docente
- Información académica del docente
- Evaluación del docente

La base de datos del Sistema de Publicaciones de la UNACH contiene:

- Publicaciones de Docentes

**b) Descripción de los datos**

Los datos recopilados se dividen en varias tablas como son estudiante, Estudiante_rendimiento, docente, docente, _infAcadémica, y publicaciones, la descripción de cada tabla se encuentra en el Anexo 2.

**c) Exploración de datos**

Se explora los datos mediante ilustraciones con el fin de hacer un análisis estadístico y conocer la distribución que existe por cada variable ver Anexo 3.

**d) Verificación de la calidad de los datos**

En esta tarea fueron analizados mediante la herramienta Talend Data Quality Online para verificar la calidad de los datos, es decir verificar errores de codificación, como la cantidad de valores válidos, inválidos y valores nulos que se detallan a continuación.

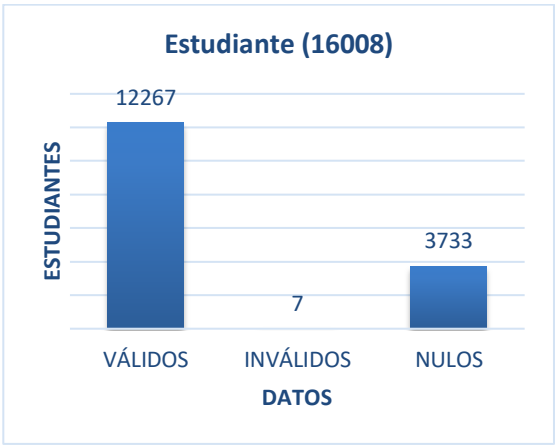



*Ilustración 12. Calidad de datos de la tabla Estudiante*

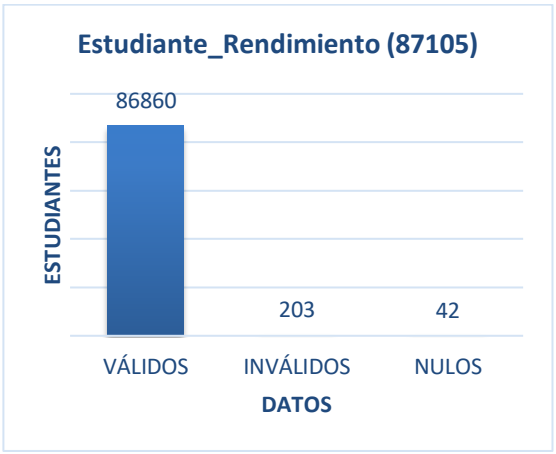



*Ilustración 13. Calidad de datos de la tabla Estudiante_rendimiento*

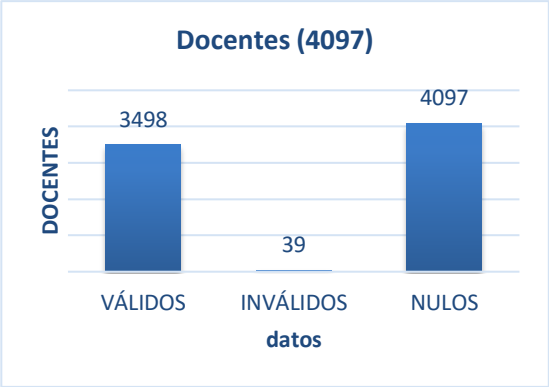



*Ilustración 14. Calidad de datos de la tabla Docente*

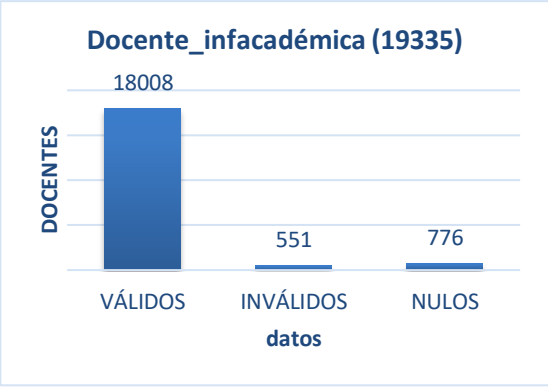



*Ilustración 15. Calidad de datos de la tabla Docente_infAcadémica*

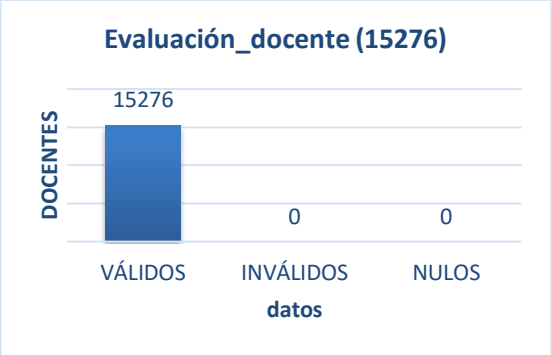



*Ilustración 16. Calidad de datos de la tabla Evaluación_Docente*


*Ilustración 17. Calidad de datos de la tabla Publicaciones.*

### 3.1.3. Fase 3: Preparación de los datos

La preparación del conjunto de datos para adaptarse a una tarea de minería de datos es la parte más lenta del proceso. Muy raramente los datos están disponibles en la forma requerida por los algoritmos de minería de datos. A continuación, se detalla el proceso específico de selección, limpieza, estructuración, integración y formateo de datos.

**a) Selección de datos**

Existen atributos que no son necesarios para nuestros objetivos de minería de datos definidos en la fase 1 (comprensión del negocio) de la metodología., por lo que se puede prescindir de algunos de ellos para conducir a un modelo más simplificado ayudando a sintetizar una explicación más efectiva del modelo. Para saber que

atributos son valiosos y cuáles no valen nada se usó la Heramienta *RapidMiner*, Construcción de Modelos predictivo,s *Auto Model*, porque ayudó a tomar una decisión indicando el valor del atributo con una burbuja de estado codificada por colores (rojo / amarillo / verde). Los detalles son proporcionados por las barras de calidad (C / I / S / M / T).

- (C) Columnas que reflejan demasiado la columna de destino,
- (I) Columnas donde casi todos los valores son diferentes,
- (S) Columnas donde casi todos los valores son idénticos,
- (M) Columnas con valores faltantes,
- (T) Columnas que parecen contener texto libre.

Según Kotu & Deshpande. (2014), Como regla general, es una buena idea anular la selección de al menos los Atributos que tienen una burbuja de estado roja. Por tal razón en la entrada del modelo de aprendizaje automático solo incluirá los atributos seleccionados por defecto es decir los de color amarillo y verde. Ver anexo 4.

**b) Limpieza de los datos**

Los modelos de redes neuronales para tareas de clasificación no funcionan bien con atributos faltantes y, por lo tanto, la preparación y selección de datos es esencial para en esta tarea. Con la finalidad de dar tratamiento a las inconsistencias encontradas, y poder generar un modelo de calidad se realizó la eliminación de registros duplicados, para construir un modelo representativo, se ignoró todos los registros de datos con valor perdido o registros con calidad de datos deficiente, este método redujo el

tamaño del conjunto de datos. Por ende, los datos ignorados se describen a continuación.

*Tabla 6. Preparación de datos*

| TABLA | ATRIBUTOS | | |
|---|---|---|---|
| | Entrada | Derivados | Eliminados |
| Estudiantes | EstudianteID<br>EstadoCivil<br>Género | ActividadDeportiva, ActividadCultural, EsForáneo TieneHermanos, Trabaja, TieneHijos, Promedio, PonderaciónPromedio | Institución Educativa, Enfermedad Catastrófica Extraña, Tipo Discapacidad, Porcentaje Discapacidad, País Nacimiento, Cantón Nacimiento, Parroquia, Tipo Vivienda, Tipo construcción, Servicio Agua Potable, Servicio electricidad, Servicio Teléfono, Servicio Internet, Servicio TV Pagada, Valor Mensual Servicios, Tiene Vehículo, Ocupación Conyugue, Ingresos Conyugue, Personas Dependen Ingresos, Situación Actual, Nivel. |
| Docente | Cedula<br>EstadoCivil<br>Género<br>NivelInstrucción | TieneHijos, EventosNacionales EventosInternacionales, HorasActividadAcadémica, Horas Clase, ResultadoFinalEvaluación Docente | Tipo Sangre, Grupo GLBTI, País, Cantón, Parroquia, Tiempo Estudio, Modalidad, Área, Subárea, Campo, Está Cursando, Institución Educativa, Titulo, Experiencia Privada, Experiencia Pública, Familiar Sustituto, Enfermedad Catastrófica, Tiene Discapacidad, Gestación Lactancia, Numero Documento, Actividad Académica, Usuario Evaluado, Tipo Evaluación, Componente |
| Publicaciones | Cedula<br>EstadoCivil<br>Género<br>NivelInstrucción | EventosNacionales, HorasActividadAcadémica, HorasClase, EventosInternacionales, Eventos Nacionales, PublicacionProduccionCientifica2, PublicacionesRegionalRevista3, PublicacionesLibro, PubCapítuloLibro, PublicacionesPonencia, TienePublicaciones | Título, Cédula, Rol Institución, Sexo, Tipo Autor, Orden Autor, Nombres, Apellido Materno, Apellido Paterno, Año, Año Mes Publicación, Año mes Registro, ciudad de publicación, Existe Comité Científico u Organizador, Existe Comité Editorial, Existe Procedimiento Selectivo, Existe Revisión por Pares Externos, Listado de Revistas SENESCYT, Estado Personal Académico, Organismo de afiliación |

### c) Estructuración de los datos

En esta tarea se realizó la derivación de campos, por ejemplo, para el campo Tiene hermanos de la tabla estudiante, consistió en asignar un valor de "Si" a todos los estudiantes que posean hermanos sin importar la cantidad, y "No" aquellos que tienen una cantidad de hermanos de cero. De la misma manera se realizó para todos los campos derivados que muestra el Anexo 5.

A continuación, la tabla 7 y 8, describen ponderaciones de la calificación del estudiante y evaluación docente respectivamente.

*Tabla 7. Ponderación promedio de la tabla Estudiante_rendimiento*

| Rango de calificación | Ponderación Promedio |
|---|---|
| 9.00 - 10.00 | Excelente |
| 7.00 - 8.99 | Bueno |
| Menos 7.00 | Regular |

*Tabla 8. Ponderación resultado final de la tabla Evaluación_docente*

| Rango de calificación | Equivalencia Calificación |
|---|---|
| 90.00 - 100.00 | Excelente |
| 70.00 - 89.99 | Muy Bueno |
| Menos de 70 | Insuficiente |

**d) Integración de los datos**

Se realizó la integración entre tablas de las bases de datos, Quedando: Estudiantes, Docentes y Publicaciones, como se detalla a continuación.

*Tabla 9. Integración de datos.*

| Tablas Final | Tablas de la Base de Datos |
|---|---|
| Estudiante | Estudiante, Estudiante_rendimiento. |
| Docente | Docente, Docente_infAcadémica y Evaluación_Docente |
| Publicaciones | Docente y Publicaciones |

**e) Formateo de datos**

Los tipos de datos son relevantes para comprender más sobre los datos y cómo se obtienen los datos. No todas las tareas de minería de datos se pueden realizar en todos los tipos de datos. Por ejemplo, el algoritmo de red neuronal no funciona con datos categóricos. Sin embargo, podemos convertir datos de un tipo de datos a otro mediante un proceso de conversión de tipos. Es por eso por lo que se codificó una puntuación numérica para cada valor de categórico como se detalla a continuación.

*Tabla 10. Formateo de Datos Tabla Estudiante*

| Campo | Descripción | Valor |
|---|---|---|
| Estado Civil | Soltero (a) | 1 |
| | Casado (a) | 2 |
| | Unión libre | 3 |
| | Divorciado (a) | 4 |
| | Viudo (a) | 5 |
| Género | Masculino | 1 |

| | | |
|---|---|---|
| | Femenino | 0 |
| Practica Actividad Deportiva, Practica Actividad Cultural, Foráneo, Tiene Hermanos, Trabaja, Tiene Hijos | SI | 1 |
| | NO | 0 |

*Tabla 11. Formateo De Datos Tabla Docente*

| Campo | Descripción | Valor |
|---|---|---|
| Estado Civil | Soltero (a) | 1 |
| | Casado (a) | 2 |
| | Unión libre | 3 |
| | Divorciado (a) | 4 |
| | Viudo (a) | 5 |
| Género | Masculino | 1 |
| | Femenino | 0 |
| Nivel Instrucción | Posgrado PhD | 1 |
| | Posgrado Maestría | 2 |
| | Posgrado Especialidad Área Salud | 3 |
| | Especialidad | 4 |
| | Superior Universitaria Completa | 5 |
| Tiene Horas de Clase, Tiene Hijos Tiene Eventos Nacionales, Tiene Eventos Internacionales | SI | 1 |
| | NO | 0 |

*Tabla 12. Formateo de datos Tabla Publicaciones*

| Campo | Descripción | Valor |
|---|---|---|
| Estado Civil | Soltero (a) | 1 |
| | Unión libre | 2 |
| | Casado (a) | 3 |
| | Divorciado (a) | 4 |
| | Viudo (a) | 5 |
| Género | Masculino | 1 |
| | Femenino | 0 |
| Nivel Instrucción | Posgrado PhD | 1 |
| | Posgrado Maestría | 2 |
| | Posgrado Especialidad Área Salud | 3 |
| | Especialidad | 4 |
| | Superior Universitaria Completa | 5 |
| Tiene Hijos, Tiene Eventos Nacionales, Tiene Eventos, Internacionales, Tiene Publicaciones, Publicación Producción Cientifica2, Publicaciones RegionalRevista3, Publicaciones Libro, PubCapítuloLibro, PublicacionesPonencia | SI | 1 |
| | NO | 0 |

### 3.1.4. Fase 4: Modelado

En esta fase se selecciona las técnicas de modelado, se genera un plan de prueba con el método de validación cruzada, se construye el modelo en *RapidMiner*, y se evalúa el modelo mediante RStudio, a continuación, se describe cada tarea.

**a) Selección de la técnica de modelado**

De los modelos que nos ofrece *RapidMiner*, el que mejor se adapta a nuestros objetivos son las redes neuronales artificiales, puesto que los problemas que queremos resolver son de clasificación. Se aplicó el algoritmo Neural Net, con una topología de tipo multicapas multilayer *perceptrón* (*MLP*) el algoritmo *Backpropagation* (BP).

**b) Generación del plan de prueba**

El modelo se realizará mediante el operador *(cross validation)* es un operador que se utiliza para estimar el rendimiento estadístico de un modelo de aprendizaje, especialmente para validar modelos predictivos. Consiste Enel cual toma los datos originales y crear dos subprocesos: uno de entrenamiento y otro de validación. Este proceso se repetirá k veces, y en cada iteración se elegirá un conjunto diferente de prueba, mientras que los datos restantes se utilizarán como conjunto de entrenamiento. Una vez terminadas las iteraciones, se calcula la precisión y el error para cada modelo producido (*RapidMiner*, 2019).

**c) Construcción del Modelo**

En este apartado, se tiene los datos necesarios y preparados para generar los modelos. En el Anexo 6, se describe el proceso completo de la implementación del modelo de clasificación predictiva de la ANN.

**d) Evaluación del modelo**

Para cumplir con el objetivo 3 referente a la verificación de la confiabilidad del modelo se utilizó la herramienta R Studio, con las mismas configuraciones de *RapidMiner*, representándose así los mismos resultados, toda la información se encuentra descrita en el Anexo 7.

### 3.1.5. Fase 5: Evaluación

En esta fase se evaluarán los resultados del modelo teniendo en cuenta el cumplimiento de los objetivos, a continuación, se describe cada tarea.

**a) Evaluación de los resultados**

Los resultados de este algoritmo backpropagation clasificaron cada atributo en base a valores establecidos al atributo objetivo que es PonderacionPromedio, los cuales obtuvieron un valor de 1 cuando eran clasificadas correctamente en cualquiera de estos casos como bueno, malo y excelente y 0 en caso contrario, es decir sino eran clasificadas correctamente, y aquellas que tienen valores reales, la que más se aproxime a 1, ver anexo 8.

**b) Evaluación de Modelos**

Para verificar la confiabilidad del modelo se utilizó accuracy, porque expresa el porcentaje que ha sido correctamente clasificado, error cuadrado medio raíz (RMSE): mide la cantidad de error al comparar un valor predicho y un valor observado, un valor de RMSE = 0 indica un ajuste perfecto (Ritter, Muñoz & Regalado, 2011), y Kappa para medir coincidencia de la predicción con la clase real. Según (Correa,

Bielza, Pamies-Texeira, & Alique, 2008) la escala de medición para la índice muestra la tabla 13. Además de eso se utilizó la matriz de confusión descrito en el capítulo IV para cada tabla.

*Tabla 13. Valor de coeficiente Kappa*

| Fuerza de concordancia | Coeficiente Kappa |
|---|---|
| Mala | 0.00 |
| Pobre | < 0.20 |
| Débil | 0.21 a 0.40 |
| Aceptable | 0.41 a 0.60 |
| Bueno | 0.61 a 0.80 |
| Excelente | 0.81 a 1.00 |

**c) Determinación de futuras fases**

Se presentará los resultados obtenidos de acuerdo con los objetivos de negocio y minería de datos.

### 3.1.6. Fase 6: Implementación

El objetivo en la última fase de la metodología CRISP-DM es el de implementar de modelo, para Estudiantes, Evaluación Docente y Publicaciones de Docentes, mismos que se mostrarán en el capítulo IV Resultados y Discusión.

## CAPÍTULO IV:

## 4. RESULTADOS Y DISCUSIÓN

En el desarrollo de esta investigación se utilizaron datos proporcionados a través de la Coordinación de Desarrollo de Sistemas Informáticos (CODESI) de Universidad Nacional de Chimborazo. La información fue proporcionada en un archivo de Excel mismas que se detallan a continuación.

**Rendimiento académico de los estudiantes**

La base de datos estudiante contenía un total de 16008 registros, comprendido entre septiembre 2012 - marzo 2013 y octubre 2018- marzo 2019, una vez realizada la calidad de datos se disminuyó a 15793 registros, correspondiendo al 98.66% del total de los datos, los campos utilizados se detallaron en la fase preparación de datos.

**Rendimiento académico de los estudiantes de la FI (917 registros)**

Después de varios entrenamientos se determinó que la configuración de la tabla 14 es la más adecuada, porque representa un 97,85% de instancias clasificadas correctamente y 02,16% clasificadas incorrectamente, el índice de kappa es alto con el 0.96 que significa que la coincidencia de la predicción con la clase real está en excelente ajuste, y el RSME de 0,13 cuanto más cercano sea el valor del error a cero, mayor será la precisión de la predicción.

Tabla 14. Parámetros óptimos para estudiantes de la FI.

| Entradas | Salidas | Parámetros de entrenamiento | | Resultados | |
|---|---|---|---|---|---|
| | | Descripción | Valor | Descripción | Valor |
| Estudiante ID, Estado Civil, Género, Actividad Deportiva, Actividad Cultural, Es Foráneo, Tiene Hermanos, Trabaja, Tiene Hijos Promedio, Ponderación Promedio | Excelente<br>Bueno<br>Regular | Training cycles<br>Learning rate<br>Momentum<br>Hidden Layer<br>Error Epsilon | 1000<br>0.2<br>0.9<br>7<br>1.0E-4 | Acuraccy<br>EC<br>Kappa<br>RMSE | 97,85%<br>2,16%<br>0,96<br>0,13 |

La tabla 15 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 3.917 estudiantes de la FI, 34 fueron mal clasificados.

*Tabla 15. Matriz de Confusión para Estudiantes de la FI*

| | true Regular | true Bueno | true Excelente |
|---|---|---|---|
| pred. Regular | 1126 | 1 | 0 |
| pred. Bueno | 10 | 2274 | 9 |
| pred. Excelente | 0 | 7 | 490 |

Las características de los estudiantes de la FCS según su promedio académico se describen en la Tabla 16. Dadas las condiciones indicadas en la tabla, el rendimiento académico de los estudiantes es excelente cuando los estudiantes son de género masculino, estado civil soltero, no sea foráneo, no debe trabajar, no debe tener hijos, debe tener hermanos, de practicar actividad deportiva y cultural.

*Tabla 16. Características - Estudiantes de la FI según su promedio*

| Promedio | Género | Estado Civil | Foráneo | Trabaja | Tiene hijos | Tiene hermanos | Actividad Deportiva | Actividad Cultural |
|---|---|---|---|---|---|---|---|---|
| Excelente | Masculino | Soltero | NO | NO | NO | SI | SI | SI |
| Bueno | Masculino | Soltero | NO | NO | NO | SI | SI | SI |
| Regular | Masculino | Soltero | NO | NO | NO | NO | SI | NO |

**Rendimiento académico de los estudiantes de la FCS (4901 registros).**

La tabla 17 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 4.901 estudiantes de la FCS, 139 fueron mal clasificados.

*Tabla 17. Matriz de Confusión de los estudiantes de la FCS*

| | true Regular | true Bueno | true Excelente |
|---|---|---|---|
| pred. Regular | 450 | 4 | 0 |
| pred. Bueno | 13 | 2499 | 78 |
| pred. Excelente | 0 | 44 | 1813 |

Las características de los estudiantes de la FCS según su promedio académico se describen en la Tabla 18. Dadas las condiciones indicadas en la tabla, el rendimiento académico de los estudiantes sería excelentes cuando: el género sea femenino, estado civil soltero, debe ser foráneo, no debe trabajar, no debe tener hijos, debe tener hermanos, de practicar actividad deportiva y cultural.

*Tabla 18. Características - Estudiantes de la FCS según su promedio*

| Promedio | Género | Estado Civil | Foráneo | Trabaja | Tiene hijos | Tiene hermanos | Actividad Deportiva | Actividad Cultural |
|---|---|---|---|---|---|---|---|---|
| Excelente | Femenino | Soltero | SI | NO | NO | SI | SI | SI |
| Bueno | Femenino | Soltero | SI | NO | NO | NO | SI | NO |
| Regular | Femenino | Soltero | SI | NO | NO | NO | SI | NO |

**Rendimiento académico de los estudiantes de la FCPYA (3698 registros).**

La tabla 19 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 3698 estudiantes de la FCPYA, 124 fueron mal clasificados.

*Tabla 19. Matriz de Confusión de los estudiantes de la FCPYA.*

| | true Regular | true Bueno | true Excelente |
|---|---|---|---|
| pred. Regular | 495 | 11 | 0 |
| pred. Bueno | 14 | 1574 | 42 |
| pred. Excelente | 0 | 57 | 1405 |

Las características de los estudiantes de la FCPYA según su promedio académico se describen en la Tabla 20. Dadas las condiciones indicadas en la tabla, el rendimiento académico de los estudiantes sería excelentes cuando: el género sea femenino, estado civil soltero, no debe ser foráneo, no debe trabajar, no debe tener hijos, debe tener hermanos, de practicar actividad deportiva y cultural.

*Tabla 20. Características - Estudiantes de la FCPYA según su promedio*

| Promedio | Género | Estado Civil | Foráneo | Trabaja | Tiene hijos | Tiene hermanos | Actividad Deportiva | Actividad Cultural |
|---|---|---|---|---|---|---|---|---|
| Excelente | Femenino | Soltero | NO | NO | NO | SI | SI | SI |
| Bueno | Masculino | Soltero | NO | NO | NO | NO | SI | SI |
| Regular | Masculino | Soltero | NO | NO | NO | NO | SI | NO |

**Rendimiento académico de los estudiantes de la FCEHYT (3277 registros).**

La tabla 21 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias

son las clasificadas incorrectamente. De 3277 estudiantes de la FCEHYT, 57 fueron mal clasificados.

*Tabla 21. Matriz de Confusión de los estudiantes de la FCEHYT*

| | true Regular | true Bueno | true Excelente |
|---|---|---|---|
| pred. Regular | 267 | 4 | 0 |
| pred. Bueno | 6 | 590 | 22 |
| pred. Excelente | 0 | 25 | 2363 |

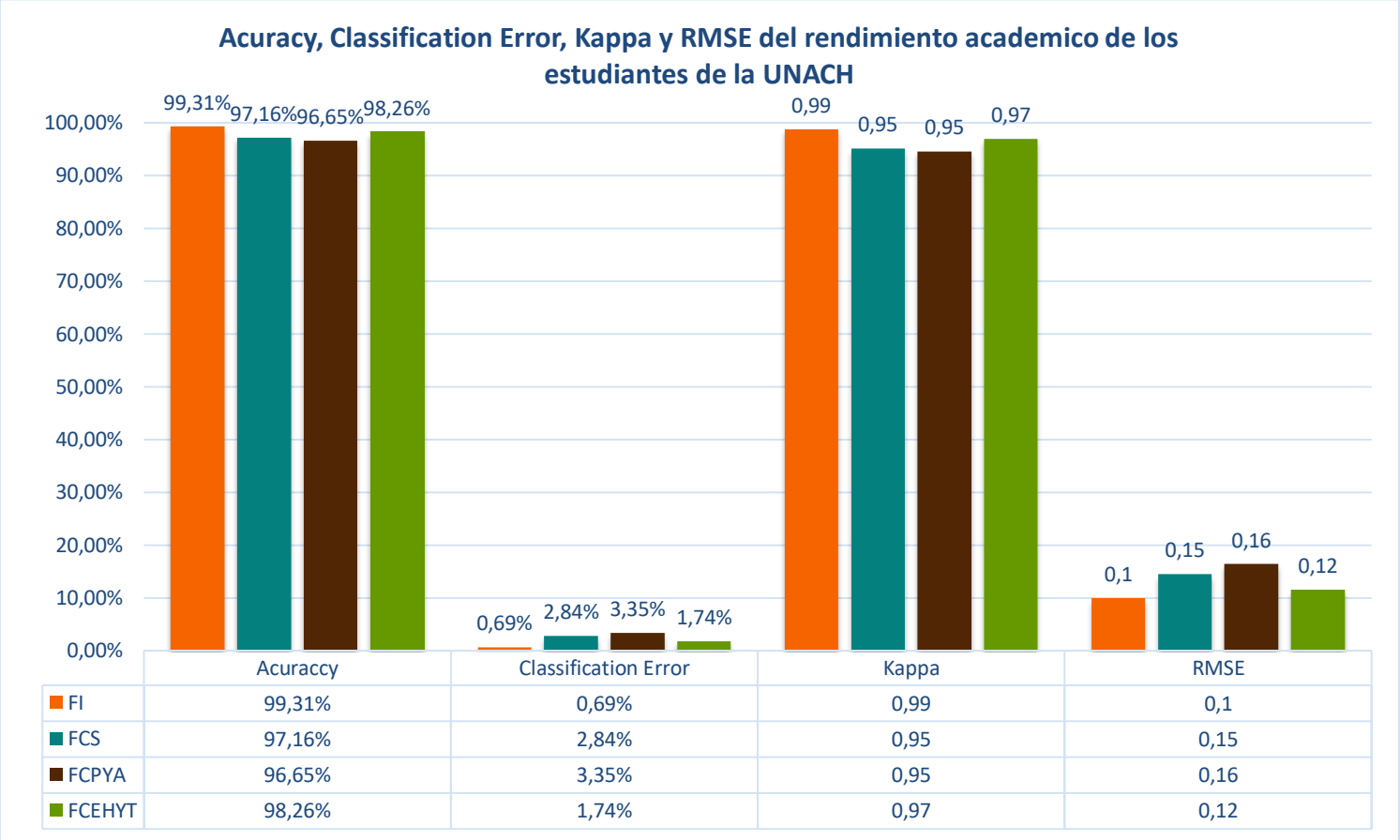
Las características de los estudiantes de la FCEHYT según su promedio académico se describen en la Tabla 22. Dadas las condiciones indicadas en la tabla, el rendimiento académico de los estudiantes sería excelentes cuando: el género sea femenino, estado civil soltero, no debe ser foráneo, no debe trabajar, no debe tener hijos, debe tener hermanos, de practicar actividad deportiva y cultural.

*Tabla 22. Características -Estudiantes de la FCEHYT según su promedio*

| Promedio | Género | Estado Civil | Foráneo | Trabaja | Tiene hijos | Tiene hermanos | Actividad Deportiva | Actividad Cultural |
|---|---|---|---|---|---|---|---|---|
| Excelente | Femenino | Soltero | NO | NO | NO | SI | SI | SI |
| Bueno | Masculino | Soltero | NO | NO | NO | NO | SI | SI |
| Regular | Masculino | Soltero | NO | NO | NO | NO | SI | NO |

Aplicando los parámetros de entrada, salida y entrenamiento de la tabla 14, los resultados de Acuracy, Classification Error, Kappa, RMSE, para el rendimiento académico de los estudiantes de las 4 facultades de la UNACH se visualiza en la ilustración 18. En donde la FI tiene la mejor confiabilidad del modelo con el 99,31% por ende, el error de clasificación es el menor en comparación con las demás facultades el 0,69%. El coeficiente Kappa de 0,99 es excelente porque está entre un rango de 0.81 a

1.00 según lo establecido por la tabla 13 y la cantidad del RMSE al comparar un valor predicho y un valor observado es de 0,1 aproximándose a 0 que indica un ajuste perfecto.




| | Acuraccy | Classification Error | Kappa | RMSE |
|---|---|---|---|---|
| FI | 99,31% | 0,69% | 0,99 | 0,1 |
| FCS | 97,16% | 2,84% | 0,95 | 0,15 |
| FCPYA | 96,65% | 3,35% | 0,95 | 0,16 |
| FCEHYT | 98,26% | 1,74% | 0,97 | 0,12 |

*Ilustración 18. Rendimiento Académico de los estudiantes de la UNACH*

La ilustración 19 muestra la probabilidad en obtener un promedio excelente, bueno y regular en el rendimiento académico de los estudiantes de las 4 facultades de la UNACH. La FCEHYT tiene mayor probabilidad en obtener un promedio excelente, con el 72,9% y la FI es la menor en obtener un promedio excelente con el 12.7%.

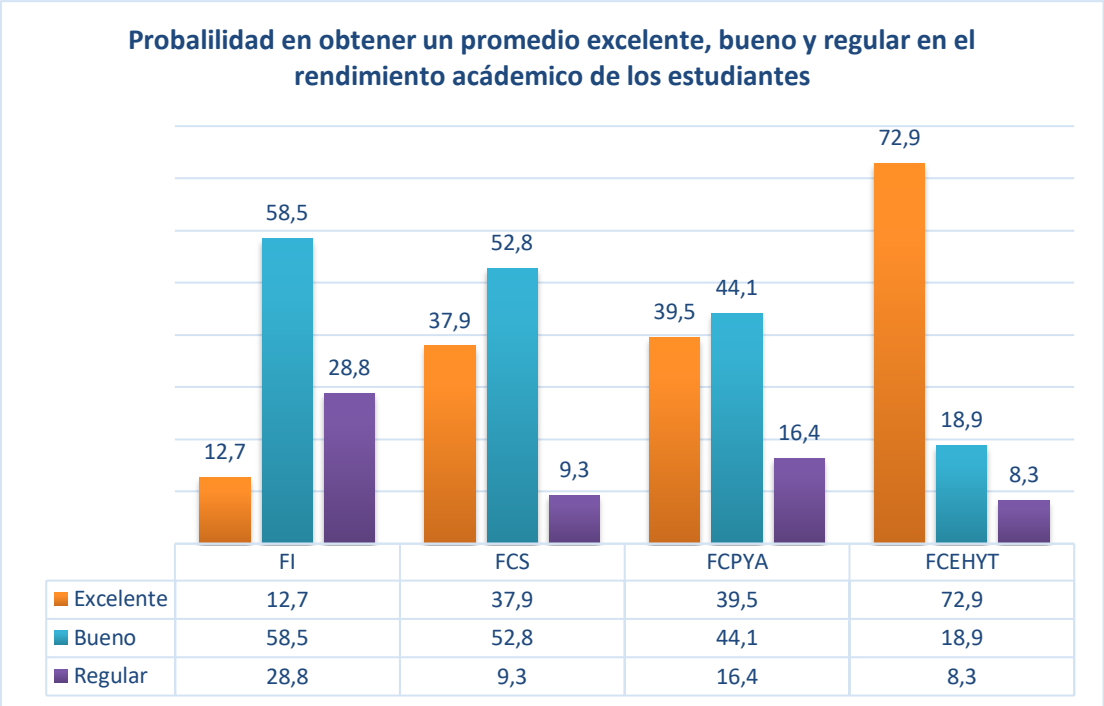



*Ilustración 19. Probabilidad de estudiantes de la UNACH en obtener un promedio*

**Proceso de Evaluación Integral en el desempeño de los docentes**

La base de datos docente contenía un total de 4097 registros, comprendido entre septiembre 2012 - marzo 2013 y octubre 2018- marzo 2019, estos datos estaban duplicados varias veces y al eliminar los registros duplicados se obtuvo un total de 826 registros, una vez realizada la calidad de datos se disminuyó a 419 registros, correspondiendo al 50.73% del total de los datos, los campos utilizados se detallaron en la fase preparación de datos.

**Evaluación Integral de los docentes de la FI (137 registros).**

La configuración de la tabla 23 es la más adecuada, porque representa un 93,83% de instancias clasificadas correctamente y 6,67% clasificadas incorrectamente, el índice de kappa es 0,69 que significa que la coincidencia de la predicción con la clase real está en ajuste aceptable, y el RSME de 0,21 cuanto más cercano sea el valor del error a cero, mayor será la precisión de la predicción.

*Tabla 23. Parámetros óptimos del modelo ANN para docentes de la FI.*

| Entradas | Salidas | Parámetros de entrenamiento | | Resultados | |
|---|---|---|---|---|---|
| | | Descripción | Valor | Descripción | Valor |
| Cedula, Estado Civil, Género, Nivel Instrucción, Tiene Hijos, Eventos Nacionales, Eventos Internacionales, Horas Actividad Académica, Horas Clase, Resultado Final Evaluación Docente | Excelente<br>Bueno<br>Insuficiente | Training cycles<br>Learning rate<br>Momentum<br>Hidden Layer<br>Error Epsilon | 1500<br>0.02<br>0.9<br>7<br>1.0E-4 | Acuraccy<br>EC<br>Kappa<br>RMSE | 93,83%<br>6,67%<br>0,69<br>0,21 |

La tabla 24 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 137 docentes de la FI, 8 fueron mal clasificados.

*Tabla 24. Matriz de Confusión de los docentes de la FI*

| | true Insuficiente | true Bueno | true Excelente |
|---|---|---|---|
| pred. Insuficiente | 0 | 1 | 0 |
| pred. Bueno | 2 | 7 | 0 |
| pred. Excelente | 0 | 5 | 122 |

Las características de los docentes de la FI según el resultado de evaluación integral se describen en la Tabla 25. Dadas las condiciones indicadas en la tabla, el resultado de evaluación al docente sería excelentes cuando: el género sea masculino, estado civil soltero, Nivel de Instrucción Maestría, debe tener Horas Clase, Horas de actividad académica, debe tener hijos, tiene eventos nacionales e internacionales.

*Tabla 25. Características - Docentes de la FI según el resultado de evaluación*

| Evaluación | Género | Estado Civil | Nivel de Instrucción | Horas Clase | Horas Act. Academica | Tiene hijos | Eventos internacionales | Eventos Nacionales |
|---|---|---|---|---|---|---|---|---|
| Excelente | Masculino | Soltero | Maestría | SI | SI | SI | NO | SI |
| Bueno | Masculino | Casado | Maestría | SI | SI | SI | NO | SI |
| Insuficiente | Masculino | Casado | Maestría | SI | SI | NO | NO | NO |

La tabla 26 determina que la evaluación a los docentes de la FI tiene alta probabilidad de obtener un promedio excelente, con el 92,70%.

*Tabla 26. Probabilidad del resultado de evaluación de los docentes de la FI*

| Promedio | Recuento Absoluto | Probabilidad |
|---|---|---|
| Excelente | 127 | 92,70 |
| Bueno | 9 | 6,60 |
| Insuficiente | 1 | 0.70 |

**Evaluación Integral de los docentes de la FCS (88 registros).**

La tabla 27 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 88 docentes de la FCS, 6 fueron mal clasificados.

*Tabla 27. Matriz de Confusión de los docentes de la FCS*

| | true Insuficiente | true Bueno | true Excelente |
|---|---|---|---|
| pred. Insuficiente | 0 | 0 | 0 |
| pred. Bueno | 4 | 14 | 0 |
| pred. Excelente | 0 | 2 | 68 |

Las características de los docentes de la FCS según el resultado de evaluación integral se describen en la Tabla 28. Dadas las condiciones indicadas en la tabla, el resultado de evaluación al docente sería excelentes cuando: el género sea masculino, estado civil soltero, Nivel de Instrucción Maestría, debe tener Horas Clase, Horas de actividad académica, debe tener hijos, debe tener eventos nacionales y no debe tener eventos internacionales.

*Tabla 28. Características - Docentes de la FCS según el resultado de evaluación*

| Evaluación | Género | Estado Civil | Nivel de Instrucción | Horas Clase | Horas Act. Academica | Tiene hijos | Eventos internacionales | Eventos Nacionales |
|---|---|---|---|---|---|---|---|---|
| Excelente | Masculino | Soltero | Maestría | SI | SI | SI | NO | SI |
| Bueno | Masculino | Casado | Maestría | SI | SI | SI | NO | SI |

**Evaluación Integral de los docentes de la FCPYA (92 registros).**

La tabla 29 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias

son las clasificadas incorrectamente. De 92 docentes de la FCPYA, 4 fueron mal clasificados.

*Tabla 29. Matriz de Confusión de los docentes de la FCPYA*

| | true Insuficiente | true Bueno | true Excelente |
|---|---|---|---|
| pred. Insuficiente | 0 | 0 | 0 |
| pred. Bueno | 1 | 8 | 1 |
| pred. Excelente | 0 | 2 | 80 |

Las características de los docentes de la FCPYA según el resultado de evaluación integral se describen en la Tabla 30. Dadas las condiciones indicadas en la tabla, el resultado de evaluación al docente sería excelentes cuando: el género sea masculino, estado civil soltero, Nivel de Instrucción Maestría, debe tener Horas Clase, Horas de actividad académica, debe tener hijos, debe tener eventos nacionales y no debe tener eventos internacionales.

*Tabla 30. Características - Docentes de FCPYA según el resultado de evaluación*

| Evaluación | Género | Estado Civil | Nivel de Instrucción | Horas Clase | Horas Act. Academica | Tiene hijos | Eventos internacionales | Eventos Nacionales |
|---|---|---|---|---|---|---|---|---|
| Excelente | Masculino | Soltero | Maestría | SI | SI | SI | NO | SI |
| Bueno | Masculino | Casado | Maestría | SI | SI | SI | NO | SI |

**Docentes Facultad Ciencias de la Educación FCEHYT (102 registros).**

La tabla 31 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 102 docentes de la FCEBHYT, 4 fueron mal clasificados.

*Tabla 31. Matriz de Confusión de los docentes de la FCEHYT.*

| | true Insuficiente | true Bueno | true Excelente |
|---|---|---|---|
| pred. Insuficiente | 4 | 0 | 0 |
| pred. Bueno | 0 | 7 | 3 |
| pred. Excelente | 0 | 5 | 83 |

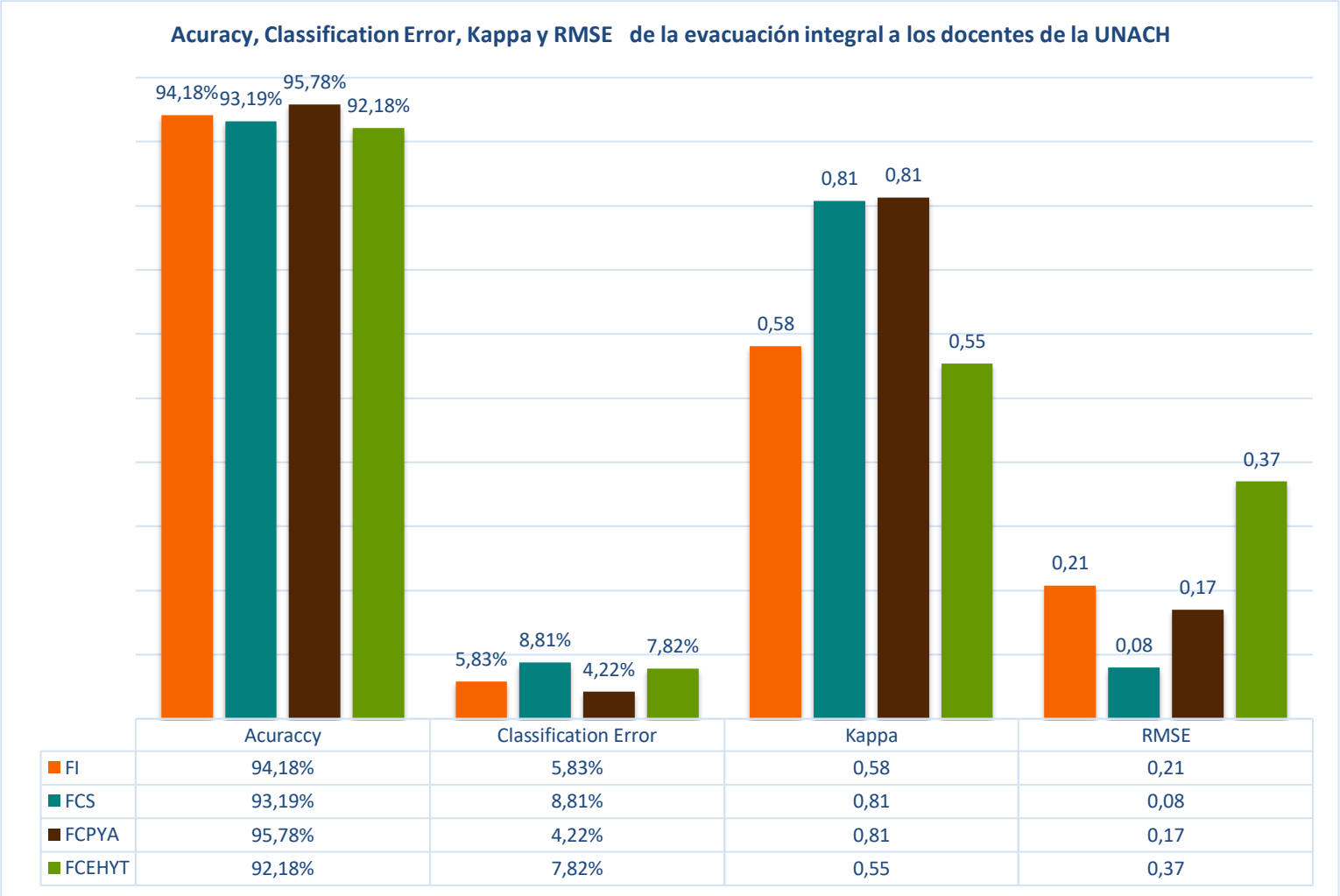
Las características de los docentes de la FCEHYT según el resultado de evaluación integral se describen en la Tabla 32. El resultado de evaluación al docente sería excelente

cuando: el género sea masculino, estado civil soltero, Nivel de Instrucción Maestría, debe tener Horas Clase, Horas de actividad académica, debe tener hijos, debe tener eventos nacionales e internacionales.

*Tabla 32. Características - Docentes de FCEHYT según el resultado de evaluación*

| Evaluación | Género | Estado Civil | Nivel de Instrucción | Horas Clase | Horas Act. Academica | Tiene hijos | Eventos internacionales | Eventos Nacionales |
|---|---|---|---|---|---|---|---|---|
| Excelente | Masculino | Soltero | Maestría | SI | SI | SI | SI | SI |
| Bueno | Masculino | Soltero | Maestría | SI | SI | SI | NO | SI |
| Insuficiente | Masculino | Solteto | Maestría | SI | SI | NO | NO | SI |

Aplicando los parámetros de entrada, salida y entrenamiento de la tabla 23, los resultados de Acuracy, Classification Error, Kappa, RMSE, para la evaluación integral de los docentes de las 4 facultades de la UNACH se visualiza en la ilustración 20. En donde la FCPYA tiene la mejor confiabilidad del modelo con el 95,78% por ende, el error de clasificación es el menor en comparación con las demás facultades el 4,22%. El coeficiente Kappa de 0,81 es excelente porque está entre un rango de 0.81 a 1.00 según lo establecido por la tabla 13 y la cantidad del RMSE al comparar un valor predicho y un valor observado es de 0,17 aproximándose a 0 que indica un ajuste perfecto.




| | Acuraccy | Classification Error | Kappa | RMSE |
|---|---|---|---|---|
| FI | 94,18% | 5,83% | 0,58 | 0,21 |
| FCS | 93,19% | 8,81% | 0,81 | 0,08 |
| FCPYA | 95,78% | 4,22% | 0,81 | 0,17 |
| FCEHYT | 92,18% | 7,82% | 0,55 | 0,37 |

*Ilustración 20. Evaluación integral de los docentes de la UNACH*

La ilustración 21 visualiza la probabilidad en obtener una calificación excelente, bueno e insuficiente en la evaluación integral de los docentes de las 4 facultades de la UNACH. La FI tiene mayor probabilidad en obtener una calificación excelente, con el 92,7% y la FCS es la menor en obtener una calificación excelente con el 79,5%.

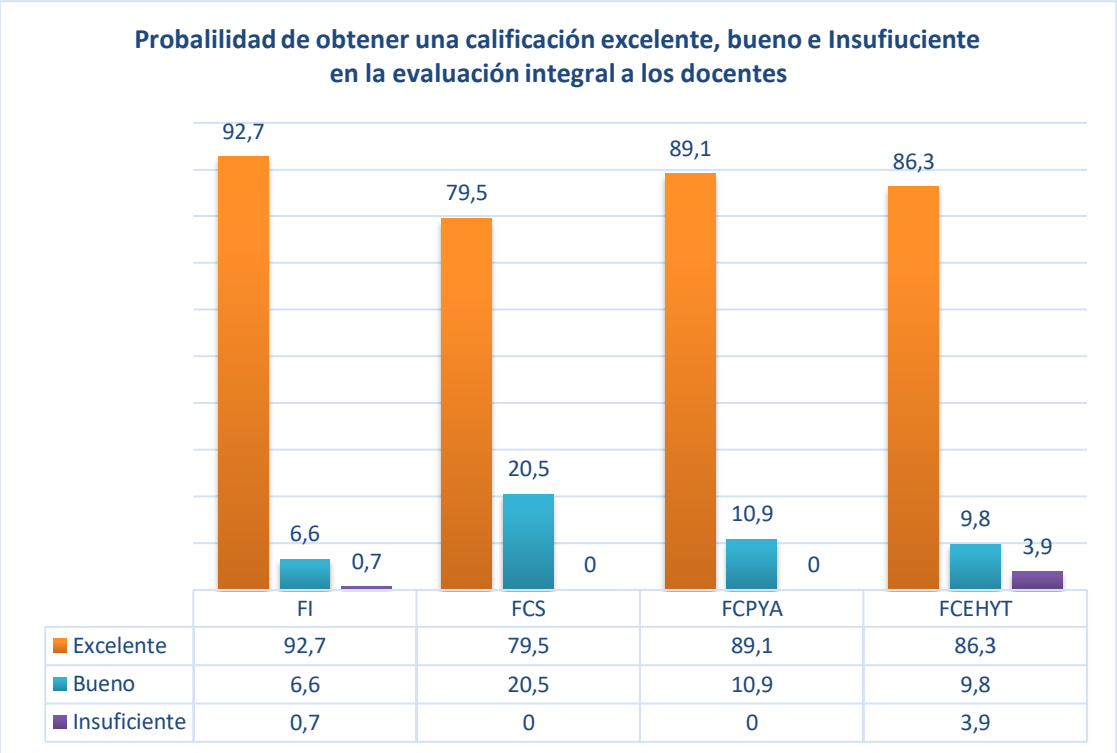



*Ilustración 21. Probabilidad de docentes de la UNACH en obtener una calificación*

**Docentes que publican y los que no publican**

La base de datos investigación contenía un total de 2051 registros, comprendido desde diciembre de 2013 hasta abril de 2018, una vez realizada la calidad de datos se disminuyó a 1150 registros, correspondiendo al 56.07% del total de los datos, los campos utilizados se detallaron en la fase preparación de datos.

**Publicaciones de docentes de la FI (295 registros)**

Después de varios entrenamientos se determinó que la configuración de la tabla 32 es la más adecuada, porque representa un 96,77% de instancias clasificadas correctamente y 3,23% clasificadas incorrectamente, el índice de kappa es alto con el 0,93 que significa que la coincidencia de la predicción con la clase real está en buen ajuste, y el RSME de 0,13 cuanto más cercano sea el valor del error a cero, mayor será la precisión de la predicción.

*Tabla 32. Parámetros óptimos del modelo ANN para docentes la FI.*

| Entradas | Salidas | Parámetros de entrenamiento | | Resultados | |
|---|---|---|---|---|---|
| | | Descripción | Valor | Descripción | Valor |
| Cedula, Estado Civil Género, Nivel Instrucción, Eventos Nacionales, Horas Actividad Académica, Horas Clase, Eventos Internacionales, Eventos Nacionales, Publicación ProduccionCientifica2, Publicaciones RegionalRevista3, Publicaciones Libro, Publicaciones Capítulo Libro, Publicaciones Ponencia, Tiene Publicaciones. | SI PUBLICA<br>NO PUBLICA | Training cycles<br>Learning rate<br>Momentum<br>Hidden Layer<br>Error Epsilon | 1000<br>0.02<br>0.9<br>9<br>1.0E-4 | Acuraccy<br>EC<br>Kappa<br>RMSE | 96.77%<br>3,23%<br>0.93<br>0,13 |

La tabla 33 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 295 publicaciones de la FI, 6 fueron mal clasificados.

*Tabla 33. Matriz de Confusión de publicaciones de los docentes de la FI.*

| | true NO | true SI |
|---|---|---|
| pred. NO | 197 | 5 |
| pred. SI | 1 | 92 |

Las características de los docentes de la FI que realizan publicaciones se describen a en la tabla 34. Los docentes que realizan publicaciones son: género Masculino, estado civil casado, Nivel de instrucción Maestría, debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica, No publica en capítulo de libro, no publica en libro, publica en ponencia, publica en revista regional y no publica en producción científica.

*Tabla 34. Características - Publicaciones de docentes de la FI.*

| Atributos | Publican | |
|---|---|---|
| | SI | NO |
| Género | Masculino | Masculino |
| Estado Civil | Casado | Casado |
| Nivel Instrucción | Maestría | Maestría |
| Tiene Hijos | Si | No |
| Eventos Nacionales | Si | Si |
| Eventos Internacionales | Si | No |
| Horas Clase | Si | Si |
| Horas Actividad Académica | Si | Si |
| Capítulo de Libro | No | No |
| Libro | No | No |
| Ponencia | Si | No |
| Revista Regional | Si | No |
| Producción Científica | No | No |

**Publicaciones de docentes de la FCS (420 registros).**

La tabla 35 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 420 publicaciones de la FCS, 7 fueron mal clasificados.

*Tabla 35. Matriz de Confusión de publicaciones de los docentes de la FCS.*

| | true NO | true SI |
|---|---|---|
| pred. NO | 313 | 6 |
| pred. SI | 1 | 100 |

Las características de los docentes de la FCS que realizan publicaciones se describen a en la tabla 36. Los docentes que realizan publicaciones son: género Femenino, estado civil casado, Nivel de instrucción Maestría, debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica, no publica en capítulo de libro, no publica en libro, publica en ponencia, publica en revista regional y no publica en producción científica

*Tabla 36. Características - Publicaciones de docentes de la FCS.*

| Atributos | Publican | |
|---|---|---|
| | SI | NO |
| Género | Femenino | Femenino |
| Estado Civil | Casado | Casado |
| Nivel Instrucción | Maestría | Maestría |
| Tiene Hijos | Si | No |
| Eventos Nacionales | Si | Si |
| Eventos Internacionales | Si | No |
| Horas Clase | Si | Si |
| Horas Actividad Académica | Si | Si |
| Capítulo de Libro | No | No |
| Libro | No | No |
| Ponencia | Si | No |
| Revista Regional | Si | No |
| Producción Científica | No | No |

**Publicaciones de docentes de la FPYA (191 registros)**

La tabla 37 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 191 publicaciones de la FI políticas, 9 fueron mal clasificados.

*Tabla 37. Matriz de Confusión de publicaciones de los docentes de la FCPYA*

| | true NO | true SI |
|---|---|---|
| pred. NO | 126 | 6 |
| pred. SI | 3 | 56 |

Las características de los docentes de la FCPYA que realizan publicaciones se describen a en la tabla 38. Los docentes que realizan publicaciones son: género Masculino, estado civil casado, Nivel de instrucción Maestría, debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica, no publica en capítulo de libro, no publica en libro, publica en ponencia, publica en revista regional y no publica en producción científica.

*Tabla 38. Características - Publicaciones de docentes de la FCPYA.*

| Atributos | Publican | |
|---|---|---|
| | SI | NO |
| Género | Masculino | Femenino |
| Estado Civil | Casado | Casado |
| Nivel Instrucción | Maestría | Maestría |
| Tiene Hijos | Si | No |
| Eventos Nacionales | Si | Si |
| Eventos Internacionales | Si | Si |
| Horas Clase | Si | No |
| Horas Actividad Académica | Si | Si |
| Capítulo de Libro | No | No |
| Libro | No | No |
| Ponencia | Si | No |
| Revista Regional | Si | No |
| Producción Científica | No | No |

**Publicaciones de docentes de la FCEHYT (244 registros).**

La tabla 39 muestra la matriz de confusión que permite observar las instancias clasificadas correctamente, las cuales se encuentran en la diagonal principal y las demás instancias son las clasificadas incorrectamente. De 244 docentes de la FCEHYT, 11 fueron mal clasificados.

*Tabla 39. Matriz de Confusión de publicaciones de los docentes de la FCEHYT*

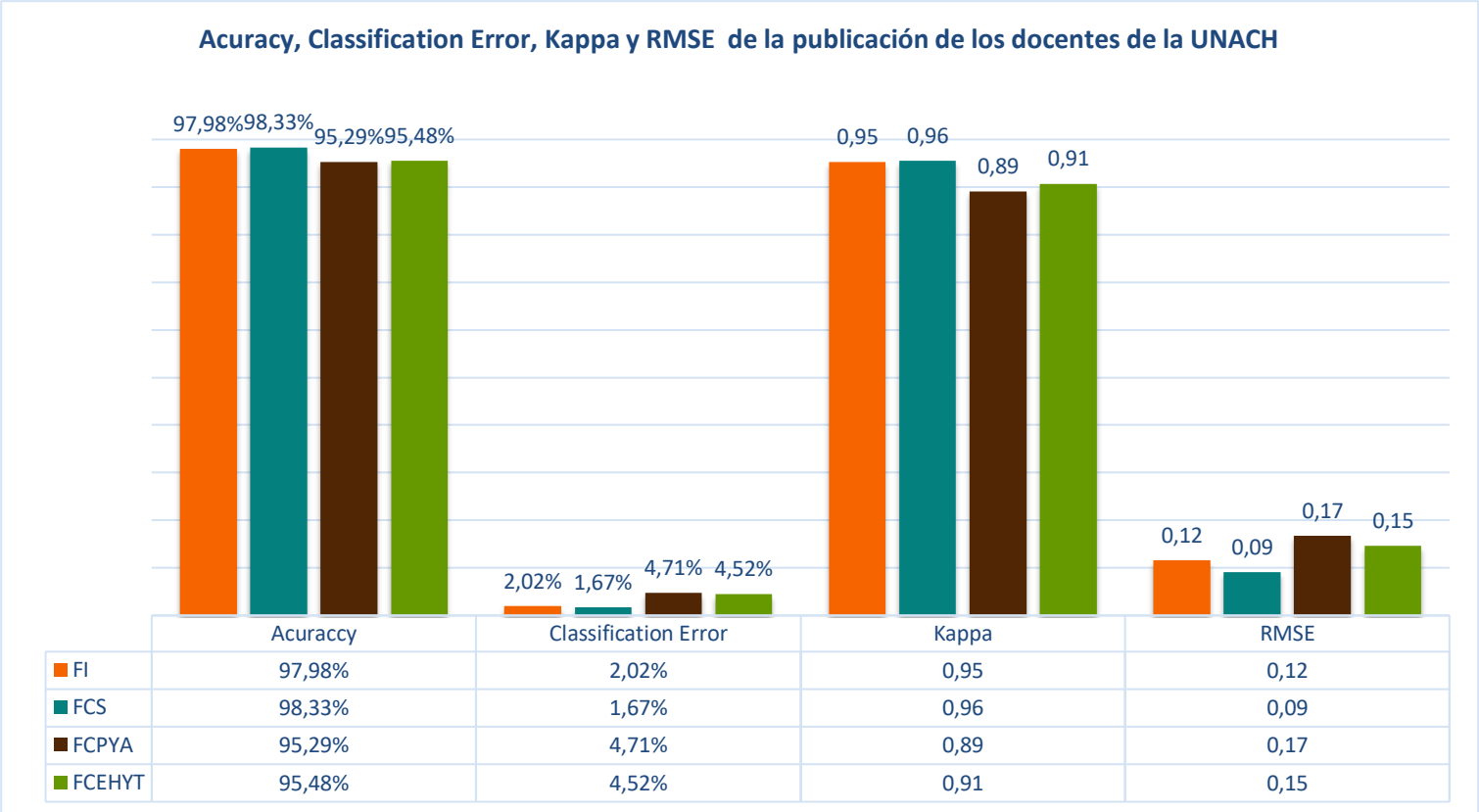
| | true NO | true SI |
|---|---|---|
| pred. NO | 131 | 7 |
| pred. SI | 4 | 102 |

Las características de los docentes de la FCPYA que realizan publicaciones se describen a en la tabla 40. Los docentes que realizan publicaciones son: género Masculino, estado civil casado, Nivel de instrucción Maestría, debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica, publica en capítulo de libro, no publica en libro, publica en ponencia, publica en revista regional y no publica en producción científica.

*Tabla 40. Características - Publicaciones de docentes de la FCPYA.*

| Atributos | Publican | |
|---|---|---|
| | SI | NO |
| Género | Masculino | Femenino |
| Estado Civil | Casado | Casado |
| Nivel Instrucción | Maestría | Maestría |
| Tiene Hijos | Si | No |
| Eventos Nacionales | Si | Si |
| Eventos Internacionales | Si | No |
| Horas Clase | Si | No |
| Horas Actividad Académica | Si | Si |
| Capítulo de Libro | Si | No |
| Libro | No | No |
| Ponencia | Si | No |
| Revista Regional | Si | No |
| Producción Científica | No | No |

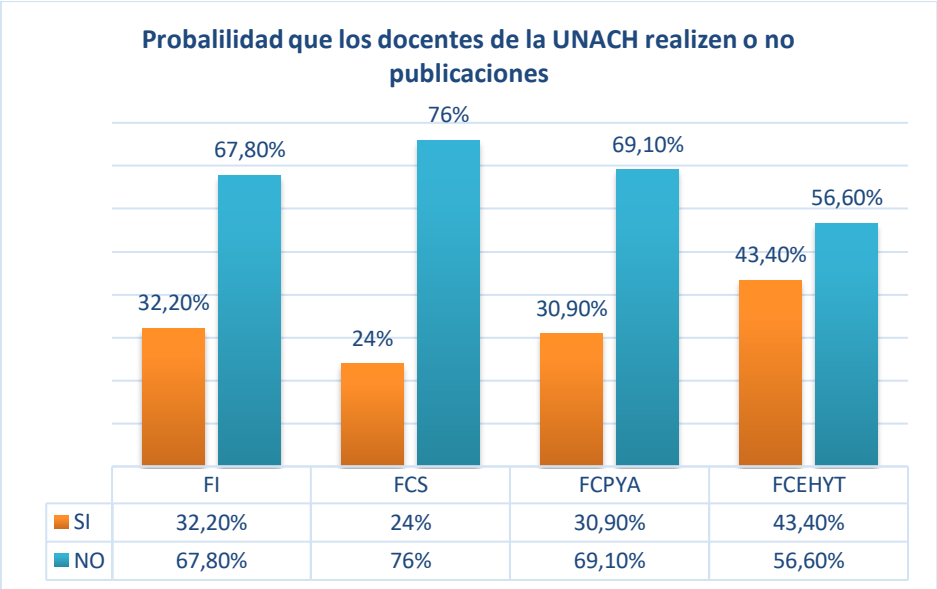
Aplicando los parámetros de entrada, salida y entrenamiento de la tabla 32, los resultados de Acuracy, Classification Error, Kappa, RMSE, para la publicación de docentes de las 4 facultades de la UNACH se visualiza en la ilustración 22. En donde la FCS tiene la mejor confiabilidad del modelo con el 98,33% por ende, el error de clasificación es el menor en comparación con las demás facultades el 1,67%. El coeficiente Kappa de 0,96 es excelente porque está entre un rango de 0.81 a 1.00 según lo establecido por la tabla 13 y la cantidad del RMSE al comparar un valor predicho y un valor observado es de 0,09 aproximándose a 0 que indica un ajuste perfecto.




| | Acuraccy | Classification Error | Kappa | RMSE |
|---|---|---|---|---|
| FI | 97,98% | 2,02% | 0,95 | 0,12 |
| FCS | 98,33% | 1,67% | 0,96 | 0,09 |
| FCPYA | 95,29% | 4,71% | 0,89 | 0,17 |
| FCEHYT | 95,48% | 4,52% | 0,91 | 0,15 |

*Ilustración 22. Publicación de los docentes de la UNACH*

La ilustración 23 muestra la probabilidad para saber si un docente de la UNACH realiza o no publicaciones. La FCEHYT tiene mayor probabilidad para realizar publicaciones con el 43,4% y la FCS es la menor en realizar publicaciones con el 76%.




| | FI | FCS | FCPYA | FCEHYT |
|---|---|---|---|---|
| SI | 32,20% | 24% | 30,90% | 43,40% |
| NO | 67,80% | 76% | 69,10% | 56,60% |

*Ilustración 23. Probabilidad de los docentes de la UNACH en realizar publicaciones*

**Tipo de publicación.**

**Producción alto impacto, revistas regionales y capitulo de libro de la FI.**

De 1150 docentes que solo 374 publican, los campos utilizados son Estado Civil, género. Nivel Instrucción, Tiene Hijos, Tiene Eventos Nacionales, Tienen Eventos Internacionales, Horas Actividad Académica.

Dadas las condiciones indicadas en la tabla 41, los docentes que realizan publicaciones en alto impacto son: género Femenino, estado civil casado, Nivel de instrucción PhD, no debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica.

*Tabla 41. Características según el tipo de publicaciones de los docentes de la FI.*

| Atributos | Producción | | |
|---|---|---|---|
| | Alto Impacto | Revista Regional | Capítulo de Libro |
| Género | Femenino | Masculino | Femenino |
| Estado Civil | Casado | Casado | Solteros |
| Nivel Instrucción | PhD | Maestría | Maestría |
| Tiene Hijos | No | Si | No |
| Eventos Nacionales | Si | Si | Si |
| Eventos Internacionales | Si | No | Si |
| Horas Clase | Si | Si | Si |
| Horas Actividad Académica | Si | Si | Si |

**Producción alto impacto, revistas regionales y capitulo de libro FCS.**

Las características de los docentes de la FCS que realizan publicaciones en alto impacto se describen a en la tabla 42. Dadas las conficiones indicadas en la tabla, los docentes que realizan publicaciones de alto impacto son: género Femenino, estado civil casado, Nivel de instrucción Maestría, si debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica.

*Tabla 42. Características según el tipo de publicaciones de los docentes de la FCS.*

| Atributos | Producción | | |
|---|---|---|---|
| | Alto Impacto | Revista Regional | Capítulo de Libro |
| Género | Femenino | Masculino | Femenino |
| Estado Civil | Casado | Casado | Casados |
| Nivel Instrucción | Maestría | Maestría | PhD |
| Tiene Hijos | Si | Si | No |
| Eventos Nacionales | Si | Si | Si |
| Eventos Internacionales | Si | Si | Si |
| Horas Clase | Si | Si | Si |
| Horas Actividad Académica | Si | Si | Si |

**Producción alto impacto, revistas regionales y capitulo de libro FCPYA.**

Las características de los docentes de la FCPYA que realizan publicaciones en alto impacto se describen a en la tabla 43. Dadas las conficiones indicadas en la tabla, los docentes que realizan publicaciones son: género Femenino, estado civil casado, Nivel de instrucción Maestría, si debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica.

*Tabla 43. Características según el tipo de publicaciones de docentes de la FCPYA*

| Atributos | Producción | | |
|---|---|---|---|
| | Alto Impacto | Revista Regional | Capítulo de Libro |
| Género | Femenino | Masculino | Masculino |
| Estado Civil | Casado | Casado | Casados |
| Nivel Instrucción | Maestría | Maestría | Maestría |
| Tiene Hijos | Si | Si | Si |
| Eventos Nacionales | Si | Si | Si |
| Eventos Internacionales | Si | No | No |
| Horas Clase | Si | Si | Si |
| Horas Actividad Académica | Si | Si | Si |

**Producción alto impacto, revistas regionales y capitulo de libro de PCEHYT.**

Las características de los docentes de la PCEHYT que realizan publicaciones en alto impacto se describen a en la tabla 44. Dadas las conficiones indicadas en la tabla, los

docentes que realizan publicaciones son: género Femenino, estado civil casado, Nivel de instrucción Maestría, si debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica.

*Tabla 44. Características según el tipo de publicaciones de docentes de la PCEHYT*

| Atributos | Producción | | |
|---|---|---|---|
| | Alto Impacto | Revista Regional | Capítulo de Libro |
| Género | Femenino | Femenino | Masculino |
| Estado Civil | Casado | Casado | Casados |
| Nivel Instrucción | Maestría | Maestría | Maestría |
| Tiene Hijos | Si | Si | Si |
| Eventos Nacionales | Si | Si | Si |
| Eventos Internacionales | Si | Si | Si |
| Horas Clase | Si | Si | Si |
| Horas Actividad Académica | Si | Si | Si |

# CAPÍTULO IV:

# 5. CONCLUSIONES Y RECOMEDACIONES

## CONCLUSIONES

Al finalizar el trabajo de titulación se obtuvieron las siguientes conclusiones:

- Una vez realizada la calidad de datos en la herramienta *Talend Data Quality* donde se identificó valores válidos, inválidos y nulos, se usó la herramienta *Rapid Miner*, Vista *AutoModel* y se identificó parámetros representativos para cada uno de los objetivos de minería de datos. La reducción del número de atributos en el conjunto de datos condujo a un modelo más simplificado ayudó a sintetizar una explicación más efectiva del modelo. Por tal razón en la entrada del modelo de aprendizaje automático solo se incluyó los atributos seleccionados por defecto es decir los de color amarillo y verde, porque los rojos se anularon. También para los pesos de las variables de

entrada se tomó en cuanta el operador estadístico *Weight by Chi Squared* de *RapidMiner*, el cual coincidía con los parámetros de *AultoModel* dados por defecto.

- La aplicación del algoritmo *Backpropagation* se llevó acabo sobre los datos de estudiantes, docentes y publicaciones y tipos de producción científica de las 4 facultades de la UNACH, para cada estudio se verificó la confiabilidad del modelo se utilizó *accuracy*, porque expresa el porcentaje que ha sido correctamente clasificado, *kappa* que mide la coincidencia de la predicción con la clase real en la interpretación de la estadística, se detallaron las principales reglas de clasificación, se entrenó la red neuronal con topología multicapas, y distintos parámetros de aprendizaje como capa oculta, ciclos de entrenamiento, tasa de aprendizaje, *momentum*, decaimiento y error épsilon de 1.0E-4, hasta alcanzar un mejor rendimiento en el modelo.

  La confiabilidad que tiene el modelo de red neuronal usando *RapidMiner* e implementado la metodología CRISP-DM para el rendimiento académico de los estudiantes de la FI fue 99,31%, estudiantes de la FCS 97,16%, estudiantes FCPYA 96,65%, estudiantes FCEHYT 98,26%. Y se llegó a la conclusión que los estudiantes de la FCS poseen una mayor probabilidad de obtener un promedio Excelente con el 72,90%, deben poseer las siguientes características; género sea femenino, estado civil soltero, debe ser foráneo, no debe trabajar, no debe tener hijos, debe tener hermanos, de practicar actividad deportiva y cultural.

  Para evaluación integral de los docentes de la FI tiene una confiabilidad de modelo de 94,18%, para docentes de la FCS 93,19%, para docentes de la FCPYA 95,78% y para docentes de la FCEHYT el 92,18%. Los docentes de la FI tienen una mayor posibilidad de obtener una calificación excelente en evaluación integral del docente con el 92,70%, deben poseer las siguientes características; género masculino, estado

civil soltero, Nivel de Instrucción Maestría, debe tener Horas Clase, Horas de actividad académica, debe tener hijos, tiene eventos nacionales e internacionales.

Para Publicaciones de los docentes de la FI 97,98%, FCS 98,33%, FCPYA 95,29%, FCEHYT 95,48%. Los docentes de la FCS con el 43,40% son los que más probabilidad presentan de realizar publicaciones. Poseen las siguientes características género femenino, estado civil casado, Nivel de instrucción Maestría, debe tener hijos, debe tener eventos nacionales e internaciones, Horas de Clase, Horas de Actividad académica, no publica en capítulo de libro, no publica en libro, publica en ponencia, publica en revista regional y no publica en producción científica

Los docentes que publican en la FI son hombres, su nivel de instrucción es Maestría, sin embargo, los que publican en alto impacto son mujeres con nivel de instrucción PhD. Los docentes que publican en la FCS son mujeres, su nivel de instrucción es Maestría, sin embargo, los que publican capítulo de libro son mujeres con nivel de instrucción PhD. Los docentes que publican en la FCPYA son hombres, su nivel de instrucción es Maestría, sin embargo, los que publican en alto impacto son mujeres con nivel de instrucción Maestría. Los docentes que publican en la FCEHYT, su nivel de instrucción es Maestría, sin embargo, los que publican en alto impacto y en revista regional son mujeres con nivel de instrucción Maestría.

Para determinar la confiabilidad del modelo se realizó el mismo proceso de *RapidMiner* en *Rstudio*, en *RapidMiner* se realizó dos análisis para cada una de las facultades de las tablas estudiante, docente e investigación una con Split Validation y otra con Validación Cruzada (*Coss Validation*), dando así la mejor confiabilidad la Validación cruzada. Es por eso que en *Rapid Miner* se usó la validación cruzada. De acuerdo con los resultados explicados en este proyecto de investigación se realizó un

promedio para cada una de las tablas, Se concluye que la confiabilidad del modelo para el rendimiento académico de los estudiantes es de 97,80% para la evaluación integral de los docentes es 93,83% y para publicación de docentes el 96,80% los cuales son rangos de precisión muy buenos. Y la confiabilidad del Modelo de todas las facultades de la UNACH es de 94,72%.

## RECOMENDACIONES

Luego de finalizar el Trabajo de Titulación son importantes las siguientes recomendaciones:

- La preparación del conjunto de datos para adaptarse a una tarea de minería de datos es la parte más lenta del proceso. Muy raramente los datos están disponibles en la forma requerida por los algoritmos de minería de datos. Se recomienda establecer principios de diseño para procesos ETL en la base de datos SICOA y en la de Publicaciones.
- Los valores faltantes se pueden reemplazar con valores promedio, corregirlo a través de la herramienta Talend Data Quality o cualquier valor predeterminado para minimizar el error.
- Una ANN no maneja datos de entrada numéricos. Si los datos tienen valores nominales, deben convertirse a valores binarios o reales, se recomienda realizar mediante el operador Nominal to Numerical de *Rapidminer.*
- La Herramienta *RapidMiner* fue fundamental en el desarrollo de este proyecto, tiene operadores que ayudan a facilitar el desarrollo de los procesos para crear modelos aplicables para el análisis.

- Construir un buen modelo ANN con parámetros optimizados lleva tiempo. Depende de la cantidad de registros de entrenamiento e iteraciones. No hay líneas de guía consistentes sobre el número de capas y nodos ocultos dentro de cada capa oculta. Por lo tanto, necesitaríamos probar muchos parámetros para optimizar la selección de parámetros. Sin embargo, una vez que se construye un modelo, es fácil de implementar para los siguientes ejemplos y su clasificación será bastante rápido.
- Para tener una validación y poderlo comparar con otros algoritmos se recomienda usar todo el proceso de AutoModel y elegir los algoritmos con los que desea comparar.
- Extender el presente estudio, con la finalidad de contribuir de manera sostenible la toma de decisiones en las actividades académicas y de investigación de la UNACH.
- Se recomienda profundizar la problemática en investigaciones futuras aplicando a otras IES de forma que la investigación sirva como base para los estudios.

## REFERENCIAS BIBLIOGRÁFICAS

Acevedo, G. L., Caicedo, E. F., & Loiza, H. C. (2007). Selección de personal mediantes redes. *Revista de Matemática. Teoría y Aplicaciones, 14*(1).

Aguilar, J., & Estrada, C. (2012). Minería de Datos (Data Mining).

Aluja, T. (2001). LA MINERÍA DE DATOS, ENTRE LA ESTADÍSTICA Y LA INTELIGENCIA ARTIFICIAL. *Qüestiió: quaderns d'estadística i investigació operativa*, 479-498.

Amaury, C., Gabriel, V., & Pardo García, A. (2013). Differentiations of objects in diffuse databases. *Revista colombiana de tecnologías de Avanzada,, 2*(22), 131-137.

Benalcázar Tamayo, J. B. (2017). Análisis comparativo de metodologías de minería de datos y su aplicabilidad a la industria de servicios. *Master's thesis.* Universidad de las Américas, Quito.

Bermúdez, J. A., & Acevedo, R. Á. (2010). Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies. *Doctoral dissertation.* Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación. Ingeniería de Sistemas y Computación, Pereira.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley & Sons.

Betancur Betancur, D., Vélez Gómez, M., & Peña Palacio, A. (2013). AUTOMATIC TRANSLATION OF THE DACTILOLOGIC LANGUAGE OF HEARING IMPAIRED BY ADAPTIVE SYSTEMS. *Revista Ingeniería Biomédica, 7*(13), 18-30.

Cataldi, Z., Salgueiro, F. A., & Lage, F. J. (2007). *Fundamentos para el submódulo evaluador en sistemas tutores inteligentes: Diagnóstico, predicción y autoevaluación.* In XIII Congreso Argentino de Ciencias de la Computación.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., . . . others. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc}, 16*.

Colina, A. M. (2017). *RETOS Y PERSPECTIVAS DE LAS TECNOLOGÍAS DE INFORMACIÓN.* Samborondón-Ecuador: Departamento de Publicaciones Universidad ECOTEC.

Correa, M., Bielza, C., Pamies-Texeira, J., & Alique, J. R. (2008). Redes Bayesianas vs redes neuronales en modelos para la predicción del acabado superficial. *Ivema*.

datascience. (21 de enero de 2014). *datascience.berkeley.edu*. Obtenido de University of Berkeley School of Information: https://datascience.berkeley.edu/wef-online-education-infographic/

Duro Novoa, V., & Gilart Iglesias, V. (2016). La competitividad en las instituciones de educación superior. Aplicación de filosofías de gestión empresarial: LEAN, SIX SIGMA y BUSINESS PROCESS MANAGEMENT (BPM). *Economía y Desarrollo, 157*(2), 166-181. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0252-85842016000200012&lng=es&tlng=es.

Duro, V. N., & Gilart, V. I. (2016). La competitividad en las instituciones de educación superior. Aplicación de filosofías de gestión empresarial: LEAN, SIX SIGMA y BUSINESS PROCESS MANAGEMENT (BPM). *Economía y Desarrollo, 157*(2), 166-181. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0252-85842016000200012&lng=es&tlng=es.

Gallant, S. I. (1993). *Neural network learning and expert systems.* MIT press.

gartner. (1 de Agosto de 2019). *https://www.gartner.com/*. Obtenido de https://www.gartner.com/doc/reprints?id=1-1OA35PNQ&ct=190715&st=sb&mkt_tok=eyJpIjoiTVdFNE9HVmxNamM0T1dNeSIsInQiOiJnVytuTE1yVlRIZFdJZDBuYVNiTzY1UTJONElGUzQ2Y0tKV2JRc0lSU1F0dGtUSGZYdWtNTXN0b2cwcUd6R0NZR2JKb2h2dDNjWCtPOStxdngxNDd0aVNIMnVBMWFsZ3I4endZK0pma09X

Gerrero, A. H. (2019). *www.unach.edu.ec*. Obtenido de http://www.unach.edu.ec/images/pdf/reglam%20de%20evaluacion%20integral%20al%20desempeno%20del%20personal%20academico.pdf

Gomez, A. R. (2014). Modelo de detección de estudiantes excluidos en carreras de ingeniería utilizando Minería de Datos. *Revista Ingenio Universidad Francisco de Paula Santander Ocaña, 6*(1), 8.

González, E. G. (2017). Factores que inciden en el rendimiento académico de los estudiantes de la Universidad Politécnica del Valle de Toluca. *Revista Latinoamericana de Estudios Educativo, 47*(1).

González, M. V. (2015). El uso del Perceptrón Multicapa para la clasificación de patrones en conductas adictivas.

Gutiérrez-Priego, R. (26 de abril de 2015). *https://www.oei.es/*. Obtenido de https://www.oei.es/historico/divulgacioncientifica/?Learning-analytics-instrumento

Hammerstrom, D. (1993). Neural networks at work. *IEEE spectrum, 30*(6), 26-32.

Hammerstrom, D. (1993). Working with neural networks. *IEEE spectrum, 30*(7), 46-53.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining (adaptive computation and machine learning).* MIT Press.

Haykin, S. (1999). Neural Networks, A comprehensive Foundation Second Edition by Prentice-Hall. *Macmillan College Publishing*.

Hernández, R. S., Fernández, C. C., & Pilar, B. L. (2014). *Metodología de la Investigación. Sexta edición.* . Santa Fe : McGRAW-HILL / Interamericana Editores, S.A. de C.V.

Herrera Díaz, C. A. (2016). Implementación de un módulo de análisis estadístico y predictivo para agricultura utilizando bigdata y machine learning, integrado al sistema iotmach. *Bachelor's thesis.* UTMACH, Machala.

Hilera, J. R., Martínez, V. J., & others. (2000). *Redes neuronales artificiales: fundamentos, modelos y aplicaciones.* Mexico: Alfaomega.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks, 2*(5), 359-366.

Huapaya, C. R., Lizarralde, F. Á., Arona, G., & Massa, S. M. (2012). Minería de datos educacional en ambientes virtuales de aprendizaje. *XIV Workshop de Investigadores en Ciencias de la Computación* (págs. 996-1000). Red de Universidades con Carreras en Informática (RedUNCI).

Jaramillo, A., & Arias, H. P. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. *Revista Tecnológica-ESPOL, 28*(1).

kdnuggets. (Agosto de 2007). *www.kdnuggets.com*. Obtenido de https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

kdnuggets. (29 de marzo de 2014). *https://www.kdnuggets.com*. Obtenido de https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

kdnuggets. (enero de 2017). *https://www.kdnuggets.com/*. Obtenido de https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer.* Waltham, Usa: Morgan Kaufmann.

Larranaga, P., Inza, I., & Moujahid, A. (1997). Tema 8. redes neuronales. *Redes Neuronales, U. del P. Vasco*, 12-17.

Llinás-Audet, F. J., Girotto, M., & Solé, F. P. (mayo - agosto de 2011). La dirección estratégica universitaria y la eficacia de las herramientas de gestión: el caso de las universidades españolas. *Revista de Educación, 255*, 33--54.

López, C. P. (2007). *Minería de datos: técnicas y herramientas.* Editorial Paraninfo.

López, J. M., & Herrero, J. G. (2006). Técnicas de análisis de datos. *Aplicaciones prácticas utilizando Microsoft Excel y Weka*.

López, R. F., & Fernandez, J. M. (2008). *Las redes neuronales artificiales.* Netbiblo.

Maithili, A., Kumari, R. V., & Rajamanickam, S. (2012). Neural network towards business forecasting. *IOSR journal of engineering*, *2*(04), 831-836.

Manjarrez, L. F. (2014). Relaciones Neuronales Para Determinar la Atenuación del Valor de la Aceleración Máxima en Superficie de Sitios en Roca Para Zonas de Subducción. *phdthesis.* UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO, Mexico.

Marqués, M. P. (2014). *Minería de datos a través de ejemplos.* RC Libros.

Martín-del-Brío, B., & Serrano-Cinca, C. (2019). Fundamentos de las redes neuronales artificiales: hardware y software. *Scire: Representación y organización del conocimiento, 1*(1), 103-125.

Mata, H. A. (2019). Análisis de las técnicas de suavizado para series temporales aplicadas a la base de datos del sistema académico de la UNACH. *Bachelor's thesis.* Universidad Nacional de Chimborazo UNACH, Riobamba. Obtenido de http://dspace.unach.edu.ec/handle/51000/6256

Mejia, C. R., Valladares-Garrido, M. J., & Valladares-Garrido, D. (2018). Baja publicación en revistas científicas de médicos peruanos con doctorado o maestría: Frecuencia y características asociadas. *Educación Médica, 19*, 135-141.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (págs. 935--940).

Mitchell, T. M. (1997). Artificial neural networks. *Machine learning, 45*, 81-127.

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. *In XIII Workshop de Investigadores en Ciencias de la Computación*, (págs. 1-4). Buenos Aires.

Olabe, X. B. (1998). Redes neuronales artificiales y sus aplicaciones. *Publicaciones de la Escuela de Ingenieros*.

Orellana Parapi, J. M. (2018). Uso de una neuronal roja para determinar el rendimiento de una organización del sector bancario. *Bachelor's thesis.* Universidad de Cuenca, Cuenca.

Ospina, F. V., & Herrera, M. S. (2018). Integración de etapas tecnicas de metodologias BI con el metodo de gestion Lean Analytics: Una metodología para obtención del conocimiento. *Tesis.* Tecnológico de Antioquia, Medellin, Colombia.

Palmer, A., Montaño, J., & Jiménez, R. (2001). Tutorial sobre redes neuronales artificiales: el Perceptrón Multicapa. *Revista electrónica de psicología, 5*(2).

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications, 41*(4), 1432--1462.

Pereira, R. T., Romero, A. C., & Toledo, J. J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Revista vínculos, 10*(1), 373-383.

Piatetsky, G. (Octubre de 2014). *www.kdnuggets.com/*. Obtenido de KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pinninghoff, M. A., Salcedo, P. A., & Contreras, R. A. (2007). Neural Networks to Predict Schooling Failure/Success. Nature Inspired Problem-Solving Methods in Knowledge Engineering. *Informatics Engineering and Computer Science Department, ,* 571-579.

Pitarque, A., Roy, J. F., & Ruiz, J. C. (1998). Redes neurales vs modelos estadísticos: Simulaciones sobre tareas de predicción y clasificación. *Psicológica, 19*, 387-400.

Pitarque, A., Ruiz, J. C., & Roy, J. F. (2000). Las redes neuronales como herramientas estadísticas no paramétricas de clasificación. *Psicothema, 12*(2), 459-463.

Pol, A. P., & Moreno, J. J. (2002). Redes neuronales artificiales aplicadas al análisis de supervivencia: un estudio comparativo con el modelo de regresión de Cox en su aspecto predictivo. *Psicothema, 14*(3), 630-636.

Quinteros, O. E., Funes, A., & Ahumada, H. C. (2016). Extracción de conocimiento en el cursado del ciclo común de articulación de carreras de Ingeniería. *In XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre RÍos, Argentina).*

Reveco, C., & Vergara, C. (2018). *www.u-cursos.cl.* Obtenido de docplayer.es: https://www.u-cursos.cl/diplomados/2011/0/DPIN-T/1/material_docente/bajar?id_material=376854

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Inteligencia Artificial. *Revista Iberoamericana de Inteligencia Artificial, 10*(29), 11-18.

Rodríguez, G. d. (2015). La ciencia de los datos y su impacto en la gestión universitaria. *Revista científíca ecociencia, 2*(1).

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(6), 601-618.

Rumelhart, D. E., & McClelland, J. L. (1986). *On learning the past tenses of English verbs.* Cambridge, MA: MIT Press.

Solines Bernardino, J. J. (2018). Minería de datos aplicada a la detección de patrones para el análisis de rendimiento académico de los estudiantes de la carrera de Ingeniería en Sistemas Computacionales de la Universidad Católica Santiago de Guayaquil. *Tesis de pregrado.* UNIVERSIDAD CATÓLICA DE SANTIAGO DE GUAYAQUIL, Guayaquil.

u-planner. (2018). *www.u-planner.com*. Obtenido de Al auge del big data en la educación superior: https://www.u-planner.com/es/blog/al-auge-de-los-grandes-datos-en-la-educacion-superior

Viñuela, I. P., & Leon, G. (2004). Redes de neuronas artificiales: un enfoque práctico. *Pearson*.

Werbos, P. (1974). Beyond regression:" new tools for prediction and analysis in the behavioral sciences". *Ph. D. dissertation, Harvard University.*

Zaidi, E., Heudecker, N., & Thoo, E. (1 de Agosto de 2019). Magic quadrant for data integration tools. *Gartner RAS Core Research Note G*. Obtenido de

https://www.gartner.com/en/documents:

https://b2bsalescafe.files.wordpress.com/2019/09/gartner-magic-quadrant-for-data-integration-tools-august-2019.pdf

Zhang, G. P. (2006). Avoiding pitfalls in neural network research. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *37*(1), 3-16.

# ANEXOS

## **ANEXO 1:** Producción de un plan del proyecto

*Tabla 45. Plan del proyecto*

| Id | Modo de tarea | Nombre de tarea | Duración | Comienzo | Fin | Predecesoras | Nombres de los recursos |
|---|---|---|---|---|---|---|---|
| 1 | | **Fase 1: Comprensión del Negocio o Problema** | **16 días** | **mié 02/10/19** | **mié 23/10/19** | | **Investigador** |
| 2 | | 1. Determinar el objetivo del negocio | 5 días | mié 02/10/19 | mar 08/10/19 | | Microsoft Word |
| 3 | | 2. Evaluación de la situación | 5 días | mié 09/10/19 | mar 15/10/19 | 2 | Microsoft Word |
| 4 | | 3. Determinación de los objetivos de DM | 4 días | mié 16/10/19 | lun 21/10/19 | 3 | Microsoft Word |
| 5 | | 4. Producción de un plan del proyecto | 2 días | mar 22/10/19 | mié 23/10/19 | 4 | Microsoft Project |
| 6 | | **Fase 2: Comprensión de los datos** | **16 días** | **jue 24/10/19** | **jue 14/11/19** | | **Investigador** |
| 7 | | 1. Recolección de datos iniciales | 2 días | jue 24/10/19 | vie 25/10/19 | 5 | Microsoft Excel |
| 8 | | 2. Descripción de los datos | 4 días | lun 28/10/19 | jue 31/10/19 | 7 | Microsoft Excel |
| 9 | | 3. Exploración de datos | 3 días | vie 01/11/19 | mar 05/11/19 | 8 | Microsoft Excel |
| 10 | | 4. Verificación de la calidad de los datos | 7 días | mié 06/11/19 | jue 14/11/19 | 9 | Microsoft Excel |
| 11 | | **Fase 3: Preparación de los datos** | **26 días** | **vie 15/11/19** | **vie 20/12/19** | | **Investigador;Tutor de Tesis** |
| 12 | | 1. Selección de datos | 3 días | vie 15/11/19 | mar 19/11/19 | 10 | talend data quality online |
| 13 | | 2. Limpieza de los datos | 9 días | mié 20/11/19 | lun 02/12/19 | 12 | talend data quality online |
| 14 | | 3. Estructuración de los datos | 5 días | mar 03/12/19 | lun 09/12/19 | 13 | talend data quality online |
| 15 | | 4. Integración de los datos | 5 días | mar 10/12/19 | lun 16/12/19 | 14 | talend data quality online |
| 16 | | 5. Formateo de datos | 4 días | mar 17/12/19 | vie 20/12/19 | 15 | talend data quality online |
| 17 | | **Fase 4: Modelado** | **17 días** | **lun 23/12/19** | **mar 14/01/20** | | **Investigador;Tutor de Tesis** |
| 18 | | 1. Selección de la técnica de modelado | 2 días | lun 23/12/19 | mar 24/12/19 | 16 | RapidMiner |
| 19 | | 2. Generación del plan de prueba | 4 días | mié 25/12/19 | lun 30/12/19 | 18 | RapidMiner |
| 20 | | 3. Construcción del Modelo | 6 días | mar 31/12/19 | mar 07/01/20 | 19 | RapidMiner |
| 21 | | 4. Evaluación del modelo | 4 días | mié 08/01/20 | lun 13/01/20 | 20 | RapidMiner |
| 22 | | **Fase 5: Evaluación** | **17 días** | **mié 15/01/20** | **jue 06/02/20** | | **Investigador;Tutor de Tesis** |
| 23 | | 1. Evaluación de los resultados | 6 días | mié 15/01/20 | mié 22/01/20 | 21 | RapidMiner |
| 24 | | 2. Proceso de revisión | 6 días | jue 23/01/20 | jue 30/01/20 | 23 | RapidMiner |
| 25 | | 3. Determinación de futuras fases | 4 días | vie 31/01/20 | mié 05/02/20 | 24 | RapidMiner |
| 26 | | **Fase 6: Generación del Proyecto** | **39 días** | **vie 20/12/19** | **mié 12/02/20** | | **Investigador;Tutor de Tesis** |
| 27 | | 1. Informe Final | 27 días | vie 20/12/19 | lun 27/01/20 | 25 | Microsoft Word |
| 28 | | 2. Revisión del proyecto | 12 días | mar 28/01/20 | mié 12/02/20 | 27 | Microsoft Word |

## **ANEXO 2:** Descripción de los datos

*Tabla 46. Descripción de la tabla estudiante*

| Atributo | Tipo | Descripción |
|---|---|---|
| Estudiante ID | Numérico | Identificador único del estudiante |
| Fecha Nacimiento | Fecha | Fecha de nacimiento del estudiante |
| Estado Civil | Texto | Estado Civil del estudiante (casado, soltero, divorciado, viudo, unión libre) |
| Orientación Sexual | Texto | Atracción afectiva del estudiante. |
| Genero | Texto | Tipo de Género del estudiante (Masculino, Femenino) |
| Etnia | Texto | Etnia en la cual se considera el estudiante (Indígena, Afroecuatoriano, Negro, Mulato, Montubio, Mestizo, Blanco, Otro) |

| | | |
|---|---|---|
| Nacionalidad Indígena | Texto | Indica si el estudiante proviene de alguna nacionalidad indigna del país (Achuar, Chachi, Chibuleo, Kañari, Karanki, Kayambi, Kichwa, Kisapincha, Otavalo, Panzaleo, Puruhá, Salasaka, Saraguro, Shuar, Tomabela, Waranka, no aplica ). |
| Institución Educativa | Texto | Institución secundaria de donde proviene el estudiante |
| Tipo | Texto | Tipo de Institución Educativa (Beneficencia, Extranjero, Fiscal, Fiscomisional, Municipal y Particular) |
| Enfermedad Catastrófica Extraña | Texto | Indica si el estudiante posee alguna enfermedad. |
| Tipo Discapacidad | Texto | Indica si el estudiante tiene algún tipo de discapacidad. |
| Actividad Cultural | Texto | Indica la actividad cultural que realiza el estudiante |
| NumeroIntegrantesHogar | Numérico | Número de integrantes en la familia del estudiante |
| País Nacimiento | Texto | País de nacimiento del estudiante. |
| Provincia Nacimiento | Texto | Provincia de nacimiento del estudiante |
| Cantón Nacimiento | Texto | Cantón de nacimiento del estudiante |
| País Procedencia | Texto | País de donde procede el estudiante |
| Provincia Procedencia | Texto | Provincia de donde procede el estudiante |
| Cantón Procedencia | Texto | Cantón de procedencia del estudiante. |
| País Dirección | Texto | País donde reside el estudiante. |
| Dirección Provincia | Texto | Provincia donde reside el estudiante. |
| Dirección Cantón | Texto | Cantón donde reside el estudiante. |
| Parroquia | Texto | Parroquia donde reside el estudiante. |
| Tipo Parroquia | Texto | Tipo de parroquia de donde procede el estudiante (rural, urbana) |
| Numero Hermanos | Numérico | Numero de hermanos que tiene el estudiante |
| Ingresos Padre | Numérico | Ingresos mensuales del padre del estudiante. |
| Ingresos Madre | Numérico | Ingresos mensuales de la madre del estudiante. |
| Ocupación Madre | Texto | Ocupación de la madre del estudiante |
| Ocupación Padre | Texto | Ocupación del padre del estudiante |
| TotalIngresosPadres | Numérico | Total, de ingresos mensuales de los padres del estudiante. |
| NúmeroDependenIngresos | Numérico | Número de personas que dependen de los ingresos de los padres del estudiante. |
| TipoVivienda | Texto | Indica el tipo de vivienda del estudiante (Arrendada, Prestada, Propia). |
| Tipo Construcción | Texto | Indica si la vivienda del estudiante es de construcción mixta, Ladrillo, Bloque, Caña, Madera, Mixta, Otro. |
| ServicioAguaPotable | Texto | El estudiante posee Servicios de Agua Potable (SI NO) |
| ServicioElectricidad | Texto | El estudiante posee servicio de electricidad (SI NO) |
| ServicioTelefono | Texto | El estudiante posee servicio de teléfono (SI NO) |
| ServicioInternet | Texto | El estudiante posee servicio de internet (SI NO) |
| ServicioTVPagada | Texto | El estudiante posee servicio de TV Pagada (SI NO) |
| ValorMensualServicios | Numérico | Valor Mensual de los Servicio que posee el estudiante |
| TieneVehículo | | El estudiante posee vehículo (SI NO) |
| Ocupación | Texto | Indica si el estudiante hace actividades extra aparte de estudiar |
| Total, Ingresos | Numérico | Ingresos mensuales del estudiante. |
| Numero Hijos | Numérico | Número de hijos que posee el estudiante |
| Ocupación Cónyuge | Texto | Ocupación del conyugue del estúdiate. |
| Ingresos Cónyuge | Numérico | Ingresos mensuales del cónyuge del estudiante. |
| TotalIngresosEstudiante | Numérico | Total, de ingresos mensuales del estudiante. |
| PersonasDependen Ingresos | Numérico | Número de personas que dependen de los ingresos del estudiante. |

*Tabla 47. Descripción de la tabla Estudiante_rendimiento*

| Atributo | Tipo | Descripción |
|---|---|---|
| EstudianteID | Numérico | Identificador único del estudiante. |
| Facultad | Texto | Indica la facultad a la que pertenece el estudiante |
| Carrera | Texto | Indica la carrera a la que pertenece el estudiante |
| Situación Actual | Texto | Situación del estudiante (Graduado, No Graduado) |

| | | |
|---|---|---|
| Nivel | Texto | Semestre del estudiante. |
| Periodo | Texto | Periodo en el que se matriculo el estudiante |
| Promedio | Numérico | Promedio general que obtuvo al finalizar el semestre |

*Tabla 48. Descripción de la tabla Docente*

| Atributo | Tipo | Descripción |
|---|---|---|
| Cedula | Texto | Número de cédula del docente. |
| País | Texto | País de procedencia. |
| Nacionalidad | Texto | Nacionalidad según el país de procedencia. |
| Fecha Nacimiento | Fecha | Fecha de nacimiento del docente. |
| Estado Civil | Texto | Estado civil del docente |
| Sexo | Texto | Sexo del docente |
| Etnia | Texto | Etnia del docente. |
| Tipo Sangre | Texto | Tipo de sangre del docente. |
| GrupoGLBTI | Texto | Grupo LGBTI al que pertenece el docente. |
| Nacionalidad Indígena | Texto | Nacionalidad indígena del docente. |
| País | Texto | País en el que radica actualmente. |
| Cantón | Texto | Cantón en el que radica actualmente. |
| Parroquia | Texto | Parroquia en la que radica actualmente. |
| Número Hijos | Numérico | Número de hijos que tiene el docente. |
| NivelIsntrucción | Texto | Tipo de título académico del docente. |
| País | Texto | País en el que se obtuvo el título. |
| TiempoEstudio | Texto | Tiempo que demoró en obtener el título. |
| Modalidad | Texto | Modalidad de estudio. |
| Área | Texto | Área en la que obtuvo el título. |
| Subárea | Texto | Subárea en la que obtuvo el título. |
| Campo | Texto | Campo en el que obtuvo el título. |
| EstáCursando | Texto | Si se encuentra estudiando actualmente, |
| InstituciónEducativa | Texto | Institución educativa en la que se obtuvo el |
| Título | Texto | Título que obtuvo. |
| NoEventosAprobados | Numérico | Numero de eventos aprobados del docente. |
| NoEventosAsistidos | Numérico | Numero de eventos asistidos por el docente. |
| HorasEventosAprobados | Numérico | Horas de eventos aprobados del docente. |
| HorasEventosAsistidos | Numérico | Horas de eventos asistidos por el docente. |
| NoEventosNacionales | Numérico | Numero de eventos nacionales del docente. |
| NoEventosInternacionales | Numérico | Numero de eventos internacionales docente. |
| ExperienciaPrivada | Texto | Experiencia del docente en entidades privadas. |
| ExperienciaPública | Texto | Experiencia del docente en entidades públicas. |
| FamiliarSustituto | Texto | Familiar sustituto en el trabajo. |
| EnfermedadCatastrófica | Texto | El docente posee una enfermedad catastrófica. |
| TieneDiscapacidad | Texto | El docente posee alguna discapacidad. |
| GestaciónLactancia | Texto | Se encuentra en estado de gestación o lactancia |

*Tabla 49. Descripción de la tabla Docente_infAcadémica*

| Atributo | Tipo | Descripción |
|---|---|---|

| | | |
|---|---|---|
| NumeroDocumento | Texto | Contiene el número de cedula del docente. |
| Facultad | Texto | Facultad a la que pertenece el docente. |
| Carrera | Texto | Carrera a la que pertenece el docente. |
| Periodo | Texto | Periodo en que dio clases el docente |
| ActividadAcadémica | Texto | Actividad académica del docente. |
| HorasActividadAcadémica | Numérico | Horas di actividad académica del docente |
| HorasClase | Numérico | Horas de clase impartidas por el docente. |

*Tabla 50. Descripción de la tabla Evaluación_Docente*

| Atributo | Tipo | Descripción |
|---|---|---|
| UsuarioEvaluado | Texto | Número de usuario del docente evaluado. |
| Cedula | Texto | Numero de cedula del docente evaluado |
| Tipo Evaluación | Texto | Tipo de evaluación que se realizó al docente. (Autoevaluación, Coevaluación Directivos, Coevaluación Pares, Heteroevaluación) |
| Componente | Texto | Componente en el que fue evaluado el docente. (docencia, gestión, investigación) |
| Resultado Final | Texto | Calificación de la evaluación al docente |
| Periodo | Texto | Periodo que fue evaluado el docente |

*Tabla 51. Descripción de la tabla Publicación*

| Atributo | Tipo | Descripción |
|---|---|---|
| ESTADO PUBLICACION | Texto | Describe el estado actual de la publicación. |
| TIPO PUBLICACION | Texto | Tipo de publicación realizada por el docente |
| TITULO | Texto | Título del proyecto de investigación |
| REVISTA | Texto | Nombre de la revista en la que fue publicada. |
| CEDULA | Texto | Número de cedula del docente publicador |
| ROL INSTITUCION | Texto | Rol del docente publicador en la institución. |
| SEXO | Texto | Sexo del docente publicador. |
| TIPO AUTOR | Texto | Si el autor es docente o no. |
| ORDEN AUTOR | Numérico | Orden de autor en la investigación. |
| NOMBRES | Texto | Nombres del docente. |
| APELLIDO MATERNO | Texto | Apellido materno del docente. |
| APELLIDO PATERNO | Texto | Apellido paterno del docente. |
| AREA DE INVESTIGACION | Texto | Área en la que se realizó la investigación. |
| LINEADEINVESTIGACION | Texto | Línea en la que se realizó la investigación. |
| CAMPO AMPLIO | Texto | Campo amplio en la que se realizó la investigación |
| CAMPO DETALLADO | Texto | Campo detallado en la que se realizó la investigación |
| CAMPO ESPECIFICO | Texto | Campo específico en la que se realizó la investigación. |
| AÑO | Fecha | Año en el que se realizó la investigación. |
| AÑO-MES DE PUBLICACION | Fecha | Año y mes que se realizó la investigación |
| AÑO-MES DE REGISTRO | Fecha | Año y mes de registro de la investigación. |
| AÑO-MES REGISTRO MOD | Fecha | Año y mes de registro de la investigación. |
| FECHA ACEPTACION | Fecha | Fecha que fue aceptada la investigación. |
| FECHA ACTUALIZACION | Fecha | Fecha que fue actualizada la investigación. |
| FECHA DE REGISTRO | Fecha | Fecha en la que fue registrada la investigación |
| FECHA PUBLICACION | Fecha | Fecha en la que fue publicada la investigación |
| FACULTAD | Texto | Facultad en la que se realizó la investigación. |
| CARRERA | Texto | Carrera en la que se realizó la investigación. |
| Código Carrera | Numérico | Código de la carrera en la que se realizó la investigación |
| CIUDAD DE PUBLICACION | Texto | Ciudad en la que se realizó la publicación. |
| COMITE CIENTIFICO U ORGANIZADOR | Texto | Comité científico u organizador. |
| COMITE EDITORIAL O EXPERTO | Texto | Comité editorial o experto. |
| CONGRESO O SEMINARIO | Texto | Congreso o seminario. |
| ES EDITORIAL DE PRESTIGIO | Texto | Indica si la editorial es de prestigio o no. |

| | | |
|---|---|---|
| ES EDITORIAL DE PRESTIGIO | Numérico | Indica si la editorial es de prestigio o no. |
| EXISTE APROBACION DE COMISION | Numérico | Indica si fue aprobado o no por una comisión. |
| EXISTE COMITE CIENTIFICO U ORGANIZADOR | Numérico | Indica el un comité científico u organizador |
| EXISTE COMITÉ EDITORIAL | Numérico | Indica el comité editorial. |
| EXISTE PROCEDIMIENTO SELECTIVO | Numérico | Indica un procedimiento selectivo |
| EXISTE REVISION POR PARES EXTERNOS | Numérico | Indica la revisión por parte de pares externos |
| FORMA PUBLICACION | Texto | Señala las características de la publicación. |
| FORMA PUBLICACION EN ARTICULO COMPLETO | Texto | Indica si forma o no publicación en artículo completo |
| OBSERVACIONES AUTOR | Texto | Muestran las observaciones del autor. |
| OBSERVACIONES DE COMISION | Texto | Muestra las observaciones de la comisión. |
| OBSERVACIONES GENERALES | Texto | Muestra las observaciones generales. |
| LISTADO DE REVISTAS SENESCYT | Texto | Listado revistas del listado de la SENESCYT. |
| OBSERVACIONES PUBLICACION | Texto | Señalan las observaciones realizadas en la publicación |
| DOAJ | Texto | La publicación se encuentra disponible en DOAJ |
| DOI | Texto | Identificador digital de la publicación. |
| EBSCO | Texto | La publicación se encuentra disponible en EBSCO |
| ESTADO PERSONAL ACADEMICO | Numérico | Estado personal académico del docente. |
| ISBN | Texto | Código ISBN que lo identifica. |
| ISI WEB KNOWLEDGE | Texto | La publicación está disponible en isi web Knowledge |
| ISSN | Texto | Código ISNN que lo identifica. |
| JSTOR | Texto | La publicación se encuentra disponible en JSTOR |
| LATINDEX | Texto | La publicación se encuentra en LATINDEX |
| LIBROS O CAPITULOS DE LIBROS REVISADOS POR PARES | Numérico | Indica el número de libros o capítulos revisados por pares. |
| LILACS | Numérico | Indica si la publicación se encuentra disponible en LILACS. |
| NACIONAL O INTERNACIONAL | Texto | La publicación es de tipo nacional o internacional |
| OAJI | Texto | La publicación se encuentra disponible en OAJI. |
| PAIS | Texto | País en el que se realizó la investigación. |
| PROCEDIMIENTO SELECTIVO | Texto | Señala la resolución final sobre la investigación. |
| PROQUEST | Texto | Indica si la publicación se encuentra disponible en PROQUEST |
| REDALYC | Texto | Indica si la publicación se encuentra disponible en REDALYC. |
| REVISION POR PARES EXTERNOS | Texto | Resolución que tomaron los pares externos. |
| SCIELO | Texto | Indica si la publicación se encuentra disponible en Scielo. |
| SCIMAGO JOURNAL RANK | Texto | Indica si la publicación se encuentra disponible en Scielo |
| SJR | Numérico | Indica el índice de impacto de la investigación |
| PAGINAS | Texto | Número de páginas de la publicación. |
| VOLUMEN | Numérico | Volumen de la revista en el que fue publicado |
| ORGANISMO DE AFILIACION | Texto | Organismo al que se encuentra afiliado el docente. |

**ANEXO 3:** Exploración de los datos

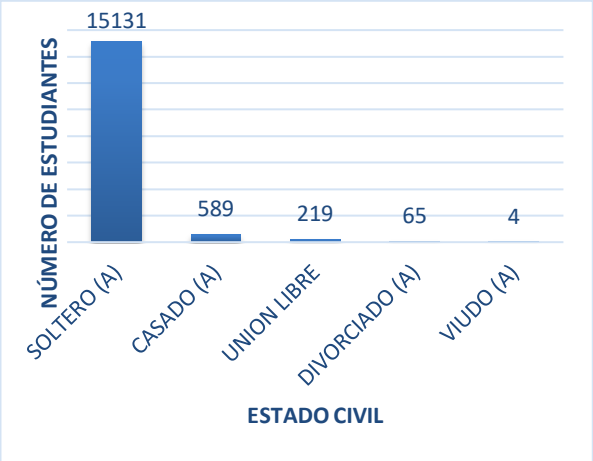



*Ilustración 24. Distribución de estudiantes por Estado Civil.*

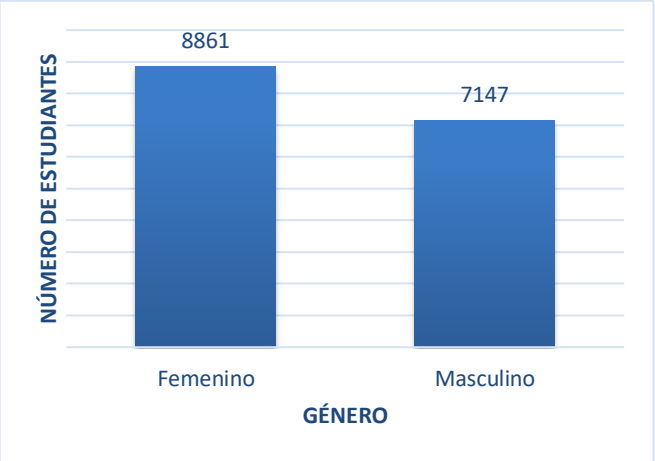



*Ilustración 25. Distribución de los estudiantes por Género*

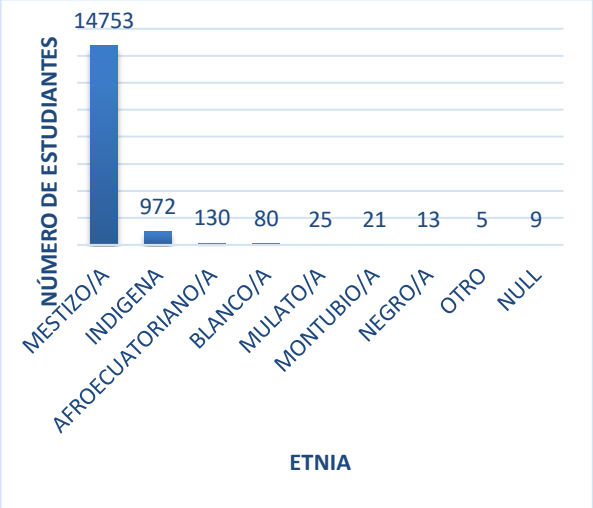



*Ilustración 26. Distribución de los estudiantes por Etnia*


*Ilustración 27. Distribución de los estudiantes por Nacionalidad Indígena*

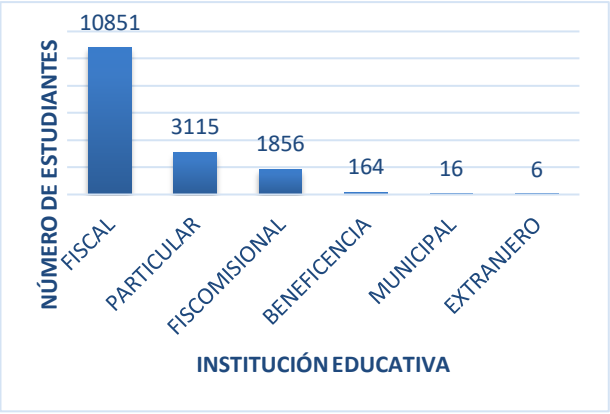



*Ilustración 28. Distribución de los estudiantes por Institución Educativa*

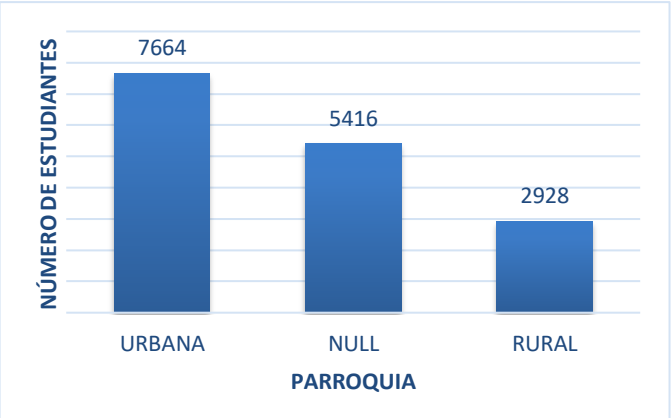



*Ilustración 29. Distribución de los estudiantes por Tipo de Parroquia*

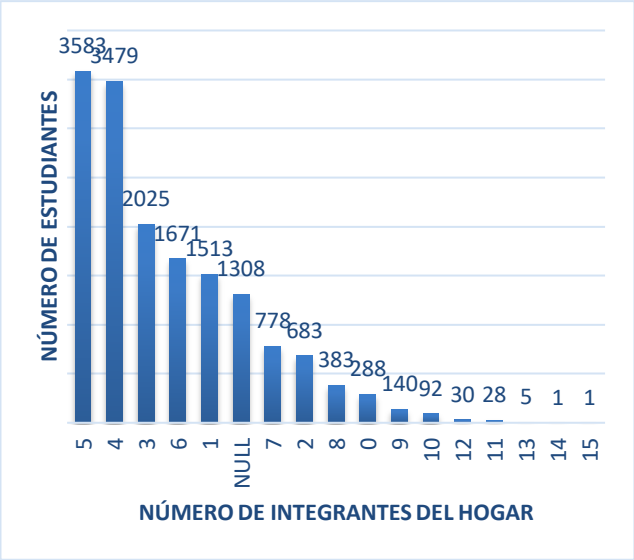



*Ilustración 30. Distribución de los estudiantes por número de Integrantes de Hogar*

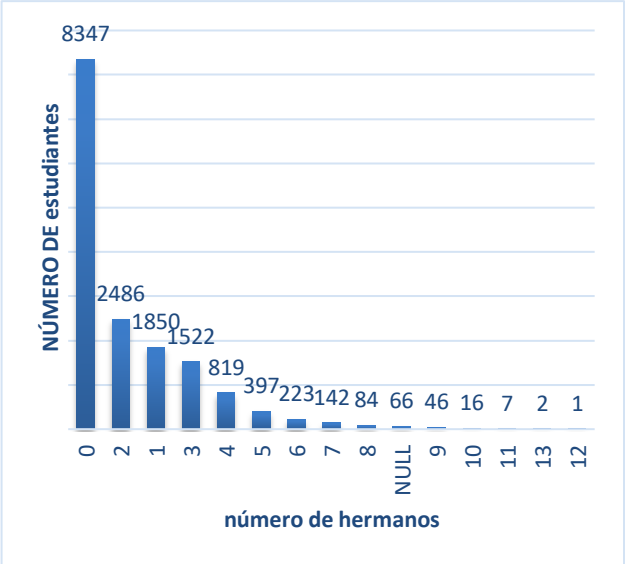



*Ilustración 31. Distribución de los estudiantes por número de hermanos*

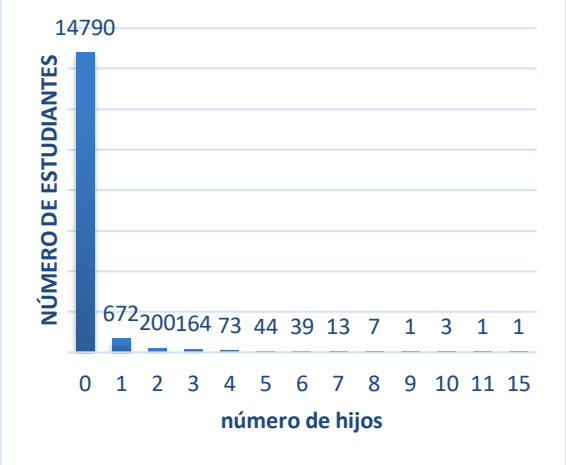



*Ilustración 32. Distribución de los estudiantes por número de Hijos*

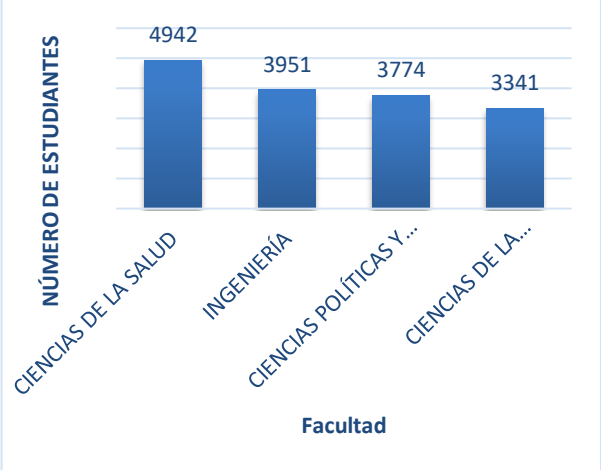



*Ilustración 33. Distribución de los estudiantes por Facultad*

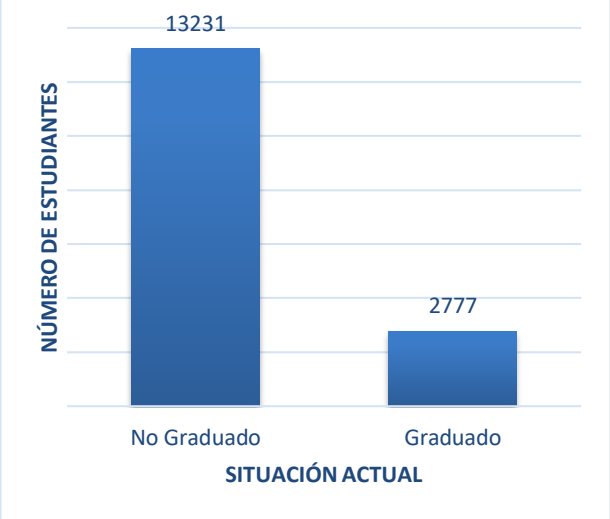



*Ilustración 34. Distribución de los estudiantes por Situación Actual*

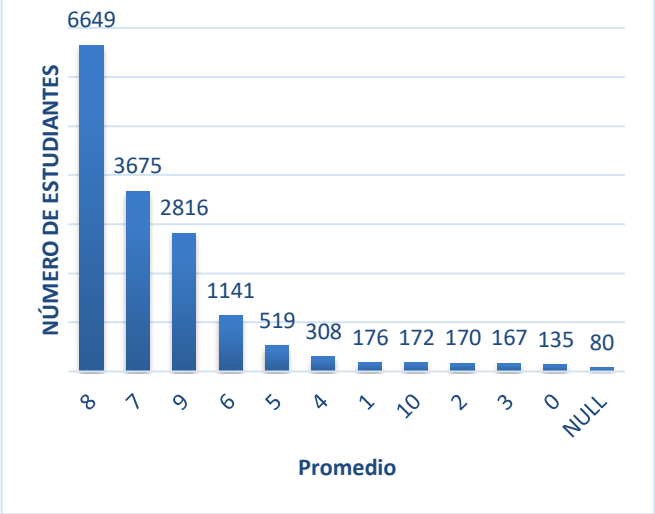



*Ilustración 35. Distribución de los estudiantes por el promedio (redondeado)*

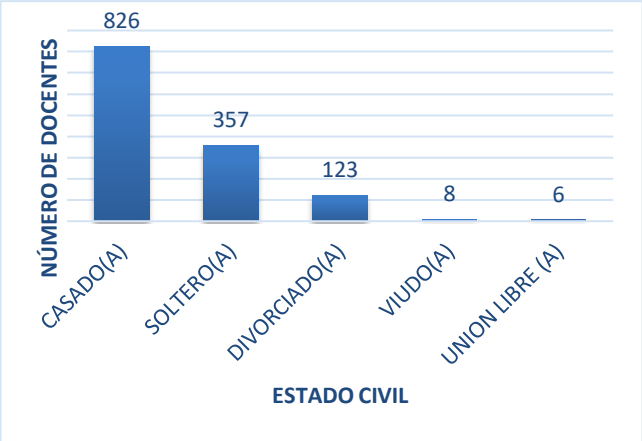



*Ilustración 36. Distribución de los docentes por el Estado Civil*

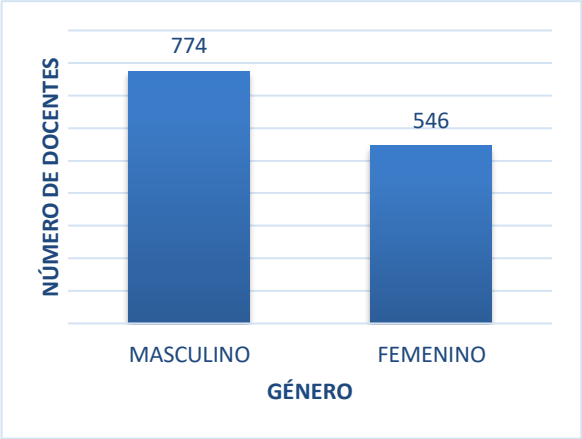



*Ilustración 37. Distribución de los docentes por el Género*

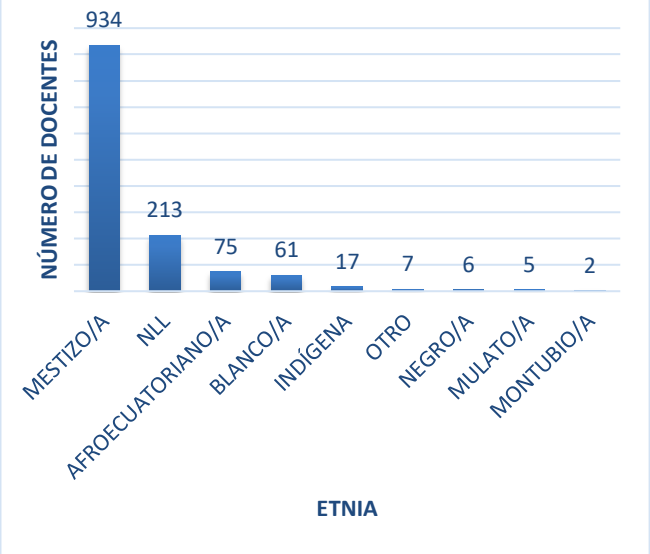



*Ilustración 38. Distribución de los docentes por Etnia*

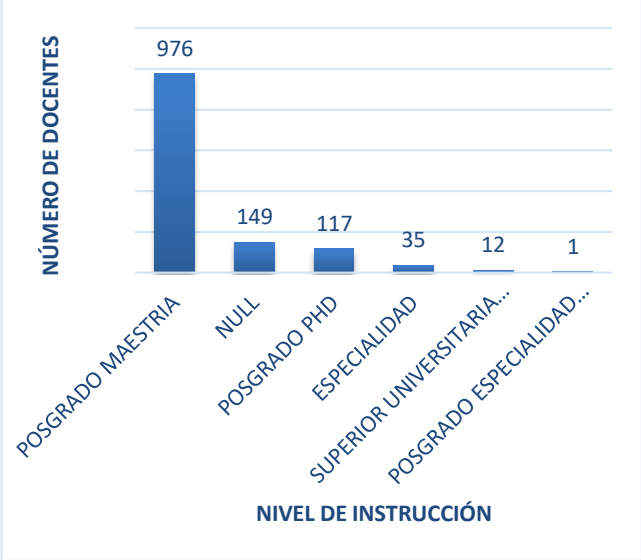



*Ilustración 39. Distribución de los docentes por Nivel de Instrucción*

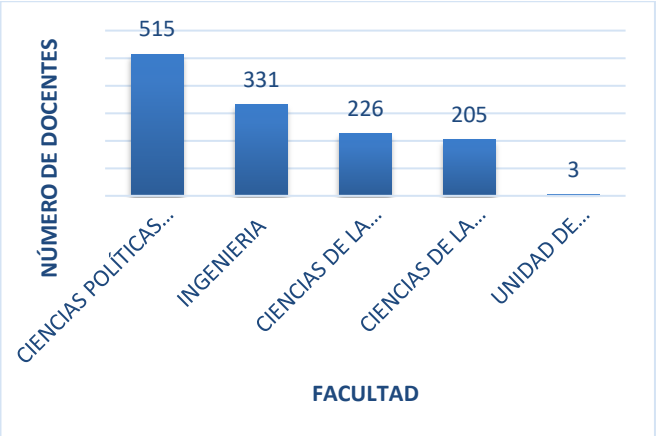



*Ilustración 40. Distribución de los docentes por Facultad*

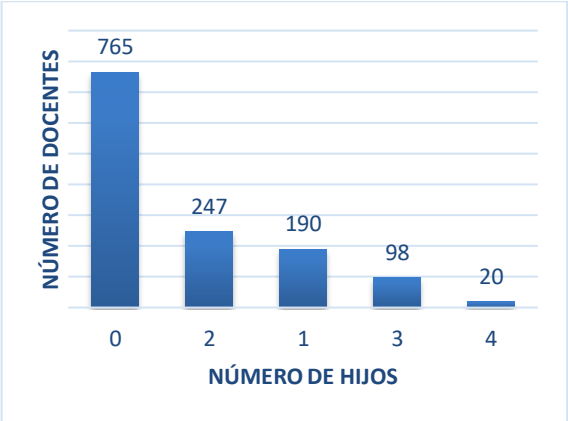



*Ilustración 41. Distribución de los docentes por Facultad*

**ANEXO 4:** Selección de los Datos

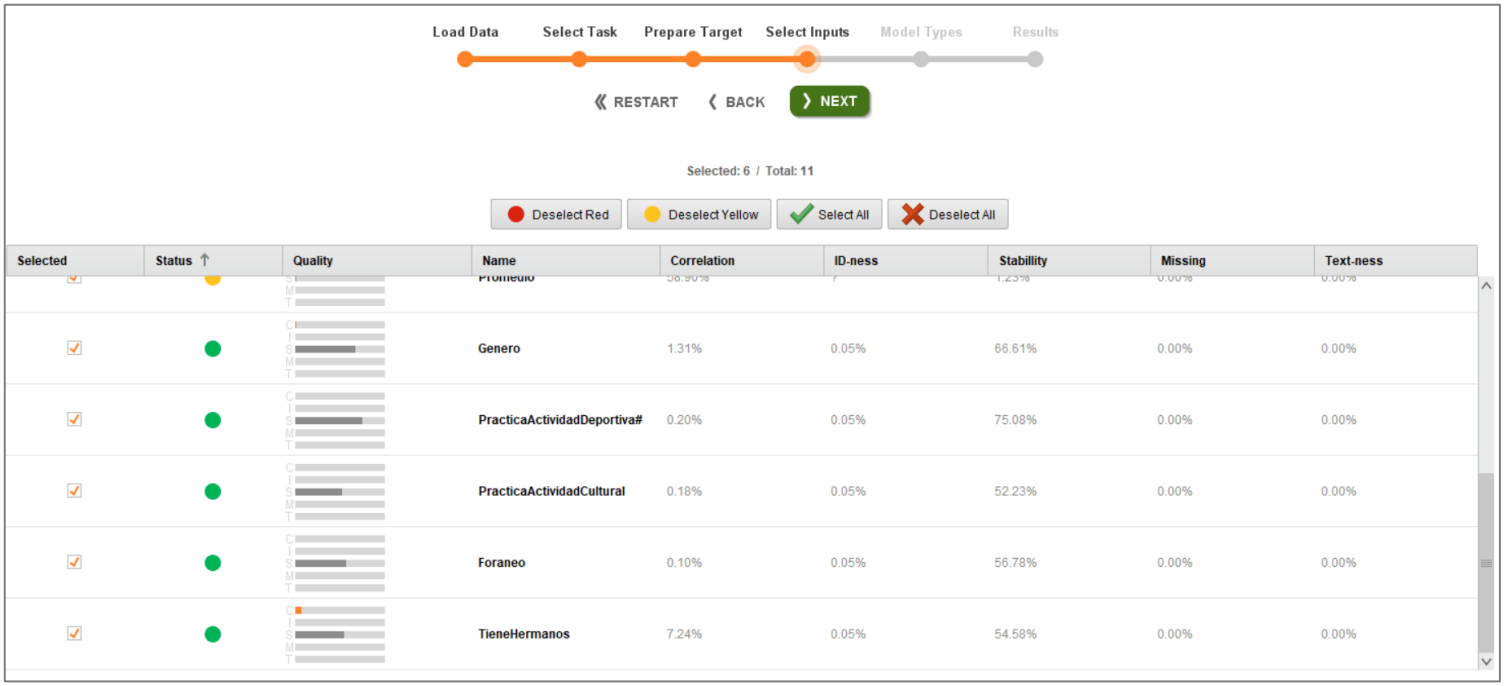



*Ilustración 42. Barras de calidad (C / I / S / M / T) para el rendimiento académico.*

**ANEXO 5:** Campos derivados

*Tabla 52. Campos Derivados*

| Tabla | Campo | Descripción |
|---|---|---|
| Estudiantes | Tiene Hermanos | Se asignó SI a todos los estudiantes que posean hermanos sin importar la cantidad, y NO a los que tienen cero. |
| | Promedio | Contiene el promedio general de todo el periodo de estudio, en la base solo estuvo el promedio por semestre. |
| Docente | Horas Actividad Académica | Si tiene horas de clase se coloca SI y sino No. |
| | Resultado Final Evaluación | Es el promedio general de la evaluación a los docentes. |
| Publicaciones | Tiene Publicaciones | Si tiene publicaciones de cualquier tipo se colocó SI y sino NO. |
| | Horas Clase | Si tiene horas clase tiene Si y sino NO. |

**ANEXO 6:** Construcción del Modelo ANN

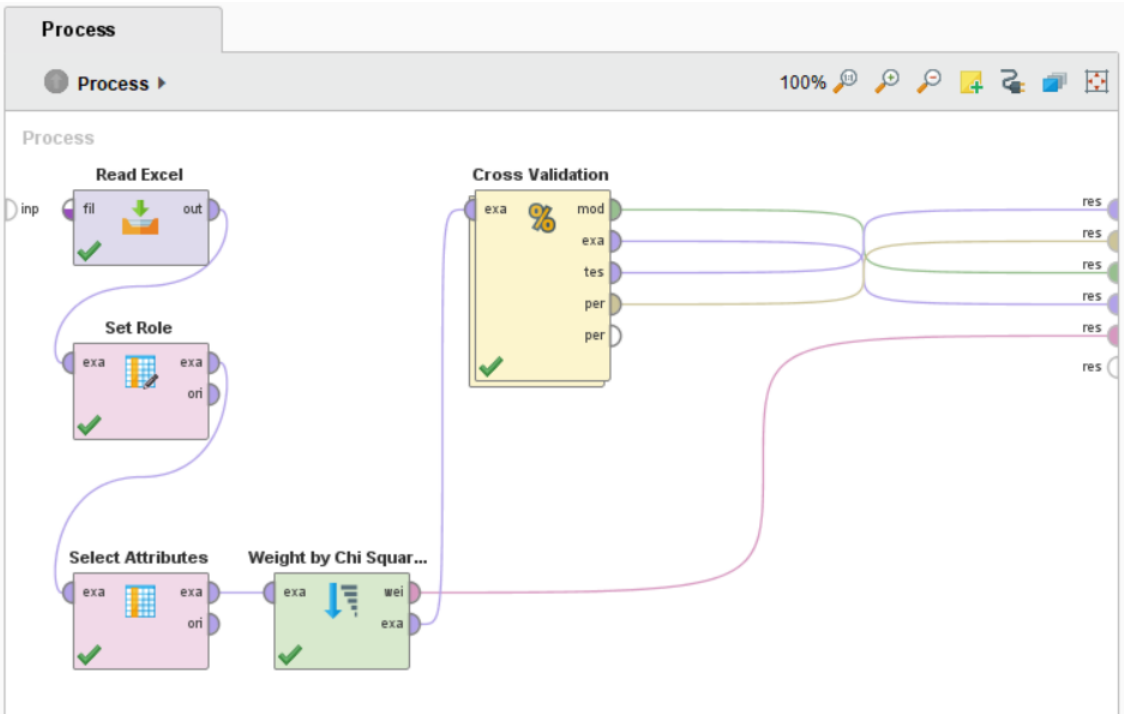



*Ilustración 43. Construcción del modelo de clasificación para ANN.*

A continuación, se detallan los pasos de la construcción del modelo

**Paso 1: preparación de datos**

**Read Excel:** Este operador lee un ExampleSet del archivo de Excel especificado, para cada proceso de Estudiante, Docente e Investigación se usa una tabla diferente. El modelo ANN no funcionará con tipos de datos categóricos o nominales. Y En nuestro caso poseemos algunos valores categóricos por tal motivo en el Formateo de Datos se convirtió a numérico.

**Set Role:** Este operador se usa para cambiar la función de uno o más atributos. En la tabla Estudiante se selecciona la variable objetiva "Ponderación Promedio", el cual contiene la ponderación del promedio final del estudiante; en la tabla Docente la "Equivalencia Calificación", el cual señala la ponderación de la calificación final de la evaluación aplicada al docente; y para Investigación "Tiene Publicaciones", en el que se indica si un docente ha realizado publicaciones o no, además se seleccionó producción de alto

impacto, revistas regionales y capitulo de libro. Estos campos se utilizaron de acuerdo con lo empleados en el apartado Formateo de Datos de la Metodología.

**Select Attributes:** Este operador selecciona un subconjunto de atributos de un conjunto de ejemplos y elimina los otros atributos. Estos campos se utilizaron de acuerdo con lo empleados en el apartado Formateo de Datos de la Metodología.

**Cross Validación:** Este operador realiza una validación cruzada para estimar el rendimiento estadístico de un modelo de aprendizaje.

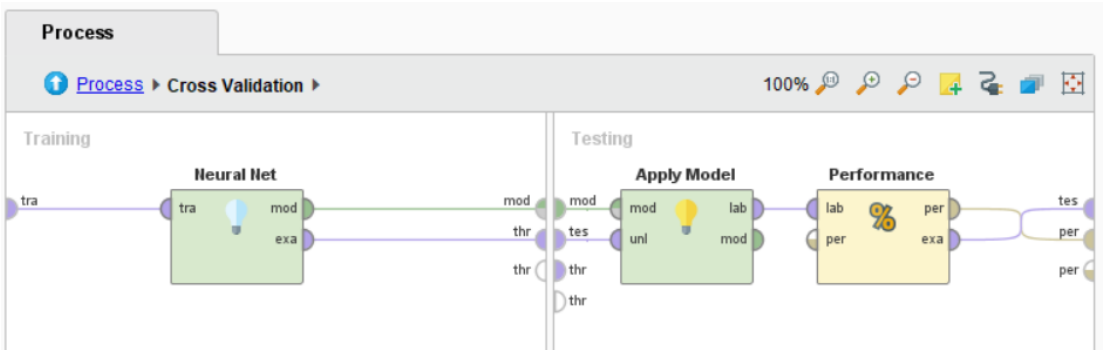



*Ilustración 44. Entrenamiento y Prueba para Redes Neuronales.*

**Paso 2: Operador de modelado y parámetros**

El conjunto de datos de entrenamiento está conectado al operador de la red neuronal (Modelado>Clasificación y regresión> Entrenamiento de redes neuronales).

El operador **Neural Net** acepta valores reales y luego los convierte en el rango normalizado -1 a 1 y genera un modelo ANN estándar. Los siguientes parámetros están disponibles en ANN para que los usuarios cambien y personalicen en el modelo.

**Capa oculta (Hidden layer):** determina el número de capas, el tamaño de cada capa oculta y los nombres de cada capa para una fácil identificación en la pantalla de salida. El tamaño predeterminado del nodo es -1, que en realidad se calcula por (número de atributos + número de clases) / 2 + 1. El valor predeterminado se puede sobrescribir especificando un número entero de nodos, sin incluir un nodo umbral sin entrada por capa.

**Ciclos de entrenamiento (Training cycles):** este número de veces que se repite un ciclo de entrenamiento; en el presente trabo de investigación se utilizó los valores de 500 y 1000 ciclos de aprendizaje. La disminución reduce el valor de la tasa de aprendizaje y la acerca a cero para el último registro de entrenamiento.

**Tasa de aprendizaje (Learning rate):** el valor de λ determina cuán sensible debe ser el cambio de peso al considerar el error del ciclo anterior. Toma un valor de 0 a 1. Un valor más cercano a 0 significa que el nuevo peso se basará más en el peso anterior y menos en la corrección de errores. Un valor más cercano a 1 se basaría principalmente en la corrección de errores. En el presente trabajo de investigación se utilizó tasas de 0,2; 0,02, 0,3, entre otros.

**Momentum:** este valor se usa para prevenir máximos locales y busca obtener resultados optimizados globalmente al agregar una fracción del peso anterior al peso actual. Para todo el proceso se usó un momento de 0,9.

**Normalizar (Normalize):** los nodos que utilizan una función de transferencia sigmoidea esperan una entrada en el rango de -1 a 1. Cualquier valor real de la entrada debe normalizarse en un modelo ANN.

**Error épsilon**: el objetivo del modelo ANN debería ser minimizar el error, pero no lo hace cero, en el cual el modelo memoriza el conjunto de entrenamiento y degrada el rendimiento. Podemos detener el proceso de construcción del modelo cuando el error es menor que un umbral llamado error épsilon.

**ANEXO 7:** ANN en RStudio

```
#Cargar Datos
library(readxl)
PFI <- read_excel("PFI.xlsx")
View (PFI)
```

```
PFI$EstadoCivil <- (PFI$EstadoCivil - min (PFI$EstadoCivil)) / (max (PFI$EstadoCivil)
- min (PFI$EstadoCivil))
PFI$Genero <- (PFI$Genero - min(PFI$Genero))/(max(PFI$Genero)-min(PFI$Genero))
PFI$NivelInstruccion <- (PFI$NivelInstruccion -
min(PFI$NivelInstruccion))/(max(PFI$NivelInstruccion) - min(PFI$NivelInstruccion))
PFI$TieneHijos <- (PFI$TieneHijos - min(PFI$TieneHijos))/(max(PFI$TieneHijos) -
min(PFI$TieneHijos))
PFI$NoEventosNacionales <- (PFI$NoEventosNacionales -
min(PFI$NoEventosNacionales))/(max(PFI$NoEventosNacionales) -
min(PFI$NoEventosNacionales))
PFI$NoEventosInternacionales <- (PFI$NoEventosInternacionales -
min(PFI$NoEventosInternacionales))/(max(PFI$NoEventosInternacionales)-
min(PFI$NoEventosInternacionales))
PFI$HorasActividadAcademica <- (PFI$HorasActividadAcademica -
min(PFI$HorasActividadAcademica))/(max(PFI$HorasActividadAcademica) -
min(PFI$HorasActividadAcademica))
PFI$HorasClase <- (PFI$HorasClase - min(PFI$HorasClase))/(max(PFI$HorasClase) -
min(PFI$HorasClase))
PFI$PublicacionProduccionCientifica2 <- (PFI$PublicacionProduccionCientifica2 -
min(PFI$PublicacionProduccionCientifica2))/(max(PFI$PublicacionProduccionCientifi
ca2) - min(PFI$PublicacionProduccionCientifica2))
PFI$PublicacionesRegionalRevista3 <- (PFI$PublicacionesRegionalRevista3 -
min(PFI$PublicacionesRegionalRevista3))/(max(PFI$PublicacionesRegionalRevista3) -
min(PFI$PublicacionesRegionalRevista3))
PFI$PublicacionesLibro <- (PFI$PublicacionesLibro -
min(PFI$PublicacionesLibro))/(max(PFI$PublicacionesLibro) -
min(PFI$PublicacionesLibro))
PFI$PubCapituloLibro <- (PFI$PubCapituloLibro -
min(PFI$PubCapituloLibro))/(max(PFI$PubCapituloLibro)-
min(PFI$PubCapituloLibro))
PFI$PublicacionesPonencia <- (PFI$PublicacionesPonencia -
min(PFI$PublicacionesPonencia))/(max(PFI$PublicacionesPonencia)-
min(PFI$PublicacionesPonencia))
#cargar el conjunto de datos del iris
ind <- sample(2, nrow(PFI), replace = TRUE, prob = c(0.8,0.2))
trainset = PFI[ind==1,]
testset = PFI[ind==2,]
#Mostrar Entrenamiento
trainset
nrow(testset)
#Mostrar Pruebas
testset
```

```
nrow(trainset)
#Instalar y cargar el packete Neural Net
library(neuralnet)
set.seed(333)
#agregue tres columnas SI PUBLICA O NO
trainset$NO = trainset$TienePublicaciones == "NO"
trainset$SI = trainset$TienePublicaciones == "SI"
#Entrenar la red neuronal usando la funcion neuralnet con tres neuronas ocultas en cada capa
neurona = neuralnet( SI + NO ~ EstadoCivil + Genero + NivelInstruccion + TieneHijos + NoEventosNacionales + NoEventosInternacionales+ HorasActividadAcademica + HorasClase + PublicacionProduccionCientifica2 + PublicacionesRegionalRevista3 + PubCapituloLibro + PublicacionesLibro + PublicacionesPonencia,
                data=trainset,
                hidden = 9,
                linear.output = FALSE
                , threshold=0.03,
                algorithm = "rprop+")
plot(neurona)
# Predicción. Se crea un data frame con las probabilidades y los nombres de las especies

prediccion <- data.frame(predict(neurona, trainset), TienePublicaciones=predict(neurona,trainset, type="response"))
prediccion
# Matriz de confusión
mc <- table (trainset$TienePublicaciones,trainset$TienePublicaciones, dnn = c("Asignado","Real"))
mc
#Mostrar Detalles de la Red
neurona
plot(neurona)
neurona$weights
neurona$net.result
neurona$result.matrix
```

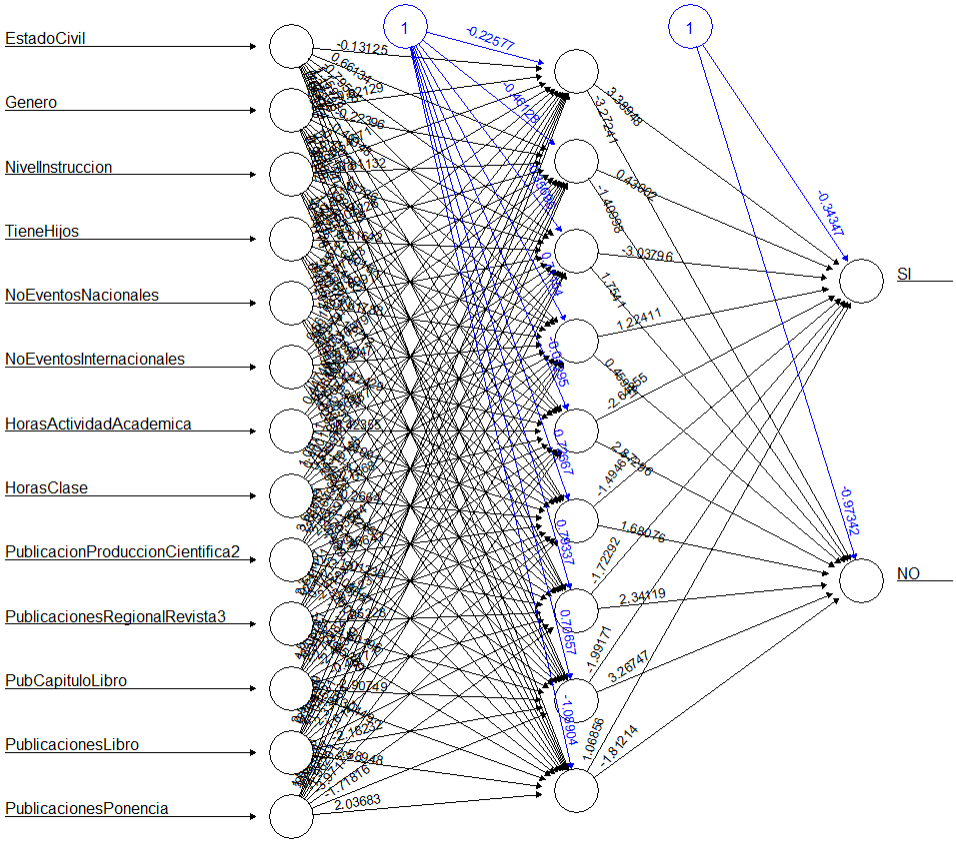



*Ilustración 45. ANN en R Studio*

**ANEXO 8:** Evaluación de los resultados

Result History | AttributeWeights (Weight by Chi Squared Statistic) | ExampleSet (Select Attributes)

Data | Statistics | Visualizations | Annotations

Open in Turbo Prep | Auto Model

Filter (3,917 / 3,917 examples): all

| EstudianteID | Ponderacio... | confidence(Excelente) | confidence(Regular) | confidence(Bueno) | prediction(PonderacionPromedio) | EstadoCivil |
|---|---|---|---|---|---|---|
| 46700 | Bueno | 0.000 | 0.000 | 1.000 | Bueno | 1 |
| 45851 | Bueno | 0.000 | 0.001 | 0.999 | Bueno | 1 |
| 45768 | Regular | 0.000 | 0.992 | 0.007 | Regular | 1 |
| 45767 | Bueno | 0.000 | 0.000 | 1.000 | Bueno | 1 |
| 45674 | Bueno | 0.000 | 0.279 | 0.721 | Bueno | 1 |
| 45523 | Bueno | 0.000 | 0.000 | 1.000 | Bueno | 1 |
| 45243 | Bueno | 0.000 | 0.007 | 0.993 | Bueno | 1 |
| 45216 | Regular | 0.000 | 0.998 | 0.002 | Regular | 1 |
| 45147 | Regular | 0.000 | 0.995 | 0.005 | Regular | 1 |
| 45087 | Bueno | 0.000 | 0.000 | 1.000 | Bueno | 1 |
| 45082 | Regular | 0.000 | 0.998 | 0.002 | Regular | 1 |
| 45015 | Regular | 0.000 | 0.998 | 0.002 | Regular | 1 |
| 44980 | Excelente | 0.999 | 0.000 | 0.001 | Excelente | 1 |
| 44934 | Regular | 0.000 | 0.998 | 0.002 | Regular | 1 |
| 44907 | Bueno | 0.054 | 0.000 | 0.946 | Bueno | 1 |
| 44888 | Excelente | 1.000 | 0.000 | 0.000 | Excelente | 1 |
| 44881 | Regular | 0.000 | 0.997 | 0.002 | Regular | 1 |

ExampleSet (3,917 examples, 6 special attributes, 10 regular attributes)

*Ilustración 46. Respuestas calificadas para construir el gráfico de elevación*