UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS Y COMPUTACIÓN

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

TRABAJO DE TITULACIÓN

ANÁLISIS DE EXACTITUD DE LOS MÉTODOS DE REGRESIÓN APLICADOS EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH.

Autora:

Thalia Maricela Veloz Chunata

Tutor:

MsC. Jorge Delgado

Riobamba – Ecuador

2019

PÁGINA DE ACEPTACIÓN

Los miembros del Tribunal de Graduación del proyecto de investigación de título: **"ANÁLISIS DE EXACTITUD DE LOS MÉTODOS DE REGRESIÓN APLICADOS EN LA BASE DE DATOS DEL SISTEMA ACADÉMICO DE LA UNACH",** presentado por la Srta. Thalia Maricela Veloz Chunata y dirigida por: MsC. Jorge Delgado.

Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación con fines de graduación escrito en el cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la UNACH.

Para constancia de lo expuesto firman:

MsC. Jorge Delgado

Tutor del proyecto **Firma**

PhD. Lida Barba

Miembro del Tribunal **Firma**

MsC. Lady Espinoza

Miembro del Tribunal **Firma**

DERECHOS DE AUTORÍA

La responsabilidad del contenido de este proyecto de graduación corresponde exclusivamente a la Srta. Thalia Maricela Veloz Chunata bajo la dirección del MsC. Jorge Delgado y al patrimonio intelectual de la Universidad Nacional de Chimborazo.

Autor:

Thalia Maricela Veloz Chunata

060405394-2

DEDICATORIA

El presente proyecto de investigación está dedicado a Dios, quien siempre me acompaña, me guía y me permitió seguir adelante, dándome salud y vida para concluir con una meta más. A mis padres que siempre me han apoyado en muchos de mis logros, en las buenas y en malas, que me inculcaron valores y reglas para ser una mejor persona con ciertas libertades que me motivaron a alcanzar mis sueños. A mi hermana que también estuvo presente durante las situaciones que más la he necesitado para culminar con mi etapa universitaria. A mis amigos de la carrera con quienes compartí varios momentos inolvidables y fueron unos excelentes compañeros de trabajo.

Thalia Maricela Veloz Chunata

AGRADECIMIENTO

Agradezco a Dios por acompañarme y protegerme todos los días durante mi camino, a mi familia y amigos por el apoyo que siempre recibí por parte de ellos. A la escuela de Sistemas y Computación de la Universidad Nacional de Chimborazo, que me abrieron sus puertas para continuar con mis estudios, en donde adquirí nuevas experiencias y me formé profesionalmente. A mis docentes de la carrera que compartieron varios de sus conocimientos y fueron un gran complemento en mi vida universitaria. Al docente MsC. Cristian Morales que fue una gran guía en la culminación del presente proyecto de investigación. Gracias a todas aquellas personas que siempre me apoyaron para lograr que este sueño se haga realidad.

Thalia Maricela Veloz Chunata

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN

La Universidad Nacional de Chimborazo (UNACH) a través de la Coordinación de Desarrollo de Sistemas Informáticos (CODESI) gestiona información en el Sistema de Control Académico (SICOA), la cual almacena grandes volúmenes de datos y se desconoce el potencial que esta información puede albergar en la toma de decisiones, por lo que en esta investigación se hace uso de la minería de datos para intentar explicar el comportamiento de ciertas variables académicas a partir de los datos personales de estudiantes y docentes, por medio de los modelos de regresión.

Con la finalidad de encontrar el mejor modelo de regresión, se realizó el análisis de exactitud de los métodos de regresión lineal y polinomial a partir de los datos cuantitativos de estudiantes y docentes. El proceso de minería de datos se realizó a través la metodología CRISP-DM. Se diseñaron y analizaron veinte modelos, que involucraron las variables independientes y dependientes y sus correlaciones. Al aplicar los modelos, fue evaluada su exactitud mediante las métricas: raíz del error cuadrático medio (RMSE), error absoluto (AE), coeficiente de correlación (R) y coeficiente de determinación (R^2).

De acuerdo con los análisis de exactitud realizados se determinó que los veinte modelos generados con los datos de estudiantes y docentes revelaron correlaciones débiles y considerables, errores altos y bajos; pero el que más se ajusta a los datos con mayor exactitud es el método de regresión lineal. Con estos resultados obtenidos, los directivos de la institución podrán tomar decisiones respecto al proyecto: "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

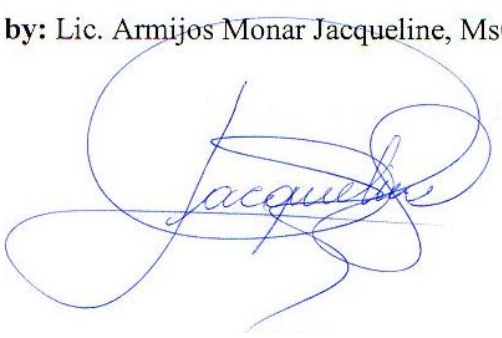
Palabras clave: metodología CRISP-DM, regresión lineal, regresión polinomial, medidas de exactitud.

ABSTRACT

The National University of Chimborazo (UNACH) through Computer Systems Development Coordination (CODESI) manages information at the Academic Control System (SICOA). It stores a big deal of potential data. The potentiality of the information for making decision is unknown. Therefore, this research has taken into account data mining usage in order to explain the behavior of certain academic variables, based on the personal data of students and professors through regression models. In order to find the best regression model, an accurate analysis of the linear and polynomial regression methods was performed. It was based on the quantitative data of students and professors. The data mining process was carried out through the CRISP-DM methodology. Twenty models were designed and analyzed, which involved independent and dependent variables and their correlations. When applying the models, their accuracy was evaluated using metrics: Root Mean Square Error (RMSE), Absolute Error (AE), Correlation Coefficient (R) and determination coefficient R-Squared (R^2). According to the accuracy analyzes performed, it was determined that the twenty models generated by using the data of students and professors revealed weak and considerable correlations, high and low errors; but the best one that fits data accurately is the linear regression method. By using the obtained results, the administrators of the institution will be able to make decisions regarding a project: "Designing strategies for continuous improvement in the academic and research management of UNACH by using data mining".

Keywords: CRISP-DM methodology, linear regression, polynomial regression, accuracy measures.

Review and corrected by: Lic. Armijos Monar Jacqueline, MsC.



INTRODUCCIÓN

En minería de datos se utilizan diversas técnicas para analizar información, mismas que descienden de la inteligencia artificial y la estadística, las que se aplican sobre un conjunto de datos para obtener resultados y descubrir patrones. El proceso de minería de datos se está convirtiendo en un elemento importante para las instituciones académicas, debido al soporte que ofrece en la toma de decisiones (Menes, Arcos, Moreno & Gallegos, 2015).

Diversas dificultades en el descubrimiento de patrones y aprendizaje automático han dado lugar al progreso de las técnicas de minería de datos, una de ellas son las predictivas que permiten obtener pronósticos de comportamientos futuros a partir de los datos recopilados (Aranda & Solotongo, 2013). Por ello los métodos de regresión se utilizan para establecer relaciones entre una variable dependiente y una o más variables independientes, con el fin de construir un modelo de regresión ya sea para fines explicativos o predictivos.

Existen algunos métodos de regresión y los que se destacan por su facilidad de aplicación e interpretación son: regresión lineal y regresión polinomial. También se encuentra diversas técnicas para validar los métodos de regresión, como son la comparación de los parámetros obtenidos mediante modelos físicos teóricos o el uso de técnicas de validación cruzada (Pérez, Delegido, Rivera & Verrelst, 2015); por medio de la validación cruzada se evalúan los resultados de un análisis estadístico, se utiliza en entornos donde el propósito primordial es estimar la exactitud de un modelo que se llevará a la práctica. Para evaluar la exactitud de un método de regresión, las medidas típicas son: raíz del error cuadrático medio (RMSE), error absoluto (AE), coeficiente de correlación (R) y coeficiente de determinación (R^2).

Por otro lado, existen algunos trabajos relacionados con el tema de investigación, donde las técnicas de minería de datos están siendo utilizadas para el análisis de información en diferentes áreas como: finanzas, marketing, medicina, entre otras; incluso en los últimos años la minería de datos está teniendo un gran impacto en el área educativa (García, 2016).

La mayoría de estudios se enfocan en aspectos socioeconómicos, personales y académicos; también se han centrado en hallar los factores que influyen en la excelencia académica y buscan aportar una mejora en el sistema educativo (Porcel, Dapozo & López, 2010). Sus resultados han permitido identificar factores que favorecen o limitan el desempeño académico al tratar de descubrir el comportamiento de variables.

En este trabajo se analiza la exactitud de los métodos de regresión: lineal y polinomial, aplicados a los datos personales y académicos de estudiantes y docentes de la base de datos del Sistema de Control Académico (SICOA) de la Universidad Nacional de Chimborazo, teniendo como variables dependientes al rendimiento académico y la evaluación final del docente. Se ha elegido la metodología CRISP-DM, la cual actualmente es una de las más utilizadas en proyectos de minería de datos y la herramienta software RapidMiner para el análisis de los métodos de regresión mencionados.

Este trabajo de investigación se organiza de la siguiente manera: en el Capítulo I se describe el problema, justificación y objetivos, en el Capítulo II se desarrolla el marco teórico, en el Capítulo III se describe la metodología de investigación, en el Capítulo IV se presenta los resultados y discusión, y finalmente conclusiones, recomendaciones y referencias bibliográficas.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1.Problema y Justificación

Gutiérrez y Molina (2015) mencionan que muchas organizaciones generan grandes cantidades de información. En muchos escenarios son tan voluminosas y complejas de analizar, que varias instituciones educativas no cuentan con personal especializado y tiempo para realizar esta labor; incluso desconocen el potencial que esta información puede albergar en la toma de decisiones, por lo que surge la necesidad de la ayuda de técnicas de minería de datos para analizar información y extraer conocimiento (Gutiérrez & Molina, 2015).

La UNACH a través de CODESI gestiona la información académica en el SICOA, el que cuenta con gran cantidad de información almacenada en una base de datos que contiene información personal y académica de estudiantes y docentes. Sin embargo, no existe la aplicación de herramientas de minería de datos efectivas para la generación de conocimiento y que a la vez contribuyan al desarrollo del proyecto: "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

Por medio de los métodos de regresión lineal y polinomial es factible realizar el análisis de los datos, mismos que relacionan una o más variables predictoras y una variable de respuesta; son utilizados para la predicción, explicación o previsión en diversas áreas relacionadas con la ciencia y la ingeniería. Estos métodos han alcanzado gran importancia en muchos entornos como: biología, finanzas, procesos industriales, policiales y políticos; sin embargo, en la gestión académica de la educación superior son escasas, lo que dificulta la toma de decisiones (Aranda & Solotongo, 2013).

En algunas instituciones de educación superior ha recibido una especial atención en encontrar los mejores modelos generados (García, Alvarado & Jiménez, 2000). Por ello, los resultados obtenidos de los métodos de regresión deben ser validados en base a su exactitud, los cuales pueden ser evaluados mediante métricas tradicionales tales como: RMSE, AE, R y R^2.

El objetivo de esta investigación es analizar la exactitud de los métodos de regresión lineal y polinomial aplicados en la base de datos del SICOA de la UNACH, partiendo de datos cuantitativos personales de estudiantes y docentes, así como también académicos para evaluar el comportamiento de las variables en los siguientes ámbitos: rendimiento académico de estudiantes y evaluación docente; lo cual apoyará a la toma de decisiones de los directivos de la institución.

1.2.Objetivos

1.2.1. Objetivo General

Identificar el mejor método de regresión aplicado en la base de datos del Sistema Académico de la UNACH por medio del análisis comparativo de la exactitud, para apoyar al proyecto: "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

1.2.2. Objetivos Específicos

- Utilizar la metodología CRISP-DM para el análisis, preparación y construcción de los datos personales y académicos de estudiantes y docentes.
- Aplicar los métodos de regresión lineal y polinomial para determinar las correlaciones entre los datos cuantitativos de estudiantes y docentes.
- Analizar los resultados de las medidas de exactitud: RMSE, AE, R y R^2 de los métodos de regresión lineal y polinomial para la determinación del mejor método.

CAPÍTULO II

MARCO TEÓRICO

2.1.Minería de datos

La minería de datos o exploración de datos es un proceso que permite obtener patrones, tendencias, relaciones y conocimiento útil a partir de los datos recopilados para ayudar a la toma de decisiones.

El objetivo principal del proceso consiste en extraer información de un gran conjunto de datos que no es descubierta a simple vista y transformarla en un modelo entendible para su uso posterior, utilizando métodos de la inteligencia artificial, estadística, aprendizaje automático y visualización gráfica, dependiendo del tipo de entorno al cual se aplica el proceso (Aranda & Solotongo, 2013).

2.2.Metodologías para minería de datos

Según Moine (2013) algunas de las metodologías dominantes son:

- **Catalyst:** Es una metodología para el proceso de extracción de conocimiento en bases de datos. Está formada por dos partes: metodología para el modelado del negocio y metodología para la minería de datos.
- **CRISP-DM:** (CRoss Industry Satandard Process for Dara Mining) Es un proceso jerárquico que propone seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Cada fase se divide en un conjunto de actividades genéricas de segundo nivel.

- **KDD:** (Knowdlege Discovery in Databases) Es el primer modelo que define el descubrimiento de conocimiento en bases de datos como un proceso iterativo e interactivo con cinco fases: selección, preprocesamiento, transformación, minería de datos y evaluación e implementación.
- **SEMMA:** (Sample, Explore, Modify, Model, Assess) Es un proceso de selección, explotación y modelamiento de grandes conjuntos de datos para descubrir patrones. El nombre corresponde a las cinco fases: muestreo, exploración, modificación, modelado y evaluación.

Para desarrollar el trabajo de investigación se eligió la metodología CRISP-DM, la cual es una de las más usadas en proyectos de análisis de datos que tiene la mayor aceptación en la última encuesta de KDNuggets (Piatetsky, 2014).

2.2.1. Fases de CRISP-DM

CRISP-DM es una metodología estándar para trabajos de minería de datos que incluye un modelo y una guía, la cual se desarrolla en seis fases no esencialmente rígidas que funcionan de manera cíclica e iterativa, donde cada fase se subdivide con tareas específicas. A continuación, se describe las seis fases (Cortina, 2015):

Comprensión del negocio: Se determinan los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el propósito de definir el problema de minería y el plan de trabajo. Las tareas de esta fase son:

- Determinar los objetivos del negocio
- Evaluación de la situación
- Determinar los objetivos de DM
- Realizar el plan de proyecto

Comprensión de los datos: Consiste en la recolección inicial de datos que se utilizarán en el proyecto, determinar su calidad y familiarizarse con ellos para que ayuden a concretar las primeras hipótesis. Las tareas de esta fase son:

- Recolectar los datos iniciales
- Descripción de los datos
- Exploración de los datos
- Verificar la calidad de los datos

Preparación de los datos: Comprende aquellas tareas de tratamiento de los datos para elaborar un conjunto final de datos sobre el cual se aplicarán las técnicas de minería de datos. Las tareas de esta fase son:

- Seleccionar los datos
- Limpiar los datos
- Construir los datos
- Integrar los datos
- Formateo de los datos

Modelado: Se elige las técnicas de minería de datos que sean más convenientes para resolver el problema. Las tareas de esta fase son:

- Escoger la técnica de modelado
- Generar el plan de prueba
- Construir el modelo
- Evaluar el modelo

Evaluación: Se evalúa el modelo y se analiza los resultados obtenidos en función de los objetivos del negocio. Las tareas de esta fase son:

- Evaluar los resultados
- Revisar el proceso
- Determinar los próximos pasos

Implementación: Se transforma el conocimiento adquirido en acciones dentro del proceso de negocio, por ejemplo, cuando se recomienda acciones basadas en la observación del modelo y sus resultados, también aplicando el modelo a otros conjuntos de datos. Por lo general un proyecto de minería de datos no concluye en esta fase, porque se debe documentar y mostrar los resultados de forma entendible al usuario. Las tareas de esta fase son:

- Planear la implementación
- Planear la monitorización y mantenimiento
- Producir el informe final
- Revisar el proyecto

2.3.Herramientas para minerías de datos

Existen diversas herramientas tecnológicas que sirven de apoyo para la minería de datos y ayudan a la toma de decisiones a las organizaciones, algunas de las más populares son:

- **R:** Es un software gratuito y de código abierto, desarrollado por Robert Gentleman y Ross Ihaka. Permite realizar análisis de datos, mostrar cálculos estadísticos y gráficas (Oviedo, Oviedo, & Vélez, 2015).

- **RAPIDMINER:** Es un software gratuito distribuido bajo licencia GPL, desarrollado en Java. Provee de una interfaz gráfica con más de 500 operadores orientados al análisis de datos, preprocesamiento de datos, métodos de entrenamiento, prueba de modelos, visualización de datos, etc., incluso permite utilizar los algoritmos comprendidos en weka (Jaramillo & Paz, 2015).
- **SPSS MODELER:** Es un software de pago desarrollada por SPSS Inc., una compañía de IBM. Contiene herramientas que permiten realizar tareas de minería de datos como: el análisis de datos, desarrollar modelos predictivos, visualización de datos, etc. (Cortina, 2015).
- **WEKA:** Es un software libre que se distribuye bajo licencia GNU-GLP, desarrollado en Java. Provee de una interfaz gráfica, una colección de métodos para el análisis de datos y herramientas de visualización de datos (Cortina, 2015).

La herramienta elegida para desarrollar el trabajo de investigación es RapidMiner, la cual es una plataforma de análisis de datos usada ampliamente y la más popular según la encuesta de KDNuggets (Piatetsky, 2014).

2.4.Técnicas de minerías de datos

Existen diferentes técnicas de minería de datos, pero para su elección se debe tomar en cuenta el tipo de variables y el objetivo de minería de datos; estas técnicas tienen como meta primaria la predicción de datos desconocidos y la descripción de patrones. Generalmente se pueden agrupar de la siguiente manera (Oviedo et al., 2015):

- **Técnicas supervisadas o predictivas:** Utilizadas para el análisis predictivo. Por ejemplo: Métodos de Regresión, Árboles de Decisión, Redes Neuronales, etc.

- **Técnicas no supervisadas o descriptivas:** Utilizadas para el análisis descriptivo. Por ejemplo: Método Jerárquico, Clustering, Reglas de Asociación, etc.

2.5.Regresión

Es un método para investigar y modelar la relación entre variables, dependiente e independiente, para que, a partir de un atributo de entrada o un valor de predicción, obtener el valor estimado o de salida de acuerdo con un valor de error permitido (Medina & Gómez, 2014).

2.5.1. Tipos de regresión

Según Vinuesa (2016) existen muchas variantes especializadas de regresión, algunas de ellas son:

- **Lineal simple:** predicción de una variable de respuesta cuantitativa a partir de una variable predictora cuantitativa.
- **Polinomial:** predicción de una variable de respuesta cuantitativa a partir de una variable predictora cuantitativa, donde la relación se modela como una función polinomial de orden n.
- **Lineal múltiple:** predicción de una variable de respuesta cuantitativa a partir de dos o más variables predictoras cuantitativas.
- **Multivariada:** predicción de más de una variable de respuesta cuantitativa a partir de una o más variables predictoras cuantitativas.
- **Logística de Poisson:** predicción de una variable categórica a partir de una o más predictoras y predicción de una variable de respuesta que representa un conteo a partir de una o más predictoras.

- **No lineal:** predicción de una variable de respuesta cuantitativa a partir de una o más predictoras, donde el modelo no es lineal.
- **Robusta:** predicción de una variable de respuesta cuantitativa a partir de una o más predictoras, usando una aproximación resistente al efecto de observaciones influyentes.

2.5.2. Regresión lineal simple

Se trata de una técnica que analiza la relación entre dos variables cuantitativas, de tal manera se pueden hacer predicciones sobre los valores de la variable Y, a partir de los valores de X, es decir, su objetivo es explicar el comportamiento de una variable dependiente (respuesta o Y), a partir de una variable independiente (predictora o X) (Laguna, 2014). Se basa en modelos lineales con la fórmula general:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

donde:

- **Y:** variable dependiente
- **X:** variable independiente
- **β_0, β_1**: coeficientes de regresión
- **ε:** error del modelo

2.5.3. Regresión polinomial

Los métodos de regresión polinomial son aplicables cuando la dependencia entre la variable de respuesta cuantitativa continua y la variable predictora cuantitativa muestran un comportamiento curvilíneo o no lineal, es decir la relación se modela como una función polinomial de orden k (Astorga, 2014):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon$$

2.6.Validación cruzada

La validación cruzada (cross validation) es un operador que se utiliza para estimar el rendimiento estadístico de un modelo de aprendizaje, especialmente para validar modelos predictivos. Consiste en tomar 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).

Para evaluar los métodos de regresión se pueden utilizar las siguientes medidas: error cuadrado medio raíz, error absoluto, correlación y correlación al cuadrado.

- **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).
- **error absoluto** (AE): es la diferencia entre el valor real y el valor aproximado (RapidMiner, 2019).
- **correlación** (R): el coeficiente de correlación o R es un método para evaluar una posible relación entre dos variables continuas. Puede tomar un valor en el rango de -1 a +1. Un coeficiente de correlación de cero indica que no existe una relación entre variables, mientras que un coeficiente de correlación de -1 o +1 indica una relación perfecta. Cuanto más fuerte es la correlación, más se aproxima el coeficiente de correlación a ± 1. Si el coeficiente es un número positivo quiere decir que a medida que aumenta el valor de una variable, el valor de la otra también tiende a hacerlo. Si, por otro lado, el coeficiente es un número negativo quiere decir que a medida que aumenta el valor de una variable, el valor de la otra tiende a disminuir (Mukaka, 2012).

- **correlación al cuadrado** (R^2): el coeficiente de determinación o R^2 es la porción de variabilidad de Y que queda explicada por su dependencia de la variable X (Dagnino, 2014). R^2 es una medida adimensional, debido a su recorrido acotado entre 0 y 1; un $R^2 = 1$ significa un ajuste lineal perfecto, muestra que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media y un $R^2 = 0$ muestra que el modelo no explica la variabilidad de los datos de respuesta en torno a su media. Mientras más se aproxime a 1 el R^2, mejor se ajustará el modelo a los datos (Martínez, 2005).

CAPÍTULO III

METODOLOGÍA

3.1.Tipo de Investigación

El desarrollo de esta investigación tuvo un enfoque cuantitativo debido a que el proceso es secuencial y se utiliza datos cuantitativos para relacionar variables, además es utilizado en casos donde se realiza mediciones numéricas mediante el análisis estadístico con el fin de establecer modelos de comportamiento (Hernández, Fernández & Baptista, 2014). Por tratarse de minería de datos, no existe hipótesis preconcebidas de la investigación y como metodología de minería de datos se implementa un proceso basado en CRISP-DM, que consta de seis fases: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación e implementación (Castorena, Silva, Domínguez & Rodríguez, 2018). Esta investigación sigue el siguiente proceso metodológico:

1. **Revisión y búsqueda de información:** la revisión bibliográfica se realizó por medio de la investigación documental a través de tesis, artículos, libros, etc.; la cual permitió tener una visión global de los estudios relacionados con el tema de investigación, como es acerca de los métodos de regresión, metodologías y herramientas para minería de datos. En este caso también se utilizó como fuente de información la base de datos del SICOA de la UNACH.
2. **Recolección y análisis de datos:** la información se recolectó de la base de datos del SICOA de la UNACH en un archivo .xlsx y se analizó la información personal del estudiante, el rendimiento del estudiante, la información personal del docente, la información académica del docente y la evaluación final del docente, dicha información obtenida se encuentra almacenada en la base de datos desde el año 2012 hasta la actualidad.

Se utilizó el método de investigación analítico y sintético, teniendo como resultado la emisión de una interpretación, descubriendo nuevos significados sobre el tema y obteniendo nuevos conocimientos, que permitirá sintetizar el comportamiento del fenómeno de estudio.

3. **Preparación de los datos:** se realizó a través de la investigación exploratoria debido a que se efectuó la exploración, limpieza, construcción, integración y calidad de datos; se seleccionaron solo los campos cuantitativos necesarios para los distintos análisis. Los datos fueron analizados mediante la herramienta Talend Data Quality Online.
4. **Implementación y evaluación de los métodos:** fueron implementados los métodos de regresión lineal y regresión polinomial. Esta investigación es observacional debido a que se analiza la correlación entre las variables cuantitativas de estudiantes y docentes de la base de datos del SICOA de la UNACH, a través del uso de la herramienta RapidMiner, con la finalidad de estimar el comportamiento y las relaciones entre las variables de estudio y evaluar los modelos generados mediante las siguientes métricas: error cuadrado medio raíz, error absoluto, correlación y correlación al cuadrado. Además, se hará uso de un análisis descriptivo debido a que se observará el comportamiento del fenómeno a través de tablas y gráficos. Para explicar los resultados encontrados durante esta investigación se lo realizará por medio del informe de minería de datos.
5. **Inducción científica:** la investigación es aplicada debido a que se desea analizar la exactitud de los métodos de regresión lineal y polinomial aplicados en la base de datos del SICOA de la UNACH partiendo de los datos cuantitativos de docentes y estudiantes que apoyará a la toma de decisiones de los directivos de la institución.

3.2.Unidad de análisis

La investigación se enfoca en los métodos de regresión (lineal – polinomial) y en las variables cuantitativas de estudiantes y docentes, datos que fueron obtenidos de la base de datos del SICOA de la UNACH.

3.3.Técnicas de recolección de datos

Las fuentes de recopilación de información son de fuentes secundarias; la obtención de información fue a partir de artículos, documentos, textos e información almacenada en la base de datos del SICOA de la UNACH.

Los instrumentos de la investigación utilizados son: fichas de registros, observación directa y procedimientos experimentales.

3.4.Técnicas de análisis e interpretación de la información

Se analizaron los datos con la herramienta Talend Data Quality Online, donde se procedió a seleccionar solo los datos cuantitativos que fueron limpiados, construidos e integrados para aplicar los métodos de regresión lineal y polinomial. Se utilizó la herramienta RapidMiner para aplicar los métodos con los datos cuantitativos de estudiantes y docentes de la base de datos del SICOA de la UNACH; analizando su exactitud mediante la validación cruzada que contiene las siguientes medidas: RMSE, AE, R y R^2.

3.5.Aplicación de la metodología CRISP-DM

3.5.1. Comprensión del negocio

3.5.1.1.Determinar los objetivos del negocio

El objetivo principal en esta investigación es analizar la exactitud de los métodos de regresión lineal y polinomial aplicados en la base de datos del SICOA de la UNACH y determinar la relación entre variables cuantitativas de estudiantes y docentes con datos personales y académicos, para obtener resultados confiables que contribuya al proyecto de: "Diseño de estrategias de mejoramiento continuo en la gestión académica e investigativa de la UNACH, utilizando minería de datos".

3.5.1.2.Evaluación de la situación

La fuente de datos es la base de datos SQL y las tablas entregadas contienen: información personal del estudiante, rendimiento del estudiante, información personal, académica y evaluación docente desde el año 2012 hasta la actualidad de la UNACH. Se cuenta con una cantidad de datos suficientes para aplicar la minería de datos.

Se dispone de los siguientes softwares libres:

- Talend Data Quality Online
- Rapid Miner Studio 9.3

También se dispone con una laptop con las siguientes características:

- Modelo: Inspiron 5567 Signature Edition
- Procesador: Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz 2.90 GHz
- Memoria RAM: 16 GB
- Disco Duro: 1T
- Sistema operativo: Windows 10

3.5.1.3.Determinar los objetivos de DM

- Determinar las correlaciones entre las variables cuantitativas y el rendimiento académico de los estudiantes de la Universidad Nacional de Chimborazo.
- Determinar las correlaciones entre las variables cuantitativas y el resultado de la evaluación final de los docentes de la Universidad Nacional de Chimborazo.

3.5.1.4.Realizar el plan del proyecto

El plan de proyecto consta de 6 pasos con un tiempo total de 11 semanas distribuido de la siguiente manera:

Paso 1: Análisis y selección de los datos. **Tiempo:** 3 semanas

Paso 2: Análisis de las propiedades de los datos. **Tiempo:** 1 semana

Paso 3: Preparación de los datos y transformación del conjunto de datos. **Tiempo:** 2 semanas

Paso 4: Aplicación de los métodos de regresión seleccionados a los datos. Construir los modelos predictivos. **Tiempo:** 2 semanas

Paso 5: Análisis de los resultados obtenidos y extracción de conocimiento. **Tiempo:** 2 semanas

Paso 6: Presentación del informe final. **Tiempo:** 1 semana

3.5.2. Comprensión de los datos

3.5.2.1.Recolectar los datos iniciales

Los datos fueron recolectados en un archivo Excel generado desde la base de datos del SICOA de la UNACH, que se encuentran almacenados desde el periodo académico Septiembre 2012 – Marzo 2013 con información personal de estudiantes y docentes, así como también información académica, de evaluación docente se consiguió desde el periodo académico Octubre 2017 – Marzo 2018; las tablas proporcionadas son las siguientes:

- Estudiante Información Personal, Estudiante Rendimiento, Docente Información Personal, Docente Información Académica y Evaluación Docente

3.5.2.2.Descripción de los datos

En este apartado se realiza la descripción de todos los campos de las tablas proporcionadas de estudiantes y docentes, en el ANEXO 1 se puede observar: nombre, tipo y descripción de cada uno de los campos.

3.5.2.3.Exploración de los datos

Los primeros resultados de la exploración de los datos muestran información estadística para determinar su consistencia y completitud mediante gráficos de distribución. Al realizar un primer análisis se puede observar en el ANEXO 2.

3.5.2.4.Verificar la calidad de los datos

Los datos fueron proporcionados de la base de datos del SICOA de la Universidad Nacional de Chimborazo, datos que fueron analizados mediante la herramienta Talend Data Quality Online para determinar la estabilidad de los datos, como la cantidad de valores nulos, valores válidos y valores no válidos, los resultados se muestran en el ANEXO 3.

3.5.3. Preparación de los datos

3.5.3.1.Seleccionar los datos

Los campos que se han seleccionado de cada tabla son solo cuantitativos, y la razón de la exclusión de los restantes campos es por los objetivos de negocio (apartado 4.1.1) y de minería de datos (apartado 4.1.3); también ciertos campos que al momento de realizar la calidad de los datos (tabla 19), estos en su mayoría son valores nulos. Los campos seleccionados son los siguientes:

Tabla Estudiante

- Estudiante ID, Número Integrantes Hogar, Número Hermanos, Número Dependen Ingresos, Total Ingresos, Número Hijos

Tabla Estudiante Rendimiento

- Estudiante ID, Promedio

Tabla Docente

- Cédula, Número Hijos, Nº Eventos Aprobados, Nº Eventos Asistidos, Horas Eventos Aprobados, Horas Eventos Asistidos, Nº Eventos Nacionales, Nº Eventos Internacionales

Tabla Docente Información Académica

- Número Documento, Horas Actividad Académica, Horas Clase

Tabla Evaluación Docente

- Usuario Evaluado, Resultado Final

3.5.3.2.Limpiar los datos

La base de datos con la que se cuenta contiene la información necesaria para poder cumplir con los objetivos de negocio y de minería de datos, no se han encontrado un número elevado de problemas en los datos seleccionados.

Se tuvo algunos campos con datos faltantes, los cuales se rellenaron con 0 en el caso de los campos cuantitativos, también se tuvo datos atípicos y duplicados, los cuales fueron eliminados. También se eliminaron los estudiantes que no tenían promedio y los docentes que no tenían el resultado final de la evaluación docente.

3.5.3.3.Construir los datos

Para construir los datos en este apartado se toma en cuenta los siguientes campos:

Tabla 1. Construcción de datos

Tabla	Campo	Descripción de construcción
Estudiante Rendimiento	Promedio	Este campo contiene el promedio de todos los estudiantes por periodo académico y por nivel, por lo cual se crea un nuevo campo con el promedio general por cada estudiante. También se crea nuevos campos con los promedios por nivel (quinto, sexto y séptimo) por cada estudiante. Se han seleccionado solo esos 3 niveles porque desde quinto semestre por lo general los estudiantes no se retiran de sus carreras y se puede obtener datos más confiables.
Docente Información Académica	Horas Actividad Académica Horas Clase	Estos campos contienen el número de horas de actividad académica y el número de horas clase respectivamente de todos los docentes por periodo académico y por carrera, por lo cual se crea un nuevo campo con el número de horas de actividad académica general y un nuevo campo con el número de horas clase general por cada docente.
Evaluación Docente	Resultado Final	Este campo contiene el resultado final de la evaluación de todos los docentes por periodo académico, por lo cual se crea un nuevo campo con el resultado final general por cada docente. También se crea nuevos campos con los resultados de evaluación por componente (docencia, gestión e investigación) de cada docente.

3.5.3.4.Integrar los datos

Los campos extraídos y combinados a partir de las tablas relacionadas por sus respectivas claves son los siguientes:

Tabla 2. Integración de datos

Tabla	Campo	Fusión
Estudiante Rendimiento	Promedio General	Tabla Estudiante
Docente Información Académica	Horas Actividad Académica General	Tabla Docente
	Horas Clase General	
Evaluación Docente	Resultado Final General	
Promedios por niveles Estudiantes (nueva tabla)	Promedio Quinto Semestre	Campo Promedio General
	Promedio Sexto Semestre	
	Promedio Séptimo Semestre	
Evaluaciones por componentes Docentes (nueva tabla)	Evaluación Componente Docencia	Campo Resultado Final General
	Evaluación Componente Gestión	
	Evaluación Componente Investigación	

3.5.3.5.Formateo de los datos

La base de datos que se dispone después de seleccionar, limpiar y construir los datos; todos los datos son numéricos. No hubo la necesidad de transformación de datos, tampoco de cambiar el orden de los campos, ni la reordenación de los registros, ni cambiar el formato de los campos.

Dicho esto, se tiene como variables independientes y dependientes a las siguientes:

Tabla 3. Variables independientes y dependientes

	Variables Independientes	**Dato original**	**Dato calculado**	**Variable dependiente**	**Dato original**	**Dato calculado**
Estudiante	Número Integrantes Hogar	✓		Promedio General		✓
	Número Hermanos	✓				
	Número Dependen Ingresos	✓				
	Total Ingresos	✓				
	Número Hijos	✓				
	Promedio Quinto Semestre	✓				
	Promedio Sexto Semestre	✓				
	Promedio Séptimo Semestre	✓				
Docente	Número Hijos	✓		Resultado Final General de Evaluación Docente		✓
	Nº Eventos Aprobados	✓				
	Nº Eventos Asistidos	✓				
	Horas Eventos Aprobados	✓				
	Horas Eventos Asistidos	✓				
	Nº Eventos Nacionales	✓				
	Nº Eventos Internacionales	✓				
	Horas Actividad Académica	✓				
	Horas Clase	✓				
	Evaluación C. Docencia	✓				
	Evaluación C. Gestión.	✓				
	Evaluación C. Investigación	✓				

El Promedio General se calculó de los promedios que tiene por semestre y por periodo académico cada estudiante y el Resultado Final General de Evaluación Docente se calculó de las evaluaciones que tiene por componente y por periodo académico cada docente.

3.5.4. Modelado

3.5.4.1.Escoger la técnica de modelado

Para el modelado se utilizará la herramienta RapidMiner, la cual tiene incorporadas una serie de técnicas de modelado, y de acuerdo con los objetivos de negocio y de minería de datos que se menciona en el apartado 4.1 (Comprensión del negocio), se debe utilizar algunos de los modelos de regresión que posee esta herramienta. De todos los modelos de regresión que nos ofrece RapidMiner, los que mejor se adaptan a los objetivos son: regresión lineal y regresión polinomial

3.5.4.2.Generar el plan de prueba

Para probar la validez y la calidad de los modelos se utilizará la validación cruzada que consiste en tomar los datos originales y crear dos subprocesos: uno de entrenamiento y otro de prueba; utilizando las siguientes medidas de exactitud: error cuadrado medio raíz (RMSE), error absoluto (AE), correlación (R) y correlación al cuadrado (R^2); métricas que son generadas gracias a los operadores que posee la herramienta RapidMiner.

3.5.4.3.Construir el modelo

En este apartado, se tiene los datos necesarios y preparados para generar los modelos. Debido a que se han definido dos objetivos de minería de datos, en el ANEXO 4, se describe la parametrización, descripción y ejecución de los modelos con los métodos de regresión lineal y polinomial.

3.5.4.4.Evaluar el modelo

Para evaluar los modelos se toma en cuenta los valores de las diferentes métricas que se establecieron en el apartado 4.4.2 (Generar el plan de prueba), evaluaciones que se pueden observar en el ANEXO 5.

3.5.5. Evaluación

3.5.5.1.Evaluar los resultados

Según los objetivos de negocio y en base a los resultados obtenidos de las métricas mediante la herramienta RapidMiner, a continuación, se hace un análisis de exactitud de los métodos de regresión aplicados en los distintos modelos.

Modelos para el objetivo 1

Los ajustes de los modelos con regresión lineal y polinomial para este objetivo no son óptimos, los valores del RMSE están en el rango 0,991 y 1,000, y los valores de AE entre 0,679 y 0,680, los cuales son errores altos que no se acercan a 0; también existe correlaciones débiles con valores de R entre 0,000 y 0,153, en cuanto a valores de R^2 están entre 0,000 y 0,024, los que representan una mínima explicación del fenómeno.

Modelos para el objetivo 2

Los ajustes de los modelos con regresión lineal y polinomial para este objetivo tampoco son óptimos, los valores del RMSE están entre 0,918 y 0,932, AE entre 0,498 y 0,518; como en el caso anterior, existen correlaciones débiles con valores de R entre 0,000 y 0,185 y R^2 entre 0,000 y 0,055, lo que representa una mínima explicación del fenómeno.

Modelos adicionales estudiantes

Los ajustes de estos modelos con regresión lineal y polinomial son óptimos, con valores RMSE entre 0,384 y 0,522, y valores de AE entre 0,297 y 0,652; en cuanto a correlaciones son considerables con valores de R entre 0,507 y 0,770 y de R^2 entre 0,267 y 0,594.

Modelos adicionales docentes

Los ajustes de estos modelos con regresión lineal y polinomial son mejores que los objetivos 1 y 2, con valores RMSE entre 0,696 y 0,931 y AE entre 0,475 y 0,517, los cuales no son errores muy altos; también existe correlaciones débiles y considerables con valores de R entre 0,150 y 0,655 y de R^2 entre 0,029 y 0,457.

3.5.5.2.Revisar el proceso

Los resultados obtenidos se analizaron de manera detallada de cada uno de los modelos generados para verificar que estén cumpliendo con el objetivo de negocio; los resultados logrados para los modelos propuestos para el objetivo 1 y 2 fueron deficientes para las medidas de exactitud: RMSE, AE, R y R^2, por lo que se ha considerado modelos adicionales para los objetivos de negocio, consiguiendo así mejores resultados, los cuales presentaron correlaciones considerables y errores más pequeños.

3.5.5.3.Determinar los próximos pasos

El siguiente paso por realizar en este proyecto es presentar los resultados obtenidos de acuerdo con los objetivos de negocio y minería de datos.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

5.1. Resultados

Durante el desarrollo de esta investigación se ha considerado cada una de las fases de la metodología CRISP-DM: compresión del negocio, comprensión de datos, preparación de datos, modelado y evaluación, donde para la segunda fase, la información fue proporcionada en un archivo Excel con las siguientes tablas:

Tabla 4. Información proporcionada del SICOA

Tabla	Número Registros	Tiempo
Estudiante Información Personal	16008	Desde el periodo académico Septiembre 2012 – Marzo 2013
Estudiante Rendimiento	87105	
Docente Información Personal	4097	
Docente Información Académica	19335	
Evaluación Docente	15276	Desde el periodo académico Octubre 2017 – Marzo 2018

Una vez seleccionado, limpiado, construido e integrado los datos en la fase 3 (preparación de datos) se trabajó con las siguientes tablas: información personal y académica de estudiantes (15793 registros) y docentes (448 registros). A continuación, se muestra los resultados obtenidos de las medidas de exactitud de los modelos generados y evaluados con los métodos de regresión lineal y polinomial.

Medidas de exactitud para la tabla Estudiante

En este parte se muestra los resultados obtenidos de las medidas de exactitud de los modelos generados con los métodos de regresión lineal y polinomial aplicados en los datos cuantitativos personales y académicos de estudiantes.

Variable dependiente: Promedio General

Tabla 5. Modelos para estudiante - Medidas de exactitud

	Modelo 1 RL/RP	**Modelo 2** RL/RP	**Modelo 3** RL/RP	**Modelo 4** RL/RP	**Modelo 5** RL/RP
Variables independientes	Número Hermanos	Número Hijos	Número Integrantes Hogar	Número dependen Ingresos	Total Ingresos
RMSE	0,988/0,991	0,998/0,998	1,000/1,000	0,993/0,995	1,000/1,000
AE	0,679/0,679	0,680/0,680	0,679/0,679	0,679/0,679	0,679/0,679
R	0,153/0,125	0,063/0,063	0,017/0,019	0,112/0,099	0,000/0,009
R^2	0,024/0,018	0,004/0,004	0,000/0,001	0,013/0,011	0,000/0,000

RL: Regresión Lineal, **RP:** Regresión Polinomial

Tabla 6. Promedio Medidas de Exactitud-Estudiante

Métodos	**Promedio RMSE**	**Promedio AE**	**Promedio R**	**Promedio R^2**
Regresión Lineal	0,995	0,679	0,069	0,008
Regresión Polinomial	0,996	0,679	0,063	0,006

Análisis: De acuerdo a las medidas de exactitud de la tabla 5 se asume como mejor al modelo 1 y como peor al modelo 5, es decir que el número de hermanos se relaciona más con el promedio general del estudiante y el total de ingresos se relaciona menos. También se calculó un promedio general de las medidas de exactitud (tabla 6) analizando las más importantes: R y RMSE; para estos 5 modelos se tiene correlaciones débiles (RL: 0,069/RP: 0,063) y no son factibles para realizar predicciones confiables porque los valores de RMSE son altos (RL: 0,995/RP: 0,996), pero se puede deducir que las variables independientes influyen mínimamente en el promedio general de los estudiantes; y el que más se ajusta a estos datos con mayor exactitud es el método de regresión lineal con un R de 0,069 y un RMSE de 0,995, en comparación con el método de regresión polinomial que tiene un R de 0,063 y un RMSE de 0,996.

También se ha considerado los siguientes modelos solo con el rendimiento académico del estudiante por semestre en relación con el promedio general final.

Variable dependiente: Promedio General

Tabla 7. Modelos para estudiante por semestre - Medidas de exactitud

	Modelo 1 RL/RP	**Modelo 2** RL/RP	**Modelo 3** RL/RP
Variables independientes	Promedio Quinto Semestre	Promedio Sexto Semestre	Promedio Séptimo Semestre
RMSE	0,522/0,869	0,384/0,701	0,391/0,717
AE	0,392/0,652	0,297/0,544	0,303/0,558
R	0,516/0,507	0,770/0,704	0,762/0,633
R^2	0,274/0,267	0,594/0,515	0,582/0,464

RL: Regresión Lineal, **RP:** Regresión Polinomial

Tabla 8. Promedio Medidas de Exactitud-Estudiante por semestre

Métodos	**Promedio RMSE**	**Promedio AE**	**Promedio R**	**Promedio R^2**
Regresión Lineal	0,432	0,330	0,682	0,483
Regresión Polinomial	0,762	0,584	0,614	0,415

Análisis: De acuerdo a las medidas de exactitud de la tabla 7 se asume como mejor al modelo 2 y como peor al modelo 1, es decir que el promedio de los estudiantes de sexto semestre de la UNACH se relaciona más con el promedio general y el promedio de los de quinto semestre se relaciona menos. También se calculó un promedio general de las medidas de exactitud (tabla 8) analizando las más importantes: R y RMSE; donde las correlaciones son considerables (RL: 0,682/RP: 0,614) y son factibles para realizar predicciones confiables porque los valores para RMSE son bajos (RL: 0,432/RP: 0,762), estos modelos son aceptables para hacer predicciones del promedio general que puede un estudiante obtener al terminar sus estudios cuando estén en quinto, sexto o séptimo semestre.

El que más se ajusta a estos datos con mayor exactitud es el método de regresión lineal con un R de 0,682 y un RMSE de 0,432, en comparación con el método de regresión polinomial que tiene un R de 0,614 y un RMSE de 0,762.

Medidas de exactitud para la tabla Docente

En este parte se muestra los resultados obtenidos de las medidas de exactitud de los modelos generados con los métodos de regresión lineal y polinomial aplicados en los datos cuantitativos personales y académicos de docentes.

Variable dependiente: Resultado Final de Evaluación Docente

Tabla 9. Modelos para docentes - Medidas de exactitud

	Modelo 1 RL/RP	**Modelo 2** RL/RP	**Modelo 3** RL/RP	**Modelo 4** RL/RP	**Modelo 5** RL/RP	**Modelo 6** RL/RP	**Modelo 7** RL/RP	**Modelo 8** RL/RP	**Modelo 9** RL/RP
Variables independientes	Horas Actividad Académica	Horas Clase	Horas Eventos Aprobados	Horas Eventos Asistidos	Nº Eventos Aprobados	Nº Eventos Asistidos	Nº Eventos Internacionales	Nº Eventos Nacionales	Número Hijos
RMSE	0,920/ 0,920	0,929/ 0,930	0,923/ 0,923	0,932/ 0,931	0,921/ 0,924	0,927/ 0,930	0,926/ 0,927	0,921/ 0,921	0,924/ 0,926
AE	0,498/ 0,498	0,514/ 0,516	0,513/ 0,513	0,517/ 0,518	0,515/ 0,514	0,513/ 0,516	0,516/ 0,518	0,514/ 0,514	0,516/ 0,518
R	0,185/ 0,185	0,000/ 0,011	0,162/ 0,162	0,010/ 0,026	0,182/ 0,164	0,109/ 0,105	0,127/ 0,118	0,179/ 0,179	0,164/ 0,164
R^2	0,055/ 0,055	0,000/ 0,003	0,031/ 0,031	0,001/ 0,009	0,042/ 0,037	0,018/ 0,016	0,020/ 0,018	0,043/ 0,043	0,044/ 0,051

RL: Regresión Lineal, **RP:** Regresión Polinomial

Tabla 10. Promedio Medidas de Exactitud-Docente

Métodos	**Promedio RMSE**	**Promedio AE**	**Promedio R**	**Promedio R^2**
Regresión Lineal	0,924	0,512	0,124	0,028
Regresión Polinomial	0,925	0,513	0,123	0,029

Análisis: Según los resultados de las medidas de exactitud de la tabla 9 se asume como mejor al modelo 1 y como peor al modelo 4, es decir que las horas de actividad académica se relaciona más con la evaluación final del docente y las horas de eventos asistidos se relaciona menos. También se calculó un promedio general de las medidas de exactitud (tabla 10) analizando las más importantes: R y RMSE; para estos 9 modelos se tiene correlaciones débiles (RL: 0,124/RP: 0,123), tampoco son factibles para realizar predicciones confiables porque los valores para RMSE son altos (RL: 0,924/RP: 0,925), pero se puede deducir que las variables independientes influyen mínimamente en la evaluación final de los docentes; y el que más se ajusta a estos datos con mayor exactitud es el método de regresión lineal con un R de 0,124 y un RMSE de 0,924, en comparación con el método de regresión polinomial que tiene un R de 0,123 y un RMSE de 0,925.

Se ha considerado también los siguientes modelos solo con las evaluaciones del docente por componente en relación con el resultado de la evaluación final.

Variable dependiente: Resultado Final de Evaluación Docente

Tabla 11. Modelos para docentes por componente - Medidas de exactitud

	Modelo 1 RL/RP	**Modelo 2** RL/RP	**Modelo 3** RL/RP
Variables independientes	Evaluación Componente Docencia	Evaluación Componente Gestión	Evaluación Componente Investigación
RMSE	0,708/0,696	0,894/0,931	0,923/0,920
AE	0,482/0,475	0,485/0,517	0,506/0,503
R	0,642/0,655	0,318/0,237	0,150/0,164
R^2	0,439/0,457	0,115/0,078	0,029/0,034

RL: Regresión Lineal, **RP:** Regresión Polinomial

Tabla 12. Promedio Medidas de Exactitud-Docente por componente

Métodos	**Promedio RMSE**	**Promedio AE**	**Promedio R**	**Promedio R^2**
Regresión Lineal	0,841	0,491	0,370	0,218
Regresión Polinomial	0,849	0,498	0,352	0,189

Análisis: De acuerdo a las medidas de exactitud de la tabla 11 se asume como mejor al modelo 1 y como peor al modelo 3, es decir que la evaluación por docencia se relaciona e influye más en el resultado final de la evaluación del docente y la evaluación por investigación se relaciona menos. También se calculó un promedio general de las medidas de exactitud (tabla 12) analizando las más importantes: R y RMSE; donde las correlaciones son débiles (RL: 0,370/RP: 0,352) y tampoco son factibles para realizar pronósticos confiables porque los valores para RMSE son altos (RL: 0,841/RP: 0,849), en este caso el que más se ajusta a estos datos con mayor exactitud es el método de regresión lineal con un R de 0,370 y un RMSE de 0,841, en comparación con el método de regresión polinomial que tiene un R de 0,352 y un RMSE de 0,849.

A continuación, se muestra en ilustración 1 y 2 los valores observados vs los pronosticados del mejor modelo, que tiene como variable independiente al promedio de los estudiantes de sexto semestre de la UNACH y como variable dependiente al promedio general, el cual presentó los mejores resultados con un RMSE de 0,384 (38,4%), un AE de 0,297 (29,7%), un R de 0,770 (77%) y un R^2 de 0,594 (59,4%) con el método de regresión lineal.

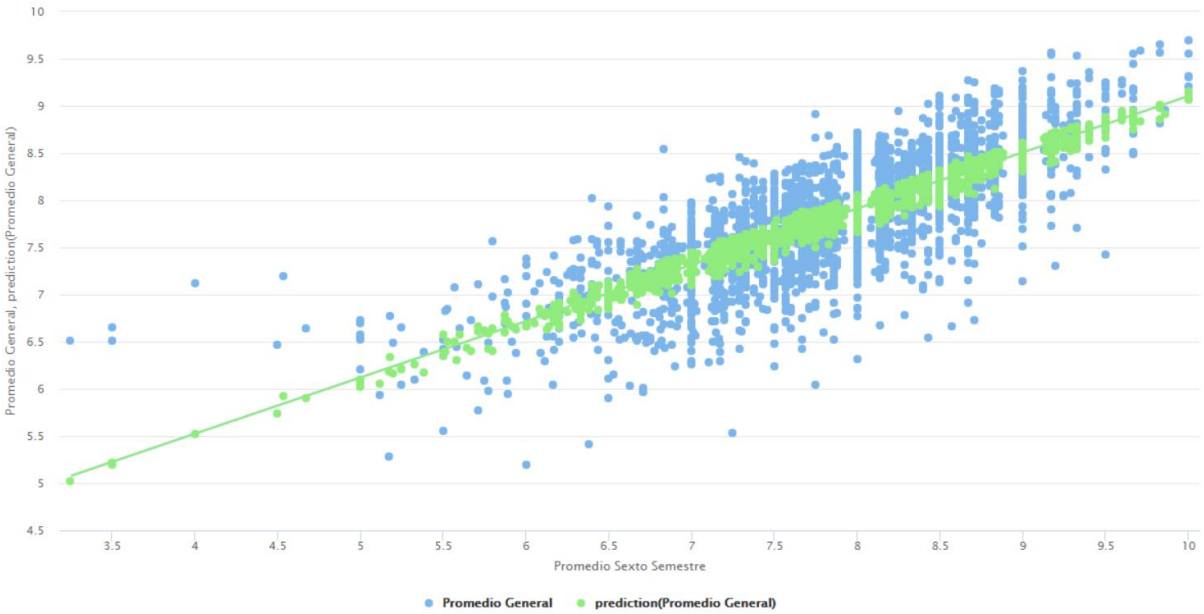


Ilustración 1. Regresión lineal - Mejor modelo

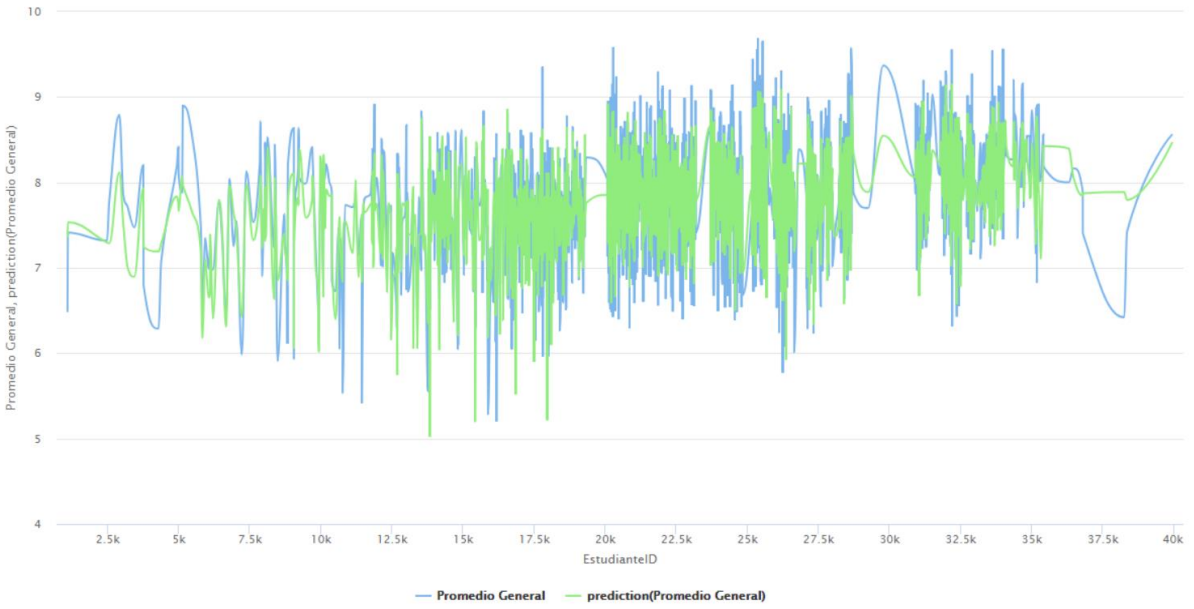


Ilustración 2. Valores observados y Valores pronosticados

Resultado final de las medidas de exactitud de los métodos

Finalmente, se calculó los promedios finales de las medidas de exactitud: RMSE, AE, R y R^2 para los métodos de regresión lineal y polinomial.

Todos los análisis realizados con los datos cuantitativos de estudiantes y docentes favorecen al método de regresión lineal, el cual muestra mayor exactitud en comparación con el método de regresión polinomial como se puede observar en la tabla 13.

Tabla 13. Promedio Final de los Métodos- Medidas de Exactitud

Métodos	**Promedio RMSE**	**Promedio AE**	**Promedio R**	**Promedio R^2**
Regresión Lineal	0,798	0,503	0,311	0,184
Regresión Polinomial	0,883	0,568	0,288	0,159

5.2.Discusión

Para dar respuesta al objetivo principal, los resultados obtenidos de los análisis de exactitud de los métodos de regresión lineal y polinomial sobre los datos cuantitativos personales de estudiantes y docentes, así como también académicos; con las variables metas: promedio general de estudiantes y resultado final de evaluación docentes revelaron que:

Los resultados de las medidas de exactitud para los 5 modelos generados para estudiantes (tabla 6) se obtuvieron correlaciones débiles (RL: 0,069/RP: 0,063) y errores altos (RL: 0,995/RP: 0,996) donde las variables independientes (Número Hermanos, Número Hijos, Número Integrantes Hogar, Número Personas dependen de Ingresos y el Total Ingresos) tienen poca relación e influyen mínimamente en el promedio de los estudiantes, pero el que más se relaciona es el número de hermanos. En 2 modelos es mejor el método de regresión lineal, en 1 el método de regresión polinomial y en 3 ninguno.

También se consideró solo el rendimiento académico del estudiante por semestre en relación con el promedio general final, donde los resultados de las medidas de exactitud para los 3 modelos generados (tabla 8) se obtuvieron correlaciones considerables (RL: 0,682/RP: 0,614) y errores bajos (RL: 0,432/RP: 0,762) donde las variables independientes (Promedio de Quinto,

Sexto y Séptimo) se relacionan más y son aceptables para hacer predicciones del promedio general de los estudiantes. En todos los modelos es mejor el método de regresión lineal.

En los 9 modelos generados para docentes, los resultados de las medidas de exactitud (tabla 10) se obtuvieron correlaciones débiles (RL: 0,124/RP: 0,123) y errores altos (RL: 0,924/RP: 0,925) donde las variables independientes (Horas Actividad Académica, Horas Clase, Horas Eventos Aprobados, Horas Eventos asistidos, Nº Eventos Aprobados, Nº Eventos Asistidos, Nº Eventos Internacionales, Nº Eventos Nacionales y Número Hijos) tienen poca relación e influyen mínimamente en la evaluación final de los docentes, pero el que más se relaciona es las horas de actividad académica. En 5 modelos es mejor el método de regresión lineal, en 1 el método de regresión polinomial y en 3 ninguno.

También se consideró las evaluaciones por componente en relación con la evaluación final, donde los resultados de las medidas de exactitud para los 3 modelos generados (tabla 12) se obtuvieron correlaciones débiles (RL: 0,370/RP: 0,352) y errores altos (RL: 0,841/RP: 0,849) donde las variables independientes (Evaluación por Componente Docencia, Gestión e Investigación) influyen mínimamente en la evaluación final del docente, aunque la evaluación por docencia se relaciona e influye más que la evaluación por gestión e investigación. En 2 modelos es mejor el método de regresión polinomial y en 1 el método de regresión lineal.

Entonces los valores finales de las medidas de exactitud favorecen con mayor exactitud al método de regresión lineal porque tiene un RMSE de 0,798, un AE de 0,503, un R de 0,311 y un R^2 de 0,184 que en comparación con el método de regresión polinomial son mejores (tabla 13) y también de acuerdo con los análisis realizados de los 20 modelos, 11 favorecen a la regresión lineal, 4 a la regresión polinomial y 5 a ninguno.

CONCLUSIONES

Se logró seguir el proceso de minería de datos mediante la utilización de la metodología CRISP-DM, teniendo como la fase más difícil a la preparación de los datos, porque para aplicar cualquier técnica de minería de datos se debe contar con información confiable, pero dada la naturaleza de los datos personales y académicos proporcionados de estudiantes y docentes de la base de datos del SICOA de la UNACH se tuvo que realizar una exploración y verificación de los datos, donde se encontró que algunos no fueron completos por lo que se determinó datos nulos y no válidos, por esta razón se tuvo que depurar y construir los datos para que sean más confiables, estables e íntegros.

Una vez aplicado los métodos de regresión a los datos cuantitativos personales y académicos de estudiantes y docentes de la UNACH se obtuvo en promedio correlaciones débiles (RL:0,311 y RP:0,288), deduciendo que para los datos de estudiantes: número de hijos, número de integrantes del hogar, número de personas que dependen de los ingresos, total de ingresos, promedio de quinto y séptimo semestre influyen mínimamente en el promedio general de los estudiantes, mientras que el número de hermanos y el promedio de sexto semestre se relaciona e influye más. Para los datos de docentes: horas de actividad académica, horas clase, horas de eventos aprobados, horas de eventos asistidos, número de eventos aprobados, número de eventos asistidos, número de eventos internacionales, número de eventos nacionales, número de hijos, evaluación por gestión e investigación también influyen mínimamente en la evaluación final del docente, mientras que las horas de actividad académica y la evaluación por docencia se relaciona e influye más.

Al realizar los análisis de las medidas de exactitud: RMSE, AE, R y R^2 obtenidos de los 20 modelos generados con los métodos de regresión lineal y polinomial sobre los datos cuantitativos de estudiantes y docentes se determinaron en promedio errores altos: RMSE (RL: 0,798/RP: 0,883) y AE (RL: 0,503/RP: 0,568) y correlaciones débiles: R (RL: 0,311/RP: 0,288) y R^2 (RL: 0,184/RP: 0,159) donde el método de regresión lineal es el que tiene mayor exactitud y se elige como el mejor que se ajusta a los datos cuantitativos de estudiantes y docentes de la UNACH, incluso en 11 de los modelos los resultados de las medidas favorecen a la regresión lineal, mientras que solo 4 a la regresión polinomial y 5 a ninguno.

RECOMENDACIONES

Escoger una metodología de minería de datos como guía, de preferencia CRISP-DM que es la más completa y utilizada por varios autores; también elegir una herramienta que ayude al análisis de los datos como por ejemplo RapidMiner que tiene una interfaz gráfica de usuario intuitiva. También se debe seleccionar solo los datos necesarios, los que pueden ser útiles para el estudio; si se trata de métodos de regresión de preferencia datos cuantitativos continuos y tener la base de datos depurada para no tener errores al aplicar los respectivos métodos, inclusive para ahorrar tiempo al momento de generar los modelos.

Para realizar un proyecto de investigación sobre minería de datos con métodos de regresión en cualquier tipo de entorno, es necesario efectuar un estudio previo de los métodos disponibles para poder determinar aquellos que mejor se ajusten a las necesidades del proyecto para posteriormente aplicarlos y evaluarlos mediante medidas de exactitud previamente establecidas como, por ejemplo: RMSE, AE, R y R^2. También se recomienda utilizar la regresión lineal en entornos educativos, porque en esta investigación fue el método con mayor exactitud que se ajusta a los datos cuantitativos personales y académicos de la base de datos del SICOA de la UNACH.

REFERENCIAS BIBLIOGRÁFICAS

Porcel, E. A., Dapozo, G. N., & López, M. V. (2010). Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa. *Revista electrónica de investigación educativa*, *12*(2), 1-21.

Aranda, Y. R., & Solotongo, A. R. (2013). INTEGRACIÓN DE LOS ALGORITMOS DE MINERÍA DE DATOS 1R, PRISM E ID3 A POSTGRESQL. *Journal of Information Systems and Technology Management, 10*(2), 389-406. doi:10.4301/S1807-17752013000200012

Astorga Gómez, J. M. (2014). Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente. *Ingeniería Energética, 35*(3), 234-241.

Castorena Peña, J. A., Silva Ávila, A. E., Domínguez Lugo, A. J., & Rodríguez Montelongo, D. L. (2018). El uso de herramientas tecnológicas de minería de datos en el análisis de datos climatológicos. *Revista Iberoamericana de las Ciencias Computacionales e Informática, 7*(13), 18. doi:10.23913/reci.v7i13.75

Cortina, V. G. (2015). *APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO.* Universidad Carlos III de Madrid, Madrid.

Dagnino, J. (2014). Coeficiente de correlacion lineal de pearson. *Chil Anest*, *43*, 150-153.

García Jiménez, V., Alvarado Izquierdo, J., & Jiménez Blanco, A. (2000). LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO: REGRESIÓN LINEAL VERSUS REGRESIÓN LOGÍSTICA. *Psicothema*, *12*(2), 248-252.

García Saiz, D. (2016). *García Saiz, D. (2016). Minería de datos aplicada a la enseñanza virtual: nuevas propuestas para la construcción de modelos y su integración en un entorno amigable para el usuario no experto.* Tesis Doctoral, Universidad de Cantabria, Departamento de Ingeniería Informática y Electrónica, Cantabria.

Gutiérrez, J., & Molina, B. (2015). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. *Revista Ontare, 3*(2), 33-51. doi:https://doi.org/10.21158/23823399.v3.n2.2015.1440

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. P. (2014). *Metodología de la Investigación.* México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.

Jaramillo, A., & Paz Arias, H. (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), 64-90.

Laguna, C. (2014). *Correlación y regresión lineal.* Instituto Aragonés de Ciencias de la Salud.

Martínez Rodríguez, E. (2005). Errores frecuentes en la interpretación del coeficiente de determinación lineal. *Anuario jurídico y económico escurialense*(38), 315-331.

Medina Rojas, F., & Gómez Santamaría, C. (2014). Funcionalidades de la minería de datos. *Revista Ingeniería y Región*, 10.

Menes Camejo, I., Arcos Medina, G., Moreno Beltrán, P., & Gallegos Carrillo, K. (14 de 09 de 2015). Desempeño de algoritmos de minería en indicadores académicos: Árbol de Decisión y Regresión Logística. *Revista Cubana de Ciencias Informáticas, 9*(4), 104-117.

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo.* UNIVERSIDAD NACIONAL DE LA PLATA, Argentina. Obtenido de http://hdl.handle.net/10915/29582

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Oviedo Carrascal, E. A., Oviedo Carrascal, A. I., & Vélez Saldarriaga , G. L. (2015). MINERÍA DE DATOS: APORTES Y TENDENCIAS EN EL SERVICIO DE SALUD DE CIUDADES INTELIGENTES. *Revista politécnica, 11*(20), 111-120.

Pérez Planells, L., Delegido, J., Rivera Caicedo, J., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista Española de Teledetección*, 55-65. doi:https://doi.org/10.4995/raet.2015.4153

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

Piatetsky, G. (2014). *KDnuggets*. Obtenido de KDnuggets: https://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

RapidMiner. (2019). *RapidMiner Documentation*. Obtenido de RapidMiner Documentation: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html

Ritter, A., Muñoz Carpena, R., & Regalado, C. M. (2011). Capacidad de predicción de modelos aplicados a la ZNS: herramienta informática para la adecuada evaluación de la bondad de ajuste con significación estadística. *Estudios en la Zona no Saturada del Suelo, 10*, 259-264.

Vinuesa, P. (2016). *Tema 9 - Regresión lineal simple y polinomial: teoría y práctica.* Universidad Nacional Autónoma de México, México.

ANEXOS

ANEXO 1: Descripción de los datos

Tabla Estudiante

Esta tabla registra toda la información personal necesaria de cada estudiante de la Universidad Nacional de Chimborazo.

Tabla 14. Descripción de los datos de la Tabla Estudiante

Campo	Tipo	Descripción
Estudiante ID	Numérico	Campo que identifica a cada estudiante
Porcentaje Discapacidad	Numérico	Campo que contiene el porcentaje de discapacidad que tiene algún estudiante
Numero Integrantes Hogar	Numérico	Campo que contiene el número de integrantes en el hogar de cada estudiante
Numero Hermanos	Numérico	Campo que contiene el número de hermanos que tiene cada estudiante
Ingresos Padre	Numérico	Campo que contiene los ingresos económicos del padre de cada estudiante
Ingresos Madre	Numérico	Campo que contiene los ingresos económicos de la madre de cada estudiante
Total Ingresos Padres	Numérico	Campo que contiene el ingreso económico total de los padres de cada estudiante
Numero Dependen Ingresos	Numérico	Campo que contiene el número de personas que dependen del total de ingresos de los padres de cada estudiante
Valor Mensual Servicios	Numérico	Campo que contiene el valor mensual que gasta en servicios (agua, luz, teléfono, internet, tv Pagada) cada estudiante
Total Ingresos	Numérico	Campo que contiene el total de ingresos económicos de cada estudiante
Numero Hijos	Numérico	Campo que contiene el número de hijos que tiene cada estudiante
Ingresos Cónyuge	Numérico	Campo que contiene el ingreso económico del cónyuge del estudiante
Total Ingresos Estudiante	Numérico	Campo que contiene el total de ingresos económicos que tiene el estudiante

Campo	Tipo	Descripción
Personas Dependen Ingresos	Numérico	Campo que contiene el número de personas que dependen de los ingresos económicos de cada estudiante
Fecha Nacimiento	Date	Campo que contiene la fecha de nacimiento de cada estudiante
Estado Civil	Texto	Campo que contiene el estado civil de cada estudiante (soltero/a, casado/a, unión libre, divorciado/a, viudo/a)
Orientación Sexual	Texto	Campo que contiene la orientación sexual de cada estudiante (heterosexual, bisexual, homosexual, intersex, transgénero, transexual, ninguna)
Sexo	Texto	Campo que contiene el sexo de cada estudiante (mujer, hombre)
Género	Texto	Campo que contiene el género de cada estudiante (masculino, femenino)
Etnia	Texto	Campo que contiene la etnia a cuál pertenece cada estudiante (mestizo/a, indígena, afroecuatoriano/a, blanco/a, mulato/a, montubio/a, negro/a, otro)
Nacionalidad Indígena	Texto	Campo que contiene la nacionalidad indígena a la cual pertenece cada estudiante (Kichwa, Puruha, Achaur, Shuar, Saraguro, etc.)
Institución Educativa	Texto	Campo que contiene el nombre de la institución educativa a la que perteneció el estudiante
Tipo	Texto	Campo que contiene el tipo de institución educativa a la que perteneció el estudiante (fiscal, particular, fiscomisional, beneficencia, municipal, extranjero)
Enfermedad Catastrófica Extraña	Texto	Campo que contiene el nombre de la enfermedad catastrófica o extraña que tiene el estudiante
Tipo Discapacidad	Texto	Campo que contiene el tipo de discapacidad que tiene el estudiante (físico motora, visual, auditiva, intelectual, ninguna)
Actividad Deportiva	Texto	Campo que contiene la actividad deportiva que practica el estudiante (futbol, baloncesto, atletismo, tenis, etc.)
Actividad Cultural	Texto	Campo que contiene la actividad cultural que hace el estudiante (actuación, baile, canto, pinta, etc.)
País Nacimiento	Texto	Campo que contiene el país de nacimiento de cada estudiante

Campo	Tipo	Descripción
Provincia Nacimiento	Texto	Campo que contiene la provincia de nacimiento de cada estudiante
Cantón Nacimiento	Texto	Campo que contiene el cantón de nacimiento de cada estudiante
País Procedencia	Texto	Campo que contiene el país de procedencia de cada estudiante
Provincia Procedencia	Texto	Campo que contiene la provincia de procedencia de cada estudiante
Cantón Procedencia	Texto	Campo que contiene el cantón de procedencia de cada estudiante
Tipo Parroquia	Texto	Campo que contiene el tipo de parroquia en donde reside el estudiante (rural, urbana)
Ocupación	Texto	Campo que contiene la ocupación del estudiante (chofer, mesera, asistente, costurera, taxista, etc.)

Tabla Estudiante Rendimiento

Esta tabla registra la información académica sobre el rendimiento de cada estudiante de la Universidad Nacional de Chimborazo.

Tabla 15. Descripción de los datos de la Tabla Estudiante Rendimiento

Campo	Tipo	Descripción
Estudiante ID	Numérico	Campo que identifica a cada estudiante
Facultad	Texto	Campo que contiene la facultad a la cual pertenece el estudiante (Ciencias de la Salud, Ingeniería, Ciencias Políticas y Administrativas, Ciencias de la Educación, Humanas y Tecnológicas)
Carrera	Texto	Campo que contiene la carrera o escuela a la que pertenece cada estudiante
Situación Actual	Texto	Campo que contiene la situación en la que se encuentra el estudiante (graduado o no graduado)
Nivel	Texto	Campo que contiene el nivel o semestre de cada estudiante en los diferentes periodos académicos
Período	Texto	Campo que contiene el periodo académico que cursó cada estudiante
Promedio	Numérico	Campo que contiene el promedio de cada estudiante por semestre y período académico

Tabla Docente

Esta tabla registra toda la información personal necesaria de cada docente de la Universidad Nacional de Chimborazo.

Tabla 16. Descripción de los datos de la Tabla Docente

Campo	Tipo	Descripción
Cédula	Texto	Campo que identifica a cada docente
País	Texto	Campo que contiene el nombre del país de nacimiento de cada docente
Nacionalidad	Texto	Campo que contiene la nacionalidad a la que pertenece cada docente
Fecha Nacimiento	Date	Campo que contiene la fecha de nacimiento de cada docente
Número Hijos	Numérico	Campo que contiene el número de hijos de cada docente
Estado Civil	Texto	Campo que contiene el estado civil de cada docente (casado/a, soltero/a, divorciado/a, viudo/a, unión libre)
Sexo	Texto	Campo que contiene el sexo de cada docente (hombre, mujer)
Etnia	Texto	Campo que contiene la etnia a la cual pertenece cada docente (mestizo/a, blanco/a, indígena, negro/a, etc.)
Tipo Sangre	Texto	Campo que contiene el tipo de sangre de cada docente
Grupo GLBTI	Texto	Campo que contiene dos opciones (SI y NO): si el docente pertenece o no a algún grupo de GLBTI
Nacionalidad Indígena	Texto	Campo que contiene la nacionalidad indígena a la cual pertenece cada docente (kichwa, achuar, puruha, etc.)
Cantón	Texto	Campo que contiene el cantón de procedencia de cada docente
Parroquia	Texto	Campo que contiene la parroquia de procedencia de cada docente
Nivel Instrucción	Texto	Campo que contiene el nivel de instrucción de cada docente (Posgrado PHD, Posgrado Maestría, Tecnología, Diplomado, etc.)

Campo	Tipo	Descripción
Modalidad	Texto	Campo que contiene la modalidad en la que estudió cada docente (años o semestres)
Área	Texto	Campo que contiene el área en la que se especializó cada docente (Educación, Salud y Bienestar, Administración, Negocios y Legislación, Ingeniería, Industria y Construcción, etc.)
Subárea	Texto	Campo que contiene la subárea en la que se especializó cada docente (Educación, Salud, Negocios y Administración, TIC, etc.)
Campo	Texto	Campo que contiene el campo en la que se especializó cada docente (Ciencias de la educación, Medicina, Derecho, Economía, Base de datos, diseño y administración de redes, etc.)
Está Cursando	Texto	Campo que contiene dos opciones (SI y NO): si el docente está o no cursando sus estudios
Institución Educativa	Texto	Campo que contiene el nombre de la institución educativa en donde cursó o está cursando el docente
Titulo	Texto	Campo que contiene el titulo obtenido por el docente
Experiencia Privada	Texto	Campo que contiene dos opciones (SI y NO): si el docente tiene o no experiencia privada
Experiencia Pública	Texto	Campo que contiene dos opciones (SI y NO): si el docente tiene o no experiencia pública
Familiar Sustituto	Texto	Campo que contiene dos opciones (SI y NO): si el docente tiene o no un familiar sustituto
Enfermedad Catastrófica	Texto	Campo que contiene dos opciones (SI y NO): si el docente tiene o no una enfermedad catastrófica
Tiene Discapacidad	Texto	Campo que contiene dos opciones (SI y NO): si el docente tiene o no algún tipo de discapacidad
Gestión Lactancia	Texto	Campo que contiene dos opciones (SI y NO): si el docente está o no en gestión de lactancia
Tiempo Estudio	Numérico	Campo que contiene el tiempo de estudio (años o semestres) de cada docente

Campo	Tipo	Descripción
Nº Eventos Aprobados	Numérico	Campo que contiene el número de eventos aprobados de cada docente
Nº Eventos Asistidos	Numérico	Campo que contiene el número de eventos asistidos de cada docente
Horas Eventos Aprobados	Numérico	Campo que contiene el número de horas de eventos aprobados de cada docente
Horas Eventos Asistidos	Numérico	Campo que contiene el número de horas de eventos asistidos de cada docente
Nº Eventos Nacionales	Numérico	Campo que contiene el número de eventos nacionales de cada docente
Nº Eventos Internacionales	Numérico	Campo que contiene el número de eventos internacionales de cada docente

Tabla Docente Información Académica

Esta tabla registra toda la información académica necesaria de cada docente de la Universidad Nacional de Chimborazo.

Tabla 17. Descripción de los datos de la Tabla Docente Información Académica

Campo	Tipo	Descripción
Numero Documento	Texto	Campo que identifica a cada docente
Facultad	Texto	Campo que contiene la facultad a la cual pertenece cada docente
Carrera	Texto	Campo que contiene la carrera o escuela a la cual pertenece cada docente
Periodo	Texto	Campo que contiene el periodo académico que estuvo en actividad académica cada docente
Actividad Académica	Texto	Campo que contiene el nombre de la activad académica que hizo cada docente en los diferentes periodos académicos
Horas Actividad Académica	Numérico	Campo que contiene el número de horas de actividad académica de cada docente en los diferentes periodos académicos
Horas Clase	Numérico	Campo que contiene el número de horas de clase de cada docente en los diferentes periodos académicos

Tabla Evaluación Docente

Esta tabla registra la información de los resultados de los diferentes tipos de evaluación realizada a cada docente de la Universidad Nacional de Chimborazo.

Tabla 18. Descripción de los datos de la Tabla Evaluación Docente

Campo	Tipo	Descripción
Usuario Evaluado	Numérico	Campo que identifica a cada docente evaluado
Tipo Evaluación	Texto	Campo que contiene el tipo de evaluación realizada a cada docente (Coevaluación directivos, Coevaluación pares, Autoevaluación, Heteroevaluación)
Componente	Texto	Campo que contiene el componente a la cual pertenece los diferentes tipos de evaluación, estos son: docencia, gestión o investigación
Periodo	Texto	Campo que contiene el periodo académico en el cual se evaluó a cada docente
Resultado Final	Numérico	Campo que contiene el resultado final de los diferentes tipos de evaluaciones en los periodos académicos de cada docente.

ANEXO 2: Exploración de los datos

La Ilustración 3 muestra la distribución de los estudiantes de la UNACH por estado civil.

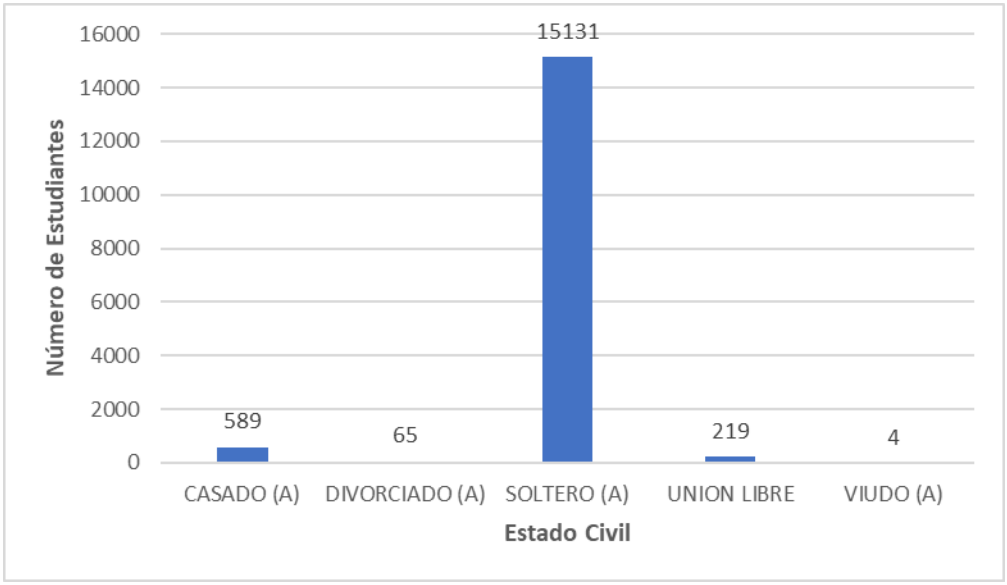


Ilustración 3. Cantidad de estudiantes por estado civil

La Ilustración 4 muestra la distribución de los estudiantes de la UNACH por género.

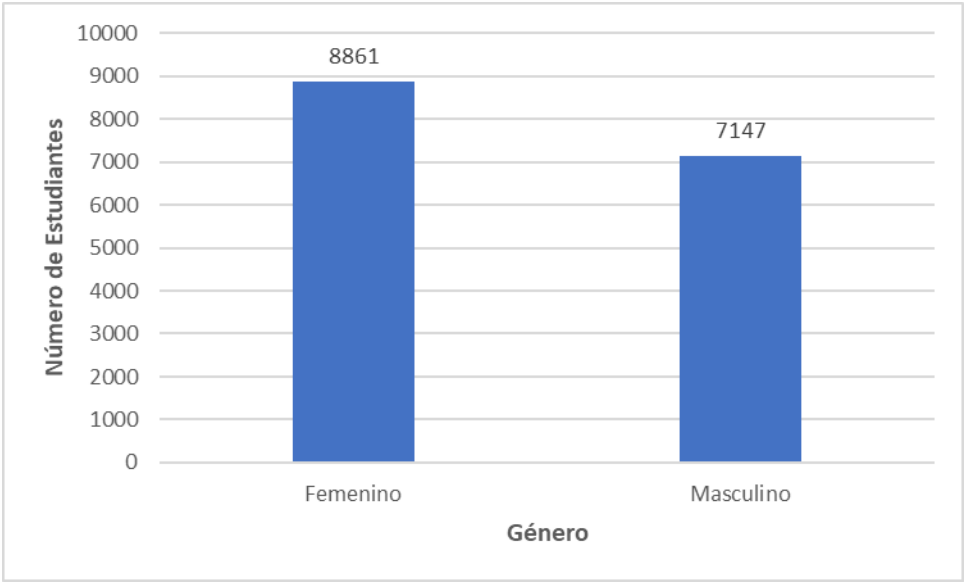


Ilustración 4. Cantidad de estudiantes por género

La Ilustración 5 muestra la distribución de los estudiantes de la Universidad Nacional de Chimborazo por etnia.

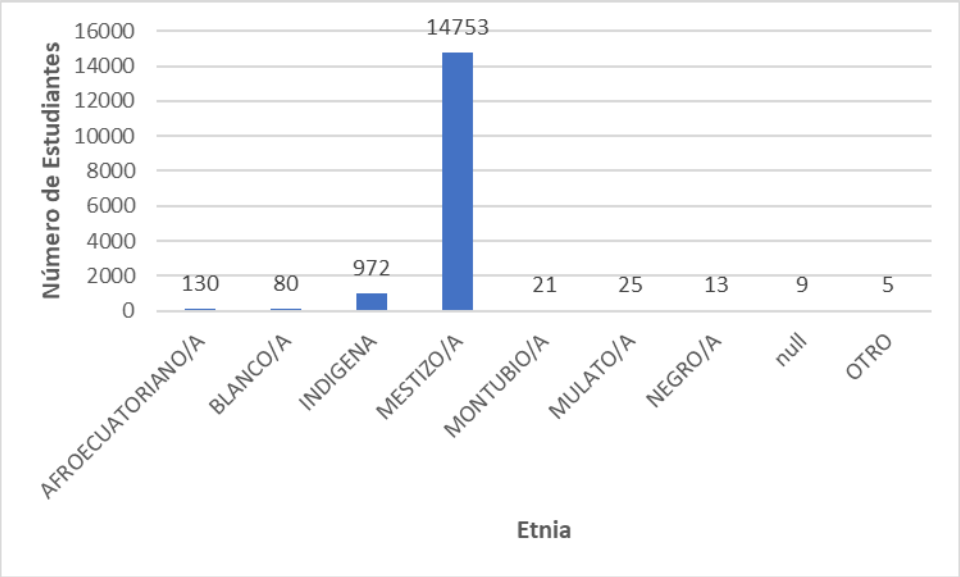


Ilustración 5. Cantidad de estudiantes por etnia

La Ilustración 6 muestra la distribución de los estudiantes de la UNACH por nacionalidad indígena.

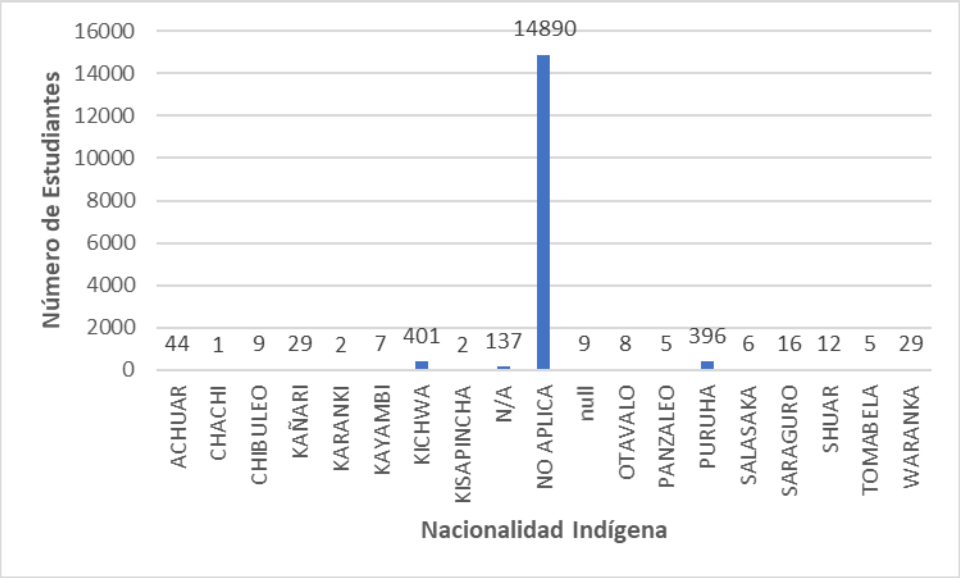


Ilustración 6. Cantidad de estudiantes por nacionalidad indígena

La Ilustración 7 muestra la distribución de los estudiantes de la UNACH por tipo de parroquia.

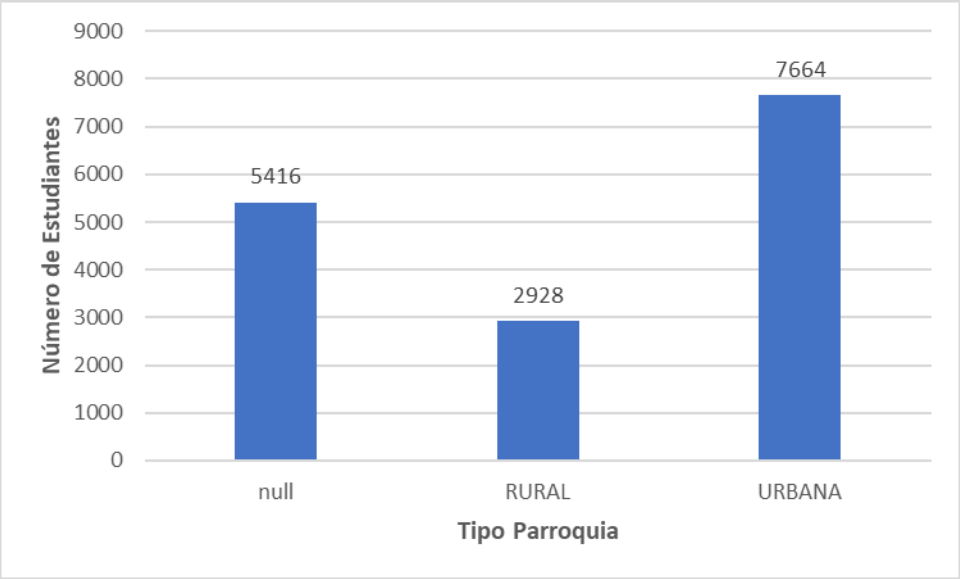


Ilustración 7. Cantidad de estudiantes por tipo de parroquia

La Ilustración 8 muestra la distribución de los estudiantes de la UNACH por el número de integrantes en el hogar.

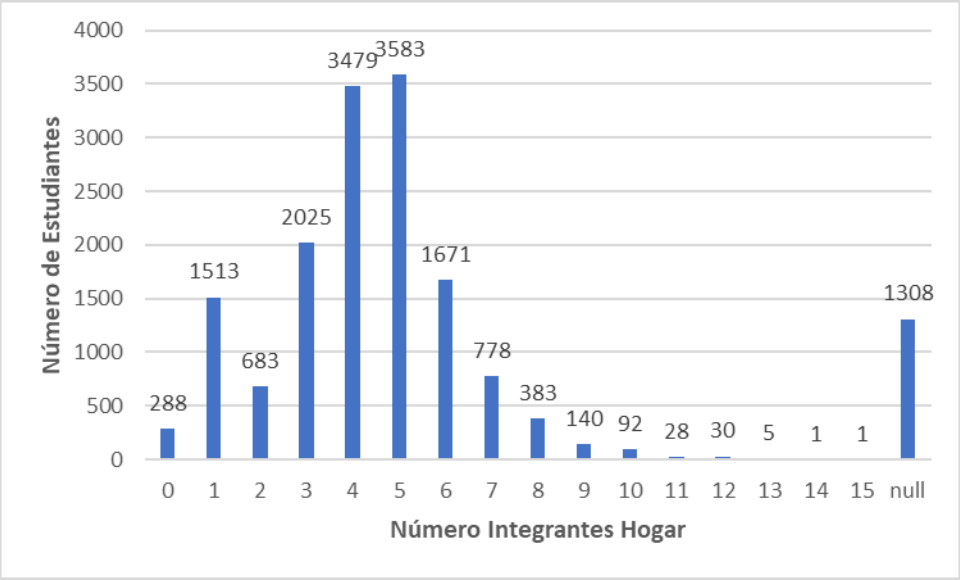


Ilustración 8. Cantidad de estudiantes por número de integrantes en el hogar

La Ilustración 9 muestra la distribución de los estudiantes de la UNACH por el número de hermanos.

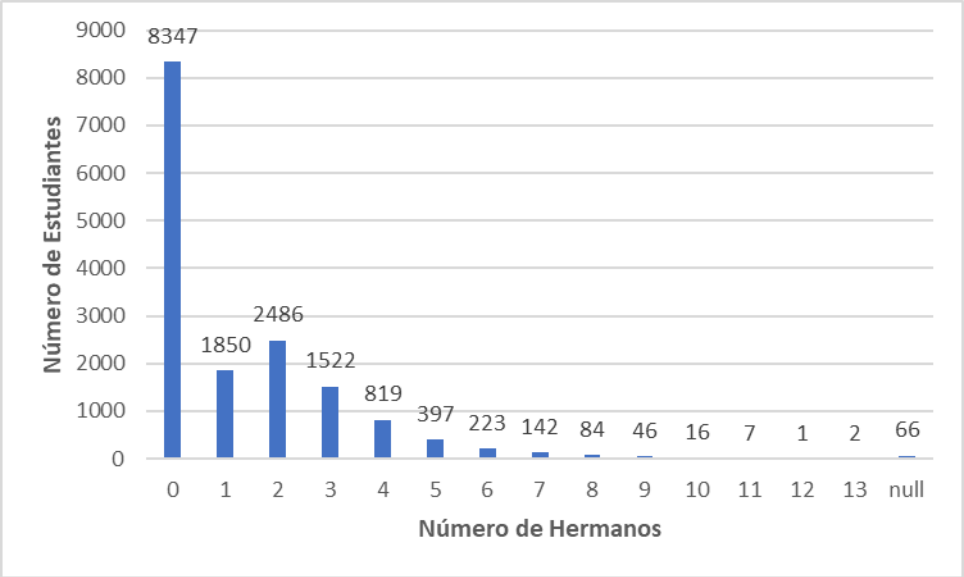


Ilustración 9. Cantidad de estudiantes por número de hermanos

La Ilustración 10 muestra la distribución de los estudiantes de la UNACH por el número de personas que dependen de ingresos.

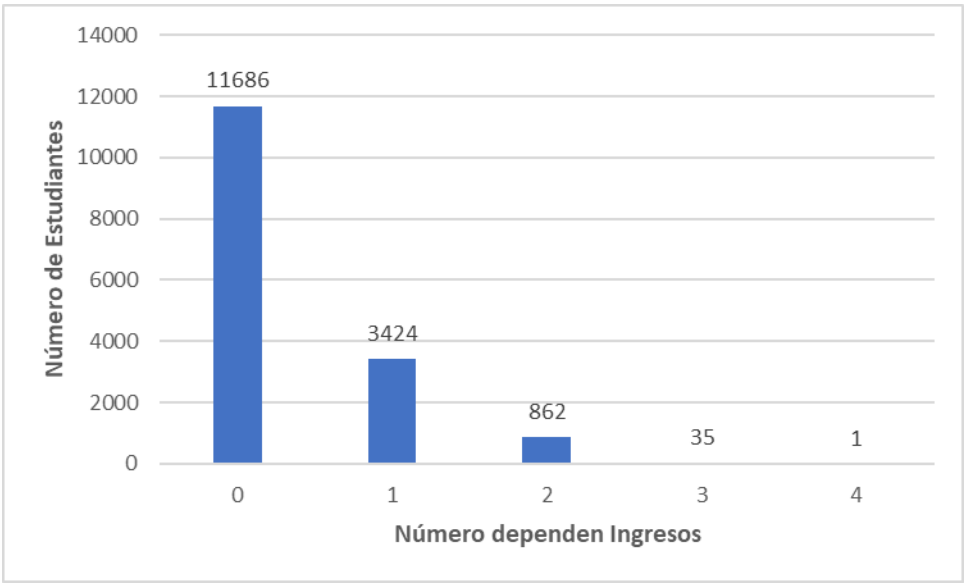


Ilustración 10. Cantidad de estudiantes por número de personas que dependen de ingresos

La Ilustración 11 muestra la distribución de los estudiantes de la UNACH por el número de hijos.

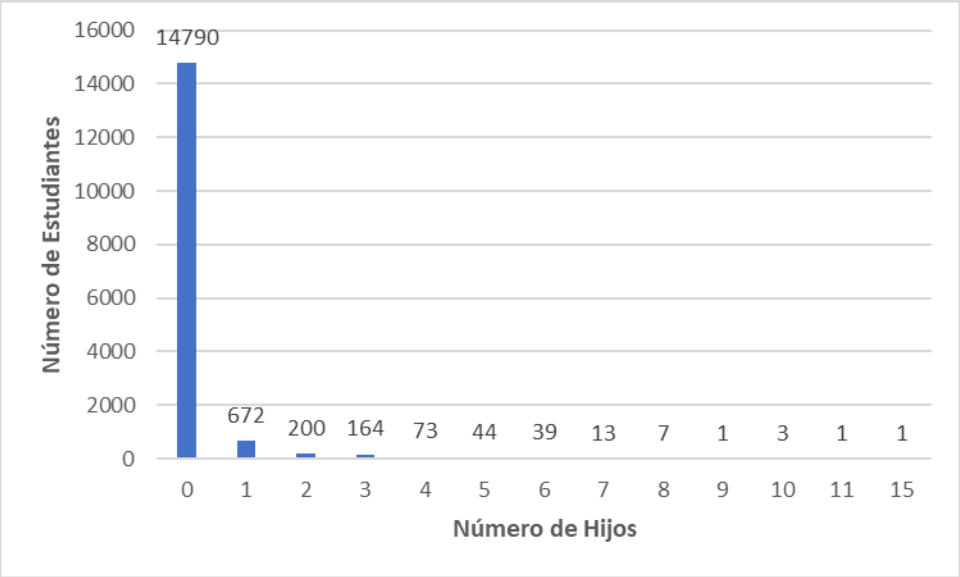


Ilustración 11. Cantidad de estudiantes por número de hijos

La Ilustración 12 muestra la distribución de los estudiantes de la UNACH por facultad.

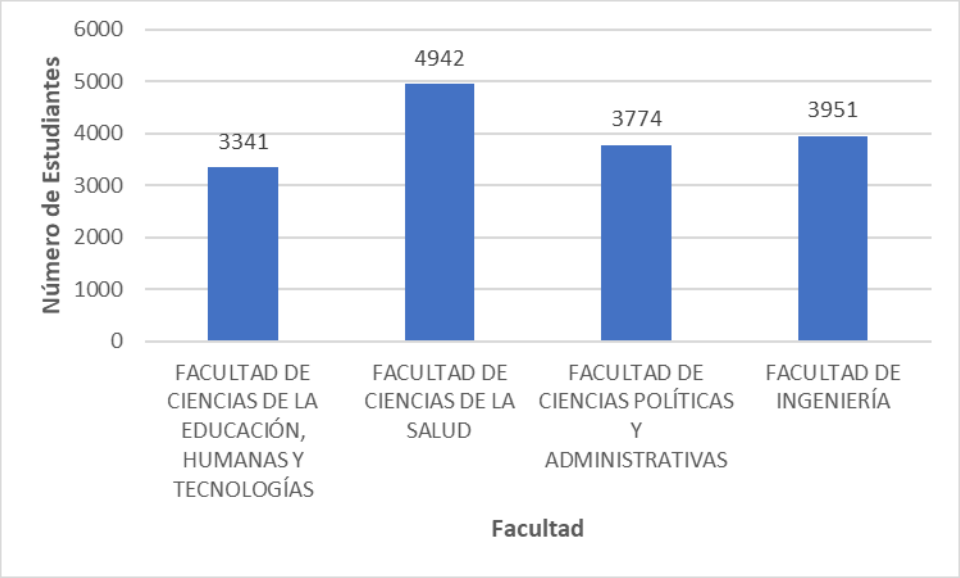


Ilustración 12. Cantidad de estudiantes por facultad

La Ilustración 13 muestra la distribución de los estudiantes de la UNACH por el promedio general (redondeado).

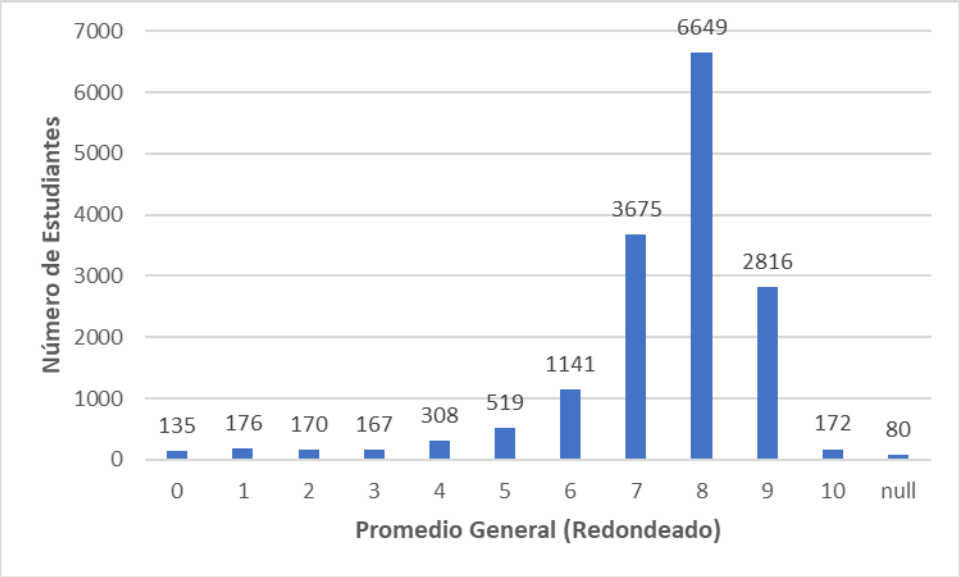


Ilustración 13. Cantidad de estudiantes por promedio general

La Ilustración 14 muestra la distribución de los docentes de la UNACH por país.

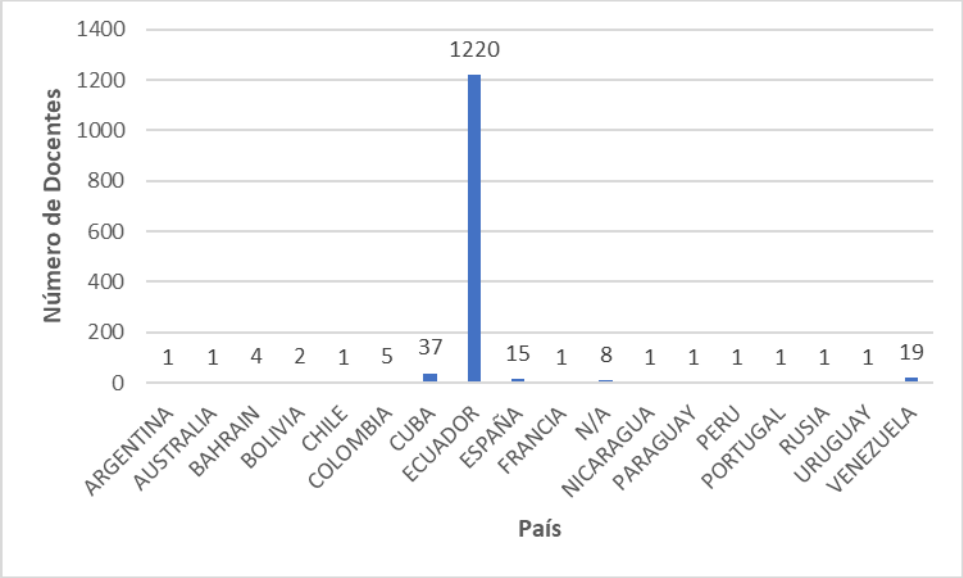


Ilustración 14. Cantidad de docentes por país

La Ilustración 15 muestra la distribución de los docentes de la UNACH por estado civil.

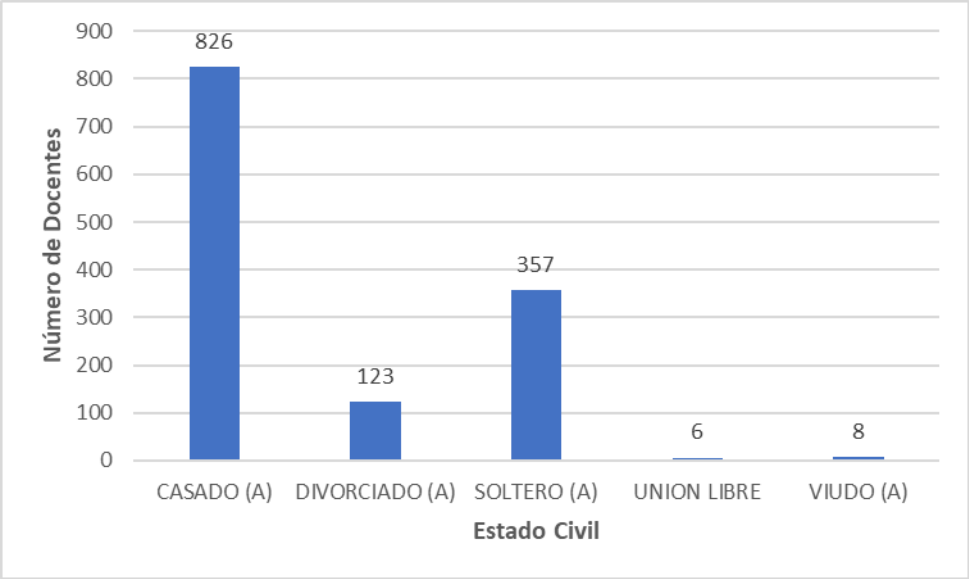


Ilustración 15. Cantidad de docentes por estado civil

La Ilustración 16 muestra la distribución de los docentes de la UNACH por género.

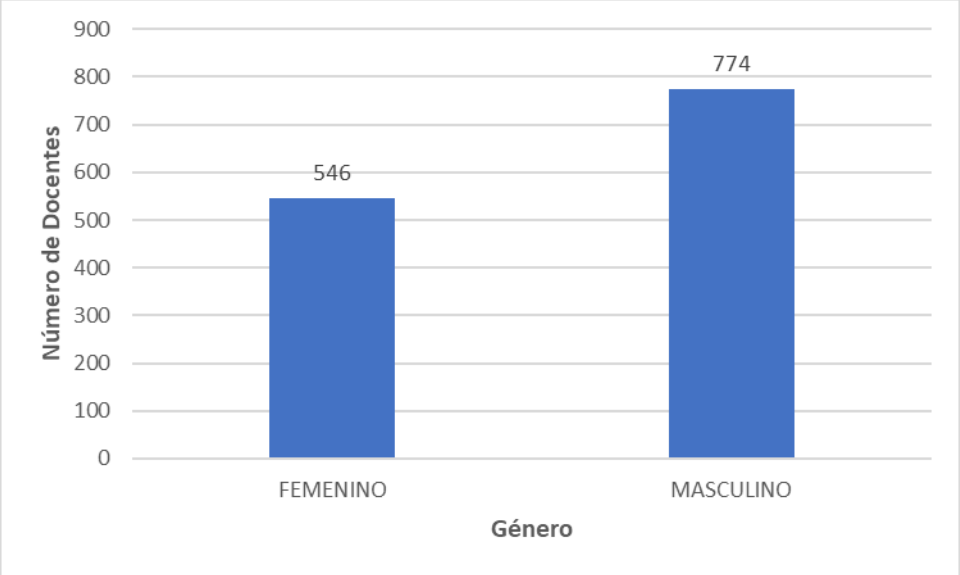


Ilustración 16. Cantidad de docentes por género

La Ilustración 17 muestra la distribución de los docentes de la UNACH por etnia.

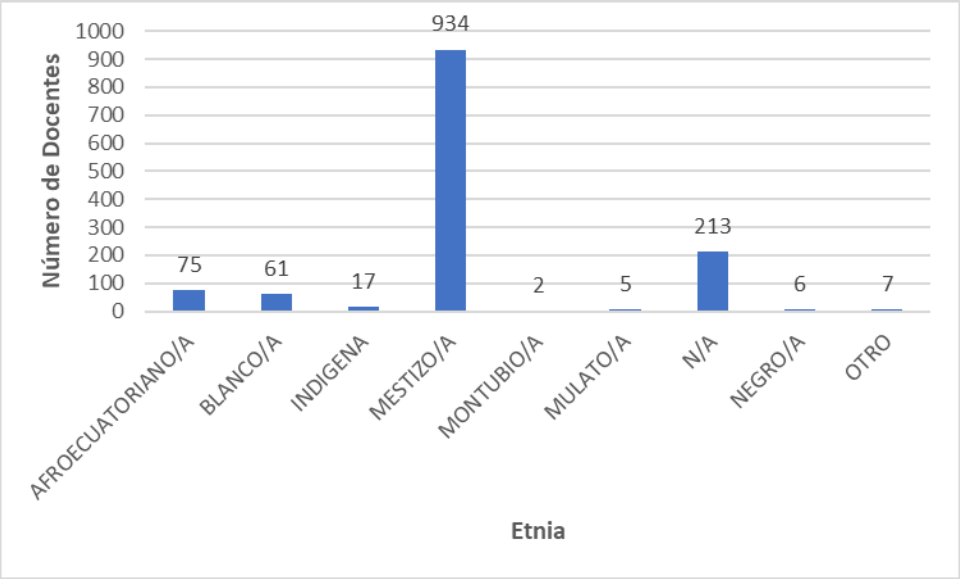


Ilustración 17. Cantidad de docentes por etnia

La Ilustración 18 muestra la distribución de los docentes de la UNACH por nivel de instrucción.

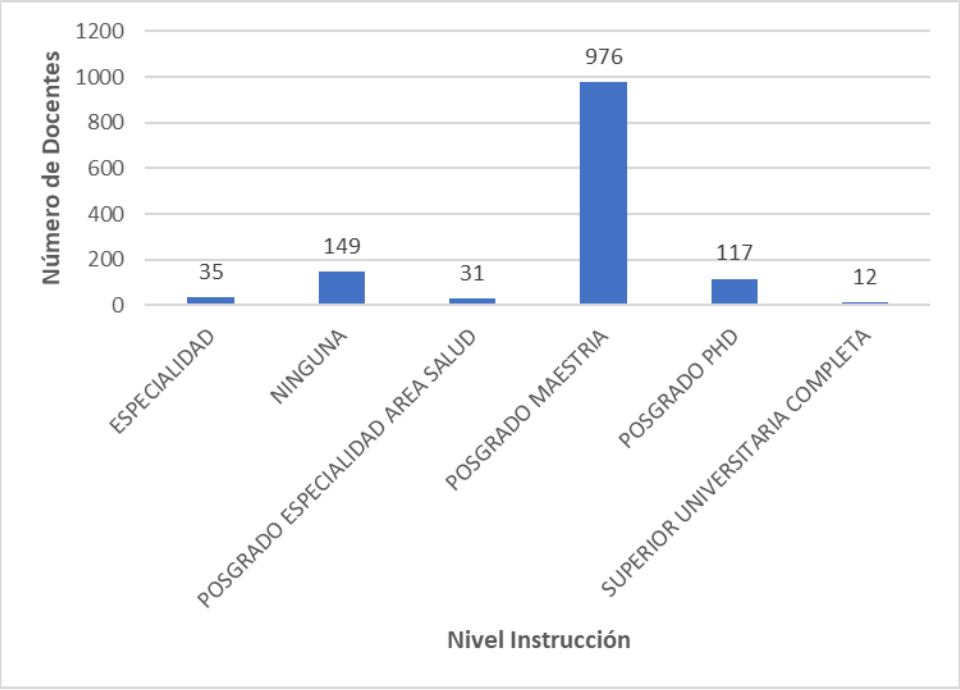


Ilustración 18. Cantidad de docentes por nivel de instrucción

La Ilustración 19 muestra la distribución de los docentes de la UNACH por facultad.

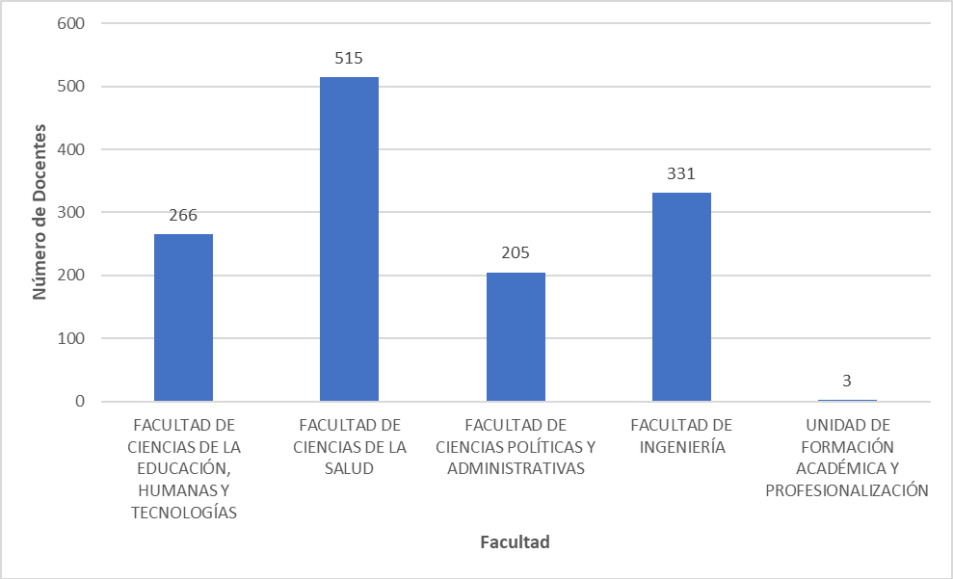


Ilustración 19. Cantidad de docentes por facultad

La Ilustración 20 muestra la distribución de los docentes de la UNACH por el número de hijos.

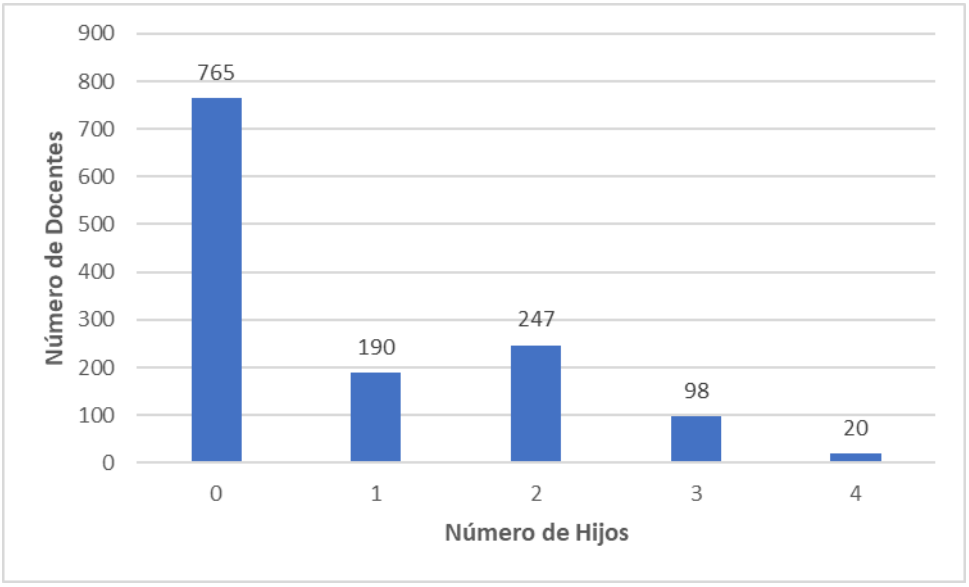


Ilustración 20. Cantidad de docentes por número de hijos

ANEXO 3: Verificar la calidad de datos

Tabla 19. Calidad de datos

Tabla	Campo	Valores nulos		Valores válidos		Valores no válidos	
		#	%	#	%	#	%
Estudiante	Estudiante ID	0	0,00	15766	98,49	242	1,51
16008 registros	Porcentaje Discapacidad	15901	99,33	107	0,67	0	0,00
	Numero Integrantes Hogar	1308	8,17	14700	91,83	0	0,00
	Numero Hermanos	66	0,41	15942	99,59	0	0,00
	Ingresos Padre	0	0,00	16008	100,00	0	0,00
	Ingresos Madre	0	0,00	16008	100,00	0	0,00
	Total Ingresos Padres	0	0,00	16008	100,00	0	0,00
	Numero Dependen Ingresos	0	0,00	16008	100,00	0	0,00
	Valor Mensual Servicios	0	0,00	16008	100,00	0	0,00
	Total Ingresos	0	0,00	16008	100,00	0	0,00
	Numero Hijos	0	0,00	16008	100,00	0	0,00
	Ingresos Cónyuge	0	0,00	16008	100,00	0	0,00
	Total Ingresos Estudiante	0	0,00	16008	100,00	0	0,00
	Personas Dependen Ingresos	0	0,00	16008	100,00	0	0,00
	Fecha Nacimiento	0	0,00	16008	100,00	0	0,00
	Estado Civil	0	0,00	16008	100,00	0	0,00
	Orientación Sexual	9	0,06	15999	99,94	0	0,00
	Sexo	0	0,00	16008	100,00	0	0,00
	Género	0	0,00	16008	100,00	0	0,00
	Etnia	9	0,06	15999	99,94	0	0,00
	Nacionalidad Indígena	9	0,06	15999	99,94	0	0,00
	Institución Educativa	0	0,00	16008	100,00	0	0,00
	Tipo	0	0,00	16008	100,00	0	0,00
	Enfermedad Catastrófica Extraña	0	0,00	16008	100,00	0	0,00
	Tipo Discapacidad	9	0,06	15999	99,94	0	0,00
	Actividad Deportiva	1560	9,75	14448	90,25	0	0,00
	Actividad Cultural	1694	10,58	14314	89,42	0	0,00
	País Nacimiento	0	0,00	15974	99,79	34	0,21
	Provincia Nacimiento	0	0,00	16008	100,00	0	0,00
	Cantón Nacimiento	5740	35,86	10268	64,14	0	0,00
	País Procedencia	5519	34,48	10406	65,00	83	0,52
	Provincia Procedencia	5629	35,16	10379	64,84	0	0,00
	Cantón Procedencia	5629	35,16	10379	64,84	0	0,00
	Tipo Parroquia	5416	33,83	10592	66,17	0	0,00
	Ocupación	14896	93,05	16008	100,00	0	0,00
Estudiante Rendimiento	Estudiante ID	0	0,00	85685	98,37	1420	1,63
87105 registros	Facultad	0	0,00	87105	100,00	0	0,00
	Carrera	0	0,00	87105	100,00	0	0,00
	Situación Actual	0	0,00	87105	100,00	0	0,00
	Nivel	0	0,00	87105	100,00	0	0,00

Tabla	Campo	Valores nulos		Valores válidos		Valores no válidos	
	Período	0	0,00	87105	100,00	0	0,00
	Promedio	296	0,34	86809	99,66	0	0,00
Docente 4097 registros	Cédula	0	0,00	3253	79,40	844	20,6
	País	13	0,32	4046	98,76	38	0,93
	Nacionalidad	3	0,07	4094	99,93	0	0,00
	Fecha Nacimiento	0	0,00	4097	100,00	0	0,00
	Número Hijos	0	0,00	4097	100,00	0	0,00
	Estado Civil	0	0,00	4097	100,00	0	0,00
	Sexo	0	0,00	4097	100,00	0	0,00
	Etnia	550	13,42	3547	86,58	0	0,00
	Tipo Sangre	556	13,57	3541	86,43	0	0,00
	Grupo GLBTI	0	0,00	4097	100,00	0	0,00
	Nacionalidad Indígena	560	16,67	3537	86,33	0	0,00
	Cantón	556	13,57	3541	86,43	0	0,00
	Parroquia	556	13,57	3541	86,43	0	0,00
	Nivel Instrucción	16	0,39	4081	99,61	0	0,00
	Modalidad	1058	25,82	2725	66,51	314	7,66
	Área	1847	45,08	2250	54,92	0	0,00
	Subárea	1847	45,08	2250	54,92	0	0,00
	Campo	1847	45,08	2250	54,92	0	0,00
	Está Cursando	1058	25,82	2725	66,51	0	0,00
	Institución Educativa	1	0,02	4096	99,98	0	0,00
	Titulo	1	0,02	4096	99,98	0	0,00
	Experiencia Privada	0	0,00	4097	100,00	0	0,00
	Experiencia Pública	0	0,00	4097	100,00	0	0,00
	Familiar Sustituto	0	0,00	4097	100,00	0	0,00
	Enfermedad Catastrófica	2116	51,65	1981	48,35	0	0,00
	Tiene Discapacidad	115	2,81	3982	97,19	0	0,00
	Gestión Lactancia	3248	79,28	849	20,72	0	0,00
	Tiempo Estudio	0	0,00	4097	100,00	0	0,00
	Nº Eventos Aprobados	0	0,00	4097	100,00	0	0,00
	Nº Eventos Asistidos	0	0,00	4097	100,00	0	0,00
	Horas Eventos Aprobados	930	22,70	3167	77,30	0	0,00
	Horas Eventos Asistidos	2191	53,48	1906	46,52	0	0,00
	Nº Eventos Nacionales	0	0,00	4097	100,00	0	0,00
	Nº Eventos Internacionales	0	0,00	4097	100,00	0	0,00
Docente Información Académica 19335 registros	Número Documento	0	0,00	15479	80,00	3857	20,0
	Facultad	0	0,00	19335	100,00	0	0,00
	Carrera	2717	14,00	16619	86,00	0	0,00
	Periodo	0	0,00	19335	100,00	0	0,00
	Actividad Académica	2717	14,00	16619	86,00	0	0,00
	Horas Actividad Académica	0	0,00	19335	100,00	0	0,00

Tabla	Campo	Valores nulos		Valores válidos		Valores no válidos	
	Horas Clase	0	0,00	19335	100,00	0	0,00
Evaluación Docente	Usuario Evaluado	0	0,00	15276	100,00	0	0,00
15276 registros	Tipo Evaluación	0	0,00	15276	100,00	0	0,00
	Componente	0	0,00	15276	100,00	0	0,00
	Periodo	0	0,00	15276	100,00	0	0,00
	Resultado Final	0	0,00	15276	100,00	0	0,00

ANEXO 4: Construir el modelo

La Ilustración 21 muestra el modelo general para aplicar los métodos de regresión.

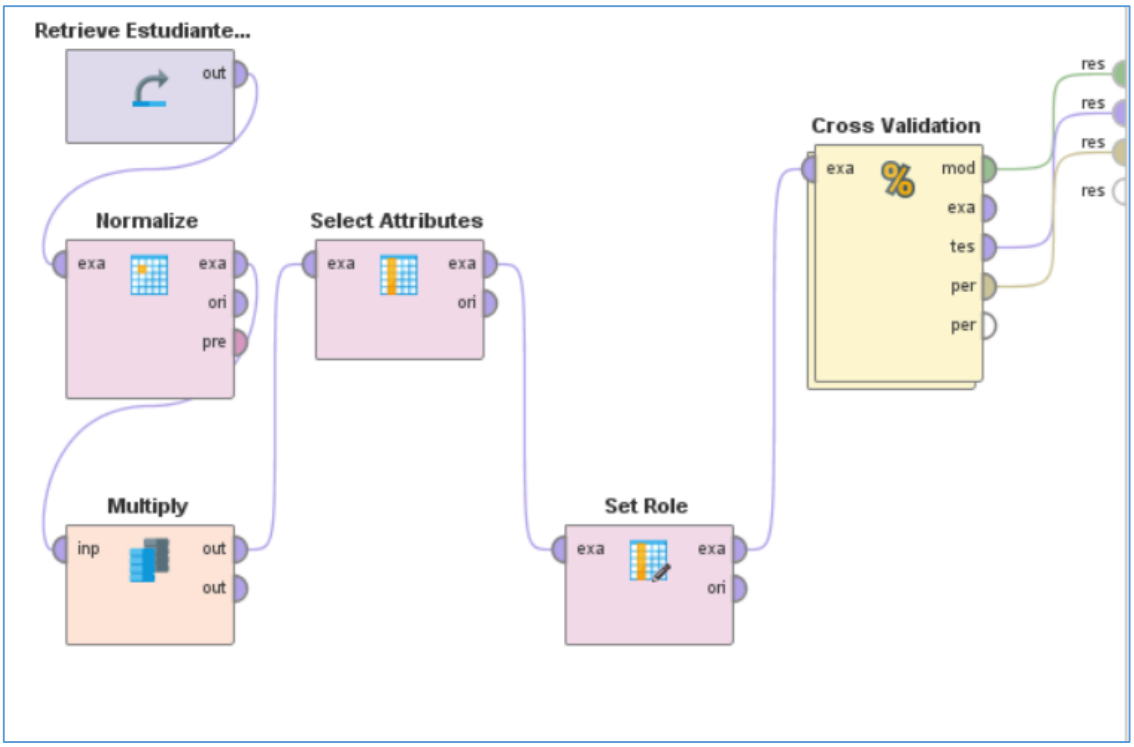


Ilustración 21. Modelo de Regresión

Donde:

- **Tabla de datos:** se utiliza una tabla de datos diferente para Estudiantes y Docentes
- **Normalize:** normaliza los valores de los atributos
- **Multiply:** crea copias de un objeto RapidMiner
- **Select Attributes:** selecciona un subconjunto de atributos de un conjunto de ejemplos y elimina los otros atributos.
- **Set Role:** cambia la función de uno o más atributos. En este caso se selecciona los atributos: Promedio de la tabla Estudiante, Resultado Final Evaluación Docente de la tabla Docente y Total Publicaciones de la tabla Investigación como variables objetivas.
- **Cross Validation:** realiza una validación cruzada para evaluar el rendimiento estadístico de un modelo, en el operador Performance se selecciona las métricas de evaluación. También se elige los métodos de regresión a aplicar, en este caso son:

regresión lineal y regresión polinomial como se muestra en la Ilustración 22 y 23 respectivamente.

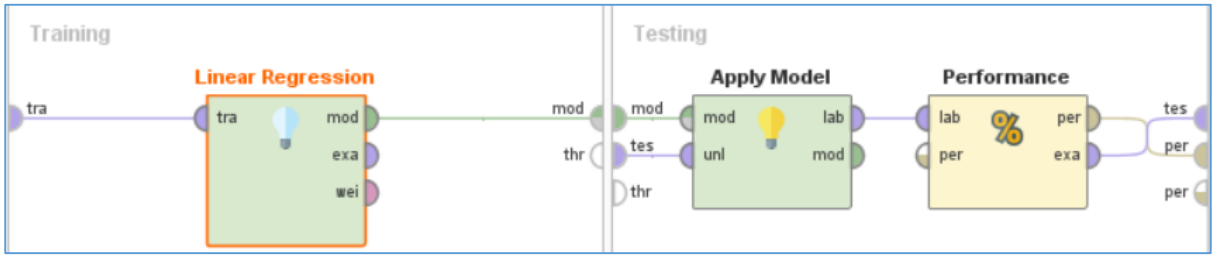


Ilustración 22. Regresión Lineal

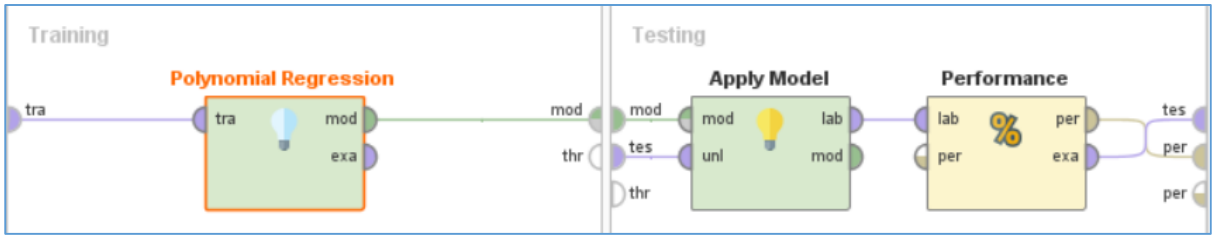


Ilustración 23. Regresión Polinomial

A continuación, se muestra los resultados de la ejecución de los diferentes modelos según los objetivos de minería de datos.

Objetivo 1: Determinar las correlaciones entre las variables cuantitativas y el rendimiento académico de los estudiantes de la Universidad Nacional de Chimborazo.

Modelo 1.1 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de hermanos y la variable promedio.

- **Variable dependiente (Y):** Promedio
- **Variable independiente (X):** Número Hermanos

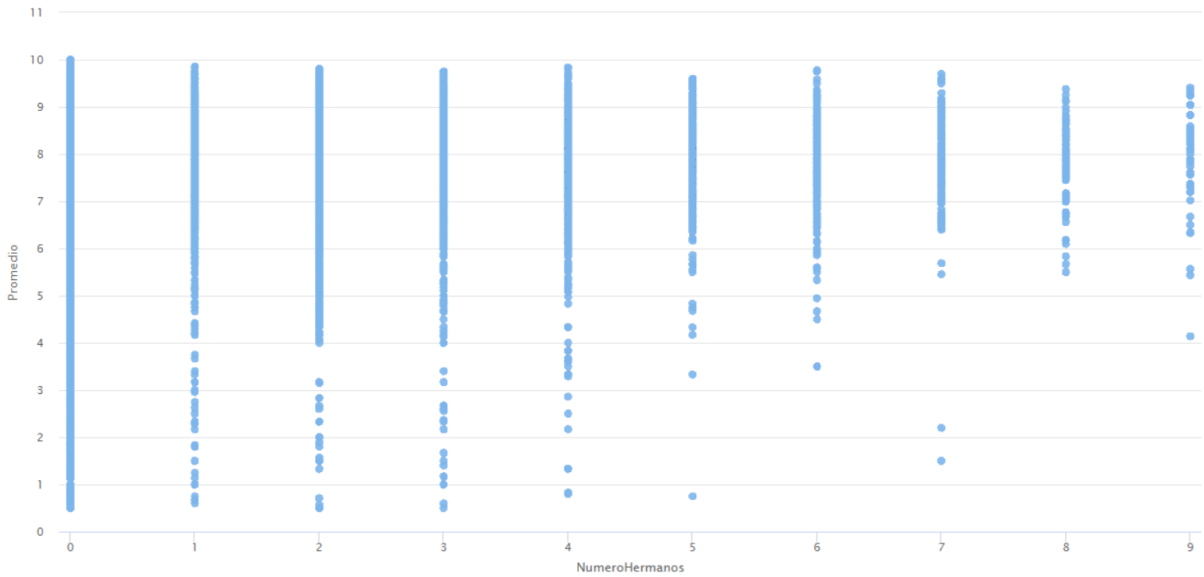


Ilustración 24. Comportamiento datos - Promedio vs Número Hermanos

En la tabla 20 se muestra los resultados que genera este modelo.

Tabla 20. Resultados Modelo 1.1

Métodos	RMSE	AE	R	R^2
Regresión Lineal	0,988	0,679	0,153	0,024
Regresión Polinomial	0,991	0,679	0,125	0,018

Modelo 1.2 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de hijos y la variable promedio.

- **Variable dependiente (Y):** Promedio
- **Variable independiente (X):** Número Hijos

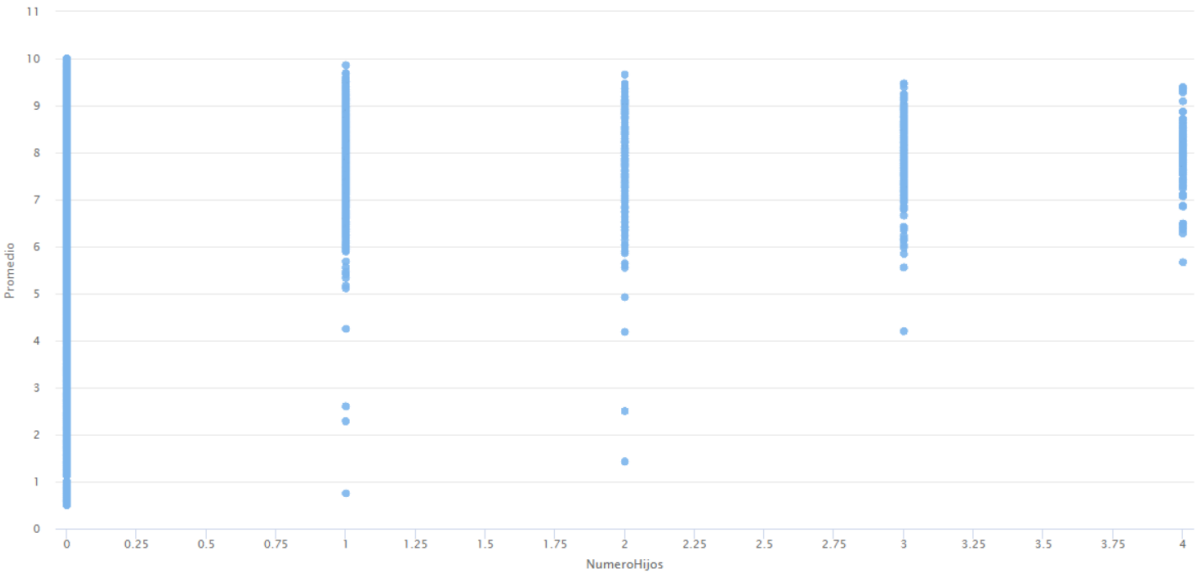


Ilustración 25. Comportamiento datos - Promedio vs Número Hijos

En la tabla 21 se muestra los resultados que genera este modelo.

Tabla 21. Resultados Modelo 1.2

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,998	0,680	0,063	0,004
Regresión Polinomial	0,998	0,680	0,063	0,004

Modelo 1.3 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de integrantes en el hogar y la variable promedio.

- **Variable dependiente (Y):** Promedio
- **Variable independiente (X):** Número Integrantes Hogar

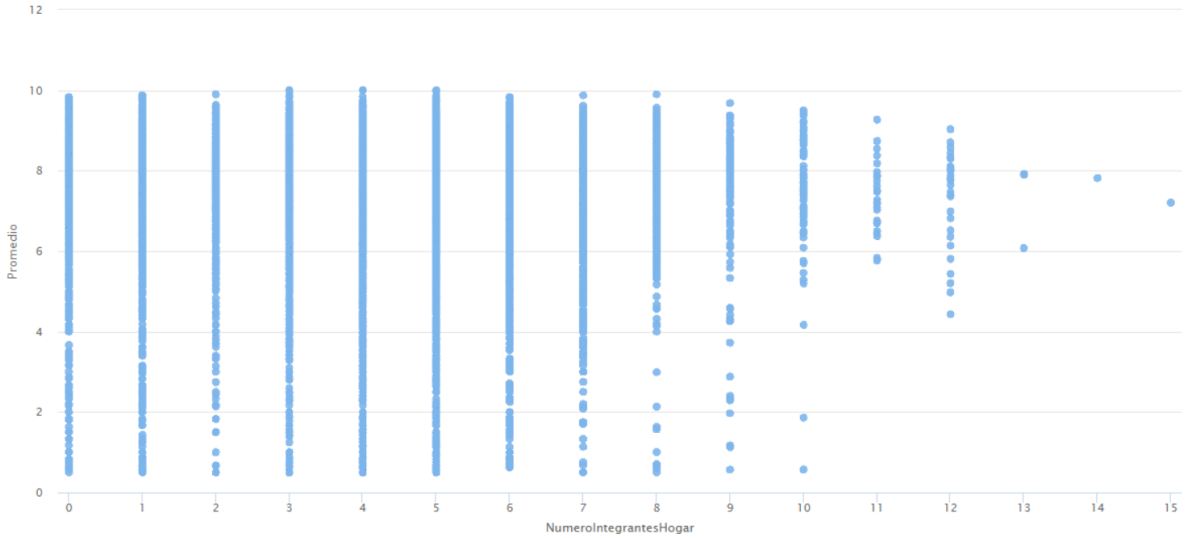


Ilustración 26. Comportamiento datos - Promedio vs Número Integrantes Hogar

En la tabla 22 se muestra los resultados que genera este modelo.

Tabla 22. Resultados Modelo 1.3

Métodos	RMSE	AE	R	R^2
Regresión Lineal	1,000	0,679	0,017	0,000
Regresión Polinomial	1,000	0,679	0,019	0,001

Modelo 1.4 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de personas que dependen de los ingresos y la variable promedio.

- **Variable dependiente (Y):** Promedio
- **Variable independiente (X):** Número Dependen Ingresos

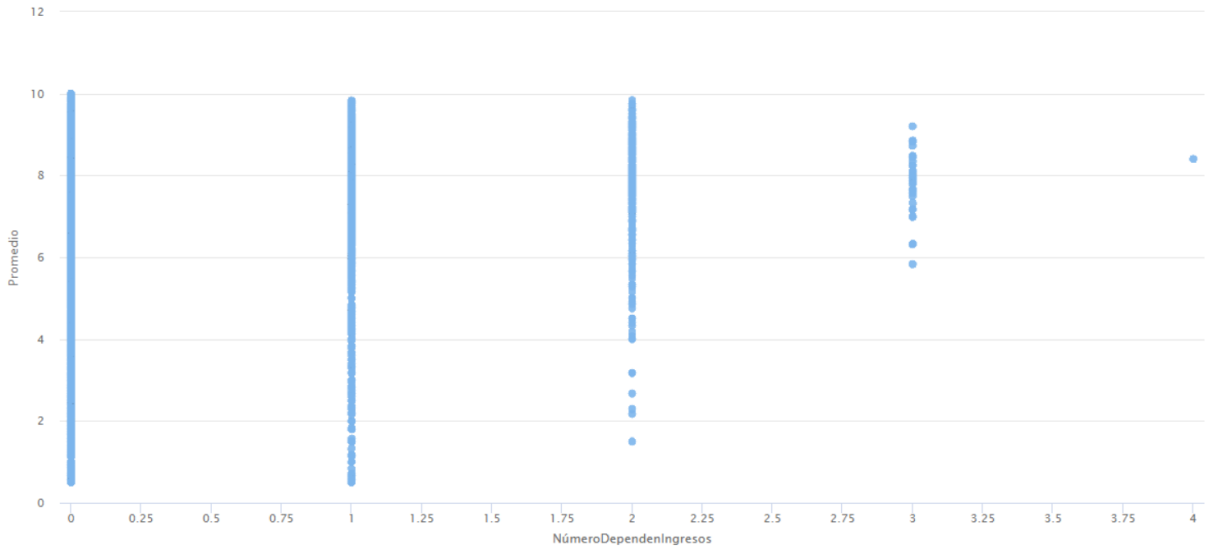


Ilustración 27. Comportamiento datos - Promedio vs Número Dependen Ingresos

En la tabla 23 se muestra los resultados que genera este modelo.

Tabla 23. Resultados Modelo 1.4

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,993	0,679	0,112	0,013
Regresión Polinomial	0,995	0,679	0,099	0,011

Modelo 1.5 - Construcción de un modelo de regresión para determinar la correlación entre la variable total de ingresos y la variable promedio.

- **Variable dependiente (Y):** Promedio
- **Variable independiente (X):** Total Ingresos

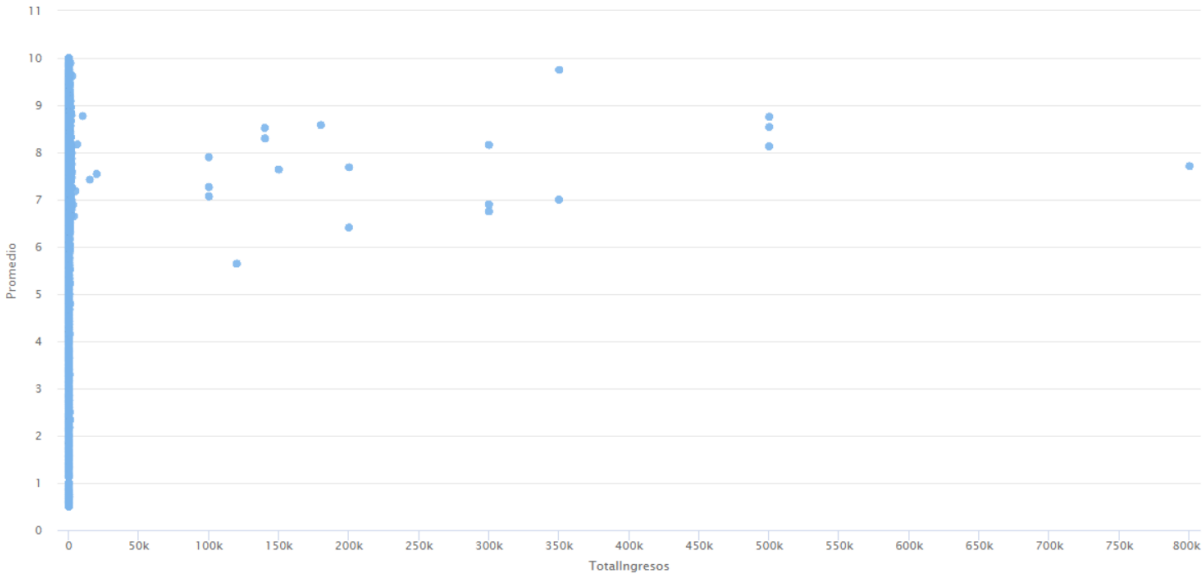


Ilustración 28. Comportamiento datos - Promedio vs Total Ingresos

En la tabla 24 se muestra los resultados que genera este modelo.

Tabla 24. Resultados Modelo 1.5

Métodos	RMSE	AE	R	R^2
Regresión Lineal	1,000	0,679	0,000	0,000
Regresión Polinomial	1,000	0,679	0,009	0,000

Objetivo 2: Determinar las correlaciones entre las variables cuantitativas y el resultado de la evaluación final de los docentes de la Universidad Nacional de Chimborazo.

Modelo 2.1 - Construcción de un modelo de regresión para determinar la correlación entre la variable horas de actividad académica y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Horas Actividad Académica

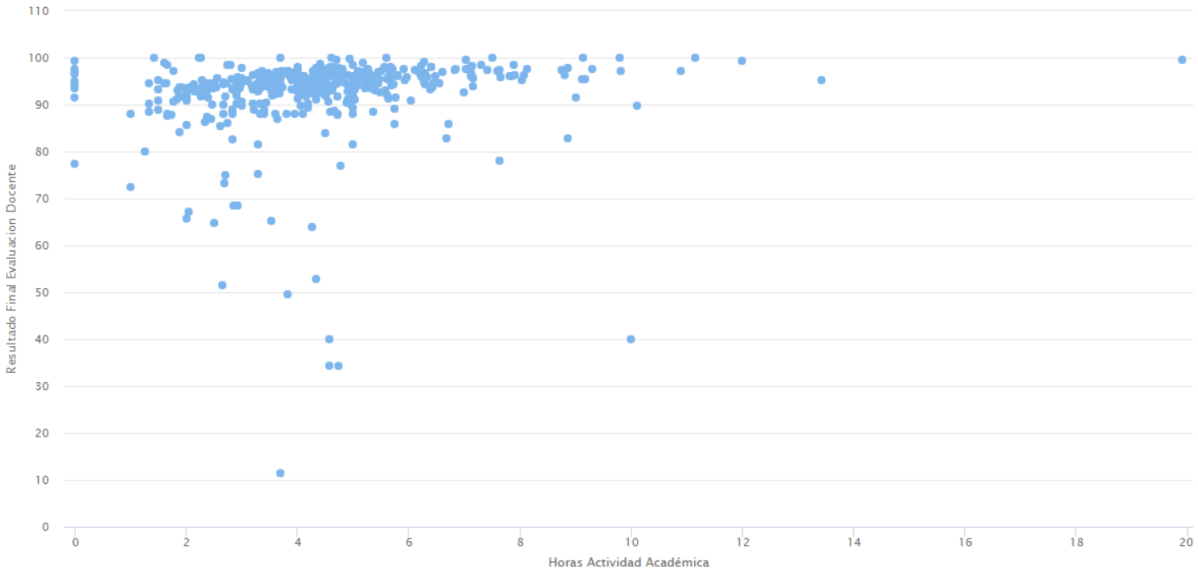


Ilustración 29. Comportamiento datos – Evaluación Docente vs Horas Actividad Académica

En la tabla 25 se muestra los resultados que genera este modelo.

Tabla 25. Resultados Modelo 2.1

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,920	0,498	0,185	0,055
Regresión Polinomial	0,920	0,498	0,185	0,055

Modelo 2.2 - Construcción de un modelo de regresión para determinar la correlación entre la variable horas de clase y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Horas Clase

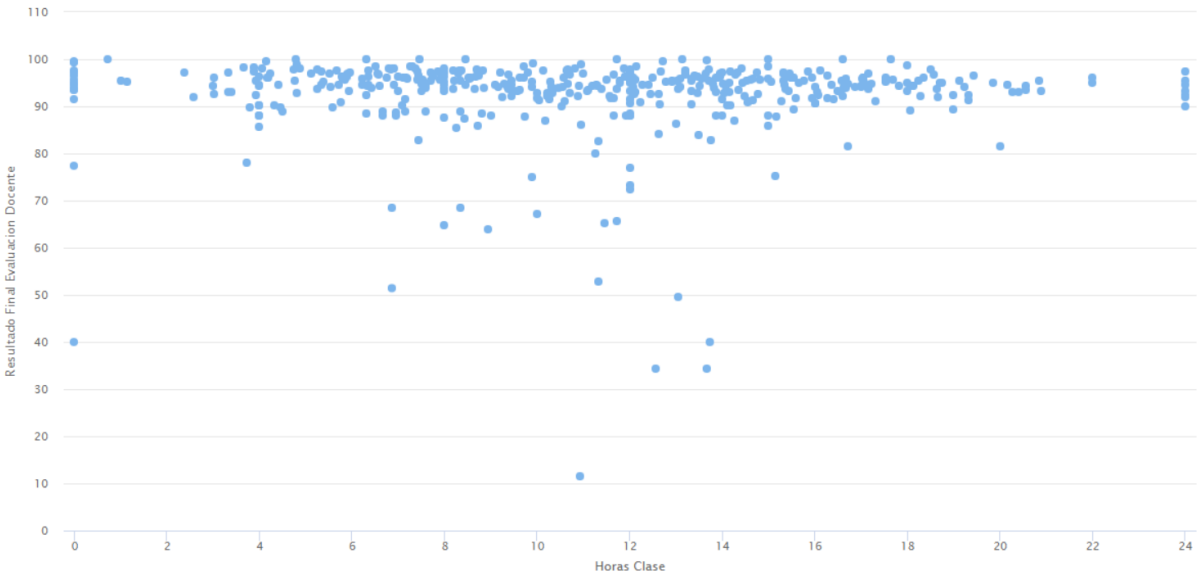


Ilustración 30. Comportamiento datos - Evaluación Docente vs Horas Clase

En la tabla 26 se muestra los resultados que genera este modelo.

Tabla 26. Resultados Modelo 2.2

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,929	0,514	0,000	0,000
Regresión Polinomial	0,930	0,516	0,011	0,003

Modelo 2.3 - Construcción de un modelo de regresión para determinar la correlación entre la variable horas de eventos aprobados y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Horas Eventos Aprobados

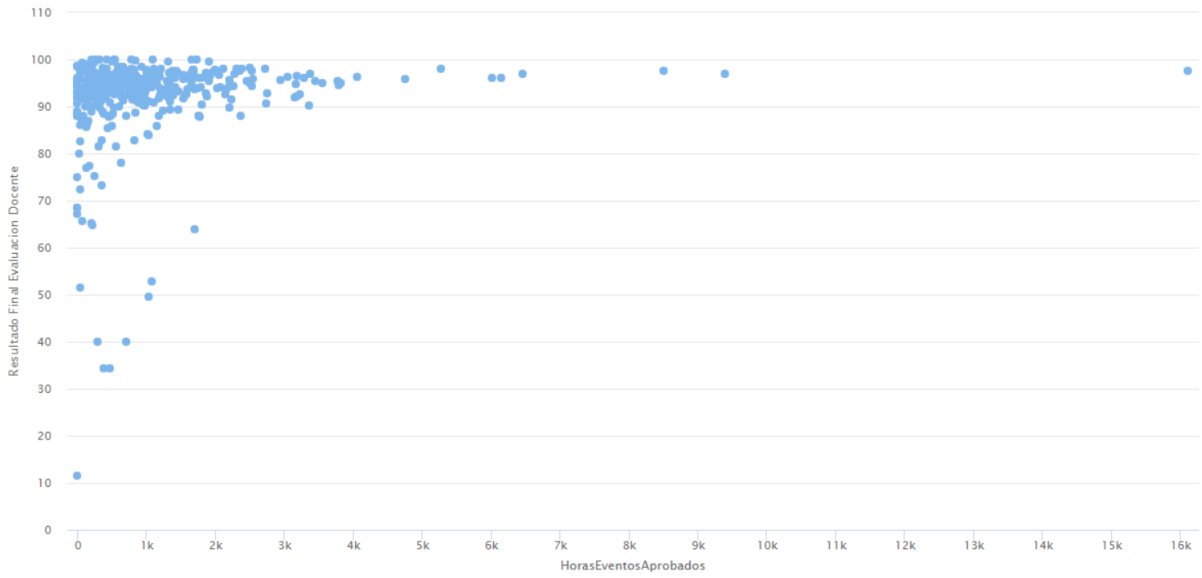


Ilustración 31. Comportamiento datos - Evaluación Docente vs Horas Eventos Aprobados

En la tabla 27 se muestra los resultados que genera este modelo.

Tabla 27. Resultados Modelo 2.3

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,923	0,513	0,162	0,031
Regresión Polinomial	0,923	0,513	0,162	0,031

Modelo 2.4 - Construcción de un modelo de regresión para determinar la correlación entre la variable horas de eventos asistidos y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Horas Eventos Asistidos

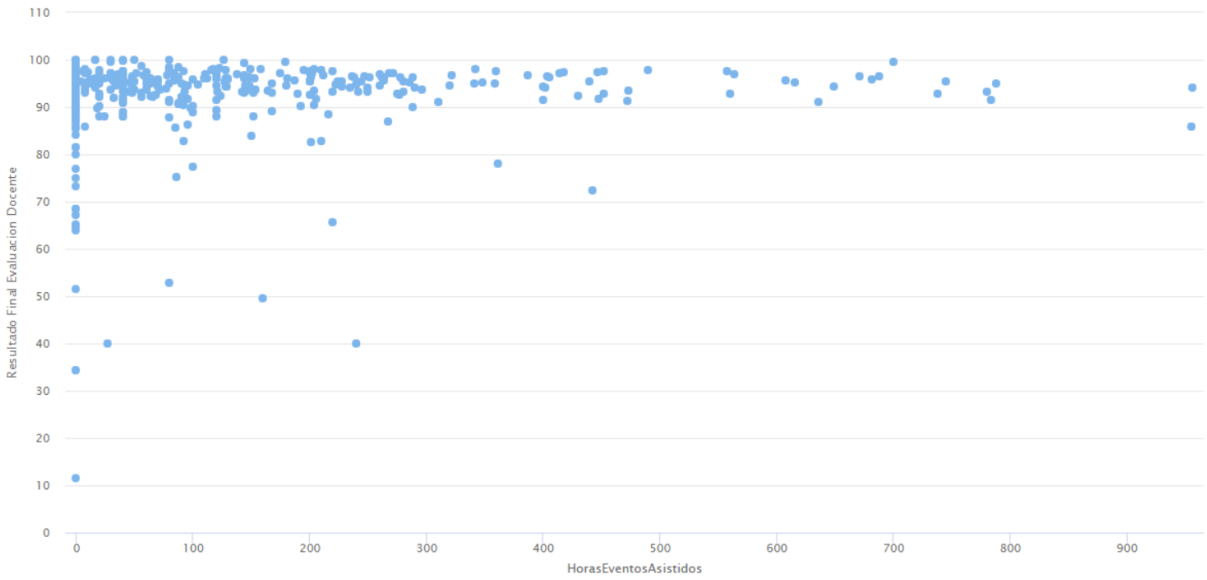


Ilustración 32. Comportamiento datos – Evaluación Docente vs Horas Eventos Asistidos

En la tabla 28 se muestra los resultados que genera este modelo.

Tabla 28. Resultados Modelo 2.4

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,932	0,517	0,010	0,001
Regresión Polinomial	0,931	0,518	0,026	0,009

Modelo 2.5 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de eventos aprobados y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Nº Eventos Aprobados

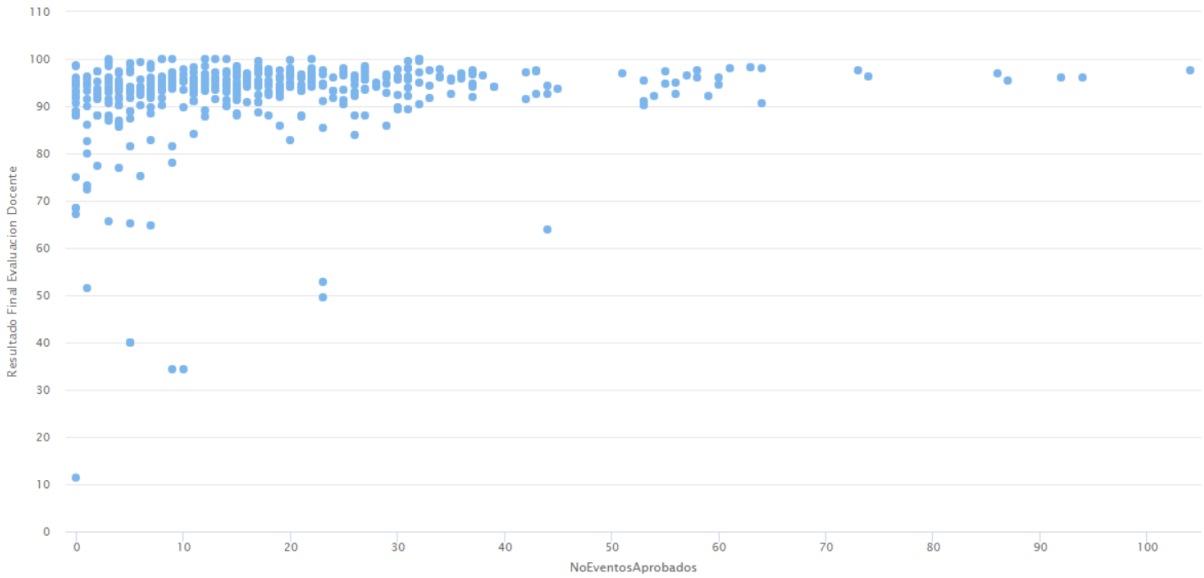


Ilustración 33. Comportamiento datos - Evaluación Docente vs Nº Eventos Aprobados

En la tabla 29 se muestra los resultados que genera este modelo.

Tabla 29. Resultados Modelo 2.5

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,921	0,515	0,182	0,042
Regresión Polinomial	0,924	0,514	0,164	0,037

Modelo 2.6 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de eventos asistidos y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Nº Eventos Asistidos

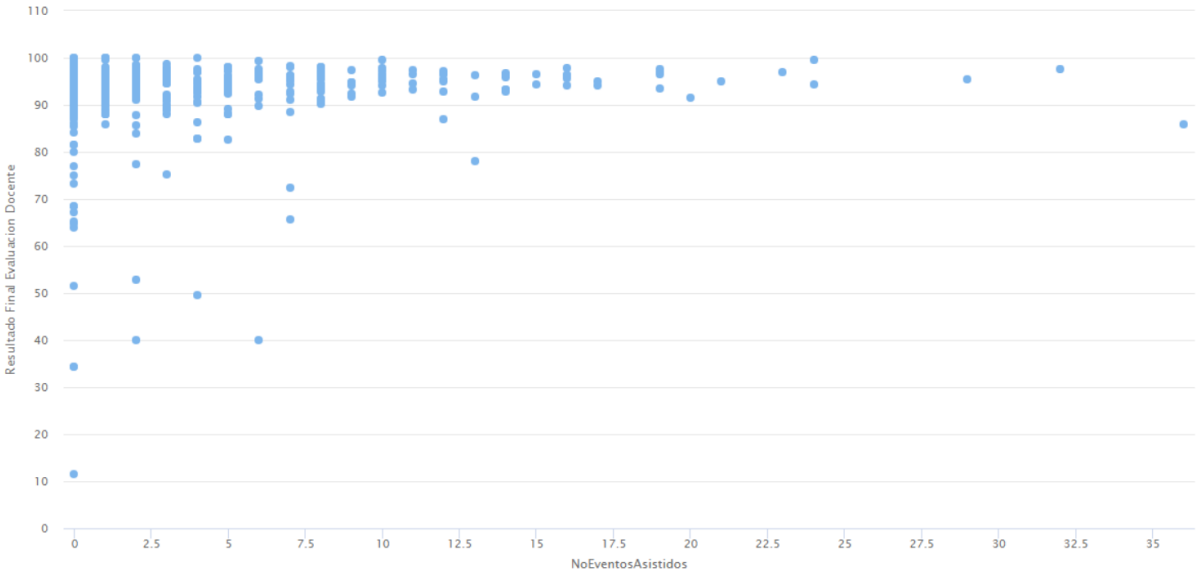


Ilustración 34. Comportamiento datos – Evaluación Docente vs Nº Eventos Asistidos

En la tabla 30 se muestra los resultados que genera este modelo.

Tabla 30. Resultados Modelo 2.6

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,927	0,513	0,109	0,018
Regresión Polinomial	0,930	0,516	0,105	0,016

Modelo 2.7 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de eventos internacionales y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Nº Eventos Internacionales

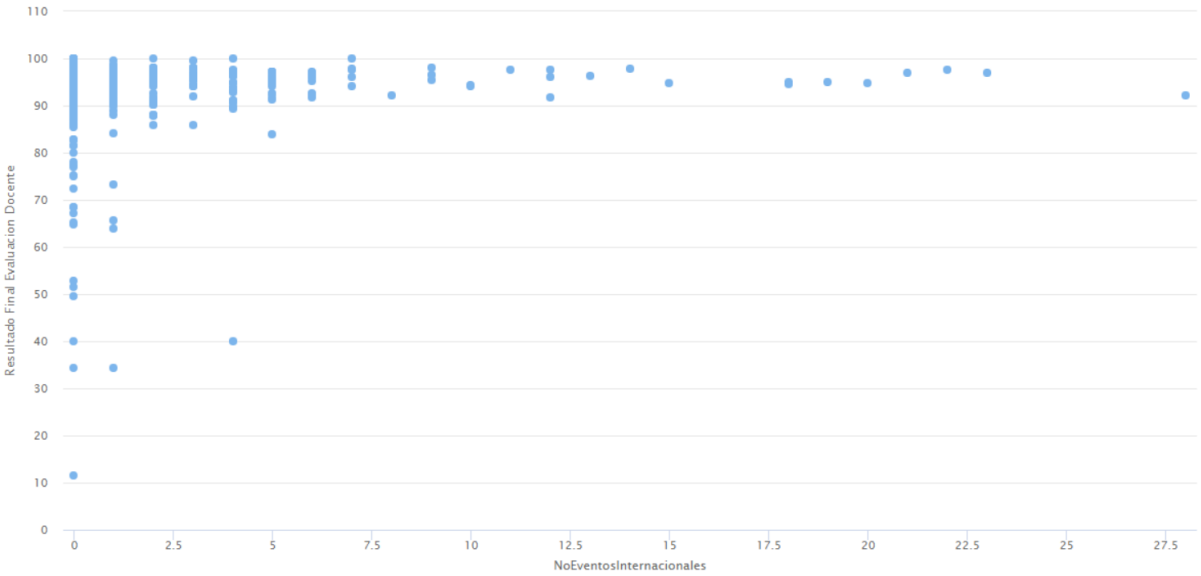


Ilustración 35. Comportamiento datos - Evaluación Docente vs Nº Eventos Internacionales

En la tabla 31 se muestra los resultados que genera este modelo.

Tabla 31. Resultados Modelo 2.7

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,926	0,516	0,127	0,020
Regresión Polinomial	0,927	0,518	0,118	0,018

Modelo 2.8 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de eventos nacionales y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Nº Eventos Nacionales

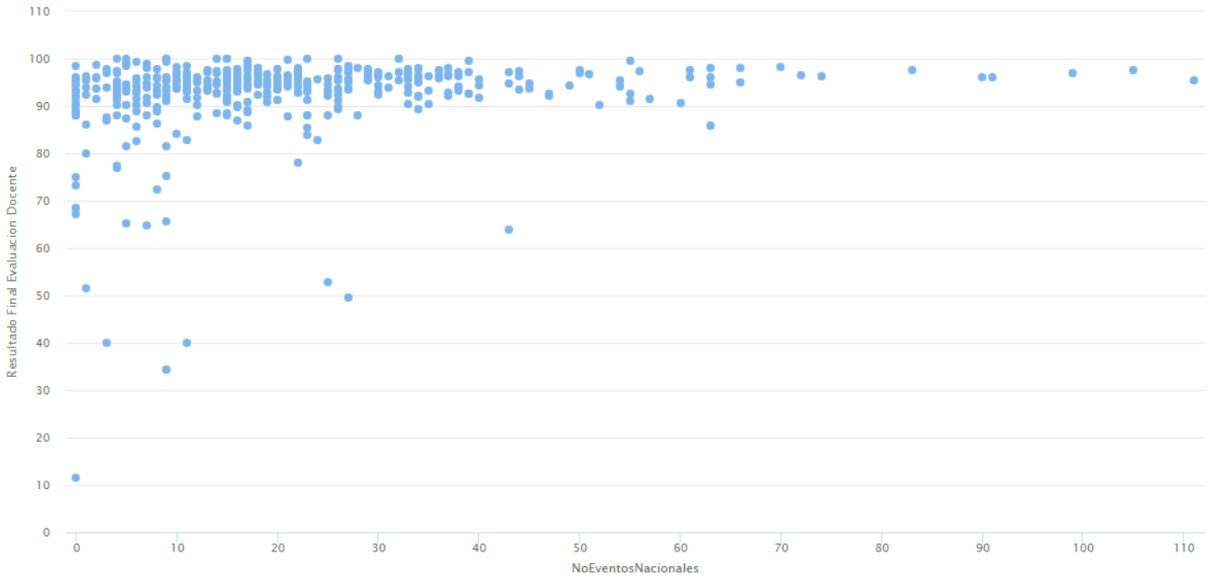


Ilustración 36. Comportamiento datos - Evaluación Docente vs Nº Eventos Nacionales

En la tabla 32 se muestra los resultados que genera este modelo.

Tabla 32. Resultados Modelo 2.8

Métodos	RMSE	AE	R	R^2
Regresión Lineal	0,921	0,514	0,179	0,043
Regresión Polinomial	0,921	0,514	0,179	0,043

Modelo 2.9 - Construcción de un modelo de regresión para determinar la correlación entre la variable número de hijos y la variable resultado final de evaluación.

- **Variable dependiente (Y):** Resultado Final Evaluación Docente
- **Variable independiente (X):** Número Hijos

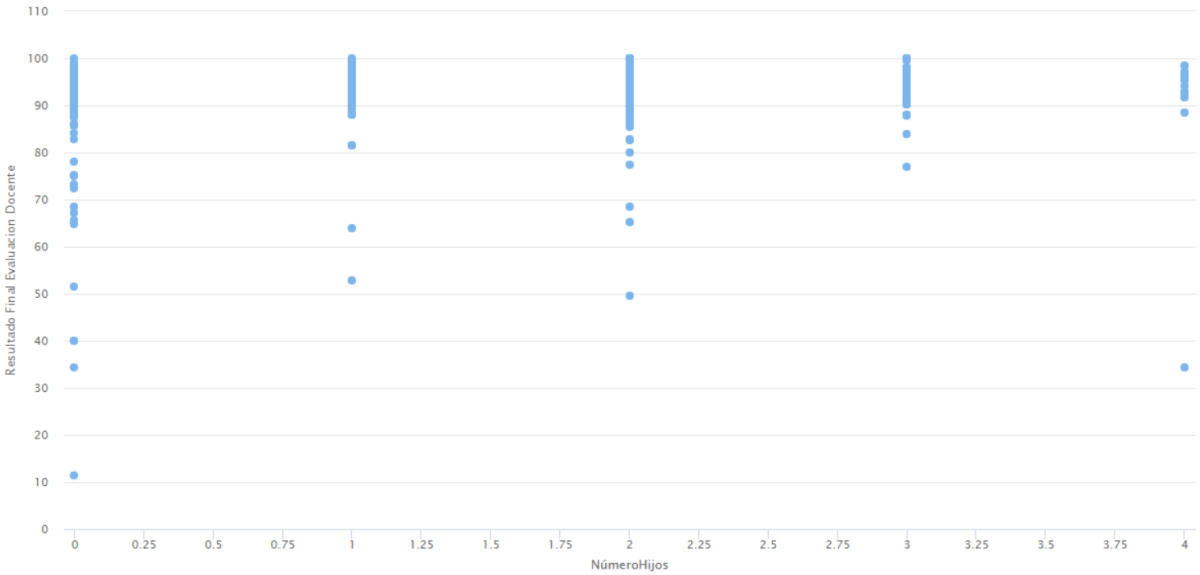


Ilustración 37. Comportamiento datos - Evaluación Docente vs Número Hijos

En la tabla 33 se muestra los resultados que genera este modelo.

Tabla 33. Resultados Modelo 2.9

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,924	0,516	0,164	0,044
Regresión Polinomial	0,926	0,518	0,164	0,051

Los siguientes modelos se han considerado como adicionales para estudiantes:

Modelo 3.1 - Construcción de un modelo de regresión para determinar la correlación entre la variable promedios de los estudiantes de quinto semestre y la variable promedio general.

- **Variable dependiente (Y):** Promedio General
- **Variable independiente (X):** Promedio Quinto Semestre

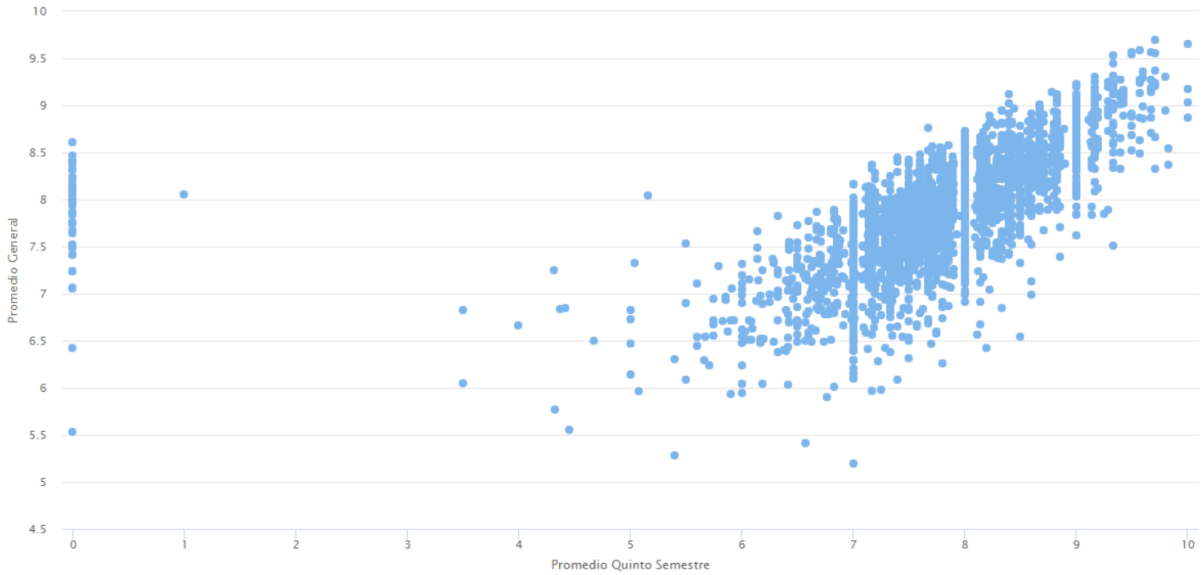


Ilustración 38. Comportamiento datos - Promedio General vs Promedio Quinto Semestre

En la tabla 34 se muestra los resultados que genera este modelo.

Tabla 34. Resultados Modelo 3.1

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,522	0,392	0,516	0,274
Regresión Polinomial	0,869	0,652	0,507	0,267

Modelo 3.2 - Construcción de un modelo de regresión para determinar la correlación entre la variable promedios de los estudiantes de sexto semestre y la variable promedio general.

- **Variable dependiente (Y):** Promedio General
- **Variable independiente (X):** Promedio Sexto Semestre

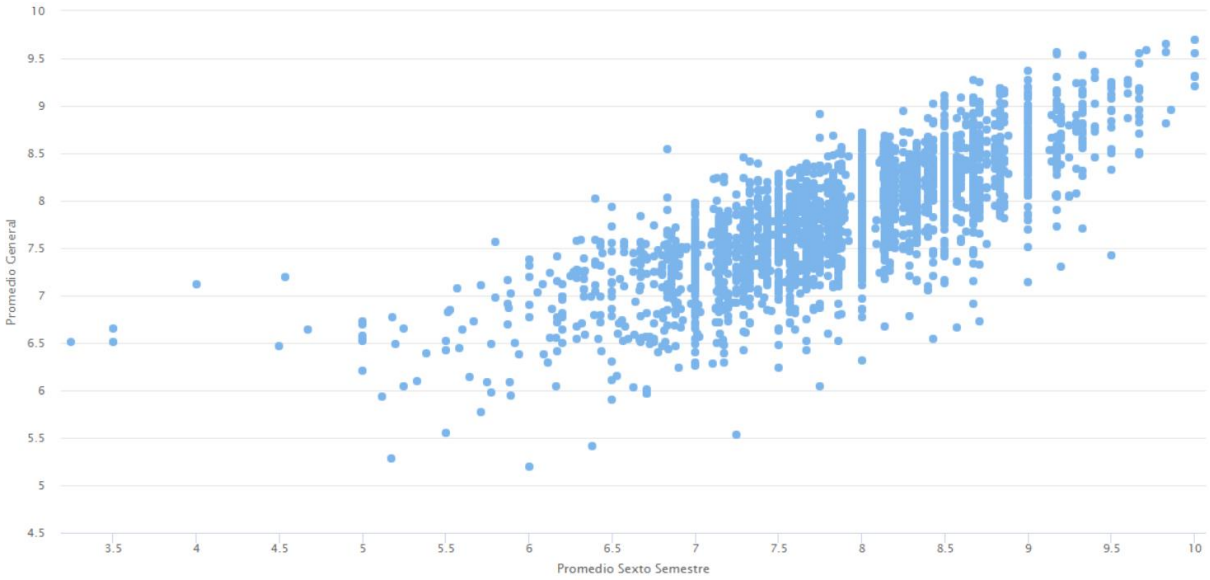


Ilustración 39. Comportamiento datos – Promedio General vs Promedio Sexto Semestre

En la tabla 35 se muestra los resultados que genera este modelo.

Tabla 35. Resultados Modelo 3.2

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,384	0,297	0,770	0,594
Regresión Polinomial	0,701	0,544	0,704	0,515

Modelo 3.3 - Construcción de un modelo de regresión para determinar la correlación entre la variable promedios de los estudiantes de séptimo semestre y la variable promedio general.

- **Variable dependiente (Y):** Promedio General
- **Variable independiente (X):** Promedio Séptimo Semestre

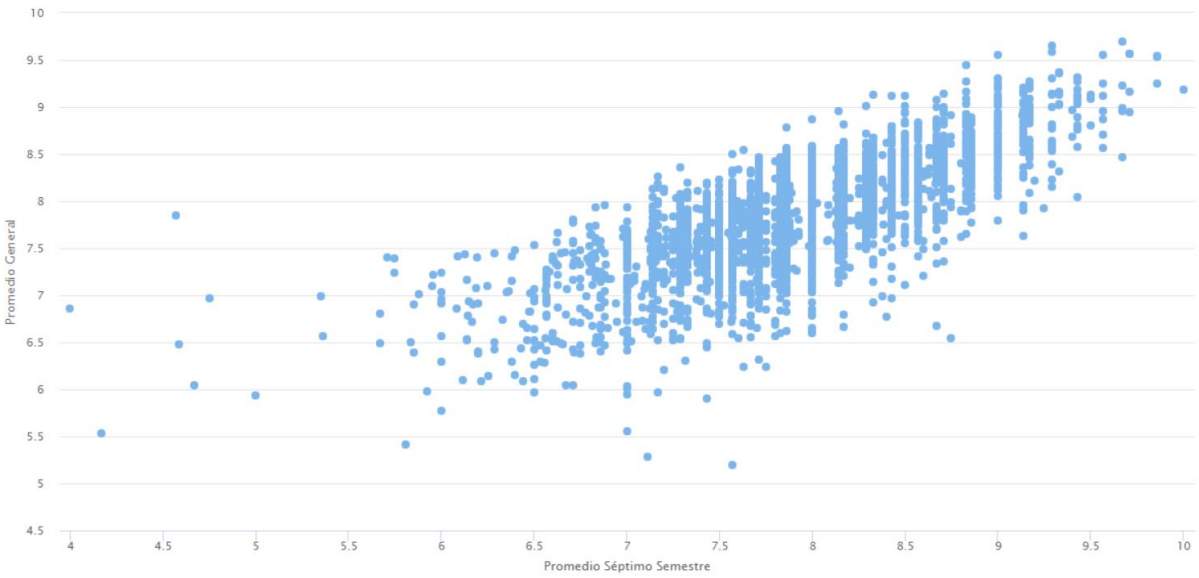


Ilustración 40. Comportamiento datos – Promedio General vs Promedio Séptimo Semestre

En la tabla 36 se muestra los resultados que genera este modelo.

Tabla 36. Resultados Modelo 3.3

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,391	0,303	0,762	0,582
Regresión Polinomial	0,717	0,558	0,633	0,464

Los siguientes modelos se han considerado como adicionales para docentes:

Modelo 4.1 - Construcción de un modelo de regresión para determinar la correlación entre la variable evaluación por componente docencia y la variable resultado final de evaluación docente.

- **Variable dependiente (Y):** Resultado Final de Evaluación Docente
- **Variable independiente (X):** Evaluación Componente Docencia

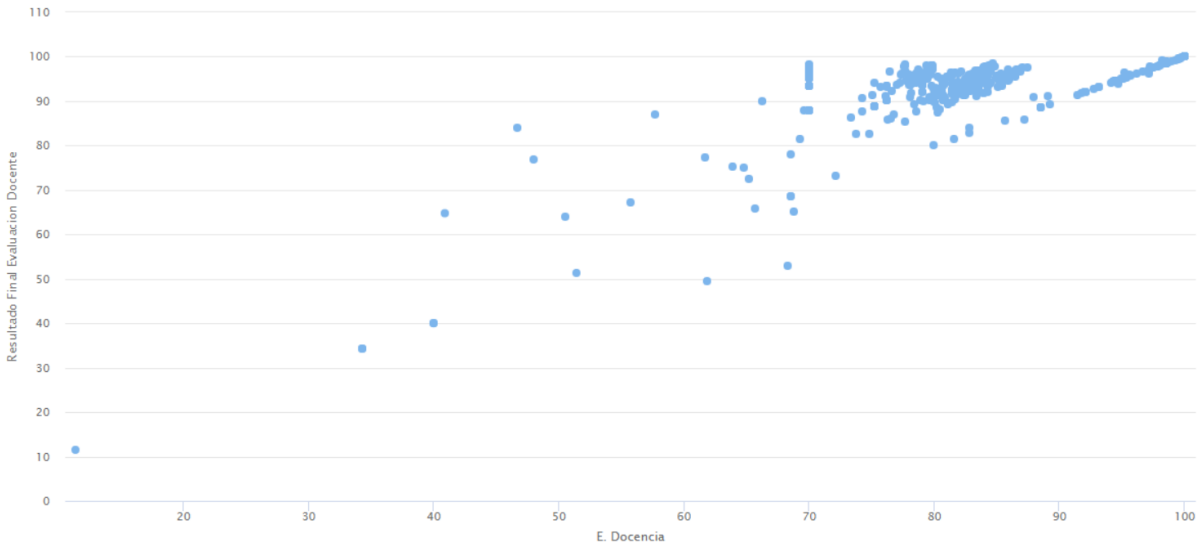


Ilustración 41. Comportamiento datos - Evaluación Final Docente vs Evaluación Docencia

En la tabla 37 se muestra los resultados que genera este modelo.

Tabla 37. Resultados Modelo 4.1

Métodos	**RMSE**	**AE**	**R**	**R^2**
Regresión Lineal	0,708	0,482	0,642	0,439
Regresión Polinomial	0,696	0,475	0,655	0,457

Modelo 4.2 - Construcción de un modelo de regresión para determinar la correlación entre la variable evaluación por componente gestión y la variable resultado final de evaluación docente.

- **Variable dependiente (Y):** Resultado Final de Evaluación Docente
- **Variable independiente (X):** Evaluación Componente Gestión

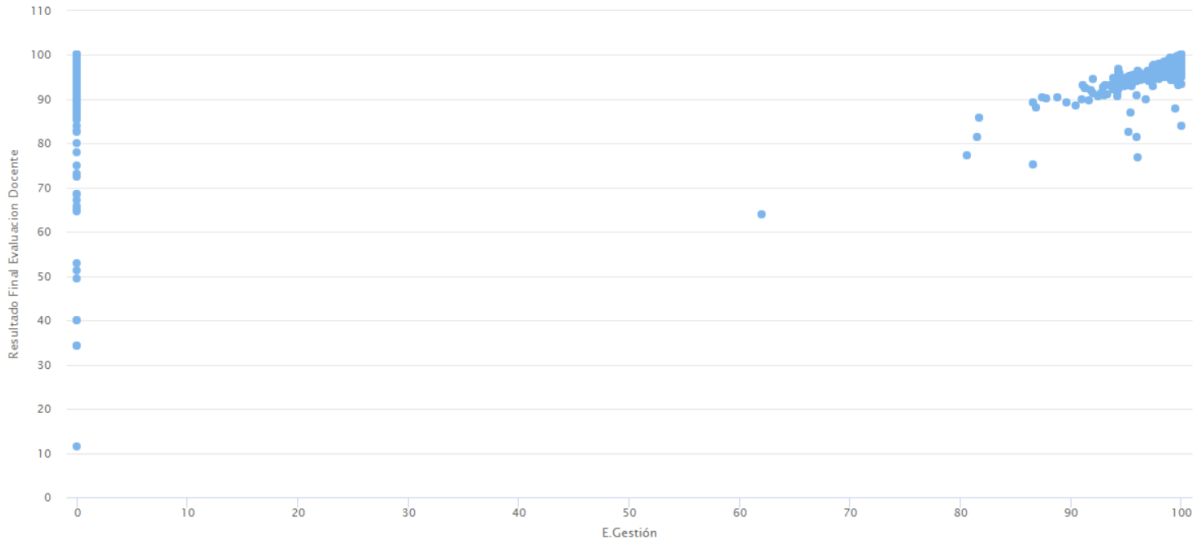


Ilustración 42. Comportamiento datos – Evaluación Final Docente vs Evaluación Gestión

En la tabla 38 se muestra los resultados que genera este modelo.

Tabla 38. Resultados Modelo 4.2

Métodos	RMSE	AE	R	R^2
Regresión Lineal	0,894	0,485	0,318	0,115
Regresión Polinomial	0,931	0,517	0,237	0,078

Modelo 4.2 - Construcción de un modelo de regresión para determinar la correlación entre la variable evaluación por componente investigación y la variable resultado final de evaluación docente.

- **Variable dependiente (Y):** Resultado Final de Evaluación Docente
- **Variable independiente (X):** Evaluación Componente Investigación

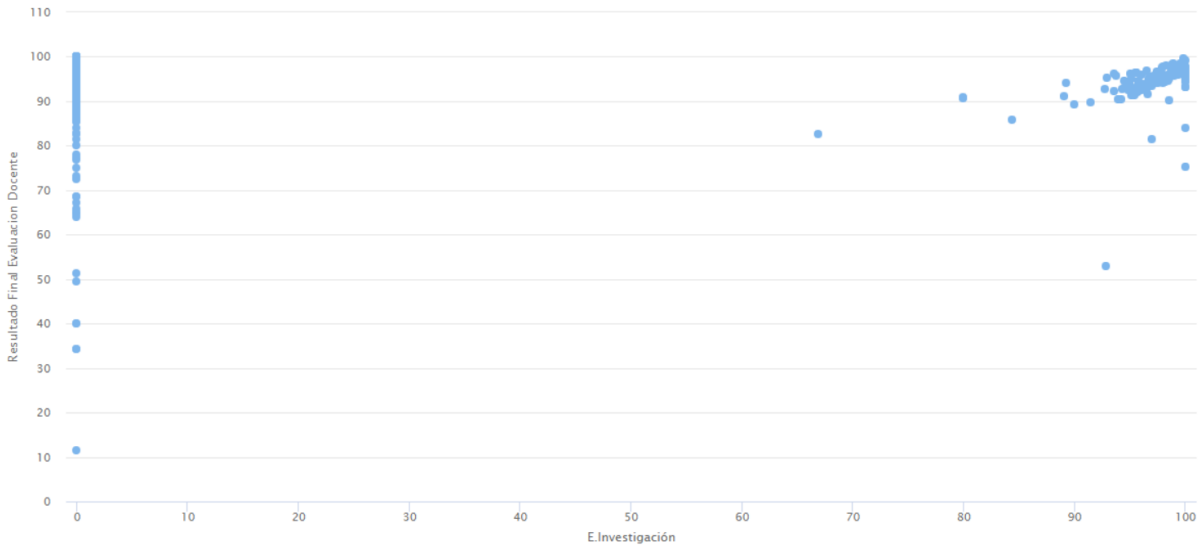


Ilustración 43. Comportamiento datos – Evaluación Final Docente vs Evaluación Investigación

En la tabla 39 se muestra los resultados que genera este modelo.

Tabla 39. Resultados Modelo 4.3

Métodos	RMSE	AE	R	R^2
Regresión Lineal	0,923	0,506	0,150	0,029
Regresión Polinomial	0,920	0,503	0,164	0,034

ANEXO 5: Evaluar el modelo

Tabla 40. Valores de las diferentes métricas y modelos

	RMSE RL/RP	**AE** RL/RP	**R** RL/RP	**R^2** RL/RP	**Mejor método**
Modelo 1.1	0,988/0,991	0,679/0,679	0,153/0,125	0,024/0,018	RL
Modelo 1.2	0,998/0,998	0,680/0,680	0,063/0,063	0,004/0,004	
Modelo 1.3	1,000/1,000	0,679/0,679	0,017/0,019	0,000/0,001	RP
Modelo 1.4	0,993/0,995	0,679/0,679	0,112/0,099	0,013/0,011	RL
Modelo 1.5	1,000/1,000	0,679/0,679	0,000/0,009	0,000/0,000	
Modelo 2.1	0,920/0,920	0,498/0,498	0,185/0,185	0,055/0,055	
Modelo 2.2	0,929/0,930	0,514/0,516	0,000/0,011	0,000/0,003	RL
Modelo 2.3	0,923/0,923	0,513/0,513	0,162/0,162	0,031/0,031	
Modelo 2.4	0,932/0,931	0,517/0,518	0,010/0,026	0,001/0,009	RP
Modelo 2.5	0,921/0,924	0,515/0,514	0,182/0,164	0,042/0,037	RL
Modelo 2.6	0,927/0,930	0,513/0,516	0,109/0,105	0,018/0,016	RL
Modelo 2.7	0,926/0,927	0,516/0,518	0,127/0,118	0,020/0,018	RL
Modelo 2.8	0,921/0,921	0,514/0,514	0,179/0,179	0,043/0,043	
Modelo 2.9	0,924/0,926	0,516/0,518	0,164/0,164	0,044/0,051	RL
Modelo 3.1	0,522/0,869	0,392/0,652	0,516/0,507	0,274/0,267	RL
Modelo 3.2	0,384/0,701	0,297/0,544	0,770/0,704	0,594/0,515	RL
Modelo 3.3	0,391/0,717	0,303/0,558	0,762/0,633	0,582/0,464	RL
Modelo 4.1	0,708/0,696	0,482/0,475	0,642/0,655	0,439/0,457	RP
Modelo 4.2	0,894/0,931	0,485/0,517	0,318/0,237	0,115/0,078	RL
Modelo 4.3	0,923/0,920	0,506/0,503	0,150/0,164	0,029/0,034	RP

RL= Regresión Lineal, **RP**=Regresión Polinomial