UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Modelos de Machine Learning para la predicción del salario en docentes peruanos de educación básica regular

Tesis para obtener el Título Profesional de Ingeniero de Sistemas

Autor:

José Luis Tinoco Ramos
Jhoset Yamiel Yupanqui Arellano

Asesor:

Dr. Juan Jesús Soria Quijaite

Lima, junio de 2024

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Dr. Juan Jesús Soria Quijaite, docente de la Facultad de Ingeniería y Arquitectura, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que la presente investigación titulada: **"Modelos de Machine Learning para la predicción del salario en docentes peruanos de educación básica regular"** de los autores José Luis Tinoco Ramos y Jhoset Yamiel Yupanqui Arellano tiene un índice de similitud de 5% verificable en el informe del programa Turnitin, y fue realizada en la Universidad Peruana Unión bajo mi dirección.

En tal sentido asumo la responsabilidad que corresponde ante cualquier falsedad u omisión de los documentos como de la información aportada, firmo la presente declaración en la ciudad de Lima a los 26 días del mes de marzo del año 2024.

Dr. Soria Quijaite Juan Jesús

235

ACTA DE SUSTENTACIÓN DE TESIS

En Lima, Ñaña, Villa Unión, a 05 día(s) del mes de junio del año 2024 siendo las 08:30 horas, se reunieron los miembros del jurado en la Universidad Peruana Unión Campus Lima, bajo la dirección del (de la) presidente(a): Mg. Geraldine Verónica Alvizuri Llerena, el (la) secretario(a): Mg. Fernando Manuel Asin Gomez y los demás miembros: Mg. Nemias Saboya Rios Mg. Ferdinan Edgardo Pineda Anco y el (la) asesor(a) Dr. Juan Jesus Soria Quijaite con el propósito de administrar el acto académico de sustentación de la tesis titulado: "Modelos de Machine Learning para la predicción del salario en docentes peruanos de educación básica regular" del(los) bachiller/es: a) Jhoset Yamiel Yupanqui Arellano b) Jose Luis Tinoco Ramos c) conducente a la obtención del título profesional de: Ingeniero de Sistemas (Denominación del Título Profesional)

El Presidente inició el acto académico de sustentación invitando al (a la) / a (los) (las) candidato(a)/s hacer uso del tiempo determinado para su exposición. Concluida la exposición, el Presidente invitó a los demás miembros del jurado a efectuar las preguntas, y aclaraciones pertinentes, las cuales fueron absueltas por al (a la) / a (los) (las) candidato(a)/s. Luego, se produjo un receso para las deliberaciones y la emisión del dictamen del jurado.

Posteriormente, el jurado procedió a dejar constancia escrita sobre la evaluación en la presente acta, con el dictamen siguiente:

Bachiller (a): Jhoset Yamiel Yupanqui Arellano

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	16.9	B	Bueno	Muy bueno

Bachiller (b): Jose Luis Tinoco Ramos

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	16.9	B	Bueno	Muy bueno

Bachiller (c):

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	

(*) Ver parte posterior

Finalmente, el Presidente del jurado invitó al (a la) / a (los) (las) candidato(a)/s a ponerse de pie, para recibir la evaluación final y concluir el acto académico de sustentación procediéndose a registrar las firmas respectivas.

Presidente/a

Secretario/a

Asesor/a

Miembro

Miembro

Bachiller (a)

Bachiller (b)

Bachiller (c)

* Esta sustentación fue realizada de manera virtual u online sincrónica según conforme al Reglamento General de Grados y Títulos

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La investigación proporciona un análisis profundo de la predicción del salario docente peruano, utilizando datos de la UGEL Ventanilla en Lima Perú y aplicando varios algoritmos predictivos de aprendizaje automático. A pesar del desafiante contexto de la variabilidad salarial en las organizaciones educativas, el estudio logró un alto grado de precisión, con el modelo de regularización Elastic Net a la cabeza. La investigación recopiló 108 317 registros docentes nombrados en cinco años correspondientes de 2018 - 2023, tomando el 80% (86 654) para el entrenamiento y el 20% (21 663) para el testeo de los modelos en estudio, con el objetivo de identificar la precisión de los algoritmos predictivos de machine learning Regresión lineal, Lasso, Ridge y Elastic Net a partir del análisis del salario docente. La investigación analizó la edad, el nivel educativo, el tiempo de servicio, la escala docente y las horas laborales como regresoras y el salario docente como predictor en un entorno normalizado por la exigencia de los supuestos inferenciales que fueron significativas estadísticamente, encontrando un salario promedio de 2771.80 soles peruanos y un modelo de regresión lineal múltiple significativo con pvalue menor a 2.2e-16, un RMSE=895.3793, MAE=619.7701, regresión Ridge con un RMSE=896.5645, MAE=622.6167, regresión Lasso con un RMSE=895.3673, MAE=619.8510, regresión Elastic Net, con un RMSE=895.3870 y MAE=619.8605. Los resultados indican que el algoritmo predictivo óptimo fue el modelo Elastic Net con $\alpha = 0.5555556$ y $\lambda_{13} = 0.20$ con coeficientes $\beta_0 = -3092.582975$; $\beta_1 = -4.824496$; $\beta_2 = 22.972778$; $\beta_3 = 17.623234$; $\beta_4 = -88.511756$; $\beta_5 = 191.104877$ y un RMSE de 895.3870 aplicado en un entorno del salario docente.

Palabras Clave: Regularización, Machine learning, Salario Docente, Lasso, Ridge, Elastic net.

ABSTRACT

This paper presents an analysis of machine learning (ML) models to predict the salaries of 108,317 appointed teachers in Ventanilla, Lima, Peru, using recent data. The focal point of the study is appointed teachers, deliberately excluding salaries of hired teachers' from the analysis. A significant result of this research is the identification of a new ML model capable of predicting teacher salaries with considerable accuracy, based on regressor variables closely related to salary. This finding is noteworthy because it fills a gap in existing ML applications for salary prediction, indicating a promising direction for future research in this area. The methodology used to analyze the wage data, while comprehensive, does not account for gender differences, which may affect wage variation over the five-year period considered. This oversight suggests that future research should include a wider range of variables, including gender, to improve the accuracy and applicability of salary predictions for both appointed and contract faculty. Such an approach could provide more nuanced information on the factors influencing teacher salaries and help develop more equitable and effective salary models. One of the key contributions of the article is the detailed examination of the factors influencing salaries of appointed teachers, including age, educational level, length of service, teaching scale, and hours worked. The use of linear regression, Ridge, Lasso, and Elastic Net models yielded accurate metrics for choosing the best model for salary prediction. This research not only advances our understanding of the determinants of teacher salaries in Peru, but also provides a valuable framework for similar studies in other contexts. Comparison with other research highlights the robustness of the chosen ML models, underscoring the potential of ML in educational administration and policymaking.

Keywords: Regularization, Machine learning, Teaching Salary, Lasso, Ridge, Elastic net.

1 INTRODUCCIÓN

El salario de un profesional desempeña un papel fundamental en la existencia de los empleados de toda organización, ya que constituye la única compensación por sus labores, abarcando todos los aspectos que representan un beneficio económico. Este ingreso se rige como la exclusiva fuente que posibilita afrontar con dignidad las necesidades esenciales de índole diversa, tanto las propias como las de su núcleo familiar [1]. Asimismo en todo el mundo las disparidades de ingresos se han convertido en un problema importante [2]. En ese sentido hay diversas razones por las que en la actualidad los empleados cambian de empleos, uno de los principales factores es el salario, más aun en un mundo competitivo donde los empleados tienen aspiraciones altas y objetivos definidos, el cual ocasiona pérdidas en las empresas u organizaciones [3].

En el Perú el salario de los empleados no es ajeno a las restricciones económicas, con respecto al salario de los docentes según una encuesta realizada el año 2019 el consejo nacional de educación CNE [4], menciona que entre el 65% y 68% de los docentes se sienten insatisfechos por su salario; los docentes públicos lo están entre 78% y 80%. Los salarios de los docentes en el estado peruano se rigen en función a ocho escalas remunerativas [5], a la condición del docente del tipo de servicio que brinda el estado, a los años de servicio que aporta como docente, las capacitaciones e investigaciones lo que permite incrementar incentivos salariales a los docentes del magisterio peruano. En la actualidad existen modelos predictivos de machine learning [2], [6], en los cuales el modelo de regresión lineal, lasso, ridge y elastic net, permiten organizar grandes cantidades de datos con penalizaciones lambda (λ), generando predicciones en muy alto grado de precisión, además pueden adaptarse rápidamente a los cambios, más aún cuando se requiere predecir el salario de los docentes peruanos. Además los modelos multinivel jerárquicos mixtos permiten asociar el salario docente con sus variables regresaras buscando predicciones. Existen pocas investigaciones que utilicen técnicas y modelos de aprendizaje automático para predecir el nivel salarial de un docente conociendo los factores que incluyen el salario [7].

Numerosas investigaciones utilizan modelos de machine learning para predecir el salarios en organizaciones que usan modelos predictivos, tal como se muestra en la investigación [8], se analizó datos desde el año 2003 hasta el año 2006 sobre los salarios de docentes de "Education at a Glance" y datos nacionales sobre el rendimiento en matemáticas el cual se analizó 30 países encontrando un modelo de regresión múltiple que relaciona el cambio salarial de los nuevos profesores y profesores con 15 años de experiencia y el rendimiento nacional en matemáticas. El modelo que se encontró fue: Salario = 2.86(New teacher salary change) + 1.28(Educational expenditure) + 0.21(GDP per capita), donde el coeficiente de determinación R^2 es 0.29, lo que significa que el 29% de las variables regresoras explican el salario docente. En el mismo contexto McKinley L. Blackburn[9], analizó datos de los años 2012 al 2016 de una encuesta sobre la comunidad estadounidense (ACS) para determinar diferencias salariales entre los docentes y no docentes, para lo cual se utilizó un modelo de regresión logarítmica.

Emma García y Eunice S. Han [10], utilizo datos representativos a nivel nacional entre los años 2009 al 2015 para examinar la relación en los sueldos de los profesores y el rendimiento académico de los alumnos, para lo cual utilizo modelos de regresión lineal y modelos multinivel de efectos mixtos, obteniendo como resultado que el modelo muestra una asociación significativa positiva entre el salario base de los docentes y el desempeño en matemáticas de los estudiantes.

Yasser T. Matbouli[11], realizo un estudio utilizando modelos de machine learning para predecir los salarios en base a características ocupacionales y organizacionales de toda economía de arabia saudita, utilizaron datos de los años 2010 al 2017 y más de dos mil actividades económicas, obteniendo como resultado que la regresión bayesina gaussiano mostro una mejora con respecto a la regresión lineal múltiple R (de 0.50 a 0.98). Asimismo, se mostraron niveles bajos, de error cuadrático medio donde se disminuyó en un 80% en comparación con la regresión lineal múltiple que se redujo en un 90%. Del mismo modo Rupashri Barik [12], realizo un estudio para predecir el salario de una persona en un determinado periodo de tiempo, y posteriormente desarrollar un sistema que permita evidenciar todos los datos del trabajo diario y del crecimiento salarial de una persona, los datos que serán utilizados son el historial de registros de salarios de un empleado de una organización, para lo cual se utilizó modelos de regresión lineal y regresión polinómica. Asimismo En la actualidad las industrias están

utilizando algoritmos de machine para resolver diversos problemas, Sarala, V. [3], se aborda los salarios de los empleados de los recién graduados y tiene como objetivo predecir el salario, según un campo en particular y sus calificaciones, para la presente investigación se utilizó algoritmos de predicción como regresión lineal, también se utilizó algoritmos de árbol de decisión y el regresor de bosque aleatorio, el algoritmo de regresión lineal da mejor precisión, el modelo predijo una precisión de 95.68% en el entrenamiento, y una precisión de 95.33% en los de prueba.

Mucha de la literatura no abarca predicciones sobre salarios utilizando algoritmos de regresión Lineal, regresión Ridge y regresión Elastic Net, pero utilizan estos modelos para solucionar diversas situaciones concretas, así como menciona Xiangning Dong [13], donde presenta una nueva técnica para predecir la incidencia de los programas de retrasos en tierra por el mal tiempo o problemas de capacidad aeroportuaria, utilizando técnicas de aprendizaje automático. Estos métodos podrían mejorar la seguridad de los pasajeros y los beneficios económicos de los vuelos, para lo cual se tomaron datos de operaciones de vuelos entre el 1 de enero del 2021 al 30 de junio del 2021 del aeropuerto internacional Nanjing Lukou, el segundo más grande en china. Se utilizó modelos de regresión lineal como Ridge y LASSO, obteniendo como resultado que los modelos de regresión mostraron un mejor rendimiento para cuantificar y generar una puntuación.

Devesh Singh[14] para atraer inversión pública los factores regionales juegan un papel importante, esta investigación utiliza datos de 18 años a nivel de condado de Hungría, encontrando un modelo de regresión Ridge, LASSO y Elastic Net, los resultados muestran que el modelo de Elastic Net es el mejor método para determinar la predicción de inversión pública a escala regional.

Por otro lado, Fadhil M. Basysyar[15], realizo una investigación, con rango de precios de viviendas, busca pronosticar el valor de las viviendas de forma acertada, los modelos utilizados para esta investigación son regresión lineal, regresión Ridge, Regresión Lasso, y Regresión Elastic Net, se utilizaron 1460 registros y 81 características de viviendas, obteniendo como resultado mostro que el modelo de regresión Lasso tuvo un mejor rendimiento al realizar los análisis superando a los demás modelos.

Frank Emmert[16], en su investigación estudia y compara los modelos de regresión Lasso, Elastic Net, Ridge, los datos que se utilizaron para el análisis consta de

93 variables de 156 muestras de índices de inflación de la economía en el país de Brasil, se discutieron términos de regularización para disminuir coeficientes que conduzcan a mejores métricas de rendimiento, se encontró que los modelos de regresión son muy eficientes en analizar datos de alta dimensión, cada uno de ellos afronta problemas en particular, y que en la actualidad estos modelos son muy populares para análisis de datos.

Este articulo tiene como objetivo predecir el salario de docentes peruanos de educación básica regular utilizando modelos de regresión con machine learning, utilizando como variables predictora el salario docente y como variables regresaras la edad, el nivel educativo, el tiempo de servicio, la escala docente y las horas laboradas.

El estudio realizado será precedente y referente teórico a futuras investigaciones dentro de los modelos predictivos con machine learning, porque ayuda a analizar las relaciones, conceptos y alcance de las variables regresaras del estudio con la variable predictora que es el salario del docente bajo modelos de regresión lineal, Lasso, Ridge y Elastic Net que son muy precisas en su predicción. Los resultados nos ayudaran a determinar el sueldo futuro de docentes que cumplan con sus estándares de escalas, ya que a mayor escala dichos docentes tendrán mejor salario.

2 METODOLOGÍA

2.1 Recojo de la información

Para tener la información de los datos de las remuneraciones de los Docentes de la UGEL Ventanilla - Callao, se solicitó mediante una carta dirigida al representante legal de la institución solicitando hacer uso de los datos de la planilla pagos de los docentes de educación básica regular nombrados, comprendidos entre los años 2018 al 2023 con la finalidad de ser estudiados y analizados. Así mismo se realizó el método de la investigación como se muestra en la figura 1.

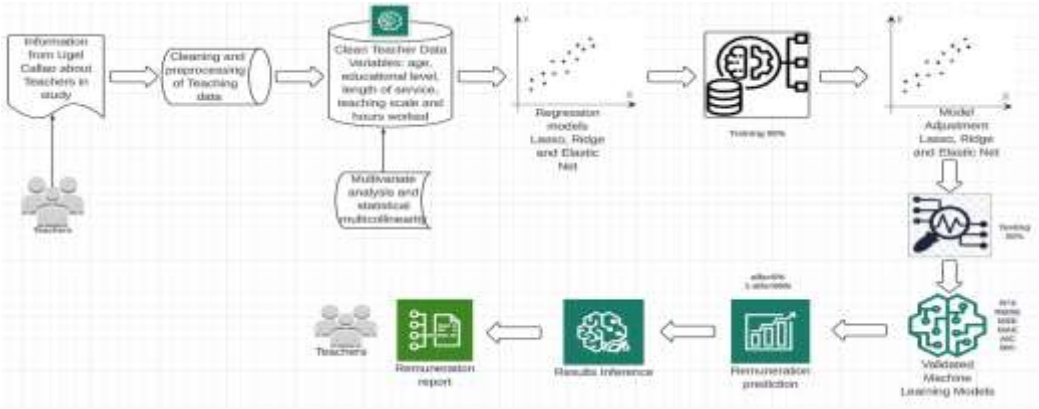


Fig. 1. Método de la Investigación.

2.2 Modelos de Machine Learning

Los modelos predictivos de machine learning utilizados en la investigación regresión lineal, regresión lasso, regresión Ridge y la regresión Elastic Net permitieron realizar un análisis muy efectivo en la predicción de las remuneraciones de los docentes de la UGEL Ventanilla, Perú mostrando mediante las métricas eficacia y un excelente performance.

La regresión lineal es una generalización del modelo de regresión poblacional de k variables (FRP) con la variable dependiente Y y k-1 variables explicativasX_2, $X_3, X_4, X_5, \cdots, X_k$ [17] y se escribe como se muestra en la ecuación 1.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \cdots + \beta_k X_{ki} + u_i \;\; ; \;\; i = 1, \cdots, n \tag{1}$$

Donde β_1 es el intercepto, β_2 a β_k son los coeficientes parciales de pendientes, u es el término de perturbación estocástica e i es la i-ésima observación, con n como tamaño de la población.

La ecuación (1) es una expresión abreviada para el conjunto de n ecuaciones simultáneas [18] mostradas en la ecuación 2 y 3.

$$\left.\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \beta_4 X_{41} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \beta_4 X_{42} + \cdots + \beta_k X_{k2} + u_2 \\ Y_3 &= \beta_1 + \beta_2 X_{23} + \beta_3 X_{33} + \beta_4 X_{43} + \cdots + \beta_k X_{k3} + u_3 \\ &\cdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \beta_4 X_{4n} + \cdots + \beta_k X_{k3} + u_n \end{aligned}\right\} \tag{2}$$

El sistema de ecuaciones 2 se escribe en forma matricial como se muestra en la ecuación 3 y 4.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & x_{41} & \cdots & x_{k1} \\ 1 & x_{22} & x_{32} & x_{42} & \cdots & x_{k2} \\ 1 & x_{23} & x_{33} & x_{43} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & x_{3n} & x_{4n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} \tag{3}$$

$$Y_{nx1} = X_{nxk}.\beta_{kx1} + u_{nx1} \tag{4}$$

Donde Y es el vector columna de nx1 de observaciones sobre la variable dependiente Y, la variable X es la matriz nxk, con n observaciones sobre k-1 variables X_2 a X_k y la primera columna de números 1 representa el término del intercepto, el valor de β es el vector columna kx1 de los parámetros desconocidos $\beta_1, \beta_2, \cdots, \beta_k$ y el valor de u es el vector columna nx1 de n perturbaciones u_i

El modelo de regresión LASSO está representado matemáticamente por ecuación 5, cuyo objetivo es minimizar los parámetros α; β en los errores tomando la penalización de los coeficientes del modelo [19].

$$\underset{\alpha,\beta}{Minimize} \left\{ \frac{1}{2n} \sum_{i=1}^{N} \left(y_i - \alpha - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \tag{5}$$

Por lo tanto, el Modelo LASSO considerando las variables anteriormente mencionadas, queda representado por la ecuación 6 y 7.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ 1 & x_{31} & x_{32} & x_{33} & x_{34} & x_{35} \\ 1 & x_{41} & x_{42} & x_{43} & x_{44} & x_{45} \\ 1 & x_{51} & x_{52} & x_{53} & x_{54} & x_{55} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \tag{6}$$

$$Y_i = X_{kx5}.\beta_{5x1} + u_i \; ; \; i = \overline{1,5} \tag{7}$$

El modelo de regresión lineal se muestra en la ecuación 8 y 9.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + u_i \quad (8)$$

$$u_i = Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \beta_4 X_{i4} + \beta_5 X_{i5} \quad (9)$$

Donde u_i es el error que se quiere minimizar.

La técnica de mínimos cuadrados realiza la minimización de los errores y su representación es:

$$D = \sum_{i=1}^{n} u_i^2$$

Al realizar la regresión Lasso, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada por la ecuación 10.

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} [u_i^2 + \lambda(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5|)] \quad (10)$$

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} [(Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \beta_4 X_{i4} - \beta_5 X_{i5})^2 + \lambda(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5|)]$$

Las ecuaciones normales por la teoría de mínimos cuadrados para la regresión múltiple Lasso se muestran en la ecuación 11.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^{n} X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i5} = \sum_{i=1}^{n} y_i \\ \alpha \sum_{i=1}^{n} X_{i1} + \beta_1 \sum_{i=1}^{n} X_{i1}^2 + \beta_2 \sum_{i=1}^{n} X_{i1}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i1}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i1}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i1}X_{i5} = \sum_{i=1}^{n} X_{i1}y_i - \lambda n \\ \alpha \sum_{i=1}^{n} X_{i2} + \beta_1 \sum_{i=1}^{n} X_{i2}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2}^2 + \beta_3 \sum_{i=1}^{n} X_{i2}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i2}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i2}X_{i5} = \sum_{i=1}^{n} X_{i2}y_i - \lambda n \\ \alpha \sum_{i=1}^{n} X_{i3} + \beta_1 \sum_{i=1}^{n} X_{i3}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i3}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i3}^2 + \beta_4 \sum_{i=1}^{n} X_{i3}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i3}X_{i5} = \sum_{i=1}^{n} X_{i3}y_i - \lambda n \\ \alpha \sum_{i=1}^{n} X_{i4} + \beta_1 \sum_{i=1}^{n} X_{i4}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i4}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i4}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i4}^2 + \beta_5 \sum_{i=1}^{n} X_{i4}X_{i5} = \sum_{i=1}^{n} X_{i4}y_i - \lambda n \\ \alpha \sum_{i=1}^{n} X_{i5} + \beta_1 \sum_{i=1}^{n} X_{i5}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i5}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i5}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i5}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i5}^2 = \sum_{i=1}^{n} X_{i5}y_i - \lambda n \end{cases} \quad (11)$$

El modelo de regresión Ridge a diferencia de la regresión LASSO reduce los coeficientes de los predictores correlacionados entre sí, lo que les permite tomar prestada la fuerza de los demás. Asimismo, desde la perspectiva bayesiano, la penalización del modelo RIDGE

es apropiado en el caso de haber varios predictores y todos tienen coeficientes distintos de cero, es decir, son extraídos de una distribución gaussiana [16]. Además, es de prioridad considerar en el análisis las propiedades del error cuadrático medio de la regresión de Ridge como la varianza y sesgo del estimador, el teorema sobre la función del cuadrado medio y los comentarios hechos sobre la función del error cuadrático medio [20]. La representación matemática del modelo de regresión Ridge se representa en la ecuación 12.

$$\beta^{ridge} = \underset{\beta \epsilon \mathbb{R}}{argmin} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \tag{12}$$

Al realizar la regresión Ridge, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada en la ecuación 13.

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} \left[\frac{1}{2n}\sum_{i=1}^{N} u_i^2 + \lambda(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2)\right] \tag{13}$$

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} \left[\frac{1}{2n}(Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \beta_4 X_{i4} + \beta_5 X_{i5})^2 + \lambda(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2)\right]$$

Las ecuaciones normales de la teoría de mínimos cuadrados para la regresión múltiple Ridge se representa en la 14.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^{n} X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i5} = \sum_{i=1}^{n} y_i \\ \alpha \sum_{i=1}^{n} X_{i1} + \beta_1 \sum_{i=1}^{n} X_{i1}^2 + \beta_2 \sum_{i=1}^{n} X_{i1}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i1}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i1}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i1}X_{i5} = \sum_{i=1}^{n} X_{i1}y_i - 2\lambda n\beta_1 \\ \alpha \sum_{i=1}^{n} X_{i2} + \beta_1 \sum_{i=1}^{n} X_{i2}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i2}^2 + \beta_3 \sum_{i=1}^{n} X_{i2}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i2}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i2}X_{i5} = \sum_{i=1}^{n} X_{i2}y_i - 2\lambda n\beta_2 \\ \alpha \sum_{i=1}^{n} X_{i3} + \beta_1 \sum_{i=1}^{n} X_{i3}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i3}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i3}^2 + \beta_4 \sum_{i=1}^{n} X_{i3}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i3}X_{i5} = \sum_{i=1}^{n} X_{i3}y_i - 2\lambda n\beta_3 \\ \alpha \sum_{i=1}^{n} X_{i4} + \beta_1 \sum_{i=1}^{n} X_{i4}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i4}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i4}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i4}^2 + \beta_5 \sum_{i=1}^{n} X_{i4}X_{i5} = \sum_{i=1}^{n} X_{i4}y_i - 2\lambda n\beta_4 \\ \alpha \sum_{i=1}^{n} X_{i5} + \beta_1 \sum_{i=1}^{n} X_{i5}X_{i1} + \beta_2 \sum_{i=1}^{n} X_{i5}X_{i2} + \beta_3 \sum_{i=1}^{n} X_{i5}X_{i3} + \beta_4 \sum_{i=1}^{n} X_{i5}X_{i4} + \beta_5 \sum_{i=1}^{n} X_{i5}^2 = \sum_{i=1}^{n} X_{i5}y_i - 2\lambda n\beta_5 \end{cases} \tag{14}$$

La penalización de la regresión Elactic Net es una adaptación de los mínimos cuadrados y permite abordar el problema de estimación produciendo un estimador β sesgado pero

con varianzas pequeñas [21]. La representación matemática del modelo de regresión Elactic Net se representa en la ecuación 15 se define:

$$\beta^{EN} = \underset{\beta \epsilon \mathbb{R}}{argmin} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1] \quad (15)$$

Al realizar la regresión Elastic Net, agregamos un factor de penalización a los mínimos cuadrados, que reduce la función de pérdida S a un valor mínimo, representada en la ecuación 16.

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} \left[\frac{1}{2n}\sum_{i=1}^{N} u_i^2 + \lambda\left[(1-\alpha)\left(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2\right)\right] + \alpha[|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5|]\right] \quad (16)$$

$$S = \underset{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5}{Min} \left\{\frac{1}{2n}(Y_i - \alpha - \beta_1 X_{i1} - \beta_2 X_{i2} - \beta_3 X_{i3} - \beta_4 X_{i4} + \beta_5 X_{i5})^2 + \lambda\left[(1-\alpha)\left(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2\right)\right] + \alpha[|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5|]\right\}$$

Luego las ecuaciones normales de la teoría de mínimos cuadrados para la regresión Elastic net se representa en la ecuación 17.

$$\begin{cases} \alpha n + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} + \beta_3 \sum_{i=1}^n X_{i3} + \beta_4 \sum_{i=1}^n X_{i4} + \beta_5 \sum_{i=1}^n X_{i5} = n\lambda \\ [(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2)] - [|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5|] + \sum_{i=1}^n y_i \\ \alpha \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i1}X_{i2} + \beta_3 \sum_{i=1}^n X_{i1}X_{i3} + \beta_4 \sum_{i=1}^n X_{i1}X_{i4} + \beta_5 \sum_{i=1}^n X_{i1}X_{i5} = \sum_{i=1}^n X_{i1}y_i + \lambda n[2\beta_1(\alpha-1) - \alpha] \\ \alpha \sum_{i=1}^n X_{i2} + \beta_1 \sum_{i=1}^n X_{i2}X_{i1} + \beta_2 \sum_{i=1}^n X_{i2}^2 + \beta_3 \sum_{i=1}^n X_{i2}X_{i3} + \beta_4 \sum_{i=1}^n X_{i2}X_{i4} + \beta_5 \sum_{i=1}^n X_{i2}X_{i5} = \sum_{i=1}^n X_{i2}y_i + \lambda n[2\beta_2(\alpha-1) - \alpha] \\ \alpha \sum_{i=1}^n X_{i3} + \beta_1 \sum_{i=1}^n X_{i3}X_{i1} + \beta_2 \sum_{i=1}^n X_{i3}X_{i2} + \beta_3 \sum_{i=1}^n X_{i3}^2 + \beta_4 \sum_{i=1}^n X_{i3}X_{i4} + \beta_5 \sum_{i=1}^n X_{i3}X_{i5} = \sum_{i=1}^n X_{i3}y_i + \lambda n[2\beta_3(\alpha-1) - \alpha] \\ \alpha \sum_{i=1}^n X_{i4} + \beta_1 \sum_{i=1}^n X_{i4}X_{i1} + \beta_2 \sum_{i=1}^n X_{i4}X_{i2} + \beta_3 \sum_{i=1}^n X_{i4}X_{i3} + \beta_4 \sum_{i=1}^n X_{i4}^2 + \beta_5 \sum_{i=1}^n X_{i4}X_{i5} = \sum_{i=1}^n X_{i4}y_i + \lambda n[2\beta_4(\alpha-1) - \alpha] \\ \alpha \sum_{i=1}^n X_{i5} + \beta_1 \sum_{i=1}^n X_{i5}X_{i1} + \beta_2 \sum_{i=1}^n X_{i5}X_{i2} + \beta_3 \sum_{i=1}^n X_{i5}X_{i3} + \beta_4 \sum_{i=1}^n X_{i5}X_{i4} + \beta_5 \sum_{i=1}^n X_{i5}^2 = \sum_{i=1}^n X_{i5}y_i + \lambda n[2\beta_5(\alpha-1) - \alpha] \end{cases} \quad (17)$$

2.3 Métricas de validación de modelos en estudio

Las métricas usadas en la investigación dan el performance de los modelos predictivos [22] que fueron el MAE (Error Absoluto Medio), que midió la magnitud de promedio de los errores en las predicciones, sin considerar su dirección. El MAE es el promedio en la muestra, de las diferencias absolutas entre la predicción y la observación real, donde todas las diferencias individuales tienen el mismo peso y se muestra en la ecuación 18.

$$Mean\ Absolute\ Error:\ MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (18)$$

Así mismo el RMSE (raíz del error cuadrático Medio) es la raíz cuadrada del promedio de las diferencias al cuadrado entre la predicción y la observación real (SSE/n) y está definida por la ecuación 19.

$$Root\ Mean\ Square\ Error:\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (19)$$

El coeficiente de determinación R^2 es otra métrica que usó los modelos predictivos, el cual midió la varianza en los datos explicado por el modelo y es determinado como uno meno la relación entre la suma de cuadrados de los errores (SSE) y el total (SST) el cual se muestra en la ecuación 20.

$$\text{Coefficient of determination: } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (20)$$

2.4 Normalización de los datos

La investigación utilizó 108317 registros docentes con datos de variable escalar, donde Shanker et al [23] sugiere que es necesario normalizar las características de las variables, ya que producen mejores resultados en general. Además los requisitos de los algoritmos usados requieren normalización de los datos donde Atlas et al [24] menciona que se debe usar en los casos de las variables regresoras y predictoras, por el cual se procedió a normalizar los datos mediante la técnica de normalización de min-max para garantizar la homogeneidad de las variables concentradas en un intervalo continuo de [0, 1] menciona M-Dawam [25] como se muestra en la ecuación 21.

$$\hat{X}[:, i] = \frac{X[:,i] - min(X[:,i])}{max(X[:,i]) - min(X[:,i])} \quad (21)$$

3 RESULTADOS

3.1 Estadísticos Descriptivos

La investigación realizó un análisis exploratorio descriptivo, encontrando un promedio de salario Docente de 2721.80 soles, con un mínimo de S/. 302.2 y un máximo de S/. 10 760.10, así mismo se obtuvo docentes con una edad promedio de 53.24 años, dos meses y 26 días con un mínimo de 23 años y un máximo de 71 años, el nivel educativo hubieron docentes de nivel 1, un máximo de 8 con un promedio de 2.372 además el

tiempo de servicio Docente con un promedio de 14 años, también una Escala Docente promedio de 2 y las horas laboradas con un promedio de 31H+8M+24S con un mínimo de 26 horas y un máximo de 40 hora, como se muestra en la tabla 1.

Tabla 1. Estadísticos descriptivos de las variables en estudio.

Variables de estudio	Mínimo	Mediana	Media	Máximo
Salario Docente	302.2	2620.1	2771.8	10760.1
Edad	23	53	53.24	71
Nivel Educativo	1	2	2.372	8
Tiempo de servicio	1	14	13.88	44
Escala Docente	1	2	1.919	7
Horas Laboradas	26	30	31.35	40

Así mismo se realizó un análisis exploratorio de las remuneraciones por año de los docentes nombrados de ventanilla y se encontró en el salario fue aumentado desde el año 2018 hasta el año 2023 como se muestra en la figura 2.

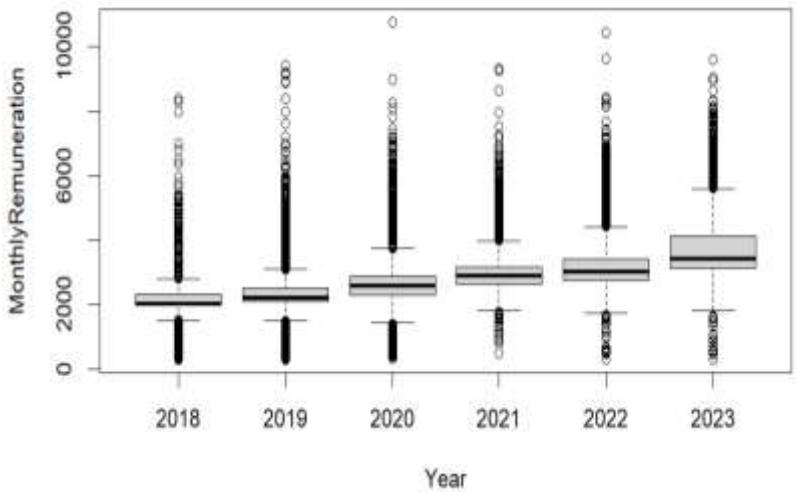


Fig. 2. Diagrama de cajas de los cinco años de los salarios docentes.

En la investigación se encontró las remuneraciones por Escala docente, tomando en primer lugar a la Escala 4, seguido de la Escala 5, luego de la Escala 6, le sigue la Escala 2, luego la Escala 1, además la Escala 3 y finalmente la escala 7, que durante el tiempo fue aumentando los salaries de los docentes como se muestra en la figura 3.

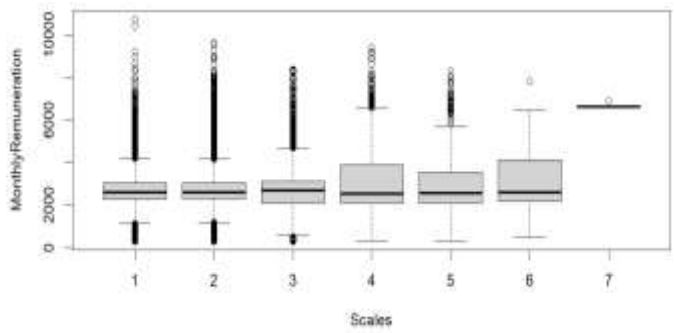


Fig. 3. Diagrama de cajas de las ocho escalas de los salarios docentes.

3.2 Estadísticos Inferenciales

En la investigación de las 108317 observaciones se tomó el 80% (86654) de las observaciones para el entrenamiento y un 20%(21663) para el testeo, encontrando en el entrenamiento una validación cruzada de remuestreo con 5 repeticiones un RMSE=895.3793, un Rsquared=0.2872618 y un MAE=619.7701

El análisis de la regresión lineal multivariado con 5 variables regresoras analizadas que son la edad (*Edad*), el nivel educativo (*NEduc*), el tiempo de servicio(*Tserv*), la Escala Docente(*EsDoc*) y las horas laboradas (*HoLab*) por los docentes, así como la variable predictora *salario Docente* cumpliendo los prerequisitos estadísticos arrojó los resultados en el software R Studio que se muestra en la figura 4.

```
> summary(lm)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4283.0  -344.6    31.1   428.6  8114.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3109.7302    38.2512 -81.298   <2e-16 ***
Edad           -5.2337     0.5378  -9.731   <2e-16 ***
NEduc          24.1794     3.2491   7.442    1e-13 ***
TServ          18.0115     0.4056  44.410   <2e-16 ***
EsDoc         -90.1259     3.0967 -29.104   <2e-16 ***
HoLab         192.1820     1.1054 173.859   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 895.4 on 86690 degrees of freedom
Multiple R-squared:  0.2873,    Adjusted R-squared:  0.2873
F-statistic:  6989 on 5 and 86690 DF,  p-value: < 2.2e-16
```

Fig. 4. Resultados de los códigos de R Studio para la regresión lineal múltiple.

*** indica el aporte de la significancia a la variable predictora

Estos resultados mostrados en la figura 1, donde la significancia para las variables regesoras fueron altamente significativas con p-value menor a 2×10^{-16} permitieron plantear el modelo de regresión multiple del Salario Docente mostrado en la ecuación 22.

$$Y = -3109.7302 - 5.2337(Edad) + 24.1794(NEduc) + 18.0115(TServ) - 90.1259(EsDoc) + 192.1820(HoLab) \quad (22)$$

Esto significa que por cada año de un docente, el salario docente disminuye en 5.2337 soles, así mismo para cada nivel educativo docente escala el salario aumenta en 24.1794 soles, además por cada año de servicio prestado el salario aumenta en 18.0115 soles, sin embargo por cada Escala docente el salario disminuye en 90.1259 soles, finalmente por cada hora laborada el salario aumenta en 192.1820 soles.

Los gráficos mostrados en la figura 5, señalan los residuos normalizados y valores ajustados, con cuantiles residuales y la raíz del error cuadrático medio, por el cual el modelo de regresión lineal cumple con todas las especificaciones y prerequisitos para la predicción.

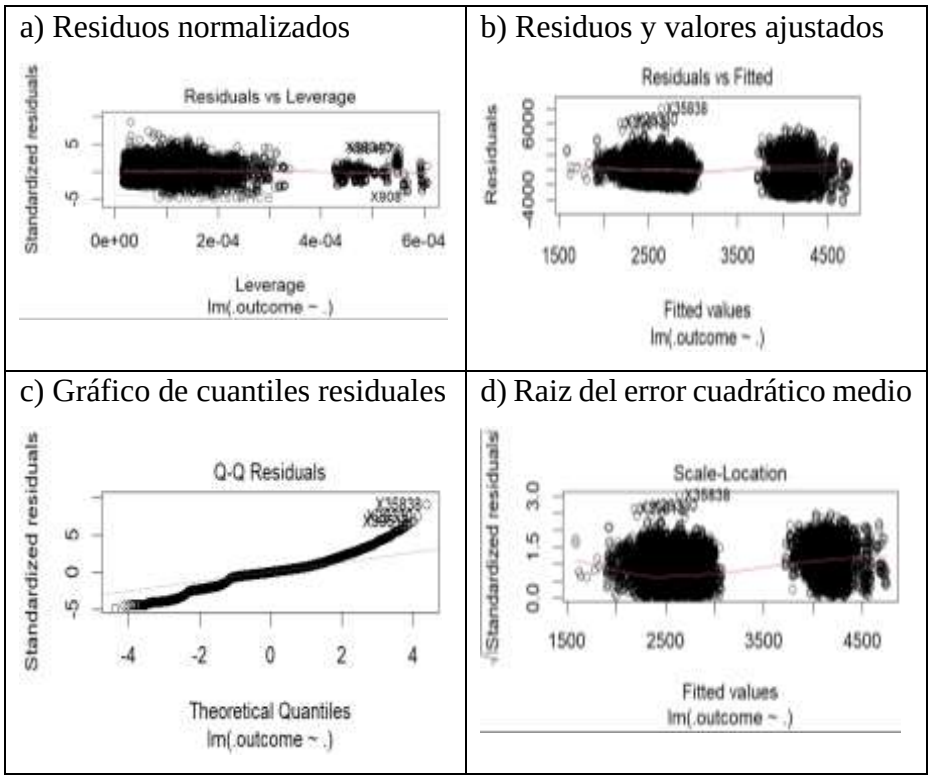


Fig. 5. Gráficos del performance del modelo de regresión lineal múltiple

Los resultados para el modelo de regresión Ridge, con un entrenamiento de 86654 observaciones y 21663 observaciones para el testeo, se encontró que las métricas del remuestreo a través de los parámetros de ajuste son las que se muestran en la tabla 2.

Tabla 2. Métricas del performance del modelo de regresión Ridge.

Lambda	RMSE	Rsquared	MAE
0.000100	896.5645	0.2866115	622.6167
0.250075	896.5645	0.2866115	622.6167
0.500050	896.5645	0.2866115	622.6167
0.750025	896.5645	0.2866115	622.6167
1.000000	896.5645	0.2866115	622.6167

En la figura 6, se muestra los gráficos para el modelo de regresión Ridge con la validación cruzada y los logaritmos de los valores de penalización Lambdas, así como los valores de los coeficientes del modelos y la importancia de las variables regresoras.

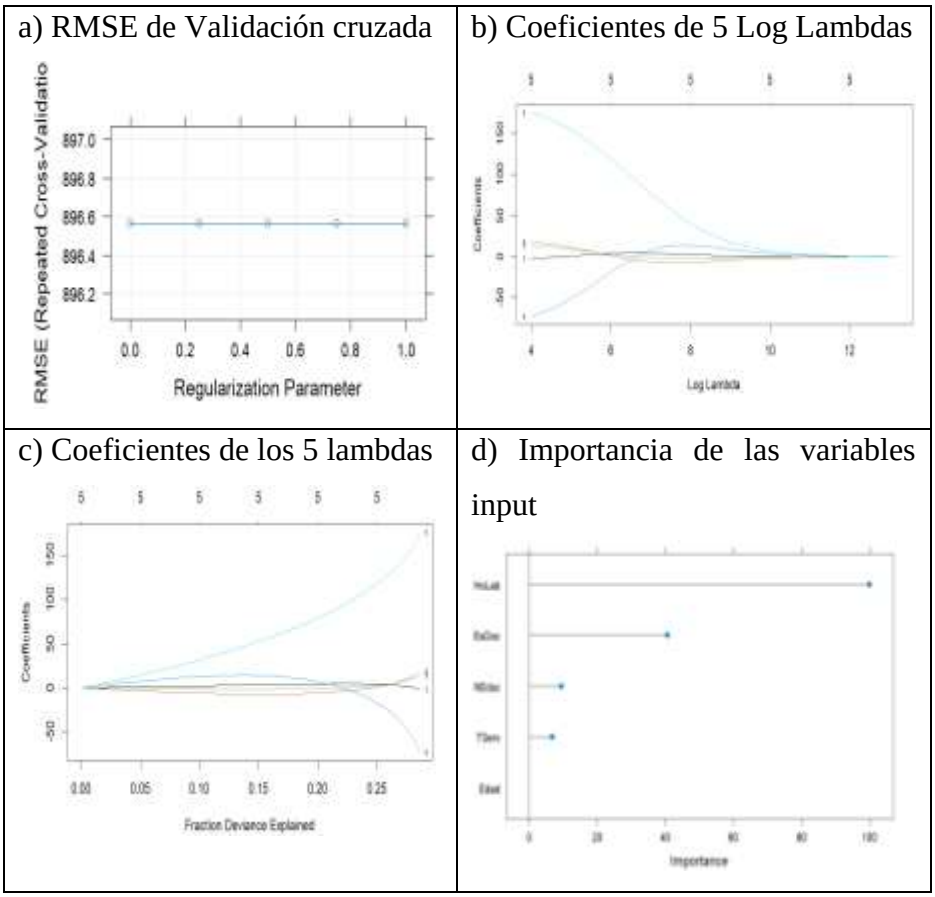


Fig. 6. Gráficos del modelo de regularización Ridge.

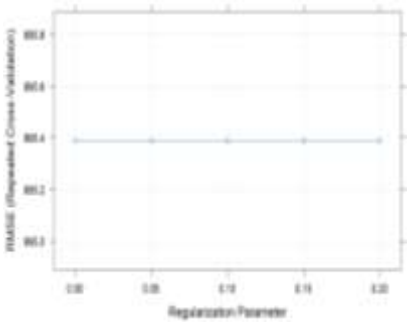
Los resultados para el modelo de regresión Lasso, con un entrenamiento de 86654 observaciones y 21663 observaciones para el testeo, se encontró que las métricas del remuestreo a través de los parámetros de ajuste son las que se muestran en la table 1.

Tabla 3. Métricas del performance de modelo de regresión Lasso.

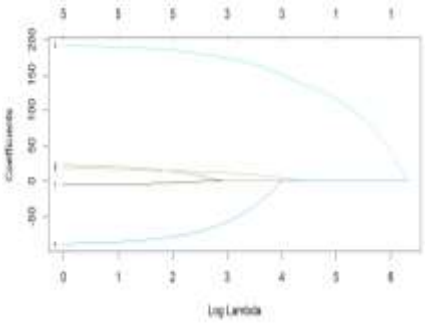
Lambda	RMSE	Rsquared	MAE
0.000100	895.3873	0.2872513	619.851
0.050075	895.3873	0.2872513	619.851
0.100050	895.3873	0.2872513	619.851
0.150025	895.3873	0.2872513	619.851
0.200000	895.3873	0.2872513	619.851

En la figura 7, se muestra los gráficos para el modelo de regresión Lasso con la validación cruzada y los logaritmos de los valores de penalización Lambdas, así como los valores de los coeficientes del modelos y la importancia de las variables regresoras.

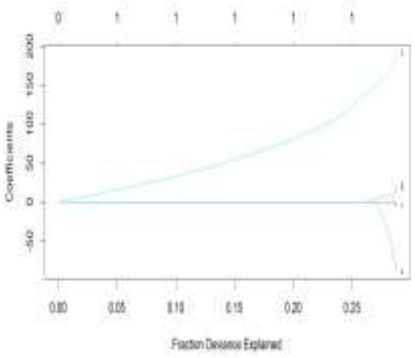
a) RMSE de Validación cruzada



b) Coeficientes de 5 Log Lambdas



c) Coeficientes de los 5 lambdas



d) Importancia de las variables input

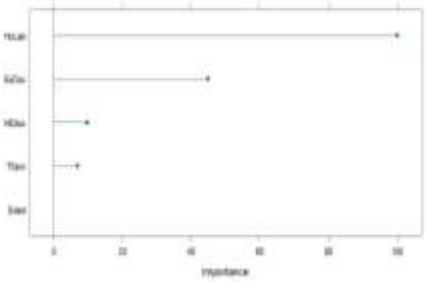


Fig. 7. Gráficos del modelo de regulazación Lasso.

Los resultados para el modelo de regresión Elastic Net, con un entrenamiento de 86654 observaciones y 21663 observaciones para el testeo, se encontró que las métricas del remuestreo a través de parámetros de ajuste son las que se muestran en la tabla 4.

Tabla 4. Métrica MAE para remuestreo de modelos.

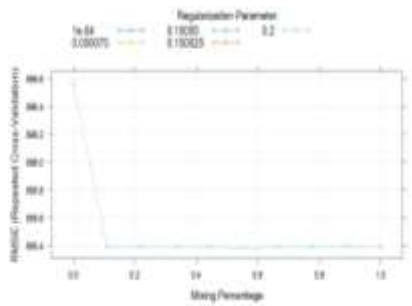
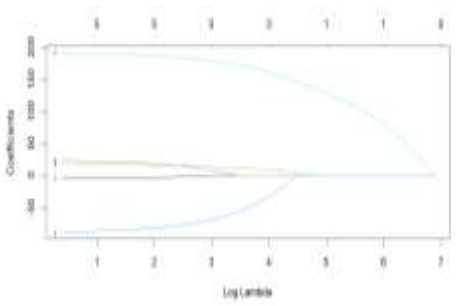
MAE	Min	Mean	Max
LinearModel	607.2400	619.7701	632.1245
Ridge	610.6784	622.6167	633.5756
Lasso	607.3899	619.8510	632.1050
ElacticNet	607.4004	619.8605	632.1108

Tabla 5. Métrica RMSE para remuestreo de modelos.

RMSE	Min	Mean	Max
LinearModel	877.5731	895.3793	912.9543
Ridge	879.1204	896.5645	912.8855
Lasso	877.6042	895.3873	912.8773
ElacticNet	877.6047	895.3870	912.8733

Tabla 6. Métrica Rsquared para remuestreo de modelos.

Rsquared	Min	Mean	Max
LinearModel	0.2558271	0.2872618	0.3196392
Ridge	0.2560430	0.2866115	0.3197616
Lasso	0.2559246	0.2872513	0.3196918
ElacticNet	0.2559196	0.2872524	0.3196939

a) RMSE de Validación cruzada	b) Coeficientes de 5 Log Lambdas
	

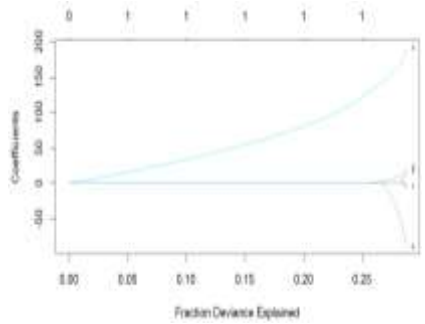
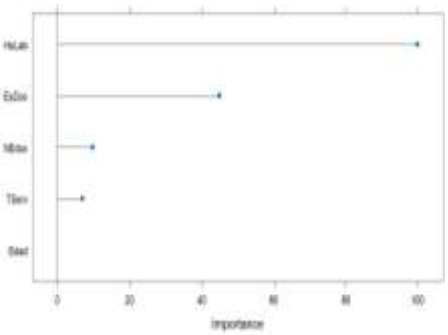
c) Coeficientes de los 5 lambdas	d) Importancia de las variables input
	

Fig. 8. Gráficos del modelo de regularación Elastic Net.

El mejor modelo que permitió predecir el salario docente es el décimo tercero con $\alpha = 0.5555556$ y un Lambda $\lambda = 0.2$ cuyo modelación se muestra en la ecuación 23.

$$Y = -3092.582975 - 4.824496(Edad) + 22.972778(NEduc) + 17.623234(TServ) - 88.511756(EsDoc) + 191.104877(HoLab) \quad (23)$$

Esto significa que por cada año de un docente, el salario docente disminuye en 5.2337 soles, así mismo para cada nivel educativo docente el salario aumenta en 24.1794 soles, además por cada año de servicio prestado el salario aumenta en 18.0115 soles, sin embargo por cada Escala docente el salario disminuye en 90.1259 soles, finalmente por cada hora laborada el salario aumenta en 192.1820 soles.

Finalmente la predicción con el mejor modelo mostrado en la ecuación (2) con los datos de entrenamiento se obtuvo un RMSE de 895.3383 y con los datos de testeo un RMSE de 906.8914, siendo una diferencia de 11.5531 que representa el 1.2739% de error con un 98.7261% de eficiencia en la predicción del modelo.

4 DISCUSIÓN Y CONCLUSIÓN

El análisis de los modelos de Machine Learning en la predicción del salario de 108317 docentes con data más reciente proporcionada por la UGEL de Ventanilla, Lima Perú produjo nuevos resultados sobre la precisión del salario de docentes nombrados. Antes de discutir los hallazgos es importante señalar que el estudio sólo analizó solo el

salario de Docentes nombrados, mas no el salario de Docentes contratados, actualmente no existe un modelo predictivo de Machine Learning con alta precisión que permita realizar la predicción del salario docente, es por ello que en estudios futuros el modelo predictivo encontrado ayuda a precisar el salario en función a variables regresoras que estén relacionadas con su variable predictora.

En segundo lugar, la data analizada con promedio salarial Docente no se toma en cuenta la invarianza de género del Docente, por lo que el monto del salario al mes se diferencia en los cinco años analizados donde la variación promedio está muy marcada para una comparación más precisa de la predicción. Los estudios futuros deben examinar la variación del salario tanto de nombrados como contratados para comprender plenamente cómo funciona la precisión en función a los modelos predictivos de machine learning analizados.

Uno de los hallazgos clave del estudio es que un docente de la escala básica regular tiene un salario promedio de S/. 2771.8 y está enmarcada en función a su edad, nivel educativo, tiempo de servicio, escala docente y horas Laboradas. Los modelos de machine Learning regresión lineal, Ridge, Lasso y Elastic Net fueron precisos en encontrar el factor de penalización del modelo tomado sus mejores métricas como la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2) que permitieron tomar el mejor modelo que predijo el salario de los Docentes nombrados de la UGEL de ventanilla.

La investigación [26] encontró un modelo de machine learning para el salario de graduados en función al campo que desempeño utilizando un modelo de regresión lineal con una precisión del 95.68% para los datos de entrenamiento y un 95.33% para los datos de prueba, obteniendo una diferencia muy pequeña afirmando un buen performance del modelo. La investigación determine el salario docente con un modelo de regresión lineal con 5 variables regresoras con una precisión del 98.7261% y un coeficiente de determinación $R^2 = 0.2072618$ el cual permite realzar un análisis profundo para otras investigaciones futuras.

La investigación [27] predijo las clases salariales de empleados utilizando aprendizaje automático, encontrando precisión y correlación de extracción de características

utilizando un Sistema PCA y un modelo de red neuronal profunda DNN con precisión de 94.9%, además Decisión Tree con 89.6% y Random Forest con 76.4% comparada con la precisión de regresión lineal de la investigación del 98.7261% existe una diferencia de 3.8226% que es mínima.

El artículo [10] encontró modelos de regresión con asociación entre el salario base de los profesores y el rendimiento de las matemáticas en diferentes distritos, donde el coeficiente de correlación entre el logaritmo de salario base y los resultados de los exámenes de matemática fue aproximadamente 10.5 lo que indica que un aumento del 10% en el salario base de los maestros se asocia al con 1.05 mayor puntuación media en los exámenes de matemáticas. A diferencia del modelo de investigación se encontró asociación entre el salario docente y el tiempo de servicio, encontrando que por cada año de servicio prestado por el docente, el salario aumenta en 18.0115 soles.

Dreher G. et al [28] encontró un modelo de regresión de la satisfacción salarial en función a las variables de percepción y de mantenimiento de la empresa como los años de servicio ($\beta = 0.01$), la formación ($\beta =$ -0.07),, la valoración de la empresa ($\beta = 0.17$),, el potencial ($\beta = 0.00$),, el género ($\beta = 0.23$),, el salario mensual ($\beta = 0.29$), y el porcentaje de aumento salarial ($\beta = 0.01$), con un coeficiente de determinación R^2 de 12.00%. La investigación halló un modelo de regresión lineal del salario docente en función a las variables de edad ($\beta =$ -5.2337),, Nivel educativo($\beta = 24.1794$),, tiempo de servicio ($\beta = 18.0115$),, Escala docente ($\beta =$ -90.1259), y horas laboradas ($\beta = 192.1820$) con un coeficiente de determinación R^2 de 28.72618%.

Mohamed A. et al [29] encontró los factores de remuneración de los recién licenciados de ingeniería en los mercados laborales de Bharat, india con algoritmos de aprendizaje automático Naïve Bayes (RMSE=0.308), Random Forest (RMSE=0.306) y Support Vector Machine (RMSE=0.305) identificando los factores como la demografía, el éxito académico, los rasgos de personalidad y la puntuación obtenida en los exámenes, el cual fueron significativos en el salario inicial. En la investigación se encontraron modelos de aprendizaje automático la regresión lasso, Ridge y Elastic net con una penalización lambda $\lambda=0.5555556$ y un valor de $\alpha=0.20$ tomando como regresoras a la edad del

docente (β=-4.824496), el nivel educativo (β=22.972778),, el tiempo de servicio (β=17.623234),, la Escala docente (β=88.511756), y las horas laboradas(β=191.104877) con un RMSE para el mejor modelo de 895.3870.

Dutta et al [30] encontró un motor de predicción para el salario adecuado de un puesto de trabajo usando modelos de machine learning como el árbol de decisión (MSE=389.64) y el modelo conjunto (MSE=329.12) con una precisión de 0.844 y 0.873 respectivamente. En la investigación se encontró para los modelo en estudio de Lasso (MAE=619.8510), Ridge (MAE=622.6167) y Elastic net (MAE=619.8605).

Geraldo-Campos et al [31] halló modelos de regularización Lasso y Ridge para la predicción del riesgo crediticio en reactiva Perú con 501 298 empresas analizadas, con sector económico, entidad otorgante, monto cubierto y departamento como regresoras y el nivel de riesgo como predictor, determinando un modelo de regresión lasso con $\lambda_{60} = 0.00038$ y un RMSE=0.3573685; así como una regresión Ridge con $\lambda_{100} = 0.00910$ y un RMSE=0.3573812, representados en la ecuación 24 y 25 respectivamente.

$$Y_{Lasso} = 0.51487 + 0.05878(Economic\ sector) - 0.19292(Credit\ granting\ entity) + 1.29671(Amount\ covered) + 0.03115(Departament) \quad (24)$$

$$Y_{Ridge} = 0.51408 + 0.05849(Economic\ sector) - 0.19071(Credit\ granting\ entity) + 1.27321(Amount\ covered) + 0.03198(Departament) \quad (25)$$

En la investigación se encontró el modelos lasso (RMSE=895.3873), Ridge (RMSE=896.5645) y Elastic net (RMSE=895.3870), donde el mejor modelo hallado con un $\alpha = 0.5555556$ y un Lambda $\lambda_{13} = 0.2$ se representa en la ecuación 26.

$$Y = -3092.582975 - 4.824496(Edad) + 22.972778(NEduc) + 17.623234(TServ) - 88.511756(EsDoc) + 191.104877(HoLab) \quad (26)$$

En conclusión la investigación obtuvo un nuevo modelo de regularización que permite predecir el salario docente en función a variables regresoras y predictoras, el cual se obtuvo eficiencia en sus métricas como se muestra en la tabla 7.

Tabla 7. Métricas de los modelos de Machine Learning.

	Linear Model	Ridge	Lasso	Elastic Net
MAE	619.7701	622.6167	619.8510	619.8605
RMSE	895.3793	896.5645	895.3673	895.3870
Rsquared	28.72618%	28.66115%	28.72513%	28.72524%

Así mismo los modelos obtenidos con sus respectivos coeficientes se muestran en la tabla 8.

Tabla 8. Coefiecientes del modelo de regresión lineal y Elastic Net.

Coeficientes del modelo	Lineal Model	Elastic Net
Intercepto	-3109.7302	-3092.582975
Edad	-5.2337	-4.824496
Nivel educativo	24.1794	22.972778
Tiempo de servicio	18.0115	17.623234
Escala Docente	-90.1259	-88.511756
Horas laboradas	192.1820	191.104877

Esto significa que por cada año de un docente, el salario docente disminuye en 4.824496 soles, así mismo por cada nivel educativo docente el salario aumenta en 22.972778 soles, además por cada año de servicio prestado el salario aumenta en 17.623234 soles, sin embargo por cada Escala docente el salario disminuye en 88.511756 soles, finalmente por cada hora laborada el salario aumenta en 191.18104877 soles, que representados matemáticamente, se muestran en las ecuaciones 27 y 28 respectivamente

Modelo de regresión lineal múltiple

$$Y = -3109.7302 - 5.2337(Edad) + 24.1794(NEduc) + 18.0115(TServ) - 90.1259(EsDoc) + 192.1820(HoLab) \quad (27)$$

Modelo de regresión Elastic net

$$Y = -3092.582975 - 4.824496(Edad) + 22.972778(NEduc) + 17.623234(TServ) - 88.511756(EsDoc) + 191.104877(HoLab) \quad (28)$$

5 REFERENCES

[1] E. C. Vieira, B. Debenedetti, L. Reyna, S. Santisteban, R. Lorena, and S. Aguirre, “Políticas remunerativas Grupo de Investigación ius et veritas.”

[2] M. Zaid and T. Rajendran, “Higher Classification Accuracy of Income Class Using Decision Tree Algorithm over Naive Bayes Algorithm,” in *Advances in Parallel Computing*, IOS Press BV, 2022, pp. 555–561. doi: 10.3233/APC220079.

[3] V. Sarala and K. Ganesh, “EMPLOYEE SALARY PREDICTION SYSTEM USING MACHINE LEARNING.”

[4] “6 DE CADA 10 DOCENTES DE ESCUELAS ESTATALES REPORTAN QUE ELLOS MISMOS SE PROVEEN DE MATERIALES PARA EL AULA 8 DE CADA 10 DOCENTES DEL ÁREA RURAL INDICAN QUE NO CUENTAN CON SALA DE PROFESORES.”

[5] “Ley de Reforma Magisterial LEY Nº 29944 (*) De conformidad con la Única Disposición Complementaria Final del Decreto Supremo N° 227-2013-EF.”

[6] J. V. Siswanto, L. A. Castilani, N. H. Winata, N. C. Nugraha, and N. T. M. Sagala, “Salary Classification & Prediction based on Job Field and Location using Ensemble Methods,” in *ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 325–330. doi: 10.1109/ICCoSITE57641.2023.10127828.

[7] J. V. Siswanto, L. A. Castilani, N. H. Winata, N. C. Nugraha, and N. T. M. Sagala, “Salary Classification & Prediction based on Job Field and Location using Ensemble Methods,” in *ICCoSITE 2023 - International Conference on Computer Science, Information Technology and Engineering: Digital Transformation Strategy in Facing the VUCA and TUNA Era*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 325–330. doi: 10.1109/ICCoSITE57641.2023.10127828.

[8] M. Akiba, Y. L. Chiu, K. Shimizu, and G. Liang, “Teacher salary and national achievement: A cross-national analysis of 30 countries,” *Int J Educ Res*, vol. 53, pp. 171–181, 2012, doi: 10.1016/j.ijer.2012.03.007.

[9] M. L. Blackburn, “Are U.S. teacher salaries competitive? Accounting for geography and the retransformation bias in logarithmic regressions,” *Econ Educ Rev*, vol. 84, Oct. 2021, doi: 10.1016/j.econedurev.2021.102169.

[10] E. García and E. S. Han, “Teachers’ Base Salary and Districts’ Academic Performance: Evidence From National Data,” *Sage Open*, vol. 12, no. 1, Mar. 2022, doi: 10.1177/21582440221082138.

[11] Y. T. Matbouli and S. M. Alghamdi, “Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations,” *Information (Switzerland)*, vol. 13, no. 10, Oct. 2022, doi: 10.3390/info13100495.

[12] Das Sayan, Barik Rupashri, and Mukherjee Ayush, “SALARY PREDICTION USING REGRESSION TECHNIQUES,” pp. 1–5, 2015.

[13] X. Dong, X. Zhu, M. Hu, and J. Bao, “A Methodology for Predicting Ground Delay Program Incidence through Machine Learning,” *Sustainability (Switzerland)*, vol. 15, no. 8, Apr. 2023, doi: 10.3390/su15086883.

[14] “10.2478_danb-2022-0017 (1)”.

[15] F. M. Basysyar and G. Dwilestari, “House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection,” *Acadlore Transactions on AI and Machine Learning*, vol. 1, no. 1, pp. 11–21, Nov. 2022, doi: 10.56578/ataiml010103.

[16] F. Emmert-Streib and M. Dehmer, “High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection,” *Machine Learning and Knowledge Extraction*, vol. 1, no. 1. MDPI, pp. 359–383, Dec. 01, 2019. doi: 10.3390/make1010021.

[17] X. Yan and X. Gang Su, “Linear Regression Analysis Theory and Computing.”

[18] Gujarati D., *Econometría*, Quinta., vol. 5, no. 1. McGRAW-HILL, 2010.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. 2014. [Online]. Available: http://www.springer.com/series/417

[20] K. Das, N. Das Chatterjee, D. Jana, and R. K. Bhattacharya, “Application of land-use regression model with regularization algorithm to assess PM2.5 and PM10 concentration and health risk in Kolkata Metropolitan,” *Urban Clim*, vol. 49, May 2023, doi: 10.1016/j.uclim.2023.101473.

[21] W. K. Härdle and Z. Hlávka, *Multivariate Statistics*. 2015.

[22] Montgomery Douglas, Peck Elizabeth, and Vining Geoffrey, *INTRODUCTION TO LINEAR REGRESSION ANALYSIS*. Arizona State University, 2012.

[23] M. Shanker, "Effect of Data Standardization on Neural Network Training," *Int. J. Mgmt Sci*, vol. 24, no. 4, pp. 385–397, 1996.

[24] L. Atlas *et al.*, "Performance Comparisons Between Backpropagation Networks and Classification Trees on Three Real-World Applications."

[25] S. R. M-Dawam and K. R. Ku-Mahamud, "Reservoir water level forecasting using normalization and multiple regression," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 1, pp. 443–449, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp443-449.

[26] V. Sarala and K. Ganesh, "EMPLOYEE SALARY PREDICTION SYSTEM USING MACHINE LEARNING," pp. 1–6, 2023.

[27] H. Aminu, B. Imam Yau, F. Umar Zambuk, E. Ramsom Nanin, A. Abdullahi, and I. Zahraddeen Yakubu, "Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm," 2023. [Online]. Available: www.ijisrt.com

[28] G. F. Dreher, "PREDICTING THE SALARY SATISFACTION OF EXEMPT EMPLOYEES," *Pers Psychol*, p. 34, 1981.

[29] A. K. Mohamed Saeed, P. Y. Abdullah, and A. T. Tahir, "Salary Prediction for Computer Engineering Positions in India," *Journal of Applied Science and Technology Trends*, vol. 4, no. 01, pp. 13–18, Feb. 2023, doi: 10.38094/jastt401140.

[30] Dutta Samanda, Halder Airiddha, and Dasgupta Kousik, "Design of a novel Prediction Engine for predicting suitable salary for a job," pp. 1–5, Nov. 2018.

[31] L. A. Geraldo-Campos, J. J. Soria, and T. Pando-Ezcurra, "Machine Learning for Credit Risk in the Reactive Peru Program: A Comparison of the Lasso and Ridge Regression Models," *Economies*, vol. 10, no. 8, Aug. 2022, doi: 10.3390/economies10080188.

6 ANEXOS

6.1 Evidencia de la Sumisión del articulo en un conference paper.

6.1.1 Carta de Aceptación de la Revista.

LETTER OF ACCEPTANCE

13th Computer Science On-line Conference 2024

Dear Jose Luis Tinoco Ramos,

Organizing & Program Committee is pleased to announce that your paper:

Machine Learning models for salary prediction in Peruvian teachers of regular basic education (Paper ID: 113022)
Author(s):Tinoco Ramos Jose Luis,

was Accepted

for the 13th Computer Science On-line Conference 2024.

For finishing your registration follow instruction, which has been already sent by e-mail to all authors of accepted papers (or follow instruction on https://csoc.openpublish.eu)

CSOC2024 is held on-line from 4/25/2024 to 4/27/2024.

Conference organization (sponsored by): OpenPublish.eu

Organization Committee Chair:
Radek Silhavy, Ph.D.

Radek Silhavy, Ph.D.
Organizing Committee Chair

6.1.2 Certificado de presentación de conference paper.

CSOC2024

CERTIFICATE OF PARTICIPATION

13th Computer Science On-line Conference 2024, April 25, 2024 - April 28, 2024

Awarded to

Tinoco Ramos Jose Luis

For the Paper presentation:

Machine Learning models for salary prediction in Peruvian teachers of regular basic education

Radek Silhavy, Ph.D.
Organising & Program Chair
OpenPublish.eu, s.r.o. Website: www.openpublish.eu

6.2 Copia de la resolución de inscripción del perfil de proyecto de tesis en formato articulo por el consejo de facultad correspondiente.

"AÑO DEL BICENTENARIO, DE LA CONSOLIDACIÓN DE NUESTRA INDEPENDENCIA, Y DE LA CONMEMORACIÓN DE LAS HEROICAS BATALLAS DE JUNÍN Y AYACUCHO"

RESOLUCIÓN N° 0202-2024/UPeU-FIA-CF-T

Lima, Ñaña 09 de abril de 2024

VISTO:

El expediente de **Jhoset Yamiel Yupanqui Arellano,** identificado(a) con código universitario N° 200310530 y **Jose Luis Tinoco Ramos,** identificado(a) con código universitario N° 200411251, de la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión;

CONSIDERANDO:

Que la Universidad Peruana Unión tiene autonomía académica, administrativa y normativa, dentro del ámbito establecido por la Ley Universitaria N° 30220 y el Estatuto de la Universidad;

Que la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, mediante sus reglamentos académicos y administrativos, ha establecido las formas y procedimientos para la designación del Comité Dictaminador del proyecto de tesis;

Que **Jhoset Yamiel Yupanqui Arellano** y **Jose Luis Tinoco Ramos,** han concluido el desarrollo de la tesis en formato artículo y con la opinión favorable de su asesor, solicitan la designación del Comité Dictaminador respectivo;

Estando a lo acordado en la sesión del Consejo de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, celebrada el 09 de abril de 2024, y en aplicación del Estatuto y el Reglamento General de Investigación de la Universidad;

SE RESUELVE:

Designar el Comité Dictaminador encargado de administrar el proceso de dictamen correspondiente a la tesis en formato artículo, titulada "Modelos de Machine Learning para la predicción del salario en docentes peruanos de educación básica regular", presentado por **Jhoset Yamiel Yupanqui Arellano** y **Jose Luis Tinoco Ramos,** otorgándoles un plazo máximo de diez (10) hábiles, posterior a la fecha de recepción de la presente resolución, para emitir el dictamen respectivo a través de la plataforma oficial.

Dictaminador 1: Mg. Nemias Saboya Rios
Dictaminador 2: Mg. Ferdinan Edgardo Pineda Anco

Regístrese, comuníquese y archívese.

Dra. Erika Inés Acuña Salinas
DECANA

Ph.D. Silvia Pilco Quesada
SECRETARIA ACADÉMICA

cc:
-Interesado
-Jurado (02)
-Archivo

6.3 Carta de Autorización de la UGEL Ventanilla para el uso de su información.

GOBIERNO REGIONAL CALLAO

DREC Dirección Regional de Educación del Callao

UGEL VENTANILLA

"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junin y Ayacucho"

Ventanilla, 08 FEB. 2024

OFICIO N° 157 -2024-UGEL VENTANILLA/DIR

Señorita:
Ing. Mag. GERALDINE VERÓNICA ALVIZURI LLERENA
Directora E.P. Ingeniería de Sistemas
Facultad de Ingeniería y Arquitectura
Universidad Peruana Unión
Presente.-

ASUNTO: ACEPTACIÓN DE ACCESO Y FACILIDADES PARA ELABORACIÓN DE PROYECTO DE TESIS.

Ref. : Carta N° 001-2024/UPeU-FIA-INGENIERIA DE SISTEMAS.

De mi especial consideración:

Es grato dirigirme a usted y expresarle mi cordial saludo; y a la vez comunicar que, de acuerdo a lo solicitado mediante la Carta de la referencia, mi Despacho gustosamente brindará las facilidades a los bachilleres:

1. **José Luis Tinoco Ramos y**
2. **Jhoset Yamiel Yupanqui Arellano**

Para desarrollar en la Unidad de Gestión Educativa Local Ventanilla su Proyecto de **Tesis Modelos de Machine Learning para la predicción de escala remunerativa en docentes peruanos de educación básica regular,** lo cual les será beneficioso en su superación profesional y el logro de sus objetivos.

Segura de contar con la atención que le brinde al presente, reitero las muestras de mi especial consideración y estima personal.

Atentamente,

MINISTERIO DE EDUCACIÓN DIRECCIÓN

Dra. OLGA LIDIA MÉNDEZ SALAS
DIRECTORA
Unidad de Gestión Educativa Local - Ventanilla

OLMS/DIR
mea/SEC