UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERIA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Predicción salarial con Machine Learning en docentes contratados de la Región del Cusco - Perú

Tesis para obtener el Título Profesional de Ingeniero de Sistemas

Autor:

Segundo Canahuire Hilari

Joel Eduardo Larico Carbajal

Asesor:

Mg. Ferdinand Edgardo Pineda Ancco

Juliaca, abril de 2024

DECLARACIÓN JURADA DE ORIGINALIDAD DE TESIS

Yo Mg. Ferdinand Edgardo Pineda Ancco, 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: **"PREDICCIÓN SALARIAL CON MACHINE LEARNING EN DOCENTES CONTRATADOS DE LA REGIÓN DEL CUSCO - PERÚ"** de los autores **Segundo Canahuire Hilari** y **Joel Eduardo Larico Carbajal**, tiene un índice de similitud de 14% 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 Juliaca, a los 08 días del mes de mayo del año 2024.

Mg. Ferdinand Edgardo Pineda Ancco
Asesor

089

ACTA DE SUSTENTACIÓN DE TESIS

En Puno, Juliaca, Villa Chullunquiani, a 19 día(s) del mes de abril del año 20[illegible] siendo las 08:00 horas, se reunieron los miembros del jurado en la Universidad Peruana Unión Campus Juliaca, bajo la dirección del (de la) presidente(a): Mg. Abel Angel Sullon Macalupu, el (la) secretario(a): Mg. David Mamani Pari y los demás miembros: Mg. Nemias Saboya Ríos Dr. Juan Tenis Soria Quijaite y el (la) asesor(a) Msc. Ferdinand Edgardo Rineda Ancco con el propósito de administrar el acto académico de sustentación de la tesis titulado: Predicción salarial con Machine Learning en docentes contratados de la Región del Cusco - Perú del(los) bachiller/es: a) Segundo Ganahuire Hilari b) Joel Eduardo Larico Garbajal 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): Segundo Ganahuire Hilari

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	18	A-	Muy Bueno	Sobresaliente

Bachiller (b): Joel Eduardo Larico Garbajal

CALIFICACIÓN	ESCALAS			Mérito
	Vigesimal	Literal	Cualitativa	
Aprobado	14	C	Aceptable	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 — Miembro — Asesor/a — Miembro — Bachiller (a) — Bachiller (b) — Bachiller (c)

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Este artículo presenta un análisis de modelos de aprendizaje automático (ML) para predecir los salarios de 11,392 docentes contrata- dos designados en la Ugel de la región Cusco-Perú´, utilizando datos recientes del sistema único de planillas. El punto focal del estudio son los docentes contratados, excluyendo deliberadamente del análisis los salarios de los docentesnombrados. Un resultado significativo de esta investigación es la identificación de un nuevo modelo de ML capaz de predecir los salarios de los docentes con considerable precisión, basado en variables regresoras estrechamente relacionadas con el salario. Este hallazgo es digno de mención porque llena un vacío en las aplicaciones de ML existentes para la predicción salarial, loque indica una dirección prometedora para futuras investigaciones en esta área. La metodología empleada para analizar los datos salariales, si bien es exhaustiva, no tiene en cuenta las diferenciasde género, que pueden afectar la variación salarial durante el periodo de tres años considerado. Este descuido sugiere que las investigaciones futuras deberían incluir una gama más amplia de variables, incluido el género, para mejorar la precisión y aplicabilidad de las predicciones salarialestanto para los docentes nombrados como para los contratados. Un enfoque de este tipo podría proporcionar información más matizada sobre los factores que influyen en los salarios de los docentes y ayudar a desarrollar modelos salariales más equitativos y eficaces. Una de las contribuciones clave del artículo es el examen detallado de los factores que influyen en los salarios de los docentes designados, incluida la edad, el cargo, el nivel educativo, el código modular, las horas semanales, el periodo y otras variables Dummy. El uso de modelos Decision Tree Regressor(DTR), Gradient Boosting Regressor(GBR), Random Forest Regressor(RFR), Red Neuronal Regresor(RNR) y Support Vector Regressor(SVR) arrojo métricas precisas para elegir el mejor modelo para la predicción salarial. Esta investigación no solo avanza nuestra comprensiónde los determinantes de los salarios docentes en la región de Cusco-Perú, sino que también ofrece un marco valioso para estudios similares en otros contextos.

Palabras Clave. *Aprendizaje automático, remuneración, aumento de gradiente, árbol de decisión, bosque aleatorio, neural rojo, vector de soporte*.

ABSTRACT

This article presents an analysis of machine learning (ML) models to predict the salaries of 11,392 contracted teachers designated in the Ugel of the Cusco-Peru region, using recent data from the unique payroll system. The focal point of the study is the contracted teachers, deliberately excluding the salaries of appointed 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 predictor variables closely related to salary. This finding is noteworthy as 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 salary data, while thorough, does not account for gender differences, which may affect salary variation during the three-year period considered. This oversight suggests that future research should include a broader range of variables, including gender, to improve the accuracy and applicability of salary predictions for both appointed and contracted teachers. Such an approach could provide more nuanced information about 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 factors influencing the salaries of designated teachers, including age, position, educational level, modular code, weekly hours, period, and other dummy variables. The use of Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), Neural Network Regressor (NNR), and Support Vector Regressor (SVR) models yielded precise metrics to choose the best model for salary prediction. This research not only advances our understanding of the determinants of teacher salaries in the Cusco-Peru region but also offers.

***Keywords*:** *Machine Learning, Remuneration, Gradient Boosting, Decision Tree, RandomForest, Red Neural, Support Vector.*

1. INTRODUCCIÓN

La estimación salarial emerge como un componente primordial para abordar una diversidad de problemas, entre los cuales incluyen el rendimiento académico del estudiante, tanto en su vertiente positiva como negativa, y el desempeño laboral del docente [4], [16], [37], [5], [6]. Estos aspectos tienen un impacto significativo tanto en el ámbito educativo como en el económico, los cuales son pilares fundamentales del desarrollo socioeconómico [1] , [2] ,[16] .En este contexto, se destaca la relevancia del salario del docente como un factor crucial que influye en la eficacia y sostenibilidad del sistema educativo [3] , [8] , [11] Además, como un elemento clave que afecta tanto al ámbito educativo como al económico mucho más aún si el docente está en calidad de contratado [5],[6]. El propósito de este estudio es desarrollar un modelo de predicción salarial identificando el algo- ritmo más eficaces y eficientes, que destaque por su elevado rendimiento según diversas métricas. en este estudio busca demostrar que el modelo desarrollado y entrenado predice el salario del docente contratado considerando los factores más relevantes.

En Sudamérica, el bajo rendimiento laboral de los docentes es relacionado con los salarios reducidos que han percibido en los últimos años [1], [15], [17] Además, se ha observado que la relación remunerativa de los docentes guarda una correlación directa con su desempeño y posteriorfortalecimiento profesional. Como también se ubicó estudios realizados se han empleado diversosmétodos convencionales [7],[1],[9],[10]. Este enfoque se centra en la estimación de los incentivossalariales a través de la consideración de la productividad laboral a base del análisis salarial, asimismo se identifica la Influencia de los salarios de los docentes sobre el rendimiento académico.[8], [28]. se ha confirmado que los salarios inciden en la calidad de la enseñanza impartida a los estudiantes [23]. En la literatura existente, se ha documentado una variedad de estudios que emplean información de individuos solicitantes de empleo con el propósito de prever de manera precisa y asegurar una compensación salarial adecuada en el ámbito laboral [23]. En este contexto,se llevó a cabo una comparación entre dos algoritmos ampliamente utilizados en la predicción decompensaciones salariales: Gradient Boosting Decision Trees (GBDT) y Random Forest (RF). Serealizó un análisis exhaustivo, empleando diversas métricas, y se determinó que el algoritmo GBDT demostró un rendimiento superior, según lo documentado en [11],[19], [34]. Asimismo, se evaluó el rendimiento del algoritmo Decision Tree (ID3) en la predicción de compensaciones salariales para estudiantes graduados, encontrando un rendimiento del 61.37% con una precisión relativamente baja del 73.96 %.[22]. Estos resultados indicaron que el algoritmo

ID3 no cumplía con los estándares necesarios para su implementación en el sistema de predicción, tal como se menciona en [22], [23]. Adicionalmente, durante la revisión de la literatura, se identificaron variosalgoritmos relevantes, incluyendo Random Forest Regressor, Decision Tree Regressor y Ridge Regression. [19],[26]. Estos hallazgos proporcionan una base sólida para la identificación de los algoritmos más robustos que ofrecen predicciones certeras, informando así el desarrollo y selección de los algoritmos en nuestro estudio.

2. METODOLOGÍA

El tipo de estudio es aplicado, ya que busca emplear diferentes algoritmos, técnicas de machine learning para realizar una predicción aproximada de la deducción del salario de un docente [44]. Esto se presenta como una herramienta útil para los usuarios interesados en realizar una adjudicación en el proceso de contratación docente en alguna Unidad de Gestión Educativa Local (UGEL). En el contexto de una investigación aplicada, resulta fundamental establecer la duración precisa de la investigación [44]. En contraste con un enfoque de alcance longitudinal, enel que la recolección de datos se hace de manera continua, en este estudio se trabajó con una muestra predefinida y recolectada [44]. El diseño metodológico adoptado se enmarca en un alcance transversal, ya que la recolección de datos se hizo en un periodo especifico [43],[44]. Lainformación utilizada en esta investigación se obtuvo directamente de la base de datos del servidordel Sistema Único de Planillas, lo que permitió acceder a los datos pertinentes de manera eficientey oportuna. La metodología adoptada en esta investigación se basa en un enfoque cuantitativo, caracterizado por su naturaleza numérica y el uso de variables cuantitativas [42], [44]. La remuneración y el enfoque implica una secuencia de pasos diseñados para probar hipótesis especificas alineadas con los objetivos del estudio, centrándose en problemas concretos y estableciendo variables que son declaradas de manera cuantitativa, utilizando valores numéricos yanálisis estadísticos [29],[30],[31]. Desde una perspectiva matemática, un salario es un tipo de datodecimal y numérico, lo que facilita su entrenamiento con diversos algoritmos y funciones matemáticas. En el caso de los datos nominales, se emplearon técnicas como el uso de diccionario,similares variables dummy y one-hot encoding, las cuales permiten categorizar los datos de maneraadecuada [22]. Las remuneraciones, es una retribución monetaria emitida por un titular de algunainstitución o entidad. En este contexto, las remuneraciones de los docentes contratados y los docentes nombrados en el sector peruano

difieren entre sí debido a la diferencia en la permanencia laboral [12],[14],[13]. Es importante destacar que un docente nombrado no puede definir su salario de manera aleatoria, sino que este está definido según la escala que ostenta. Por otro lado, un docente contratado sí puede definir anualmente su salario mensual, ya que puede seleccionar alguna vacante que este libre, en el proceso de contratación de docentes al que esté sometido es por ello que la población de este estudio es la predicción salarial de docentes contratados.

2.1. Técnicas de machine learning

2.1.1. Gradient Boosting Regressor

El algoritmo de Gradient Boosting se basa en la creación de árboles para realizar predicciones más robustas [42], la formula se visualiza en la ecuación (1).

$$F(x) = F_{\text{ant}}(x) + \eta \cdot h_i(x)$$

Figuras 1. Formula Gradiend Boosting Regressor

Este algoritmo funciona agregando secuencialmente predictores a un conjunto de árboles, cada uno corrigiendo los errores cometidos por sus predecesores, hasta simplificar y llegar a optimizar la predicción óptima.

2.1.2. Random Forest Regresor

Es un algoritmo que proporciona una medida de la importancia de las variables para la predicción [22]. La teoría detrás de cómo se calcula y como se utiliza para interpretar y comprender el modelo, que tiene la fórmula matemática, mostrada en la ecuación (2).

$$\hat{Y} = \frac{1}{M} \sum_{i=1}^{M} T(x; \Theta_i)$$

Figuras 2. Formula Random Forest Regresor

2.1.3. Decision Tree Regressor

Es un algoritmo identificado por su forma de decisión también puede ser usado para realizar funciones. Discute los conceptos teóricos de la regresión y como se aplican en el contexto de los árboles de decisión [18],[22],[27], esto puede incluir una revisión de los diferentes tipos de regresión y como se utilizan para predecir valores numéricos, lo que se muestra en la ecuación (3).

$$\hat{y}(x) = \sum_{m=1}^{M} c_m \cdot \mathbf{1}(x \in R_m)$$

Figuras 3. Formula Decision Tree Regressor

2.1.4. Red Neural Regresor

Una Red Neuronal Regresora, es un tipo específico de red neuronal artificial que se emplea para resolver problemas relacionados con la regresión. A diferencia de las redes neuronales utilizadas para clasificación, cuya función es predecir la clase a la que pertenece una entrada, las redes neuronales regresoras tienen la capacidad de predecir valores numéricos continuos [18], [40]. Es la arquitectura de una Red Neuronal Regresora, se pueden identificar capas de neuronas que están interconectadas entre sí. Cada neurona en estas capas está vinculada a una función de activación particular [20],[41]. La red recibe entradas numéricas, que pueden ser características o variables explicativas, y procesa estas entradas a través de múltiples capas ocultas para calcular una salida continua y que se muestra en la siguiente ecuación (4).

$$\hat{y} = f(W^{(L)} \cdot f(W^{(L-1)} \cdot \ldots \cdot f(W^{(1)} \cdot x + b^{(1)}) + b^{(2)}) + \ldots + b^{(L)})$$

Figuras 4. Formula Red Neural Regresor

2.1.5. Support Vector Regressor

Es un método de aprendizaje automático diseñado para abordar problemas de regresión [25],[24],[26]. Similar a su contraparte en clasificación, Support Vector Machines (SVM), el SVR se basa en el concepto de máquinas de vectores de soporte para determinar la función que mejor se ajuste a un conjunto de datos especifico el cual se muestra en la ecuación (5).

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b$$

Figuras 5. Formula Support Vector Regressor

2.1.6. Hiperparámetros Principales

Los hiperparametros seleccionados para el estudio comprenden el número de estimadores n estimators, también conocido como el número de árboles utilizados, la tasa de aprendizaje learning rate, que controla la contribución de cada árbol en cada predicción, y la profundidad máxima de cada árbol en el conjunto max depth. los tres primeros algoritmos que tienen particularidad fueron considerados con los siguientes valores.

2.2. Métrica

2.2.1. Coeficiente de Determinación

Es importante tener en cuenta que el R2 tiene limitaciones y debe considerarse junto con otras métricas de evaluación del modelo para obtener una imagen completa de su rendimiento. Además, R2 puede ser sensible a ciertos problemas, como la presencia de variables irrelevantes en el modelo o la heterocedasticidad de los residuos. Por lo tanto, es importante interpretar R2 en el contexto adecuado y complementarlo con otras métricas de evaluación del modelo. que se muestra en la ecuación (6). El coeficiente de determinación, denotado como R2, es una medida estadística que indica cuanta variabilidad en la variable de respuesta puede explicarse por el modelo estadístico, [18]. En el contexto de modelos de regresión, el coeficiente de determinación se utiliza para evaluar que tan bien el modelo se ajusta a los datos observados. Proporciona una medida de la calidad de la predicción del modelo y varía entre 0 y 1.

$$R^2 = 1 - \frac{SSR}{SST}$$

Figuras 6. Formula Coeficiente de Determinación

2.2.2. Raíz del Error Cuadrático Medio.

Esta métrica constituye una métrica de evaluación habitualmente empleada para cuantificarla discrepancia entre las observaciones reales y las predicciones generadas por un modelo[18],[20]. Se utiliza de manera extendida en el análisis de la precisión de modelos de regresión. ,[36],[38]. Su cálculo implica determinar la raíz cuadrada de la media de los cuadrados de las diferencias entre las predicciones del modelo y los valores observados como se detalla en la ecuación (7).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Figuras 6. Formula Raíz del Error Cuadrático Medio

2.2.3. Error Absoluto Medio.

Métrica utilizada para evaluar la precisión de un modelo de regresión [32],[33]. Se calcula como la media de las diferencias absolutas entre los valores predichos por el modelo y los valores observados en los datos reales [12],[18],[20]. Es una medida robusta y fácil de interpretar, ya que representa la magnitud promedio del error entre las predicciones y los valores reales, sin considerar la dirección de las diferencias, mostrado en la ecuación (8).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Figuras 7. Formula Error Absoluto Medio

3. METODOLOGÍA DE DESARROLLADO DEL MODELO

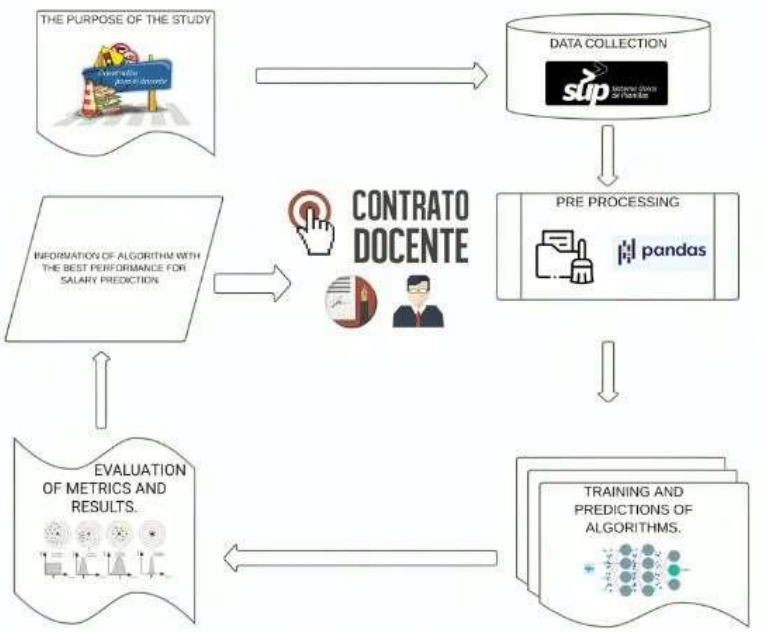
Contextualización de la Metodología. Se usó la metodología Cross-Industria Standard Process for Data Mining (CRISP-DM), reconocida por proporcionar un marco estructurado que facilita la planificación, ejecución y evaluación de proyectos de minería de datos [44]. En el contexto de este estudio, se tomó esta metodología como referencia principal y se elaboró una estructura metodológica específica para nuestro caso en la predicción salarial, mostrada en la figura 1.

3.1. Objetivo del Modelo

El empleo de técnicas de aprendizaje automático tiene como fin de prever compensaciones salariales que se posiciona como una herramienta esencial para profesionales docentes que aspiran asegurar una colocación laboral adecuada [12],[21],[19],[39].

3.2. Recolección de Datos

La recopilación de datos se realizó mediante la técnica documental, con el objetivo de extraer información del Sistema Único de Planillas (SUP), llevada a cabo con la debida autorización del titular de la entidad correspondiente. La información recopilada se exporta un documento de Microsoft Excel, identificado con el nombre de ”DATASET”, con variables dependientes(salario) y variable independiente X1, X2, X3, X4, ..., Xn.



Figuras 8. Distribución salarial según los cargos de docente

3.3. Preprocesamiento

El desarrollo del modelo, que abarca el preprocesamiento de los datos, el entrenamiento con los algoritmos seleccionados y el análisis de las métricas pertinentes, se llevó a cabo en la plataforma Anaconda. Para estos propósitos, se utilizó la librería Pandas para la manipulación y análisis de datos, así como los algoritmos disponibles en la librería Scikit-learn para el entrenamiento y la evaluación del modelo,[11] ,[34] ,[35],[36]. Además, se procedió a la identificación de las variables mencionadas, tal como se muestra en la Tabla 1.

Dado el carácter de los datos como objetos y la adopción de una metodología de investigación cuantitativa, se optó por utilizar diccionarios para categorizar las variables. En las columnas que contenían información de fechas, como la Fecha de Nacimiento, se llevó a cabo una transformación de los datos en valores numéricos, representando exclusivamente el año correspondiente. Este procedimiento resulto en una uniformidad en la cual los datos pueden ser identificados como de tipo numérico y de punto flotante. Adicionalmente, se procedió a evaluar la presencia de valores nulos, lo que condujo a la elaboración de un informe grafico que señala la ausencia de dichos valores. Esto demuestra que la recopilación de datos fue depurada previamente durante el proceso de extracción mediante la consulta SQL.

TABLA 9

VARIABLES IDENTIFICADAS EN LA INVESTIGACIÓN

Variable	Tipo de data	Tipo de Variable
GENERO	num	Independiente
FECHA NACIMIENTO	num	Independent
EDAD ACTUAL	num	Independent
CARGO	num	Independent
NIVEL EDUCATIVO	num	Independent
CODIGO DE IIEE	num	Independent
HORAS TRABAJADAS SEMANALES	num	Independent
PERIODOS	num	Independent
REGIMEN PENSIONARIO	num	Independent
APORTACIONES OBLIGATORIAS	float	Independent
PAGO DE SEGURO	float	Independent
ESSALUD	float	Independent
TRIBUTABLE	float	Independent
IMPONIBLE TRIBUTABLE	float	Independent
TOTAL HABERES	float	Dependent
TOTAL DE LOS DESCUENTOS	float	Independent
LIQUIDO	float	Independent

3.4. Distribución de los Datos

Durante el proceso de distribución de los datos, se asignó el 80% del conjunto de datos para su uso en propósitos de entrenamiento, mientras que el 20% restante fue reservado para la fase de prueba.

3.5. Entrenamiento del algoritmo

El entrenamiento de los algoritmos se llevó a cabo utilizando la cantidad de datos estimada para la fase de prueba, que consta de 11,392 filas y 17 columnas, junto con sus respectivos parámetros estimados.

3.6. Predicción y evaluación de resultados

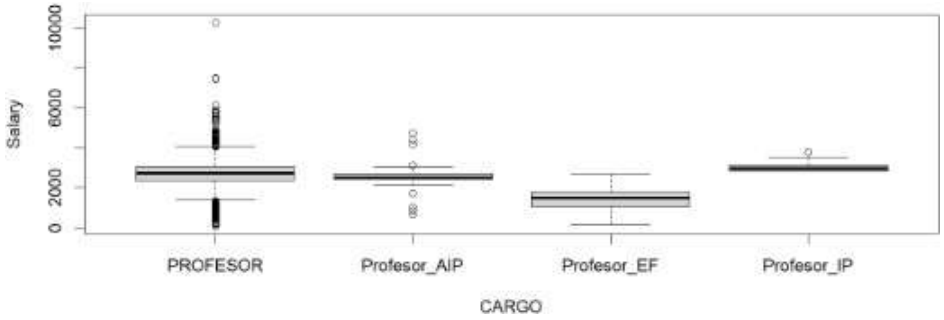
La realización de predicciones y la evaluación de resultados se llevó a cabo mediante tres enfoques diferentes, cada uno generado utilizando características específicas de hiperparámetros y evaluados de forma independiente de acuerdo a su naturaleza particular. Los tres algoritmos identificados para su utilización fueron: Random Forest Regressor, Decision Tree Regressor y Gradient Boosting. Estos presentan una característica distintiva relacionada con la homogeneidad de tres hiperparámetros específicos: la cantidad de árboles, el control de aleatoriedad

y la profundidad que limita la cantidad de nodos en que cada ´árbol divida. La predicción salarial se realizó utilizando la cantidad de datos distribuidos para la prueba, que consta de 1140 filas y 15 columnas. Las métricas utilizadas como el RMSE, MAE y R2.

4. RESULTADOS

4.1. Análisis descriptivo

Cargo Laboral Los docentes ocupan diversos cargos en función de sus especializaciones, los cuales se asocian con distintos niveles salariales, tal como se puede observar en la Figura 2.



Figuras 10 Distribución Salarial Según Cargos Docente.

Los docentes dentro de una institución son designados con diferentes nominaciones según su cargo, y cada uno de ellos desempeña funciones distintas, como se detalla en la Tabla 2.

TABLAS 2

DESCRIPCIÓN DE LA VARIABLE CARGO

Datos de la variable	Descripción	Datos de Categorización
Profesor	Docente normal de Aula Regular.	1
Profesor AIP	Docente de Aula de Innovación Pedagógica.	2
Profesor EF	Docente de Educación Fisica.	3
Profesor IP	Docente de innovación Pedagógica.	4

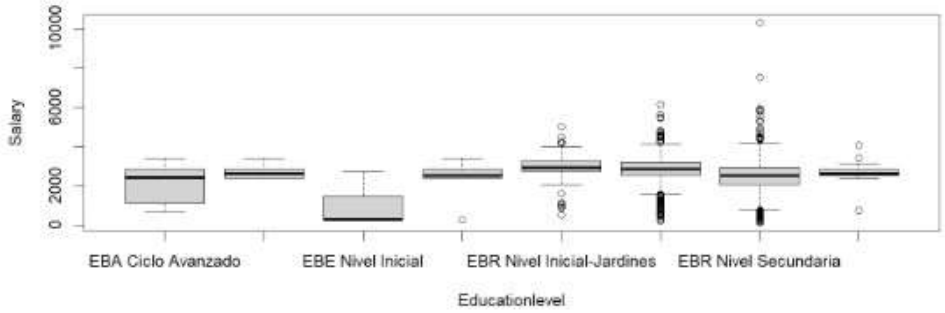
Nivel Educativo Como también identificamos los datos de la variable Nivel Educativo que identifica los diferentes modalidades y niveles de educación de en la Tabla 3.

Periodo Laboral El análisis de los salarios permitió obtener los 36 periodos analizados, equivalentes a un lapso de tres años.

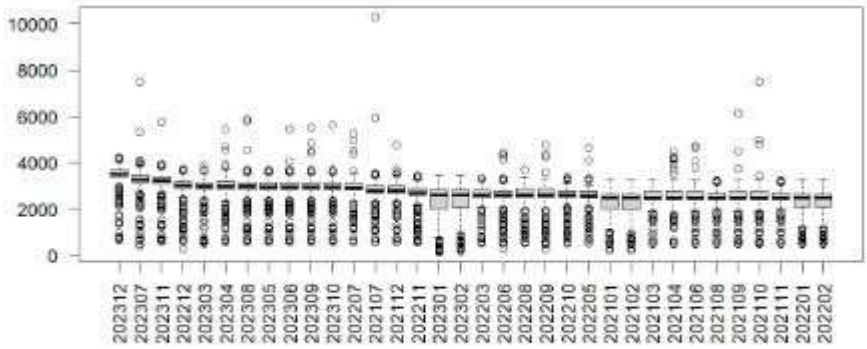
TABLAS 11

DESCRIPCIÓN DE LA VARIABLE NIVEL EDUCATIVO

Datos de la variable	Descripción	Datos de Categorización
EBA Ciclo Avanzado	Educación basica avanzada del ciclo avanzado.	1
EBE Nivel Inicial	Educación Basica Especial del Nivel Inicial.	2
EBR Nivel Inicial-Jardines	Educación Basica Regular del Nivel Inicial-Jardines.	3
EBR Nivel Secundaria	Educación Basica Regular del Nivel Secundaria.	4

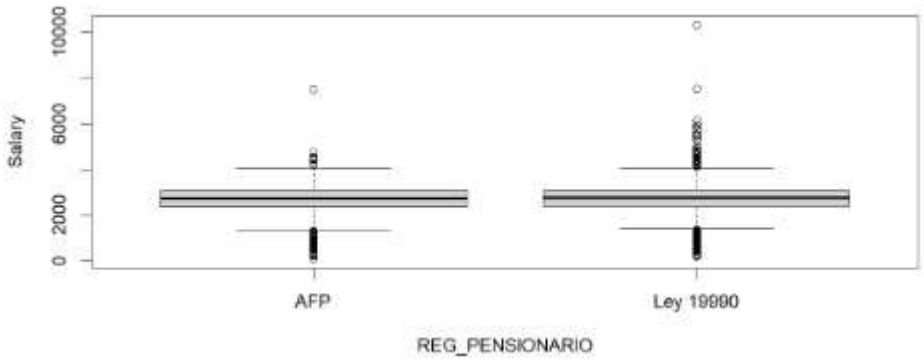


Figuras 13. Distribución salarial según los niveles educativos



Figuras 12. Distribución salarial según tiempo de servicio

Régimen Pensionario El régimen pensionario es una variable indispensable de analizar, ya que proporciona una perspectiva sobre la contribución de los docentes en el fondo de la ONP o AFP.



Figuras 14. Descripción estadística del régimen según el salario

4.2. Análisis de las predicciones

Las predicciones fueron llevadas a cabo utilizando diversos algoritmos y utilizando los datos asignados como prueba de testeo, los cuales reflejan la distribución de los datos. Estos resultados están disponibles en la tabla. 4.

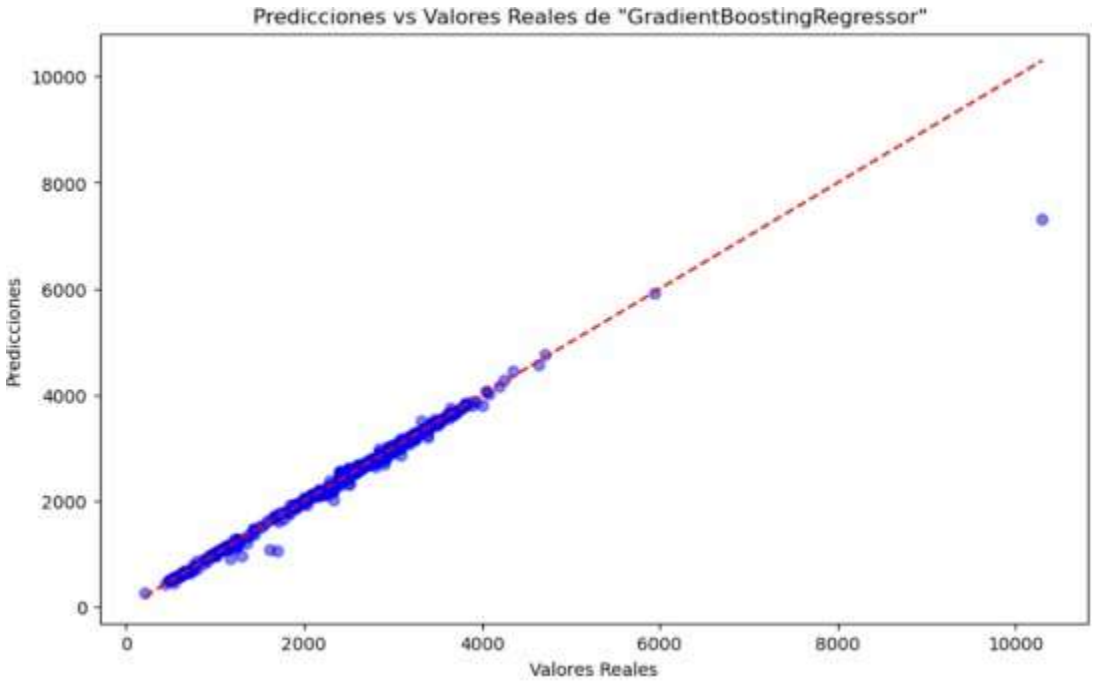
TABLAS 4

ESTIMATION OF METRICS WITH A NORMALIZATION IN THE DATA

Algoritmo	R^2	RMSE	MAE
GBR	0.9860251594377123	93.81509971476737	28.037611822746353
DTR	0.8787308456082283	276.35955726081613	170.57089135724908
RFR	0.882305219762534	272.2562853842356	163.30430451526675
RNR	0.6493418508795777	469.9387763365514	328.1931640268571
SVR	0.03650441325935483	807.9507555689243	544.097969044606

4.3. Análisis de dispersión de los resultados.

Se observa que dos algoritmos exhiben un alto índice de predicción: el Gradient Boosting Regressor y la Red Neuronal Regresora. A pesar de sus diferentes métodos de funcionamiento, ambos algoritmos se consideran óptimos para la predicción de salarios de docentes contratados. Se obtuvo un error cuadrático medio RMSE de 0.9860251594377123, un error absoluto medio MAE de 93.81509971476737 y un coeficiente de determinación de 28.037611822746353, lo que indica el mejor rendimiento en el estudio. La dispersión del algoritmo Decision Tree Regresso revela una predicción cercana a la perfección. Se identificó una dispersión similar a la predicción perfecta, lo que se refleja en un coeficiente de 0.986482765.



Figuras 15. Dispersion del Gradient Boosting Regressor

La dispersión del algoritmo Decision Tree Regressor se muestra en la Figura, indicando una aproximación a la predicción perfecta. El valor del coeficiente de determinación es de 0.8787308456082283. Este análisis proporciona una representación visual de la dispersión el algoritmo y su rendimiento en la predicción de los datos en el estudio realizado.

La dispersión del algoritmo Random Forest Regressor se muestra en la figura, indicando una aproximación a la predicción perfecta. El valor del coeficiente de determinación es de 0.870502225. Este análisis proporciona una representación visual de la dispersión del algoritmo y su rendimiento en la predicción de los datos en el estudio realizado.

La dispersión del algoritmo Red Neuronal Regresora se muestra en la figura, indicando una aproximación a la predicción perfecta. El valor del coeficiente de determinación es de 0.882305219762534. Este análisis proporciona una representación visual de la dispersión del algoritmo y su rendimiento en la predicción de los datos en el estudio realizado.

La dispersión del algoritmo Support Vector Regressor se muestra en la Figura, indicando una aproximación a la predicción perfecta en la figura.

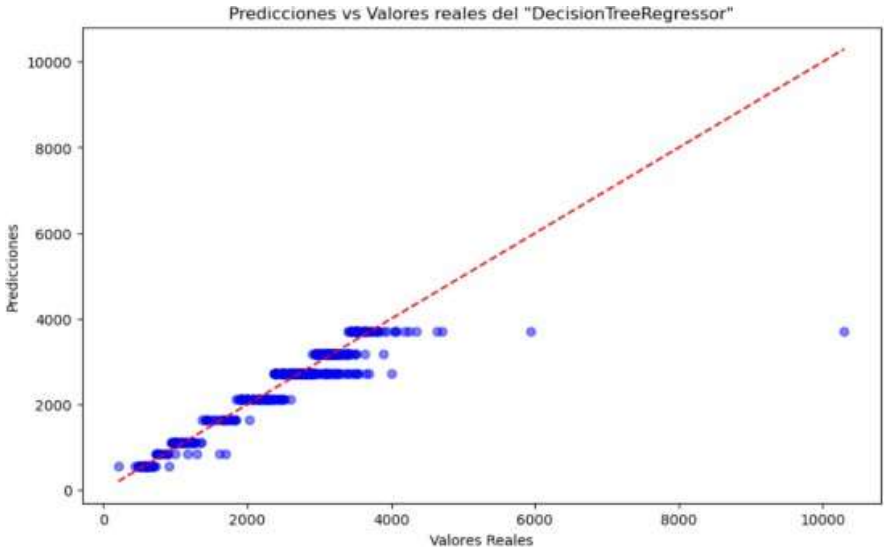


Figura 16. Dispersion del Decision Tree Regressor

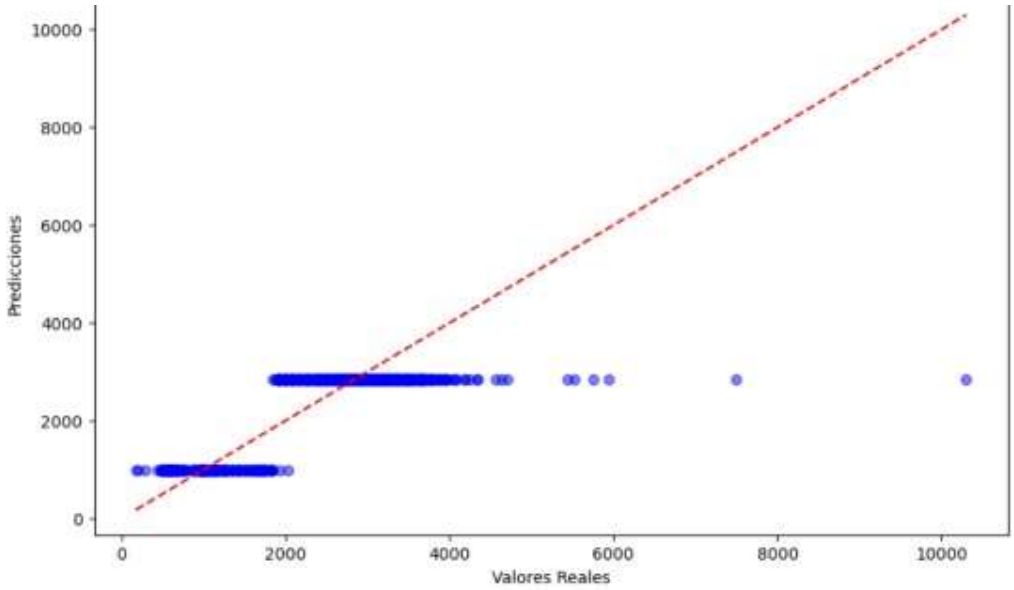


Figura 16. Dispersion del Random Forest Regressor

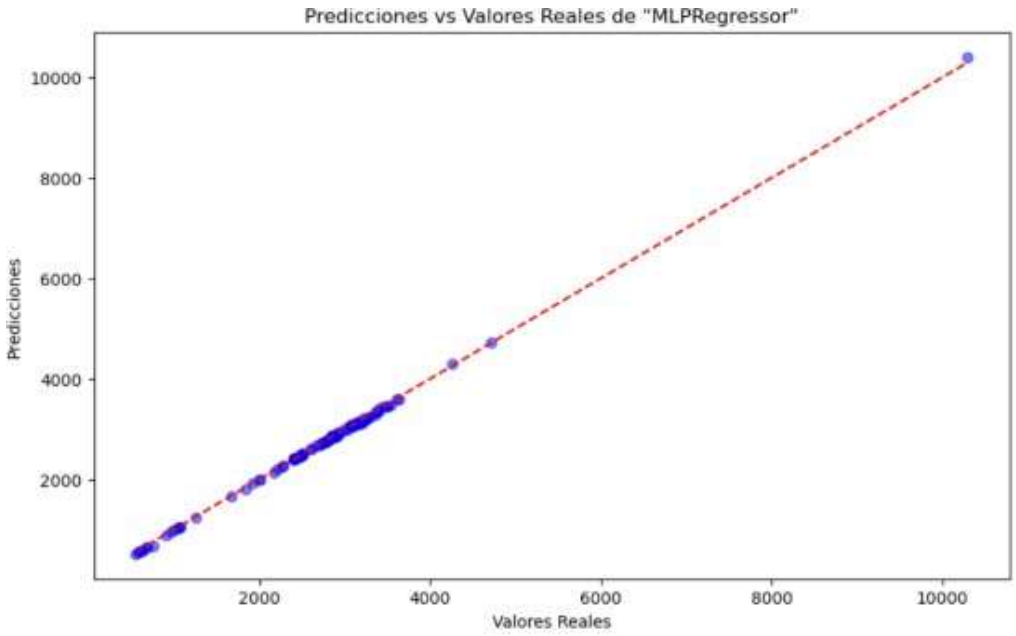


Figura 17. Dispersion de la Red Neuronal Regresora

El valor del coeficiente de determinación es de 0.03650441325935483. Este análisis proporciona una representación visual de la dispersión del algoritmo y su rendimiento en la predicción de los datos en el estudio realizado.

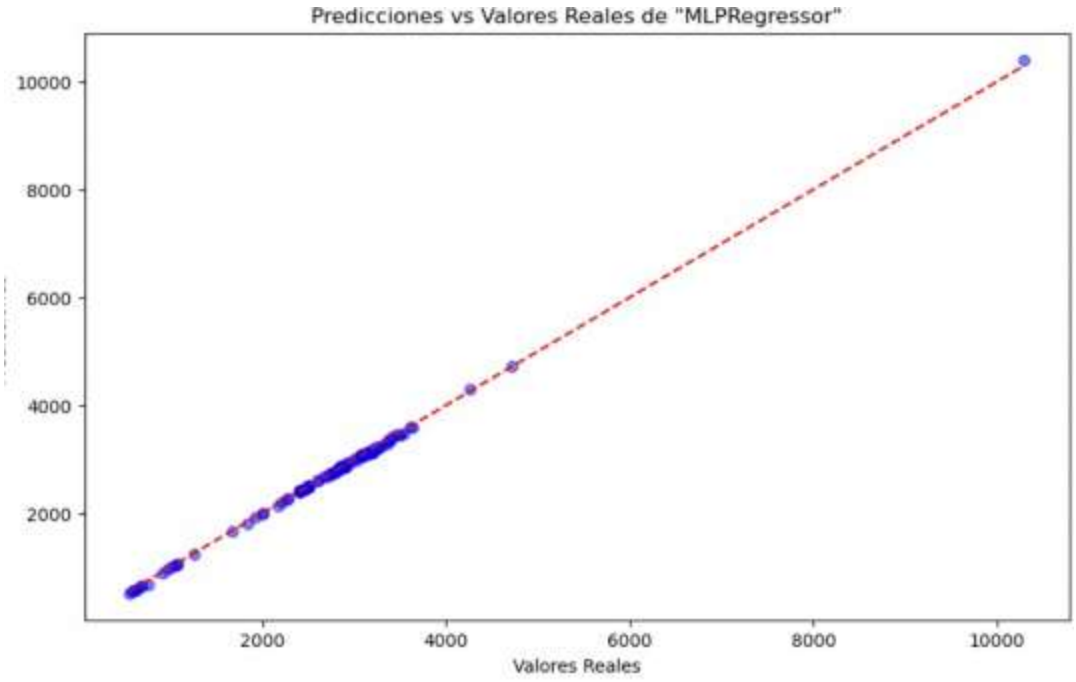


Figura 18. Dispersion del Support Vector Regresor

Es importante mencionar que los resultados en la Tabla 4 fueron normalizados para obtener una predicción perfecta. Sin embargo, como podemos ver en la Tabla 5, la Red Neuronal de Vectores de Soporte no se desempeña adecuada- mente al predecir los salarios de los profesores.

TABLAS 5

ESTIMATION OF METRICS WITHOUT NORMALIZATION IN THE DATA

Algoritmo	R^2	RMSE	MAE
GBR	0.9844852425983279	98.84889257129682	28.212597567917552
DTR	0.8787308456082288	276.3595572608155	170.5708913572491
RFR	0.882633969010152	271.875781517728	163.93307428510266
RNR	-10.259110941652075	2662.8785715408394	2541.8749240533475
SVR	-0.0519873593758835	813.9628323366182	547.9746204475647

El Gradient sigue funcionando óptimamente con una reducción mínima, manteniendo un alto índice de RMSE de 98.84889257129682 y un coeficiente de determinación de 0.9844852425983279. Esto no nos lleva a decidir tomar un único algoritmo como el óptimo para predecir los salarios de los profesores contratados.

5. CONCLUSIÓN

El estudio concluye que el mejor algoritmo para la predicción salarial para profesores contratados, con un rendimiento de 0,9860251594377123, es Gradiente Boosting, que demostró un alto nivel de efectividad. Considerando esto, cuando se implementa con datos de remuneración de docentes contratados en el sector de educación pública, supero al Decision Tree Regresor y al Random Forest Regresor en términos de entrenamiento y predicción, a pesar de compartir características como el número de árboles y la profundidad de cada árbol con sus nodos.

Se realizó un estudio comparativo de los resultados de las métricas utilizadas, como RMSE y MAE, que confirmaron el gradiente Impulsar regresor como un algoritmo robusto para la predicción salarial de docentes contratados, haciendo uso de árboles. Por otro lado, la red neuronal regresora fue identificada como el segundo mejor algoritmo. A diferencia del primer algoritmo, utiliza múltiples capas y núcleos. Se puede utilizar para predicción siempre que el conjunto de datos este normalizado.

6. REFERENCES

[1.] Cecilia Y. Cuellar and Jorge O. Moreno: Employment, wages, and the gender gap in Mexico: Evidence of three decades of the urban labor market

[2.] Cecilia Y. Cuellar and Jorge O. Moreno: Employment, wages, and the gender gap in Mexico: Evidence of three decades of the urban labor market

[3.] Mariam Al Akasheh and Esraa Faisal Malik and Omar Hujran and Nazar Zaki : A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review : 2024

[4.] Zöe Cullen and Ricardo Perez-Truglia : The salary taboo privacy norms andthe diffusion of information : 2023

[5.] Mikko Ranta and Mika Ylinen : Employee benefits and company performance: Evidence from a high-dimensional machine learning model : 2023

[6.] Salazar Vega Jose Shamir y Tamay Quilcate Esther Karol : Desempeño laboral en docentes de un colegio particular lima 2019 : 2021

[7.] Silva Escalante, Vilma del Rosario: Relación entre remuneraciones y el desempeño laboral docente en el instituto superior tecnológico privado" DIMAS : 2022.

[8.] César Álvarez Ramírez, Anginson Cervantes Irigoyen , Christian CoaquiraMamani, Edgard Oscco Bustíos , Fabricio Pacheco Calderón : Principales Causas de la Rotación Laboral en el Sector de Educación Básica Regular Privada de la Provincia de Arequipa : 2016

[9.] Luisa Güell: Estudio de la satisfacción laboral de los maestros : 2015

[10.] Peralta Valenzuela, Katherin Yuly, Zuniga Velasquez, Claudia Jacqueline : Desval-orizacion magisterial abordada desde la perspectiva de un docente:un estudio bi- ografico : 2021

[11.] Baggerly esthefanya espillco flores : Gestión de la remuneración y finanzaspersonales de los docentes de la Ugel Lucanas Ayacucho 2017 : 2019

[12.] Tee Zhen Quan Mafas Raheem and Kuala Lumpur and Malaysia Kuala Lumpur : Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits-A Literature Review : 2022

[13.] Jhon Alexander Méndez Sayago : Relaciones entre los salarios y la productividad en Colombia : 2017

[14.] León Atao Gladys Milagros : Influencia de los salarios de los docentes sobreel rendimiento

académico : 2021

[15.] Hugo Maul Rivas y Jorge Lavarreda : Análisis de las remuneraciones de losdocentes del sector público en Guatemala : 2008

[16.] Manuel Sánchez Ceron, Francisca María del Sagrario Corte Cruz : La precarización del trabajo. El caso de los maestros de educación básica en América Latina : 2012

[17.] Anélido Tello Díaz : Motivación y Desempeño Laboral en los Docentes dela In- situación Educativa del nivel secundario German Tejada Vela, Moyobamba 2017 : 2018

[18.] Gary McCulloch : Documentary Research Education, History and the SocialSci- ences : 2004

[19.] Aquib javed khan, dr cool drew, john syrinek : Aprendizaje machinelearning : 2017

[20.] Yanming Chen and Xinlong Li : Salary Prediction Based on the Resumes of the Candidates : 2023

[21.] Shiqi Yang : Automated employee salary prediction algorithm based on machine learning : 2023

[22.] Nils J Nilsson : introduction to machine learning an early draft of a proposedtextbook : 1998

[23.] Pornthep Khongchai and Pokpong Songmuang : Random Forest for Salary Predic- tion System to Improve Students' Motivation : 2017

[24.] Amirhosein Jafari and Behzad Rouhanizadeh and Sharareh Kermanshachi and Munahil Murrieum : Predictive Analytics Approach to Evaluate Wage Inequality in Engineering Organizations : 2020.

[25.] Ruksana, MS Sonia : A Review on Classification of Machine Learning : 758-767

[26.] Susmita Ray : A Quick Review of Machine Learning Algorithms

[27.] Rutuja Bhamare Vaishnavi Barve and Tushar Sharama : Salary Prediction using Machine Learning Algorithm : 2023

[28.] Leo Breiman : Random Forests : 2001

[29.] Priya santohosini d: Job salary prediction : 2021

[30.] Samuel Iorhemen Ayua and Yusuf Musa Malgwi and James Afrifa : SalaryPredic- tion Model for Non-Academic Staff Using Polynomial Regression Technique : 2023

[31.] Daniel Hentilä : The link between salary and performance: Are NBA playersover- paid

[32.] U. Bansal and A. Narang and A. Sachdeva and I. Kashyap and S. P. Panda: Empirical analysis of regression techniques by house price and salaryprediction : 2021

[33.] Birol Yildiz and S¸afak A˘gdeniz : a comparative study of machine learningalgo- rithms as an audit tool in financial failure prediction : 2019

[34.] Raghawendra Naik and Pavan N Kunchur : Census Employee SalaryPrediction using Supervised Machine Learning : 2022

[35.] Phuwadol Viroonluecha, Thongchai Kaewkiriya : Salary Predictor Systemfor Thai- land Labour Workforce using Deep Learning : ISCIT 2018.

[36.] Sananda Dutta, Airiddha Halder, Kousik Dasgupta : Design of a novel Prediction Engine for predicting suitable salary for a job : ISCIT 2018.

[37.] Sergio Delgado Quintero : Aprende Python : 2024

[38.] Tomas Gómez rodriguez, Humberto Rios Bolivar, Ali Aali Bujari : Salario eficiente y crecimiento económico para el caso de América Latina : 2017

[39.] Yasser T. Matbouli and Suliman M. Alghamdi : Statistical Machine Learning Re- gression Models for Salary Prediction Featuring Economy Wide Activities and Oc- cupations : 2022

[40.] Ignacio Martín and Andrea Mariello and Roberto Battiti and José Alberto Hernandez : Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study : 2018

[41.] Jong Yih Kuo and Hui Chi Lin and Chien Hung Liu : Building graduate salary grading prediction model based on deep learning : 2021

[42.] Ying Sun and Fuzhen Zhuang and Hengshu Zhu and Qi Zhang and Qing Heand Hui Xiong : Market-oriented job skill valuation with cooperativecomposition neural network : 2021

[43.] Ashish Pawha and Deepali Kamthania : Quantitative analysis of historical data for prediction of job salary in India - A case study : 2019

[44.] Mateo Vargas-Zapata and Marisol Medina-Sierra and Luis Fernando Galeano-Vasco and Mario Fernando Cer´on-Mun˜oz : Algoritmos de aprendizaje de máquina para la predicción de propiedades fisicoquímicas del suelo mediante información espectral: una revisión sistemática : 2022

[45.] Zoila Rosa Vargas Cordero: La investigación aplicada: una forma de conocer las realidades con evidencia científica 2009,

7. ANEXOS

7.1. Evidencia de la Sumisión del articulo en una conferencia paper.

7.1.1. Carta de Acepción de la Revista.

LETTER OF ACCEPTANCE

13th Computer Science On-line Conference 2024

Dear Segundo Canahuire Hilari,

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

Salary prediction with Machine Learning in teachers hired from the Region of Cusco - Perú. (Paper ID: 113122)
Author(s):Canahuire Hilari Segundo,

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 4/28/2024.

Conference organization (sponsored by): OpenPublish.eu

Organization Committee Chair:
Radek Silhavy, Ph.D.

Radek Silhavy, Ph.D.
Organizing Committee Chair

7.1.2. Certificado de presentación en conferencia.

CSOC2024

CERTIFICATE OF PARTICIPATION

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

Awarded to

Canahuire Hilari Segundo

For the Paper presentation:

Salary prediction with Machine Learning in teachers hired from the Region of Cusco - Perú.

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

7.2. 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° 0057-2024/UPeU-FIA-CF-T

Lima, Ñaña 20 de febrero de 2024

VISTO:

El expediente de **Joel Eduardo Larico Carbajal**, identificado(a) con Código Universitario N° 201810469 y **Segundo Canahuire Hilari**, identificado(a) con Código Universitario N° 201122625, 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 aprobación e inscripción del perfil de proyecto de tesis en formato artículo y la designación o nombramiento del asesor para la obtención del título profesional;

Que **Joel Eduardo Larico Carbajal** y **Segundo Canahuire Hilari**, han solicitado: la inscripción del perfil de proyecto de tesis titulado "Modelo de predicción salarial para docentes contratados utilizando técnicas machine learning en la región Cusco" y la designación del Asesor, encargado de orientar y asesorar la ejecución del perfil de proyecto de tesis en formato artículo;

Estando a lo acordado en la sesión del Consejo de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión, celebrada el 20 de febrero de 2024, y en aplicación del Estatuto y el Reglamento General de Investigación de la Universidad;

SE RESUELVE:

Aprobar el perfil de proyecto de tesis en formato artículo titulado **"Modelo de predicción salarial para docentes contratados utilizando técnicas machine learning en la región Cusco"** y disponer su inscripción en el registro correspondiente, designar a **Mg. Ferdinan Edgardo Pineda Anco** como ASESOR para que oriente y asesore la ejecución del perfil de proyecto de tesis en formato artículo el cual fue dictaminado por: **Mg. Nemias Saboya Rios** y **Dr. Juan Jesus Soria Quijaite**, otorgándoles un plazo máximo de doce (12) meses para la ejecución.

Regístrese, comuníquese y archívese.

Dra. Erika Inés Acuña Salinas
DECANA

Mg. Ketty Magaly Arellano Lino
SECRETARIA ACADÉMICA

cc:
-Interesado
-Asesor
-Dirección General de Investigación
-Archivo

7.3. Carta de consentimiento de la UGEL Canas para el uso de la Información.

"Decenio de la Igualdad de oportunidades para mujeres y hombres"
"Año del fortalecimiento de la soberanía nacional"

Cusco, 29 de Diciembre del 2023

CARTA N° 204 -2023-GR-C/GEREDU-C/D-UGEL-C

SEÑOR(A):

DR. Walter Murillo Antón

RECTOR DE UNIVERSIDAD PERUANA UNIÓN

PRESENTE.:

ASUNTO : Se autoriza para hacer uso de información de planillas.

REFERENCIA : **EXPEDIENTE N° 12029-2023.**

Memorandun a la oficina de Informatica.

De mi consideración

Es grato dirigirme a usted, para manifestarle que se ha recepcionado el documento de la referencia, mediante el cual el administrado Bach. Segundo CANAHUIRE HILARI, solicita realizar trabajo de investigación con datos de planillas y NEXUs de la Unidad de Gestión Educativa Local Canas, al respecto previa evaluación del mismo por parte de la Administradora de esta sede administrativa; da la conformidad para utilizar los datos en mención.

Por lo expuesto en el párrafo anterior, mi despacho comunica la autorización para el uso de estos datos de planilla de los años 2021, 2022 y 2023; aclarando que es exclusivamente para el proyecto de investigación del Bach. Segundo CANAYHUIRE HILARI; lo que pongo en su conocimiento para los tramites administrativos correspondientes.

Atentamente

GOBIERNO REGIONAL CUSCO
Dra. Gladis Nancy Huaraz Guzmán
DIRECTORA DE LA UGEL - CANAS