NCML NeutrosophicComputing and Machine Learning, Vol. 42, 2026

University of New Mexico

Análisis predictivo de salarios de desarrolladores latinoamericanos con encuesta Stack Overflow 2025
Predictive analysis of Latin American developer salaries using the Stack Overflow 2025 survey

Ing. Jonathan Kenny Vera Macias[1], MSc. Miguel Angel Quiroz Martinez[2], and MSc. Dayron Rumbaut Rangel[3]

[1]Universidad Bolivariana del Ecuador, Km 5 ½ vía Durán Yaguachi, Ecuador, jkveram@ube.edu.ec

[2]Universidad Bolivariana del Ecuador, Km 5 ½ vía Durán Yaguachi, Ecuador, maquirozm@ube.edu.ec,
[3]Universidad Bolivariana del Ecuador, Km 5 ½ vía Durán Yaguachi, Ecuador, drumbautr@ube.edu.ec

Resumen

El crecimiento del sector tecnológico en América Latina ha incrementado la demanda de desarrolladores de software, generando un mercado laboral dinámico caracterizado por una alta variabilidad salarial. Estas diferencias están influenciadas por múltiples factores, como la experiencia profesional, el nivel educativo, el país de residencia y las tecnologías dominadas, lo que dificulta su análisis mediante enfoques tradicionales. En este contexto, la analítica predictiva aplicada a los recursos humanos surge como una alternativa eficaz para modelar relaciones complejas y no lineales presentes en grandes volúmenes de datos. El presente estudio tiene como objetivo predecir los salarios de desarrolladores de software latinoamericanos a partir de datos provenientes de la encuesta Stack Overflow 2025. Para ello, se comparan dos enfoques de modelado: la Regresión Lineal, utilizada como modelo base de carácter explicativo, y el Bosque Aleatorio, seleccionado por su capacidad para manejar datos heterogéneos y capturar patrones no lineales. Los resultados permiten evaluar el desempeño predictivo de ambos modelos y aportan evidencia empírica sobre la utilidad de técnicas de aprendizaje automático en el análisis salarial del sector tecnológico en la región, sentando las bases para futuras investigaciones con modelos más avanzados.

Palabras clave: Analítica predictiva, salarios, desarrolladores de software, aprendizaje automático, América Latina, Stack Overflow.

Abstract

The growth of the technology sector in Latin America has increased the demand for software developers, generating a dynamic labor market characterized by high salary variability. These differences are influenced by multiple factors, such as professional experience, educational level, country of residence, and technologies mastered, making them difficult to analyze using traditional approaches. In this context, predictive analytics applied to human resources emerges as an effective alternative for modeling complex and nonlinear relationships present in large volumes of data. This study aims to predict the salaries of Latin American software developers using data from the 2025 Stack Overflow survey. To this end, two modeling approaches are compared: Linear Regression, used as the base explanatory model, and Random Forest, selected for its ability to handle heterogeneous data and capture nonlinear patterns. The results allow for the evaluation of the predictive performance of both models and provide empirical evidence on

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

the usefulness of machine learning techniques in salary analysis within the technology sector in the region, laying the groundwork for future research with more advanced models.

Keywords: Predictive analytics, salaries, software developers, machine learning, Latin America, Stack Overflow.

1 Introducción

El crecimiento de la industria tecnológica en América Latina ha impulsado una demanda sostenida de desarrolladores de software. Esta tendencia, motivada por la digitalización de procesos y la adopción del trabajo remoto, ha transformado el mercado laboral en uno altamente competitivo y heterogéneo [1, 24].

En este escenario, los salarios de los desarrolladores presentan una gran variabilidad, determinada por factores como el país de residencia, la experiencia, el nivel educativo y las tecnologías dominadas [2]. Comprender estas diferencias resulta crucial tanto para las organizaciones que buscan retener talento como para los profesionales que planifican su desarrollo laboral. Sin embargo, la literatura actual carece de estudios que aborden específicamente qué vacíos dejan los modelos tradicionales al explicar esta variabilidad en el contexto latinoamericano.

La analítica predictiva aplicada al ámbito de los recursos humanos permite identificar patrones y estimar resultados con base en grandes volúmenes de datos. A diferencia de los métodos estadísticos tradicionales, los enfoques de aprendizaje automático capturan relaciones no lineales y manejan variables de alta dimensionalidad [3, 21, 15].

Surge así la siguiente pregunta de investigación: *¿Qué modelo ofrece mejor capacidad predictiva de los salarios de desarrolladores latinoamericanos, considerando datos de la encuesta Stack Overflow 2025?*

Este estudio busca aplicar un análisis predictivo de los salarios de desarrolladores en la región utilizando dichos datos. Se comparan dos modelos representativos: la Regresión Lineal, como modelo base explicativo, y el Bosque Aleatorio, reconocido por su robustez frente a datos heterogéneos y no lineales [23]. Además, se discute la pertinencia de modelos avanzados (p. ej., boosting y aprendizaje profundo) como línea futura de investigación [17, 14].

2 Estudios previos

2.1 Modelos predictivos salariales

La literatura reciente muestra el uso de técnicas de aprendizaje automático para predecir salarios en diferentes contextos. Shao et al. [4] analizaron los ingresos de profesionales de TI en China y demostraron que los algoritmos basados en árboles superan a los modelos lineales en precisión. De manera similar, Alowolodu et al. [5] aplicaron estrategias de boosting en mercados emergentes, logrando mejoras significativas en la estimación de ingresos.

Otros trabajos como Kuo y Chien [6], Liu y Zhang [7] y Wang et al. [23] destacan la eficacia de los enfoques de ensamble frente a la dispersión salarial. En la frontera aplicada, Kim y Alvarez [20] exploran métodos multimodales para el pronóstico salarial, mientras que Ng et al. [22] enfatizan la interpretabilidad de los modelos para compensaciones.

En base a lo anterior, cabe agregar que, aunque estos estudios muestran la eficacia de modelos basados en árboles, pocos han explorado su comportamiento con datos específicamente latinoamericanos y encuestas abiertas como Stack Overflow. En conjunto, estos trabajos evidencian la creciente adopción de modelos de aprendizaje automático en la predicción salarial, aunque la mayoría

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

se ha centrado en contextos asiáticos o de países desarrollados.

2.2 Encuestas tecnológicas

Las encuestas de Stack Overflow son una fuente consolidada para el análisis de tendencias laborales y tecnológicas. Studien et al. [8] examinaron las respuestas de miles de desarrolladores para identificar relaciones entre habilidades técnicas y diferencias salariales. Aunque la muestra no es plenamente representativa, su periodicidad anual y cobertura global la convierten en un insumo valioso [24, 9].

2.3 Machine Learning en Recursos Humanos

El aprendizaje automático se ha consolidado en el ámbito de la analítica de recursos humanos. Saini et al. [3] sintetizaron técnicas de ML aplicadas a RR. HH., incluyendo compensaciones. En América Latina, Rivas et al. [10] aplicaron modelos predictivos a datos de empresas locales, confirmando su utilidad en la gestión de talento. Asimismo, la literatura reciente destaca la importancia de la transparencia y la equidad algorítmica en este dominio [12, 25].

3 Metodología

Diseño del estudio: se trató de una investigación *cuantitativa, no experimental, predictiva y de corte transversal*.

Población y muestra: se emplearon 49 128 registros de la encuesta Stack Overflow 2025; tras filtrar por residencia en países latinoamericanos, se analizaron 2 027 respuestas válidas correspondientes a desarrolladores activos.

Variables consideradas: se analizaron 19 campos relevantes:

- Age: edad del encuestado (demográfica).
- Country: país de residencia (contexto geográfico y macroeconómico).
- EdLevel: nivel educativo alcanzado (capital humano formal).
- WorkExp y YearsCode: años de experiencia laboral y programando, respectivamente (trayectoria).
- Employment y EmploymentAddl: tipo de contrato principal y actividades adicionales.
- DevType: rol principal (backend, frontend, fullstack, etc.).
- OrgSize: tamaño de la organización.
- ICorPM: contribuidor individual o gerente.
- RemoteWork: modalidad remota/presencial.
- Industry: sector económico.
- MainBranch: rama principal de actividad profesional.
- JobSat: satisfacción laboral.
- LanguageHaveWorkedWith, DatabaseHaveWorkedWith, PlatformHaveWorkedWith, WebframeHaveWorkedWith, DevEnvsHaveWorkedWith: tecnologías declaradas (capital específico).
- Currency: tipo de moneda reportada (control monetario cuando disponible).

Entorno de ejecución y herramientas: el procesamiento y entrenamiento se realizaron en **Python 3.12.3**, utilizando pandas, numpy, matplotlib, seaborn y scikit-learn [16, 13]. El experimento se ejecutó en un equipo con procesador **AMD Ryzen 7 5700G** (3.8 GHz), 16 GB de RAM, tarjeta **NVIDIA GeForce**

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

RTX 4060 y **Windows 11 64 bits**. Se siguieron buenas prácticas de ingeniería de características, así como validación [19, 15]; para referencia sobre ecosistemas de ML se consideran *XGBoost* y *TensorFlow* [17, 18]. Para garantizar la reproducibilidad, se fijó una semilla aleatoria (`random_state=42`) en todos los procesos estocásticos.

3.1 Adquisición y preprocesamiento de datos

Se descargaron los microdatos de la encuesta Stack Overflow 2025 y se filtraron las respuestas de residentes en América Latina. Los datos fueron depurados mediante imputación (mediana para numéricas, moda para categóricas), codificación *One-Hot* para variables textuales y recorte por cuantiles (1–99 %) para atenuar la influencia de valores extremos [19].

3.2 Modelado predictivo

Para evaluar la capacidad de generalización y la estabilidad de los modelos, se utilizó una estrategia de **validación cruzada de 5 pliegues (5-fold cross-validation)** sobre el conjunto total de datos filtrados. Posteriormente, se reservó un 20 % de los datos para la generación de gráficos de diagnóstico finales. Se compararon dos algoritmos representativos:

- **Regresión Lineal**, como modelo base explicativo (sesgos/varianzas bien caracterizados) [15].
- **Bosque Aleatorio (Random Forest)**, como modelo no lineal robusto a heterogeneidad [23].

El rendimiento se reporta mediante la media y la desviación estándar de MAE, RMSE y R^2. Los flujos se construyeron con `Pipeline` y `ColumnTransformer` para integrar preparación y modelado [16]. La discusión incorpora criterios de interpretabilidad y transparencia [22, 25].

4 Resultados

Los modelos fueron evaluados en términos de precisión y error mediante validación cruzada robusta. El Bosque Aleatorio alcanzó un MAE promedio de **29,797 USD** (± 7,441) y un RMSE de **123,191 USD**, superando ampliamente a la Regresión Lineal, cuyo MAE medio fue de 133,200 USD con alta variabilidad.

El coeficiente de determinación (R^2) obtenido fue de **0.086** para el Bosque Aleatorio. Aunque bajo, este valor es significativamente superior al modelo lineal (que presentó valores negativos, indicando un ajuste peor que una simple media horizontal) [23]. La alta desviación estándar observada en el RMSE (± 125,483 para el Random Forest) sugiere la presencia de valores atípicos extremos que afectan la estabilidad de las predicciones cuadráticas.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

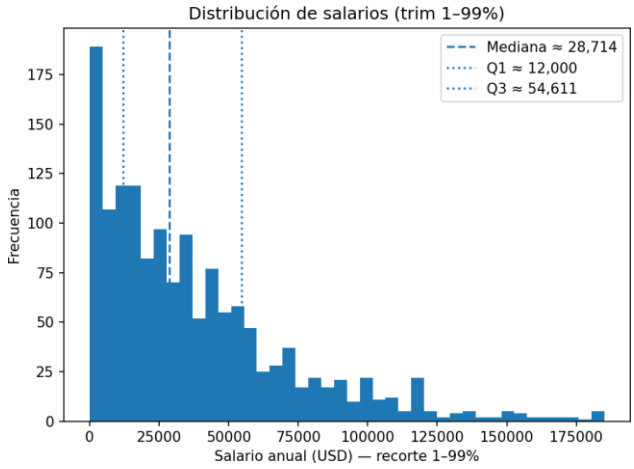

Fig. 1: Distribución de salarios (recorte 1–99%).

Análisis. La distribución presenta marcada asimetría positiva: la masa principal se concentra entre 10 000 y 60 000 USD, con cola derecha larga. El recorte 1–99 % redujo la influencia de extremos, estabilizando mediana e IQR.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

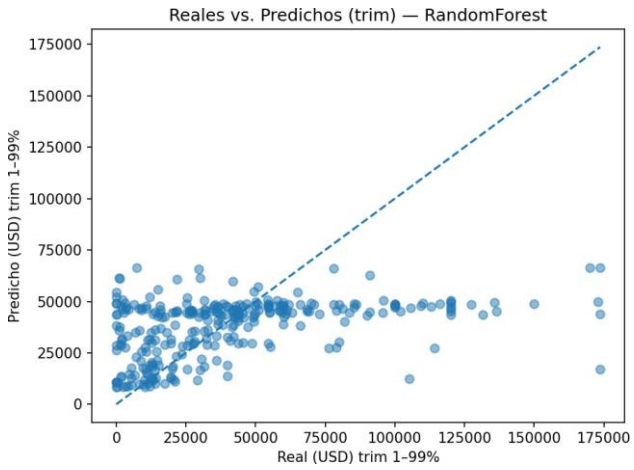

Fig. 2: Valores reales vs. predichos (RandomForest, recorte 1–99%).

Análisis. La nube de puntos se alinea parcialmente con la diagonal, lo que indica una correlación moderada entre valores reales y predichos. Se observa subestimación en salarios altos, consistente con colas pesadas.

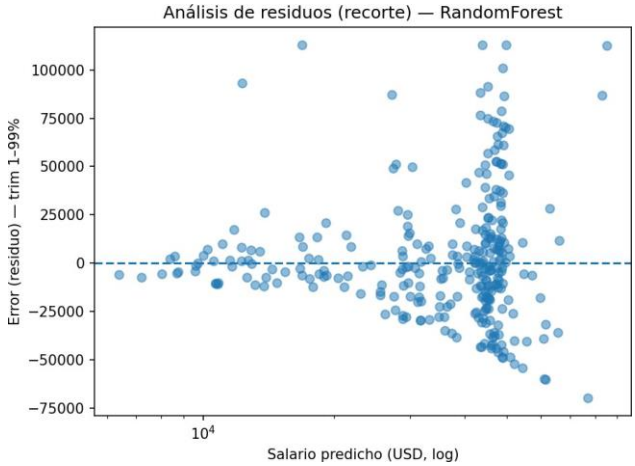

Fig. 3: Análisis de residuos del modelo RandomForest.

Análisis. Los residuos se distribuyen alrededor de cero sin patrones fuertes. La dispersión crece con el salario predicho (heterocedasticidad esperada).

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

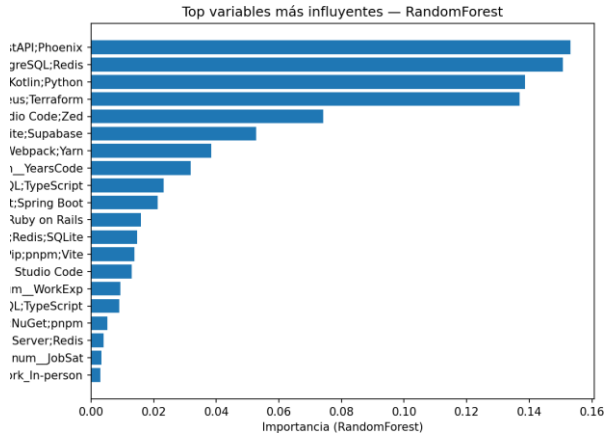

Fig. 4: Importancia de variables en el modelo RandomForest.

Análisis. Destacan experiencia (YearsCode/WorkExp), nivel educativo (EdLevel) y ciertos stacks tecnológicos. La lectura se alinea con literatura en compensaciones [22].

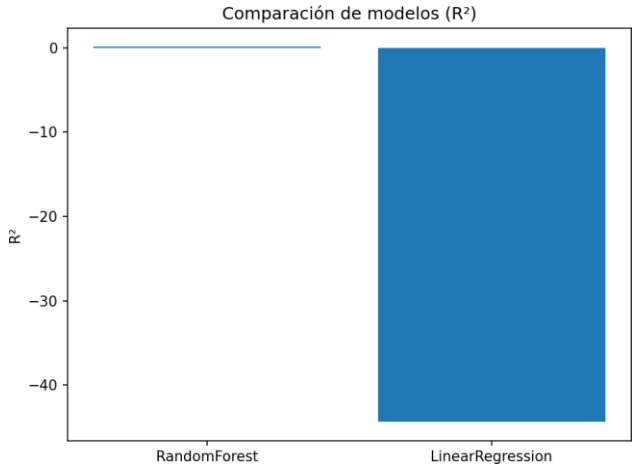

Fig. 5: Comparación de modelos (R^2).

Análisis. El Random Forest obtuvo un R^2 superior, confirmando que modelos no lineales capturan mejor la variabilidad del salario, aunque el poder explicativo global es bajo [21].

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

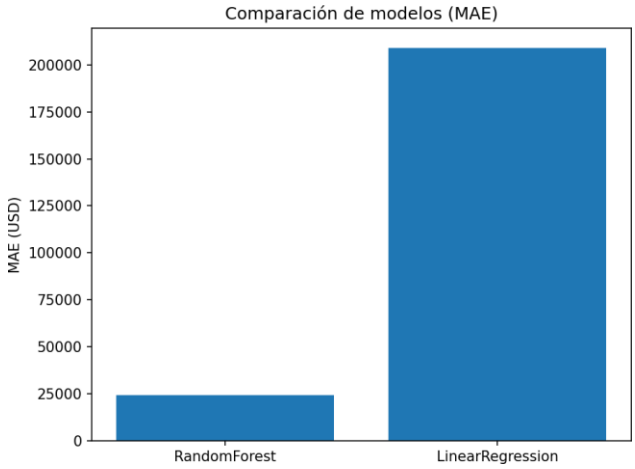

Fig. 6: Comparación de modelos (MAE).

Análisis. El MAE del Bosque Aleatorio (≈29 797 USD) fue mucho menor que el de la Regresión Lineal, reflejando mayor estabilidad ante atípicos [6].

Table 1: Comparación de desempeño entre modelos (Validación Cruzada 5-folds).

Modelo	MAE (USD)	RMSE (USD)	R^2
Bosque Aleatorio	29,797 ± 7,441	123,191 ± 125,483	0.086 ± 0.07
Regresión Lineal	133,200 ± 72,318	207,536 ± 148,549	-3.84 ± 3.08

Interpretación. Se reportan la media y la desviación estándar (±) de las métricas. El Bosque Aleatorio presenta un error medio notablemente menor y mayor estabilidad que la Regresión Lineal. Sin embargo, la alta desviación estándar en RMSE y el R^2 bajo (≈ 0.09) sugieren que variables omitidas (sector, tamaño, políticas) juegan un rol crucial, alentando a explorar enfoques más avanzados [17, 20].

5 Conclusiones

El estudio demuestra que los modelos no lineales, como el Bosque Aleatorio, ofrecen mejor desempeño que los modelos lineales para la predicción de salarios de desarrolladores en América Latina. Sin embargo, el bajo valor del coeficiente de determinación ($R^2 \approx 0.09$) indica que la predicción salarial depende de factores adicionales no capturados por la encuesta Stack Overflow 2025 [21].

Aunque el Bosque Aleatorio mostró mejor desempeño, el bajo R^2 evidencia que los factores salariales dependen de variables contextuales no contempladas en la encuesta. Por tanto, la capacidad predictiva actual es limitada.

Futuras investigaciones deberían incorporar variables macroeconómicas y organizacionales, así como explorar modelos más avanzados (p. ej., Gradient Boosting/XGBoost y aprendizaje profundo) y evaluar su interpretabilidad y equidad en RR. HH. [17, 14, 22, 25, 23, 20]. Estos hallazgos evidencian el potencial del aprendizaje automático para comprender las dinámicas salariales regionales y

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

abren la puerta a estudios que integren factores económicos y sociales en la predicción del talento tecnológico latinoamericano.

En conjunto, los resultados obtenidos evidencian que, si bien los modelos de aprendizaje automático no lineales mejoran el desempeño predictivo frente a enfoques lineales, persiste un alto nivel de incertidumbre estructural en la predicción salarial de desarrolladores latinoamericanos. Esta incertidumbre no responde únicamente a limitaciones técnicas del modelo, sino a la naturaleza compleja, contextual y heterogénea del fenómeno analizado. Desde esta perspectiva, la adopción de enfoques neutrosóficos se perfila como una alternativa prometedora para representar explícitamente componentes de verdad, falsedad e indeterminación en los procesos predictivos, complementando los modelos clásicos de machine learning y contribuyendo a una interpretación más realista y responsable de los resultados [26].

References

[1]. Zhang, L., Chen, J., Li, Y.: Advances in applied machine learning for workforce analytics. *ACM Computing Surveys* 56(4), 1–36 (2023). https://doi.org/10.1145/3587929

[2]. Choudhury, R., Singh, A.: Predictive analytics in human resources: A systematic review. *Journal of Business Research* 153, 280–295 (2022). https://doi.org/10.1016/j.jbusres.2022.08.015

[3]. Saini, R., Kumar, P., Gupta, V.: Machine learning techniques in hu- man resource analytics: A review. *Expert Systems* 40(5), e13309 (2023). https://doi.org/10.1111/exsy.13309

[4]. Shao, W., Li, J., Xu, H.: Salary prediction using machine learning algorithms. *Ap- plied Intelligence* 52, 9123–9138 (2022). https://doi.org/10.1007/s10489-022-03021-
i. 3

[5]. Alowolodu, O., Adepoju, A., Oyedele, S.: Machine learning models for income prediction in emerging markets. *Heliyon* 9(6), e18345 (2023). https://doi.org/10.1016/j.heliyon.2023.e18345

[6]. Kuo, J., Chien, C.: Predictive modeling of wages using machine learning. *Expert Systems with Applications* 227, 120320 (2023). https://doi.org/10.1016/j.eswa.2023.120320

[7]. Liu, Q., Zhang, Y.: Data-driven salary estimation via regression en- sembles. *Information Processing and Management* 59(5), 102999 (2022). https://doi.org/10.1016/j.ipm.2022.102999

[8]. Studien, F., Müller, P., Hoffmann, T.: Trends in developer technologies and com- pensation: An analysis of Stack Overflow surveys. *Information Systems Frontiers* (2023). https://doi.org/10.1007/s10796-023-10345-1

[9]. Raza, S., Kim, D., Lee, J.: Cross-country salary prediction using ensemble learning. *Journal of Computational Social Science* (2023). https://doi.org/10.1007/s42001- 023-00206-0

[10]. Rivas, M., Gómez, A., Torres, L.: Predictive analytics for HR management in Latin America: A case study. *Revista Iberoamericana de Tecnología* 29(1), 77–95 (2022).

[11]. Mishra, S., Banerjee, A., Singh, R.: Applications of data mining and machine learning in HR analytics. *IEEE Access* 10, 34500–34515 (2022). https://doi.org/10.1109/ACCESS.2022.3163459

[12]. Zhang, H., Liu, W., Wang, Z.: Responsible machine learning in HR an- alytics: Transparency and bias mitigation. *AI Ethics* 4(2), 230–243 (2023). https://doi.org/10.1007/s43681-022-00243-1

[13]. Géron, A.: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.*
i. O'Reilly Media, 3rd Edition (2023).

[14]. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016).

[15]. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

i. Springer (2009).

[16]. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011).

[17]. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD*, 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[18]. Abadi, M., Agarwal, A., Barham, P., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2016). https://www.tensorflow.org

[19]. Kuhn, M., Johnson, K.: *Feature Engineering and Selection for Predictive Modeling*.

i. CRC Press (2019).

[20]. Kim, D., Alvarez, C.: AI-driven wage forecasting using multimodal learning models. *Computers in Human Behavior Reports* 9, 100358 (2024). https://doi.org/10.1016/j.chbr.2024.100358

[21]. Shmueli, G., Koppius, O.: Predictive analytics in information systems research: Review and recommendations. *MIS Quarterly Executive* 21(1), 45–63 (2022). https://doi.org/10.25300/MISQ/2022/16542

[22]. Ng, K., Zhou, Y., Patel, T.: Interpretable AI models for employee compensation prediction. *Decision Support Systems* 165, 113950 (2023). https://doi.org/10.1016/j.dss.2023.113950

[23]. Wang, Y., Lin, F., Chen, J.: Comparative study of machine learning algorithms for wage prediction. *Applied Soft Computing* 142, 110372 (2023). https://doi.org/10.1016/j.asoc.2023.110372

[24]. Torres, M., Pereira, A.: Big Data analytics for labor market insights in Latin America. *Information Systems Frontiers* 26(2), 455–470 (2024). https://doi.org/10.1007/s10796-024-10421-9

[25]. Soto, R., Alvarez, P., Muñoz, G.: Evaluating the fairness of machine learning models in HR applications. *AI and Ethics* 5(3), 331–346 (2024). https://doi.org/10.1007/s43681-024-00283-9

[26]. Vázquez, M. L., & Smarandache, F. (2026). La integración de perspectivas contradictorias en la filosofía latinoamericana: Un enfoque MultiAlisista. Neutrosophic Computing and Machine Learning. ISSN 2574-1101, 41, 213-220.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/