Universidad Autónoma de Querétaro

Facultad de Ingeniería

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

25 de agosto de 2023

Autor: *Ing. Carlos Manuel Sánchez Martínez*

Machine Learning (curso), Dr. Marco Antonio Aceves Fernández

Practica 1: Adquisición de datos

Resumen

En esta práctica, abordaremos la adquisición y exploración de una base de datos que contiene información salarial de desarrolladores de software. Utilizaremos una base de datos pública que nos brindará una visión completa de los variados sueldos en distintos roles laborales, y cómo estos sueldos se vinculan con el índice de popularidad de las empresas. A través de técnicas de estadística descriptiva, examinaremos detenidamente esta base de datos, encontrando patrones y desafíos que puedan surgir. Dentro de este análisis, también exploraremos la viabilidad de enriquecer la base de datos mediante la incorporación de columnas adicionales que aporten información crucial que el autor original omitió. Las características que se encuentren serán esenciales para una toma de decisiones en futuros entrenamientos de aprendizaje profundo que se realizaran en las siguientes prácticas.

Palabras clave: Exploración de Datos, Visualización, Distribución, Estadística Descriptiva.

Introducción

En el mundo de las empresas de software, la exploración y el análisis de datos se convierten en herramientas fundamentales en la extracción de características y patrones valiosos, con los cuales es posible hacer selecciones estratégicas que nos ayuden a facilitar las comparaciones entre distintas series de datos. En este proyecto abordaremos temas de estadística y análisis de datos para indagar sobre un conjunto de datos con el cual nos adentraremos a encontrar posibles patrones en los sueldos de los desarrolladores en la actualidad. Esto nos puede abrir un panorama para conocer nuevas áreas de oportunidad en el mundo.

Marco teórico

La historia de la estadística tiene raíces muy antiguas, evidenciadas por registros de recolección de datos sobre población, recursos y producción en civilizaciones como la china (aproximadamente 1000 años a.C.), sumeria y egipcia. Incluso en el libro de Números de la Biblia, se hacen menciones sobre el conteo de israelitas en edad de servicio militar. Un ejemplo notable es el censo que llevó a José y María a Belén, según el Evangelio. Los censos, como institución, ya existían en el siglo IV a.C. en el Imperio Romano [2].

KDD (Knowledge Discovery in Databases), conocido como el proceso completo de extracción informativa, que no solo involucra la preparación de datos, sino también la interpretación de los resultados [3]. Apareció por primera vez a finales del siglo XX, con el cual, se trataba de interpretar grandes cantidades de datos y encontrar relaciones o patrones, esto debido a la gran cantidad de información que se fue generando al paso de los años. Actualmente, el incremento de la información se ve reflejada con el surgir de nuevos sistemas de información capaces de unir un número sin fin de dispositivos por medio de internet. Según un informe de IBM [4], el 90 % de los datos disponibles del mundo se han creado en los dos últimos años.

El análisis de datos multivariantes se dedica al estudio estadístico de varias variables medidas en elementos de una población, con el propósito de abordar la síntesis del conjunto de variables en un número reducido de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información, identificar patrones en los datos, como grupos o estructuras, en caso de que estén presentes, clasificar nuevas observaciones en grupos predefinidos, permitiendo una asignación precisa y explorar y analizar las relaciones entre dos conjuntos de variables, identificando posibles vínculos o dependencias [9].

La teoría de la evolución del software, propuesta por Meir Lehman en la década de 1970, ofrece una perspectiva conceptual que describe cómo los sistemas de software evolucionan a lo largo del tiempo. Esta teoría se basa en un conjunto de leyes que resaltan patrones recurrentes y fenómenos observables en el proceso de desarrollo, mantenimiento y adaptación del software en respuesta a las necesidades cambiantes del entorno. Estas leyes no solo proporcionan una comprensión más profunda de la dinámica de la evolución del software, sino que también tienen implicaciones significativas para la gestión y la ingeniería del software [8].

Materiales y métodos

1. Base de Datos: En esta práctica, se utilizó una base de datos pública, la cual contiene información de más de, 22700 profesionales de software con diferentes características, como sus salarios (rupias), puestos de trabajo, nombre de la empresa, calificación de la empresa, número de veces que se informan los salarios y ubicación de la empresa.

   Enlace: https://www.kaggle.com/datasets /whenamancodes/software-professional-salary-dataset

2. Entorno de Ejecución: Con la ayuda de herramientas de desarrollo como Python, utilizamos una de sus muchas librerías, Jupyter Notebooks, el cual nos ayudó en el desarrollar en un ambiente computacional, código interactivo, el cual es ejecutado en aplicaciones web. Para la lectura de datos utilizamos Pandas, la cual es una librería de programación en Python que proporciona herramientas para el análisis y la manipulación de datos. Y por último se utilizaron 2 librerías como medio

para representar los datos de manera gráfica, seaborn y matplotlib, las cuales están enfocadas al 100 % en la visualización de datos estadísticos y nos han ayudado en crear visualizaciones atractivas e informativas.

3. Métodos Empleados:
   - Valor Mínimo: Se inicializa una variable con la primera posición de la columna deseada, son recorridos cada uno de los valores hasta encontrar un valor menor al inicial, en caso de encontrar alguno, se remplaza la variable previamente inicializada para después sustituir su valor con el nuevo encontrado.
   - Valor Máximo: Se inicializa una variable con la primera posición de la columna deseada, son recorridos cada uno de los valores hasta encontrar un valor mayor al inicial, en caso de encontrar alguno, se remplaza la variable previamente inicializada para después sustituir su valor con el nuevo encontrado.
   - Media: Se recorre cada uno de los valores en la columna, estos serán sumados y se va incrementando un contador para calcular la cantidad de números iterados, al final dividimos la suma total de los elementos entre la cantidad de número iterados [Moore].
   - Valor Desviación Estándar: Con el uso del valor de la media previamente calculada, es necesario ir restándola a las a iteraciones de la columna deseada y luego esta debe ser elevada al cuadrado. Posteriormente, es necesario sumar todas las diferencias al cuadrado calculadas en el paso anterior y dividirlas entre la cantidad total de números y por último es necesario calcular la raíz cuadrada del valor obtenido [7].
   - Datos Faltantes: Recorrer la columna a evaluar y se debe contar la cantidad de valores faltantes, como None.
   - Datos Atípicos: Con el uso de la desviación estándar, definimos algún criterio para identificar los datos atípicos, ejemplo: ± 2 Esto significará que los valores que se encuentran más allá de 2 veces la desviación estándar respecto a la media serán considerados atípicos. Es necesario recorrer cada valor de la columna y comparar si está fuera de este criterio [1].
   - Histograma: Primero es necesario dividir el rango de los datos en segmentos o intervalos, colocamos los datos en los intervalos propuestos y contamos la cantidad de veces en el que aparece el número en los intervalos seleccionados [6].

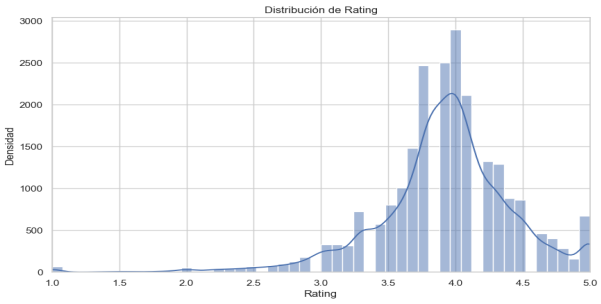
Resultados

Con el uso de las herramientas de software y metodologías para el análisis de datos, se obtuvieron los siguientes resultados después de realizar una estadística descriptiva a la base de datos propuesta.

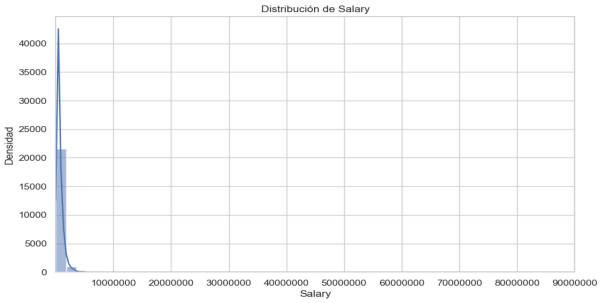
	Media	Moda	Mínimo	Máximo	Desviación Estándar	Datos Faltantes
Rating	3.918213	4.0	1.0	5.0	0.519675	0
Salary	695387.211243	300000.0	2112.0	90000000.0	884399.013676	0
Salaries Reported	1.855775	1.0	1.0	361.0	6.823668	0

- Imagen 1. Estadística descriptiva. Conteo de valores nulos, máximo, mínimo, media y desviación estándar.

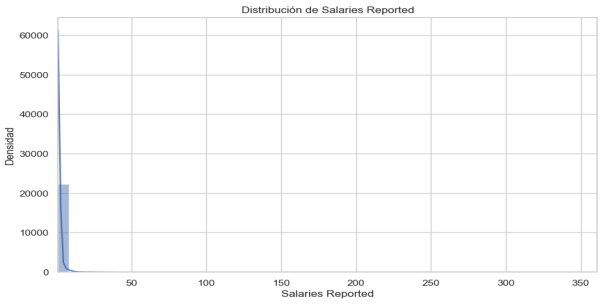
Una vez que conocemos los valores máximos y mínimos, podemos ajustar las escalas de las distribuciones siguientes.



- Imagen 2. Gráfica de Distribución. Densidad vs. Rating, se muestra la estimación de la función de densidad de probabilidad de la variable continua.

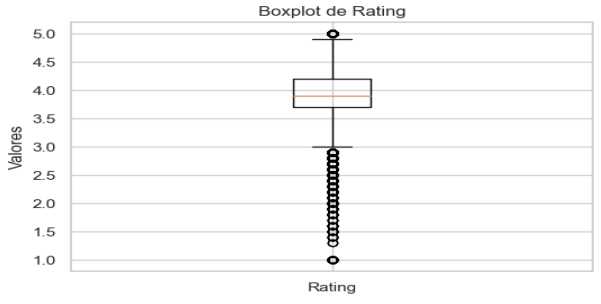


- Imagen 3. Gráfica de Distribución. Densidad vs. Salario, cambiando la notación científica en el eje x.

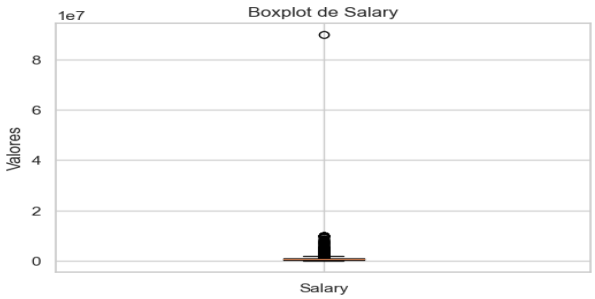


- Imagen 4. Gráfica de Distribución. Densidad vs. Salarios Reportados

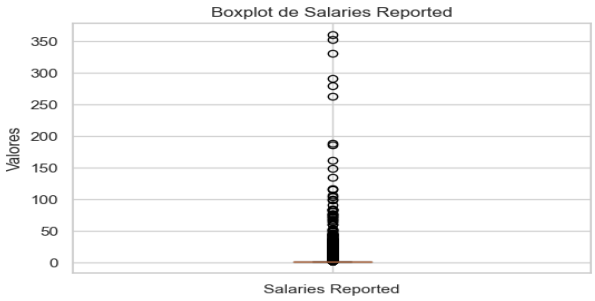
Una vez concluidas las gráficas de distribución daremos el uso de boxplots para mostrar información valiosa sobre la mediana, el rango intercuartil, los valores atípicos y la simetría de los datos.



- Imagen 5. Boxplot de Rating. Gráfica descriptiva sobre la distribución de datos.



- Imagen 6. Boxplot de Salario. Identificación de valores atípicos.



- Imagen 7. Boxplot de Salarios Reportados. Búsqueda de valores asimétricos.

Discusión

Con la implementación de la metodología propuesta para la indagación de los datos propuestos, se encontró que la distribución del rating (Imagen 2) nos indica la fuerza de la relación entre las variables. Se puede apreciar una tendencia a tener forma de campana (simetría), pero esta es ligeramente cargada a los ratings más altos. Por el contrario, tanto las figuras de salarios y salarios reportados (Imágenes 4,5), tuvieron una tendencia de mayor dispersión, a simple vista no se puede observar una buena distribución, se estuvieron modificando las escalas y la cantidad de barras en el histograma para una mejoría en la calidad de la imagen, pero al realizar los boxplot (Imagenes 5,6), se pueden notar que existen outlayer, los cuales le dieron sentido al porqué las gráficas del histograma se comportaban de esa manera, por otro lado, el boxplot que correspondía al rating, se puede apreciar su distribución de manera más cercana a la mediana, pero aun así siguen existiendo outlayer en cada uno de nuestros atributos. Esta práctica ayudó para elaborar una estrategia más concreta, en la siguiente etapa de imputación de datos.

Como tarea de buscar una mejoría a la calidad de datos propuestos por el autor de la base de datos, se proponen agregar los siguientes atributos:

- (idiomas) inglés, español, chino, etc..
- (fecha inicio) entrada a la empresa
- (fecha fin) salida de la empresa, si existe
- (motivo de salida) renuncia, despido, factores externos, etc
- (habilidades) bases de datos, estadística, programación, contabilidad, finanzas, etc
- (números de trabajos) Cantidad de puestos de trabajo por lo que ha pasado el ingeniero
- (número de proyectos) Cantidad de proyectos elaborados a lo largo de su carrera
- (años de experiencia)
- (certificaciones obtenidas) número de certificaciones dentro de la empresa
- (tamaño de la empresa) pequeña, mediana, grande
- (edad)

  Estos datos son propuestos con el objetivo de buscar el rating de la empresa dependiendo de las características que se dan a destacar en el área de informática, tanto en entrevistas por parte de recursos humanos y experiencias de empleo de egresados de la facultad de informática UAQ.

Conclusiones

El análisis de datos en sueldos de software ofrece una visión valiosa de las tendencias y patrones salariales en la industria. Con la aplicación correcta de la información, se pueden guiar futuros análisis, estrategias de recursos humanos, para identificar tendencias emergentes en la industria, como qué habilidades o tecnologías están en demanda y cómo esto se refleja en los sueldos en el área en la que estoy especializándome. En lo personal, me permitirá trazar una trayectoria de carrera más informada y establecer metas para aumentar mis ingresos a lo largo del tiempo.

Referencias

[1] BARNETT, Vic, et al. Outliers in statistical data. New York: Wiley, 1994.

[2] BATANERO, Carmen; GODINO, Juan. Análisis de datos y su didáctica. Departamento de Didáctica de la Matemática de la Universidad de Granada, 2001.

[3] GARCÍA, Jesús, et al. Ciencia de datos. Técnicas Analíticas y Aprendizaje Estadístico. Bogotá, Colombia. Publicaciones Altaria, SL, 2018.

[4], http://www.ibm.com/analytics/us/en/technology/data-science/, 2016

[5] HUBERMAN, A. Michael; MILES, Matthew B. Métodos para el manejo y el análisis de datos. Denman CA, Haro JA (comp.). Por los rincones. Antología de métodos cualitativos en la investigación social. Hermosillo: El Colegio de Sonora, 2000, p. 253-300.

[6] MENDENHALL, William; BEAVER, Robert J.; BEAVER, Barbara M. Introduction to probability and statistics. Cengage Learning, 2012.

[7] MOORE, David S. Introduction to the Practice of Statistics. WH Freeman and company, 2009.

[8] LEHMAN, Meir; FERNÁANDEZ-RAMIL, Juan C. Software evolution. Software evolution and feedback: Theory and practice, 2006, p. 7-40.

[9] PEÑA, Daniel, et al. Análisis de datos multivariantes. Cambridge: McGraw-Hill España, 2013.