Ciencia de datos: salarios bajo la lupa *

Daniela Alcázar *Universidad de Antioquia*
Esteffany Peña *Universidad de Antioquia*
Juliana Rueda *Universidad de Antioquia*

Resumen: La ciencia de datos es un campo de estudio que ha ganado importancia en los últimos años. Los salarios de los profesionales en esta área dependen de factores inherentes al puesto de trabajo. Este estudio presenta una aplicación Shiny desarrollada en R, que permite a los usuarios explorar la relación entre algunas variables de los profesionales, como el nivel de experiencia y la residencia, y de la empresa donde trabajan, como el tamaño y ubicación, y sus salarios. La aplicación presenta la visualización y el análisis descriptivo de los datos, así como la posibilidad de que el usuario prediga un salario a partir de la configuración de las demás variables. La herramienta puede ser utilizada para la toma de decisiones a nivel personal, permitiendo que estudiantes y profesionales interesados en ciencia de datos planifiquen su carrera o tomen decisiones sobre su desarrollo profesional, y a nivel empresarial, para que empresas puedan determinar el salario adecuado para un puesto de trabajo.

Keywords: Salarios, Ciencia de datos, Análisis descriptivo, Análisis predictivo

1. Introducción

La ciencia de datos es definida como un campo interdisciplinario que emplea técnicas, procesos, cálculos y algoritmos para extraer conocimiento de información masiva, ya sea estructurada o no estructurada, para el alcance de un objetivo en un escenario determinado (Goyal and Malviya, 2023; Lu, 2022; Mamatha et al., 2023). Este escenario puede tomar lugar en diversos sectores, tales como salud, financiero, manufacturero, agrícola, entretenimiento y comercio electrónico (Mamatha et al., 2023). Ha tomado gran relevancia en los sectores, especialmente por el aumento del volumen de información y la incorporación de las bases de datos en el arsenal tecnológico de las empresas (Kavitha, Jaisingh and Kaarthiekheyan, 2023; Xu et al., 2021). Sus bondades comprenden la identificación y descripción de patrones, tendencias y simetrías ocultas para el ojo humano en conjuntos de datos, dando valor a información aparentemente carente de importancia (Kavitha, Jaisingh and Kaarthiekheyan, 2023; Roman et al., 2023; Sundaram et al., 2023). De acuerdo con diversos autores, pilares de la ciencia de datos como la analítica son fundamentales para la toma de decisiones y el alcance de ventaja competitiva en las empresas (Carroll, 2023; Lo and Pachamanova, 2023).

En este contexto, la relevancia de orientar actividades científicas a asuntos relacionados con los salarios de los profesionales en ciencia de datos radica en la adecuada valoración del trabajo del recurso humano que ejecuta labores esenciales en una organización con interés en este tema. Partiendo de esa premisa, se ha desarrollado un sitio web que presenta información textual y gráfica y un modelo predictivo de salarios, con base en una recopilación de datos reales sobre salarios de 96 roles de ciencia de datos en distintos países del mundo. Se pretende que sea usado para la toma

* **Autor de contacto**: daniela.alcazar@udea.edu.co.

de decisiones en el futuro laboral de estudiantes y profesionales interesados en ciencia de datos.

La siguiente parte del documento está dividida así: la sección 2 expone la metodología usada para desarrollar el sitio web, en la sección 3 se explica de forma general las especificaciones de la página y se tratan con mayor profundidad en la sección 4; finalmente, en la sección 5 se disponen las conclusiones.

2. Metodología

La página web se crea a partir de elementos textuales y gráficos basados en un conjunto de datos abiertos obtenido de Kaggle y originado de aijobs.net, que está disponible en este vínculo. La base de datos contiene información de los salarios de profesionales del área de ciencia de datos en el periodo 2020-2023, cuenta con 3755 observaciones y 11 variables.

Para construir el la página web se emplearon 22 librerías. En la construcción estructural y estética de la página web se usaron las librerías **shiny** (Chang et al., 2023), **shinythemes** (Chang, 2021) y **wordcloud2** (Lang and tin Chien, 2018). Para la manipulación de datos se usaron las librerías **dplyr** (Wickham et al., 2023), **readr** (Wickham, Hester and Bryan, 2023) y **tidyverse** (Wickham et al., 2019). La visualización de los datos por medio de los distintos tipos de gráficos en la página web es posible usando las librerías **ggplot2** (Wickham, 2016), **ggridges** (Wilke, 2022), **ggtext** (Wilke and Wiernik, 2022), **rlang** (Henry and Wickham, 2023), **scales** (Wickham and Seidel, 2022), **sf** (Pebesma and Bivand, 2023), **leaflet** (Cheng et al., 2023) y **plotly** (Sievert, 2020); también se visualizan datos por medio de una tabla y un resumen, ambos en la sección descriptiva de la página web, para lo cual se usaron las librerías **DT** (Xie, Cheng and Tan, 2023), **gt** (Iannone et al., 2023) y **gtsummary** (Sjoberg et al., 2021). Por último, para el ajuste y validación del modelo de regresión lineal se emplearon las librerías **goftest** (Faraway et al., 2021), **lmtest** (Zeileis and Hothorn, 2002), **Metrics** (Hamner and Frasco, 2018), **MASS** (Venables and Ripley, 2002) y **nortest** (Gross and Ligges, 2015).

3. Interfaz

La aplicación web está disponible en este vínculo, y facilita el análisis de la base de datos mediante dos secciones principales: sección descriptiva y sección predictiva.

En la sección descriptiva el usuario puede interactuar con distintos elementos gráficos que le permitan entender las características de variables individuales de la base de datos. Asimismo, puede apreciar la relación entre el salario y otras (hasta dos) variables que él escoja; para algunas combinaciones de variables se tienen elementos visuales geográficos.

La sección predictiva se presenta como una herramienta para que el usuario realice predicciones de salarios con base en variables explicativas de su preferencia, también puede escoger las medidas de desempeño a las que se somete el modelo de regresión lineal. El valor predicho del salario, las características del modelo, los parámetros de predicción, las variables incluidas y los resultados de las medidas de desempeño son mostradas en esta sección.

En la Figura 1 se ilustra la ventana que se despliega al dar click al vínculo.

Figura 1: Pestaña 'Acerca del proyecto'

4. Resultados

4.1. Sección descriptiva

En la sección descriptiva se presentan tres subsecciones dispuestas en un menú de pestañas: "Análisis univariado", "Análisis multivariado" y "Datos". Esto se puede entender con mayor facilidad observando la Figura 2.

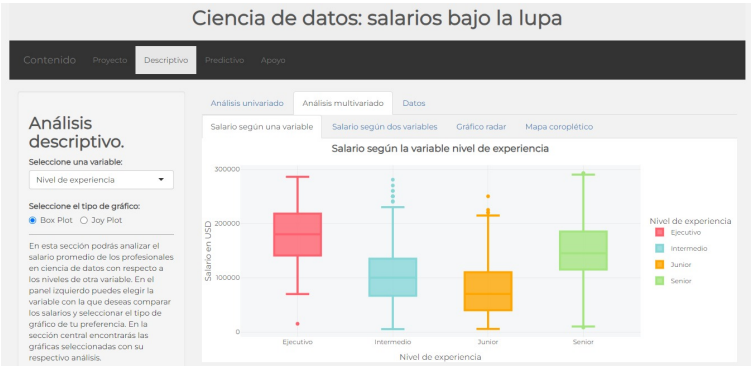

Figura 2: Sección descriptiva

La primera pestaña muestra, por medio de gráficos de barras y gráficos circulares, la información sobre las observaciones en la base de datos de las variables tipo factor: 'Año', 'Nivel de experiencia', 'Tipo de contrato', 'Rol', 'Moneda de pago', 'Residencia del empleado', 'Modalidad de trabajo, 'Ubicación de la empresa' y 'Tamaño de la compañía'.

En la pestaña "Análisis multivariado" se presentan cuatro opciones: "Salarios respecto a una variable", "Salario respecto a dos variables", "Gráfico radar" y "Mapa coroplético". La primera permite apreciar por medio de un diagrama de caja o un gráfico ridgeline la relación entre los niveles de una variable categórica elegida con el salario; la segunda opción, similar a la primera, permite al usuario observar el comportamiento del salario, esta vez el a partir de los niveles de dos variables mediante diagramas de barras en forma de múltiples paneles; finalmente, las últimas dos opciones presentan al usuario un gráfico de radar en el que puede observar los salarios medios con respecto a las variables 'Rol' y 'Nivel de experiencia', y un mapa coroplético.

La última pestaña muestra dos opciones que son “Base de datos” y “Resumen”. En “Base de datos” se muestra la base de datos usada en este estudio. El usuario puede elegir el número de datos que se muestran a la vez, filtrar la base de acuerdo con una variable que elija o por salario máximo y mínimo, e incluso descargar los datos. En la pestaña “Resumen” se presenta un resumen de las variables de la base de datos.

4.1. Sección predictiva

La sección predictiva se presenta como una herramienta para que el usuario, a partir de un modelo de regresión lineal múltiple, realice predicciones del salario en dólares con base en una configuración de las variables explicativas. De igual manera, en esta sección puede conocer las características y medidas de desempeño a las que es sometido el modelo.

5. Conclusiones

La aplicación web Shiny, desarrollada con R, proporciona apoyo a personas interesadas en desarrollarse profesionalmente en el campo de ciencia de datos. Esto se logra mediante la organización gráfica de la información de salarios recientemente devengados por profesionales del área, así como a través de un modelo de regresión lineal que estima el salario en función de especificaciones proporcionadas por el usuario.

En futuros estudios se debería considerar la inclusión de observaciones de otros países en la base de datos, a fin de tener una muestra representativa. De este modo, se tendría una visión más adecuada de las múltiples tendencias salariales en ciencia de datos y se mejoraría la aproximación de salarios para profesionales no estadounidenses.

Referencias

Carroll, Paula. 2023. "Analytics Modules for Business Students." *Operations Research Forum* 4:41.
URL: *https://doi.org/10.1007/s43069-023-00216-5*

Chang, Winston. 2021. *shinythemes: Themes for Shiny*. R package version 1.2.0.
URL: *https://CRAN.R-project.org/package=shinythemes*

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2023. *shiny: Web Application Framework for R*. R package version 1.7.5.
URL: *https://CRAN.R-project.org/package=shiny*

Cheng, Joe, Barret Schloerke, Bhaskar Karambelkar and Yihui Xie. 2023. *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. R package version 2.2.0.
URL: *https://CRAN.R-project.org/package=leaflet*

Faraway, Julian, George Marsaglia, John Marsaglia and Adrian Baddeley. 2021. *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. R package version 1.2-3.
URL: *https://CRAN.R-project.org/package=goftest*

Goyal, Priyanshi and Rishabha Malviya. 2023. "Challenges and opportunities of big data analytics in healthcare." *Health Care Science* .

Gross, Juergen and Uwe Ligges. 2015. *nortest: Tests for Normality*. R package version 1.0-4.
URL: *https://CRAN.R-project.org/package=nortest*

Hamner, Ben and Michael Frasco. 2018. *Metrics: Evaluation Metrics for Machine Learning*. R package version 0.1.4.
URL: *https://CRAN.R-project.org/package=Metrics*

Henry, Lionel and Hadley Wickham. 2023. *rlang: Functions for Base Types and Core R and 'Tidyverse' Features*. R package version 1.1.2.
URL: *https://CRAN.R-project.org/package=rlang*

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer and JooYoung Seo. 2023. *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.10.0.
URL: *https://CRAN.R-project.org/package=gt*

Kavitha, RK, W Jaisingh and V Kaarthiekheyan. 2023. Application of Artificial Intelligence and Data Science Across Domains: A Perspective Study. In *Handbook of Research on Data Science and Cybersecurity Innovations in Industry 4.0 Technologies*. IGI Global pp. 1–29.

Lang, Dawei and Guan tin Chien. 2018. *wordcloud2: Create Word Cloud by 'htmlwidget'*. R package version 0.2.1.
URL: *https://CRAN.R-project.org/package=wordcloud2*

Lo, Victor S Y and Dessislava A Pachamanova. 2023. "From Meaningful Data Science to Impactful Decisions: The Importance of Being Causally Prescriptive." *Data Science Journal* .

Lu, Jing. 2022. "Data science in the business environment: Insight management for an Executive MBA." *The International Journal of Management Education* 20:100588.
URL: *https://www.sciencedirect.com/science/article/pii/S1472811721001373*

Mamatha, T, A Balaram, B Rama Subba Reddy, C Shoba Bindu and M Niranjanamurthy. 2023. "Applications and Advancements in Data Science and Analytics." *Data Engineering and Data Science: Concepts and Applications* pp. 409–439.

Pebesma, Edzer and Roger Bivand. 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC.
URL: *https://r-spatial.org/book/*

Roman, Alexander, Roy T Forestano, Konstantin T Matchev, Katia Matcheva and Eyup B Unlu. 2023. "Oracle-Preserving Latent Flows." *Symmetry* 15.
URL: *https://www.mdpi.com/2073-8994/15/7/1352*

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.
URL: *https://plotly-r.com*

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery and Joseph Larmarange. 2021. "Reproducible Summary Tables with the gtsummary Package." *The R Journal* 13:570–580.
URL: *https://doi.org/10.32614/RJ-2021-053*

Sundaram, Jawahar, K Gowri, S Devaraju, S Gokuldev, Sujith Jayaprakash, Harishchander Anandaram, C Manivasagan and M Thenmozhi. 2023. An Exploration of Python Libraries in Machine Learning Models for Data Science. In *Advanced Interdisciplinary Applications of Machine Learning Python Libraries for Data Science*. IGI Global pp. 1–31.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer. ISBN 0-387-95457-0.
URL: *https://www.stats.ox.ac.uk/pub/MASS4/*

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wickham, Hadley and Dana Seidel. 2022. *scales: Scale Functions for Visualization*. R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=scales*

Wickham, Hadley, Jim Hester and Jennifer Bryan. 2023. *readr: Read Rectangular Text Data*. R package version 2.1.4.
URL: *https://CRAN.R-project.org/package=readr*

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo and Hiroaki Yutani. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4(43):1686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. R package version 1.1.3.
URL: *https://CRAN.R-project.org/package=dplyr*

Wilke, Claus O. 2022. *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.5.4.
URL: *https://CRAN.R-project.org/package=ggridges*

Wilke, Claus O. and Brenton M. Wiernik. 2022. *ggtext: Improved Text Rendering Support for 'ggplot2'*. R package version 0.1.2.
URL: *https://CRAN.R-project.org/package=ggtext*

Xie, Yihui, Joe Cheng and Xianying Tan. 2023. *DT: A Wrapper of the JavaScript Library 'DataTables'*. R package version 0.30.
URL: *https://CRAN.R-project.org/package=DT*

Xu, Zongben, Niansheng Tang, Chen Xu and Xueqi Cheng. 2021. "Data science: connotation, methods, technologies, and development." *Data Science and Management* 1(1):32–37.

Zeileis, Achim and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2(3):7–10.
URL: *https://CRAN.R-project.org/doc/Rnews/*