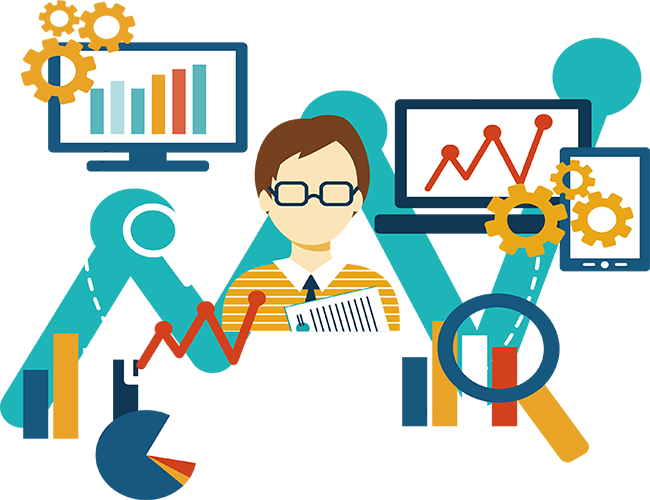


Proyecto de finalización

Análisis de los salarios de Ciencia de datos en 2023

Materia: Procesamiento de datos

Curso: 2°

Carrera: Ciencia de Datos e Inteligencia Artificial Año: 2023

Índice

[Introducción… 2](#_TOC_250002)

[Desarrollo del proyecto… 3](#_TOC_250001)

* Recolección y preparación de datos
* Análisis exploratorio de datos
* Procesamiento y análisis de datos
* Visualización de datos
* Generación de reportes Conclusiones… 8

[Referencias… 9](#_TOC_250000)

# Introducción

El presente proyecto tiene como objetivo realizar un análisis de los sueldos en el campo de la ciencia de datos durante el año 2023. Como estudiantes de la tecnicatura en Ciencia de Datos e Inteligencia Artificial, consideramos que este tema es de gran relevancia, ya que nos permite profundizar nuestro conocimiento en el área y entender las tendencias salariales en este campo en constante crecimiento.

Para llevar a cabo este análisis, utilizaremos un conjunto de datos que incluye información detallada sobre los sueldos en el campo de la ciencia de datos. Los datos que analizaremos incluyen variables como el año de trabajo, el nivel de experiencia, el tipo de empleo, el título del puesto, el monto total del salario, la moneda en la que se pagó el salario, el salario en USD, el país de residencia del empleado, la proporción de trabajo remoto, el país de ubicación de la empresa y el tamaño de la empresa en términos de número de empleados.

Nuestro objetivo principal es realizar un análisis exhaustivo de estos datos para obtener información valiosa sobre los sueldos en el campo de la ciencia de datos en el año 2023. Mediante técnicas de procesamiento y análisis de datos, exploramos patrones, tendencias y relaciones entre las variables mencionadas. También utilizaremos herramientas de visualización de datos para representar gráficamente la información obtenida, lo que nos permitirá comunicar de manera efectiva los resultados del análisis.

Este proyecto nos brindará una comprensión más profunda de los sueldos en el campo de la ciencia de datos, lo cual es fundamental para nuestra formación como profesionales en este ámbito. Además, los resultados obtenidos podrán ser de utilidad para otros estudiantes, profesionales y empresas que deseen conocer las tendencias salariales en este campo en constante evolución.

# Desarrollo del proyecto

Recolección y preparación de datos

En este proyecto, utilizamos conjuntos de datos extraídos de Kaggle, una reconocida plataforma de recursos y competiciones en ciencia de datos. Los conjuntos de datos seleccionados contienen información relevante sobre los sueldos en el campo de la ciencia de datos durante el año 2023. Es importante destacar que estos datos fueron obtenidos a través de una encuesta realizada a profesionales del campo y, si bien no son información oficial, proporcionan un buen parámetro para analizar los montos salariales en esta área.

Una vez obtenidos los conjuntos de datos de la encuesta, procedimos a realizar las tareas de limpieza, integración y transformación necesarias para preparar los datos para el análisis. Utilizando SQL como lenguaje principal, realizamos un análisis detallado de cada una de las columnas de los datos. Además, llevamos a cabo la traducción de las variables a un formato más comprensible y unificamos variables similares para facilitar el análisis.

Durante la limpieza de los datos, eliminamos aquellos registros que no eran útiles para nuestro análisis, como entradas con valores inconsistentes. Asimismo, reemplazamos acrónimos y códigos con terminología más clara y significativa. Por ejemplo, en la columna de presencialidad, reemplazamos los valores numéricos 0, 50 y 100 por "presencial", "híbrido" y "remoto", respectivamente, para facilitar el entendimiento de la base de datos.

Una vez completadas las tareas de limpieza, integración y transformación, los datos estuvieron listos para su análisis. Esta etapa de preparación de datos nos permitió contar con un conjunto de datos coherente, comprensible y optimizado para su posterior exploración y análisis de sueldos en el campo de la ciencia de datos en 2023.

Los comandos utilizados se pueden observar en la carpeta /sql de nuestro [repositorio](https://github.com/Agusgal16/Procesamiento-de-Datos) de GitHub y las bases de datos en la carpeta /datasets.

Análisis exploratorio de datos

Una vez preparados los datos, procedimos a realizar un análisis exploratorio detallado con el objetivo de comprender y explorar los datos en profundidad.

En primer lugar, realizamos un análisis descriptivo de las variables principales, como los sueldos, la experiencia laboral y el tipo de empleo. Esto nos permitió tener una visión general de las características de los sueldos en el campo de la ciencia de datos en 2023.

En nuestro proyecto, utilizamos Google Colab para realizar diversas consultas y análisis exploratorios. A continuación, se detallan algunas de las consultas que realizamos:

* Primeras 5 observaciones: Utilizamos la función head() para obtener las primeras 5 observaciones de los datos y examinar su estructura y contenido.
* Resumen de los datos: Utilizamos la función info() para obtener un resumen de las variables, incluyendo el tipo de dato, el número de valores no nulos y el uso de memoria.

Estas consultas nos permitieron obtener una visión general de los datos y asegurarnos de que la preparación de los datos se realizó correctamente. También pudimos identificar posibles

problemas, como la presencia de valores faltantes o la necesidad de realizar conversiones de tipo de datos.

Los comandos utilizados se pueden observar en la carpeta /notebooks de nuestro [repositorio](https://github.com/Agusgal16/Procesamiento-de-Datos) de GitHub.

Procesamiento y análisis de datos

En el paso de Procesamiento y análisis de datos, utilizamos Python en Jupyter Notebooks con la librerías Pandas para realizar diversas tareas, incluyendo cálculo de métricas, filtrado de datos y agrupación de información relevante.

En primer lugar, procedimos a calcular diferentes métricas relacionadas con los salarios en dólares utilizando las columnas pertinentes de nuestro conjunto de datos. Utilizamos las siguientes operaciones sobre la columna 'salario\_en\_dolares':

* Calculamos la media del salario utilizando la función mean(). Esto nos permitió obtener un valor que representa el promedio de los salarios en dólares de la muestra.
* Calculamos la mediana del salario mediante la función median(). Esta métrica representa el valor que se encuentra en el medio de la distribución de los salarios en dólares.
* Determinamos el salario máximo y mínimo utilizando las funciones max() y min(), respectivamente. Estas métricas nos proporcionaron información sobre los valores extremos de los salarios en dólares.
* Calculamos la desviación estándar del salario mediante la función std(). Esta métrica nos permitió entender la dispersión de los salarios en dólares y evaluar la variabilidad de los datos con respecto a la media.
* Luego de calcular estas métricas, las presentamos mediante la función print() para mostrar los resultados obtenidos.

Además, realizamos filtrado de datos para obtener información específica sobre la experiencia y el título de los empleados. Utilizamos la función loc[] para filtrar los registros donde la experiencia era "Semi Senior". Esto nos permitió obtener un subconjunto de datos que cumplía con este criterio.

Posteriormente, realizamos otro filtrado de datos combinando dos condiciones: experiencia igual a "Semi Senior" y posición igual a "Científico de datos". Empleamos el operador lógico "&" para combinar las condiciones y obtener un subconjunto más específico que cumplía ambas.

Por último, llevamos a cabo la agrupación de datos utilizando la función groupby(). Agrupamos los datos por la columna 'experiencia' y 'posicion', y aplicamos la función agg() a la columna 'salario\_en\_dolares'. Esto nos permitió calcular métricas específicas, como el salario mínimo y máximo, dentro de cada grupo definido por la combinación de experiencia y posición. Los resultados fueron presentados mediante la función print().

El uso de Python en Jupyter Notebooks y la librerías Pandas nos proporcionó las herramientas necesarias para realizar eficientemente el procesamiento y análisis de datos. A través del cálculo de métricas, el filtrado de datos y la agrupación, pudimos obtener

información valiosa y significativa sobre los salarios en dólares y explorar características específicas de los empleados en función de su experiencia y posición.

Los comandos utilizados se pueden observar en la carpeta /notebooks de nuestro [repositorio](https://github.com/Agusgal16/Procesamiento-de-Datos) de GitHub.

Visualización de datos

En la etapa de Visualización de datos, continuamos trabajando en Google Colab y utilizamos diversas técnicas y herramientas para representar gráficamente la información obtenida a partir de nuestros datos. A través del siguiente código en Python, empleamos la biblioteca Matplotlib para crear visualizaciones claras y significativas:

* Gráfico de barras: Salario por experiencia

Ordenamos los datos de acuerdo al salario en dólares de manera descendente y utilizamos la función bar() para crear un gráfico de barras que muestra la relación entre la experiencia y el salario. El eje x representa la experiencia, mientras que el eje y representa el salario en dólares. Agregamos etiquetas a los ejes y un título descriptivo para facilitar la comprensión de la visualización.

* Gráfico de barras: Salario por posición

Creamos un gráfico de barras utilizando la función bar() para representar la relación entre la posición y el salario en dólares. Ajustamos el tamaño de la figura y rotamos las etiquetas en el eje x para mejorar la legibilidad de los datos. También agregamos etiquetas a los ejes y un título para brindar una descripción clara de la visualización.

* Diagrama de caja: Salario por presencialidad

Utilizamos la función boxplot() para generar un diagrama de caja que muestra la distribución del salario en dólares según la modalidad de trabajo: presencial, híbrido y remoto. Los diferentes diagramas de caja se representan en el eje x y el salario en dólares en el eje y. Añadimos etiquetas a los ejes y un título descriptivo para facilitar la interpretación de la visualización.

Estas visualizaciones nos permiten explorar y comprender mejor los datos relacionados con los salarios en función de diferentes variables, como la experiencia, la posición y la presencialidad. Proporcionan una representación gráfica efectiva de la información obtenida y nos ayudan a identificar patrones, tendencias y posibles valores atípicos en los datos.

Además de las visualizaciones realizadas con Matplotlib en Python, también utilizamos Power BI para crear otros gráficos y análisis visuales. A continuación, se describen algunos de ellos:

* Cantidad de trabajadores por posición y experiencia:

Utilizando los datos disponibles, creamos un gráfico que muestra la cantidad de trabajadores en cada posición según su nivel de experiencia. Esta visualización nos permite identificar las distribuciones y proporciones de empleados en diferentes roles y niveles de experiencia dentro de la organización.

* Promedio de salario en dólares por experiencia:

Mediante Power BI, generamos un gráfico que muestra el promedio de salario en dólares para cada nivel de experiencia. Esta visualización nos ayuda a comprender cómo varía el salario promedio en función de la experiencia laboral de los empleados.

* Gráfico de torta de tipos de moneda utilizados para el pago de sueldos:

Utilizando la función de gráfico de torta en Power BI, representamos la distribución de diferentes tipos de moneda utilizados por los trabajadores para recibir su salario. Esta visualización nos permite identificar las principales monedas utilizadas y su participación relativa en la muestra de datos.

* Total de personas que trabajan en empresas medianas, chicas o grandes:

A través de un gráfico en Power BI, representamos el recuento total de personas empleadas en empresas clasificadas como medianas, chicas o grandes según ciertos criterios establecidos. Esta visualización nos brinda una visión general del tamaño de las empresas y la cantidad de empleados asociados a cada categoría.

* Tabla de promedio de salario en dólares según la modalidad de empleo:

Utilizando Power BI, creamos una tabla que muestra el promedio de salario en dólares para diferentes modalidades de empleo, como contractor, part time, freelancer o full time. Esta tabla nos permite comparar y analizar las diferencias salariales entre estas modalidades.

Los reportes se pueden observar en la carpeta /reports de nuestro [repositorio](https://github.com/Agusgal16/Procesamiento-de-Datos) de GitHub. Generación de reportes

En nuestro proyecto de análisis de sueldos de ciencia de datos en 2023, llevamos a cabo el paso de Comunicación efectiva de resultados de la siguiente manera:

Estructura del reporte: Creamos un reporte estructurado que consta de una introducción donde explicamos el objetivo del proyecto, seguido de una descripción detallada de los datos utilizados, incluyendo su origen y características. Luego, presentamos los métodos utilizados en el procesamiento y análisis de los datos. Finalmente, dedicamos una sección para presentar los resultados obtenidos.

Lenguaje claro y conciso: Redactamos el reporte utilizando un lenguaje claro y accesible para cualquier lector interesado en el tema de los sueldos de ciencia de datos. Evitamos utilizar terminología técnica compleja y proporcionamos explicaciones claras de los conceptos y resultados presentados.

Visualización de datos: Utilizamos diversas técnicas y herramientas de visualización de datos para presentar los resultados de manera efectiva. Creamos gráficos de barras para comparar los salarios según el nivel de experiencia y el tipo de empleo. También utilizamos gráficos de dispersión para visualizar la relación entre el salario y el año de trabajo. Además, incluimos diagramas de caja para mostrar la distribución de los salarios según el tipo de empleo. Estas visualizaciones ayudaron a transmitir de manera clara y concisa la información obtenida.

Interpretación de resultados: En el reporte, interpretamos los resultados obtenidos a partir de las visualizaciones y análisis realizados. Analizamos los patrones y tendencias identificados en los sueldos de ciencia de datos en 2023, y proporcionamos explicaciones claras sobre su significado. También extrajimos conclusiones relevantes, como las diferencias salariales según el nivel de experiencia y el tipo de empleo.

Audiencia objetivo: Adaptamos el contenido del reporte a nuestra audiencia objetivo, que en este caso son estudiantes de la tecnicatura en Ciencia de Datos e Inteligencia Artificial. Proporcionamos el contexto necesario y explicamos los conceptos clave de manera comprensible para aquellos que están iniciando en el campo de la ciencia de datos.

Formato del reporte: Decidimos utilizar un formato digital, como un archivo PDF, que nos permitió presentar de manera organizada y visualmente atractiva los resultados y conclusiones del proyecto.

El reporte se puede descargar desde el README de nuestro repositorio o desde el siguiente [link](https://drive.google.com/file/d/1-ukhjwdCAdTuV936GL3F4GSA2ON605YX/view?usp=sharing).

# Conclusiones

Basado en una muestra de 1000 personas, hemos llegado a las siguientes conclusiones:

* El 92% de las personas encuestadas cobra su salario en dólares. Esto puede indicar una preferencia por esta moneda en el campo de la ciencia de datos.
* La mayoría de las personas encuestadas trabajan en modalidad full time, lo que sugiere que esta es la forma de empleo más común en esta industria.
* El salario promedio en dólares por año es de 148,574. Esto proporciona una idea general de los niveles salariales en el campo de la ciencia de datos en el año 2023.
* No hay grandes diferencias salariales entre aquellos que trabajan en modalidad remoto o presencial. Esto puede indicar que el tipo de trabajo no tiene un impacto significativo en el salario.
* La mayoría de las empresas en las que trabajan las personas encuestadas son de tamaño mediano. Esto puede reflejar la distribución de empresas en la industria de la ciencia de datos.
* En cuanto a las categorías de experiencia, el promedio salarial en dólares para un junior es de 101,811, para un semi senior es de 119,869, para un senior es de 159,542 y para un líder de equipo es de 189,832. Esto muestra una tendencia de incremento salarial a medida que se adquiere más experiencia en el campo.
* Los salarios más altos se observan en las personas que se dedican a machine learning, arquitectos de datos e ingenieros de datos. Esto puede deberse a la demanda y especialización requerida en estas áreas.

Estas conclusiones proporcionan una visión general de los sueldos en la industria de la ciencia de datos en 2023, y pueden ser útiles para comprender las tendencias y patrones salariales en este campo. Es importante tener en cuenta que estas conclusiones se basan en una muestra de 1000 personas y pueden no representar la totalidad de la industria.

**Referencias**

Durante el desarrollo de nuestro proyecto, utilizamos la base de datos "Data Science Salaries 2023" que obtuvimos de Kaggle, una reconocida plataforma de código abierto que alberga diversos conjuntos de datos. Específicamente, accedimos a esta base de datos a través del siguiente [enlace](https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023).

La base de datos "Data Science Salaries 2023" fue recopilada originalmente por AIJobs.net, una plataforma dedicada a brindar información y recursos relacionados con trabajos en el campo de la inteligencia artificial y la ciencia de datos. AIJobs.net proporciona una amplia gama de datos e insights relevantes para aquellos interesados en carreras en estas áreas.

La base de datos en cuestión contiene información detallada sobre los salarios en el campo de la ciencia de datos, específicamente para el año 2023. Incluye variables como el año de trabajo, el nivel de experiencia, el tipo de empleo, el título del puesto, el monto del salario, la moneda utilizada, la equivalencia del salario en dólares estadounidenses, la residencia del empleado, el porcentaje de trabajo remoto, la ubicación de la empresa y el tamaño de la misma.

Al utilizar esta base de datos en nuestro proyecto, hemos podido analizar y explorar los salarios en el campo de la ciencia de datos en 2023, aplicando técnicas de procesamiento, análisis y visualización de datos. La información proporcionada por esta base de datos nos ha permitido obtener insights valiosos y realizar evaluaciones sobre los sueldos en esta industria.

Es importante destacar que, si bien la base de datos "Data Science Salaries 2023" proporciona un conjunto de datos significativo para nuestro análisis, debemos tener en cuenta que no es información oficial, sino que se obtiene a partir de una encuesta realizada por AIJobs.net. Sin embargo, consideramos que esta base de datos es un buen punto de partida y un parámetro útil para analizar los montos salariales en el campo de la ciencia de datos durante el año 2023.