# Introducción a la Ciencia de Datos

CIMAT 2023

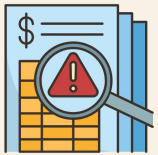
Act. Lorena Pineda Rodríguez
Act. Carla Paola Malerva Reséndiz

## ¿Qué es la ciencia de Datos?

Campo de estudio que combina la experiencia de negocio, las habilidades de programación y el conocimiento de las matemáticas y la estadística para extraer información significativa de los datos. Producir sistemas de inteligencia artificial (IA) (fuentes de información como: números, texto, imágenes, video, audio y más ) para realizar tareas que normalmente requieren inteligencia humana.

**Aplicaciones:**

- Determinar la fuga de clientes
- Mejorar la eficiencia al analizar los patrones de tráfico, las condiciones climáticas y otros factores para que las empresas de logística puedan mejorar los tiempos de entrega y reducir los costos.
- Optimizar la cadena de suministro al predecir cuándo se producirán fallos en los equipos
- Detectar los fraudes en los servicios financieros.
- Crear recomendaciones para los clientes basadas en las compras anteriores.




https://github.com/Malerva/CIMAT_2023_05

# ¿Qué es un Científico de Datos?

Además de las habilidades base mencionadas, los profesionales del rubro deberían ir un poco más allá cuando se habla de trabajar con un recurso tan sensible como lo es la información.

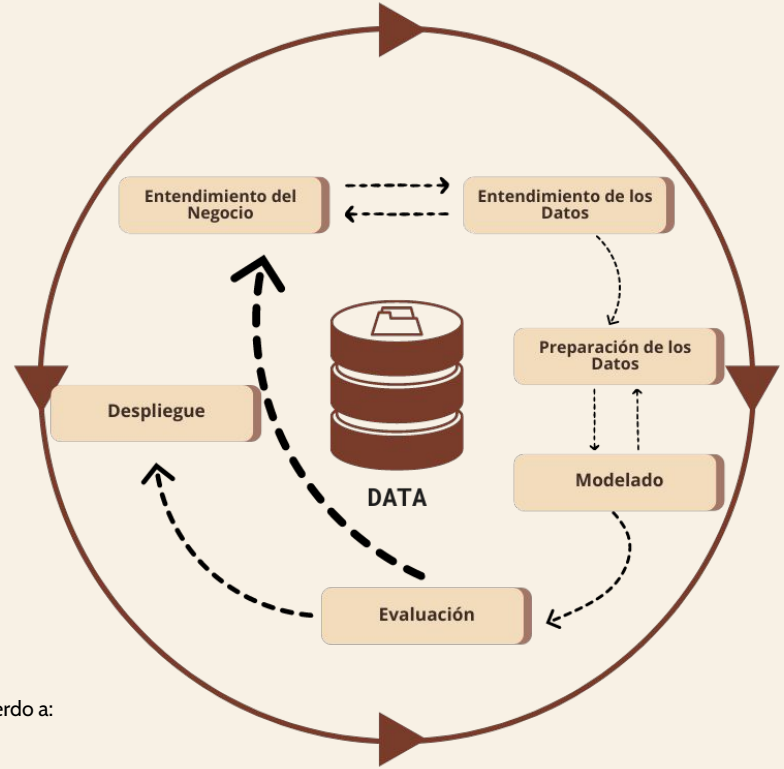

Los algoritmos aprenden sobre lo que nosotros decidimos proporcionar, pero tener una participación activa en este rubro podrá generar soluciones de maneras más comprometidas y constructivas, siempre con un enfoque en el bien común.

Consiste desde hacer las preguntas correctas a la información, imaginar las posibilidades para un problema hasta entender si el mensaje es accionable.

# ¿Cómo es el día a día de un Científico de Datos?

Las habilidades mencionadas anteriormente comenzarán a tomar sentido en el siguiente proceso, ya que un CD se involucra desde la necesidad del negocio, entendimiento de los datos, propuestas de soluciones y finalmente hasta llegar a un entregable accionable.

- Identifica la necesidad de negocio
- Problema que se quiere resolver
- Impacto en el negocio
- KPI principales
- Áreas que se impactan
- Objetivo que se desea alcanzar




- Extracción de los datos
- Diccionario
- Análisis de la información
  - Valida la información
  - Tipos de datos
  - Rangos de los mismos
- Identificar posibles desafios
-

- Limpieza y transformación de los datos
- Asegurar su calidad
- Prepararlos para ser el output de un modelo de ML

- Identificar opciones de modelos más adecuados
- Entrenamiento
- Hiperparametrización

- ETL
- Diseño de herramienta para el consumo de la solución
- Revisión de desempeño

- Seleccionar el mejor de acuerdo a:
  - Métricas
  - Características
  - Necesidades del negocio

# PROYECTO: CIENCIA TECNOLOGÍA Y GÉNERO

Objetivo: Guiarlos a través del proceso que atraviesa un Científico de Datos en su día a día mediante un ejemplo práctico, desde la ideación, generación de la información, data cleaning, ingeniería, modelado, evaluación y finalizando con el análisis de los resultados.

# CIENCIA TECNOLOGÍA Y GÉNERO : Ideación del objetivo

Actualmente en la industria tecnológica se tiene una gran disparidad en cuanto al género de las personas que se dedican a la ciencia de datos. Preocupadas y curiosas ante el por qué puede estar sucediendo esto realizamos una encuesta para conocer los puntos de vista, tanto de hombres como de mujeres.

## Definición de Objetivo

- Conocer la perspectiva actual respecto a la disparidad en la industria tech
- Analizar los factores que influyen en la desigualdad.

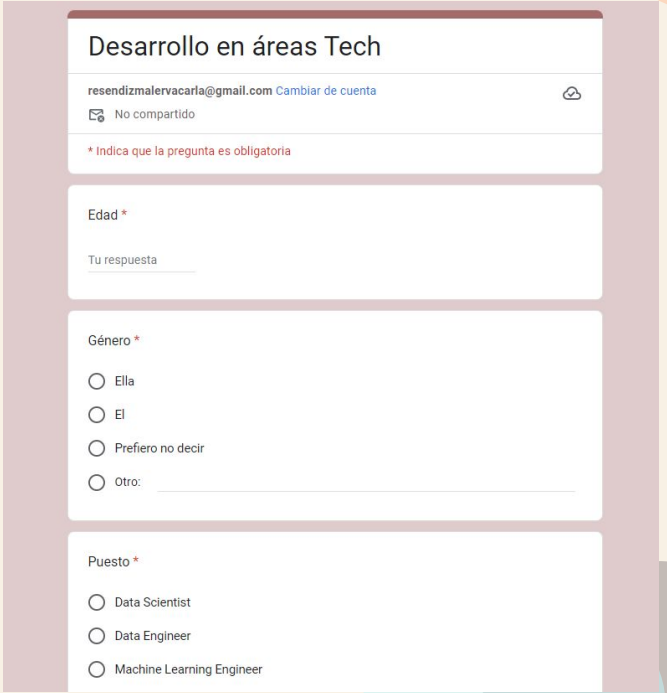
## Generación y recopilación de los Datos


- Determinar al público objetivo
- Definir lo que se quiere conocer del usuario
- Creación de las preguntas relacionadas con el objetivo
- Congruencia en el diseño de la encuesta (no ambigüedad)
- Acotar las respuestas por longitud, tipo de dato, etc.

# CIENCIA TECNOLOGÍA Y GÉNERO : Generación y Recopilación de Datos

Los puntos principales que se incluyeron en la encuesta fueron los siguientes:

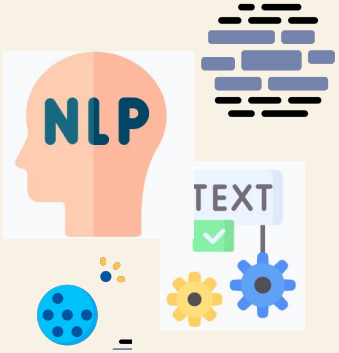
- Edad
- Género con el que se identifican
- Nivel de desarrollo en hard skills
- Nivel de desarrollo en soft skills
- Puesto
- Antigüedad en su trabajo actual
- Años de experiencia
- Ascensos salariales
- Que desearían haber sabido antes de entrar al area tech
- Sesgos en la etapa de reclutamiento
- Trato diferente con pares
- Problemas de confianza
- **Opinión respecto a la situación actual de las mujeres en la industria tech**



# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

Se recopilaron 135 respuestas de estudiantes del Diplomado en Ciencia de Datos de la Facultad de Estudios Superiores Acatlán de la UNAM.
Los datos fueron almacenados en un archivo csv a través de Google forms, el cual contiene distintos formatos dependiendo el público que contestó las preguntas.

## Limpieza




- Homologación de categorías
- Eliminación de datos atípicos
- Imputación de valores ausentes
- Transformación de columnas
- Limpieza de texto: stopwords, stemming, lematización

## Análisis

- Entendimiento de la situación actual en el sector
- Profundizar sobre las diferencias específicas que existen en las habilidades, ingresos , género, edad, etc.
- Se busca identificar si existe alguna relación entre el género y la compensación económica

# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

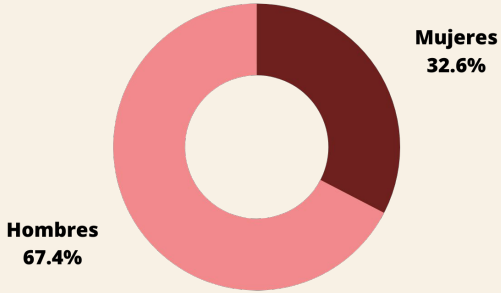
135 Registros

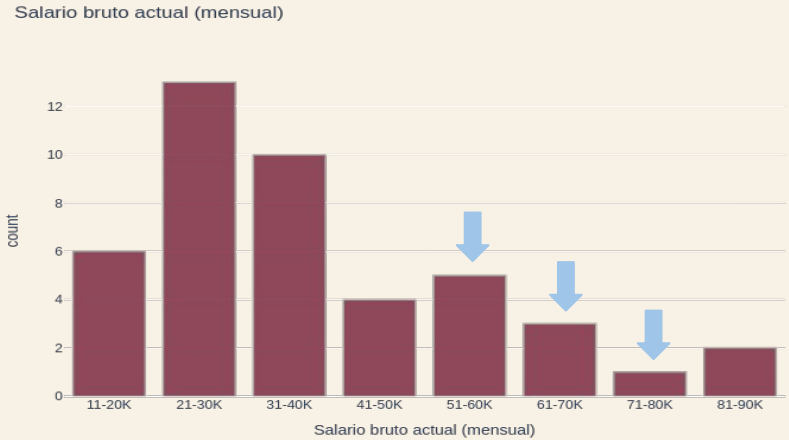
Posicion:

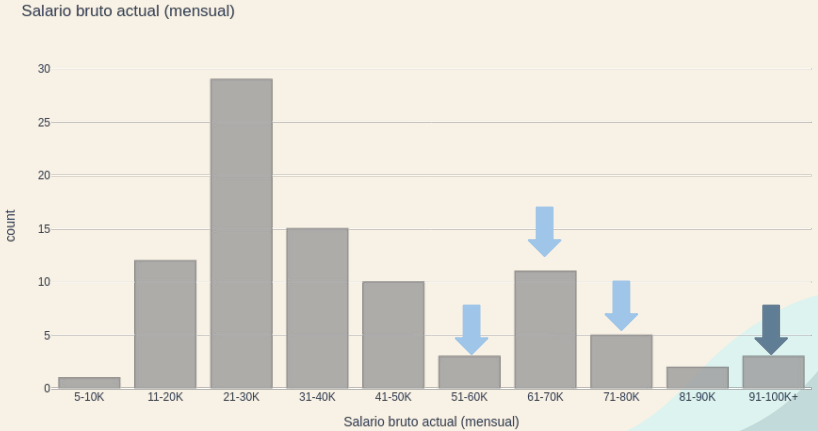
Data Scientist 34%
Data Analyst 27%
Data Engineer 5%
Devops 2%

Edad:

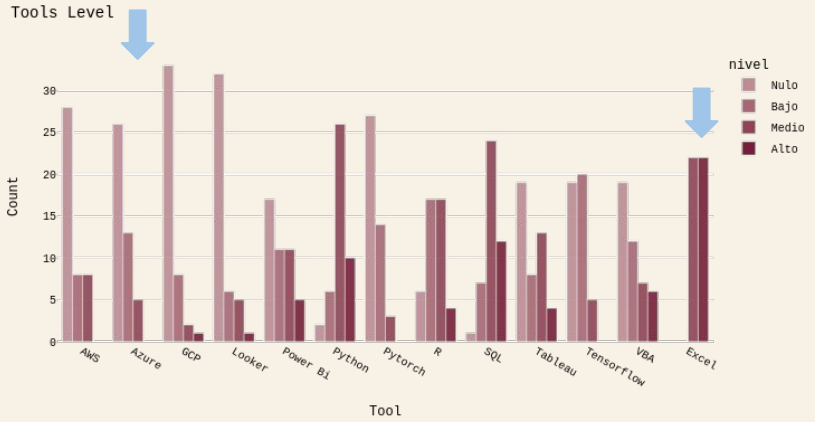
- Mínimo 22 años
- Media 29 años
- Mediana 27 años
- Max 56 años

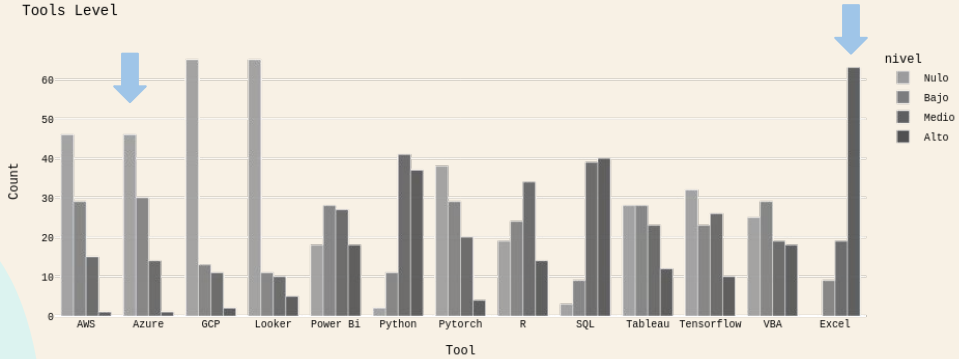


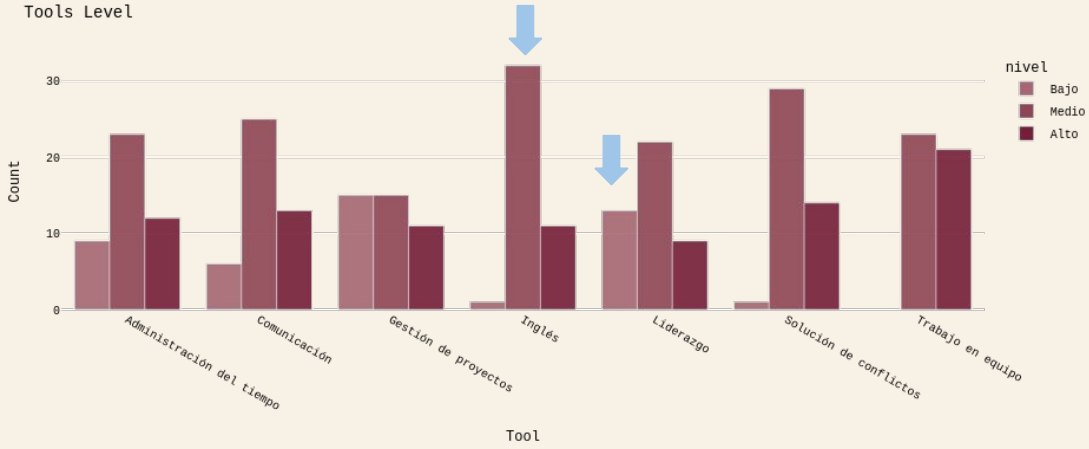
# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

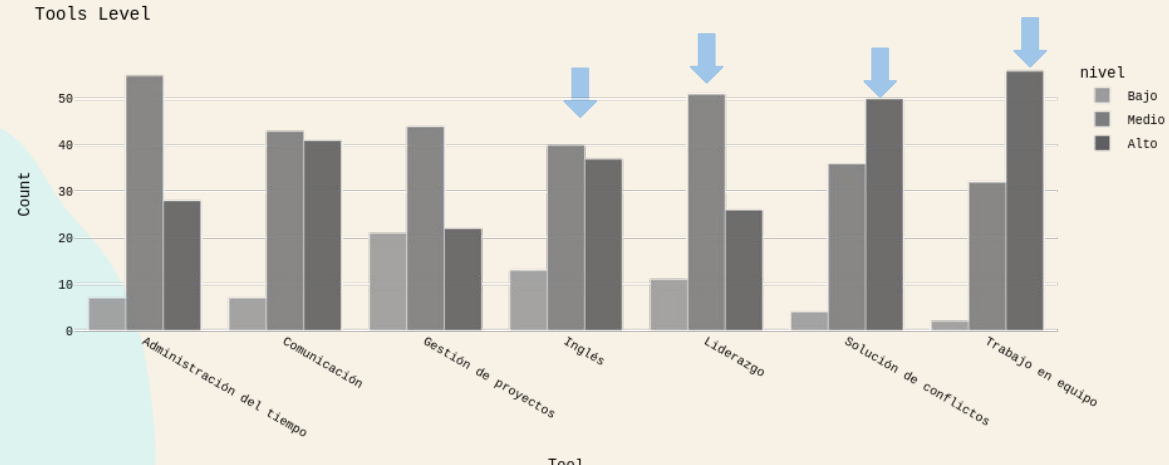



- **Excel, SQL, Python** —> Herramientas donde se cuenta con un nivel más alto de adopción
- Herramientas de la nube con bajo porcentaje de dominio
- En general el manejo de las herramientas de visualización en ambos casos es bajo
- El manejo de herramientas más especializadas como lo son Pytorch y TensorFlow es muy bajo.

# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información




- Liderazgo, hay más mujeres con un nivel bajo
- En general hay más hombres con un nivel alto en Inglés y en resolución de problemas.
- Los hombres tienen un índice más alto de trabajo en equipo

# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

## Puesto

| Mujeres | Hombres |
|---|---|
| Data Analyst (36%) | Data Scientist (37%) |
| Data Scientist (27%) | Data Analyst (28%) |
| Data Engineer (7%) | Data Engineer (4%) |

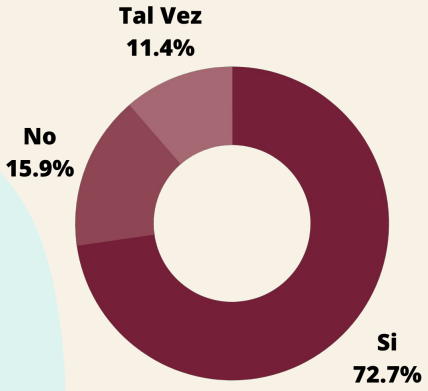
## Sector

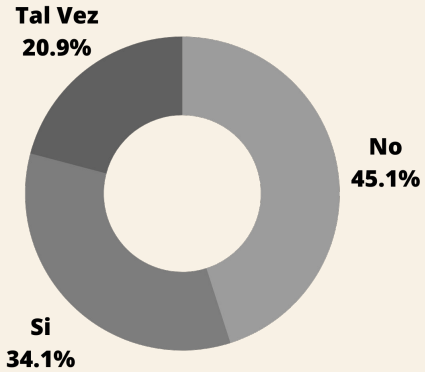
| Mujeres | Hombres |
|---|---|
| Finanzas (56%) | Finanzas (45%) |
| Informática y tecnología (14%) | Informática y tecnología (15%) |
| Retail (4.5%) | Publicidad y Marketing (5%) |

¿Crees que es importante que más mujeres se unan a la industria?

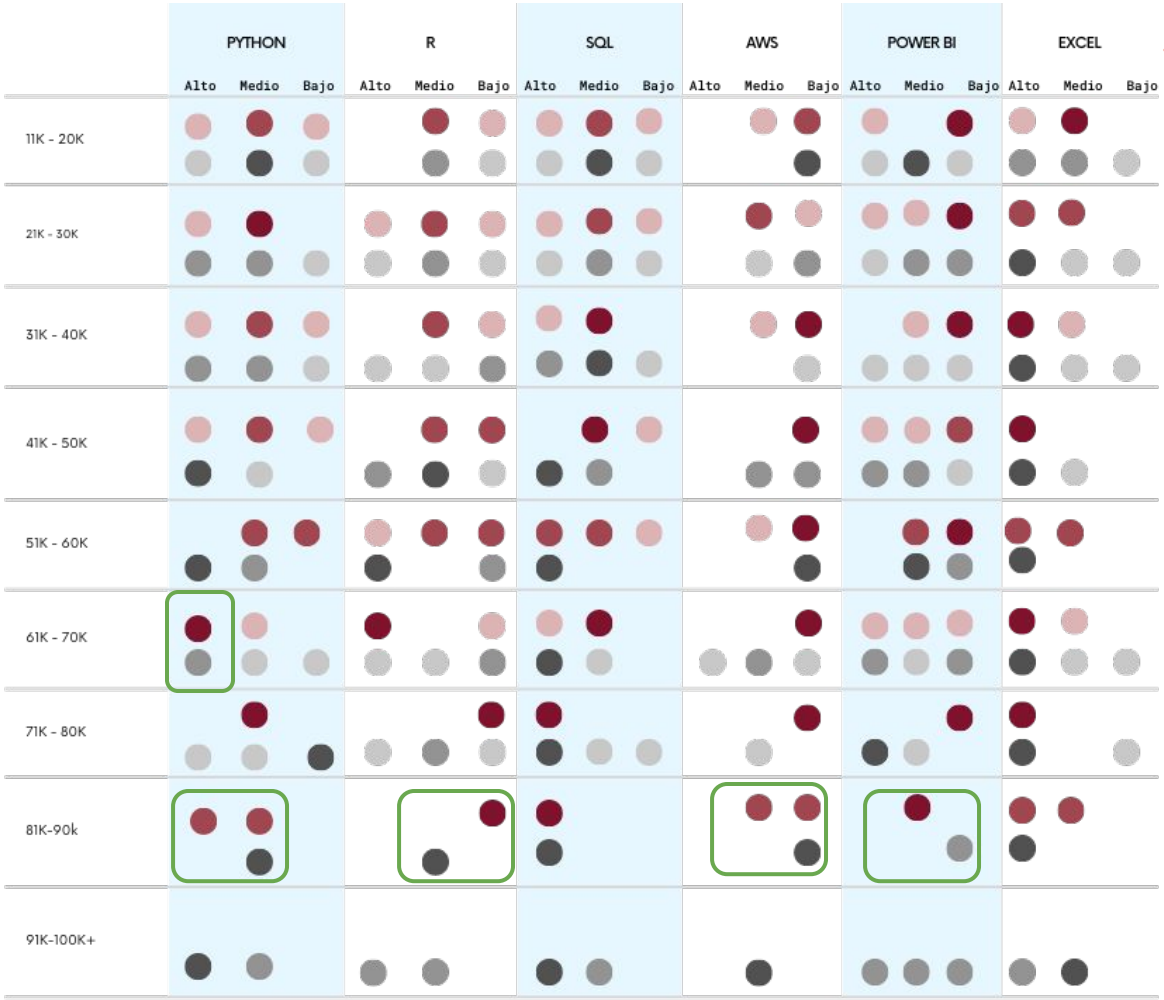
- 95% Consideran que es importante que más mujeres se involucren
- 5% Restante tiene un punto de vista diferente

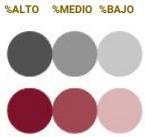
## ¿Alguna vez luchas con la confianza?



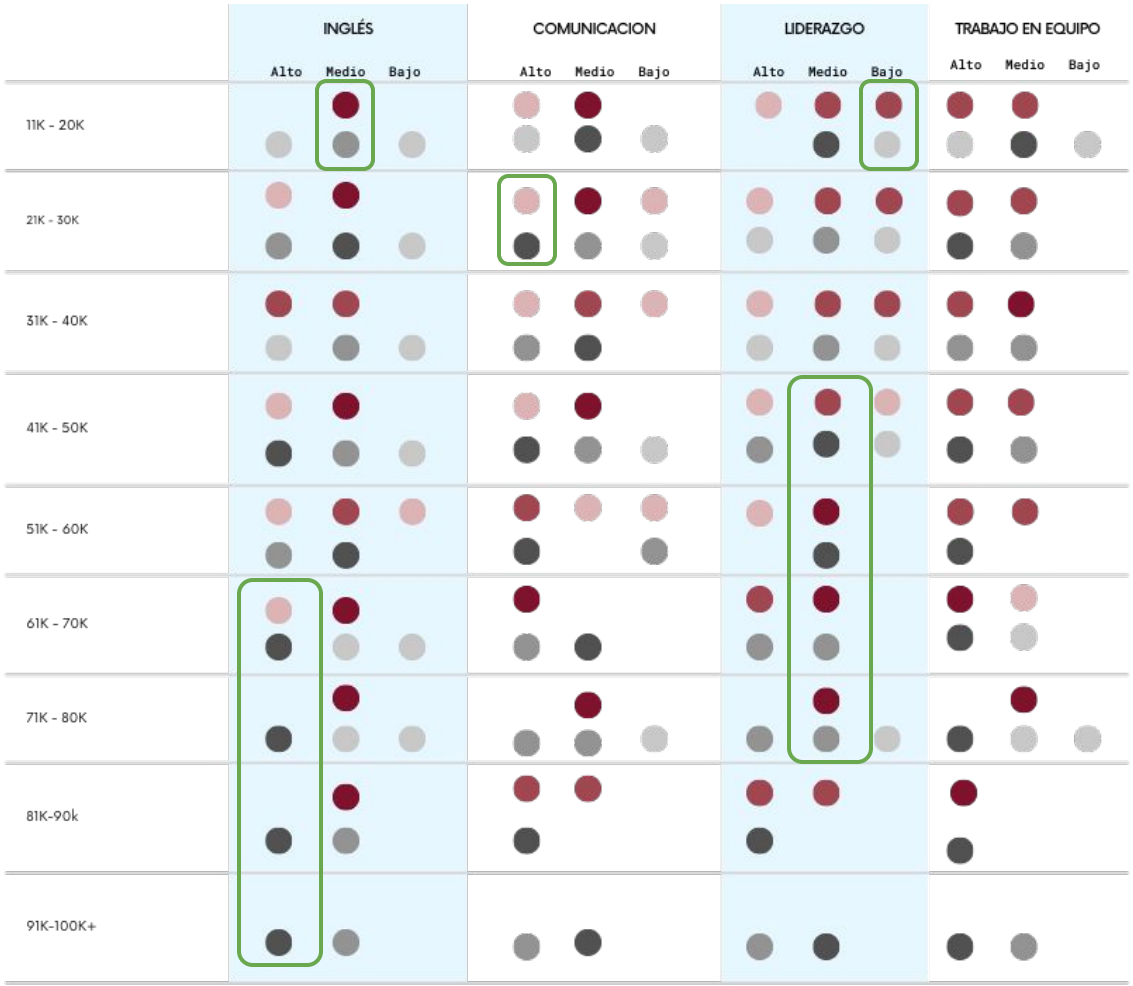




- 47% Si han luchado alguna vez con la confianza
- 35% de los encuestados no ha pasado por esto
- Finalmente un 18% de los usuarios considera que “Tal vez”




Excel es una herramienta base

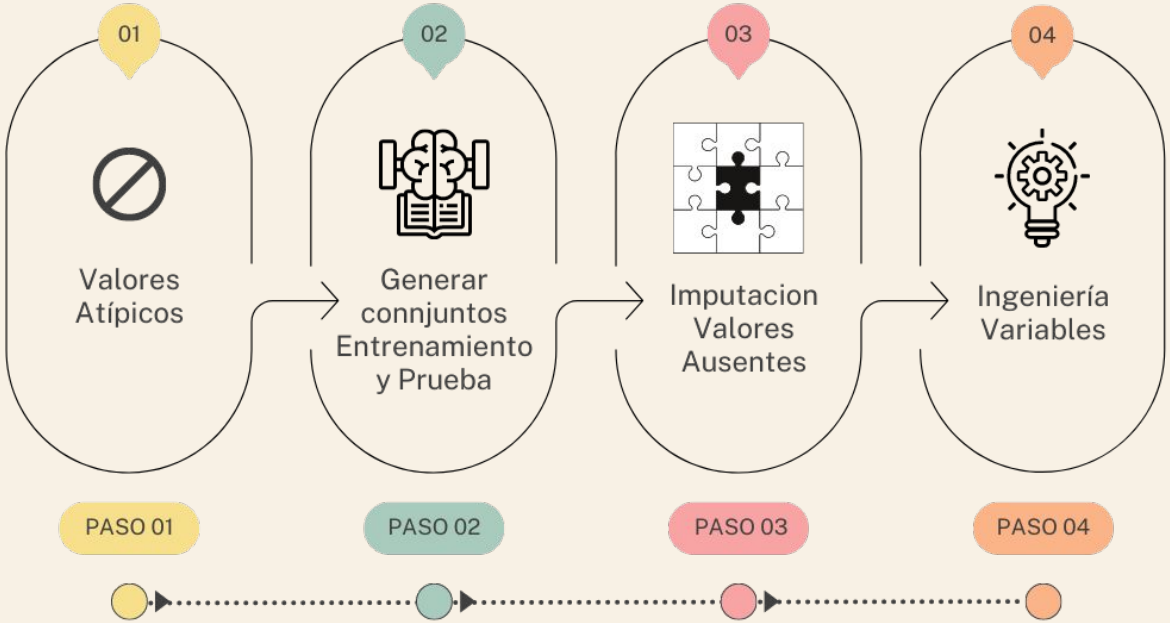




Sin registros de mujeres en este rango salarial

## CIENCIA TECNOLOGÍA Y GÉNERO : Limpieza de la información

Objetivo: Crear modelo que sea capaz de predecir el salario basado en las características del aplicante



## CIENCIA TECNOLOGÍA Y GÉNERO : Ingeniería de variables

La finalidad de esta sección es generar una representación numérica de las variables activables y seleccionar las mejores características que nos ayuden a predecir el salario potencial.

### CONTINUAS

| |
|---|
| Años Experiencia |
| Edad |
| Número de ascensos |
| Número de certificaciones |

### CATEGORICAS

#### Ordinales

| | |
|---|---|
| Nivel Tecnologias * | Nulo → 0<br>Bajo → 1<br>Medio → 2<br>Alto → 3 |
| Nivel Habilidades * | Nulo → 0<br>Bajo → 1<br>Medio → 2<br>Alto → 3 |

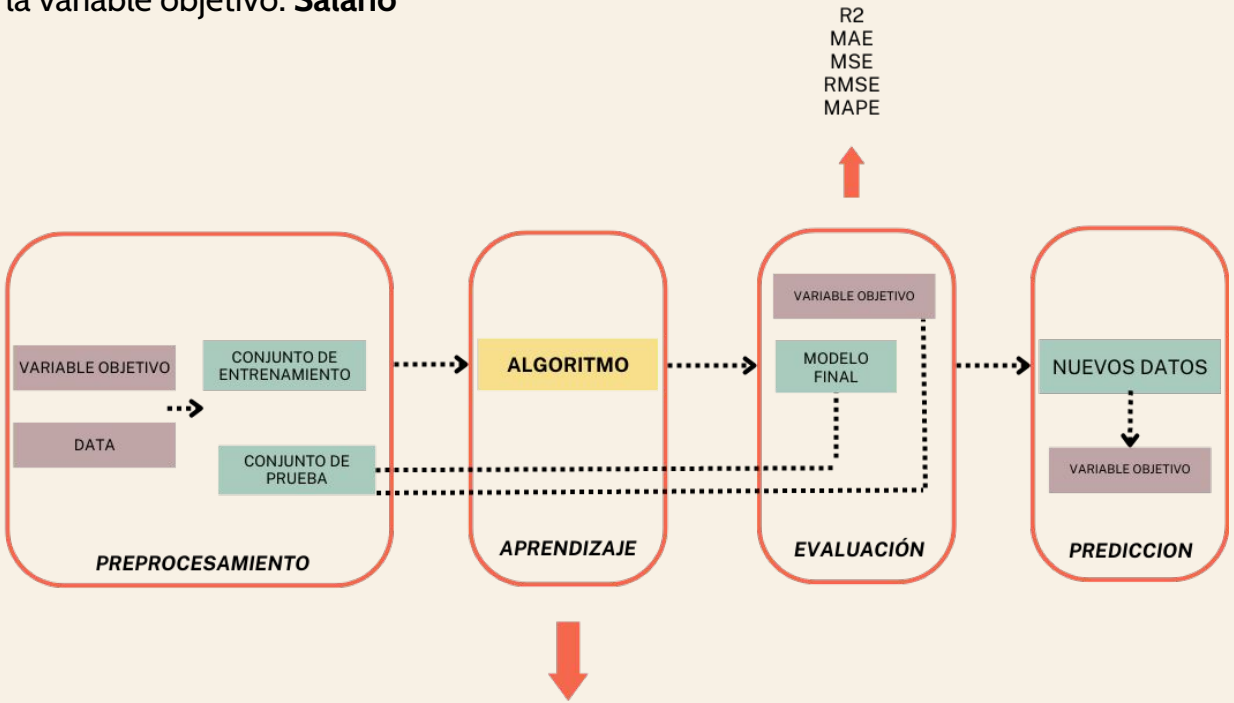
#### Nominales

| | |
|---|---|
| Posición | Data Scientist<br>Data Analyst<br>Data Engineer<br>Other |

* Habilidades: Liderazgo, Solución de conflictos, Gestión de proyectos, Inglés, Comunicación, Trabajo en Equipo
* Habilidades: Python, R, Tableau, Aws, SQL, Pytorch, PowerBI,GCP,Azure, VBA, Excel, Tensorflow, Looker

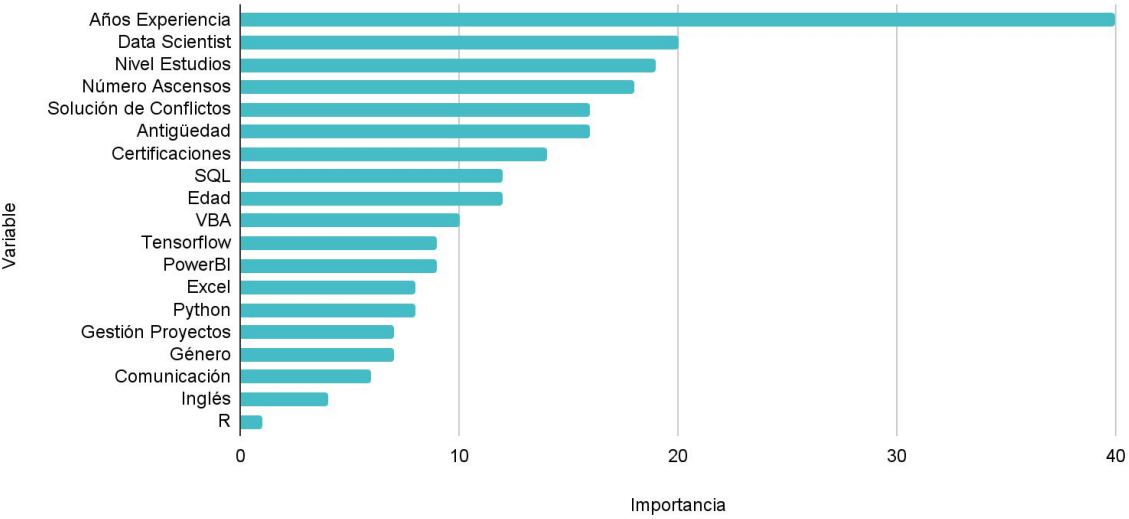
## CIENCIA TECNOLOGÍA Y GÉNERO : Modelado

En esta sección se lleva a cabo el entrenamiento del modelo que identificará las relaciones entre las variables creadas con anterioridad y la variable objetivo: **Salario**



## CIENCIA TECNOLOGÍA Y GÉNERO : Evaluación

Se determinaron las variables que de acuerdo a los resultados del mejor modelo obtenido tienen una mayor influencia en la predicción del salario, por ejemplo los años de experiencia y tener una posición como Científico de Datos son más determinantes en el resultado.




Número de entrenamientos : **14 Millones**

**Algoritmos:**

SVM
Lasso
LinearRegression
AdaBoost
XGBoost

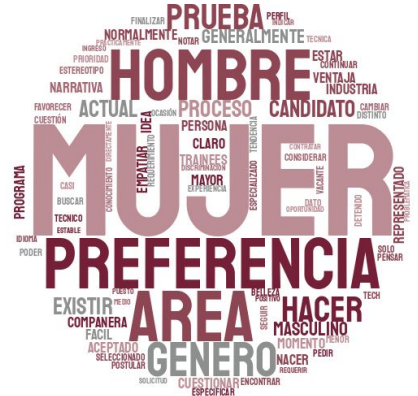
| Modelo | R2 Train | R2 Test | MSE Train | MSE Test |
|---|---|---|---|---|
| XGBoost | 78 | 75 | 122 | 187 |

# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

¿Algo que desearías haber sabido antes de comenzar tu carrera en el campo tech?



Respuestas de los candidatos que sintieron algún sesgo en la etapa de reclutamiento:


Comentarios de los usuarios que experimentaron un trato diferente con sus pares:


# CIENCIA TECNOLOGÍA Y GÉNERO : Análisis y Entendimiento de la Información

¿Cómo crees que se puede combatir los sesgos en la industria tecnológica?


¿Qué consejo le darías a las mujeres que ingresan al campo de la tecnología?


https://github.com/Malerva/CIMAT_2023_05/blob/main/Prediccion_Salario_Ejemplo.ipynb

# Conclusiones y Recomendaciones

- Extender la encuesta a un número más alto de profesionales en la industria para tener un mejor panorama de la situación actual.
- No forzar a los equipos a ser 50/50, realizar encuestas para conocer las condiciones en las que las personas trabajan, si se ha presentado algún aspecto de sesgo, distinción y/o tratos para tener medidas al respecto.
- Generar campañas de concientización sobre la desigualdad de las mujeres en la industria tech.
- Información en universidades y preparatorias para dar a conocer cuales son las labores del científico de datos.
- Impulsar el plan de carrera, continuar preparándonos para estar a la vanguardia de las nuevas tecnologías.