**Análisis Salarios Data Science**



**Alumno:**

Carlos Gabriel Leguizamon Guillaumet.

Federico Mastrascusa.

**Tutor:**

Franklin Zhunio.

Contenido

[**INTRODUCCIÓN** 3](#_Toc124156210)

[**MACHINE LEARNING** 4](#_Toc124156211)

[**ANÁLISIS Y RESULTADOS** 6](#_Toc124156212)

[Análisis Univariado 7](#_Toc124156213)

[Análisis bivariado 12](#_Toc124156214)

[**PREPARACIÓN DE LA DATA E IMPLEMENTACION DE MODELO** 14](#_Toc124156215)

[KNeighborsClassifier 15](#_Toc124156216)

[RandomForestClassifier 16](#_Toc124156217)

[DecisionTreeClassifier 17](#_Toc124156218)

[**Conclusión** 19](#_Toc124156219)



**INTRODUCCIÓN**

Estamos en un momento bisagra en el mercado laboral mundial, que obliga a las empresas a redefinir sus modelos de liderazgo, políticas de compensación, beneficios y modelos de trabajo.

Con la nueva aparición de los trabajos relacionados con el mundo IT en Latinoamérica también apareció la nueva camada de estudiantes que deciden tanto perfilar desde cero su carrera profesional al mundo de la tecnología, como también, estudiantes que deciden reinventarse y empezar a orientar su vida profesional a este nuevo mercado laboral.

Por suerte, no es necesario comenzar una carrera relacionada a la tecnología desde cero, existen ramas donde con un título orientado a otra rama laboral, ténganse en cuenta como, por ejemplo: Medicina, Matemáticas, Contabilidad, Biología, Actuarial, entre muchas otras, pueden aprender las herramientas que nos brindan el mundo de los datos para obtener un valor agregado en sus carreras profesionales, herramientas como, por ejemplo:

* Machine Learning: disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones.
* Softwares de visualización de datos: siendo estos utilizados para poder mostrar datos a través de gráficos como, por ejemplo, gráficos de barras, circulares, de dispersión, entre otros.
* Lenguajes de programación: Como lo son SQL, R y Python, el cual se utiliza para poder, escribir programas que permiten la comunicación usuario-máquina, que, junto con sus bibliotecas como Pandas, Numpy, Matplotlib, Seaborn, entre otras, nos ayudan a la hora de poder manipular datos.
* Storytelling: disciplina que se utiliza para describir como una narrativa de sucesos, que nos permite contar como una historia los datos, que si bien, no es un concepto que tiene un origen en los datos, nos permite poder explicar de una mejor manera los resultados de nuestros análisis.

Entre otras cuestiones, el mundo de los datos es un nicho relativamente nuevo por explotar, en el mundo que vivimos utilizamos los datos para poder realizar todo tipo de análisis y predicciones, por ende, ser amigo de los datos siempre nos puede resultar conveniente.

En base a esta introducción y perspectiva, he realizado el siguiente trabajo en torno a los salarios que las empresas pagan a sus trabajadores de Datos.



## **MACHINE LEARNING**

Es una rama de la inteligencia artificial, basada en la idea que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con una mínima intervención humana. Siendo así un método de análisis de datos que automatiza la construcción de “modelos analíticos”, que surgió en los años 50.

La principal característica de este tipo de algoritmo es que son capaces de reajustarse automáticamente para mejorar su rendimiento en función del número de aciertos y fallos producidos en un proceso de entrenamiento previo a su aplicación y durante la ejecución en tiempo real del mismo.

Machine Learning (ML) es una disciplina científica que maneja sistemas inteligentes, es decir que aprenden automáticamente al identificar ciertos patrones presentes en los datos. Para este aprendizaje ML usa algoritmos que se encargan de revisar datos mediante ejemplos o instrucciones predefinidas para así predecir comportamientos futuros permitiendo además la incorporación de información adicional y reajustar el resultado. ML maneja conocimiento inductivo obteniendo un enunciado general en base a enunciados que describen casos particulares (Mohri, Rostamizadeh, & Talwalkar, 2018).

El surgimiento del interés en el aprendizaje basado en máquinas se debe a:

* los volúmenes y variedad de datos disponibles
* el procesamiento computacional más económico y poderoso
* el almacenaje de dato asequible y mucho menos costoso
* y, todo hacen posible producir modelos de manera rápida y automática que pueda analizar datos grandes y complejos

Existen diferentes etapas, la primera es definir el objetivo. Es vital entender el problema a resolver y comprender la data que tenderemos a disposición. La segunda etapa es la de recolección de datos (data acquisition), de dónde y cómo conseguiremos la información. Como tercera etapa está la preparación de la data (se llama data wrangling): limpieza o formateo de datos que tiene como objetivo manipular y convertir la data en forma que produzcan mejores resultados. Algunas acciones pueden ser la categorización de los valores de las variables, normalizar datos numéricos o escalarlos para que sean compatibles.

La cuarta etapa es la selección del algoritmo más adecuado en relación con el problema que deseamos resolver, es decir el tipo de aprendizaje a implementar. Los algoritmos de ML se clasifican generalmente en:

* Aprendizaje supervisado: Su objetivo principal es predecir respuestas que habrá en el futuro, gracias al entrenamiento del algoritmo con datos conocidos del pasado (históricos). Pudiendo ser problemas de clasificación (necesitan predecir la clase más probable de un elemento en función de un conjunto de variables de entrada, siendo la variable target del tipo categórica) o de regresión (predicen valores numéricos, no categóricos, por lo cual la variable target es del tipo cuantitativa).



Son algoritmos entrenados utilizando ejemplos etiquetados, aprende comparando su resultado real con resultados correctos para encontrar errores. La salida de este algoritmo es conocida. Algunos ejemplos son predicciones, optimización de procesos, diagnósticos, detección de fraude y otros.

* Aprendizaje no supervisado: Son algoritmos que infieren patrones de un conjunto de datos sin referencia a resultados conocidos o etiquetados (es decir que sólo conocemos la variable x). Por lo cual, estos algoritmos permiten realizar tareas de procesamiento de datos más complejos en comparación al aprendizaje supervisado.

Sólo se dan los datos finales (inputs) a la máquina para que encuentre patrones interesantes a partir de esos datos. Todas las variables tienen la misma importancia, el objetivo es encontrar interdependencia entre las variables porque no hay una variable objetivos, o de salida, ni variables que ayuden a predecir la variable objetivo.

Dentro de los más populares están los de clustering o agrupamiento, reglas de asociación y algoritmos de reducción de la dimensionalidad, entre otros.

* Aprendizaje por refuerzo: Se utiliza con frecuencia para robótica, juegos y navegación, siendo un algoritmo que descubre a través de ensayo y error qué acciones producen las mayores recompensas. Tiene tres componentes principales: el agente (el que aprende o toma decisiones), el entorno (todo con lo que interactúa el agente) y acciones (lo que el agente puede hacer).

La quinta etapa es la de entrenar el modelo. El proceso de entrenamiento de un modelo de ML consiste en proporcionarle al modelo datos de entrenamiento de los cuales pueda aprender (training y test).

La sexta etapa es la validación del modelo; ya con el modelo entrenado se debe validar los resultados obtenidos según métricas utilizada para los diferentes tipos de algoritmos.

Y, por último, la séptima etapa es el deployment o despliegue del modelo, el cual consiste en la implementación en producción de nuestro modelo.



## **ANÁLISIS Y RESULTADOS**

El dataset elegido fue Data Science Job Salaries , a través de Kaggle.com ([link](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries)) cuenta con 607 filas y 11 columnas

Una vez definido el dataset comenzamos con el análisis

Se trata de una base de datos que extrae salarios de los trabajadores con empleos relacionados a la rama de los datos.

Información de atributos.

1) *work\_year* = Año en el que el salario fue pagado

2) *experience\_level* = El nivel de experiencia en el puesto durante el año con los siguientes valores posibles: EN Entry-level/Junior MI Mid-level/Intermediate SE Senior-level/Experto EX Executive-level/Director

3) *employment\_type* = El tipo de empleo para el puesto: PT Tiempo parcial FT Tiempo completo CT Contrato FL Freelance

4) *job\_title* = Rol trabajado durante el año

5) *salary* = El monto total del salario bruto pagado.

6) *salary\_currency* = La moneda del salario pagado como un código de moneda ISO 4217

7) *salary\_in\_usd* = El salario en USD (tasa de cambio dividida por la tasa promedio de USD para el año respectivo a través de fxdata.foorilla.com)

8) *employee\_residence* = El país de residencia principal del empleado durante el año laboral como código de país ISO 3166.

9) *remote\_ratio* = La cantidad total de trabajo realizado de forma remota, los valores posibles son los siguientes: 0 Sin trabajo remoto (menos del 20 %) 50 Parcialmente remoto 100 Totalmente remoto (más del 80 %)

10) *company\_location* = El país de la oficina principal del empleador o sucursal contratante como un código de país ISO 3166.

11) *company\_size* = Número promedio de personas que trabajaron para la empresa durante el año: S menos de 50 empleados (pequeño) M 50 a 250 empleados (mediano) L más de 250 empleados (grande)

Primero que nada, cargamos el dataset a una cuenta de GITHUB para poder realizar la importación desde allí

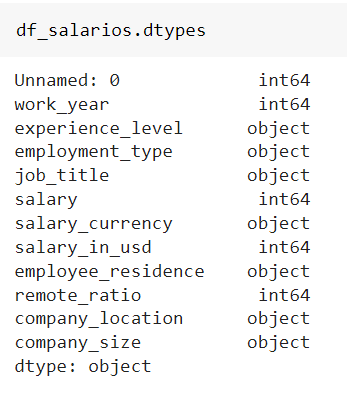
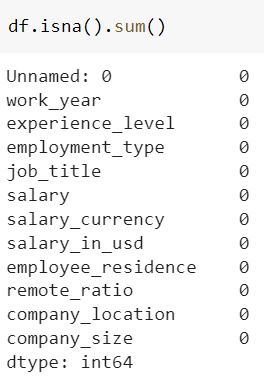
Luego importamos las librerías necesarias para la importación





Una vez importado el Dataset, comenzamos una exploración del mismo, definimos un Data Frame para poder contar los nulos, como podemos ver en el código y resultado, no tenemos ningún valor nulo en los atributos.

También, podemos visualizar el tipo de dato que se basa cada uno de los atributos.



### Análisis Univariado

En esta sección, exploraremos cada columna en el conjunto de datos para ver las distribuciones de características, principalmente dos partes: Análisis en columnas categóricas y análisis en columnas numéricas.

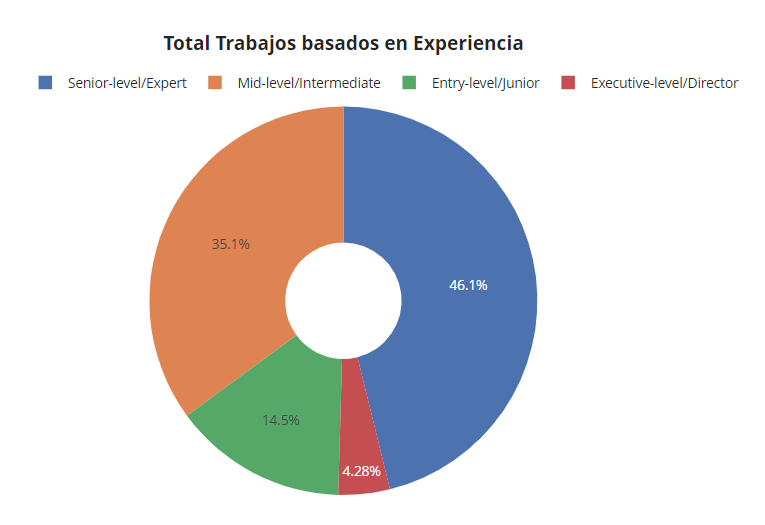
Variables Categóricas

#### **6 columnas categóricas:**

* Experience Level
* Job Titles
* Employment Type
* Employee Residence
* Company Location
* Company Size.



**Experience Level**: El nivel de experiencia en el puesto durante el año con los siguientes valores posibles: EN Entry-level/Junior MI Mid-level/Intermediate SE Senior-level/Experto EX Executive-level/Director

El cuales tenemos para

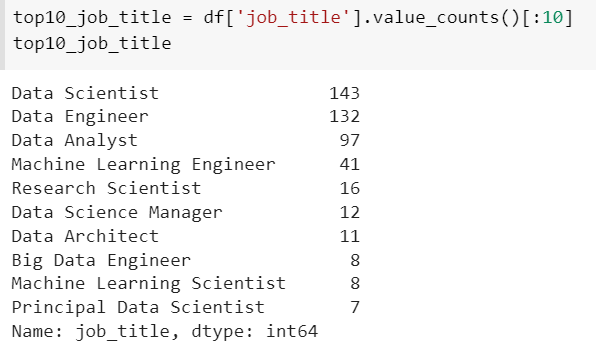
Executive level/Director: 26 datos

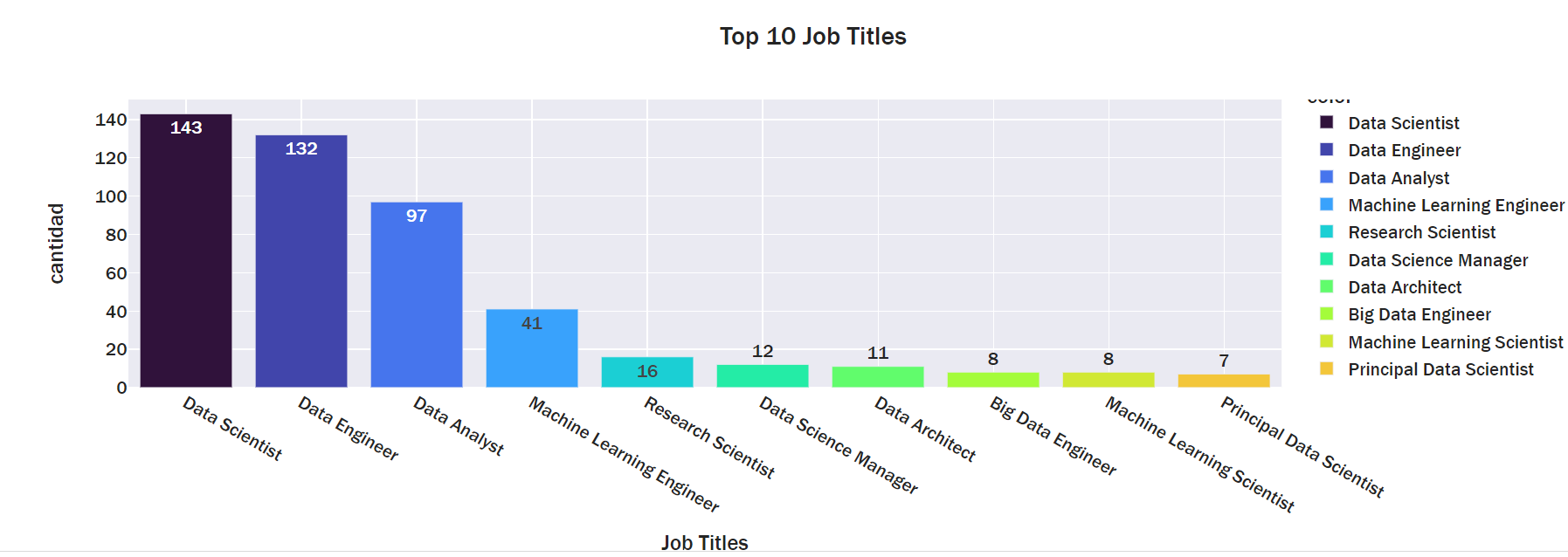
Senior level/Experto: 280 datos

MId level/Intermediate: 213 datos

Entry level/Junior: 88 datos

**Job Title**: Los títulos que tienen los trabajadores, en este dataset contamos con 50 títulos de trabajo distintos, por ello trabajamos solo con el TOP 10

. 

Podemos hacer una mejor visualización con un gráfico de barras de la biblioteca Matplotlib.



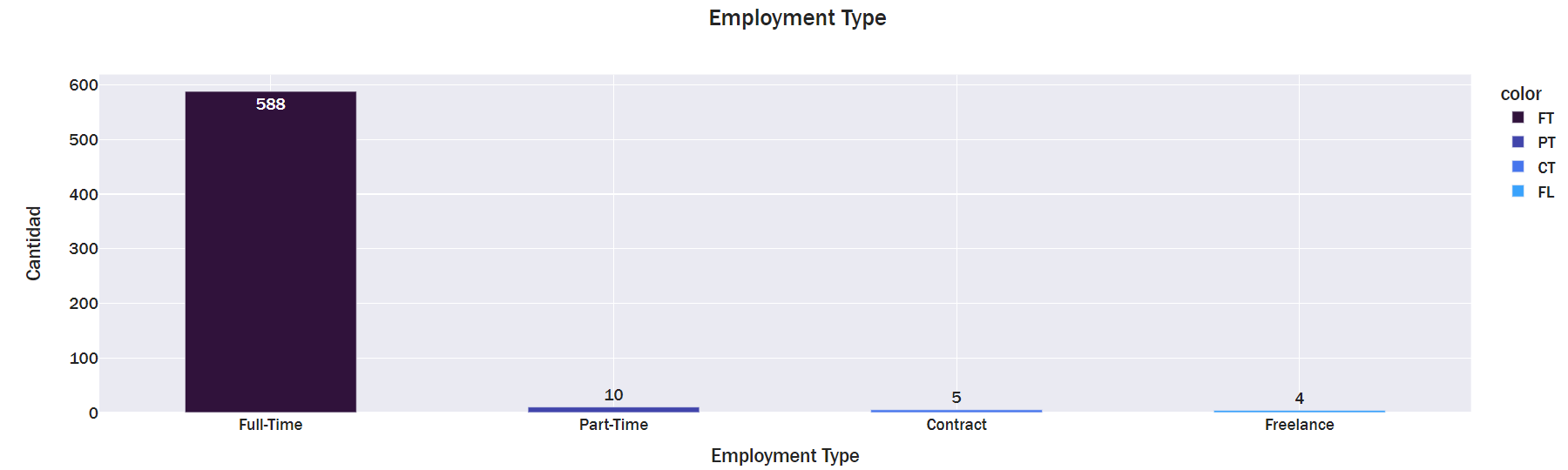
**Employment Type:** Cuatro tipos de empleo,

PT: Part Time – 10 datos

FT: Full Time – 588 datos

CT: Contrato – 5 datos

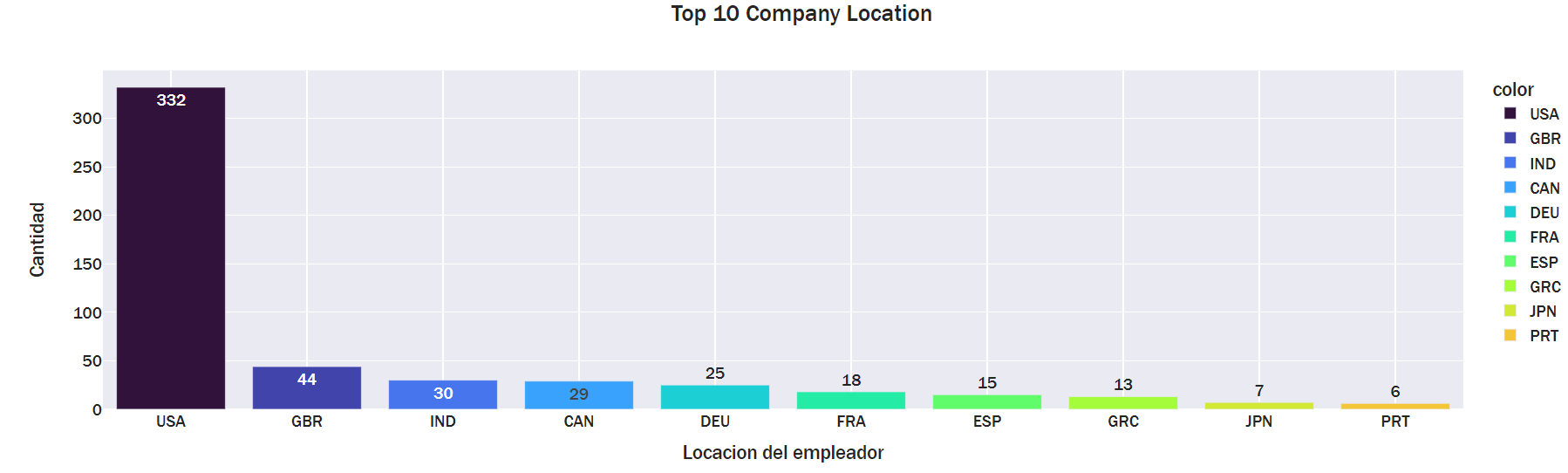
FL: Freelance – 4 datos

****

**Employee Residence:** El país de residencia principal del empleado durante el año.

**Company Location:** El país de la oficina principal del empleador o sucursal contratante.

Como tenemos datos de empresas de todo el mundo, para una mejor visualización tomamos el TOP 10



El cual nos muestra que la mayoría de los empleadores se encuentran en estados unidos ‘USA’

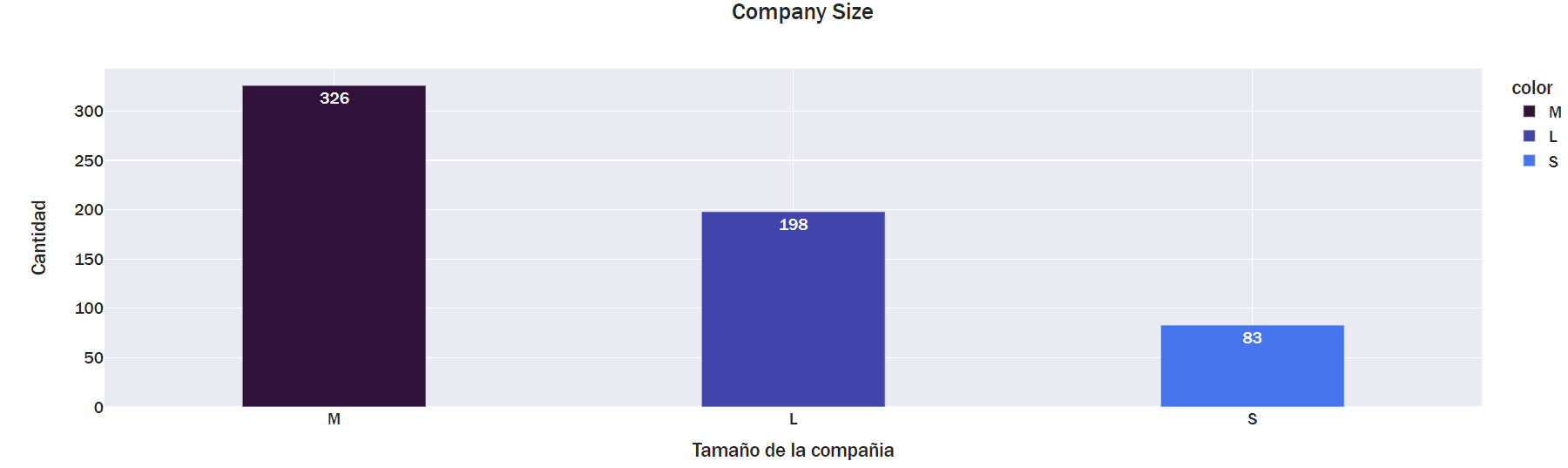


**Company Size**: Por último, el tamaño de la compañía.

L: Grande

M: Mediana

S: Pequeña



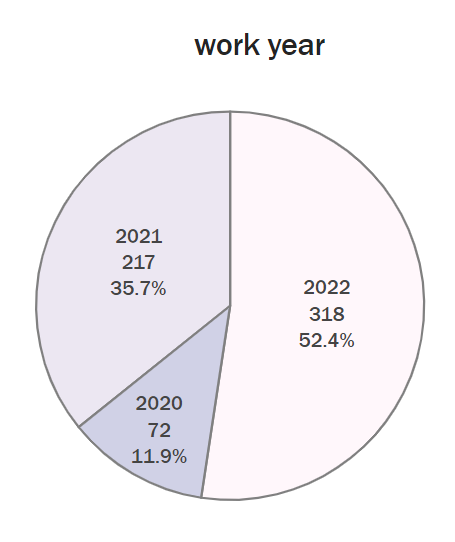
Variables Numéricas

**3 columnas numéricas**

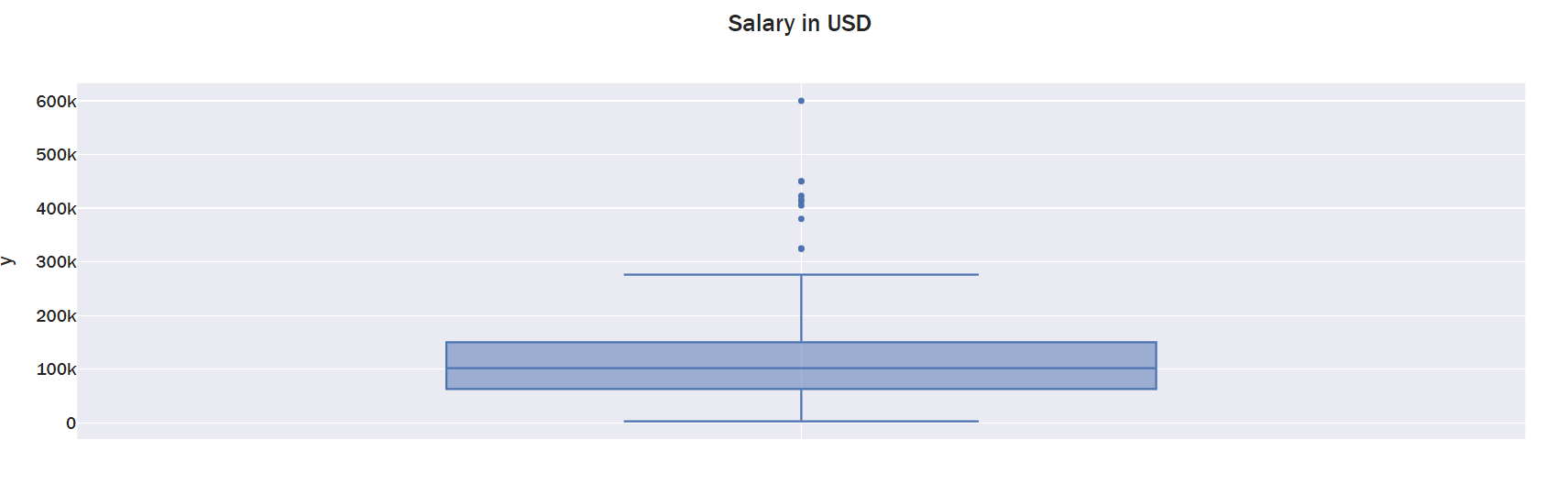
* Work\_year
* Salary\_in\_usd
* Remote\_ratio



**Work\_year**: es el año que se recopilo los datos, tenemos datos del 2020, 2021 y 2022



**Salary\_in\_usd**: el salario visto en DOLARES, si bien, tenemos datos de salarios de muchas monedas, se convirtió todo a USD para poder tener una mejor comparación



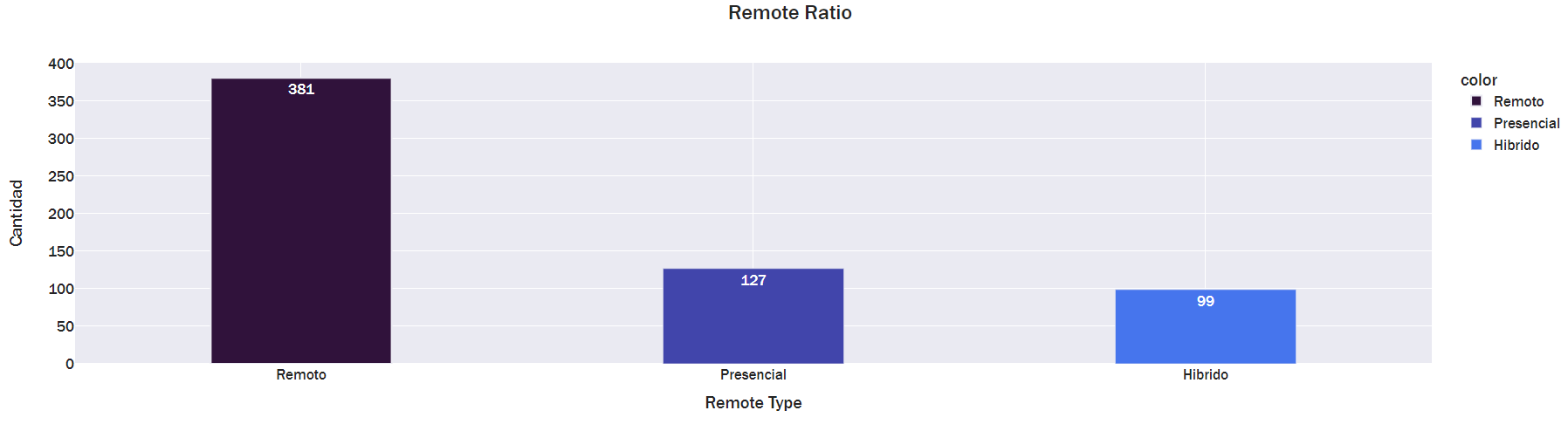


**Remote\_ratio** : La variable busca informar, el tipo de trabajo, siendo remoto, presencial o hibrido

0% Remote: Trabajo presencial –127datos

50% Remote: Trabajo hibrido, 50% presencial y 50% home office – 99 datos

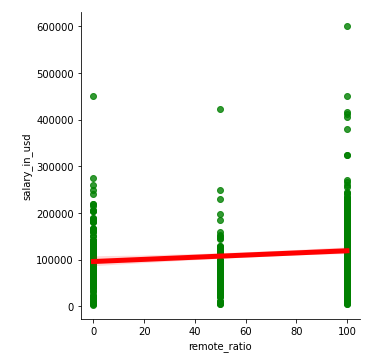
100% Remote: Trabajo home office – 381 datos



### Análisis bivariado

Para primer análisis, vamos a tener en cuenta las variables “salary\_in\_usd” y el modelo de trabajo que tienen “remote\_ratio”

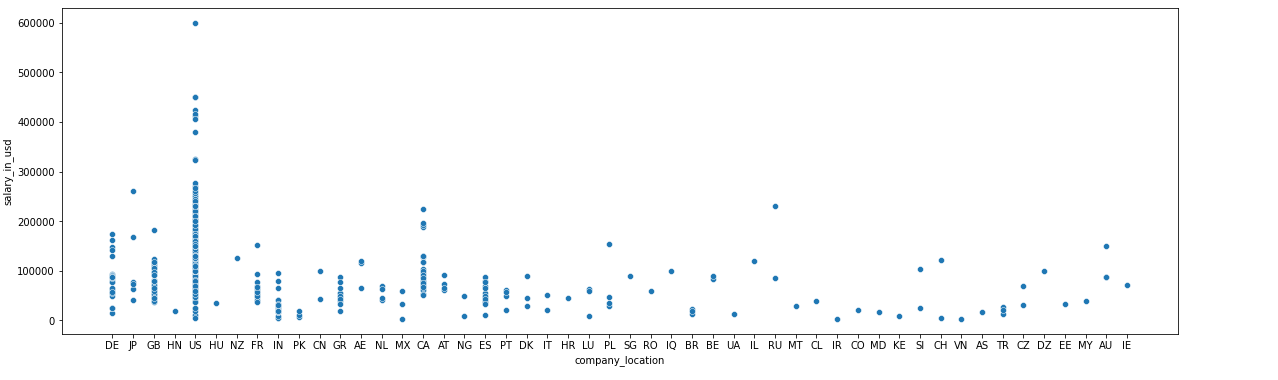
Verificamos si hay una regresión lineal entre el salario que cobran con el modelo de trabajo que tienen



Se podría decir que si, como podemos visualizar, los trabajos 100% remotos tienen mejor salario, habiendo un dato que nos muestra que pueden llegar hasta $600000

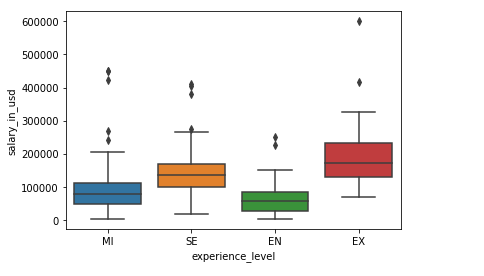


Para el segundo análisis se tuvo en cuenta las variables “Salary\_in\_usd” y la locación de la compañía “Company\_location” con un gráfico de dispersión.



Se visualiza que en el país US=Estados Unidos, hay mayor salario, se podría decir que es por varios factores como, por ejemplo, escases de datos sobre salarios de compañías de otros países, o también porque en US hay más compañías que utilizan este tipo de nuevos trabajos como los que vendrían ser los relacionados a Data Science.

Y por último, en el tercer análisis se utilizó las variables “Experience\_level” y “salary\_in\_usd” para poder verificar si existe una relación entre la experiencia del empleado con el salario que se le paga, mediante un gráfico de caja y bigotes.



En este gráfico de Caja y Bigotes se puede apreciar que según la experiencia que tiene el trabajador, puede llegar a tener un mayor salario, como lo vemos en los trabajadores con experiencia EX = Nivel directivo, llegan a tener una mediana mayor a la mediana de los otros



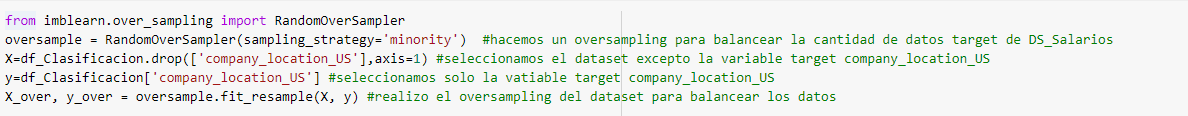
salarios, y también podemos visualizar que la mediana de los empleados con experiencia EN = entry/juniors, tiene una mediana que es menor a todas las demás medianas.

## **PREPARACIÓN DE LA DATA E IMPLEMENTACION DE MODELO**

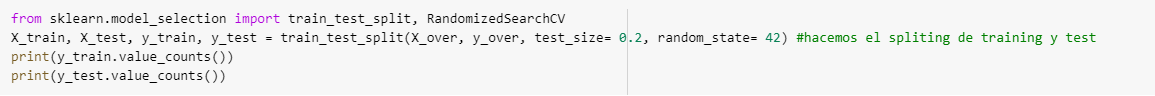
Por medio del método método .get\_dummies creamos la variable df\_clasificacion asignamos valores numéricos a columnas categóricas

En base a eso, utilizamos la localización de la compañía (en este caso de USA) para poder entrenar la máquina, quedando la variable utilizada como “Company\_location\_US”

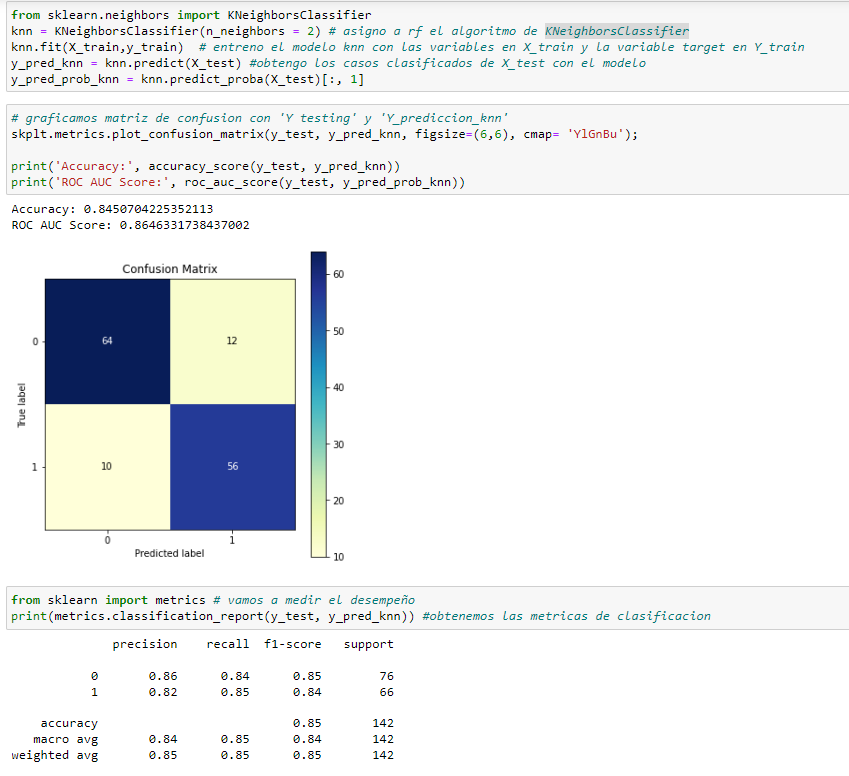
Realizamos un OverSampling para balancear los datos y poder usarlos correctamente



Realizamos el Split de la base de entrenamiento y test, asignando también la Y (nuestra variable objetivo). Entrenamos varios modelos, y comparamos los resultados.



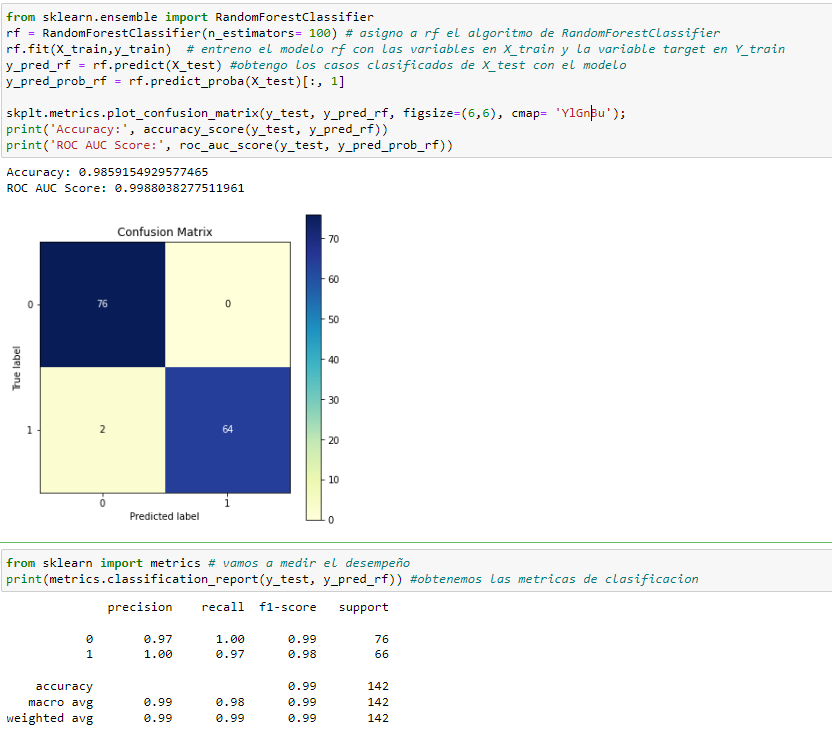
Luego, entrenamos los modelos KNeighborsClassifier, RandomForestClassifier y DecisionTreeClassifier con los datos asignados y medimos su desempeño.



### KNeighborsClassifier

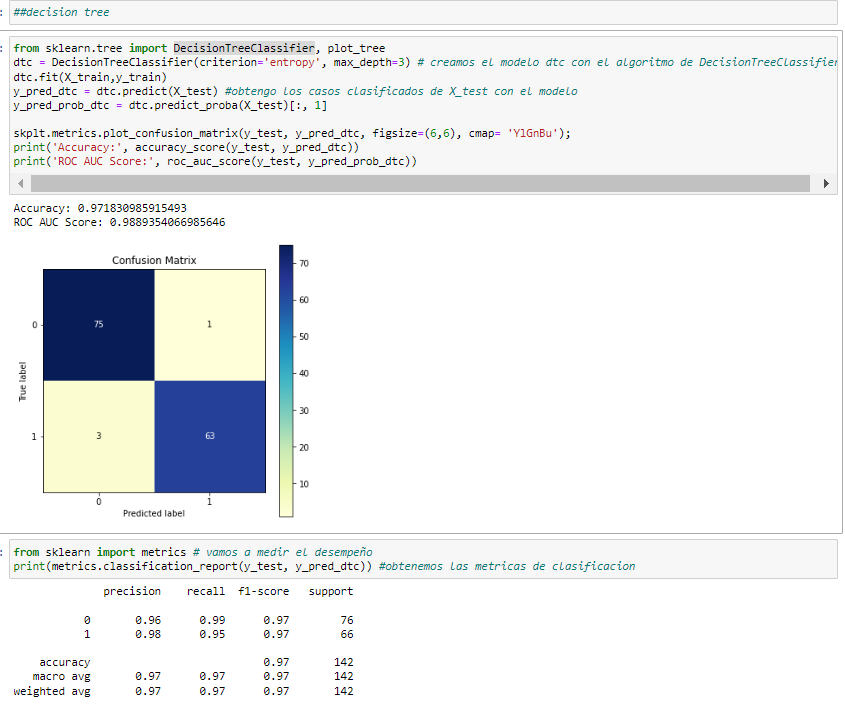


### RandomForestClassifier



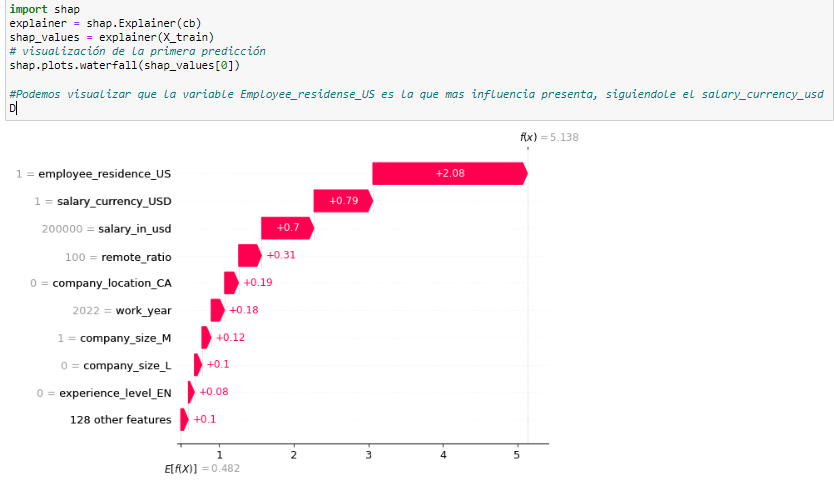
##### 

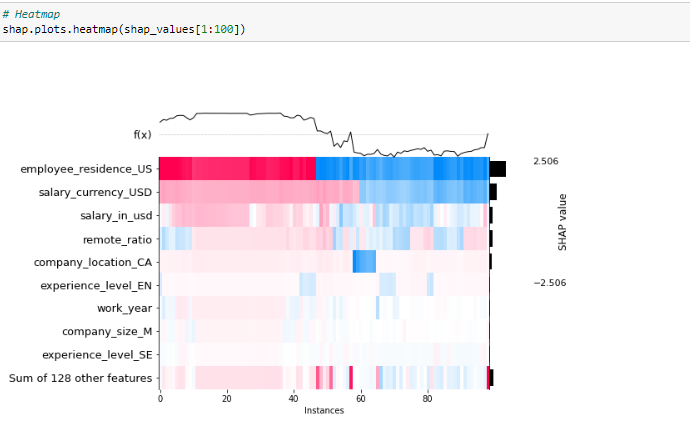
### DecisionTreeClassifier





Por último, con SHAP visualizamos cuales fueron las variables más influyentes





Siendo en esta predicción, la variable Employee\_residence\_US (Residentes de USA) variable que más influyó en la predicción.



## **Conclusión**

Del análisis anterior, el país que tiene la cantidad máxima de trabajos en el campo de la ciencia de datos es EE.UU. y los empleados de tipo contrato tienen salarios máximos en este campo. Los empleados que tienen experiencia senior y ejecutivo tienen salarios máximos mientras que los Juniors o entry level presentan los salarios más bajos.

Los 5 roles más populares del dataset en la ciencia de datos son:

1 Data Scientist

2 Data Engineer

3 Data Analyst

4 ML Engineer

5 Research Analyst

Es claro que el crecimiento de estos nuevos trabajos es exponencial, los datos son el nuevo petróleo de las nuevas generaciones, ayudan a los negocios a poder llegar a más personas y a poder realizar análisis cada vez más complejos para poder satisfacer necesidades de manera eficiente.

Estas nuevas herramientas de análisis de datos pueden ayudar a abaratar costos al largo plazo.

Es intrigante la idea de que será el día de mañana con los datos, cuáles serán los nuevos puestos que saldrán a la luz y cuáles serán las nuevas herramientas que nos van ayudar a poder entender cada vez más a cómo funciona la sociedad.

Como toda persona curiosa, espero con ansias el poder analizar los resultados del futuro.

Muchas gracias.