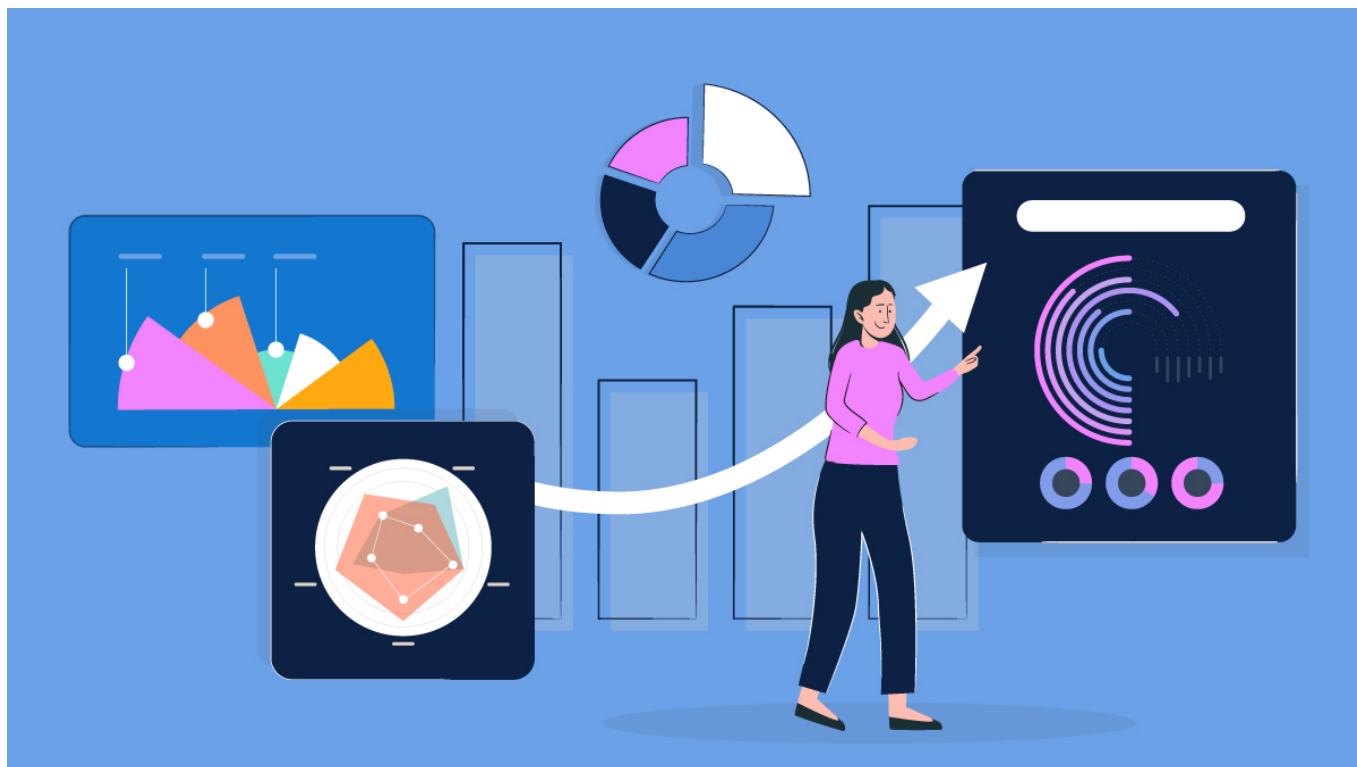


VISUALIZACIÓN DE DATOS

Trabajo Final: Periodismo de Datos (Etapa 2)



Entrega Equipo 5

Diego Rosales León	201810531-7
Alejandro Moyano Mejías	201873598-1
Benjamín Bello Gamboa	201873537-k

Descripción

Para este trabajo se escogió el paper “An overview and comparison of free Python libraries for data mining and big data analysis” [1], en el cual se realiza una comparación entre 27 librerías gratuitas para el lenguaje de programación Python, las cuales son bastante usadas en el contexto de minería de datos y análisis de big data. Para realizar esta comparación, lo hacen mediante su popularidad y contribución en GitHub, para luego categorizarlas en 6 grupos: librerías core, preparación de datos, visualización de datos, machine learning, deep learning y big data. Finalmente, se comparan las diversas funcionalidades que ofrece cada librería según su categoría.

En la categoría de machine learning, se comparan 3 librerías: scikit-learn, mlxtend y Shogun. Donde se contrastan las funciones disponibles para selección de características, transformación de características, aprendiz del árbol de decisión, clasificadores bayesianos, clasificación basada en funciones, aprendizaje basado en instancias, analisis de regresion, ANN, SVM, aprendizaje conjunto, agrupación jerárquica, agrupación centroide (partición), agrupación basada en distribución, agrupación basada en densidad, reglas de asociación (sin supervisión), métodos de evaluación y métricas.

Para estas comparativas los autores nos presentan diferentes tablas, como una tabla resumen de las cantidades de cada librería en GitHub, otras para las categorías con más funciones y librerías indicando si las librerías contrastadas tienen las funciones o no. Como grupo, creemos que la inexistencia de gráficos con colores y/o áreas, restringe la información de la información bruta de las tablas, ya que si se utilizarán estos, la comprensión de la información por las personas que lo lean, sería más entendible.

Perfiles de usuarios

Para definir los perfiles de usuarios, se utilizarán personas que estén relacionadas al área del Big Data y Data Mining, como estudiantes, juniors y seniors de diferentes áreas:




Perfiles de usuario			
Rol	Data scientist Senior 	Estudiante Ing. Civil Informática 	Data Analytics Junior 
Nombre	Oliver Atom	Hannah Montana	Martin Torres
Edad	37	24	30
Educación	Profesional con magíster y doctorado	Futura profesional	Profesional en el área de Analytics, y profesor universitario
Contexto Familiar	Padre	Hija única	Hijo menor
Estado civil	Casado con hijos	Soltera	Soltero
Domicilio	Valparaíso	Santiago	Concepción
Objetivos y metas	Obtener una oferta laboral de una empresa fuera del país	Obtener título universitario	Obtener un ascenso para convertirse en Data Analytics Senior
	Dar el mejor futuro posible para sus hijos	Cumplir con las expectativas de los padres	Ser un buen orientador para sus alumnos
	Ayudar a sus hijos a plantearse metas	Tener una profesión para ayudar a los padres en el hogar	miro

Tabla 1: Usuarios creados para los datos

El usuario elegido fue la Estudiante de Ing. Civil Informática, Hannah Montana. Toda la información del usuario puede ser encontrada en [Miro](#).

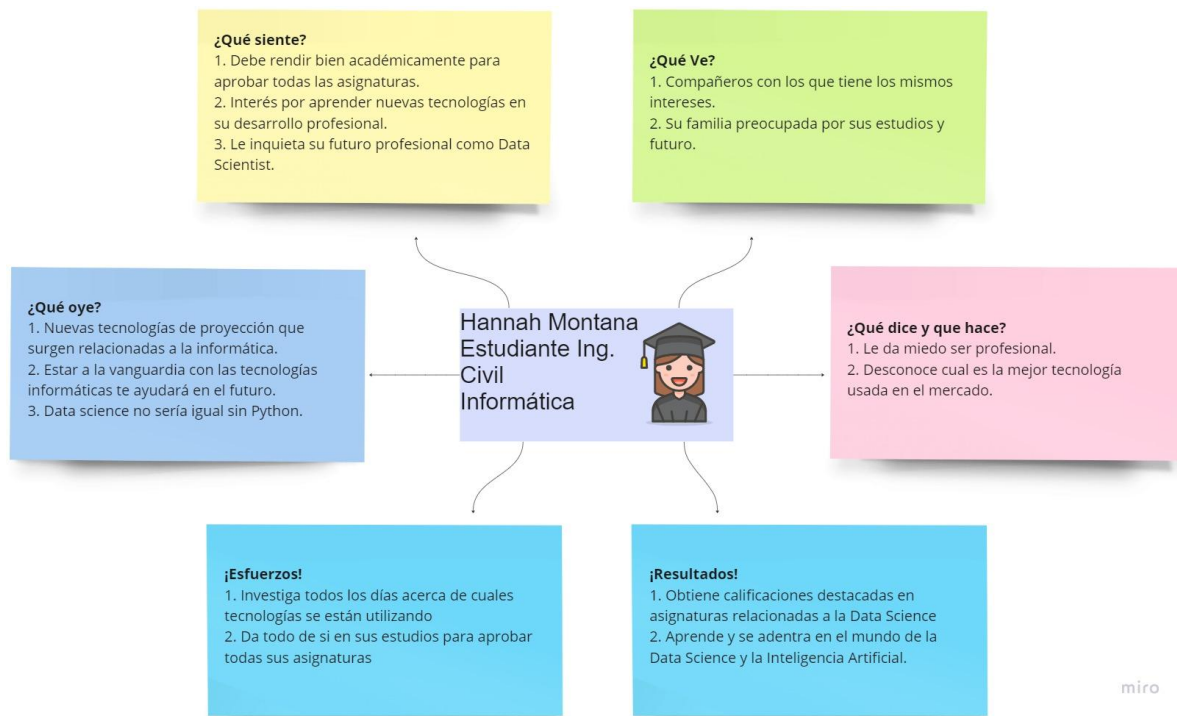
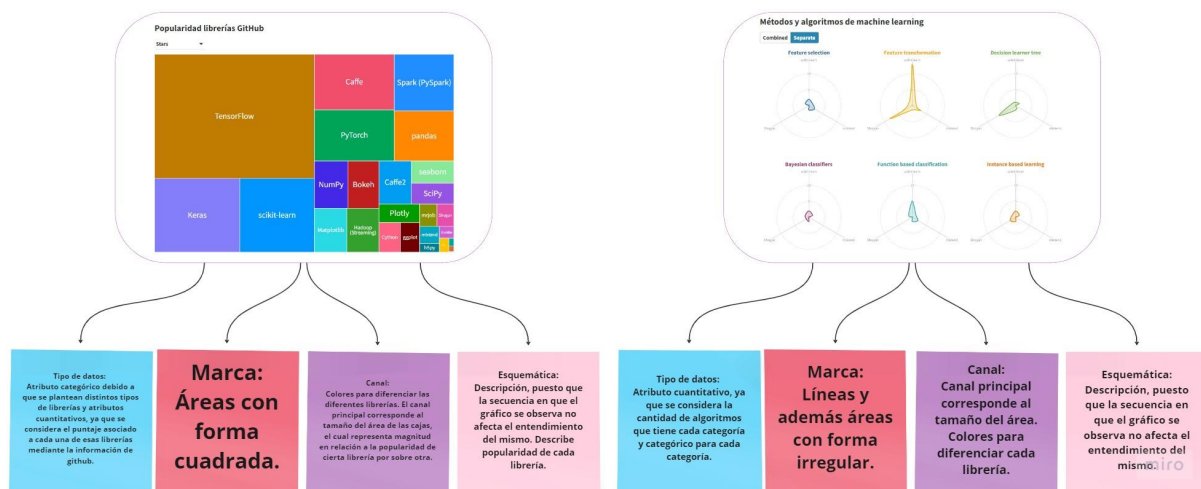


Figura 1: Usuaría elegida.

Prototipos de visualización

Se construirán tres prototipos de visualización de datos para poder complementar la información del paper elegido [1]. El paper será complementado con diferentes tipos de prototipos (con dos gráficos diferentes cada uno) con los valores provenientes del paper [2] (Los prototipos pueden ser observados de mejor forma en [Miro](#))

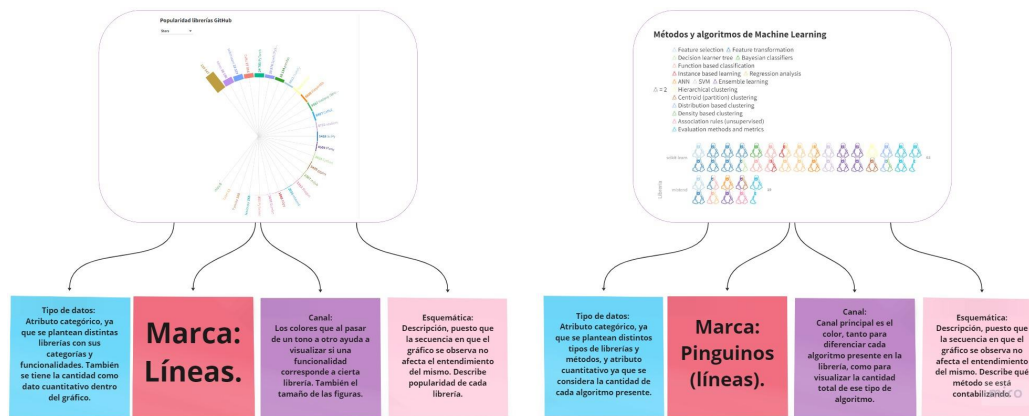
Prototipo 1



Lenguaje Visual			
Color	Paleta de colores	Dado que existen varios datos, y como es necesario hacer una comparativa entre ellos, fue necesario utilizar una paleta de colores en donde pueda distinguirse cada elemento de manera clara.	
Texturas	Lisa	No existió la necesidad de utilizar un tipo de textura que no fuera lisa. La textura lisa beneficia a la estética que tendrán los gráficos.	
Tipografía	Source Sans Pro	Es una letra legible y que se utiliza mucho en la creación de diseños, por lo que es una muy buena elección al usarla con los gráficos.	Source Sans Pro
Líneas	Líneas de 0,2 px	Para el gráfico 2, con el fin de que pueda visualizarse de manera correcta las formas irregulares de los elementos, fue necesario que cada forma tuviera un ancho de línea de 0,2 px.	 miro

Prototipo 1 con sus gráficos, fichas técnicas respectivas y lenguaje visual.

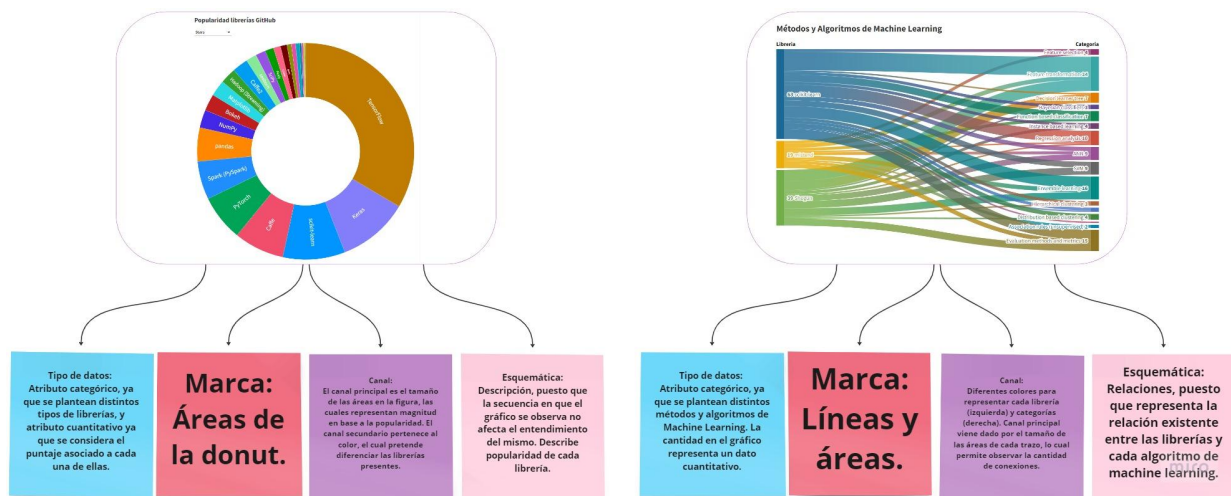
Prototipo 2



Lenguaje Visual			
Color	Paleta de Colores emparejados	Puesto que existen varios datos y como es necesario hacer una comparativa entre ellos, fue necesario utilizar una paleta de colores en donde pueda distinguirse cada elemento de manera clara, en este caso, una paleta de colores emparejados.	
Texturas	Lisa	No existió la necesidad de utilizar un tipo de textura que no fuera lisa. La textura lisa beneficia a la estética que tendrán los gráficos.	
Tipografía	Source Sans Pro	Es una letra legible y que se utiliza mucho en la creación de diseños, por lo que es una muy buena elección al usarla con los gráficos.	Source Sans Pro
Líneas	Líneas con un máximo de 40 px	Para el gráfico 1, con el fin de que pueda visualizarse de mejor manera la comparativa entre cada elemento, fue necesario que cada barra (línea) tuviera un ancho de 40 px.	 miro

Prototipo 2 con sus gráficos, fichas técnicas respectivas y lenguaje visual.

Prototipo 3



Lenguaje Visual			
Color	Paleta de Colores	Al igual que el prototipo anterior, puesto que existen varios datos, y como es necesario hacer una comparativa entre ellos, fue necesario utilizar una paleta de colores en donde pueda distinguirse cada elemento de manera clara.	
Texturas	Lisa	No existió la necesidad de utilizar un tipo de textura que no fuera lisa. La textura lisa beneficia a la estética que tendrán los gráficos.	
Tipografía	Source Sans Pro	Es una letra legible y que se utiliza mucho en la creación de diseños, por lo que es una muy buena elección al usarla con los gráficos.	
Líneas	Líneas de 0,1px a un máximo de 12px	En el gráfico dos, es necesario que las líneas sean gruesas y delgadas, permitiendo que la información sea mucho más apreciable.	

Prototipo 3 con sus gráficos, fichas técnicas respectivas y lenguaje visual.

Comparativa Gráficos

Gráfico 1

Para la creación del gráfico 1 en los tres prototipos se ocupó la tabla de popularidad en librerías GitHub. Viendo el lado favorable de la visualización en el prototipo 2, es la utilización del canal del tamaño (altura) de las figuras para mostrar los valores, lo cual es un poco más sencillo de visualizar que el área de las figuras en el caso de los prototipos 1 y 3. Además, en el prototipo 2, se muestra la cantidad exacta de la popularidad del dato, lo cual en los prototipos 1 y 3, esto no se cumple. Por otro lado, tanto el prototipo 1 como el 3 tienen la característica de tener un orden descendente en la popularidad, que en el caso del prototipo 1 es en diagonal hacia abajo, mientras que, en el prototipo 3, es en el sentido de las manecillas del reloj, lo que ayuda a observar de mejor manera la información.

Gráfico 2

En este caso, el gráfico del prototipo 3 tiene un punto a favor con respecto a los otros prototipos, ya que muestra en detalle cada una de las relaciones y la cantidad de tipos de librerías que hay, lo cual no ocurre en los prototipos 1 y 2, ya que en el primero solo ayuda a comprender las diferencias relativas entre los elementos de sus datos, y en el segundo solo muestra el número total de cada librería, perdiéndose información. En resumen, el prototipo 3 resulta ser más intuitivo. Caso contrario puede ocurrir con respecto al gráfico del prototipo 2, puesto que se debe considerar la mitad de cada pingüino para la cantidad, esto puede provocar una mezcla de muchos colores y generar confusión para la vista.

El prototipo 1 es bueno para entender los gráficos de manera visualmente atractiva, pero pueden ser difíciles de entender, debido a que no se puede visualizar la diferencia de cuán grande es un valor por sobre el otro.

El prototipo 2 es muy bueno para mostrar las cantidades exactas de la información entregada, pero aunque sea muy bueno en eso, pierde mucha información de las diferencias relativas entre los elementos.

El prototipo 3, aunque tenga una dificultad al verlo al principio, es bueno para entregar la información de manera más fácil versus los otros prototipos, pero una de las dificultades que presenta es que se pierde la información a la vista (ya que presentan mucha más información de lo que se está mostrando).

Visualización Final

Luego de la comparación, se decidió por quedarse con el gráfico 1 del prototipo 1, el gráfico 2 del prototipo 2 y el gráfico 2 del prototipo 3.

Se llegó a esta decisión ya que el grupo prefirió optar, en el caso del prototipo 1, por un gráfico más sencillo, relativamente fácil de visualizar y también más manipulable en el tema de la comparación de datos. Para el prototipo 2, se decidió utilizar tal gráfico por el hecho de que nos parecía más interesante o atractivo visualmente, cosa que no ocurría con el gráfico 1 de tal prototipo, que además podía parecer muy básico. Finalmente, para el prototipo 3, se optó por tal gráfico por el hecho de que se mostraban de una manera adecuada las relaciones entre los datos y era más atractivo visualmente, cosa que no ocurría con el gráfico 1.



Gráfico 1 del prototipo 1 con su ficha técnica respectiva y lenguaje visual.

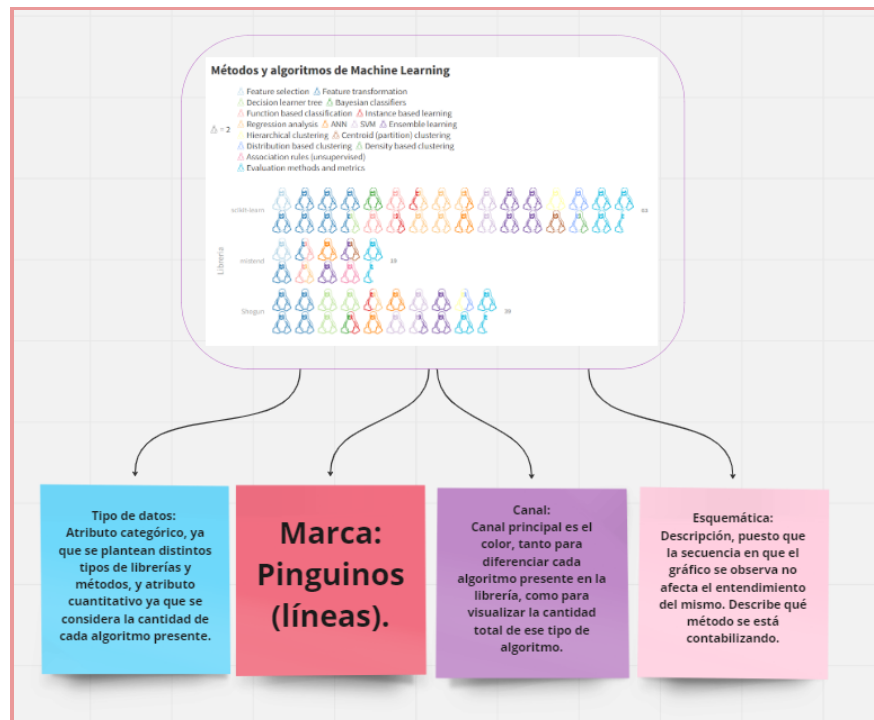


Gráfico 2 del prototipo 2 con su ficha técnica respectiva y lenguaje visual.

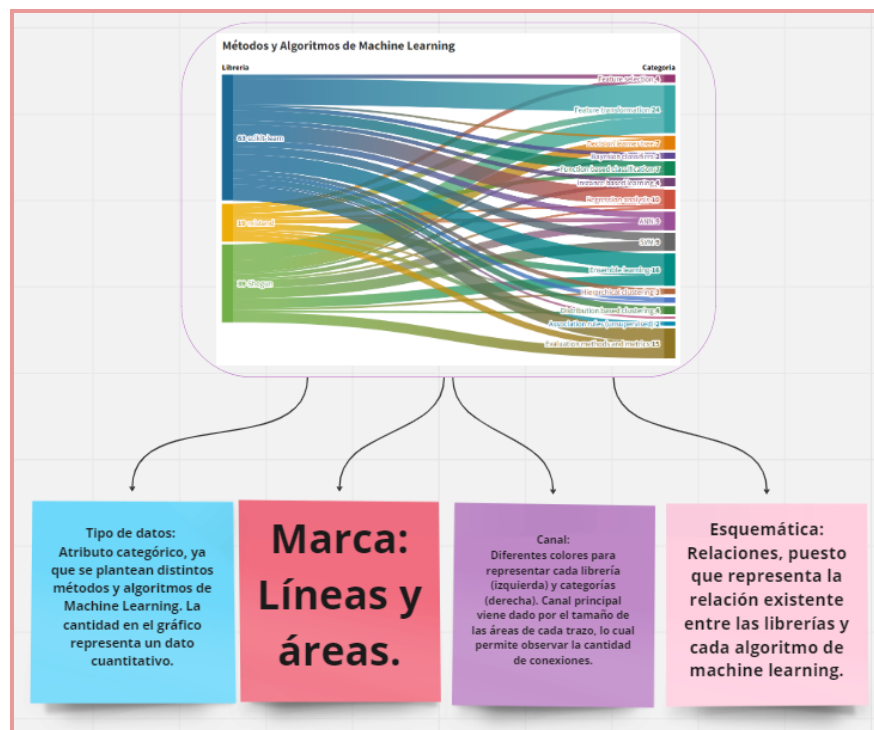



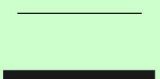


Gráfico 2 del prototipo 3 con su ficha técnica respectiva y lenguaje visual.

Lenguaje Visual			
Color	Paleta de Colores	Dado que todos los gráficos son distintos y tienen diferentes tipos de datos, fue necesario utilizar una paleta de colores en donde pueda distinguirse cada elemento de manera clara.	
Texturas	Lisa	No existió la necesidad de utilizar un tipo de textura que no fuera lisa. La textura lisa beneficia a la estética que tendrán los gráficos.	
Tipografía	Source Sans Pro	Es una letra legible y que se utiliza mucho en la creación de diseños, por lo que es una muy buena elección al usarla con los gráficos.	
Líneas	Líneas de 0,1px a un máximo de 12px	En el gráfico tres presentado, es necesario que las líneas sean gruesas y delgadas, permitiendo que la información sea mucho más apreciable.	

Infografía con todos los gráficos elegidos con su respectivo lenguaje visual.

Valor Agregado

El paper utilizado [1] presenta toda su información a través de mucho texto y el uso de tablas que son necesarias cuando el autor requiere de comparar las diferentes librerías propuestas, estas tablas son difíciles de entender en una primera instancia, debido a que presentan texto, valores numéricos y valores de aceptación o negación respectivamente. Estas cualidades suponen un problema para los lectores que quieren entender y aprender de la información que es planteada por el autor a través de todas estas comparativas.

Con nuestra propuesta para presentar toda la información de las tablas a través de diferentes prototipos y de gráficos, las cualidades que anteriormente eran un problema, se ven solucionadas, esto se debe principalmente a que le estamos agregando un canal visual, colores, formas y distintos tamaños a todo el paper, lo cual permite atraer visualmente al lector, facilitando su comprensión y aprendizaje de la información planteada de una manera más intuitiva según corresponda a cada gráfico.

Referencias

- [1] Stančin, I., & Jović, A. (2019, May). An overview and comparison of free Python libraries for data mining and big data analysis. In 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 977-982). IEEE.
- [2] Datos provenientes del paper y utilizados en los gráficos. (2022). Google Docs. <https://docs.google.com/spreadsheets/d/1cKnyhLBzOuvxqg5jYiMpC0p0Fla4yhJ4S0CBpvh1Qw/edit?usp=sharing>
- [3] Perfil de usuario y cada gráfico utilizado. https://miro.com/app/board/uXjVPLDdVA4=?share_link_id=107687776952
- [4] Gráfico 1 del prototipo 1. <https://public.flourish.studio/visualisation/11943481/>
- [5] Gráfico 2 del prototipo 1. <https://public.flourish.studio/visualisation/11941289/>
- [6] Gráfico 1 del prototipo 2. <https://public.flourish.studio/visualisation/11943529/>
- [7] Gráfico 2 del prototipo 2. <https://public.flourish.studio/visualisation/11941467/>
- [8] Gráfico 1 del prototipo 3. <https://public.flourish.studio/visualisation/11940997/>
- [9] Gráfico 2 del prototipo 3. <https://public.flourish.studio/visualisation/11941772/>
- [10] Infografía de la Visualización Final. https://www.canva.com/design/DAFUO8LRqI4/rabP-3rRti9NTKCdmRuuvw/view?utm_content=DAFUO8LRqI4&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink