Universidad Nacional de General Sarmiento

# LABORATORIO DE CONSTRUCCIÓN DE SOFTWARE

# PROYECTO PROFESIONAL 1

## TRABAJO PRÁCTICO INICIAL

## ENTREGA 4


INTEGRANTES

Guadalupe Nicole Arroyo

Lautaro Manuel Avalos

Federico Emanuel Farias

PROFESORES

Ing. Juan Carlos Monteros

Ing. Francisco Orozco De La Hoz

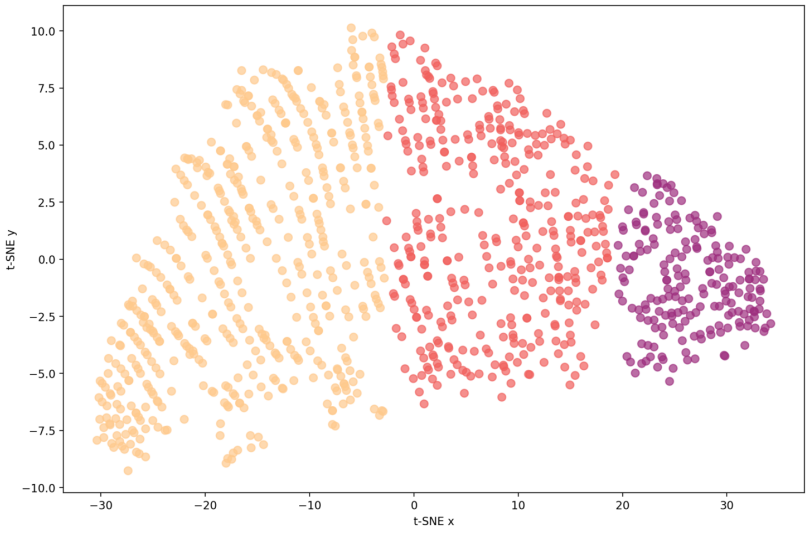
Lic. Leandro Dikenstein

# INTRODUCCIÓN

El siguiente informe se enfoca en mostrar en detalle los resultados obtenidos a partir del entrenamiento del modelo de machine learning elegido. Además se presenta la implementación de este modelo en la nube a través de la plataforma Streamlit mental-health-kmeans.streamlit.app, lo que permite un acceso más amplio para interesados, y facilita la visualización de los resultados. Esta aplicación tiene el objetivo de ser útil en el ámbito de la salud mental y la prevención del suicidio, proporcionando una herramienta útil para evaluar el riesgo potencial. También es importante recordar que se realiza por motivo de investigación y lo que se encuentra en la app debe usarse con la precaución necesaria.

# INTERPRETACIÓN DE LOS RESULTADOS

La interpretación de los resultados es fundamental para comprender cómo se relacionan las variables y cómo se agrupan los registros. La visualización final es un gráfico de dispersión que muestra los registros en función de las dos dimensiones reducidas por t-SNE. Cada punto en el gráfico se colorea según el clúster al que pertenece, lo que permite identificar visualmente cómo se agrupan los registros en el espacio bidimensional. Esto proporciona información valiosa sobre la estructura de los datos y cómo se distribuyen naturalmente en clústeres.



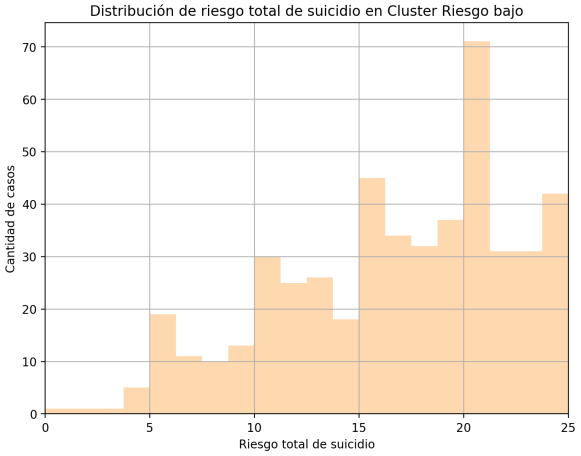

Se puede ver en el gráfico, gracias a la reducción de dimensionalidad t-SNE (80 de perplejidad), la distribución de los 3 clusters utilizados en el entrenamiento de K-Means. El color amarillo, cluster 1, nos indica los casos de bajo riesgo, el rosado, cluster 0, indica los casos de riesgo medio y el violeta, cluster 2, de riesgo alto. Esta gama de colores fue elegida

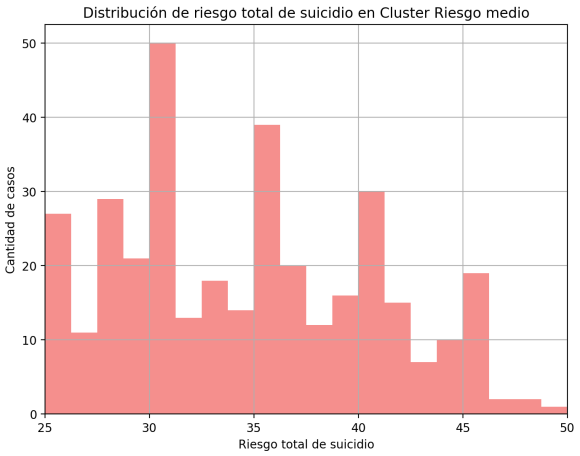
por ser cálida y suave en vez de una gama estilo semáforo (verde, amarillo y rojo), en caso de que alguna persona pueda sentirse incomoda al ver las visualizaciones y/o la app.

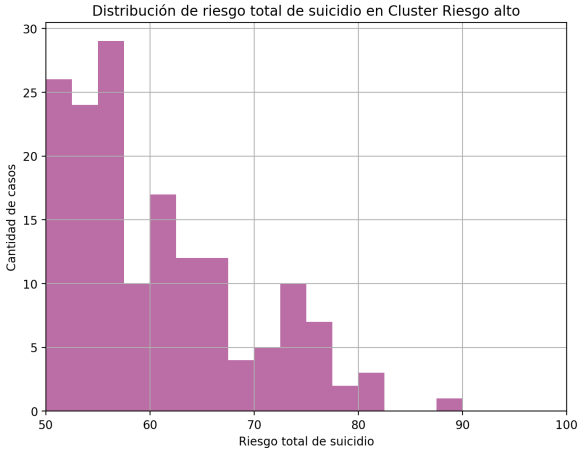
En el contexto de nuestro análisis de riesgo de suicidio, definimos que los valores aproximadamente en el rango de 50 a 100 se clasifican como "riesgo alto", en el rango de 25 a 50 se clasifican como “riesgo medio”, y en el rango de 0 a 25 se clasifican como “riesgo bajo”. Esta clasificación se basa en una evaluación de los datos, y en la determinación de que los valores en estos rangos representan una mayor o menor probabilidad de riesgo de suicidio.

Tener en cuenta que la clasificación de riesgo total puede variar dependiendo de los valores ingresados en las variables “valor de riesgo”, y “valor del promedio de riesgo”. Esta última abarca las variables depresión, ansiedad, historial de intento de suicidio, historial de trastorno mental, ingresos economicos, educación, etc., que se codificaron a valores numéricos en la entrega anterior.

A continuación se encuentran los histogramas utilizados para medir la cantidad de casos en función del riesgo de suicidio.

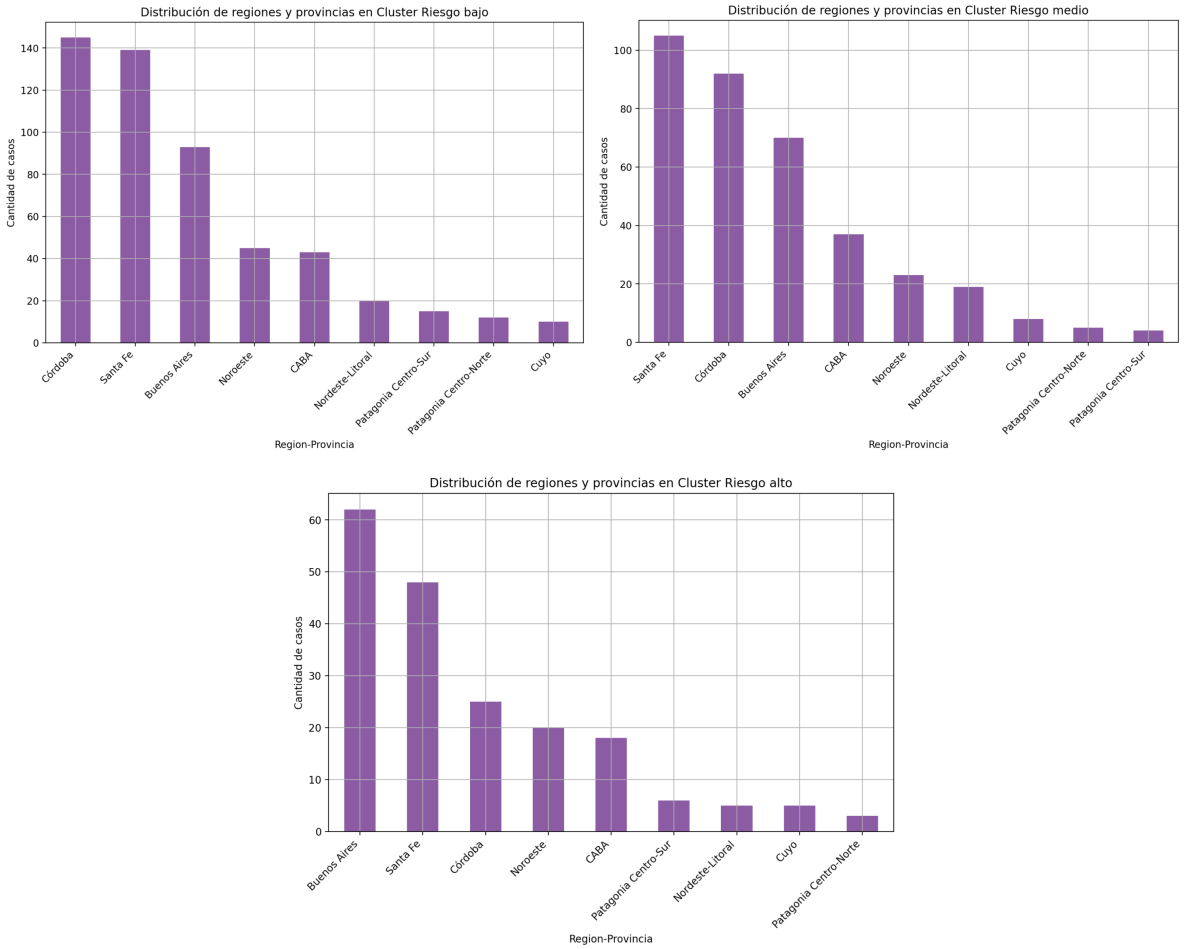


Una vez visualizados estos datos, luego del entrenamiento como se explicó en el changelog del informe y presentación de la entrega 3, se decidió unir las provincias en donde había pocos registros, según regiones oficiales de Argentina.

A continuación están las visualizaciones por cluster y por regiones-provincias de los casos.



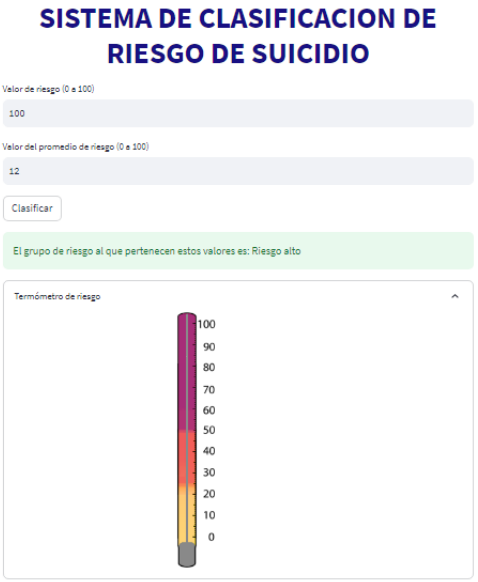

Se puede observar que Buenos Aires, Santa Fe y Córdoba tienen mayor cantidad de casos con riesgo más alto.

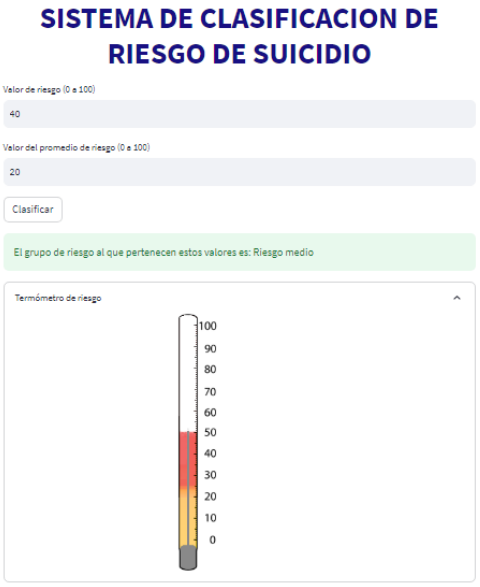
# IMPLEMENTACIÓN EN LA NUBE

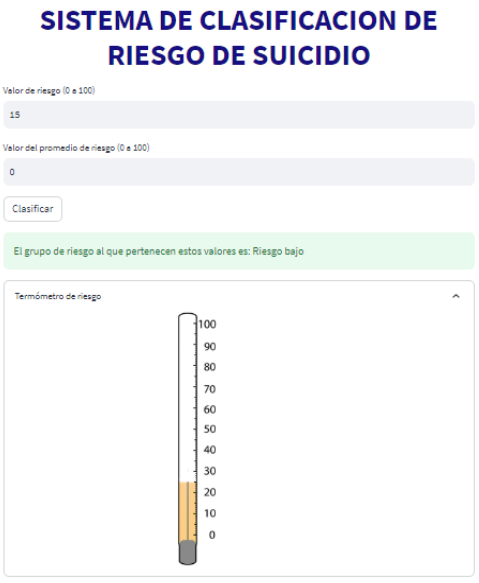
Nuestra aplicación web proporciona al usuario una forma interactiva de realizar predicciones sobre a qué grupo de riesgo de suicidio pertenece en función de los datos ingresados. También ofrecemos visualizaciones informativas para comprender mejor los resultados y la distribución de los datos en diferentes clústeres. En el proceso de implementación en la nube se involucraron los siguientes pasos:

- Descarga del modelo entrenado con las modificaciones indicadas en el Changelog con la herramienta “pickle” desde Google Cloud.
- En el desarrollo de la app frontend para permitir el ingreso y visualización de datos al usuario, se utiliza el modelo entrenado que permite clasificar a qué grupo de riego iría una persona.
- Ajuste de variables de entorno y compatibilidad de dependencias.
- Pruebas en entorno local con Google Colab.
- Deploy en en la nube de Streamlit asociado a Github.

A continuación se muestran unas imágenes de la aplicación web:



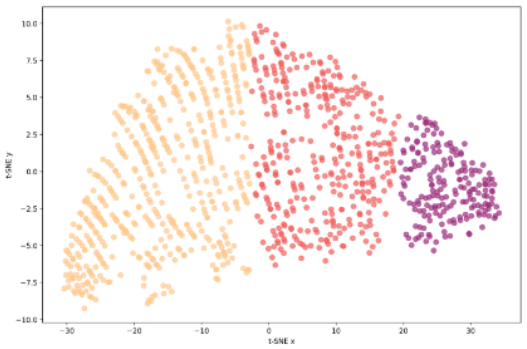



Distribución de clusters en 2D

## Distribución de clusters en 2D

Así es como se ve gracias a la reducción de dimensionalidad t-SNE (80 de perplejidad) la distribución de los 3 clusters con el entrenamiento de K-Means en dos dimensiones. Cada punto en el gráfico se colorea según el clúster al que pertenece, lo que permite identificar visualmente cómo se agrupan los registros en el espacio bidimensional. El color amarillo nos indica los casos de bajo riesgo, el rosado de riesgo medio y el violeta de riesgo alto.




Distribución de clusters en 3D

Histograma cantidad de casos según riesgo total

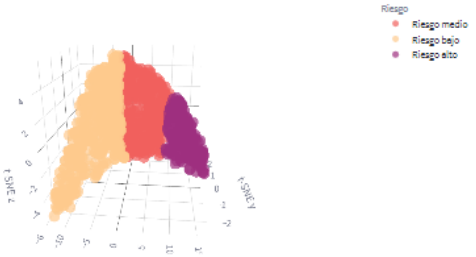
Histograma cantidad de casos según riesgo por regiones-provincias

Distribución de clusters en 2D

Distribución de clusters en 3D

## Distribución de clusters en 3D

Así es como se ve gracias a la reducción de dimensionalidad t-SNE (80 de perplejidad) la distribución de los 3 clusters con el entrenamiento de K-Means en 3 dimensiones. Se respetan los mismos colores que en el de dos dimensiones.




Histograma cantidad de casos según riesgo total

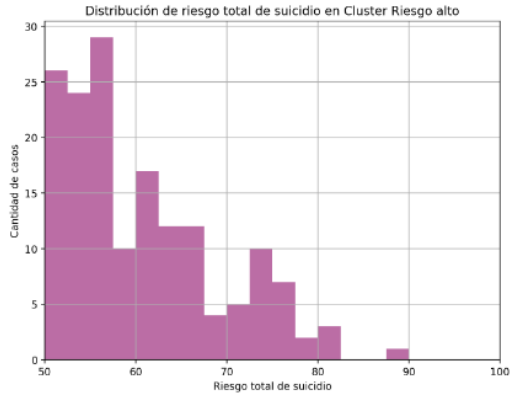
Histograma cantidad de casos según riesgo por regiones-provincias

Distribución de clusters en 2D

Distribución de clusters en 3D

Histograma cantidad de casos según riesgo total

## Histograma cantidad de casos según riesgo total



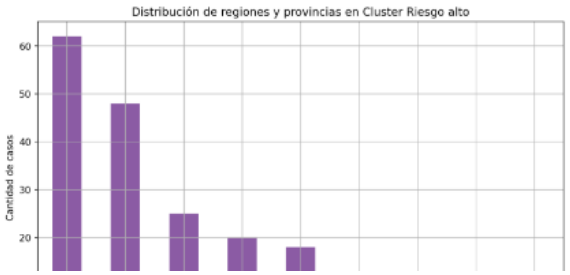

Distribución de clusters en 2D

Distribución de clusters en 3D

Histograma cantidad de casos según riesgo total

Histograma cantidad de casos según riesgo por regiones-provincias

## Histograma cantidad de casos según riesgo por regiones-provincias



# CONCLUSIONES

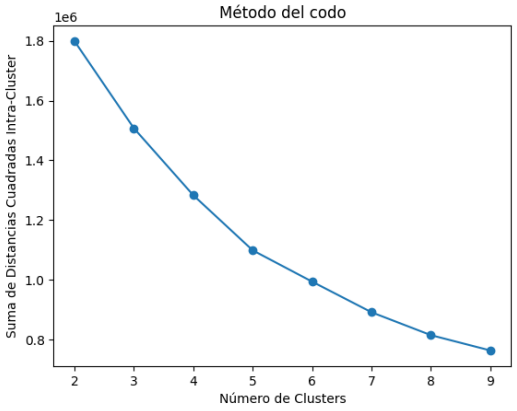
Después de las modificaciones implementadas en el modelo, se logró una mejor comprensión de los resultados. Esto permite apreciar el potencial de la inteligencia artificial en aplicaciones de salud, así como en diversas áreas que requieren el procesamiento de datos. Además, durante este proceso, se adquirieron conocimientos sobre diversas herramientas que agilizan la implementación de IA en la nube, lo que facilita la llegada de estas aplicaciones a un público más amplio y diverso.

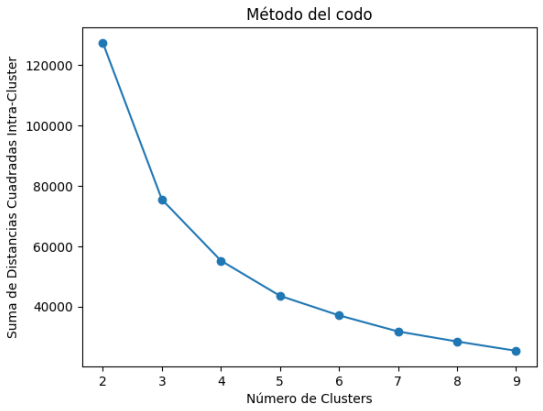
# CHANGELOG

Las principales modificaciones en esta entrega se vieron en:

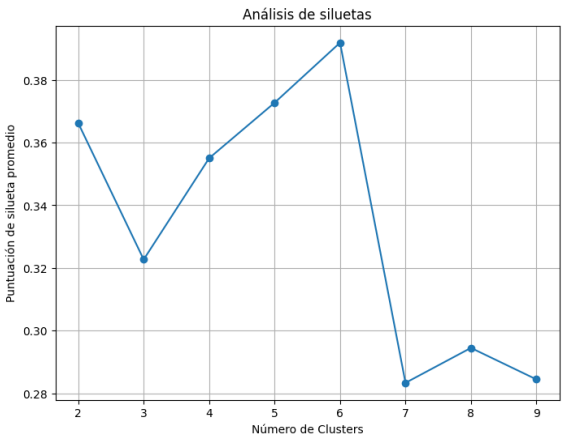
Refinamiento de datos: a diferencia de la entrega anterior, para este caso se utiliza como datos de entrada el riesgo de suicidio, y un promedio de las demás variables importantes del dataset. Esto se decidió para intentar llegar a la lógica que tuvieron las profesionales de la salud que crearon el dataset original, y llegar a los valores cercanos de riesgo de suicidio original.

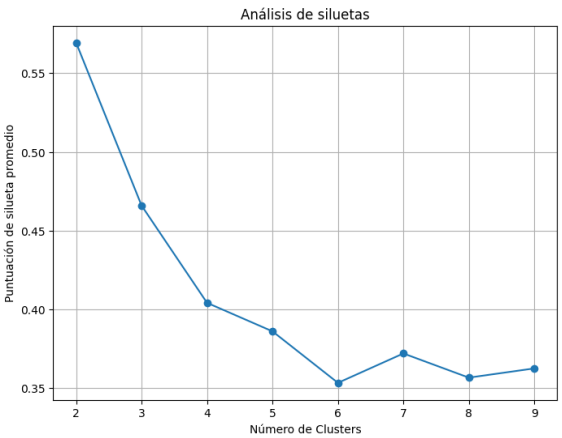
Cambio en la cantidad de clusters de K-Means: a causa de lo anterior y bajo nuevas mediciones con el método del codo y el análisis de siluetas, se decidió bajar la cantidad de clusters a 3 manteniendo la cantidad de iteraciones.




VS




VS

Se puede observar con más claridad el codo, y el aumento en el valor del análisis de siluetas, de 0.39 (6 clusters) como valor más alto en el primer análisis, a 0.47 (3 clusters) en el último análisis.

Cambios en el valor de perplejidad de t-SNE: se cambió el valor a 80 para mejorar la visualización. También se agrega una columna extra del eje z para la visualización 3D interactiva en la app.

# BIBLIOGRAFÍA

Google Colab (https://colab.research.google.com)

Pandas documentation (https://pandas.pydata.org/docs/)

Numpy documentation (https://numpy.org/doc/stable/)

Matplotlib documentation (https://matplotlib.org/stable/users/index)

Scikit-Learn (https://scikit-learn.org/stable/)

Plotly (https://plotly.com/python/)

Streamlit (https://streamlit.io/)

k-means clustering. En *Wikipedia*. (https://en.wikipedia.org/wiki/K-means_clustering)

t-distributed stochastic neighbor embedding. En *Wikipedia*.
(https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding)

López Steinmetz, L. C. (2021, May 3). *R Code and dataset for: Levels and predictors of depression, anxiety, and suicidal risk during COVID-19 pandemic in Argentina: The impacts of quarantine extensions on mental health state* [Dataset].
https://rdu.unc.edu.ar/handle/11086/20168