Universidad Nacional de General Sarmiento

LABORATORIO DE CONSTRUCCIÓN DE SOFTWARE

PROYECTO PROFESIONAL 1

TRABAJO PRÁCTICO INICIAL

ENTREGA 3

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 cómo al no obtener los resultados esperados con el modelo trabajado en la entrega anterior, el equipo pudo darse cuenta en donde estaban los errores para corregirlos utilizando otro modelo de entrenamiento. Luego se mostrará la configuración de los parámetros del nuevo modelo elegido y el re-entrenamiento posterior.

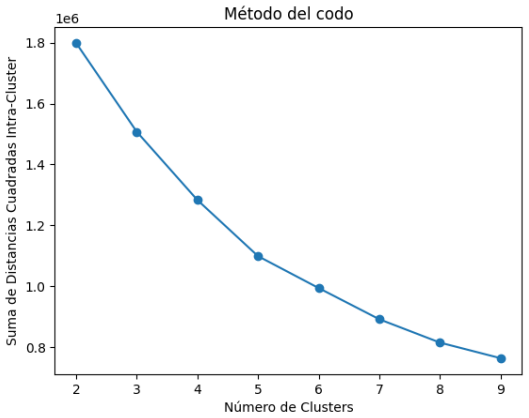
MODELO ELEGIDO

En el proyecto inicialmente se optó por utilizar el modelo de regresión logística, pero luego de un profundo análisis de los resultados obtenidos, se llegó a la conclusión de que, en la mayoría de las pruebas, este enfoque no brindó los resultados deseados (ver changelog). Por lo tanto, luego de una nueva investigación, se decidió adoptar un modelo de entrenamiento basado en la agrupación, K-Means, el cual tiene la capacidad de descubrir agrupaciones y patrones más sutiles que no pueden ser capturados fácilmente por un modelo de regresión logística.

CONFIGURACIÓN DE PARÁMETROS

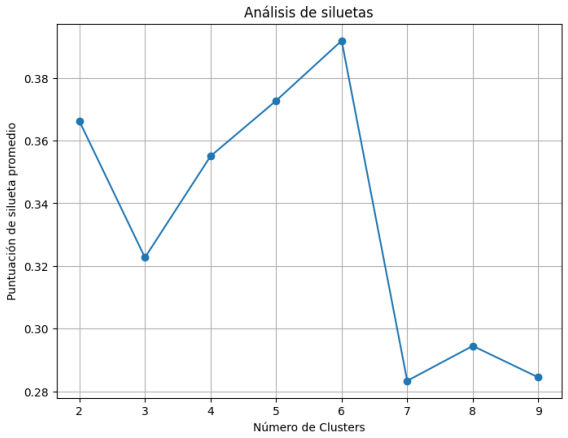
Se utilizaron dos técnicas para descubrir cuál era el k ideal, es decir, la cantidad de clusters con las que iba a trabajar el modelo.

El primero que se utilizó fue el método del codo, que consiste en determinar el número adecuado de clusters en el algoritmo de agrupamiento K-Means. Consiste en graficar la variabilidad dentro de los clusters en función del número de clusters y buscar el punto en el que la curva muestra un quiebre similar a un codo. Este punto sugiere un valor óptimo para el número de clusters, ya que agregar más clusters no mejora significativamente la reducción de la variabilidad.



Como se ve en el gráfico, el codo no es tan visible pero tanto 5 y 6 clusters parecen ser los candidatos. Al no tener una buena visualización, se utilizó otra estrategia para la validación de clusters para una decisión más precisa.

El análisis de siluetas en K-Means es una técnica para evaluar la calidad de los clusters generados en un conjunto de datos. Calcula la separación y cohesión de los objetos en cada cluster, asignando una puntuación a cada punto en función de su distancia promedio con los puntos del mismo cluster y del cluster más cercano. Las puntuaciones más altas indican una mejor separación y cohesión. Esta técnica ayuda a determinar el número óptimo de clusters y a medir la efectividad de las asignaciones.



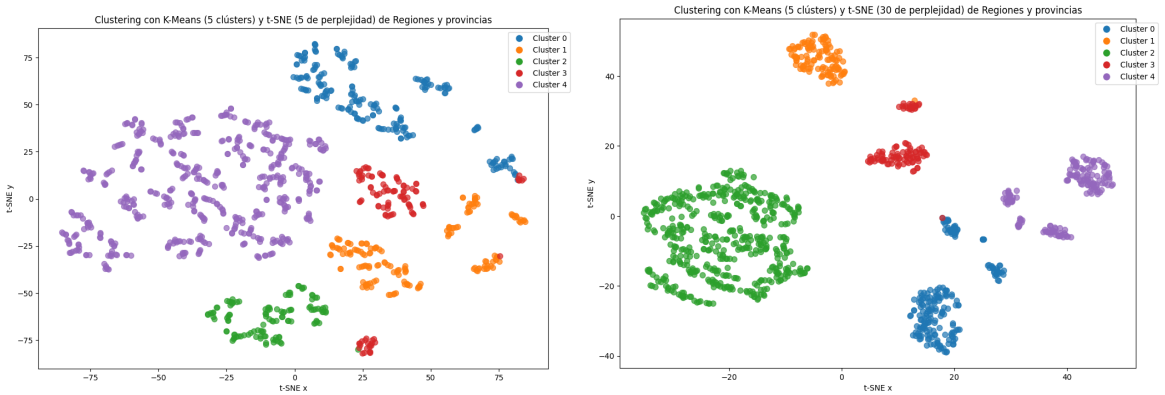
Se puede observar que los 3 valores más altos son, 2, 5 y 6. Dos clusters no sirven para el objetivo propuesto y se ve que 5 y 6 vuelven a ser candidatos. Es decir que con el método del codo y con el análisis de siluetas los mejores candidatos a cantidad de clusters fueron 5 y 6.

Otro parámetro que se usó fue el de la cantidad de iteraciones de K-Means. El entrenamiento comenzó con este número en 10 iteraciones y luego de varias pruebas se estableció en 100, ya que entregaba un resultado óptimo.

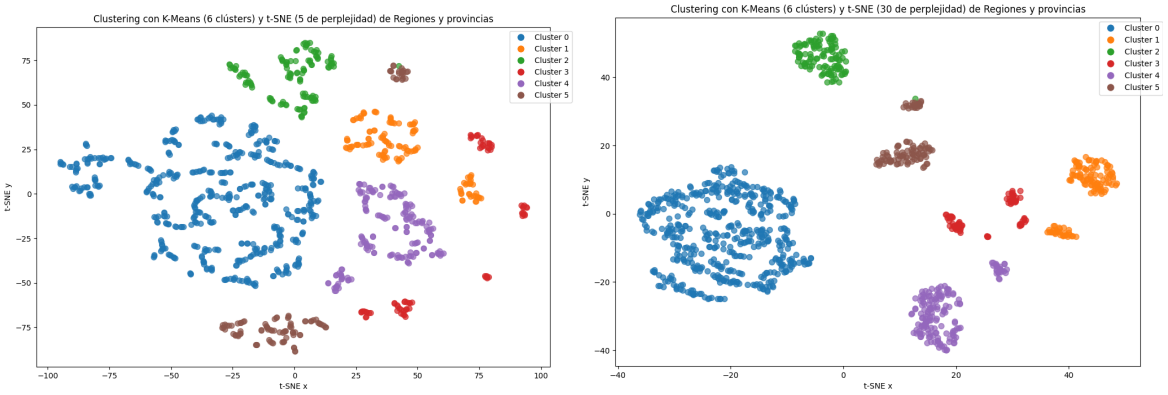
RE-ENTRENAMIENTO DEL MODELO

Como se explicó anteriormente, se realizaron las pruebas del modelo con 5 y 6 clusters, pero además, se implementó la técnica t-SNE (t-Distributed Stochastic Neighbor Embedding, en español, incrustación Estocástica de Vecinos Distribuidos t-SNE) para reducir la dimensionalidad de los datos a dos dimensiones, lo que permite visualizar las relaciones entre los puntos en un plano. Esta técnica también es parametrizable, mediante la perplejidad. Este es un valor ajustable que controla cómo se agrupan los puntos en un espacio de menor dimensión. Representa la cantidad de vecinos cercanos considerados por cada punto al calcular sus relaciones en el nuevo espacio. En el entrenamiento se decidió probar con valores de perplejidad que iban desde el 5 al 30.

Las coordenadas resultantes se agregan al conjunto de datos original para un análisis o visualización posterior. El resultado final es un gráfico que muestra la distribución de los casos totales en el espacio t-SNE, coloreados según los clusters generados por el algoritmo K-Means.



K-Means con 5 clusters y 5 de perplejidad (izq.) y 30 de perplejidad (der.)



K-Means con 6 clusters y 5 de perplejidad (izq.) y 30 de perplejidad (der.)

CONCLUSIONES

Al no obtener resultados deseados con la regresión logística, el equipo cambió al modelo de agrupación, K-Means para agrupar y clasificar a las provincias y/o regiones por riesgo de suicidio. Luego de hacer los ajustes en los parámetros de números de clústers, usando el método del codo y análisis de siluetas, y ajustar la perplejidad al aplicar t-SNE para reducir dimensionalidad, se lograron resultados más allá de lo esperado que acerca al equipo a continuar y llegar al objetivo propuesto.

CHANGELOG

En las entregas anteriores se mostró que se iba a utilizar un tipo de aprendizaje supervisado para realizar la clasificación de las provincias según el objetivo que propuesto.

Se inició con regresión logística para la clasificación y configuración de las variables en el modelo. El enfoque inicial investigó correlaciones entre variables como depresión, ansiedad y edad, excluyendo el riesgo de suicidio. Sin embargo, surgió un problema al intentar clasificar casos por provincia, ya que este modelo no permitía analizar todas simultáneamente. Esta aproximación se consideró poco práctica para analizar todas las provincias juntas, llevando a un cambio de enfoque. Se optó por una nueva investigación colaborativa para identificar un modelo más apropiado y práctico para el objetivo.

Después de una deliberación exhaustiva, se tomó la decisión de emplear la técnica de clustering con K-Means. Esto permitiría agrupar todos los casos en conjuntos coherentes, permitiendo al modelo determinar la estructura óptima de los clusters. Esta elección se basó en la capacidad de esta técnica para manejar conjuntos de datos complejos y proporcionar una visión más global y cohesionada de las relaciones entre las variables y las provincias en estudio.

Al cambiar de modelo y al tener en el dataset columnas/variables no numéricas que no fueron pensadas para el modelo anterior, se decidió darles valores acordes para que el modelo sepa/pueda clasificar y agrupar mejor los datos. Como los valores de depresión y ansiedad van en el rango de 0 a 100, siendo 100 el peor valor posible la asignación que les dió a las variables no numéricas fueron las siguientes:

```
assignment_mapping = {
    'MENTAL DISORDER HISTORY': {'no': 0, 'yes': 50},
    'EDUCATION': {
        'Completed postgraduate': 30,
        'Incomplete tertiary or university': 60,
        'Completed high school': 70,
        'Incomplete postgraduate': 40,
        'Completed tertiary or university': 50,
        'Incomplete high school': 80,
        'Incomplete elementary school': 100,
        'Completed elementary school': 90
    },
    'SUIC ATTEMPT HISTORY': {'ideation': 50, 'no': 0, 'yes': 100},
    'LIVING WITH SOMEBODY': {'no': 20, 'yes': 0},
    'ECONOMIC INCOME': {'yes': 0, 'no': 50}
}
```

Se puede observar que se le dió valores altos a lo que se vería como peor posible, como por ejemplo, si tiene un historial de intentos de suicidio, siendo 100 el peor valor ya que es un sí y 0 para no, ya que nunca tuvo intentos.

Otro tipo de transformación que se realizó fue crear una nueva columna en el dataset llamada REGION, respetando las regiones existentes en Argentina, ya que algunas provincias tenían pocas ocurrencias en el dataset y la visualización final (luego del agrupamiento en clusters) no sería la ideal.

Finalmente se dejaron de lado las columnas que no se usarían y se comenzó con el entrenamiento del modelo.

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/)

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