Speed Dating Dataset – Segunda entrega

Por:

Jose Franco Arroyave Cardona

Isaac Esteban Uribe

Profesor:

Raul Ramos Pollan

Medellín, Colombia

Universidad de Antioquia

2023

Introducción:

Este informe tiene como objetivo recopilar y explicar el progreso alcanzado en el proyecto de machine learning para la materia Inteligencia Artificial para las Ciencias e Ingenierías. La labor inicial se ha centrado en la exploración exhaustiva de un dataset extraído del portal web www.kaggle.com, comprendiendo el significado de cada una de las 195 variables, su relevancia y la forma en que se relacionan entre ellas, una etapa crucial en la que se ha puesto especial énfasis en la preparación de los datos para su posterior análisis y modelado. Es importante resaltar que, en el ámbito del machine learning, la calidad de los datos es un pilar fundamental para el éxito de cualquier proyecto. Por tanto, los pasos detallados a continuación representan el sólido cimiento sobre el cual se construirá la futura investigación y modelado.

Exploración y Preprocesamiento de Datos:

Carga de los datos:

Análisis de Variables:
El proceso se inicia con una exploración minuciosa de las variables contenidas en el dataset. Esto implicó el estudio detallado de cada variable, incluyendo su significado, su tipo de dato y su contribución potencial al objetivo del proyecto. La comprensión de estas facilita el manejo de los datos, incluso para el preprocesamiento.

Eliminación de Variables Redundantes:
Durante esta fase, se identificaron que se consideran redundantes o que no tenían relevancia y tomando como decisión eliminarlas para simplificar y limpiar el dataset.

Se realiza la eliminación de las columnas "field" y "career" debido a la presencia de columnas similares que estaban mejor categorizadas en el dataset, pues en el caso de estas dos columnas contienen cadenas de texto las cuales tiene diferentes formas de referirse a una misma carrera o campo laboral en el cual se desempeña una persona o en las cadenas de texto existen diferencias entre las mayúsculas o minúsculas, dificultando la tarea de limpiar la data, por lo tanto se opta por la opción de remover las columnas.

Además, se identifica el conjunto de columnas: 'idg', 'condtn', 'wave', 'round', 'position', 'positin1', que no aportan información valiosa para el objetivo a analizar, por lo tanto, se excluyen del dataset.

Tratamiento de Datos Faltantes:
Lidiar con los datos faltantes se convirtió en un desafío significativo. Para algunas columnas, la tarea tiene un menor grado de complejidad, ya que se pueden asignar

valores de manera directa para completar los faltantes correspondientes a categorías vacías. Sin embargo, en otras columnas existe una problemática más compleja: un alto porcentaje de datos faltantes que no pueden resolverse simplemente con imputación de valores, en particular, debido a la compleja estructura y relaciones intrínsecas entre grupos de variables. Para abordar este problema se exploran varios enfoques.

- **Segmentación del Dataset:**
  Cuando se encuentran columnas con un alto porcentaje de datos faltantes, la imputación de valores directos no es una opción viable debido a la compleja estructura del dataset y las relaciones intrínsecas entre grupos de variables. Para abordar esta problemática, se opta por la segmentación del dataset en áreas más manejables. Al dividir el dataset en subconjuntos con correlaciones menos fuertes, se puede enfocarse en grupos de variables específicos, lo que facilitó la imputación y el análisis de datos faltantes.

- **Estrategias de Imputación Avanzada:**
  Para abordar las columnas con datos faltantes de manera más avanzada, se implementan estrategias de imputación basadas en modelos. Utilizamos técnicas como la imputación por regresión lineal, donde los valores faltantes se rellenan con lo que se considera pertinente, haciendo un análisis posterior en su comportamiento para saber si esta decisión afecta la integración del dataset, primero se prueba rellenando los datos nulos con 0 (cero) y aplicar una regresión lineal para estimar los pesos de las variables, de la misma manera al observar el R^2 (R cuadrado) o Coeficiente de Determinación. Pero no se obtienen buenos resultados en la predicción del modelo y al determinar la importancia de las variables, por lo tanto se descarta este método.

- **Análisis de Correlaciones:**
  Un aspecto fundamental es el análisis de correlaciones entre variables. Este análisis permite identificar las relaciones entre las variables. Las correlaciones fuertes entre variables a menudo indican que una puede ser inferida o estimada a partir de otra, lo que facilitó la imputación y eliminación de parámetros de manera más precisa.

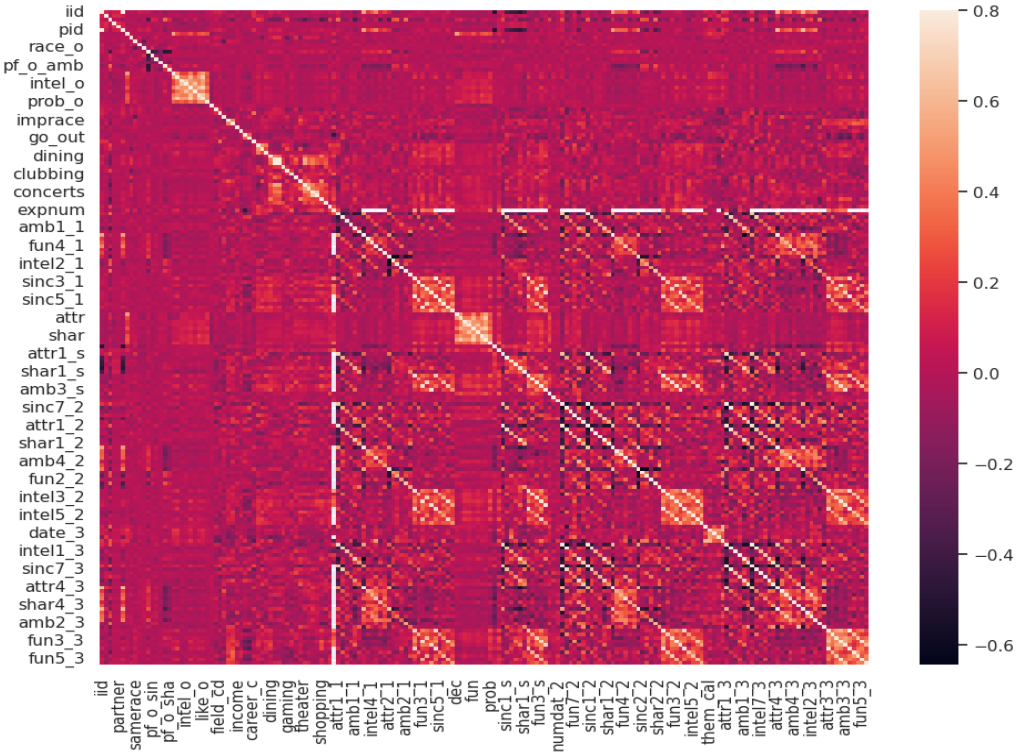

Figura 1.1 - Mapa de Correlaciones dataset en bruto

En la figura 1.1 se pueden ver las correlaciones de las variables antes de la limpieza del dataset, con esta gráfica se puede identificar que entre algunas de estas hay correlaciones perfectas o muy altas, esto permitie identificar posibles candidatas a ser eliminadas.

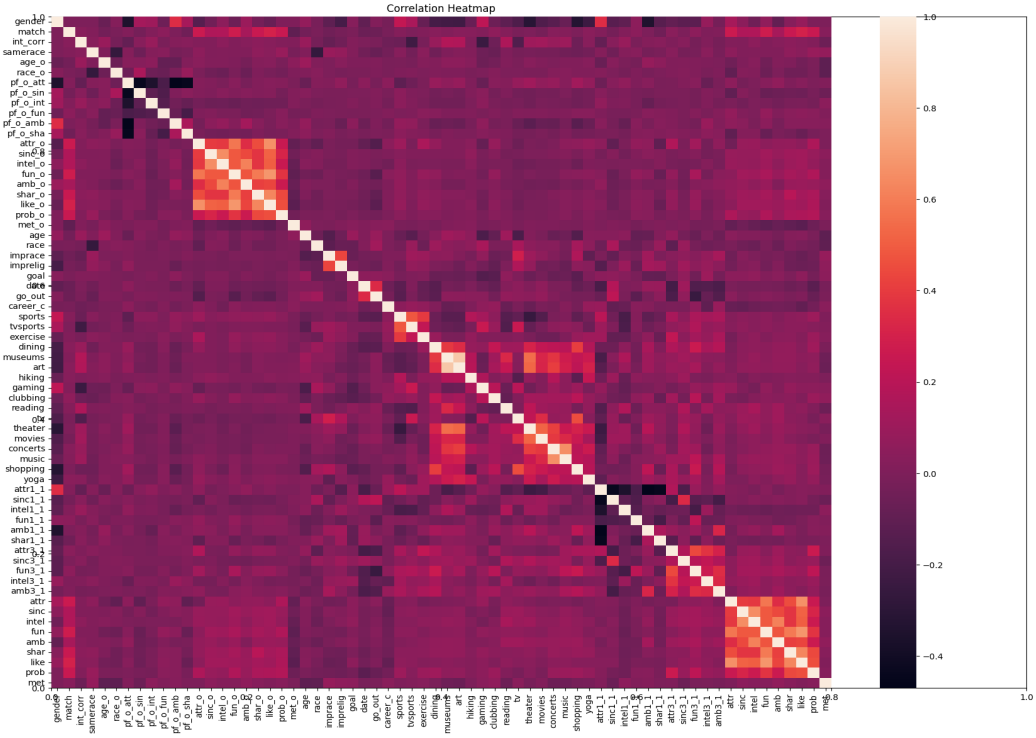

Figura 1.2 - Mapa de Correlaciones después de la limpieza

La figura 1.2 corresponde al mapa de correlaciones luego de las estrategias implementadas para la limpieza del dataset, en donde se observa una mejoría notable en la correlación entre variables.

Selección de Plantilla:
Se implementaron diferentes modelos de ML enfocados en la tarea de clasificar los datos en diferentes grupos, para la tarea puntal se busca definir si existe un match entre 2 personas por lo tanto se debe clasificar en 2 categorías, Si(1), No(0).

En la siguiente tabla se compara el desempeño de los diferentes modelos implementados, llegando a una conclusión preliminar de que no existe diferencia significativa entre estos.

Model	Accuracy
DecisionTreeClassifier	0.823
LogisticRegression	0.849
SVC	0.827
SGDClassifier	0.848
RidgeClassifier	0.841
RandomForestClassifier	0.837

Tabla 1.1 - Modelos vs Accuracy