

# Planificación, análisis y limpieza de datos de transfusión de sangre

Carvajal Diana, Castiblanco Yuranny, García Carlos.  
Fundación Universitaria Ucompensar, Bogotá D.C.  
dpcarvajal@ucompensar.edu.co  
yscastiblanco@ucompensar.edu.co  
carmandogarcia@ucompensar.edu.co

**Resumen**—Para cierta población, la inteligencia artificial (IA) es sinónimo de cualquier forma de razón lograda por sistemas no vivos, y este artículo tiene como propósito, no solo plantear las diferentes características de la inteligencia artificial, sino describir acerca de un procedimiento utilizado e implementado para simular esta ciencia, con el fin de cumplir a cabalidad y exponer de una manera funcional los conocimientos adquiridos en el transcurso del semestre académico.

**Abstract**--For a certain population, artificial intelligence (AI) is synonymous with any form of reason achieved by non -living systems, and this article has as its purpose, not only to raise the different characteristics of artificial intelligence, but to describe about a procedure used and implemented To simulate this science, in order to fully fulfill and expose in a functional way the knowledge acquired in the course of the academic semester.

## I. INTRODUCCIÓN

La presente investigación se refiere al tema K-Means en Python, que se puede definir como un algoritmo no supervisado de Clustering, el cual se hace uso de su funcionalidad cuando se cuenta con una gran cantidad de datos sin etiquetar.

El factor fundamental de este algoritmo es el de encontrar “K” grupos (clusters) entre los datos crudos (información suministrada). El cual trabaja iterativamente para asignar a cada “punto” (las filas del conjunto de entrada forman una coordenada) uno de los “K” grupos basado en sus respectivas características.

Con lo expuesto anteriormente, se estaría haciendo uso de la inteligencia artificial, dado que Es una tecnología que permite hacer automáticas una serie de operaciones con el fin de reducir la necesidad de que intervengan los seres humanos. Esto puede suponer una gran ventaja a la hora de controlar una ingente cantidad de información de un modo mucho más efectivo [1].

## II. ESTADO DEL ARTE

En la actualidad, existen numerosos estudios e investigaciones relacionados con el campo de inteligencia artificial (para el cual se encuentra abundante información sobre su base teórica, metodología de implementación, recomendaciones para su éxito), ligándose al algoritmo K-Means, que se centran en intentar resolver los problemas de agrupamiento de datos, con el fin de efectuar un aprendizaje automático no supervisado, dado que segrega los datos no

etiquetados en varios grupos, llamados clústeres, basados en características similares, patrones comunes.

El algoritmo K-means sigue siendo objeto de estudio por parte de la comunidad científica.

Desde su aparición, se han presentado muchos artículos relacionados con diferentes aspectos del algoritmo. Si se analizara la importancia del algoritmo K-means en un periodo de tiempo, no sería raro describir la perspectiva, evolución y la continua aparición de nuevos algoritmos de agrupamiento [2].

En la etapa de inicialización del algoritmo, se determina el número de grupos a crear y se seleccionan los centroides iniciales. Debido a que la elección de los centroides iniciales impacta en la solución del agrupamiento, no existe un método generalizado en la deliberación de centroides iniciales, pero si la comparación de diferentes métodos de inicialización [3].

Al detectar que a medida que incrementan las iteraciones del algoritmo K-means, algunos objetos permanecen cerca de su centroide asignado y al mismo tiempo retirados de centroides más distantes, la exclusión de centroides resulta interesante debido al concepto de desigualdad triangular en geometría [4].

El límite de distancias presenta un comportamiento similar a la desigualdad triangular debido a que su intervalo se encuentra por arriba de 1 y por debajo de k. [5]

## III. METODOLOGÍA

El actual artículo de investigación “Planificación, validación y limpieza de datos de transfusión de sangre”, corresponde a un proyecto de desarrollo por cuánto está encaminado a resolver problemas prácticos, a través de la inteligencia artificial, basados del algoritmo K-Means, pero utilizando la técnica (y librería del lenguaje de programación Python) Pandas Profiling, tomando como base el conjunto de datos del Centro de Servicio de Transfusión de Sangre en la ciudad de Hsin-Chu en Taiwán [6].

Para construir el modelo de datos, se seleccionaron 748 donantes, cada uno incluía R (Recencia - meses desde la última donación), F (Frecuencia - número total de donaciones), M (Monetario - total de sangre donada en c.c.), T (Tiempo - meses desde la primera donación), y una variable binaria que representa si donó sangre en marzo de 2007 (1 significa donar sangre; 0 significa no donar sangre).

Para proceder con el respectivo análisis exploratorio de datos, se proporcionó un nombre de variable, el tipo de variable, la unidad de medida y una breve descripción. El orden de este listado corresponde al orden de los números a lo largo de las filas de la base de datos.

\* Revista Argentina de Trabajos Estudiantiles. Patrocinada por la IEEE.

R (Reciente - meses desde la última donación), F (Frecuencia - número total de donaciones), M (Monetario - total de sangre donada en c.c.), T (Tiempo - meses desde la primera donación), y una variable binaria que representa si él/ella sangre donada en marzo de 2007 (1 significa donar sangre; 0 significa no donar sangre).

Técnicamente se aplicó la librería Pandas Profiling, dado que esta genera automáticamente informes de los conjuntos de datos contenidos en objetos DataFrame, y adicionalmente permite evitar la repetición e iteración de análisis exploratorio, es decir, comprender de una mejor manera el análisis efectuado.

#### IV. RESULTADOS

A lo largo del desarrollo de este proyecto de IA e investigación, se analizaron diferentes técnicas y procedimientos con el fin de dar cumplimiento a la actividad, y se obtuvieron los siguientes sobre la planificación, análisis y limpieza de datos de transfusión de sangre:

##### A. Uso de Pandas:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sb
```

Fig. 1. Se invoca librerías en lenguaje Python.

##### B. Llamado lista de datos:

```
5 #Llamado de la lista de datos
6 data = pd.read_csv('/archivo_csv/BD_transfucion.csv')
7 #HEAD: nos muestra los 5 primeros registros
8 data.head()
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	he/she donated
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

Fig. 2. Se realiza llamado de lista de datos con la función READ\_CSV.

##### C. Data completa:

```
data
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	he/she donated
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0
...	...	...	...	...	...
743	23	2	500	38	0
744	21	2	500	52	0
745	23	3	750	62	0
746	39	1	250	39	0
747	72	1	250	72	0

Fig. 3. Se genera toda la información con la instrucción DATA.

##### D. Uso de la instrucción Shape:

```
#SHAPE nos indica la cantidad de filas u columnas
data.shape
```

(748, 5)

Fig. 4. Se ejecuta la instrucción Shape que genera la cantidad de filas y columnas.

##### E. Análisis exploratorio:

```
# DESCRIBE analisis exploratorio (Cantidad , la media , desviacion , valor max y min y percentiles)
data.describe()
```

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	he/she donated
count	748.000000	748.000000	748.000000	748.000000	748.000000
mean	9.506684	5.514706	1378.676471	34.282086	0.237968
std	8.095396	5.839307	1459.826781	24.376714	0.426124
min	0.000000	1.000000	250.000000	2.000000	0.000000
25%	2.750000	2.000000	500.000000	16.000000	0.000000
50%	7.000000	4.000000	1000.000000	28.000000	0.000000
75%	14.000000	7.000000	1750.000000	50.000000	0.000000
max	74.000000	50.000000	12500.000000	98.000000	1.000000

Fig. 5. Se hace uso de la sentencia DESCRIBE, con el fin de efectuar el respectivo análisis exploratorio del conjunto de datos.

##### F. Identificación de variables (Features):

```
# Muestra la informacion de los datos (feature)
data.info()
```

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 748 entries, 0 to 747  
Data columns (total 5 columns):

#	Column	Non-Null Count	Dtype
0	Recency (months)	748 non-null	int64
1	Frequency (times)	748 non-null	int64
2	Monetary (c.c. blood)	748 non-null	int64
3	Time (months)	748 non-null	int64
4	he/she donated	748 non-null	int64

dtypes: int64(5)  
memory usage: 29.3 KB

Fig. 6. Se efectúa identificación clara de las variables a usar (Features) por medio de la función DATA.INFO().

##### G. Validación de frecuencia:

```
data['Frequency (times)'].value_counts(dropna=False)
```

1	158
2	112
3	87
4	62
5	62
6	52
7	43
8	31
9	24
11	22
12	14
10	14
14	13
16	13
13	9
15	6

Fig. 7. Se realiza generación de conteo de frecuencia.

##### H. Identificación de Label:

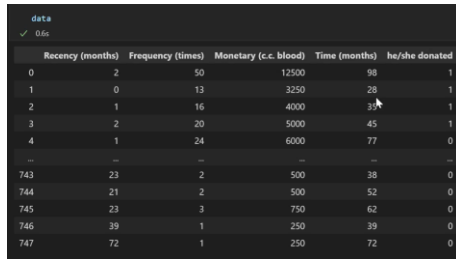
```
#Label
data[['he/she donated']]
```

he/she donated	
0	1
1	1
2	1
3	1
4	0
...	...
743	0
744	0
745	0
746	0
747	0

748 rows x 1 columns

Fig. 8. Se corrobora la donación sobre la Data cargada.

## I. Análisis de Data:



	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	he/she donated
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0
...	...	...	...	...	...
743	23	2	500	38	0
744	21	2	500	52	0
745	23	3	750	62	0
746	39	1	250	39	0
747	72	1	250	72	0

Fig. 9. Se genera la Data para realizar el rastreo correspondiente.

## J. Limpieza de los datos:

```
# analisis realizado se identifico que esta data es supervisada ya que los clasifica en 1 si Dono - 0 si no dono
data.drop(['he/she donated'],1).hist()
plt.show()

C:\Users\dcarvasa\AppData\Local\Temp\ipykernel_3896\2961748206.py:1: FutureWarning: In a future version of pandas all arguments
of DataFrame.drop except for the argument 'labels' will be keyword-only.
data.drop(['he/she donated'],1).hist()
```

Fig. 10. Se realiza uso de limpieza de datos.

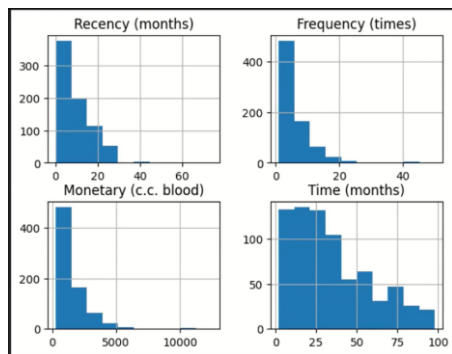


Fig. 11. Evidencia limpieza de datos.

## K. Análisis supervisado y no supervisado:

```
#agrupacion y relacion de cada uno (Analisis para supervisado y no supervisado)
sb.pairplot(data.dropna(), hue='he/she donated', size=(8,8), vars=['Recency (months)', 'Frequency (times)', 'Monetary (c.c. blood)', 'Time (months)'], kind='scatter')
plt.show()
```

Fig. 12. Agrupación y relación de la Data, que genera como resultado un análisis supervisado, dado que genera información etiquetada (Leavel, los clasifica como 1 y 0).

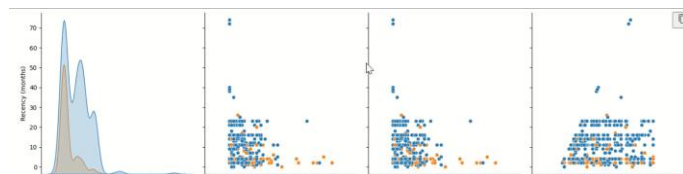


Fig. 13. Evidencia análisis supervisado y no supervisado (1).

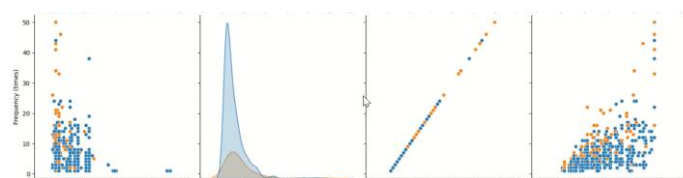


Fig. 14. Evidencia análisis supervisado y no supervisado (2).

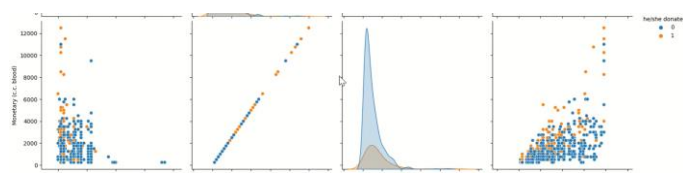


Fig. 15. Evidencia análisis supervisado y no supervisado (3).

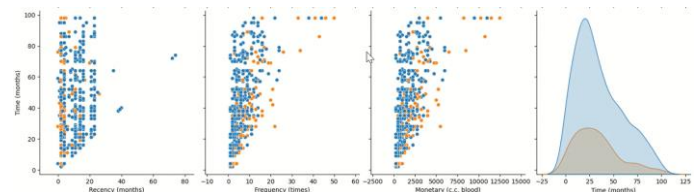


Fig. 16. Evidencia análisis supervisado y no supervisado (4).

## V. CONCLUSIONES

Con el agrupamiento y cargue de información sobre la transfusión de sangre, que inicialmente presentaba problema de clasificación, se muestra que es factible mejorar la información de acuerdo con metodologías tecnológicas, entre ellas, la técnica Pandas Profiling.

Con base en los resultados experimentales, se observó que la implementación de inteligencia artificial genera una reducción del tiempo de ejecución y efectúa una limpieza de datos, los cuales se pueden contemplar en las evidencias plasmadas en el ítem “IV. RESULTADOS”.

Finalmente, se cumple a cabalidad con la necesidad planteada en la actividad y se obtienen resultados positivos tanto en la elaboración del proyecto como en el conocimiento obtenido.

## REFERENCIAS

- [1] Redacción APD. (2019, Mar 04). ¿Qué es Machine Learning y cómo funciona?. <https://www.apd.es/que-es-machine-learning/>
- [2] Jain, A. K (2010). Data clustering: 50 years beyond K-means. <https://www.sciencedirect.com/science/article/abs/pii/S01678655090002323>
- [3] Peña, J. M., Lozano, J. A., & Larrañaga, P. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. <https://www.sciencedirect.com/science/article/abs/pii/S0167865599000690>
- [4] Phillips, S. J (2002). Acceleration of K-Means and Related Clustering Algorithms. [https://link.springer.com/chapter/10.1007/3-540-45643-0\\_13](https://link.springer.com/chapter/10.1007/3-540-45643-0_13)
- [5] Drake, J., & Hamerly, G. (2012). Accelerated k-means with adaptive distance bounds. [http://opt.kyb.tuebingen.mpg.de/papers/opt2012\\_paper\\_13.pdf](http://opt.kyb.tuebingen.mpg.de/papers/opt2012_paper_13.pdf)
- [6] Chauchan, A. (2022, Ago). Blood Transfusion Dataset. <https://www.kaggle.com/datasets/4632ef2e012b5faa60253f5444991622660fd46e7baf3c6bcd7d2b0ec90bec6?resource=download>