Clasificador de riesgo de cáncer de pulmón basado en los síntomas del paciente

Universidad Industrial de Santander

Jorge Andrey Garcia Vanegas - 2180115
Daniel Felipe Calderón Calderón - 2210052
Miguel Enrique Quintero Suarez - 2190932

INTRODUCCIÓN

EL CÁNCER DE PULMÓN, UNA ENFERMEDAD DEVASTADORA QUE AFECTA A MILLONES DE PERSONAS EN TODO EL MUNDO, PRESENTA UN DESAFÍO SIGNIFICATIVO EN TÉRMINOS DE DIAGNÓSTICO TEMPRANO Y TRATAMIENTO EFICAZ. EN ESTE CONTEXTO, EL AVANCE DE LA INTELIGENCIA ARTIFICIAL EMERGE COMO UNA HERRAMIENTA INVALUABLE PARA MEJORAR LA DETECCIÓN Y LA PRECISIÓN DIAGNÓSTICA.

NUESTRO PROYECTO SE CENTRA EN EL DESARROLLO DE UN SISTEMA DE INTELIGENCIA ARTIFICIAL CAPAZ DE IDENTIFICAR EL RIESGO DE CÁNCER DE PULMÓN MEDIANTE LA CONSIDERACIÓN DE DIVERSOS FACTORES QUE PUEDEN INFLUIR EN EL DESARROLLO DE LA ENFERMEDAD

OBJETIVOS

* DESARROLLAR UN SISTEMA PRECISO DE DETECCIÓN DEL CÁNCER DE PULMÓN MEDIANTE EL USO DE LAS CARACTERÍSTICAS DEL DATASET.

*HACIENDO USO DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL CÓMO EL ALGORITMO DE DEEP LEARNING, PCA,ECT. SE LOGRARÁ MANEJAR LA INFORMACIÓN DE MEJOR MANERA PARA DAR UN RESULTADO PRECISO .

*DEMOSTRAR MEDIANTE GRÁFICAS LOS RESULTADOS DEL PROYECTO PARA EVIDENCIAR VISUALMENTE LOS OBJETIVOS.

DATASET UTILIZADO PARA EL PROYECTO

- https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

COLUMNAS DEL DATASET

Age: The age of the patient. (Numeric)

Gender: The gender of the patient. (Categorical)

Air Pollution: The level of air pollution exposure of the patient. (Categorical)

Alcohol use: The level of alcohol use of the patient. (Categorical)

Dust Allergy: The level of dust allergy of the patient. (Categorical)

OccuPational Hazards: The level of occupational hazards of the patient. (Categorical)

Genetic Risk: The level of genetic risk of the patient. (Categorical)

chronic Lung Disease: The level of chronic lung disease of the patient. (Categorical)

Balanced Diet: The level of balanced diet of the patient. (Categorical)

Obesity: The level of obesity of the patient. (Categorical)

Smoking: The level of smoking of the patient. (Categorical)

Passive Smoker: The level of passive smoker of the patient. (Categorical)

Chest Pain: The level of chest pain of the patient. (Categorical)

Coughing of Blood: The level of coughing of blood of the patient. (Categorical)

Fatigue: The level of fatigue of the patient. (Categorical)

Weight Loss: The level of weight loss of the patient. (Categorical)

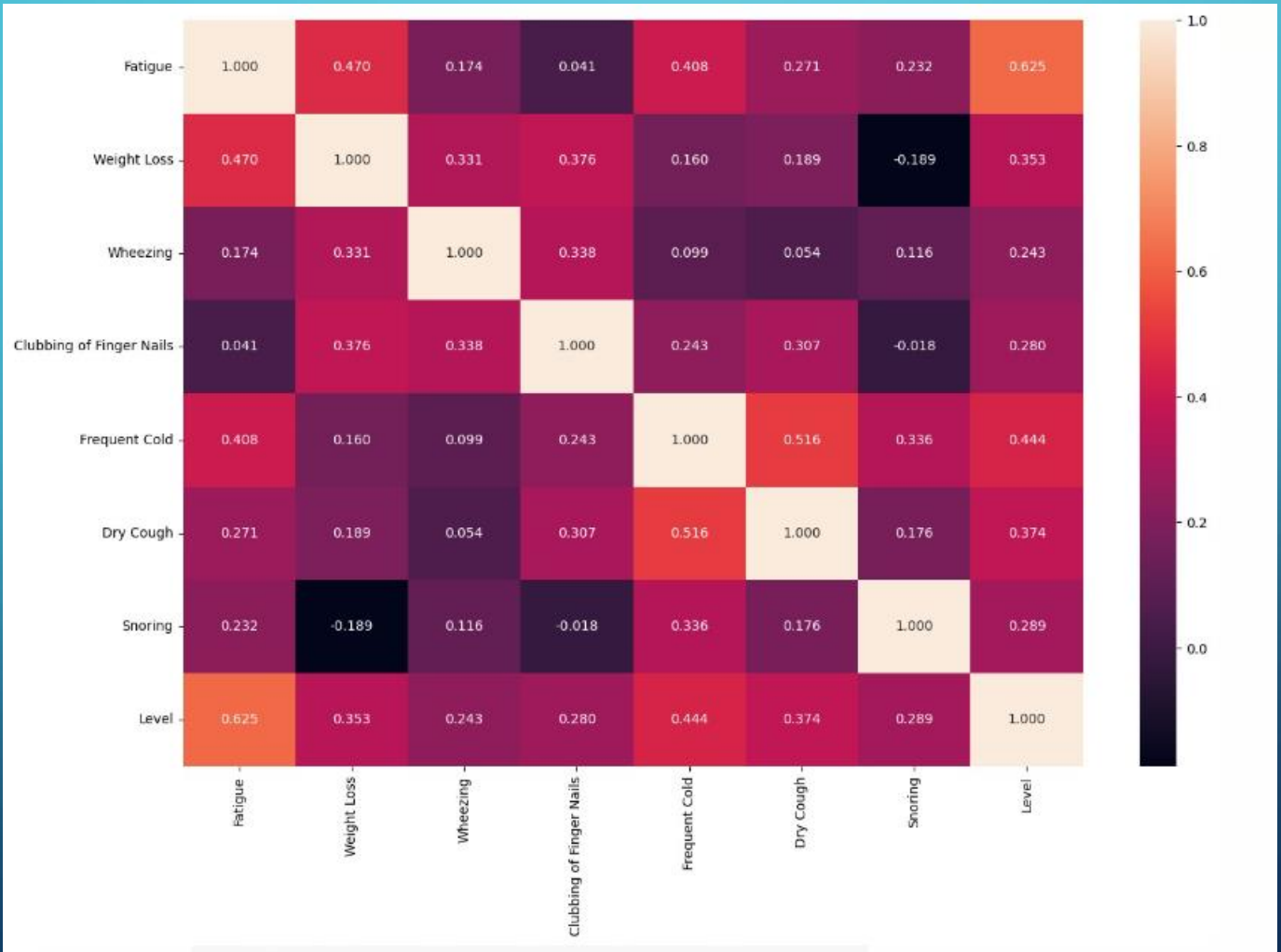
Shortness of Breath: The level of shortness of breath of the patient. (Categorical)

Wheezing: The level of wheezing of the patient. (Categorical)

Swallowing Difficulty: The level of swallowing difficulty of the patient. (Categorical)

Clubbing of Finger Nails: The level of clubbing of finger nails of the patient. (Categorical)

MATRIZ DE CORELACION



TRATAMIENTO DE DATOS

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Coughing of Blood	Fatigue
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...	1000.000000	1000.000000
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	...	4.859000	3.856000
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	...	2.427965	2.244616
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	...	3.000000	2.000000
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	...	4.000000	3.000000
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	...	7.000000	5.000000
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	...	9.000000	9.000000

8 rows × 24 columns

Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring
1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
3.856000	3.855000	4.240000	3.777000	3.746000	3.923000	3.536000	3.853000	2.926000
2.244616	2.206546	2.285087	2.041921	2.270383	2.388048	1.832502	2.039007	1.474686
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
3.000000	3.000000	4.000000	4.000000	4.000000	4.000000	3.000000	4.000000	3.000000
5.000000	6.000000	6.000000	5.000000	5.000000	5.000000	5.000000	6.000000	4.000000
9.000000	8.000000	9.000000	8.000000	8.000000	9.000000	7.000000	7.000000	7.000000

TRATAMIENTO DE DATOS

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4

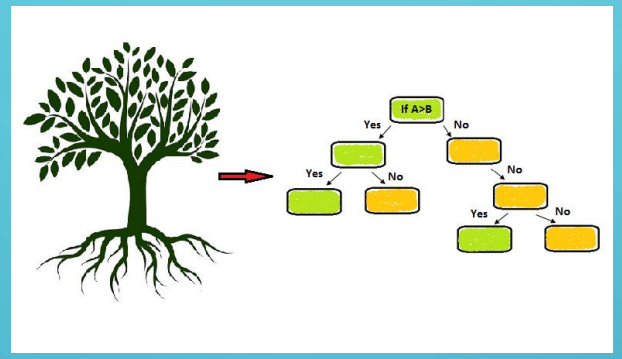
5 rows × 26 columns

Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
3	1	2	3	4	Low
6	2	1	7	2	Medium
1	4	6	7	2	High
4	5	6	7	5	High
4	2	4	2	3	High

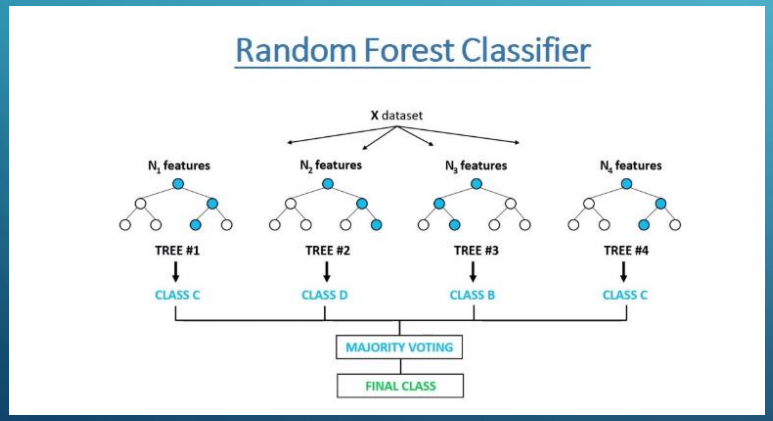
MÉTODOS DE IA IMPLEMENTADOS

MÉTODOS DE CALSIFICACIÓN (OBTENER EL RIESGO DE CONTRAER CANCER DE PULMÓN)

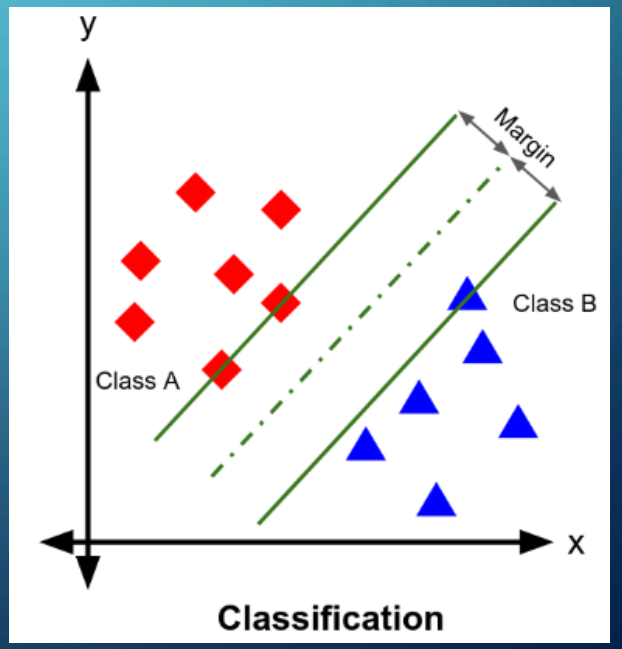
DECISION TREE CLASSIFIER



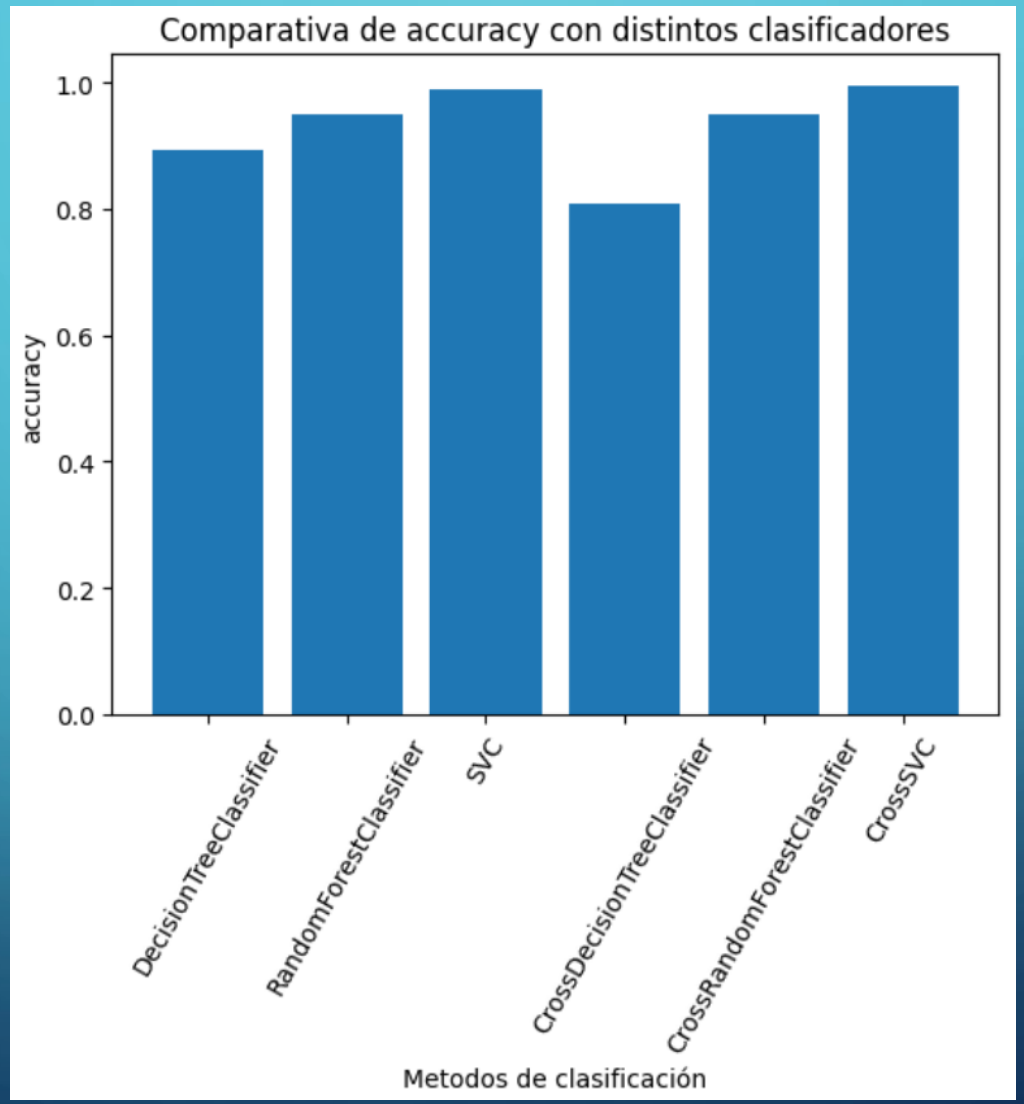
RANDOM FOREST CLASSIFIER



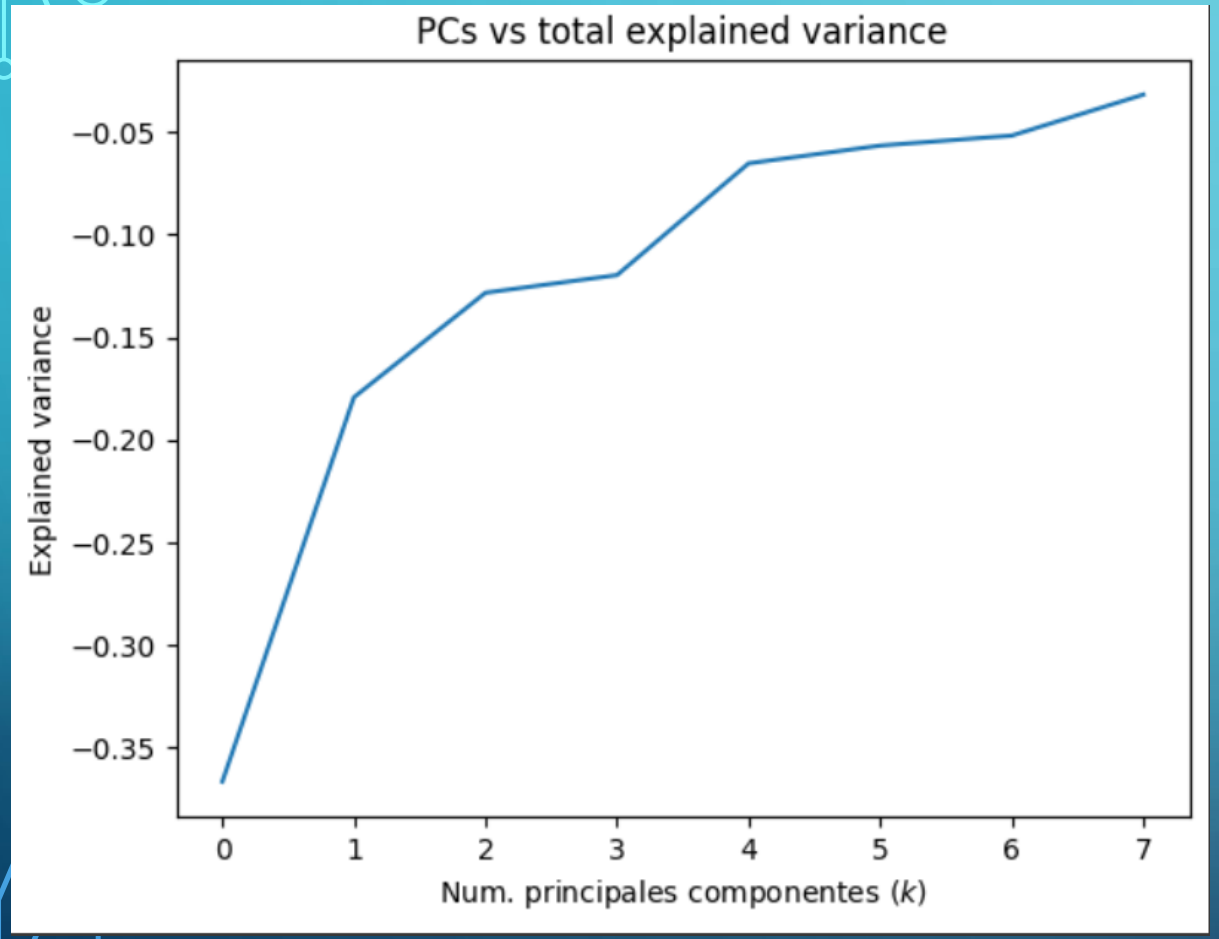
SUPPORT VECTOR MACHINE

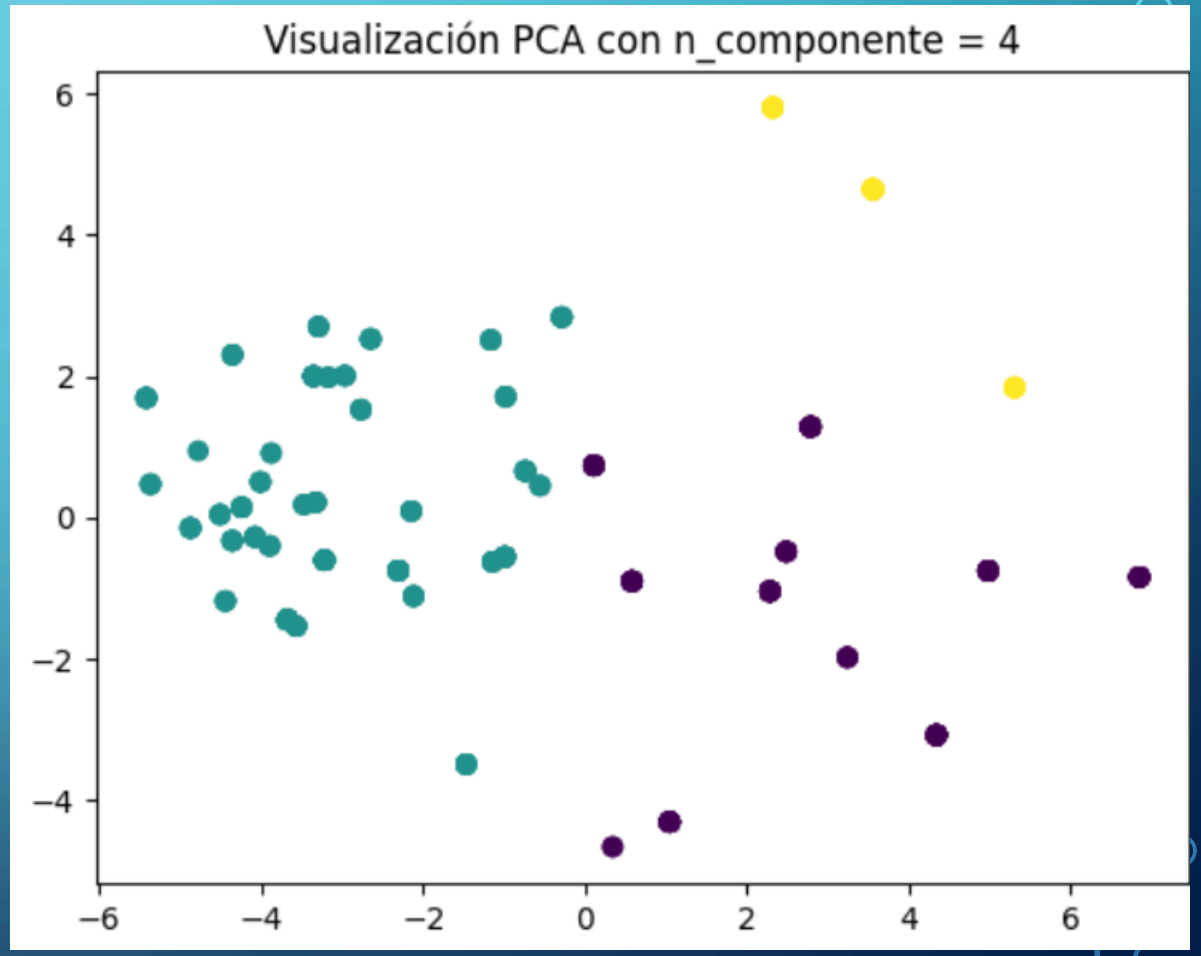


COMPARACIÓN EFECTIVIDAD DE LOS 3 MÉTODOS

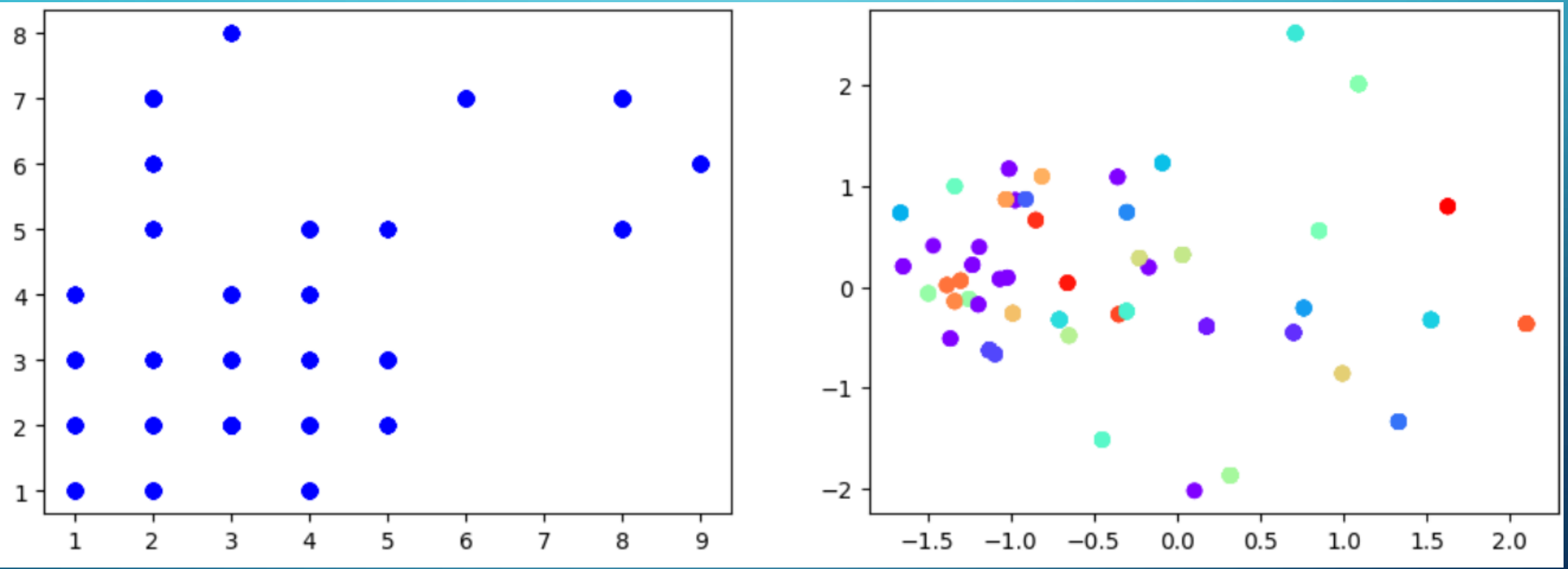


USO DE PCA PARA REDUCIR EL DATASET MEDIANTE NÚMERO DE COMPONENTES

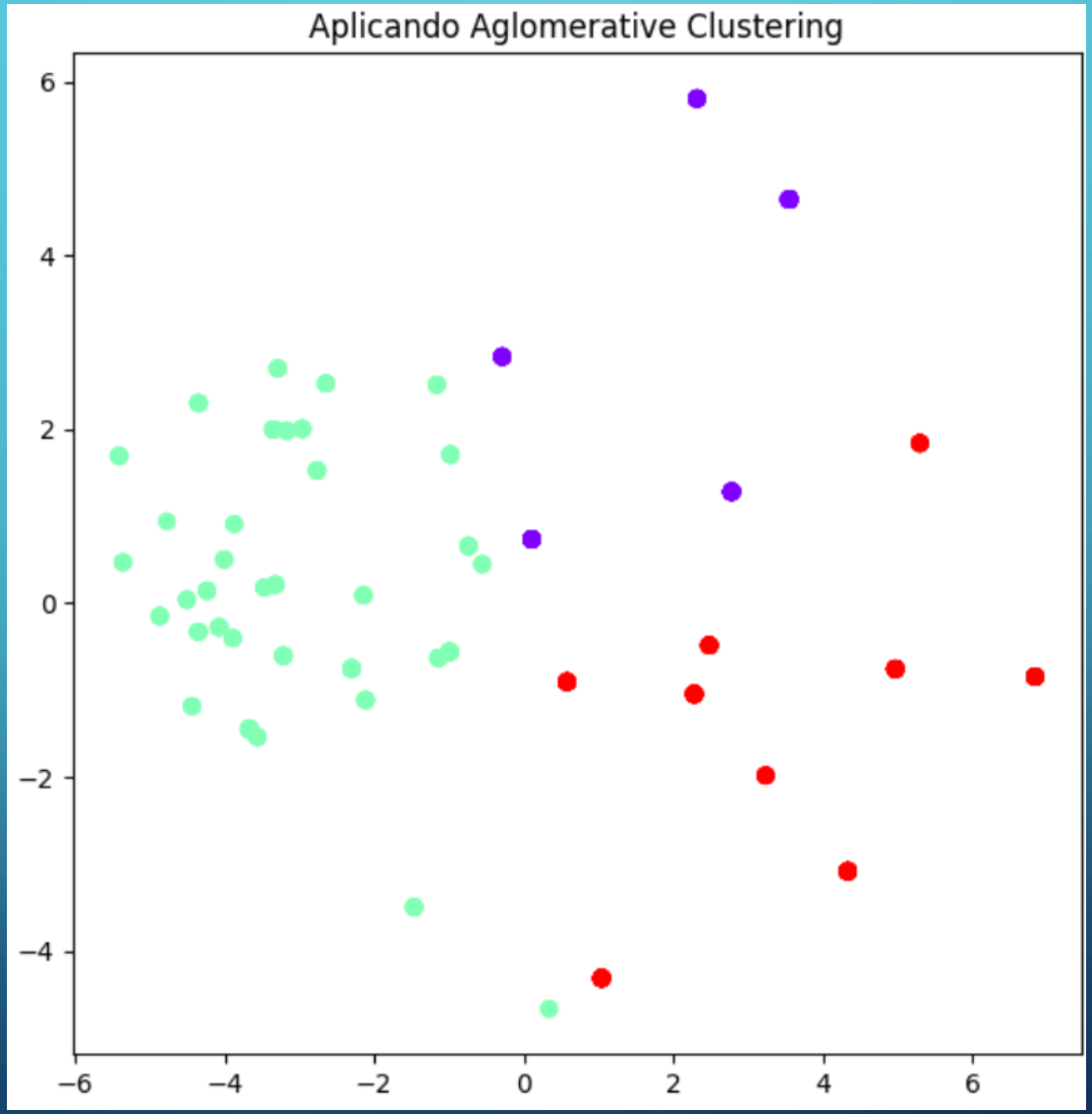




DBSCAN



AGLOMERATIVE CLUSTERING



DEEP LEARNING

```
16/16 [==============================] - 0s 4ms/step
              precision    recall  f1-score   support

           0       0.93      1.00      0.96       154
           1       0.97      0.93      0.95       162
           2       1.00      0.97      0.99       184

    accuracy                           0.97       500
   macro avg       0.97      0.97      0.97       500
weighted avg       0.97      0.97      0.97       500
```

CONCLUSION

LOS MÉTODOS DE IA ARROJAN RESULTADOS CERTEROS, EL USO DE PCA, METODOS DE CLASIFICACION, DEEP LEARNING, DBSCAN Y AGLOMERATIVE CLUSTERING FUERON NECESARIOS Y UTLIES PARA MANEJAR LA INFORMACION Y DAR UNA RESPUESTA CERTERA RESPECTO AL RIESGO DE CONTRAER CANCER DE PULMON.