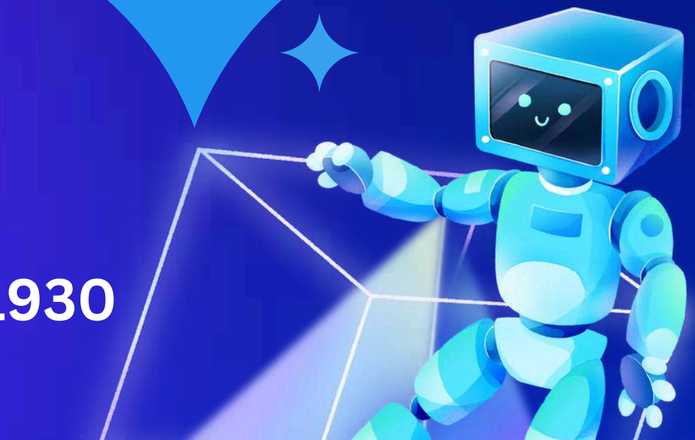
Inteligencia Artificial I - 2025-2 C1

PREDICCIÓN Y ANÁLISIS DE RENDIMIENTO ESTUDIANTIL

IA



Equipo: Los pandas

Integrantes:

- Miguel Andres Jaimes Ortiz - **2221895**
- Jeison Fernando Guarguati Anaya - **2221930**

RECAPITULANDO

PROBLEMA Y RELEVANCIA

El objetivo es identificar los factores que afectan el rendimiento académico de los estudiantes. Comprender estos factores permite implementar intervenciones tempranas, como tutorías, apoyo familiar o acceso a recursos, para mejorar los resultados académicos y reducir desigualdades en el aprendizaje.

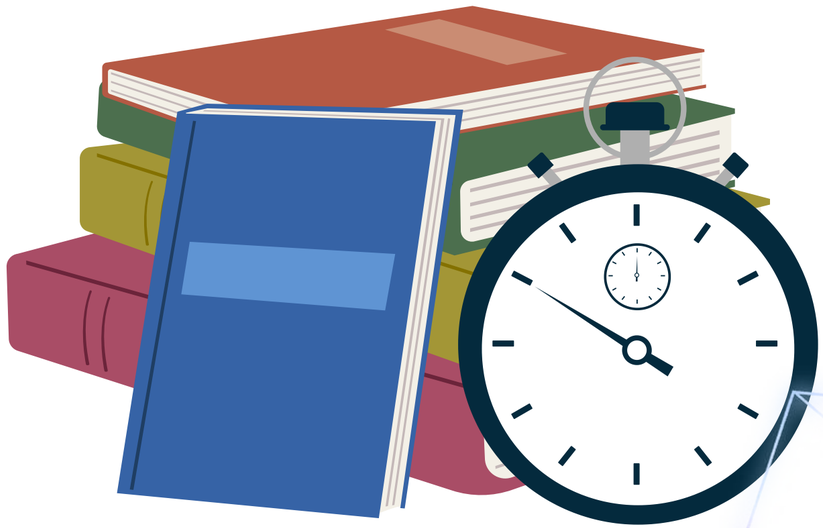
1. ANÁLISIS EXPLORATORIO DE DATOS

COLUMNAS DEL DATASET

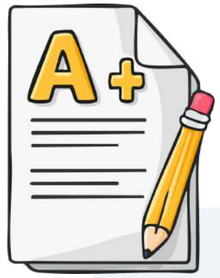
#	Columnas de características
0	Horas de Estudio
1	Asistencia
2	Participación de los padres
3	Acceso a Recursos
4	Actividades extracurriculares
5	Horas de sueño
6	Calificaciones anteriores
7	Nivel de motivación
8	Acceso a internet
9	Sesiones de tutoría

#	Columnas de características
10	Ingreso familiar
11	Calidad del profesor
12	Tipo de escuela
13	Influencia de los compañeros
14	Actividad física
15	Dificultades de aprendizaje
16	Nivel educativo de los padres
17	Distancia desde el hogar
18	Género
19	Puntaje del examen

Ground truth	
Regresión	Puntaje del Examen
Clasificación	Estudiantes Aprobados > 70

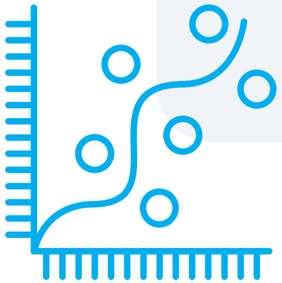


OBJETIVO DEL PROBLEMA A RESOLVER



REGRESIÓN

Buscar un modelo que permita predecir la nota de un estudiante, con base en factores como horas de estudio, asistencia, etc.



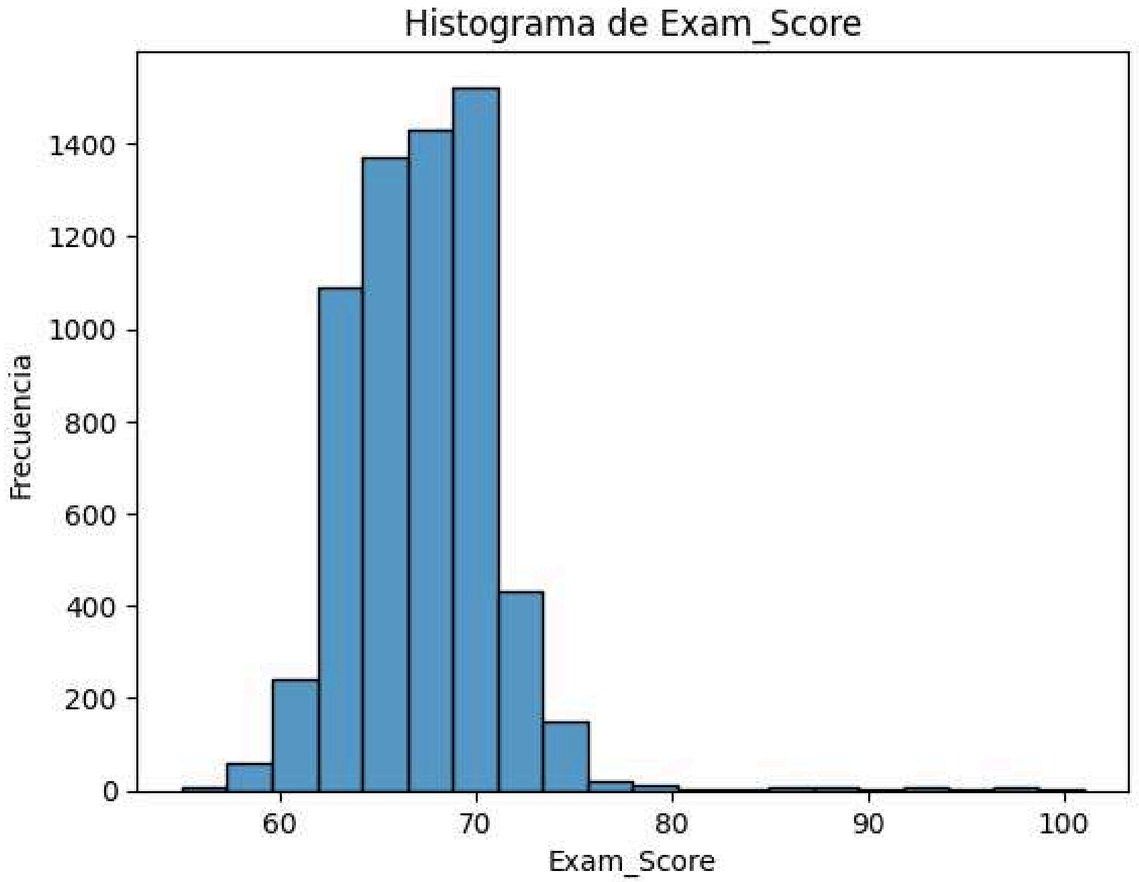
CLASIFICACIÓN

Buscar un modelo que permita predecir si un estudiante aprobará o no, con base en factores como horas de estudio, asistencia, etc.

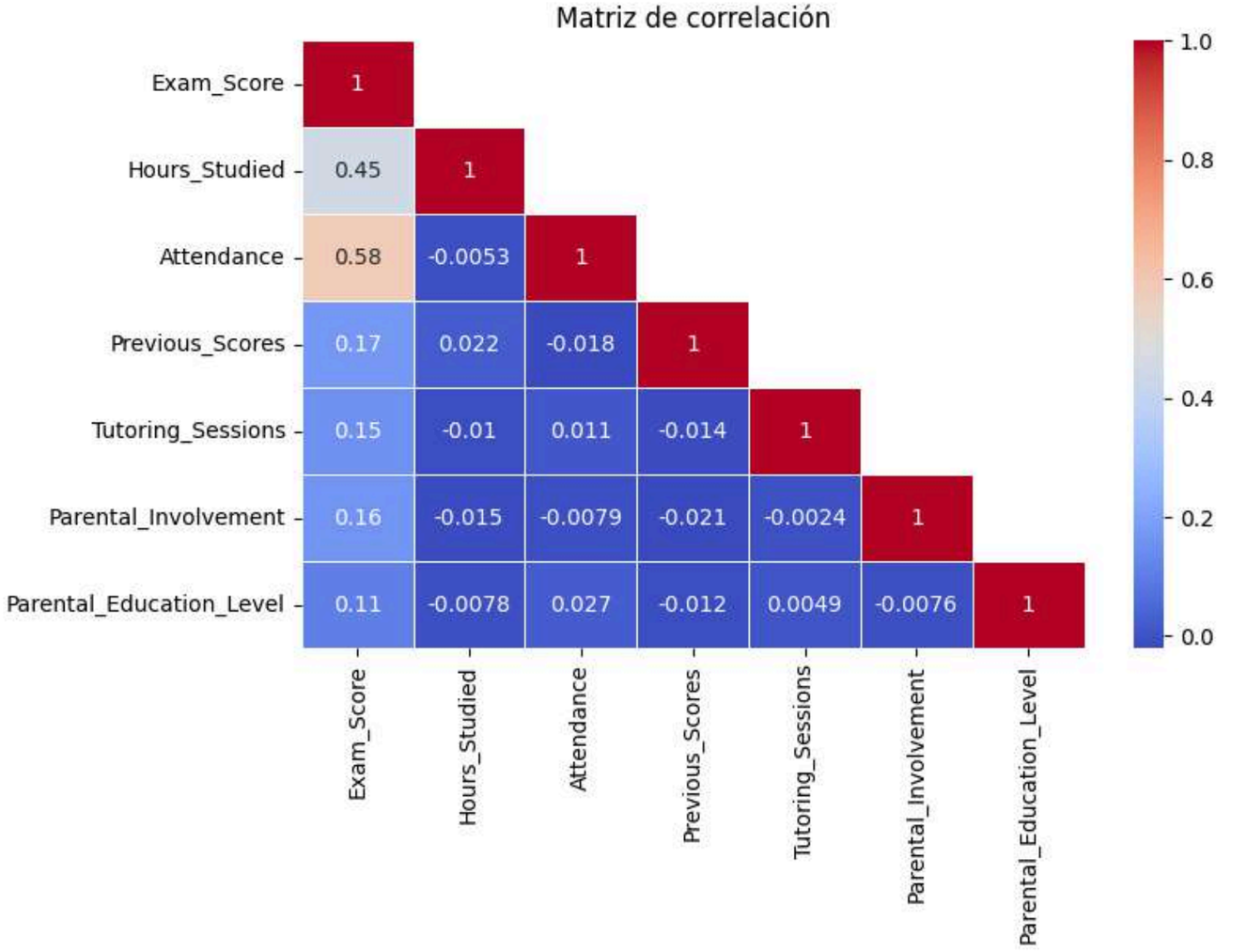
COLUMNAS DEL DATASET

Exploración de nuestra variable ‘**Exam_score’** que es importante para saber el comportamiento de las nota y ver si se pueden predecir.

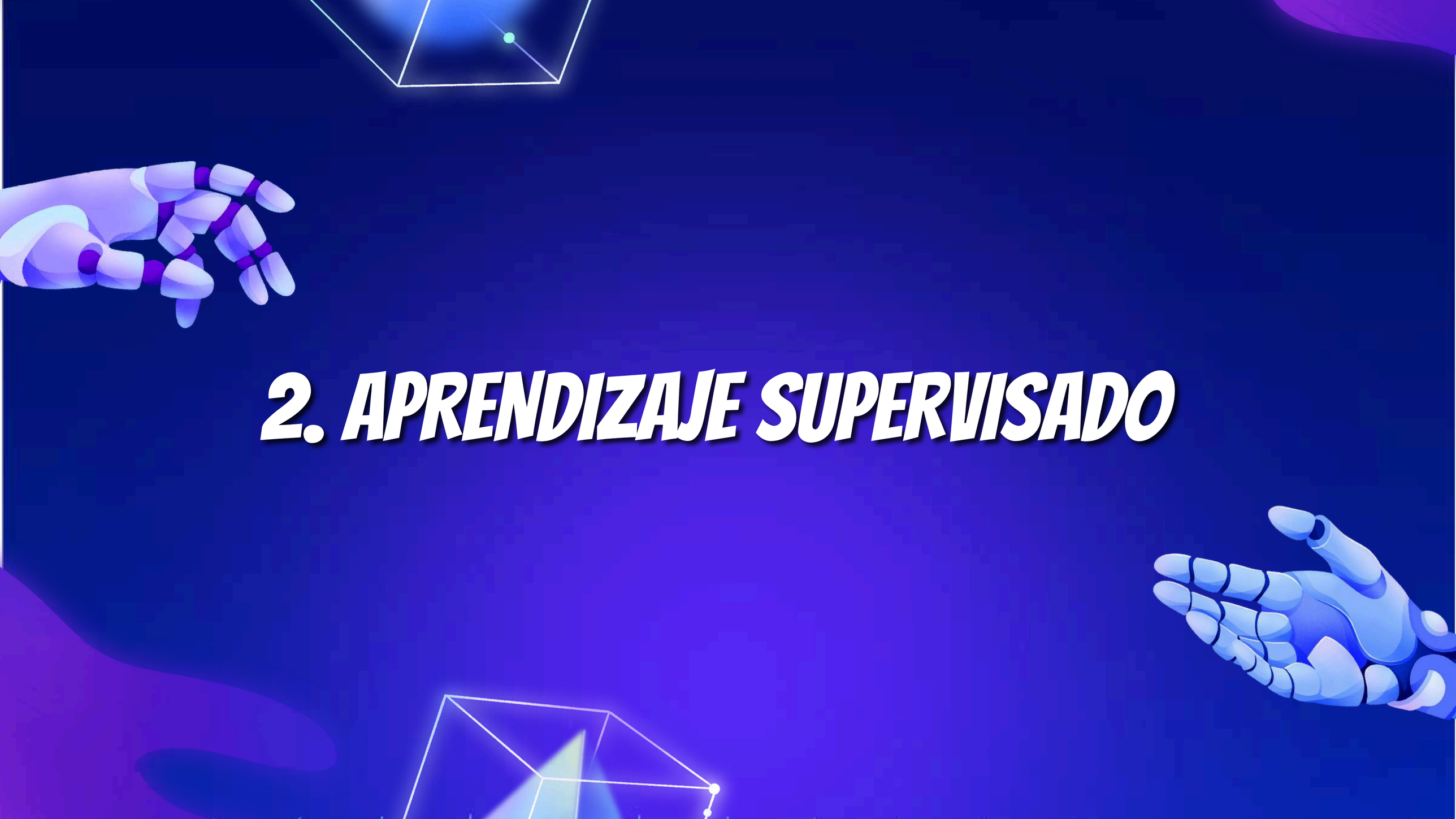
	Exam_Score
count	6378.000000
mean	67.252117
std	3.914217
min	55.000000
25%	65.000000
50%	67.000000
75%	69.000000
max	101.000000



MATRIZ DE CORRELACIÓN



	Exam_Score
Exam_Score	1.000000
Attendance	0.580259
Hours_Studied	0.445104
Previous_Scores	0.174283
Access_to_Resources	0.167856
Tutoring_Sessions	0.156829
Parental_Involvement	0.156014
Parental_Education_Level	0.105253
Peer_Influence	0.099133
Family_Income	0.094555
Motivation_Level	0.088502
Teacher_Quality	0.075107
Extracurricular_Activities_Yes	0.063063
Internet_Access_Yes	0.051124
Physical_Activity	0.025148
Gender_Male	-0.004932
School_Type_Public	-0.010868
Sleep_Hours	-0.017171
Learning_Disabilities_Yes	-0.083911
Distance_from_Home	-0.088083

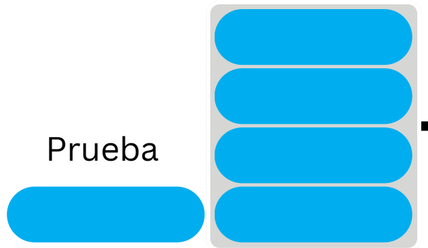

2. APRENDIZAJE SUPERVISADO

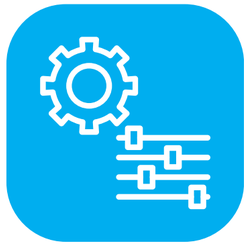
METODOLOGÍA

Partición del dataset

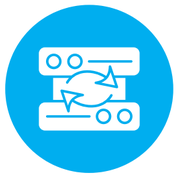
Entrenamiento

Prueba





Búsqueda de mejores hiperparámetros



Se usa CrossValidation

Entrenamiento

Predicciones

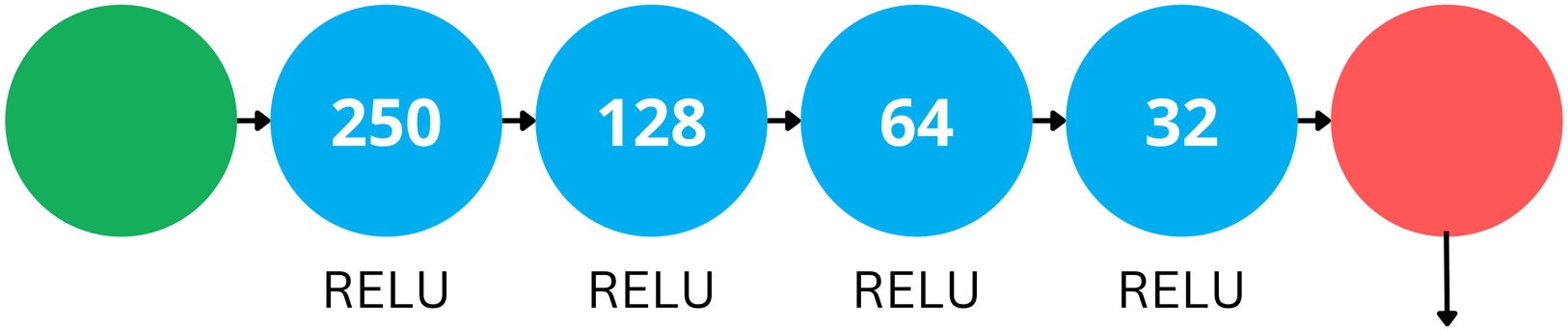
Prueba

Reporte de métricas

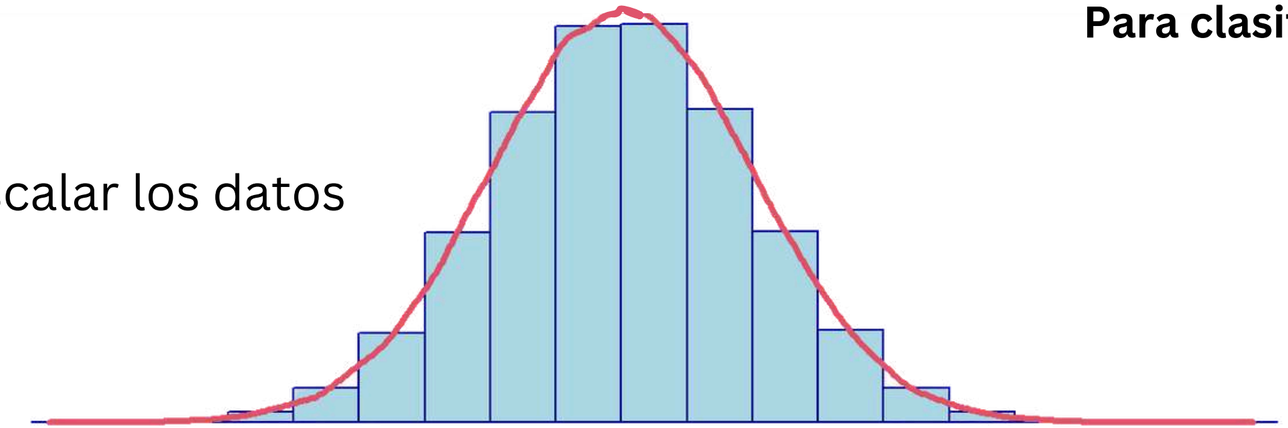
Análisis

CONFIGURACIÓN EXPERIMENTAL

PROCESO DE DEEP LEARNING



Escalar los datos

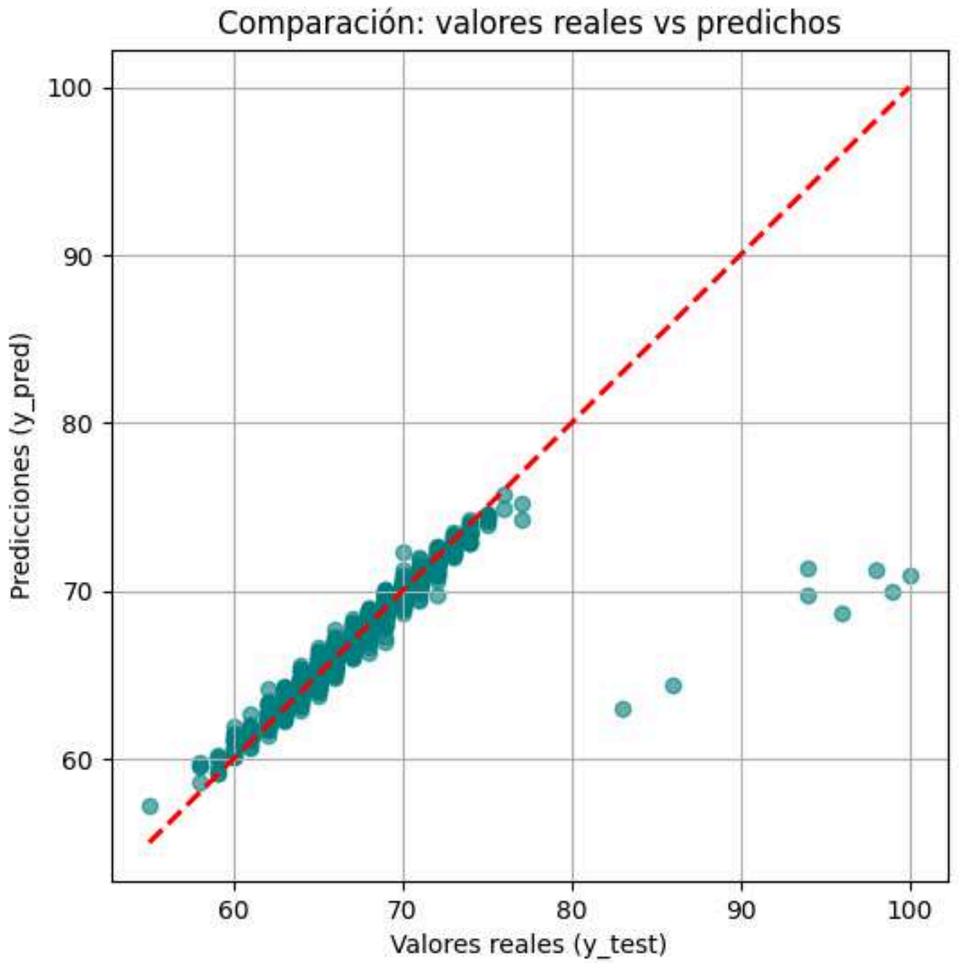


REGRESIÓN

REPORTE DE RESULTADOS

Modelos	MAE	MSE
Decisión Tree	**1.6321**	**8.4604**
Random Forest	1.0972	5.6633
Support Vector Machine	**0.5817**	**4.3163**
Deep learning	0.6443	4.4905

COMPARACIÓN DE VALORES



MEJOR MODELO

SVR

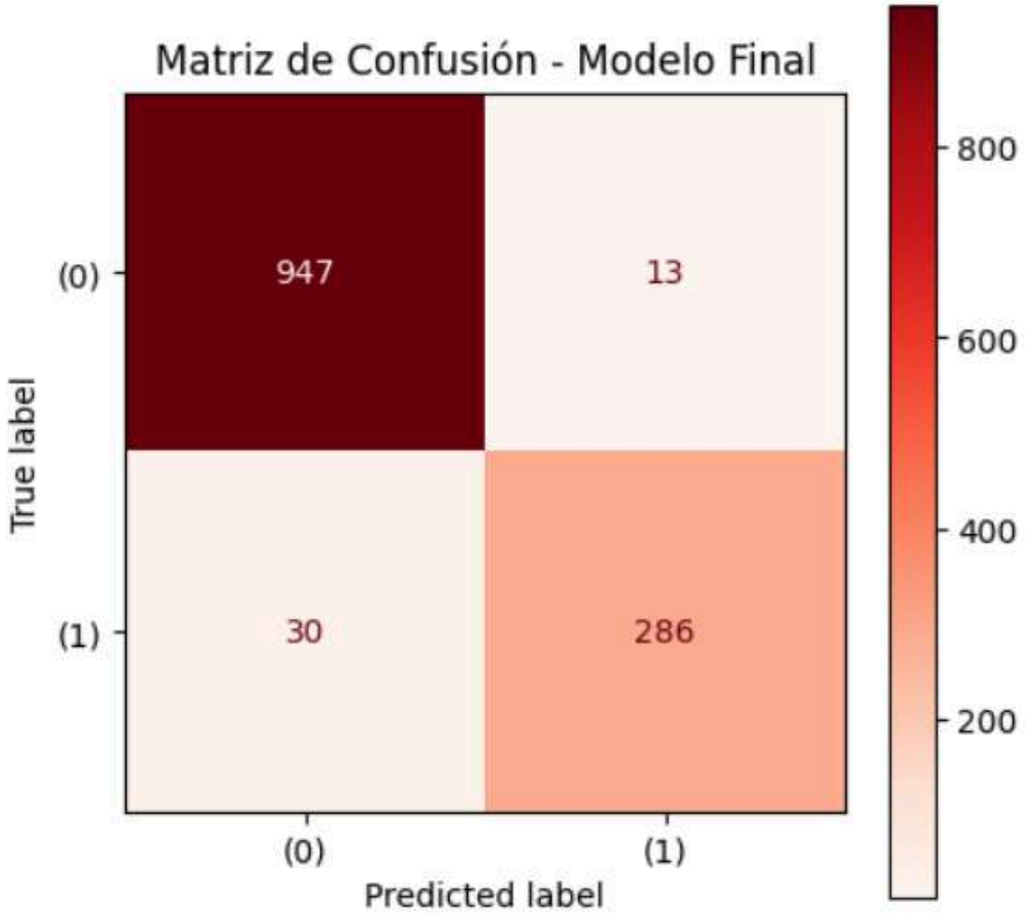
CLASIFICACIÓN

DISTRIBUCCIÓN DE CLASIFICACIÓN

REPORTE DE RESULTADOS

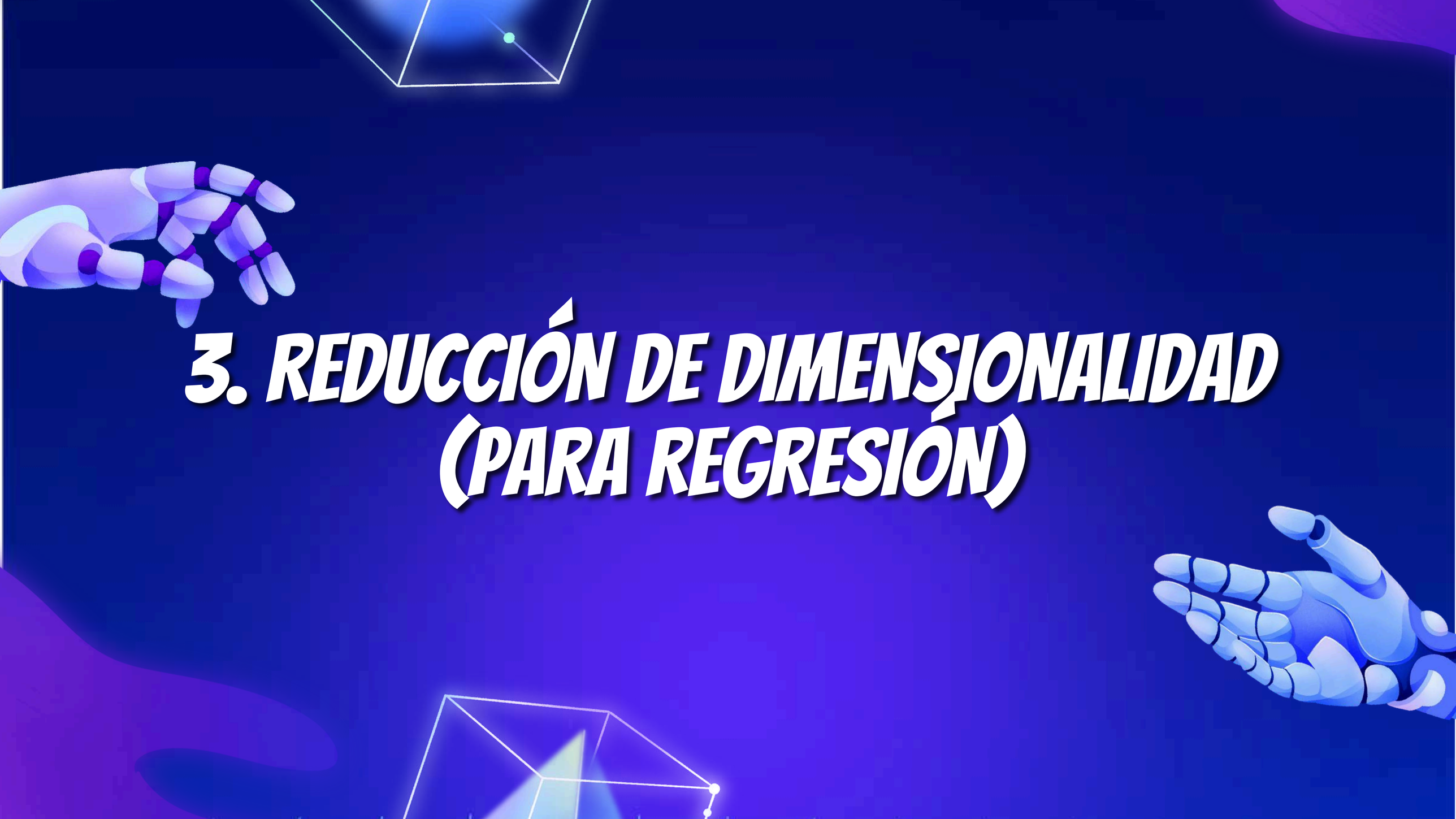
Modelos	Balanced Accuracy	Sensibilidad	Especificidad	Precisión	F1-Score
Gaussian Bayes	0.8783	0.7785	0.9781	0.9213	0.8439
Decision Tree	**0.8312**	**0.7373**	**0.925**	**0.7639**	**0.7504**
Random Forest	0.8672	0.7563	0.9781	0.9192	0.8299
Support Vector Machine	0.8858	0.8101	0.9615	0.8737	0.8407
Deep learning	**0.9458**	**0.9051**	**0.9865**	**0.9565**	**0.9301**

MATRIZ DE CONFUSIÓN

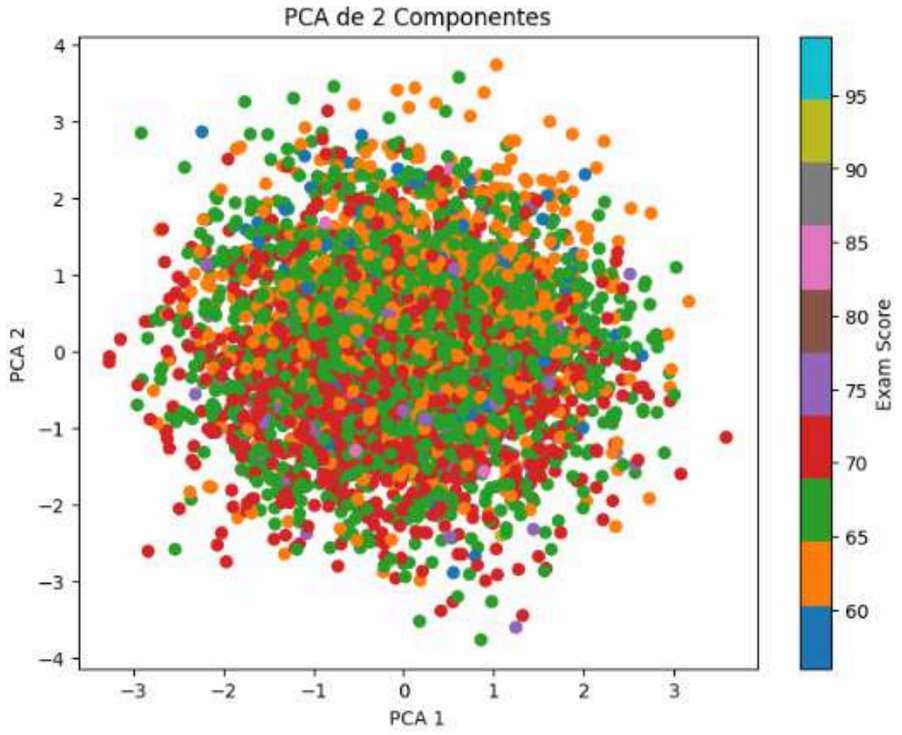


MEJOR MODELO

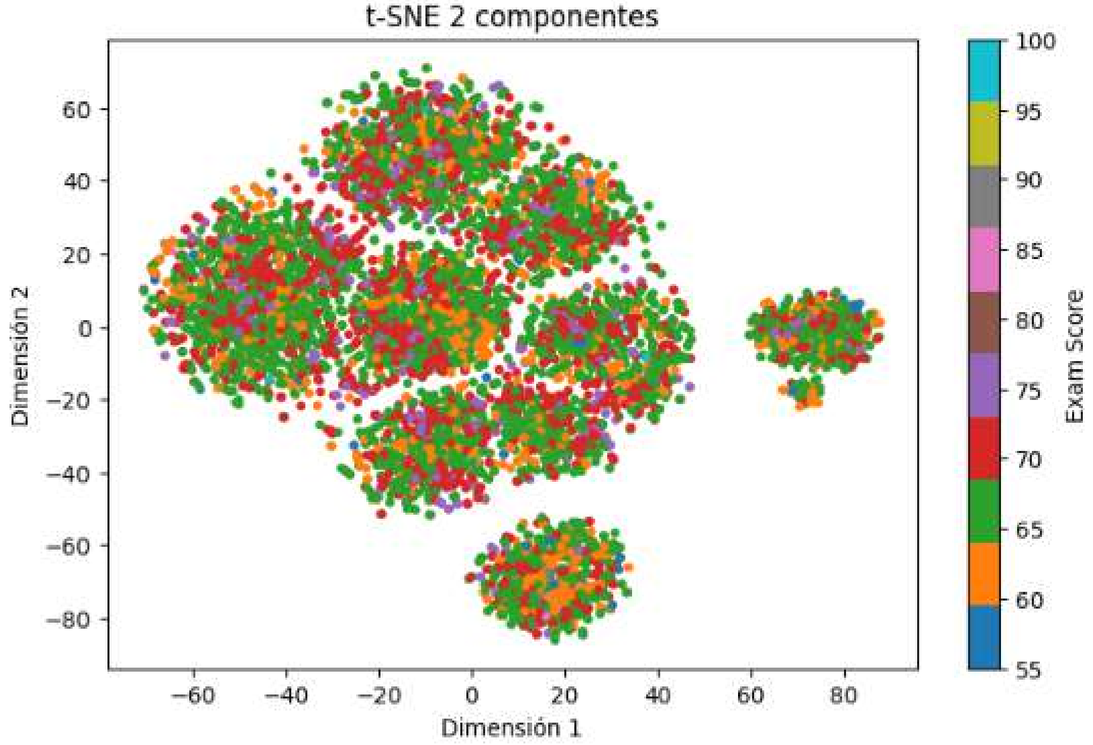
Deep learning


3. REDUCCIÓN DE DIMENSIONALIDAD
(PARA REGRESIÓN)

PCA

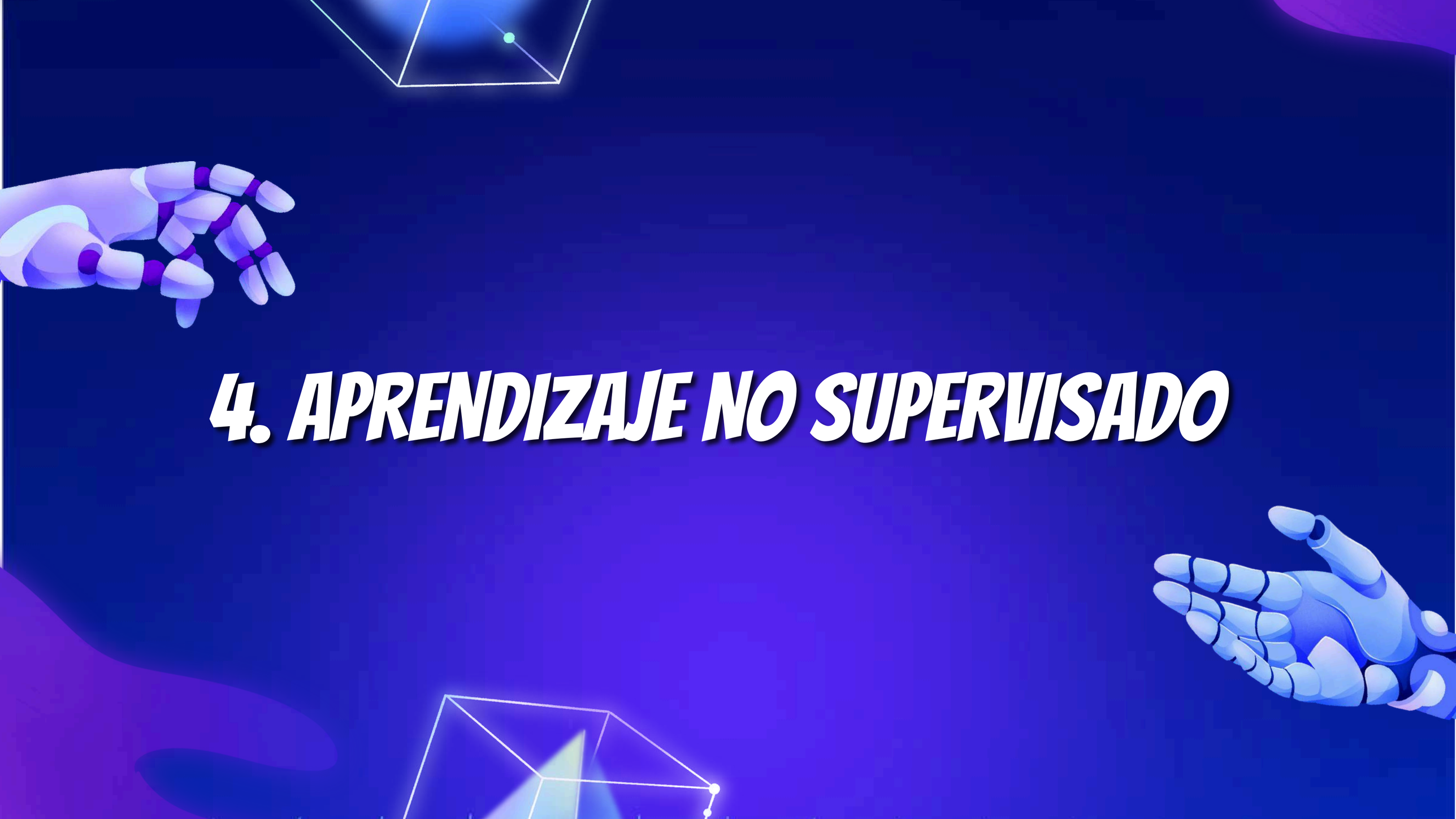

PCA de 2 Componentes
PCA 2
PCA 1
Exam Score
95
90
85
80
75
70
65
60

T- SNE


t-SNE 2 componentes
Dimensión 2
Dimensión 1
Exam Score
100
95
90
85
80
75
70
65
60
55

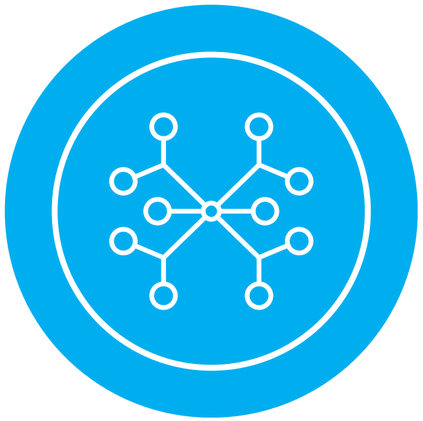
COMPARATIVA DE RESULTADOS

REDUCCIÓN UTILIZADA	MAE	MSE
Sin reducción	0.5832	4.3200
PCA (19 componentes)	0.4298	4.1260
t-SNE (2 dimensiones)	2.1927	8.2685

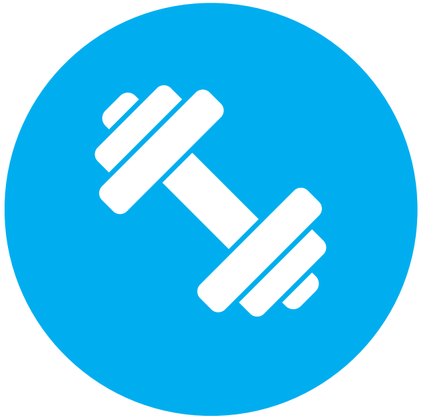

4. APRENDIZAJE NO SUPERVISADO

METODOLOGÍA

Selección de hiperparámetros



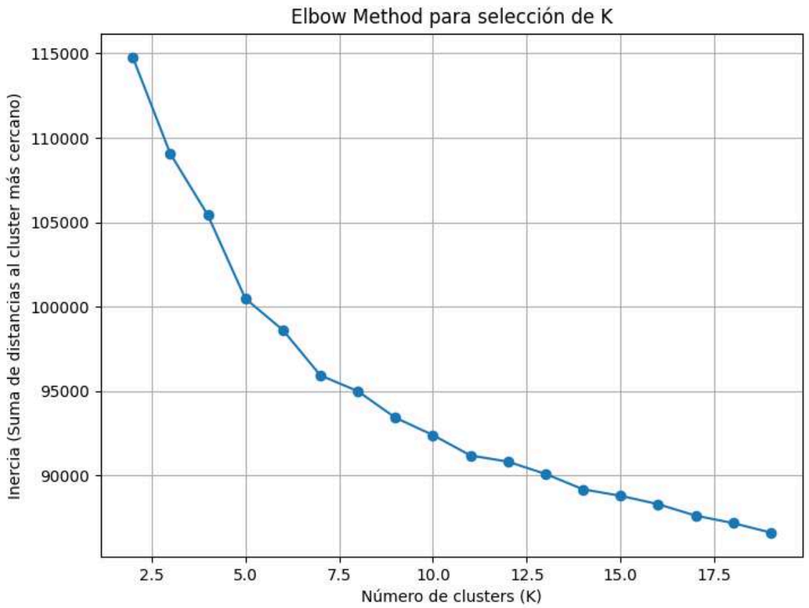
Entrenamiento de los modelos



Evaluación de los modelos

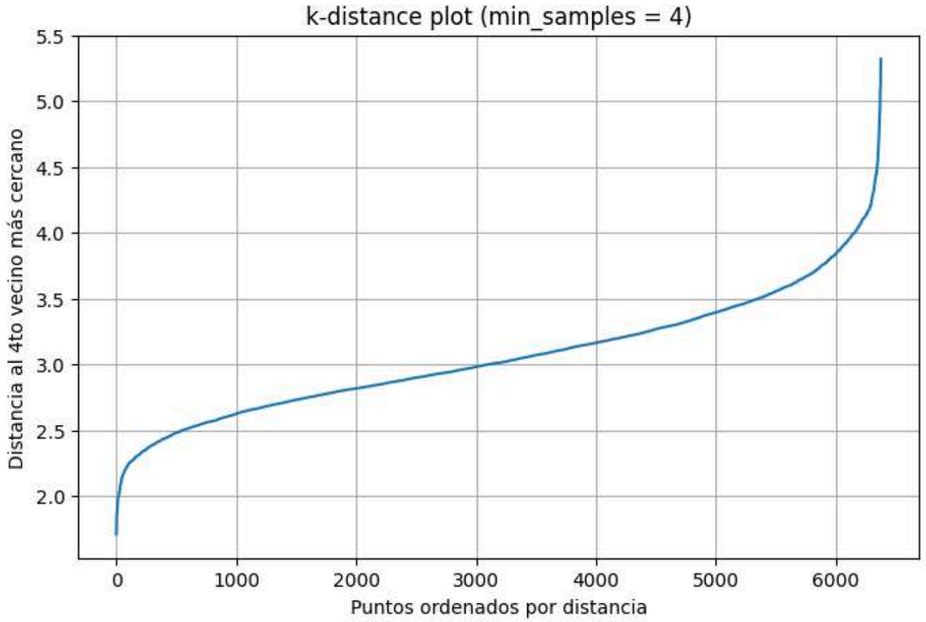
MEJORES HIPERPARÁMETROS

K - MEANS



k = 6 o 7

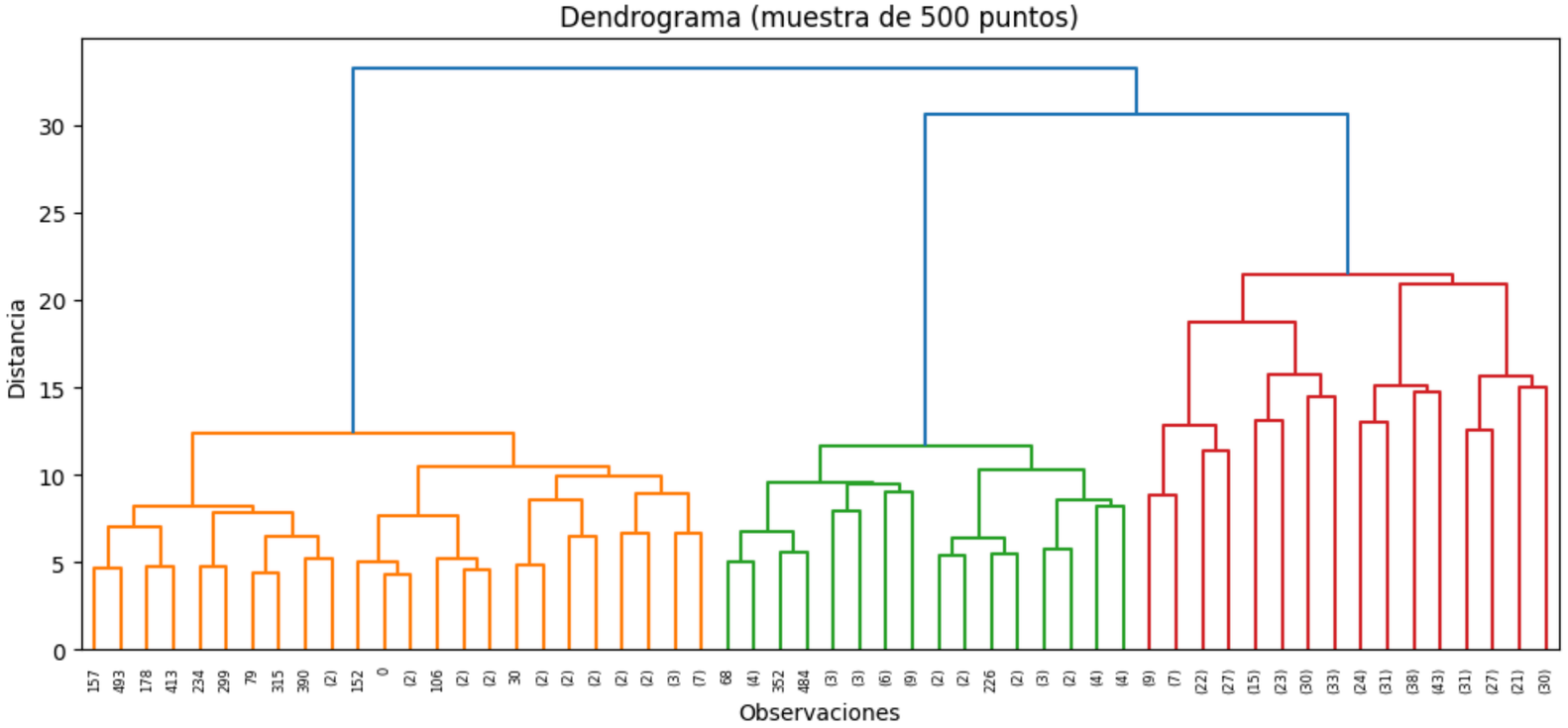
DBSCAN



k = 4

MEJORES HIPERPARÁMETROS

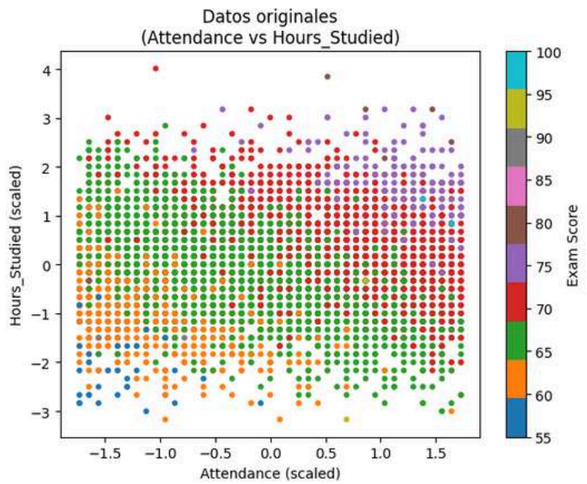
Agglomerative Clustering

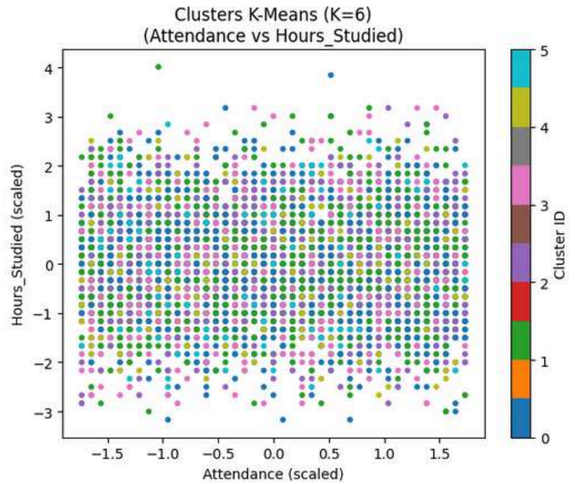


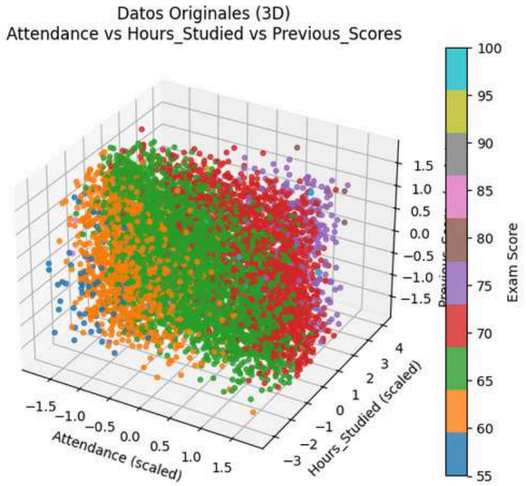
k = 3 - 6

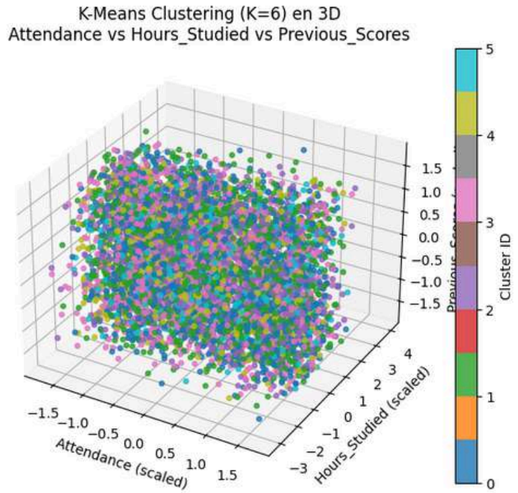
RESULTADOS DE ENTRENAMIENTO

K - MEANS



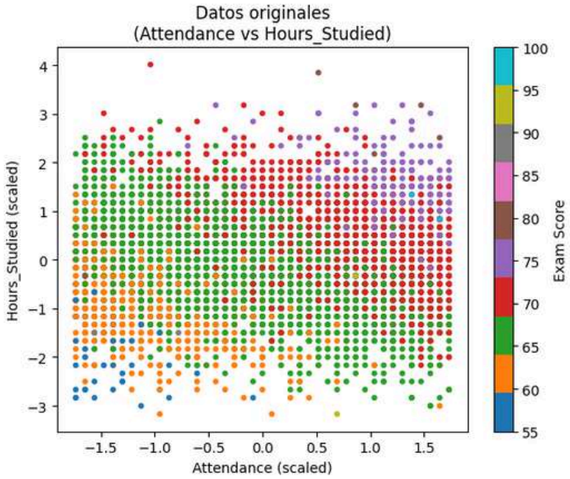


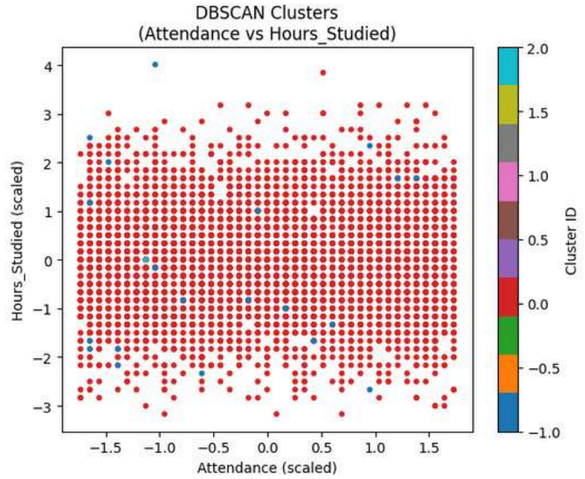


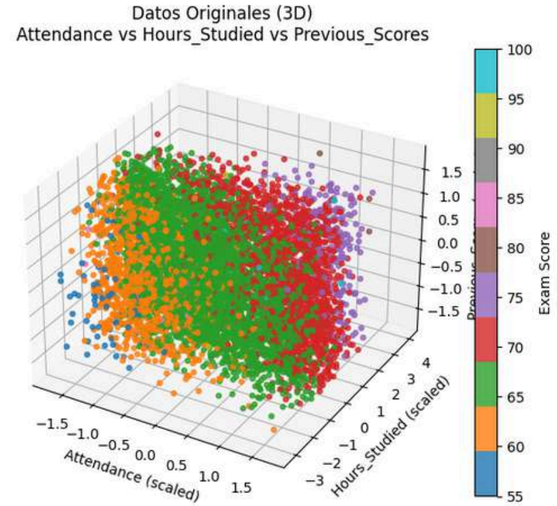


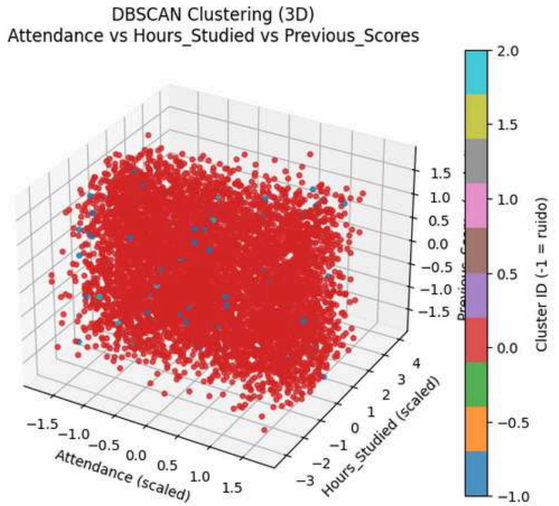
RESULTADOS DE ENTRENAMIENTO

DBSCAN



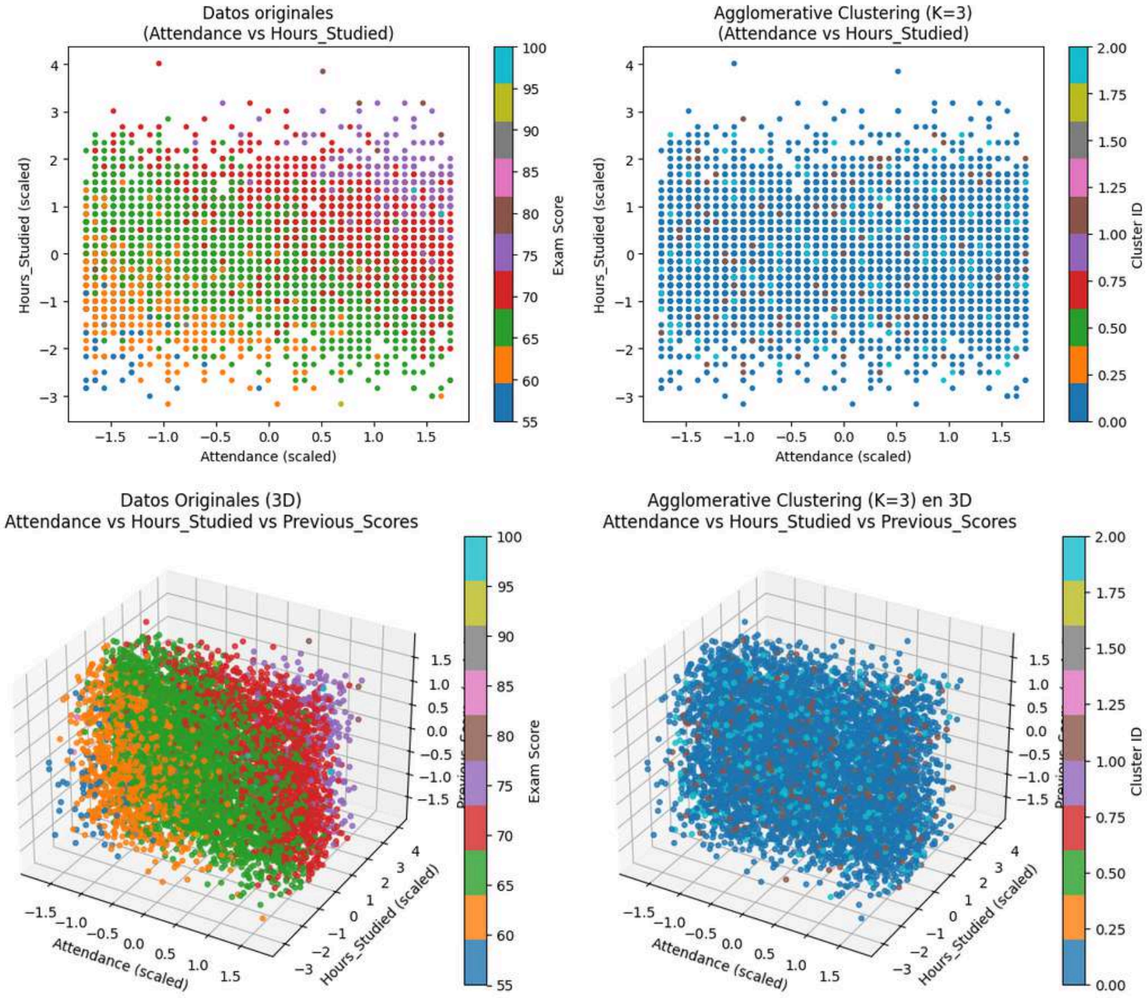






RESULTADOS DE ENTRENAMIENTO

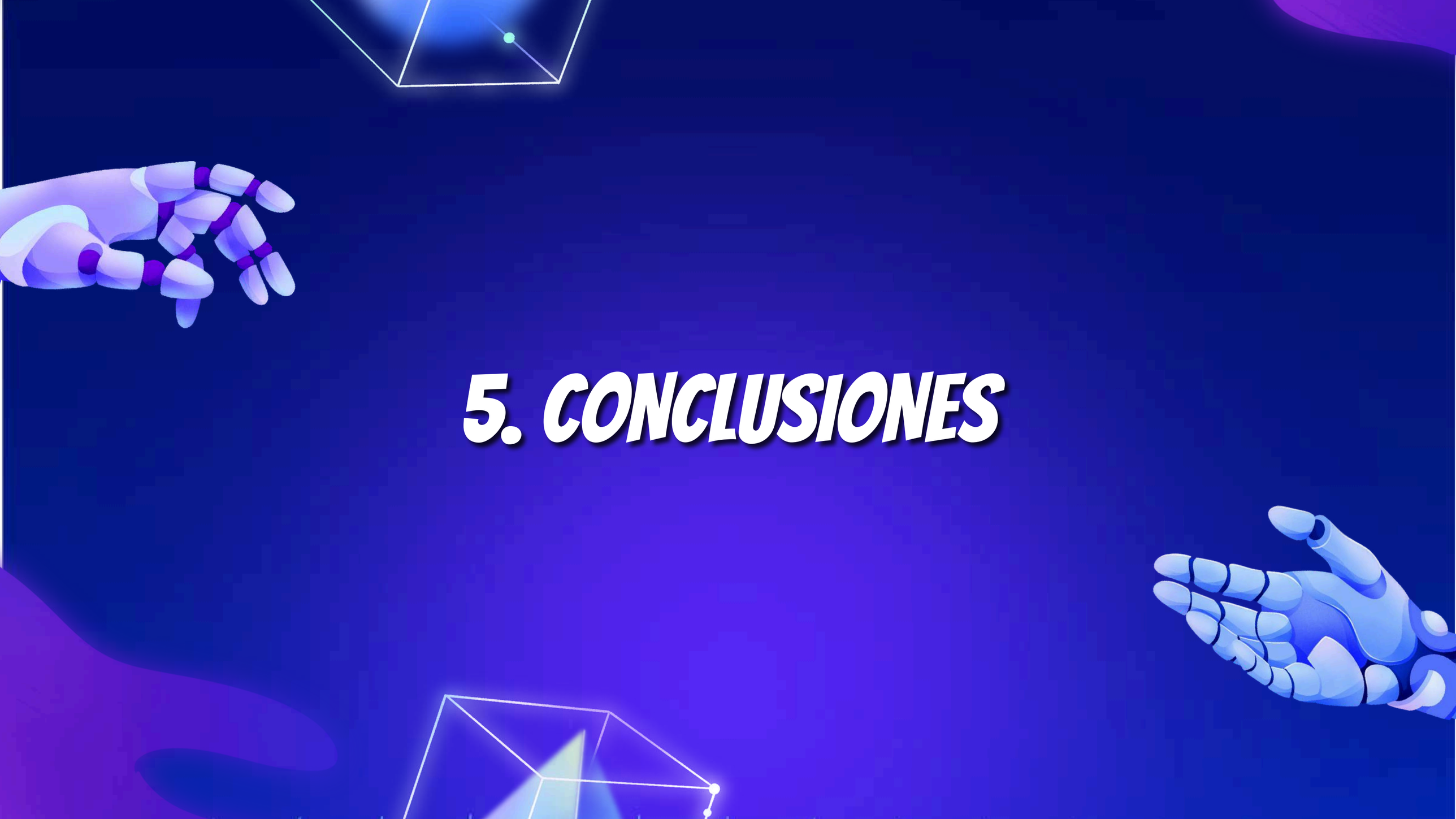
AGGLOMERATIVE



EVALUACIÓN DE LOS MODELOS

SILHOUETTE SCORE

ALGORITMO	N° DE CLUSTERS	SCORE
K-Means	6	0.0573
DBSCAN	4	0.1494
Agglomerative	3	0.1364


5. CONCLUSIONES

- Se identificaron características con una relación significativa que permitieron entrenar los modelos, como la asistencia y las horas de estudio. En la vida real, estos factores influyen directamente en los resultados obtenidos en los exámenes.

- El modelo de regresión SVR mostró un rendimiento sólido en las métricas de evaluación. Sin embargo, se requiere disponer de más datos en las regiones donde los puntajes son mayores a 80 y menores a 55.

- El modelo de clasificación basado en Deep Learning mostró un excelente rendimiento en sus métricas generales, alcanzando niveles de precisión superiores al 90%.

- La reducción dimensional no mejora el rendimiento de los modelos cuando se eliminan variables, ya que cada una aporta información relevante.

- El aprendizaje no supervisado no obtuvo un buen rendimiento porque los datos no presentan agrupaciones claras. En lugar de formar clústeres definidos, las variables cambian de manera gradual, lo que dificulta que los algoritmos identifiquen separaciones naturales dentro del conjunto de datos.


¡GRACIAS!