Predicción de banda prohibida para óxidos de alta entropía (Fluorite y Mukesh) con machine learning y deep learning

Wilhelm David Buitrago Garcia[1] José Luis Rueda Mayorga[1] y Hector Daniel Vasquez Rivera[1,2]

[1]Estudiante de pregrado en Ingeniería Mecatrónica.
[2]Estudiante de pregrado en Estadística.
Universidad Nacional de Colombia sede de La Paz.

4 de abril del 2024

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Band gap como característica del material

Es de vital importancia conocer y caracterizar las propiedades de los nuevos materiales para hacer un uso apropiado de los mismos. En esta lista de características, el band gap (Energía de banda prohibida) juega un papel importante ya que, permite clasificar los materiales dentro de tres (3) grande grupos: Conductores, Semiconductores y Aislantes.

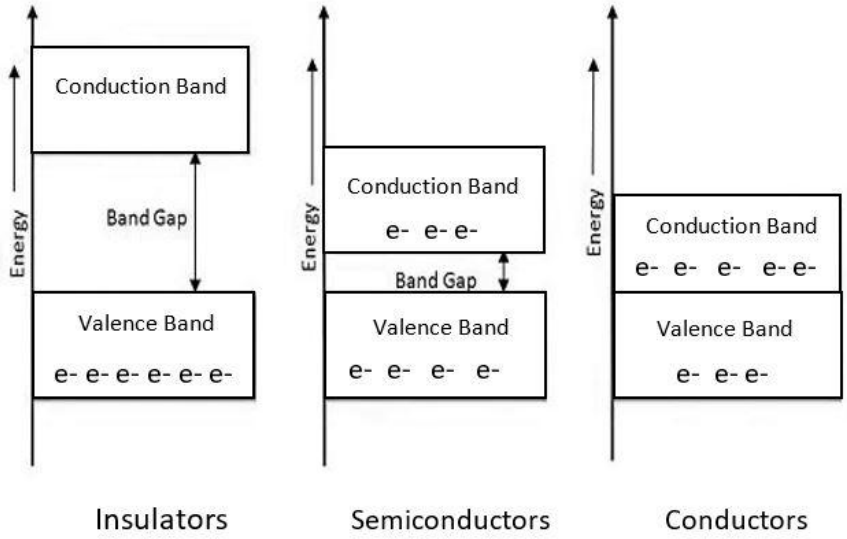


Figura 1: Clasificación de materiales segun su band gap. Tomada de Explain the formation of energy bands in solids. On the basis of energy bands distinguish between a metal, a semiconductor and an insulator (s.f.)

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Revisión de los datos

Se presentan dos datasets en forma de .csv correspondientes a información obtenida de espectroscopia UV-Vis transformada a puntos de un diagrama Tauc's realizada a dos óxidos de alta entropía: Fluorite con 530 datos y Mukesh con 2560 datos.

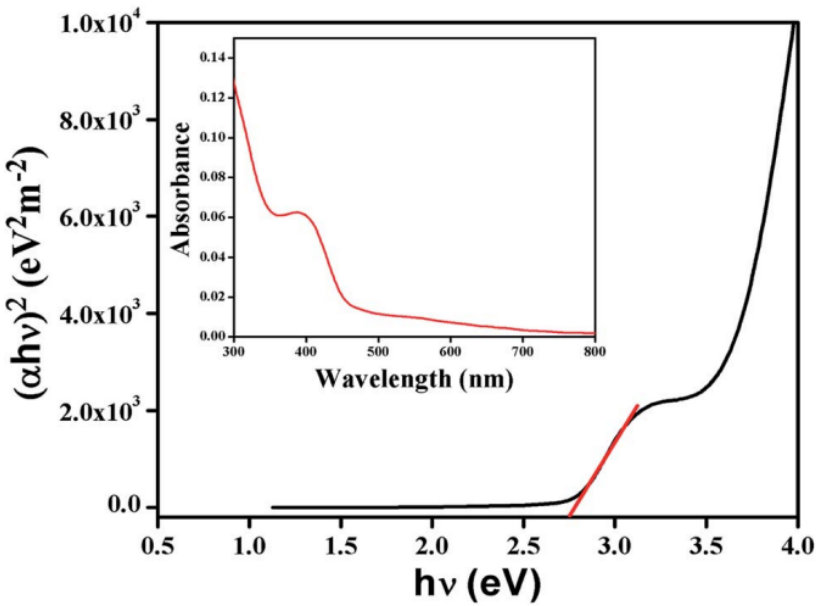
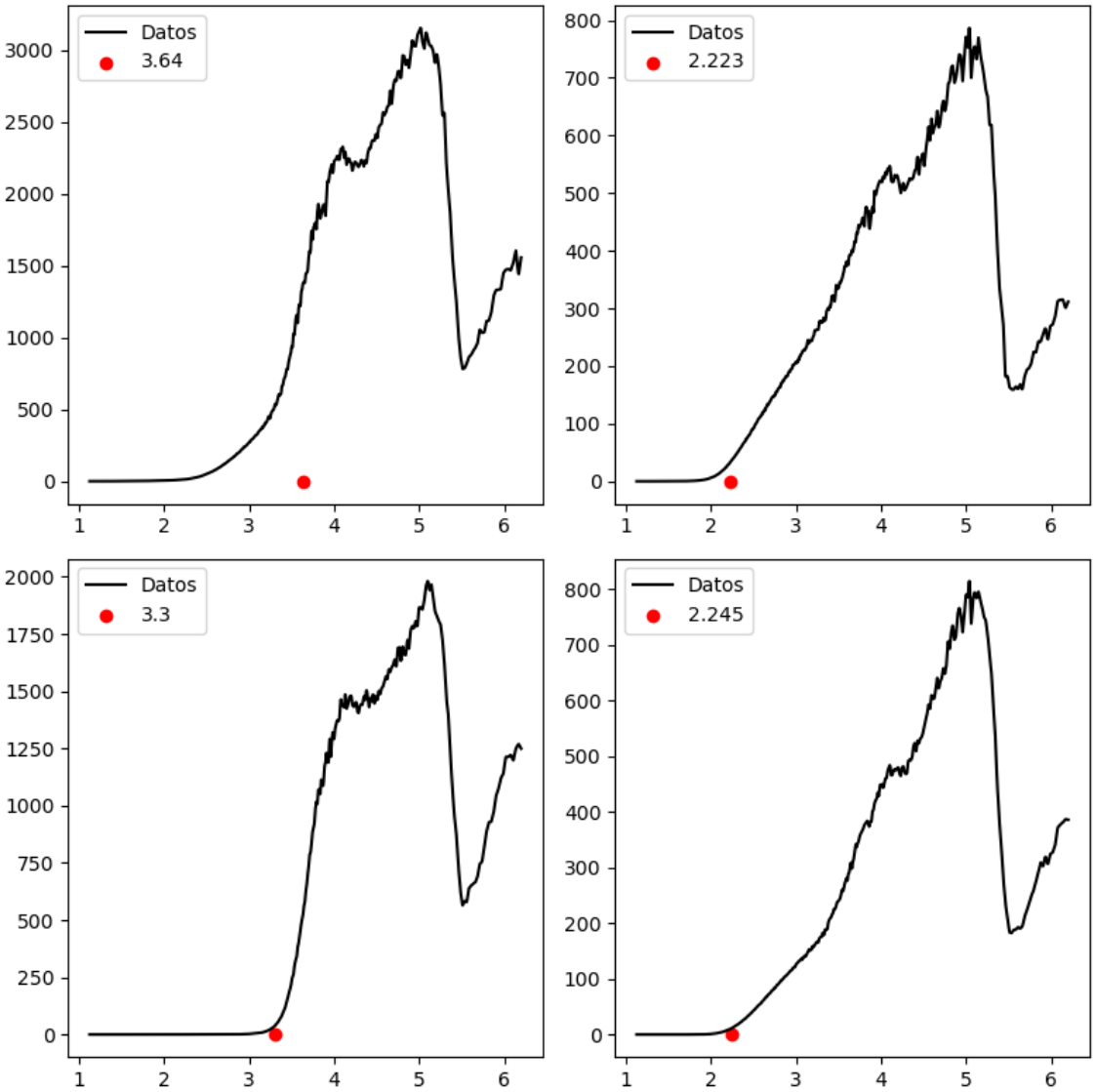


Figura 2: Ejemplo de UV-Vis (Wavelength vs Absorbance) visto como un gráfico Tauc's, Tomada de Photosensitive Schottky barrier diode behavior of a semiconducting Co(III)–Na complex with a compartmental Schiff base ligan (2019)

UNIVERSIDAD NACIONAL DE COLOMBIA

Revisión de los datos



Revisión de los datos

Cada conjunto de datos presenta dos columnas: x1,x2. Donde x1 es una lista unidimensional (1D) que representa la energia de los fotones incidentes (eV) y x2 es una lista unidimensional (1D) de datos correspondientes al absorbancia del material para fotones incidentes. Ambas en forma de cadena de texto (string).

	x1	x2
0	"[1.12726309532197, $\cdots$, 1.1293180115103607]"	"[1.12726309532197, $\cdots$, 1.1282891794029504]"

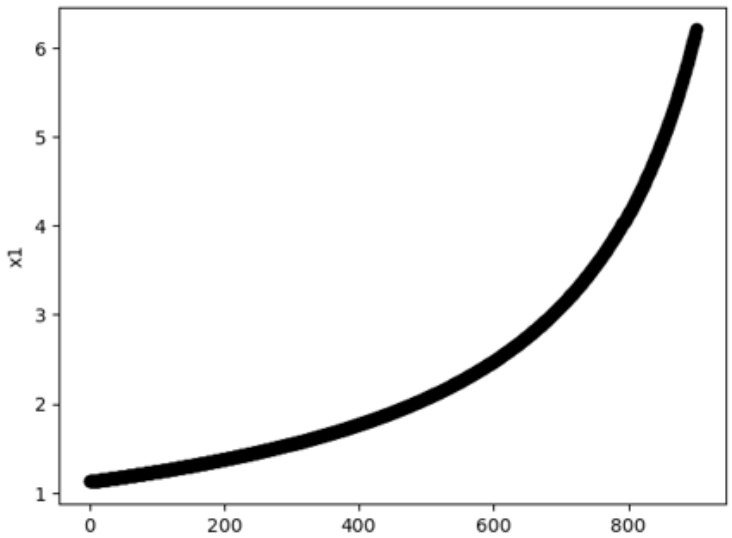
Tabla 1: Ejemplo de un dato de uno de los dataset

Cabe aclarar que adicionalmente, el dataset de Flourite presenta una columna denominada "band_gap_std", sin embargo debido a que solamente se encuentra en este dataset y es una medida hallada a partir de la predicción a realizar, se ha decidido por la eliminación de la misma.

UNIVERSIDAD NACIONAL DE COLOMBIA

Revisión de los datos

Dado que la columna x1 esta presente en ambos datasets con el mismo rango y forma se decidió eliminarla, sin embargo x2 se encuentra relacionado con x1 y este ultimo no presenta un comportamiento normalmente espaciado y eliminarlo representaría perdida de información en x2, por tanto se decide hacer una interpolación a los datos, haciendo que queden equidistantes:



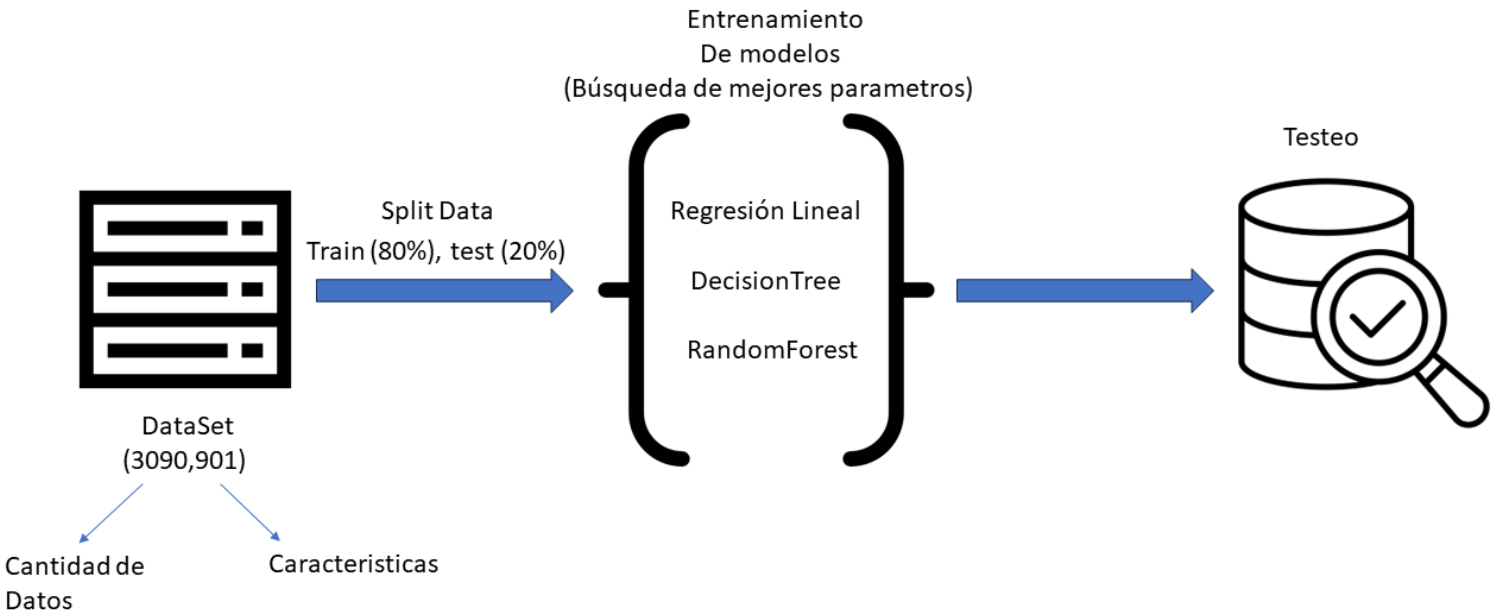

```
#Interpolacion para que los datos en x1 sean x distantes
x = np.linspace(data_orig[0][0].min(),data_orig[0][0].max(),900)
temp = np.zeros((data_orig.shape[0],900))
for i in range(data_orig.shape[0]):
    temp[i,:] = np.interp(x,data_orig[i][0][:-1],data_orig[i][1][:-1])
```

UNIVERSIDAD NACIONAL DE COLOMBIA

Método 1

Este método consiste en entrenar algoritmos de Machine Learning usando la totalidad de los datos 1D de x2 (901 elementos) de cada dato, es decir que, se tratara cada punto o valor de este arreglo 1D como una característica o predictor del problema, a priori se espera entonces que, este método tenga un buen rendimiento pero, sea costoso computacionalmente.

Se dividió el dataset en train/test(20 %) y se entrenaron distintos modelos de machine learning de regresión para conocer sus resultados. Para la búsqueda de mejores hiperparametros se utilizo el metodo RandomizedGridCV de sklearn.



Método 2

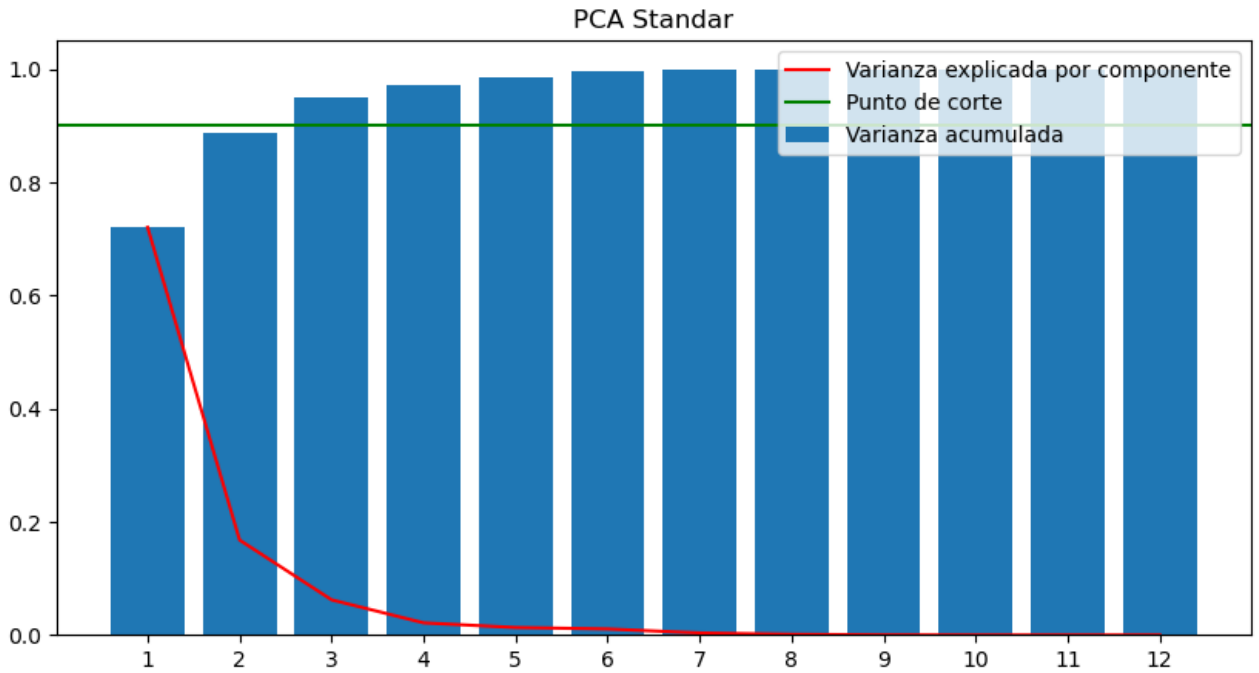
En este método se exploro una posible extracción de características a los arreglos 1D, se decidió extraer características básicas como:

- Valor Máximo
- Valor Medio
- Desviación Estándar
- Asimetría
- Curtosis
- Pendiente Mínima
- Pendiente Máxima
- Pendiente Media
- Área bajo la curva
- Cantidad de puntos hasta alcanzar valor máximo (Subida)
- Cantidad de puntos restantes después de alcanzar valor máximo (Bajada)

UNIVERSIDAD NACIONAL DE COLOMBIA

Método 2

Posteriormente, se escalan los datos usando MinMaxScaler (MMScaler), StandarScaler (StadScaler) y RobustScaler (RobScaler) para luego ser normalizandos por medio de PowerTransform con el método yeo-johnson. En este caso, se tiene todavía una moderada dimensionalidad de los datos, por lo que se realiza PCA para la reducción de la misma.

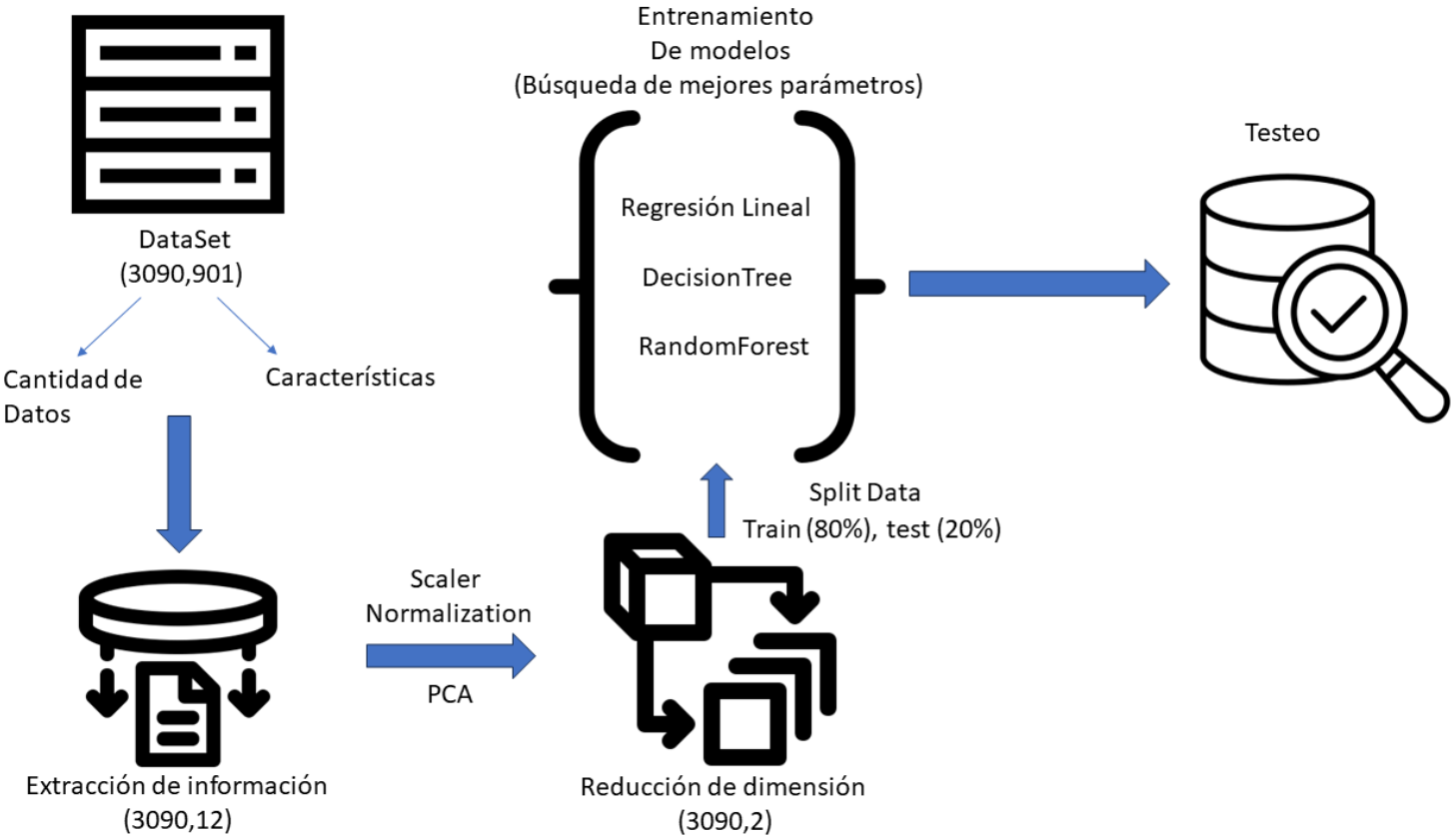


Como se observa, una gran variabilidad de los datos es encontrada en las primeras dos componentes, por lo que es aceptable reducir la dimensionalidad eligiendo representar las variables en las primeras dos componentes.

UNIVERSIDAD NACIONAL DE COLOMBIA

Metodo 2

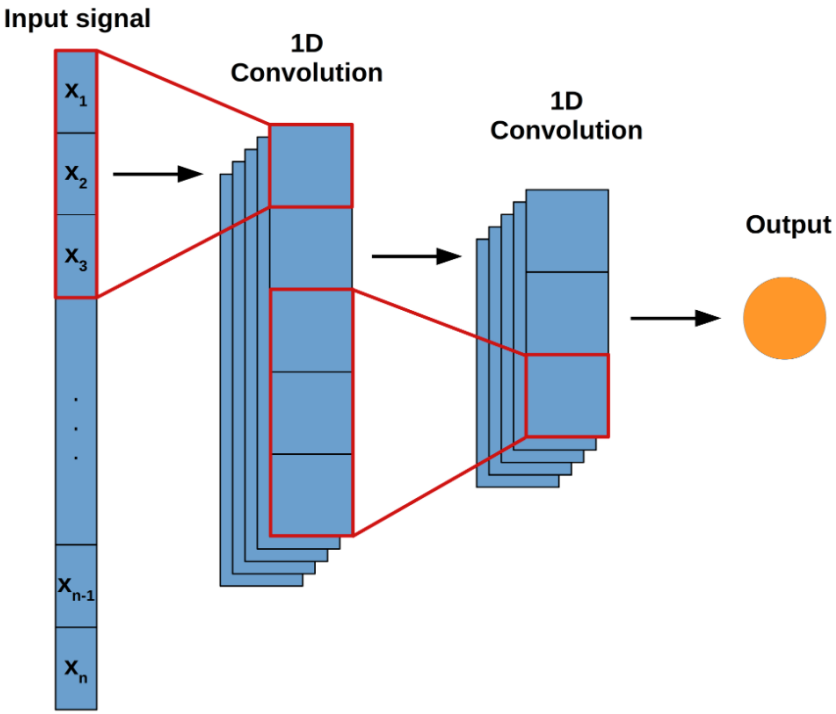
Se divide el conjunto de datos entre train/test(20 %) para el posterior entrenamiento de modelos



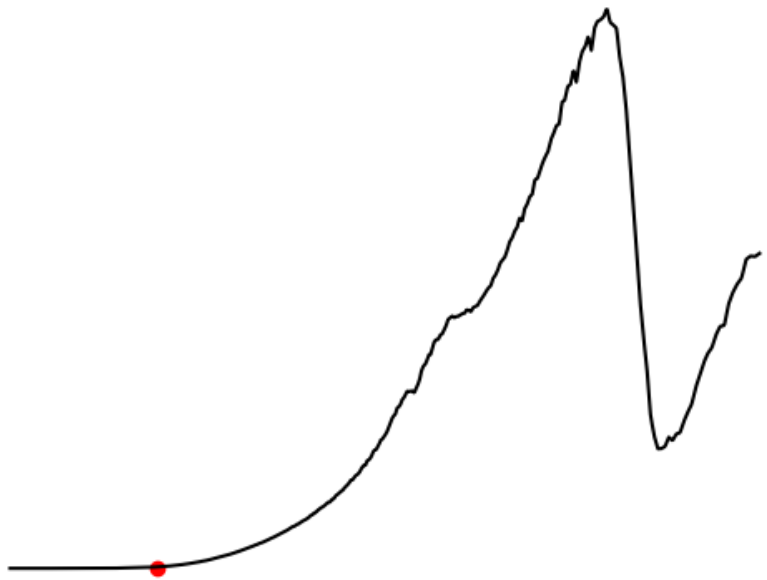
Hipótesis

El problema se intento solucionar desde dos enfoques aplicando redes neuronales convolucionales, para ambos casos los datos tendrían dos canales uno con X_2 y X_2'.

- **Convolución 1D:** Se creo una CNN con convoluciones1d para extraer las características de los arreglos debidamente pre-procesados.
- **Gráficas:** se crearon las graficas de los datos con las cuales se entrenaria la CNN.



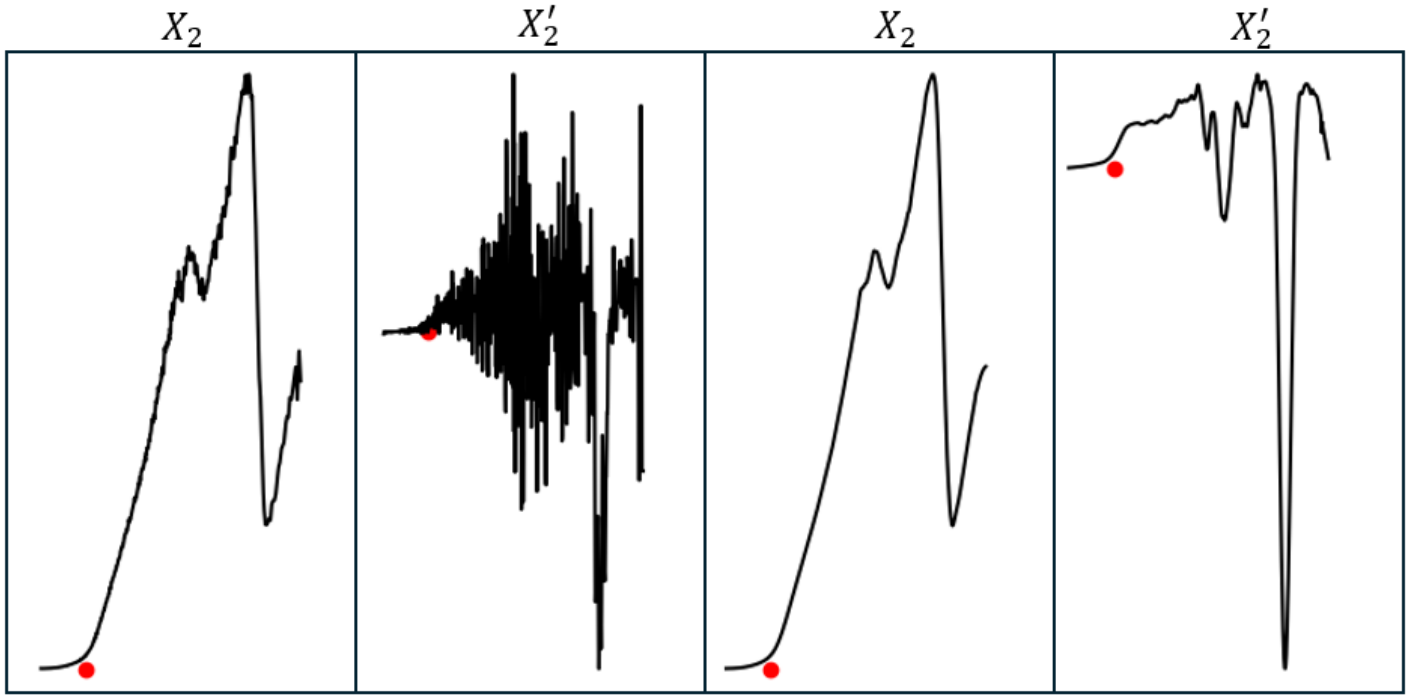
(a) Descripción de la primera imagen



(b) Descripción de la primera imagen

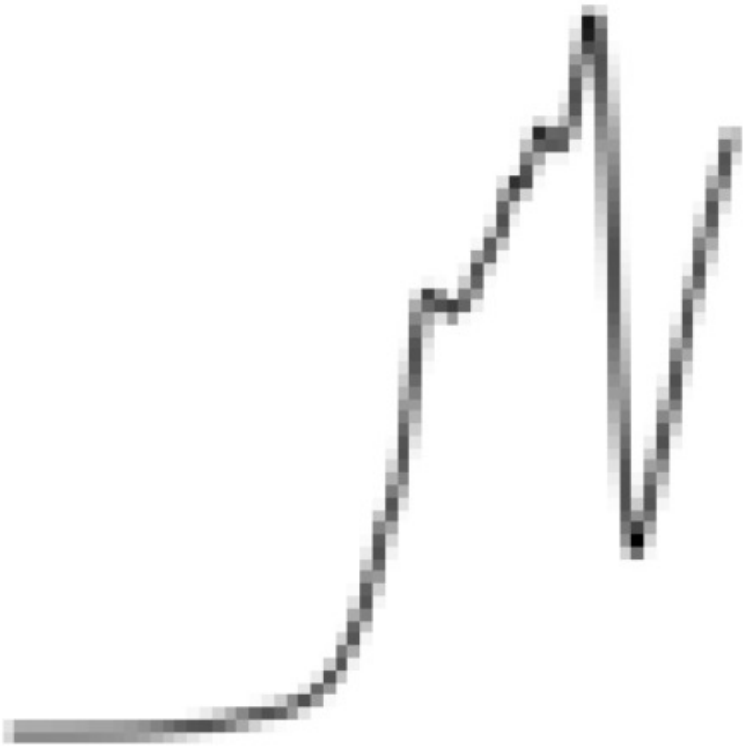
Preprocesamiento

- Para disminuir el ruido de los datos y conseguir calcular una derivada mas aproximada se aplico el filtro de Savitzky-Golay el cual permite suavizar los datos y eliminar el ruido de alta frecuencia.

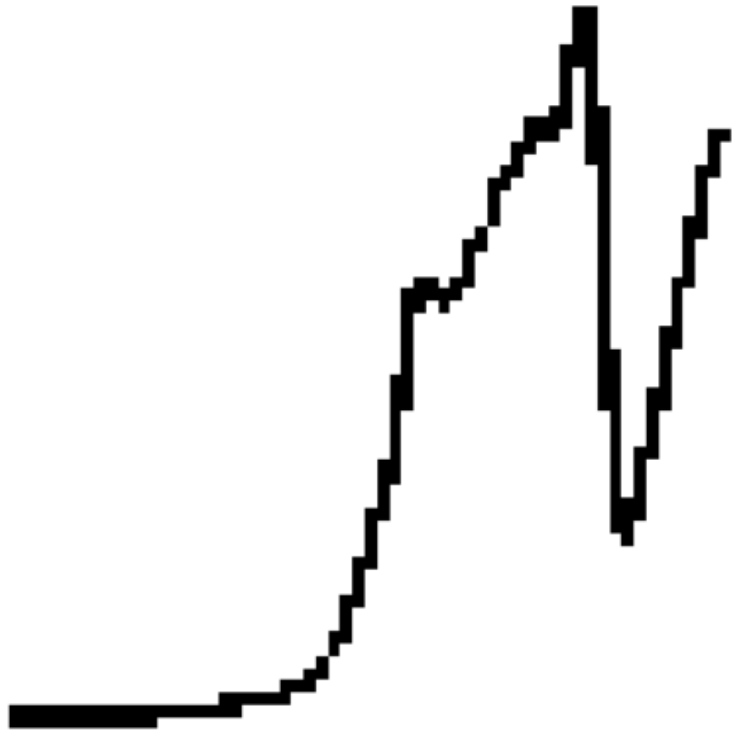


Binarización

- Para la binarización, primero se redujo las dimensiones de 270x270 a 64x64, luego se normaliza con una media y desviación estandar de 0.5. Luego se aplica un umbral (threshold) a la imagen para convertirla en binaria. Todos los píxeles por debajo del umbral se establecían como cero (negro), mientras que los pixeles por encima del umbral es 1 (blanco).



(a) Sin Binarizar



(b) Binarización

Binarización

También se realizo lo contrario, convirtiendo todos los píxeles de la imagen que son menores que 0.64 a 1.0, y el resto a 0.

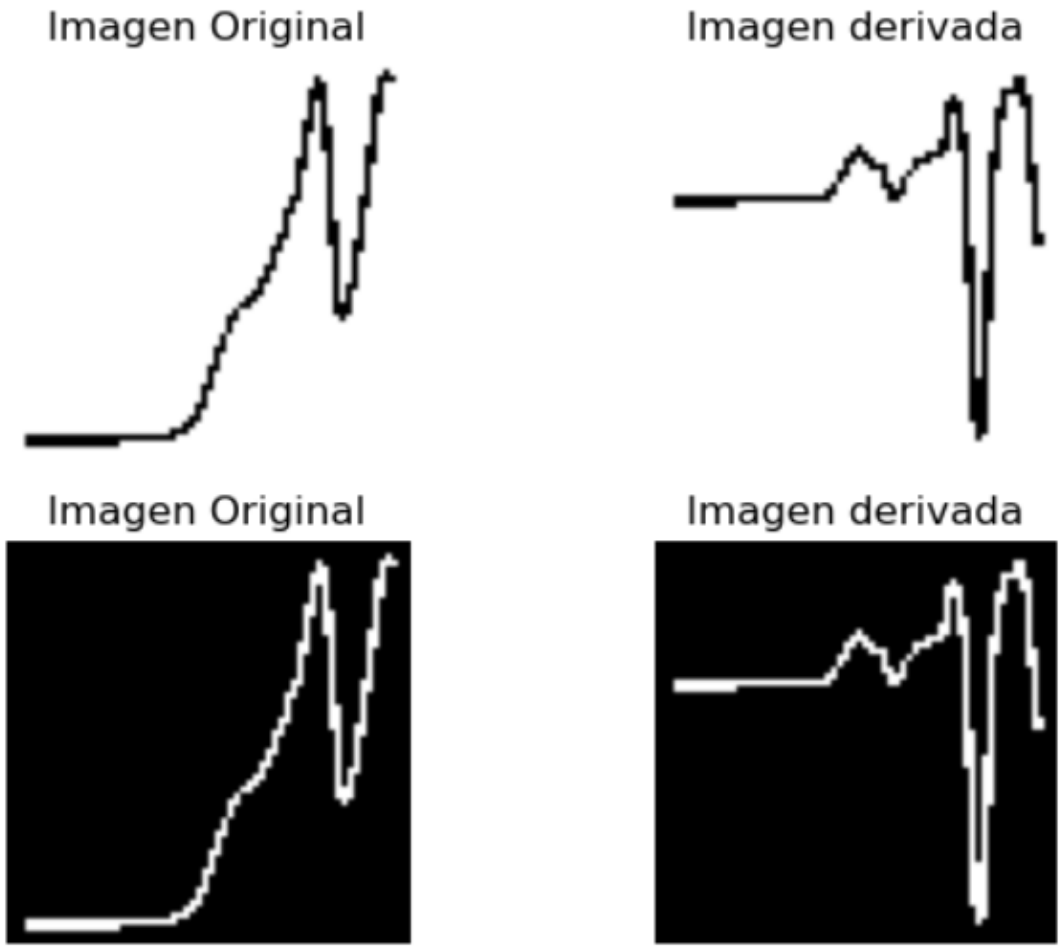


Figura 5: Ejemplo de las Binarizaciones

- **Arquitectura convolucion1d:**

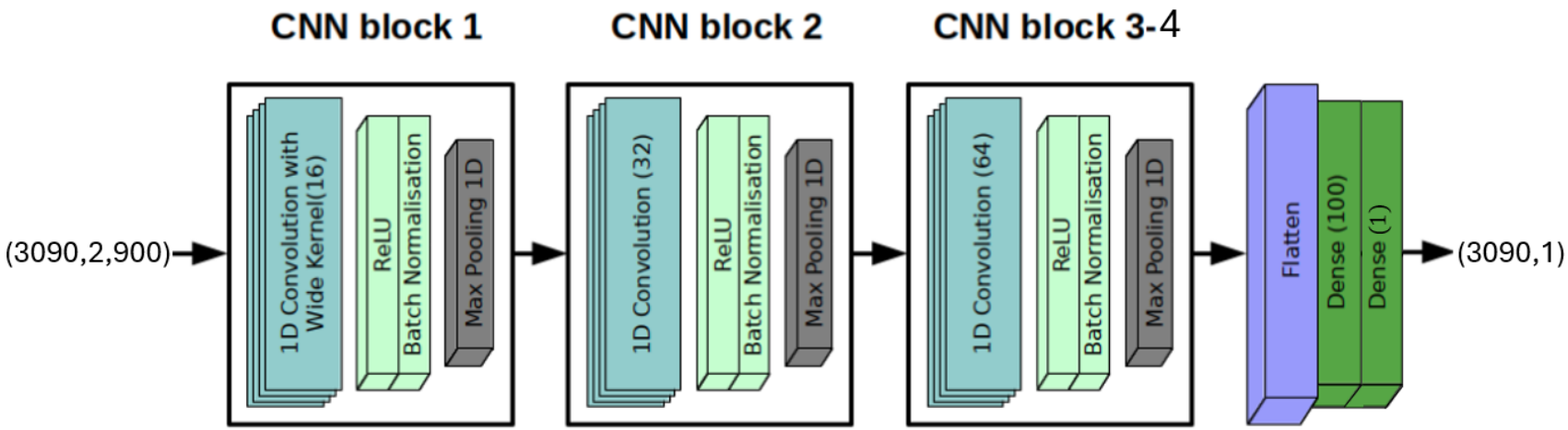


Figura 6

- **Arquitectura del metodo de la Gráfica:**

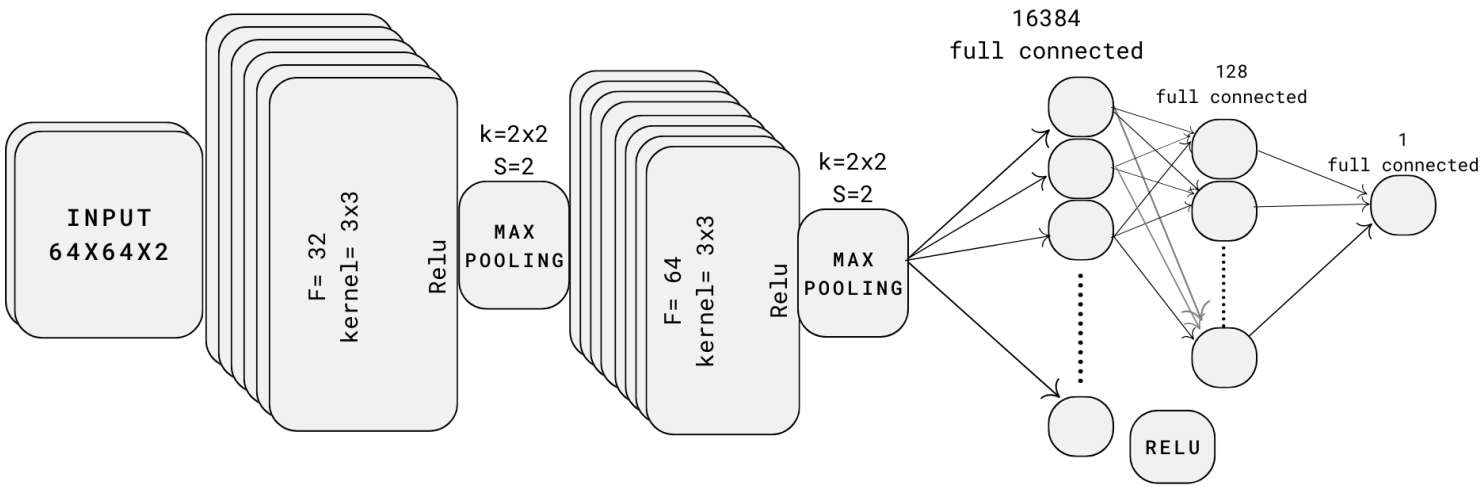


Figura 7: Representación de la arquitectura para el método de gráfica

Metricas

Las metricas utilizadas para evaluar los modelos fueron las siguientes:

- **MAE (Error absoluto medio):** El promedio de las diferencias absolutas entre las predicciones y los valores reales.
- **MSE (Error cuadrático medio):** El promedio de los cuadrados de las diferencias entre las predicciones y los valores reales.
- **RMSE (Raíz del error cuadrático medio):** La raíz cuadrada del MSE. Proporciona una medida de la dispersión de los errores en la misma unidad que los valores originales.
- R^2 **(Coeficiente de determinación):** Una medida de la proporción de la variabilidad en los datos que es explicada por el modelo.

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Búsqueda en Grid

Se realizo la búsqueda en grid de los parámetros:

- Regresión Lineal:
  - n_jobs = -1
- Decision Tree:
  - Max_depth = 5
  - min_samples_split = 5
- Random Forest:
  - n_estimators = 800
  - Max_depth = 50
  - min_samples_leaf = 2

UNIVERSIDAD NACIONAL DE COLOMBIA

MSE	**RMSE**	**MAE**
RandomForest (0.159)	RandomForest (0.399)	RandomForest (0.254)
DecisionTree (0.163)	DecisionTree (0.404)	DecisionTree (0.258)
LinearRegression (23.905)	LinearRegression (4.889)	LinearRegression (0.761)

R^2	**Max_Error**
RandomForest (0.646)	DecisionTree (2.616)
DecisionTree (0.637)	RandomForest (2.616)
LinearRegression (-52.293)	LinearRegression (114.744)

Tabla 2: Resultados obtenidos utilizando todos los puntos como características del modelo

Resultados ML - Método 2: Regresión Lineal

MSE	**RMSE**	**MAE**
RobScaler (0.284)	RobScaler (0.532)	RobScaler (0.413)
StadScaler (0.29)	StadScaler (0.538)	StadScaler (0.415)
MMScaler (0.294)	MMScaler (0.542)	MMScaler (0.415)

R^2	**Max_Error**
RobScaler (0.368)	RobScaler (2.247)
StadScaler (0.354)	StadScaler (2.366)
MMScaler (0.345)	MMScaler (2.422)

Tabla 3: Resultados obtenidos del mejor scaler para Regresión lineal

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultados ML - Método 2: DecisionForest

MSE	**RMSE**	**MAE**
StadScaler (0.224)	StadScaler (0.473)	MMScaler (0.315)
MMScaler (0.23)	MMScaler (0.479)	StadScaler (0.334)
RobScaler (0.253)	RobScaler (0.503)	RobScaler (0.348)

R^2	**Max_Error**
StadScaler (0.501)	StadScaler (1.919)
MMScaler (0.488)	RobScaler (2.225)
RobScaler (0.436)	MMScaler (3.128)

Tabla 4: Resultados obtenidos del mejor scaler para Arboles de desiciones

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultados ML - Método 2: RandomForest

MSE	**RMSE**	**MAE**
MMScaler (0.229)	MMScaler (0.479)	StadScaler (0.338)
StadScaler (0.23)	StadScaler (0.48)	MMScaler (0.342)
RobScaler (0.253)	RobScaler (0.503)	RobScaler (0.348)

R^2	**Max_Error**
MMScaler (0.488)	RobScaler (2.225)
StadScaler (0.487)	MMScaler (2.248)
RobScaler (0.436)	StadScaler (2.315)

Tabla 5: Resultados obtenidos del mejor scaler para RandomForest

UNIVERSIDAD NACIONAL DE COLOMBIA

Búsqueda en Grid

Se realizo la búsqueda en grid de los parámetros:

- Regresión Lineal:
  - n_jobs = -1
- Decision Tree:
  - Max_depth = 10
  - Max_features = 'sqrt'
  - min_samples_leaf = 8
- Random Forest:
  - Max_features = 'sqrt'
  - n_estimators = 300
  - Max_depth = 30

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultados ML - Método 2: Finales

MSE	**RMSE**	**MAE**
RandomForest (0.157)	RandomForest (0.396)	RandomForest (0.294)
DecisionTree (0.224)	DecisionTree (0.473)	DecisionTree (0.334)
LinearRegression (0.29)	LinearRegression (0.538)	LinearRegression (0.415)

R^2	**Max_Error**
RandomForest (0.65)	RandomForest (1.769)
DecisionTree (0.501)	DecisionTree (1.919)
LinearRegression (0.354)	LinearRegression (2.366)

Tabla 6: Resultado de los modelos para StandarScaler

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultado DL - Método de gráfica

Se realizo diferentes pruebas, variando las epocas, el learning rate, el batch y utilizando diferentes optimizadores como SGD y Adam.

Fondo Blanco
MAE
sgd

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	0.61	0.46	0.68	-0.36
	0.01		0.56	0.43	0.68	-0.35
	0.001		0.54	0.57	0.75	-0.3

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.39	0.290	0.54	0.3
		128	0.43	0.32	0.56	0.24
		600	0.54	0.57	0.76	-0.347

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
60	0.001	500	0.54	0.56	0.74	-0.31
100			0.56	0.6	0.7	-0.399
200			0.45	0.357	0.598	0.16

Fondo Negro
MAE
sgd

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	0.3	0.20	0.45	0.49
	0.01		0.42	0.33	0.57	0.22
	0.001		0.5	0.47	0.69	-011

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.36	0.28	0.53	0.34
		128	0.41	0.32	0.57	0.24
		600	0.51	0.49	0.7	-0.14

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
60	0.001	500	0.49	0.44	0.66	-0.02
100			0.47	0.39	0.62	0.08
200			0.4	0.32	0.56	0.24

Figura 8: MAE & SGD

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultado DL - Método de gráfica

Fondo Blanco
MSE
sgd

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	0.39	0.25	0.5	0.4
	0.01		0.38	0.24	0.498	0.42
	0.001		0.56	0.38	0.62	0.1

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.4	0.26	0.5	0.38
		128	0.44	0.3	0.55	0.29
		600	0.55	0.39	0.62	0.08

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
60	0.001	500	0.56	0.4	0.63	0.05
100			0.53	0.37	0.61	0.12
200			0.45	0.31	0.55	0.26

Fondo Negro
MSE
sgd

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	Nan	Nan	Nan	Nan
	0.01		0.44	0.3	0.54	0.29
	0.001		0.47	0.35	0.57	0.21

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.35	0.21	0.46	0.502
		128	0.39	0.25	0.5	0.4
		600	0.47	0.335	0.57	0.21

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
60	0.001	500	0.49	0.34	0.588	0.19
100			0.41	0.28	0.532	0.33
200			0.4	0.269	0.519	0.37

Figura 9: MSE & SGD

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultado DL - Método de gráfica

Fondo Blanco
MSE
adam

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	0.58	0.42	0.65	-0.0006
	0.01		0.455	0.32	0.57	0.23
	0.001		0.27	0.13	0.37	0.67

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.27	0.14	0.38	0.65
		128	0.29	0.157	0.39	0.633
		600	0.27	0.148	0.38	0.65

Fondo Negro
MSE
adam

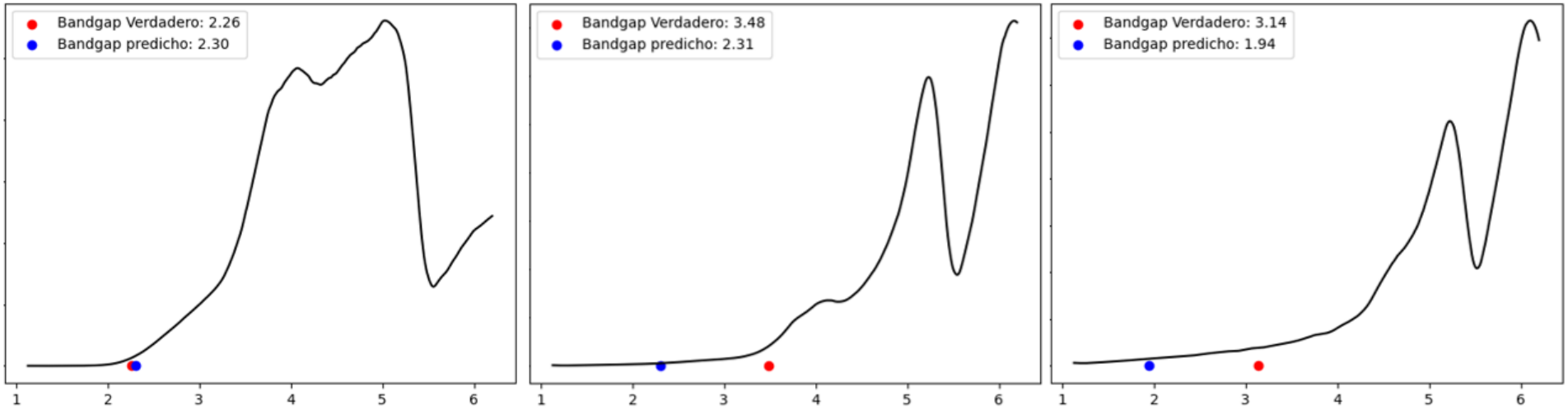
Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.1	500	0.58	0.42	0.65	-0.005
	0.01		0.55	0.4	0.6	-0.002
	0.001		0.26	0.13	0.36	0.685

Epocas	Lr	Batch	(MAE)	(MSE)	(RMSE)	R^2
50	0.001	64	0.28	0.16	0.41	0.60
		128	0.37	0.23	0.48	0.59
		600	0.3	0.16	0.4	0.61

Figura 10: MSE & ADAM

UNIVERSIDAD NACIONAL DE COLOMBIA

Resultados DL - Método de Conv1D



- Las mejores metricas obtenidas fueron las siguientes:

MAE	**MSE**
0.5854554	0.6722501
RMSE	R^2
0.8199086	-0.356794

Tabla 7: Mejores Resultados de Método Conv1D

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Conclusiones Machine Learning

En conclusión, el Random Forest se destacó como el mejor modelo en ambos experimentos, lo que está respaldado por la literatura existente. Aunque el método 2, que implicaba la extracción de características básicas de los datos, mostró una mejora la cual si bien no fue significativa en términos de métricas de rendimiento, sí resultó en tiempos de computación más bajos.

Esto se debe a que el método 2 presentaba una dimensionalidad de datos significativamente más baja en comparación con el método 1. Lo que significa que este acercamiento puede llegar a ser acertado para la predicción de band gap a través de lista 1D de puntos de diagrama Tauc's.

En resumen, se elige como mejor modelo el RandomForest del método 2 con un R^2 de 65 %, con un error aproximado de 0.39 eV y presentando un error máximo de 1.77 eV.

UNIVERSIDAD NACIONAL DE COLOMBIA

Conclusiones Deep Learning

En síntesis, la investigación encontró que al tratar x2 como una señal procesada con convoluciones 1d, el modelo resultante es muy sensible a las pequeñas variaciones de la pendiente por lo que este realiza una predicción prematura en comparación a la etiqueta real.

Por otro lado, se obtuvieron mejores resultados al entrenar al modelo con las imágenes de las gráficas de discretización de fondo negro alcanzando un R^2 de 0.68 utilizando Adam como optimizador y MSE como función de perdida. Sin embargo, con este método aparentemente se realizaba la predicción a partir de la pendiente mas pronunciada, por lo que en algunos casos difería en gran medida de la etiqueta real.

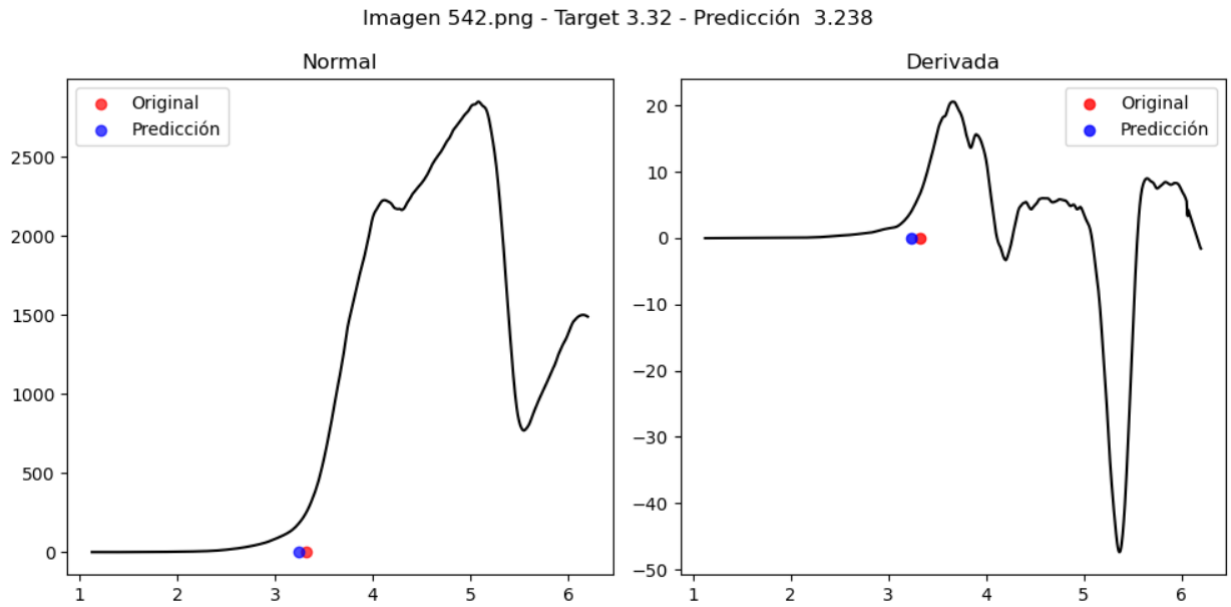


Figura 11: MSE con Adam

Conclusiones Deep Learning

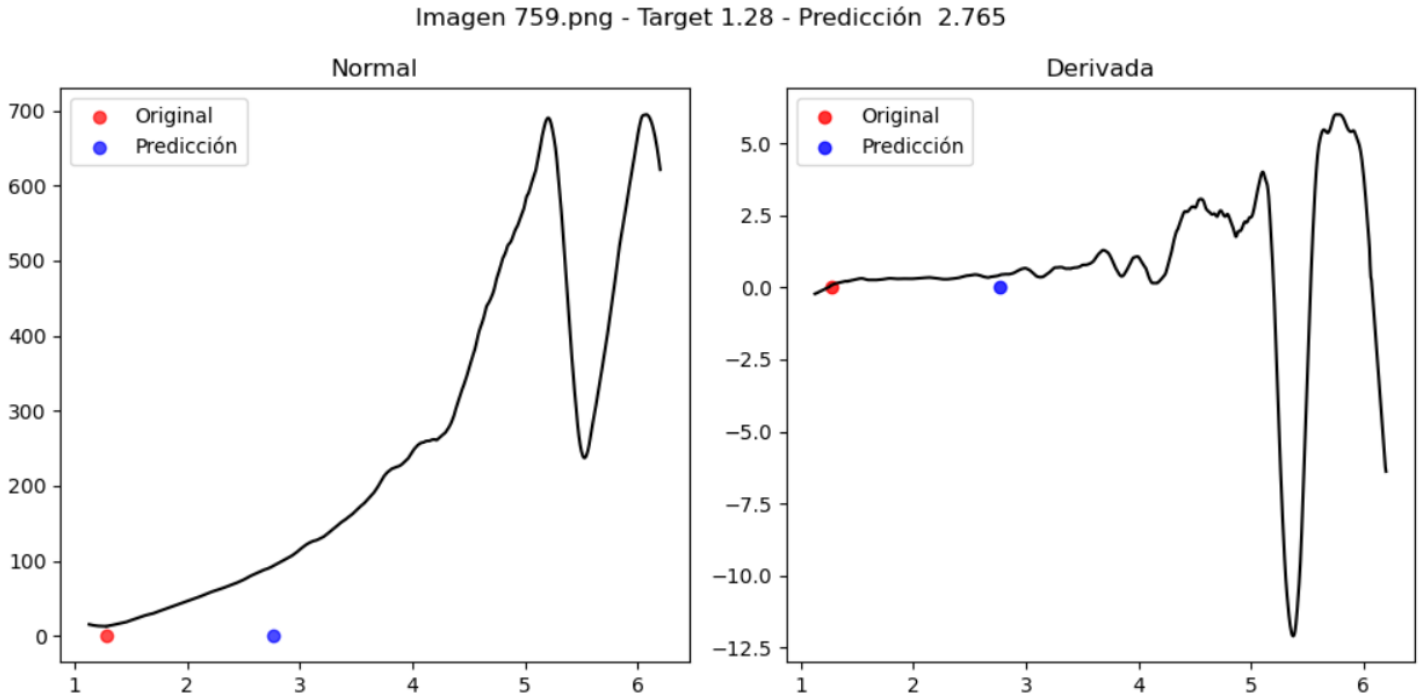


Figura 12: caso particular

Contenido

UNIVERSIDAD NACIONAL DE COLOMBIA

Referencias

- *Explain the formation of energy bands in solids. On the basis of energy bands distinguish between a metal, a semiconductor and an insulator.*(2024).
- Ghosh, K., Sil, S., Ray, P. P., Ortega-Castro, J., Frontera, A., & Chattopadhyay, S. (2019). Photosensitive Schottky barrier diode behavior of a semiconducting Co(iii)-Na complex with a compartmental Schiff base ligand. *RSC Advances*, 9(60), 34710–34719. https://doi.org/10.1039/c9ra06354d

UNIVERSIDAD NACIONAL DE COLOMBIA