

# LA FUENTE DE ENERGÍA DE LAS GALAXIAS VISTA DESDE EL MACHINE LEARNING: CLASIFICACIÓN USANDO LOS ESPECTROS DE SPITZER.

Proyecto de grado

Autora: Andrea Elneser Tejeda

Director: Francisco Carlos Calderon Bocanegra

Cliente: Juan Rafael Martínez Galarza



Fecha 22/11/2022

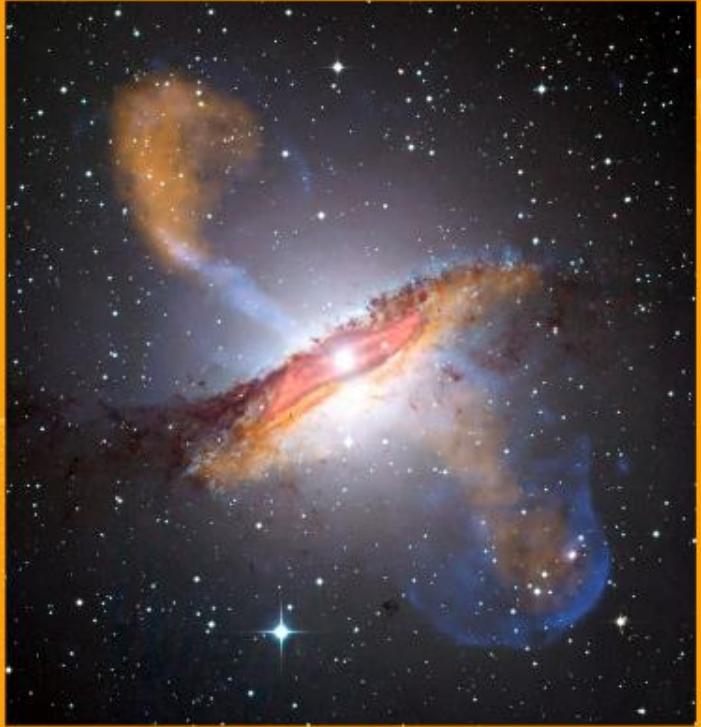


Figura 1. Galaxia AGN. Tomada de [1].

# 01. CONTEXTO



El presente proyecto consiste en la creación e implementación de un algoritmo de machine learning que pueda clasificar galáxias según su distribución energética.



Figura 1. Pilares de la creación. Toma de [2].

**01.**

# CONTEXTO: OBJETIVOS

## GENERAL

Implementar un algoritmo que a partir de los datos del telescopio Spitzer permita clasificar galaxias entre aquellas que tienen un núcleo activo y las que no.

## ESPECÍFICOS

- Recopilar y limpiar datos.
- Implementar métodos de machine learning.
- Evaluar el desempeño de los métodos.



# 01. CONTEXTO: ESTADO DEL ARTE

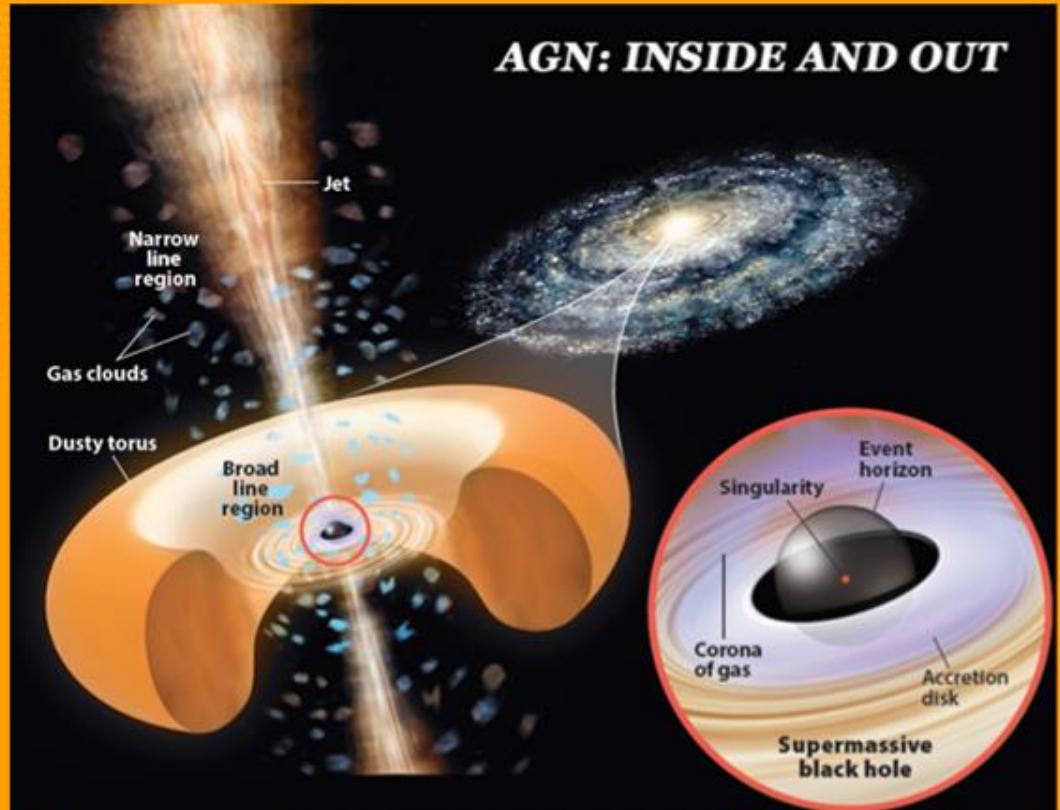


Figura 3. Modelo de AGN. Tomada de [3].

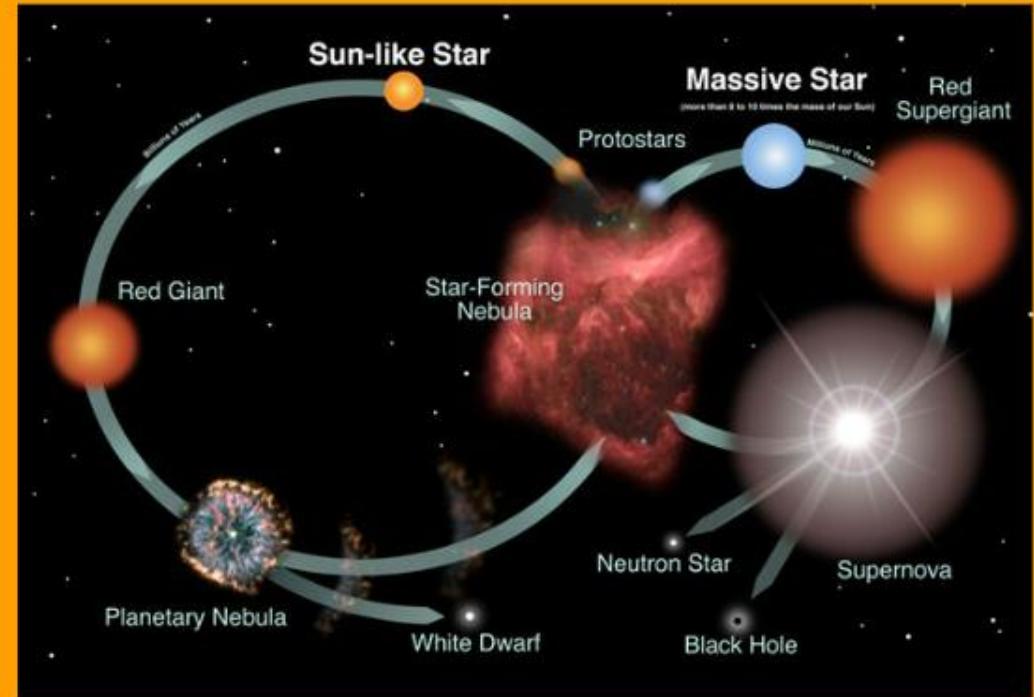


Figura 4. Formación estelar. Tomada de [4].

# 01. CONTEXTO: ESTADO DEL ARTE

Estudios que implementan directamente el machine learning para clasificar en la astronomía:

- Image feature extraction and galaxy classification: a novel and efficient approach with automated machine learning: Clasificación de imágenes de galaxias con la base de datos Dark Energy Survey.
- Research on star/galaxy classification based on stacking ensemble learning: Clasificación de las magnitudes de las fuentes del dataset Sloan Digital Sky Survey

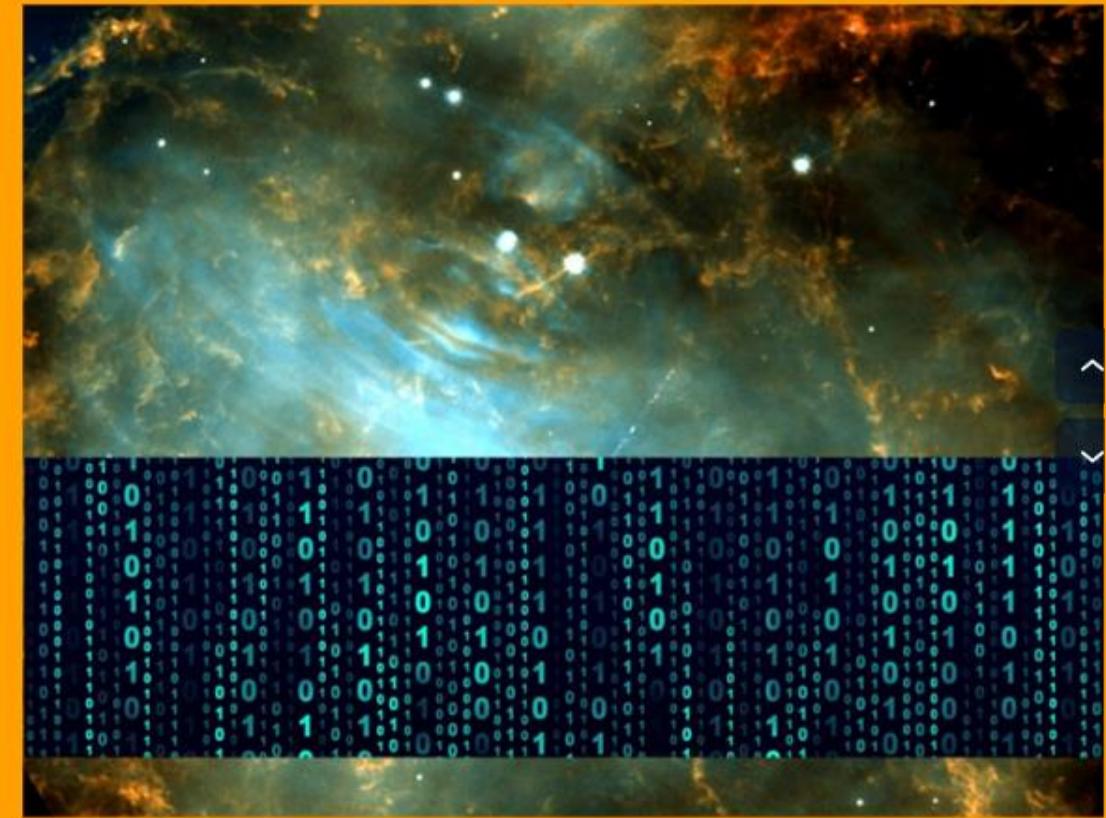


Figura 5. Tomada de [5].

# 02. DESARROLLO: DIAGRAMA DE FLUJO



Figura 6. Diagrama de flujo del proyecto

# 02. DESARROLLO: RECOPILACIÓN Y ORGANIZACIÓN DE LOS DATOS

## 1. Descarga de datos

- Código que lee todos los Aorkeys

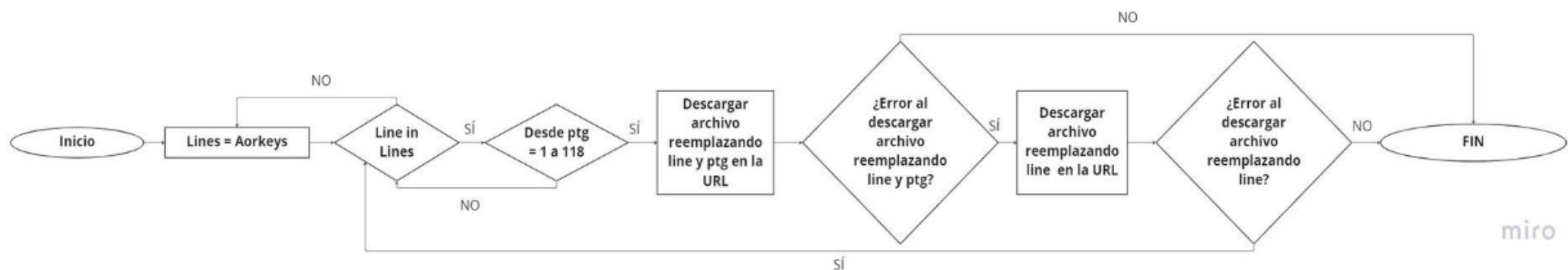


Figura 7. Diagrama de descarga de datos

# 02. DESARROLLO: RECOPILACIÓN Y ORGANIZACIÓN DE LOS DATOS

## 2. Datos a archivos .csv

- Extracción de columna de densidad de flujo y longitud de onda a archivo .csv.



Figura 8. Diagrama de datos a archivos .csv

# 02. DESARROLLO: RECOPILACIÓN Y ORGANIZACIÓN DE LOS DATOS

## 3. Etiquetar los datos

- Cruce entre datos descargados y sus categorías.
- Categorías se obtuvieron mediante web scraping de la base de datos.
- 38 categorías de 2025 ProgramIDs.
- ProgramIDs son identificadores de las categorías y esto sí se encuentran en los datos.



Figura 9. Diagrama de categorización de datos

# 02. DESARROLLO: RECOPILACIÓN Y ORGANIZACIÓN DE LOS DATOS

## 4. Vector de longitudes de onda

Se tomó el dato que más longitudes de onda tenía y se le agregó las longitudes faltantes concatenándolas en un vector que se organizó y del que se borraron los repetidos.

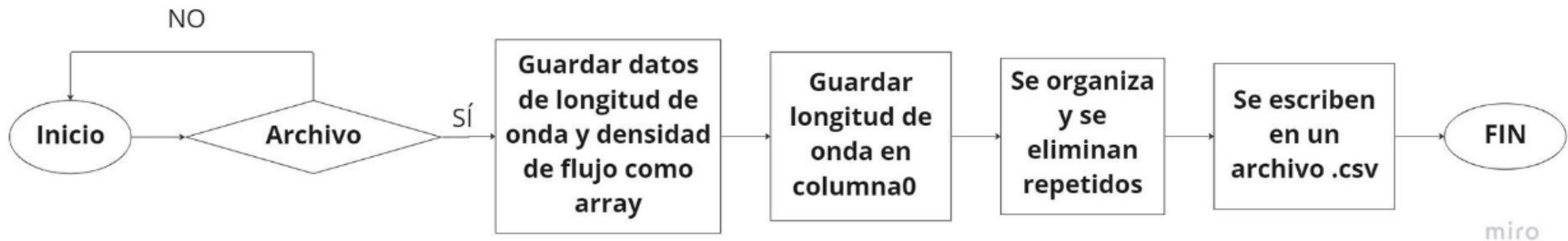


Figura 10. Diagrama de creación de vector de longitud de ondas

# 02. DESARROLLO: RECOPILACIÓN Y ORGANIZACIÓN DE LOS DATOS

## 5. Densidades de flujo respecto a longitudes de onda

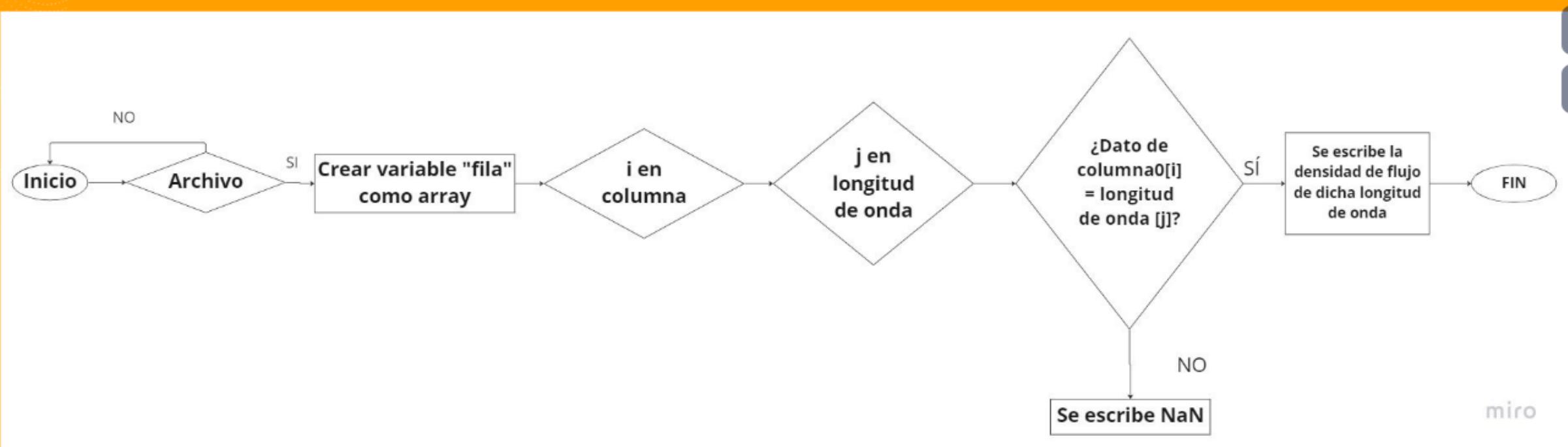


Figura 11. Diagrama de creación de tablas con densidades

miro

# 02. DESARROLLO: LIMPIEZA DE DATOS

## 6. Interpolación de los valores faltantes

Se interpolaron los valores inicialmente de manera lineal.

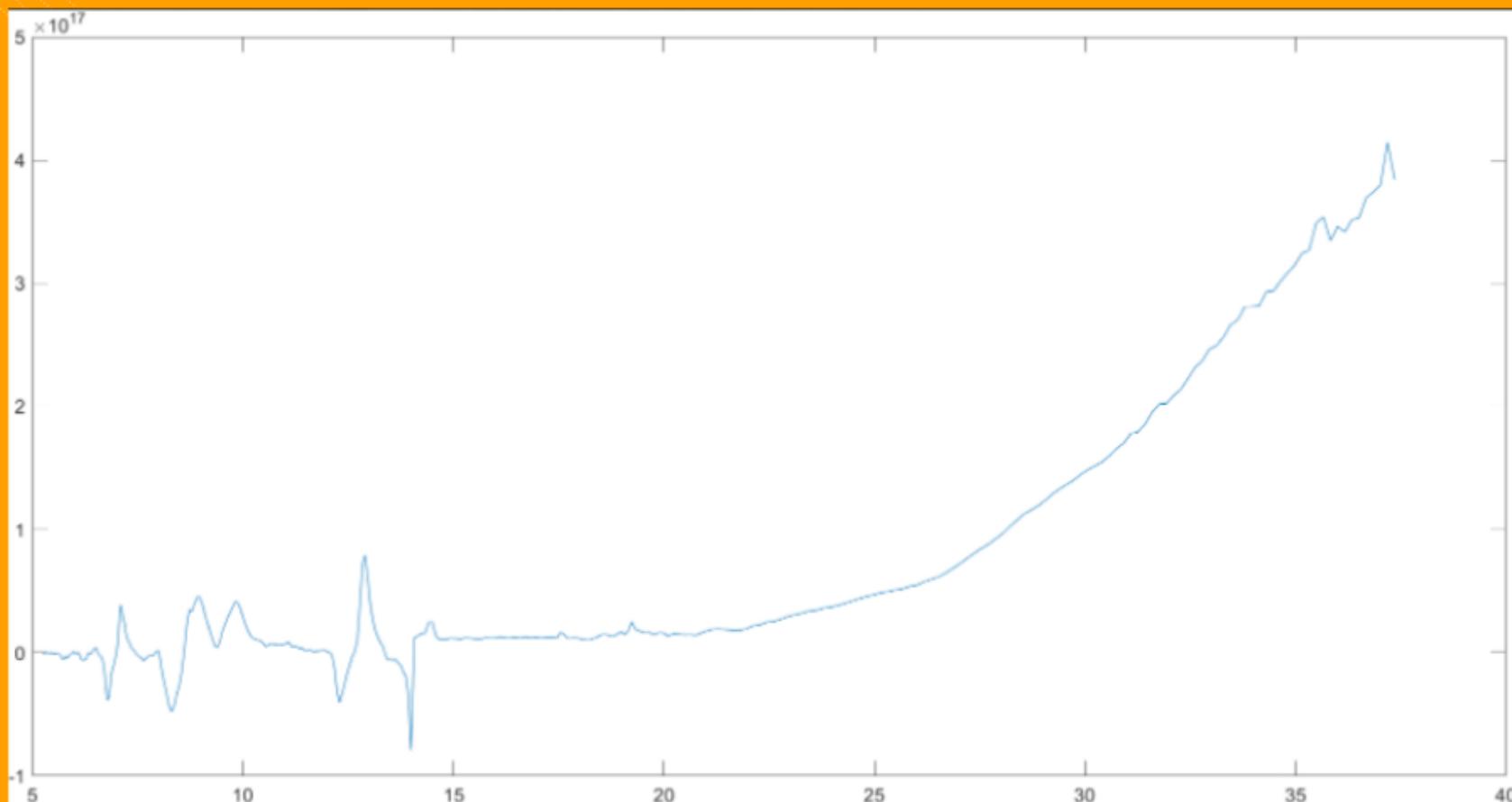


Figura 12. Gráfica de datos con valores interpolados linealmente

# 02. DESARROLLO: LIMPIEZA DE DATOS

## 6. Interpolación de los valores faltantes

Al final se optó por una mediana móvil de las columnas.

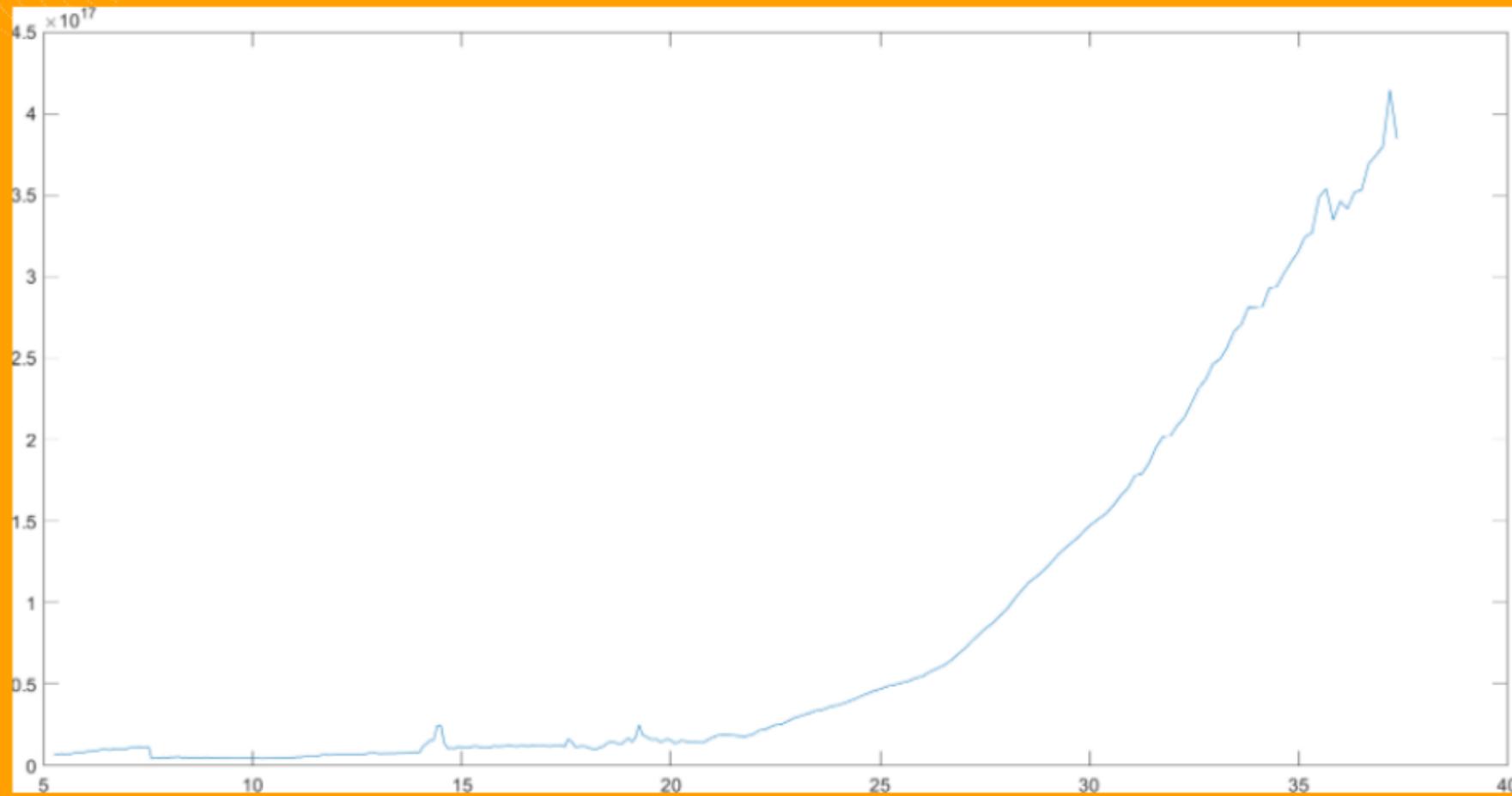


Figura 13. Gráfica de datos con valores interpolados con mediana móvil de las columnas

# 02. DESARROLLO: MÉTRICAS DE VALIDACIÓN

$$\text{Accuracy: } ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Coeficiente de correlación de Matthews: } MATTHEWS = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\text{F1-SCORE: } F1 = \frac{2TP}{2TP+FP+FN}$$

# 03. ANÁLISIS DE RESULTADOS: PCA

```
>> [a,b]=max(abs(pcaCoefficients))  
  
a =  
  
    0.3425    0.1643    0.0922  
  
b =  
  
   104    358    257
```

- Coeficientes de PCA
- Se pasó de 360 componentes a 3.
- Entre los 3 coeficientes principales, se logra una varianza explicada a 99,1 %

Figura 14. Componentes principales de PCA

# 03. ANÁLISIS DE RESULTADOS: PCA

- El primer componente principal se encuentra en 104.
- Varianza explicada de 73.6 %.

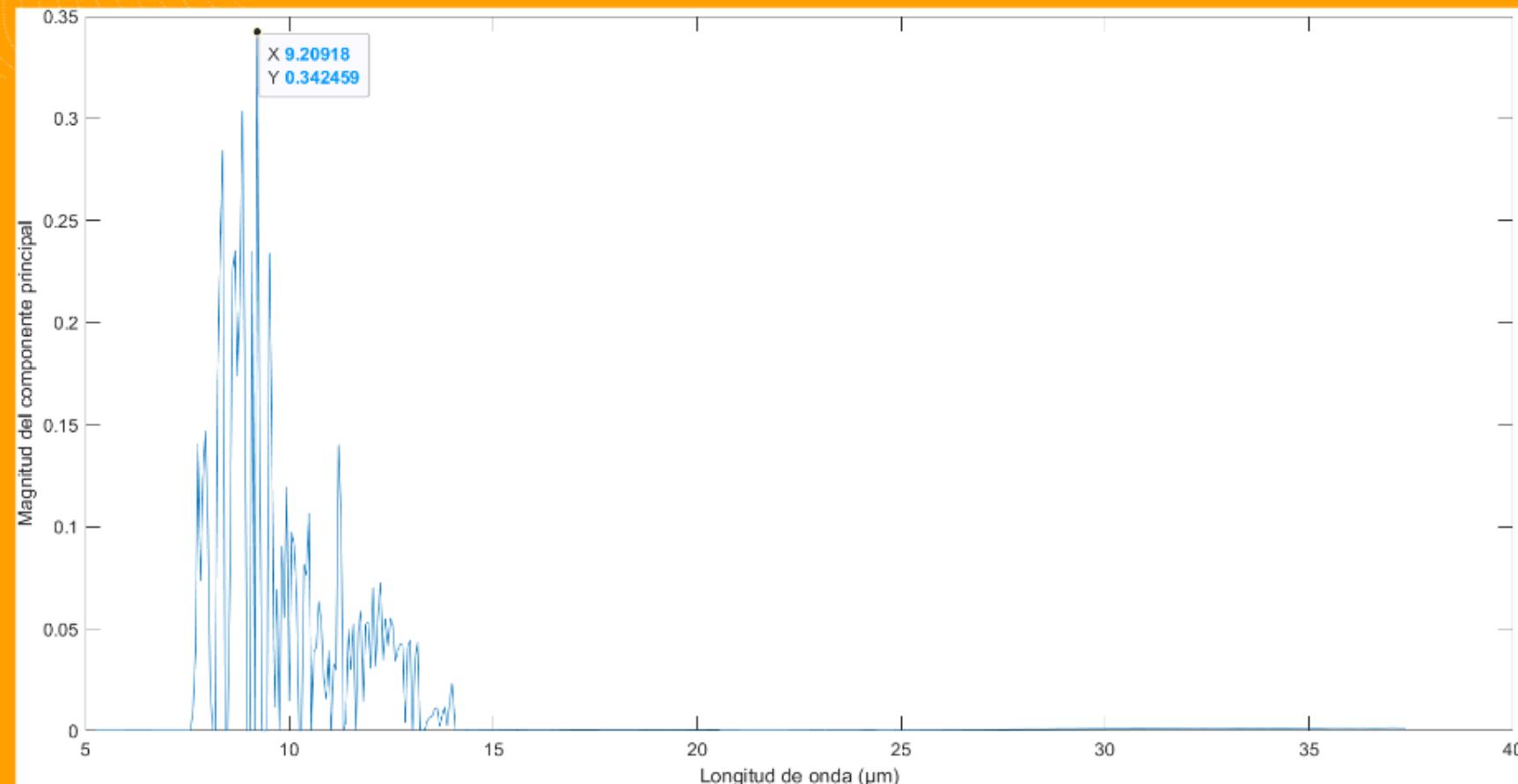


Figura 15. Primer componente principal de PCA

# 03. ANÁLISIS DE RESULTADOS: PCA

- El segundo componente se encuentra en 358.
- Varianza explicada de 24.5 %.

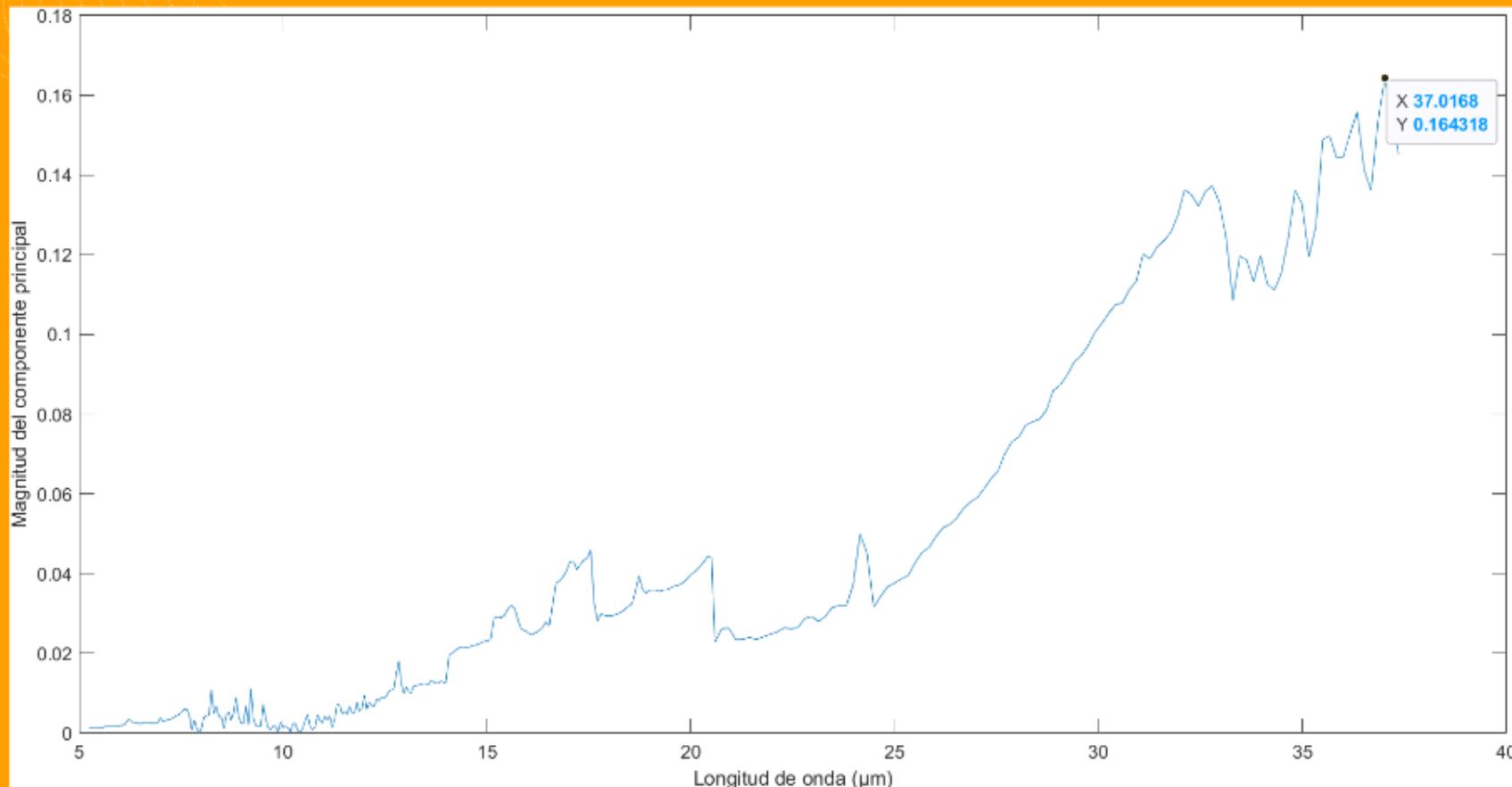


Figura 16. Segundo componente principal de PCA

# 03. ANÁLISIS DE RESULTADOS: PCA

- El último componente se encuentra en 257.
- Varianza explicada de 1.0 %..

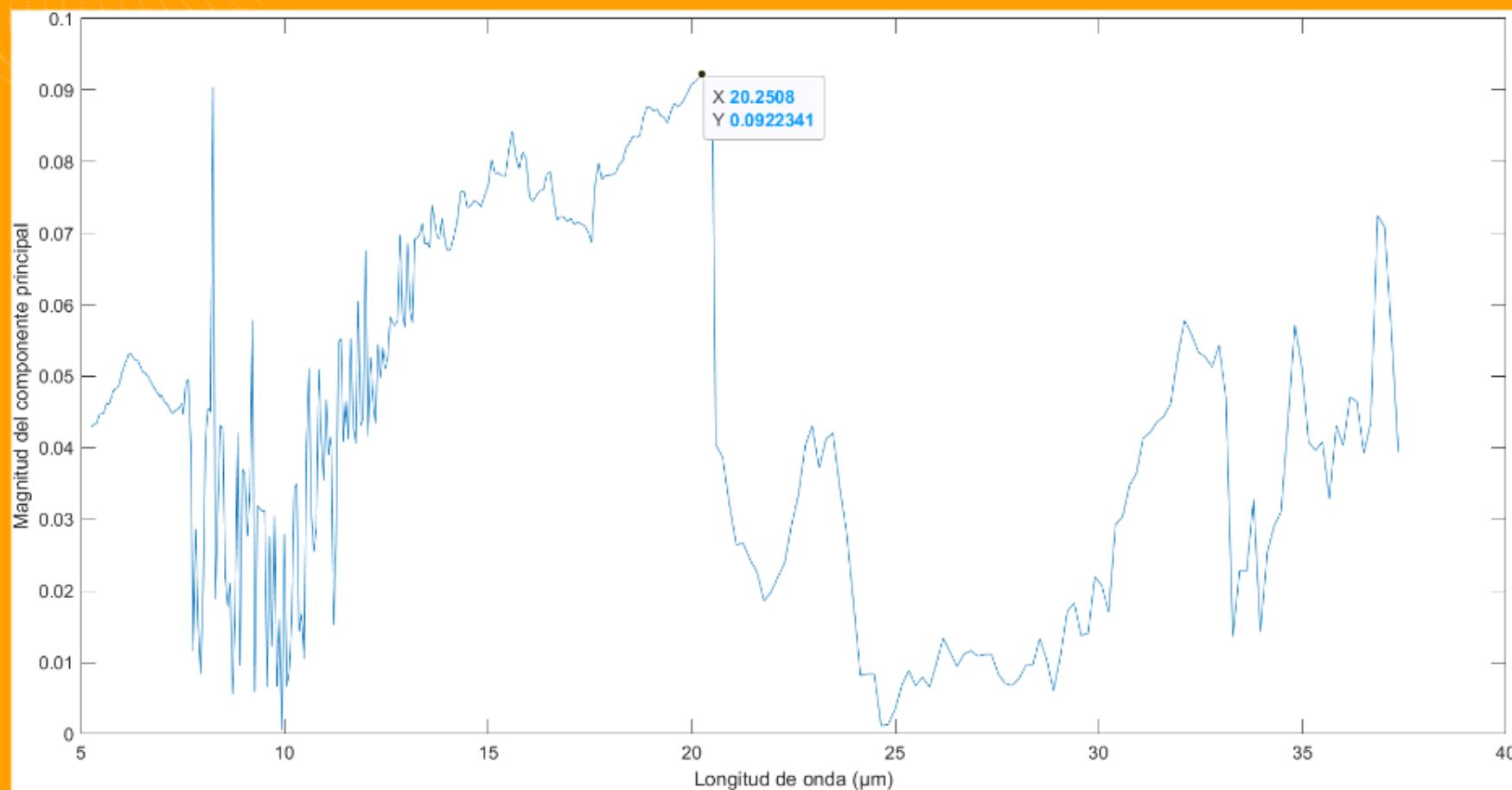


Figura 17. Tercer componente principal de PCA

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Tabla de resultados

- Mejores resultados organizados por orden de accuracy.
- El accuracy fue calculado por el classification learner.

Model Type	Accuracy % (MATLA)
Neural Network	96.09
Ensemble	94.91
Tree	94.32
KNN	93.84
SVM	91.98
Kernel	91.59
Discriminant	91.29
Naive Bayes	87.67
Logistic Regression	83.17

Figura 18. Tabla de modelos organizador por accuracy

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Tabla de resultados

- Mejores resultados organizados por orden de coeficiente de Matthews.
- El coeficiente de Matthews es la mejor métrica en este caso.

Model Type	Accuracy % (MATLA)	TP	TN	FP	FN	Accuracy calculada %	Coeficiente Matthe	F1-SCOF
Neural Network	96.09	925	57	34	6	96.09	0.73	0.98
Ensemble	94.91	915	55	36	16	94.91	0.66	0.97
Tree	94.32	908	56	35	23	94.32	0.63	0.97
KNN	93.84	922	37	54	9	93.84	0.55	0.97
SVM	91.98	930	10	81	1	91.98	0.30	0.96
Naive Bayes	87.67	879	26	65	61	87.78	0.23	0.93
Kernel	91.59	930	6	85	1	91.59	0.22	0.96
Logistic Regression	83.17	820	30	61	111	83.17	0.17	0.91
Discriminant	91.29	929	4	87	2	91.29	0.16	0.95

Figura 19. Tabla de modelos organizador por coeficiente de Matthews

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Resultados

Model Type	Neural Network
Accuracy % (MATLAB)	96.09
TP	925
TN	57
FP	34
FN	6
Accuracy calculada %	96.09
Coeficiente Matthews	0.73
F1-SCORE	0.98

- Model Hyperparameters  
Preset: Medium Neural Network  
Number of fully connected layers: 1  
First layer size: 25  
Activation: ReLU  
Iteration limit: 1000  
Regularization strength (Lambda): 0  
Standardize data: Yes

Model Type	Ensemble
Accuracy % (MATLAB)	94.91
TP	915
TN	55
FP	36
FN	16
Accuracy calculada %	94.91
Coeficiente Matthews	0.66
F1-SCORE	0.97

Preset: Bagged Trees  
Ensemble method: Bag  
Learner type: Decision tree  
Maximum number of splits: 1021  
Number of learners: 30  
Number of predictors to sample: Select All

Figura 20. Datos de neural network

Figura 21. Datos de ensemble

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Resultados

Model Type	Tree	Model Hyperparameters  Preset: Fine Tree Maximum number of splits: 100 Split criterion: Gini's diversity index Surrogate decision splits: Off
Accuracy % (MATLAB)	94.32	
TP	908	
TN	56	
FP	35	
FN	23	
Accuracy calculada %	94.32	
Coeficiente Matthews	0.63	
F1-SCORE	0.97	

Figura 22. Datos de Tree

Model Type	KNN	Model Hyperparameters  Preset: Weighted KNN Number of neighbors: 10 Distance metric: Euclidean Distance weight: Squared inverse Standardize data: Yes
Accuracy % (MATLAB)	93.84	
TP	922	
TN	37	
FP	54	
FN	9	
Accuracy calculada %	93.84	
Coeficiente Matthews	0.55	
F1-SCORE	0.97	

Figura 23. Datos de KNN

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Resultados

Model Type	SVM
Accuracy % (MATLAB)	91.98
TP	930
TN	10
FP	81
FN	1
Accuracy calculada %	91.98
Coeficiente Matthews	0.30
F1-SCORE	0.96

Model Hyperparameters  
Preset: Linear SVM  
Kernel function: Linear  
Kernel scale: Automatic  
Box constraint level: 1  
Multiclass method: One-vs-One  
Standardize data: Yes

Figura 24. Datos de SVM

Model Type	Kernel
Accuracy % (MATLAB)	91.59
TP	930
TN	6
FP	85
FN	1
Accuracy calculada %	91.59
Coeficiente Matthews	0.22
F1-SCORE	0.96

Model Hyperparameters  
Preset: SVM Kernel  
Learner: SVM  
Number of expansion dimensions: Auto  
Regularization strength (Lambda): Auto  
Kernel scale: Auto  
Multiclass method: One-vs-One  
Iteration limit: 1000

Figura 25. Datos de Kernel

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Resultados

Model Type	Discriminant
Accuracy % (MATLAB)	91.29
TP	929
TN	4
FP	87
FN	2
Accuracy calculada %	91.29
Coeficiente Matthews	0.16
F1-SCORE	0.95

Model Hyperparameters  
Preset: Linear Discriminant  
Covariance structure: Full

Model Type	Naive Bayes
Accuracy % (MATLAB)	87.67
TP	879
TN	26
FP	65
FN	61
Accuracy calculada %	87.78
Coeficiente Matthews	0.23
F1-SCORE	0.93

Model Hyperparameters  
Preset: Kernel Naive Bayes  
Distribution name for numeric predictors: Kernel  
Distribution name for categorical predictors: Not Applicable  
Kernel type: Gaussian  
Support: Unbounded

# 03. ANÁLISIS DE RESULTADOS: CLASSIFICATION LEARNER

## Resultados

Model Type	Logistic Regression	Model Hyperparameters
		Preset: Logistic Regression
Accuracy % (MATLAB)	83.17	
TP	820	
TN	30	
FP	61	
FN	111	
Accuracy calculada %	83.17	
Coeficiente Matthews	0.17	
F1-SCORE	0.91	

Figura 26. Datos de logistic regression

# 04. CONCLUSIONES

- Reducción considerable de PCA simplificó la complejidad del espacio muestral.

Name	Value
Datos	1022x360 double

Figura 27. Tamaño de datos originales

```
>> [a,b]=max(abs(pcaCoefficients))
```

```
a =
```

```
0.3425 0.1643 0.0922
```

```
b =
```

```
104 358 257
```

Figura 28. PCA

# 04. CONCLUSIONES

- Entre todas las métricas de validación, la mejor para este proyecto es el coeficiente de Matthews porque los datos son desbalanceados.

		Model 3.31	
		AGN/Quasars/Rad	Star Formation
True Class	AGN/Quasars/Rad	930	1
	Star Formation	85	6
		AGN/Quasars/Rad	Star Formation
Predicted Class			

Figura 29. Matriz de confusión Kernel

# 04. CONCLUSIONES

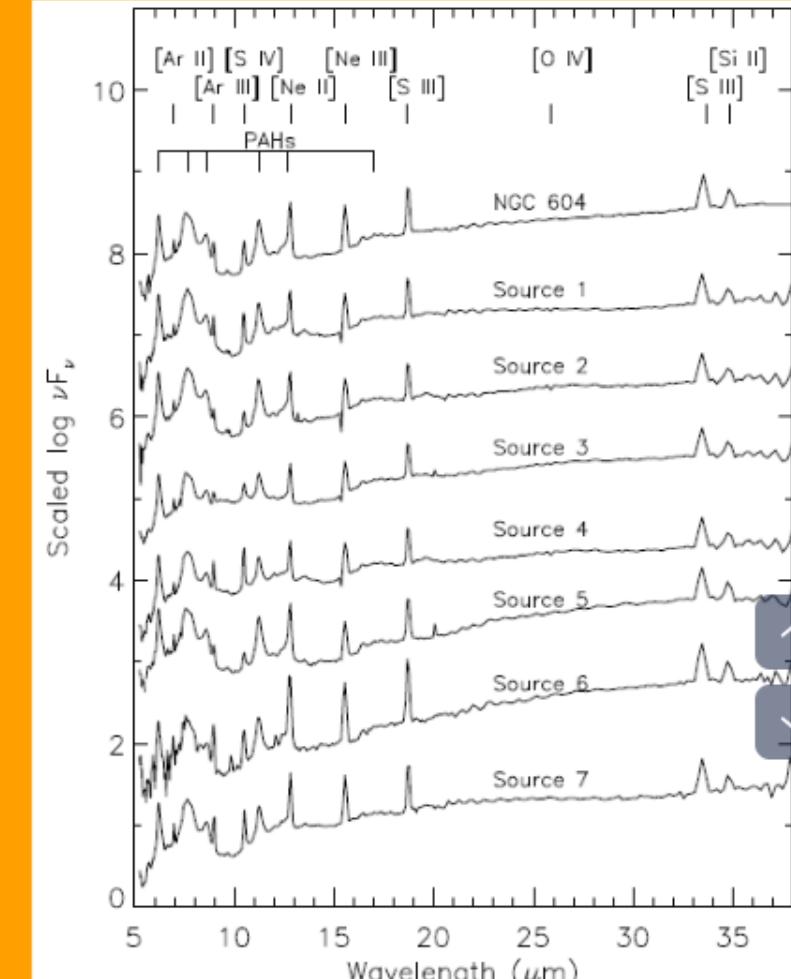
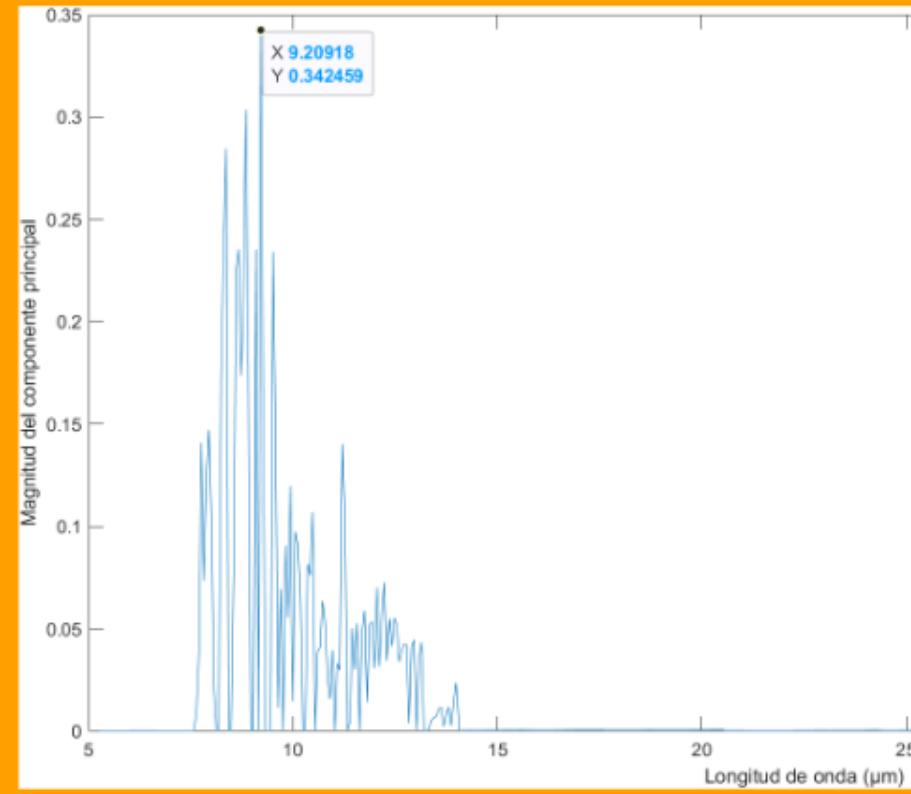
- El mejor método para este proyecto son las redes neuronales.

Model Number	Model Type	Coeficiente Matthews
2.27	Neural Network	0.73
3.22	Ensemble	0.66
1	Tree	0.63
2.2	KNN	0.55
2.9	SVM	0.30
2.8	Naive Bayes	0.23
3.31	Kernel	0.22
2.6	Logistic Regression	0.17
2.4	Discriminant	0.16

Figura 30. Tabla de datos organizados por coeficiente de Matthews

# 04. CONCLUSIONES

- El primer y tercer coeficiente del PCA son cercanos a los valores de los PAHs y el OIV, respectivamente.



# 04. CONCLUSIONES

- El primer y tercer coeficiente del PCA son cercanos a los valores de los PAHs y el OIV, respectivamente.

[O IV]  
25.89 μm  
Excitation Pot.: 54.93eV  
Ionization Pot.: 77.41eV

Figura 34. Rangos de O IV

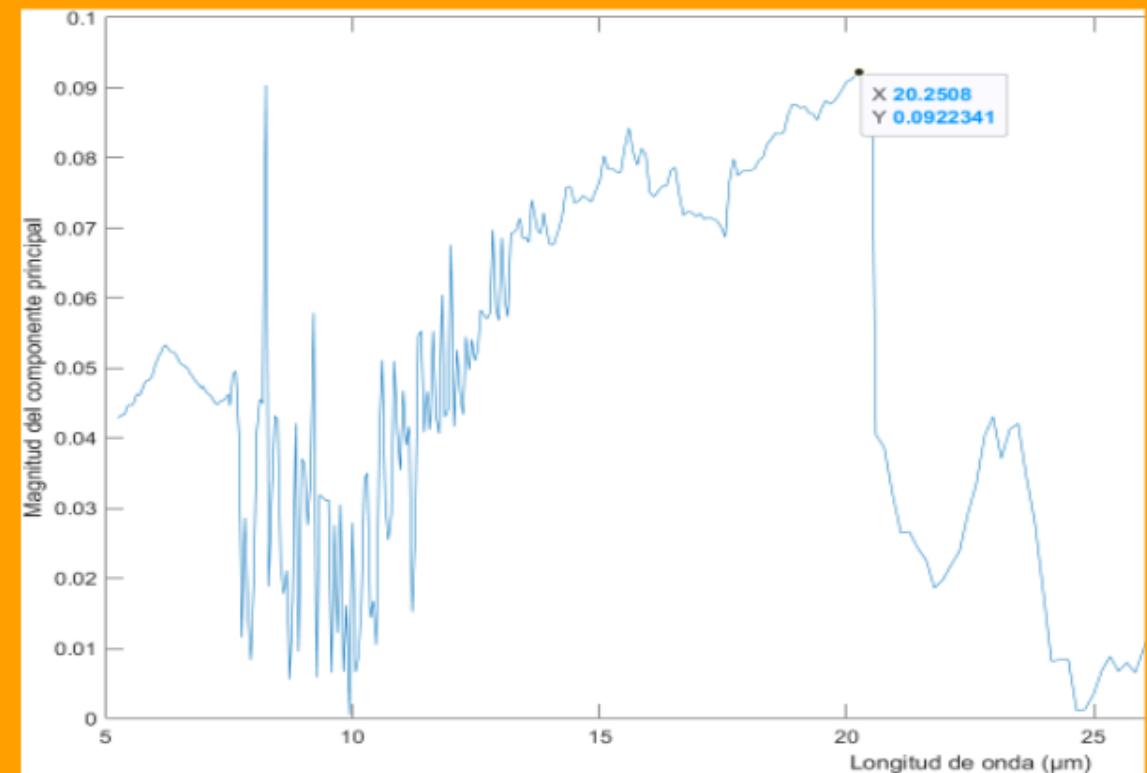


Figura 33. Gráfica de tercer componente de PCA

# 5.

# REFERENCIAS

- 
- 
1. Active Galactic Nuclei - Department of Physics and Astronomy - Uppsala University, Sweden. (s. f.). <https://www.physics.uu.se/research/astronomy-and-space-physics/research/galaxies/agn/>
  2. Rodríguez, H. (2022, 28 octubre). Los espectaculares Pilares de la Creación capturados por el Telescopio Espacial James Webb. [www.nationalgeographic.com.es. https://www.nationalgeographic.com.es/ciencia/espectaculares-pilares-creacion-capturados-por-telescopio-espacial-james-webb\\_18943](https://www.nationalgeographic.com.es/ciencia/espectaculares-pilares-creacion-capturados-por-telescopio-espacial-james-webb_18943)
  3. Why galaxy M77's active nucleus is hiding. (2022, 16 febrero). Astronomy.com. <https://astronomy.com/news/2022/02/m77s-active-nucleus-is-hiding>
  4. 8 Frequently Asked Space-Related Questions (Part 1 of 2). (2017, 1 julio). StarTalk Radio Show by Neil deGrasse Tyson. <https://startalkmedia.com/8-frequently-asked-space-related-questions-part-1-of-2/>
  5. Astronomy - ESAC Trainees - Cosmos. (s. f.). <https://www.cosmos.esa.int/web/esac-trainees/astronomy>



Gracias por su  
atención

¿Alguna pregunta?