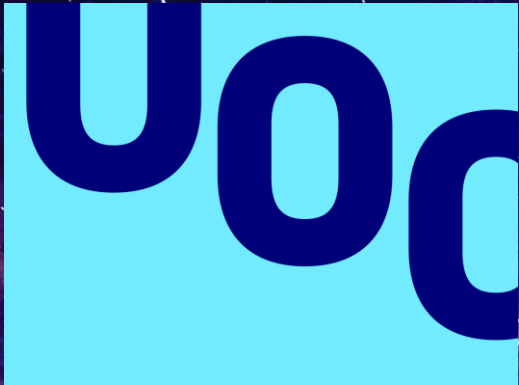


Machine Learning para la detección de exoplanetas: revisión y nuevos enfoques



Trabajo de Fin de Máster
Máster en Ciencia de Datos
Enero 2023

Autor

Javier Gómez de Diego

Tutora

Laura Ruiz Dern

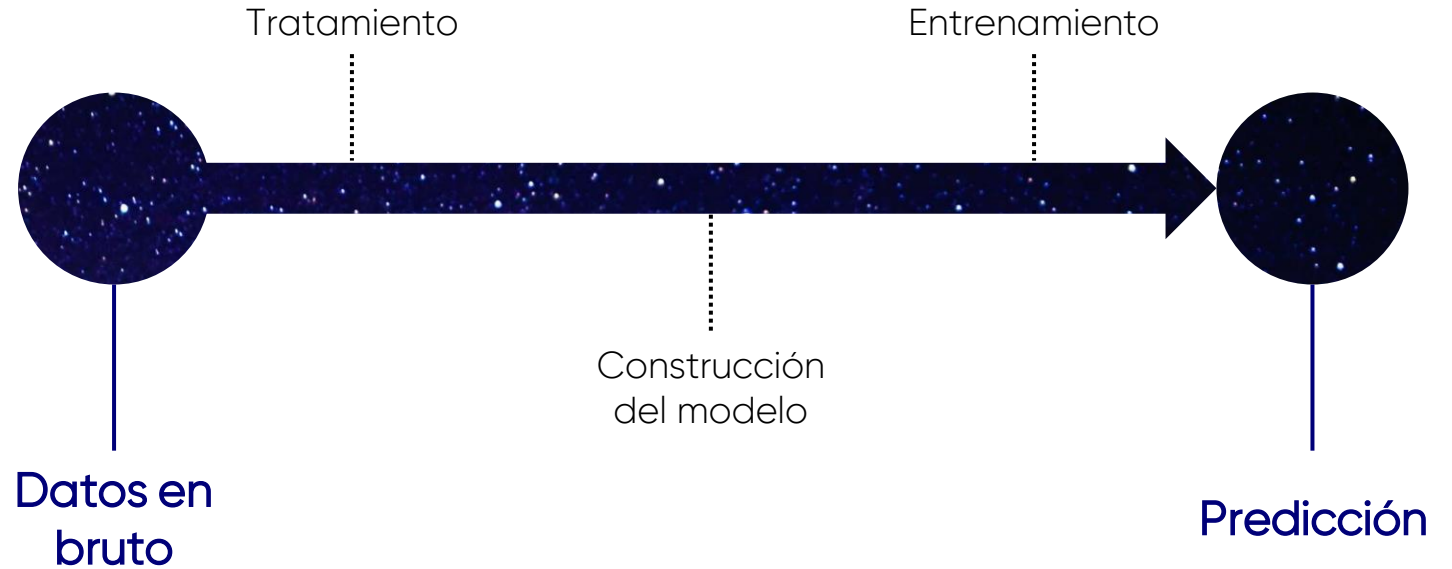
Índice

- **Introducción**
- **Preprocesamiento**
- **Data Augmentation**
- **Estrategias**
- **Resultados**



Introducción

TFG

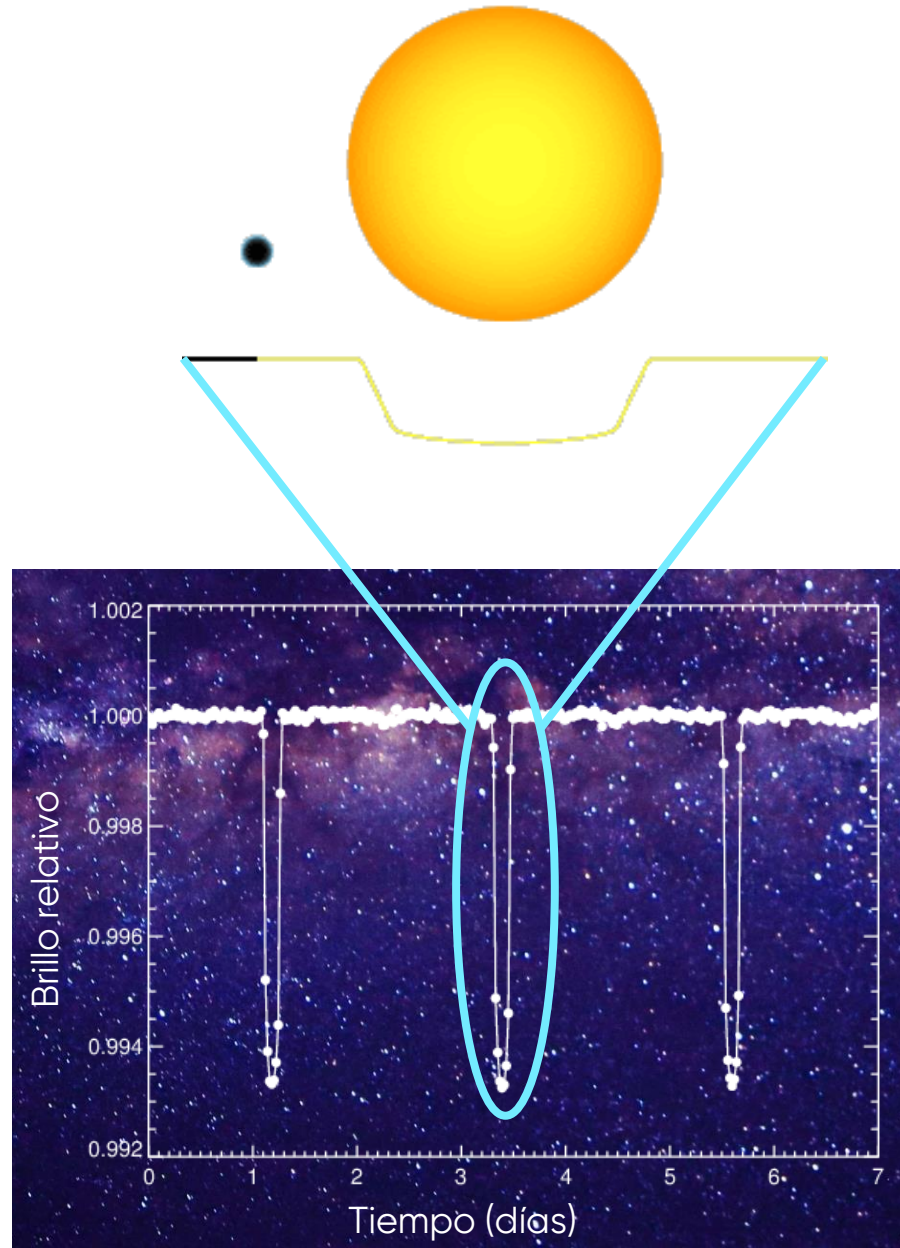


Objetivos

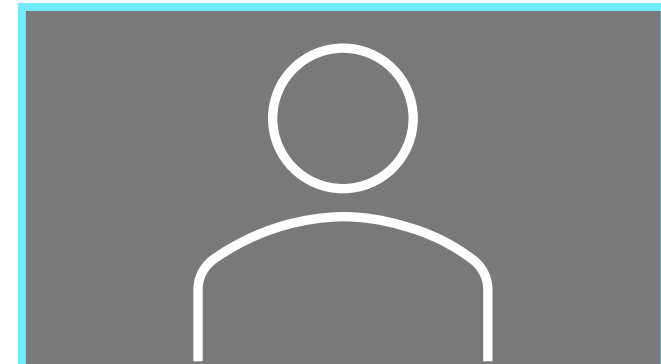
- Mejorar proceso
- Nuevos enfoques
- Mejorar resultados



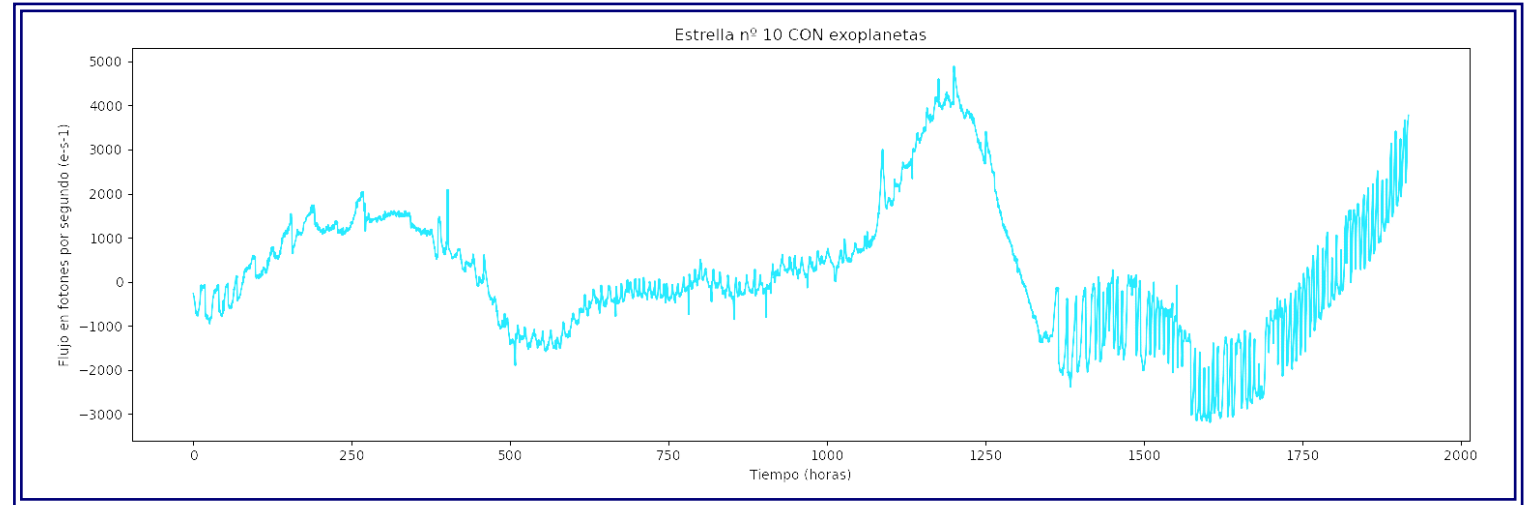
Exoplanetas



Tránsito



Ejemplo



Dimensiones

Entrenamiento

3197 dimensiones
5087 registros

37 positivos
5050 negativos

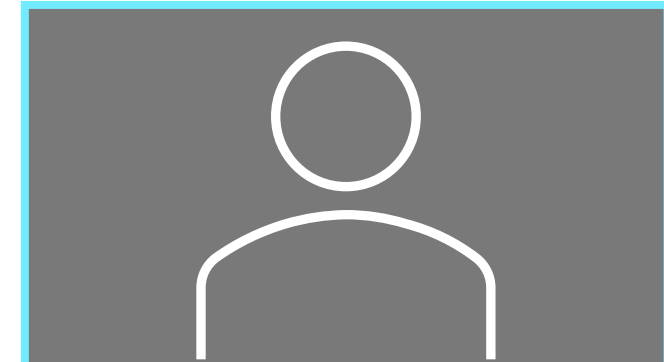
Test

3197 dimensiones
570 registros

5 positivos
565 negativos

Problemas

- Dimensionalidad
- Desbalanceamiento



TFG

Preprocesamiento ➡ PCA ➡ Data Augmentation ➡ Red neuronal

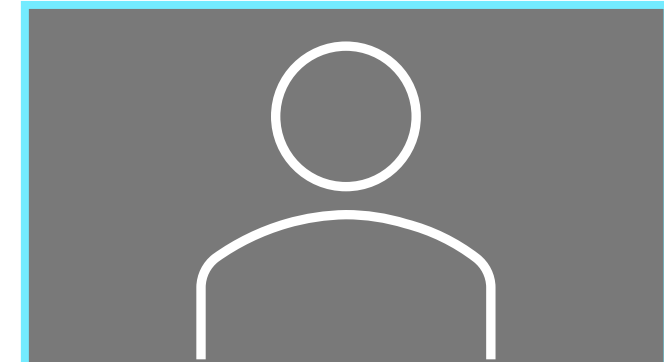
3197 ➡ 603
dimensiones

37 ➡ 2849
positivos

TFM

Preprocesamiento ➡ Data Augmentation ➡ Estrategias

37 ➡ 2849
positivos



Preprocesamiento

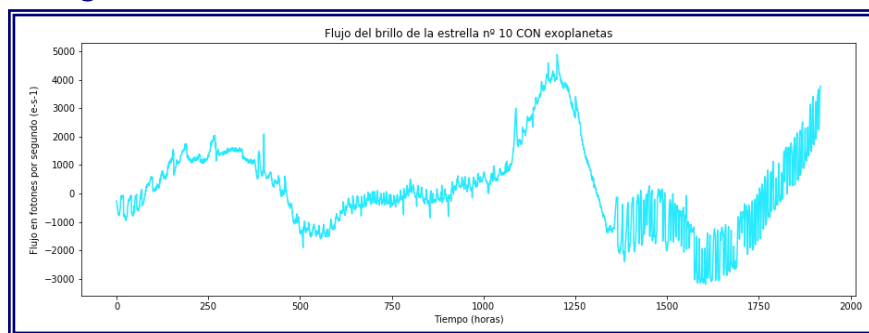
TFG

1. Abstracción de tendencia

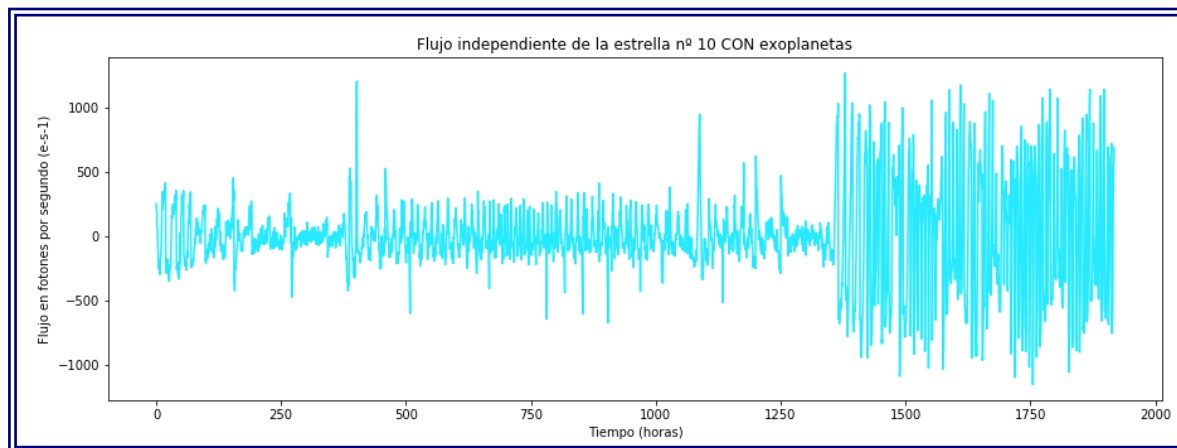
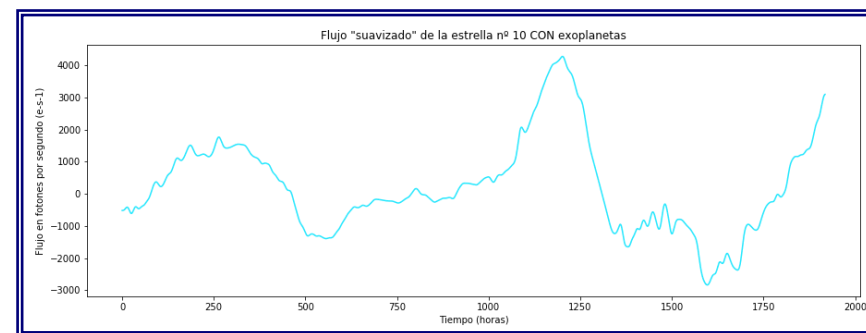
2. Normalización

3. Outliers superiores

Original



Tendencia



TFG

1. Abstracción de tendencia

2. Normalización

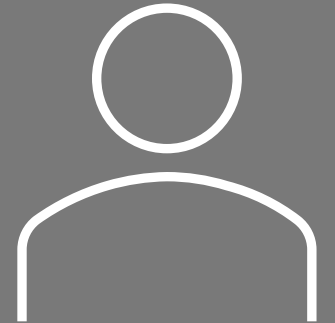
3. Outliers superiores

TFM

1. Abstracción de tendencia

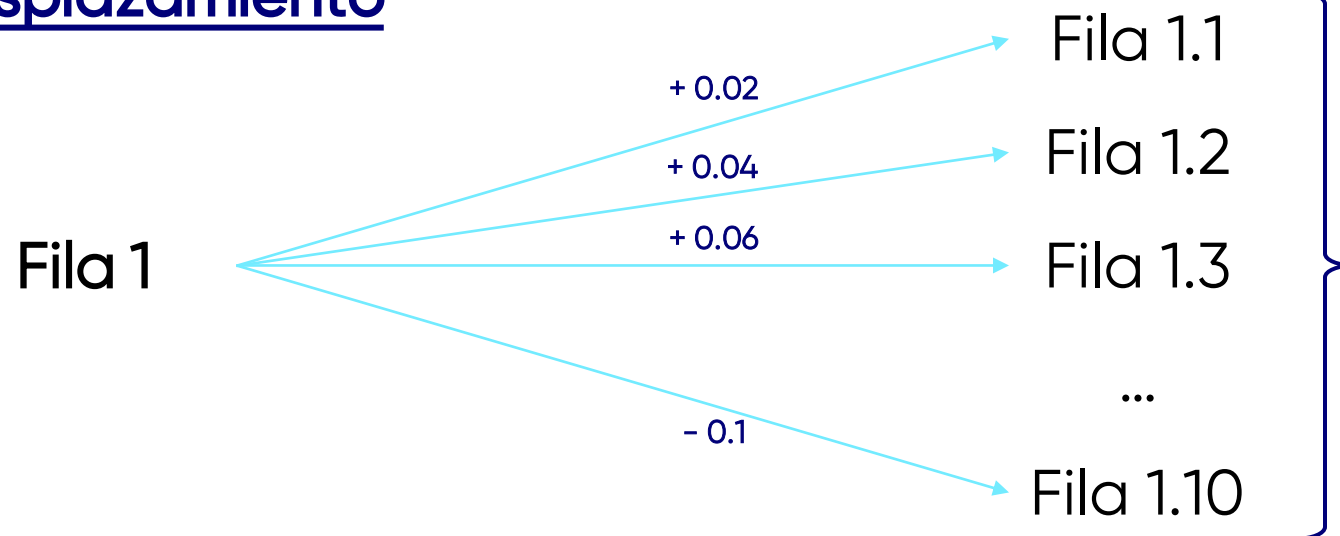
2. *Outliers superiores*

3. *Normalización [0,1]*



Data Augmentation

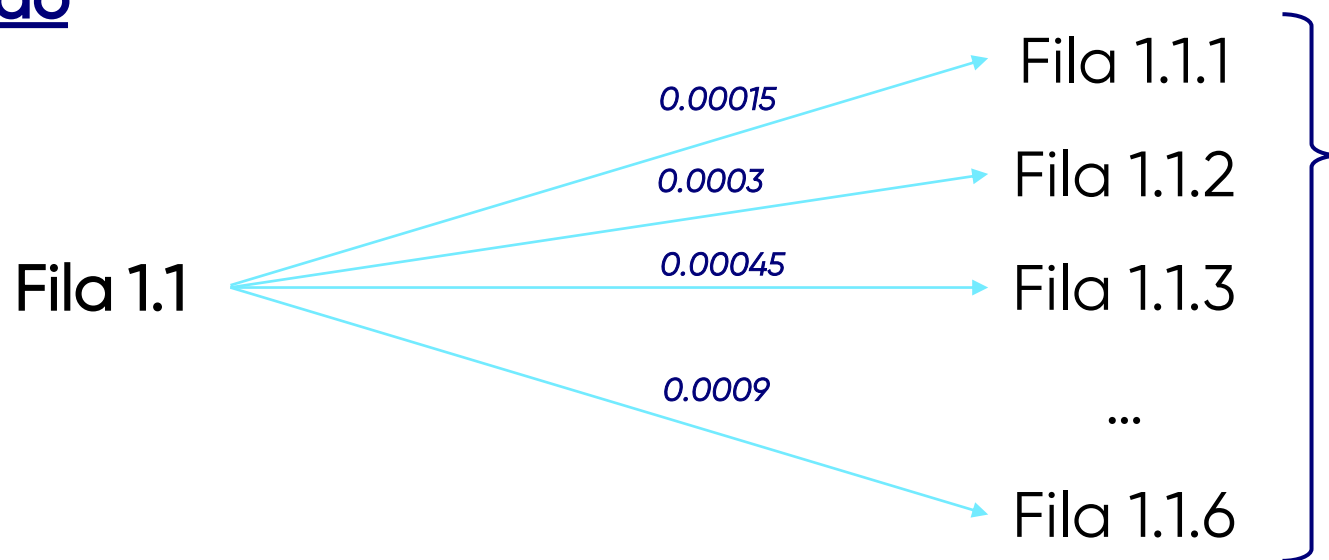
Desplazamiento



$$37 + 37 \times 10 =$$

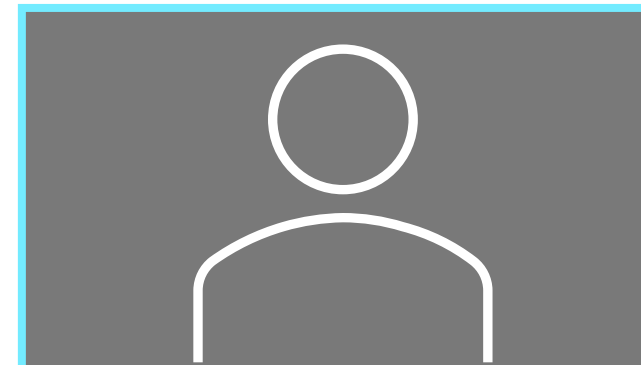
407 filas

Ruido

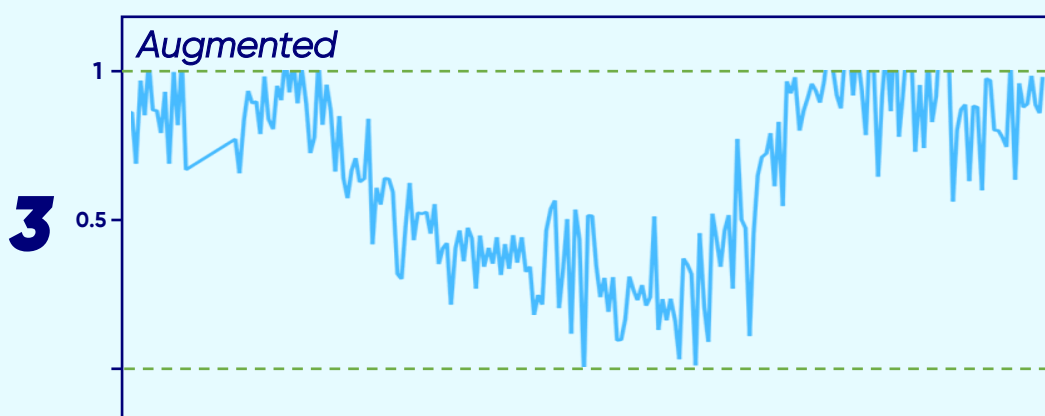
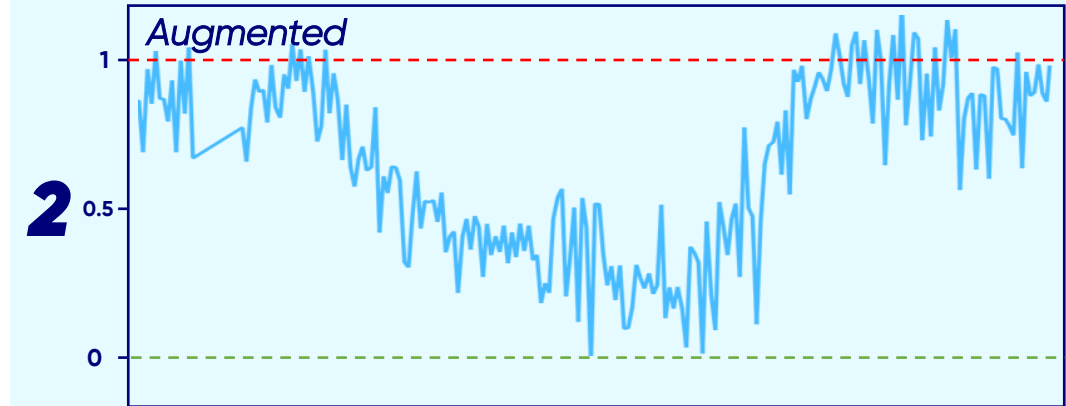
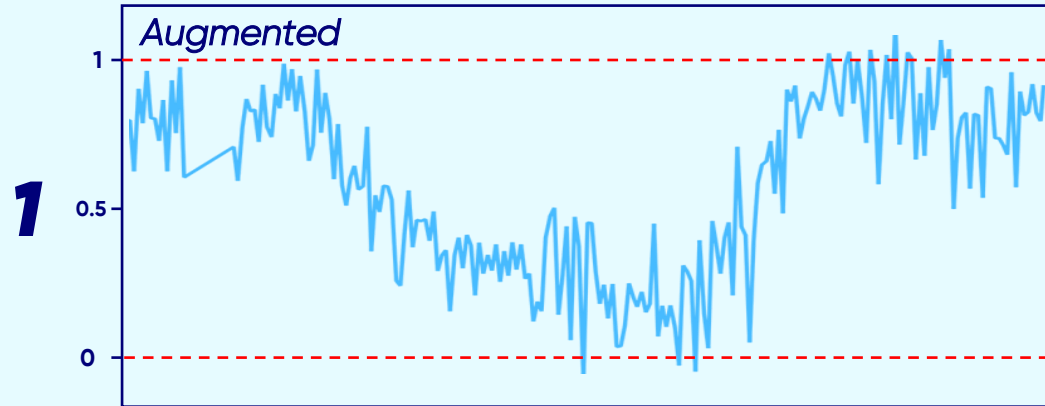
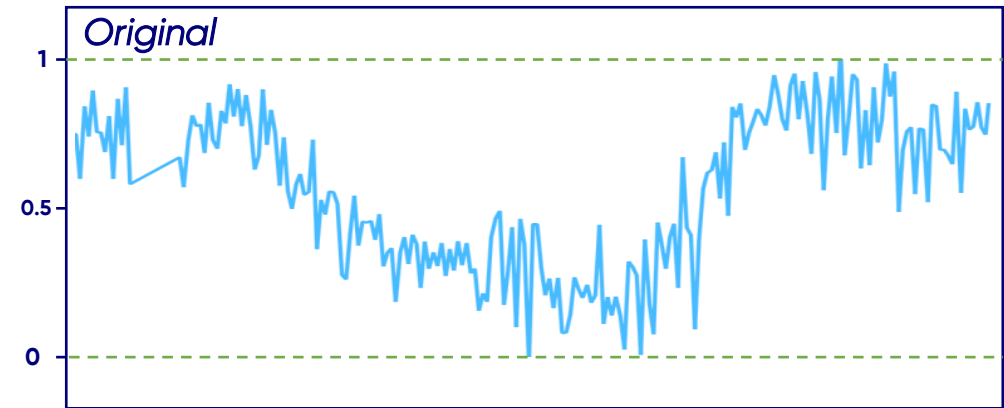


$$407 + 407 \times 6 =$$

2849 filas

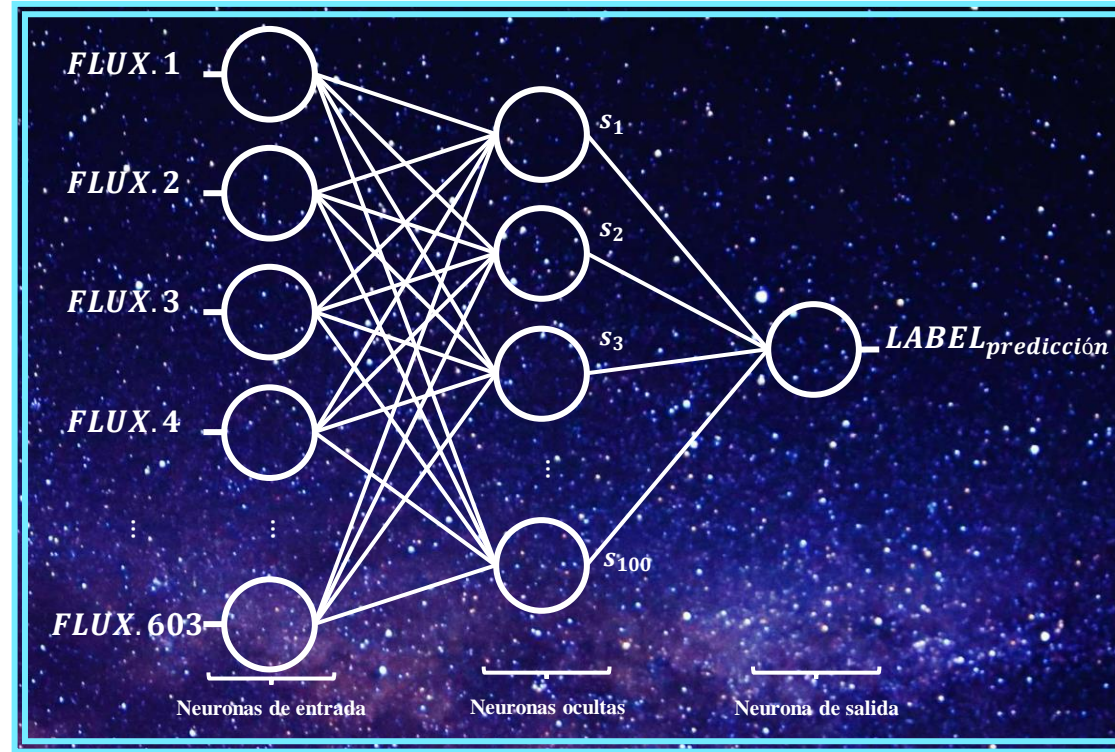


Nuevas consideraciones



Clasificación

TFG

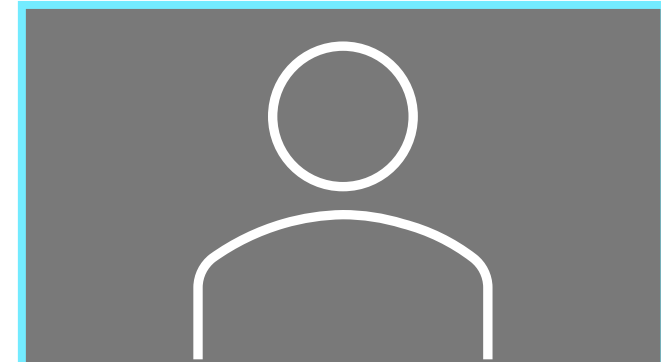


TFM

Estrategia A – Distribuciones

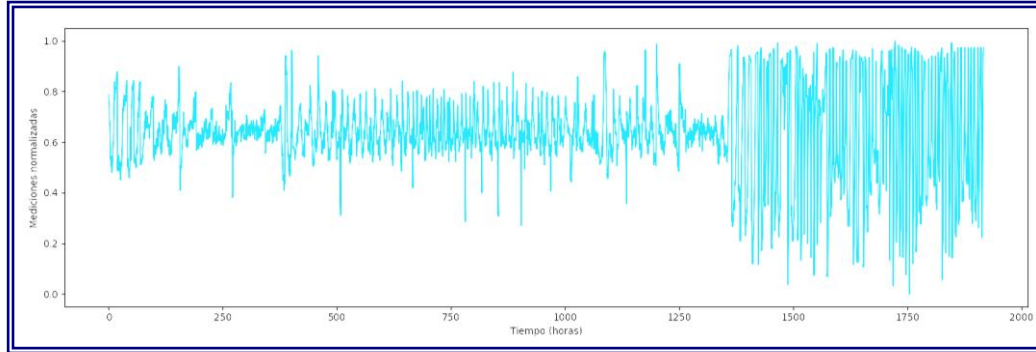
Estrategia B – Selección dimensionalidad

Estrategia C – Mecanismo de atención

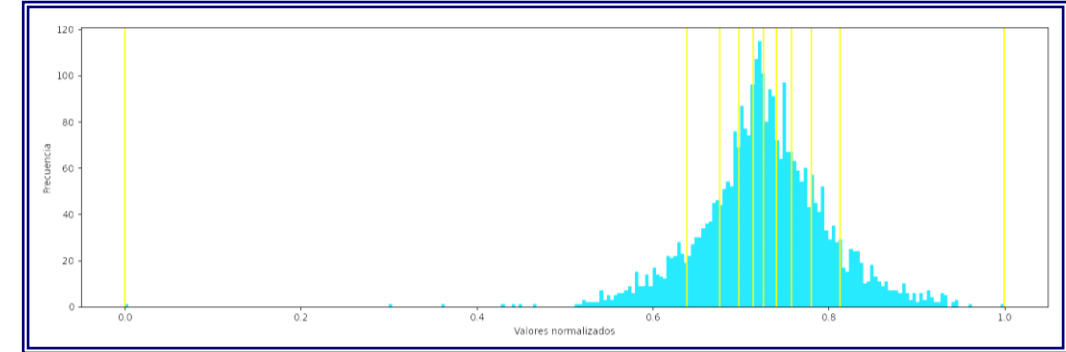


Distribuciones

Original



Distribución



Dimensiones

- Media
- Varianza
- Desviación estándar
- Quintiles 0-10

➔ 14



Estrategia A

Modelos

Capas Learning rate Batch size

5	0.01	128
----------	-------------	------------

5	0.01	256
---	------	-----

5	0.001	128
---	-------	-----

5	0.001	256
---	-------	-----

5	0.0001	128
---	--------	-----

5	0.0001	256
---	--------	-----

4	0.01	128
---	------	-----

4	0.01	256
----------	-------------	------------

4	0.001	128
---	-------	-----

4	0.001	256
---	-------	-----

4	0.0001	128
---	--------	-----

4	0.0001	256
---	--------	-----

3	0.01	128
----------	-------------	------------

3	0.01	256
---	------	-----

3	0.001	128
---	-------	-----

3	0.001	256
---	-------	-----

3	0.0001	128
---	--------	-----

3	0.0001	256
---	--------	-----

Resultados

Predicción

SIN exoplanetas	CON exoplanetas
-----------------	-----------------

Real	SIN exoplanetas	564	1
	CON exoplanetas	3	2

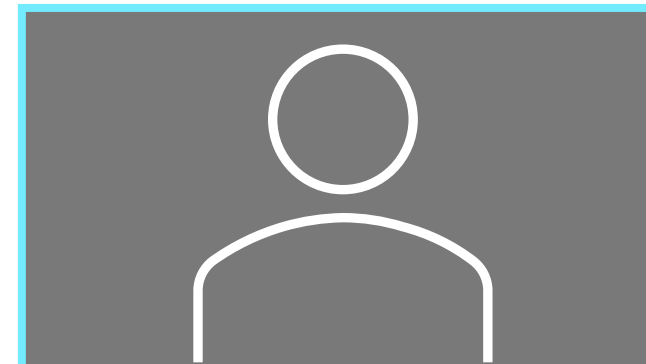
Métricas

Accuracy	99.3%
----------	-------

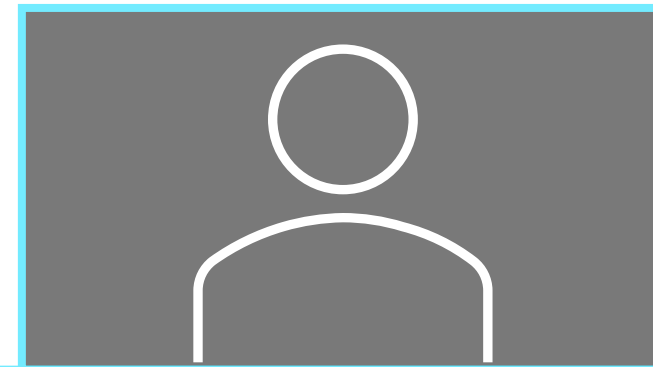
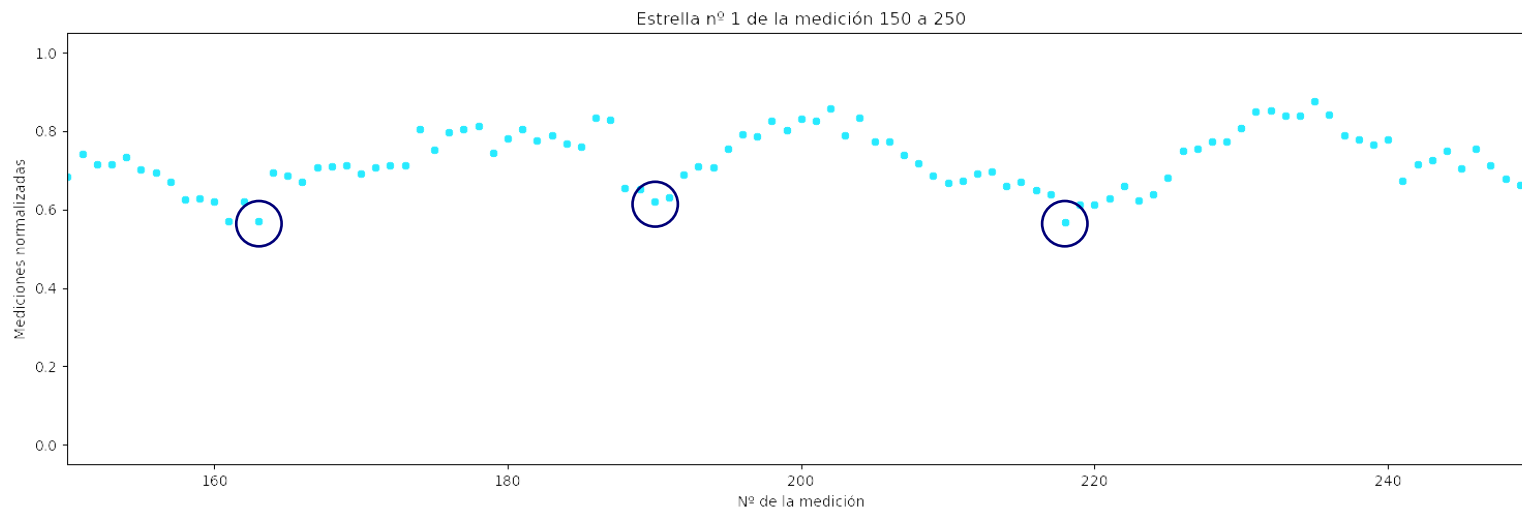
Precision	66.7%
-----------	-------

Recall	40%
--------	-----

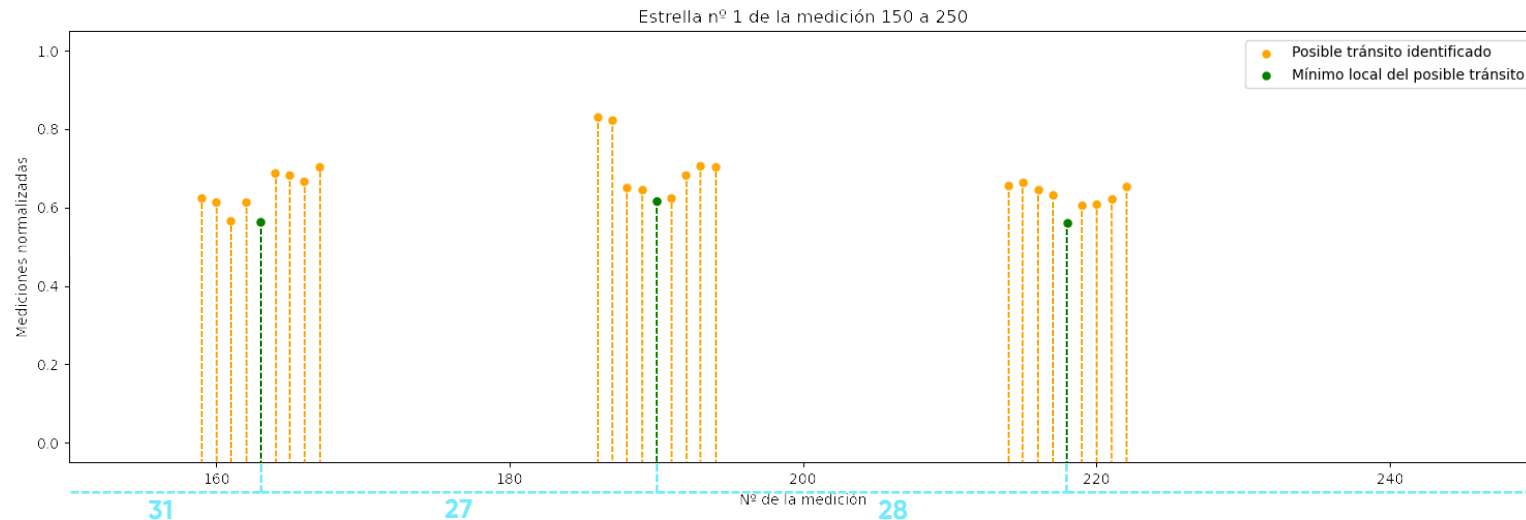
F1	50%
----	------------



Selección de dimensionalidad



Selección de dimensionalidad



Vector mediciones

...	.59	.61	.58	.71	.70	.64	.63	.61	.62	.69	.71	.63	.62	.57	.60	.61	...
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Vector posición

...	163	190	218	...
-----	-----	-----	-----	-----

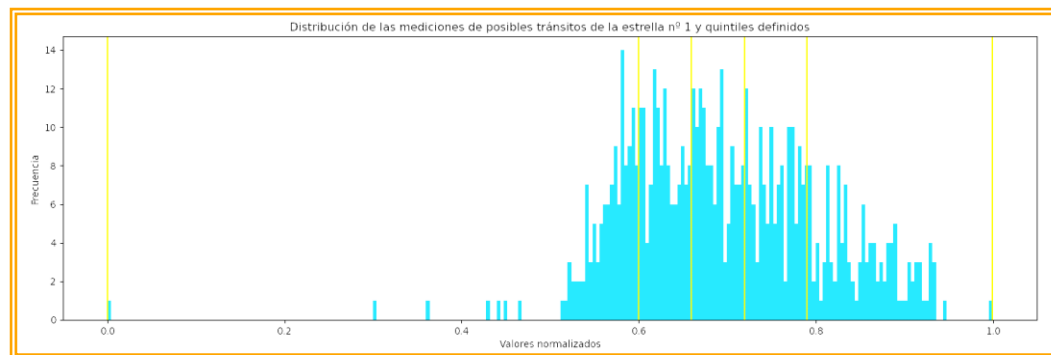
Vector frecuencia

...	31	27	28	...
-----	----	----	----	-----



Selección de dimensionalidad

Distribución mediciones



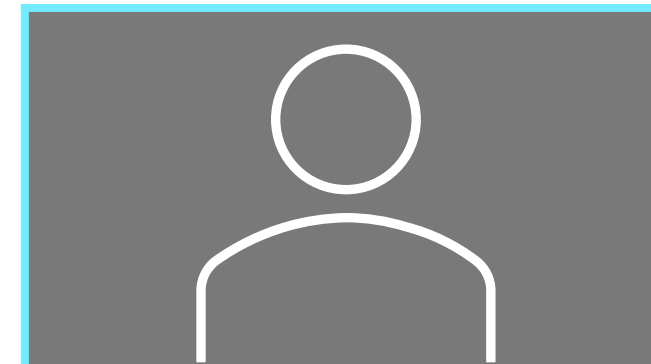
Distribución frecuencias



Dimensiones

21

- | | | |
|---------------------|---------------------|---------------------|
| Media | Media | Número de tránsitos |
| Mediana | Mediana | |
| Varianza | Varianza | |
| Desviación estándar | Desviación estándar | |
| Quintiles 0-5 | Quintiles 0-5 | |



Estrategia B

Modelos

Capas	Learning rate	Batch size
6	0.01	64
6	0.01	128
6	0.001	64
6	0.001	128
6	0.0001	64
6	0.0001	128
5	0.01	64
5	0.01	128
5	0.001	64
5	0.001	128
5	0.0001	64
5	0.0001	128
4	0.01	64
4	0.01	128
4	0.001	64
4	0.001	128
4	0.0001	64
4	0.0001	128

Real

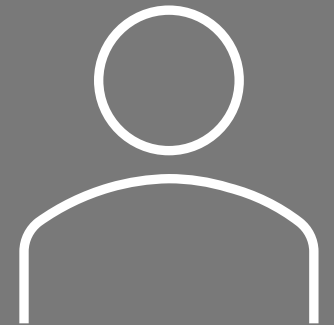
SIN exoplanetas
CON exoplanetas

Resultados

Predicción	
SIN exoplanetas	CON exoplanetas
560	5
3	2

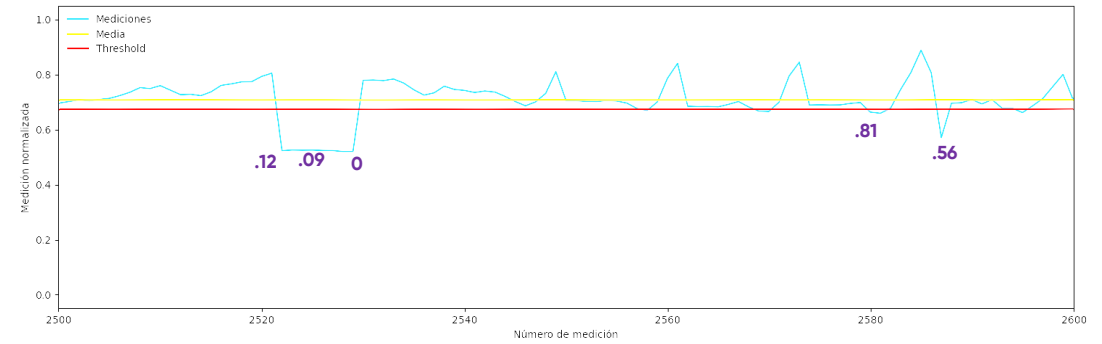
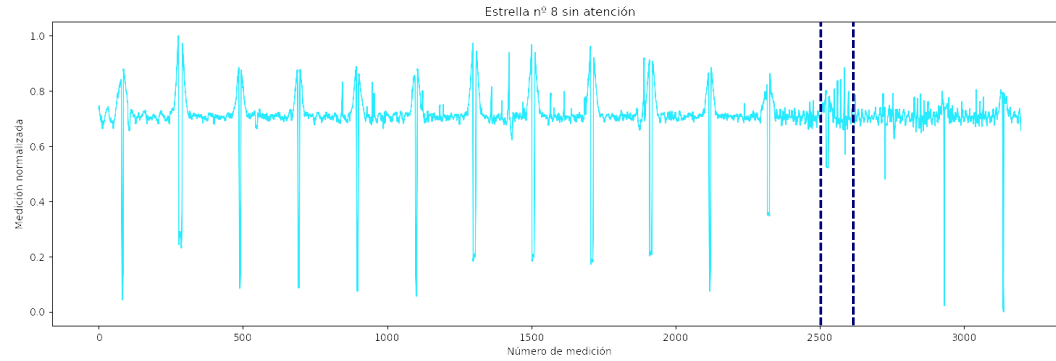
Métricas

Accuracy	98.6%
Precision	28.6%
Recall	40%
F1	33.3%

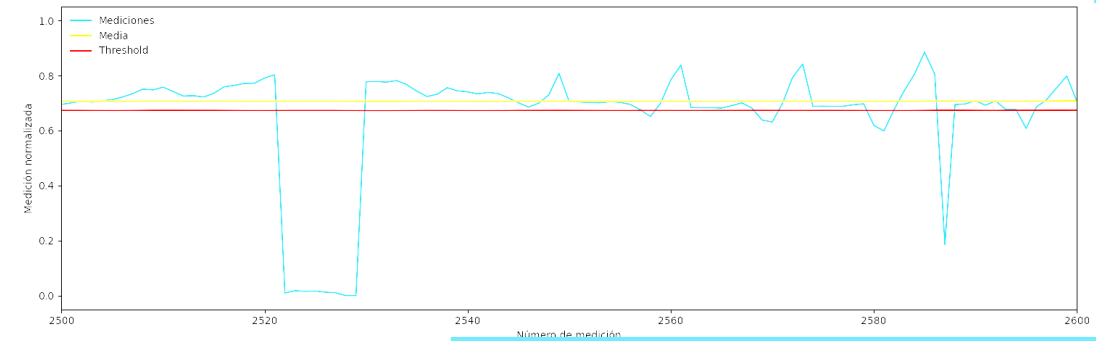
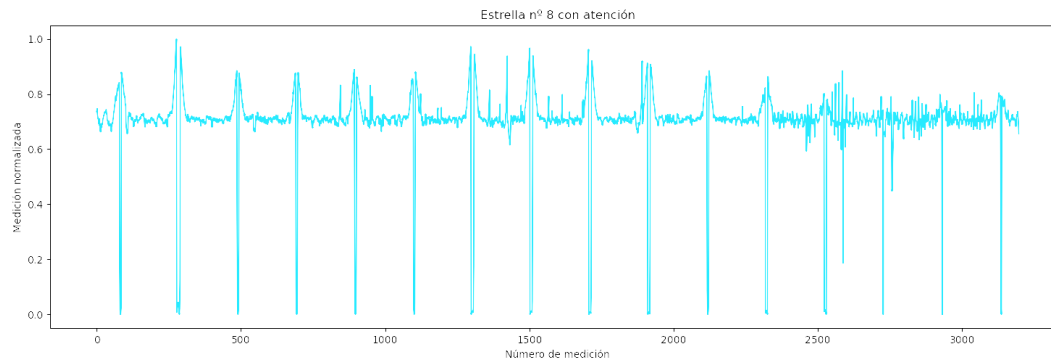


Estrategia C

Mecanismo de atención



Vector atención



Estrategia A – Distribuciones

Estrategia B – Selección dimensionalidad



Modelos

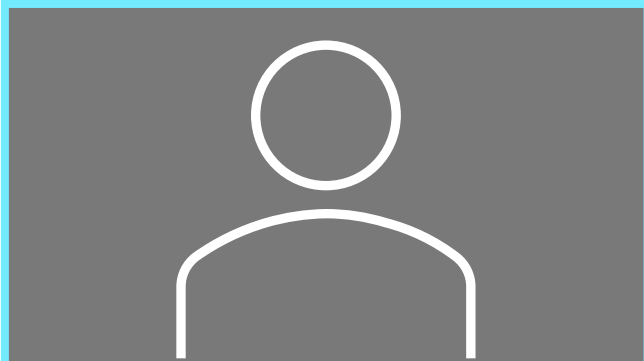
Capas	Learning rate	Batch size
5	0.01	128
5	0.01	256
5	0.001	128
5	0.001	256
5	0.0001	128
5	0.0001	256
4	0.01	128
4	0.01	256
4	0.001	128
4	0.001	256
4	0.0001	128
4	0.0001	256
3	0.01	128
3	0.01	256
3	0.001	128
3	0.001	256
3	0.0001	128
3	0.0001	256

Resultados

		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN exoplanetas	552	13
	CON exoplanetas	3	2

Métricas

Accuracy	97.2%
Precision	13.3%
Recall	40%
F1	20%



Modelos

Capas	Learning rate	Batch size
6	0.01	64
6	0.01	128
6	0.001	64
6	0.001	128
6	0.0001	64
6	0.0001	128
5	0.01	64
5	0.01	128
5	0.001	64
5	0.001	128
5	0.0001	64
5	0.0001	128
4	0.01	64
4	0.01	128
4	0.001	64
4	0.001	128
4	0.0001	64
4	0.0001	128

Real

SIN exoplanetas
CON exoplanetas

Resultados

Predicción

SIN exoplanetas	CON exoplanetas
564	1
2	3

Métricas

Accuracy	99.5%
Precision	75%
Recall	60%
F1	66.7%



Resultados

Estrategia A F1 – 50%

		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN	564	1
	CON	3	2

Estrategia B F1 – 33.3%

		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN	560	5
	CON	3	2

TFG F1 – 40%

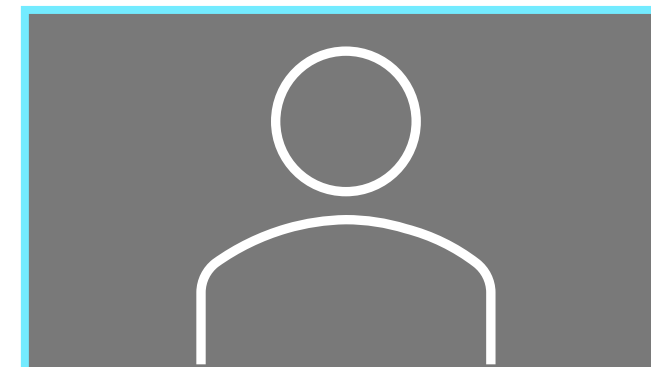
		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN	562	3
	CON	3	2

Estrategia C_A F1 – 20%

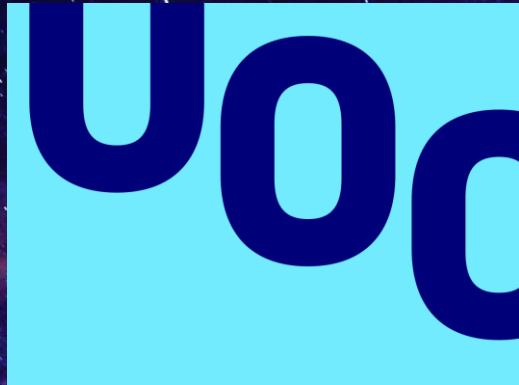
		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN	552	13
	CON	3	2

Estrategia C_B F1 – 66.7%

		Predicción	
		SIN exoplanetas	CON exoplanetas
Real	SIN	564	1
	CON	2	3



Machine Learning para la detección de exoplanetas: revisión y nuevos enfoques



Trabajo de Fin de Máster
Máster en Ciencia de Datos
Enero 2023

Autor

Javier Gómez de Diego

Tutora

Laura Ruiz Dern