Universitat Oberta de Catalunya

uoc.edu

Clasificación automática de objetos astronómicos por fotometría en series históricas recogidas por el Large Synoptic Survey Telescope (LSST)

Alumno: **Luis Enrique Arribas Zapater**
Grado Ingeniería Informática
Inteligencia Artificial

AGENDA

- **Motivación del proyecto**
- **Descripción de los datos y contexto astronómico**
- **Extracción de atributos**
- **Clasificadores**
- **Resultados**
- **Conclusiones**

OBJETIVOS

OBJETIVOS GENERALES

- **Resolver un problema de clasificación de objetos astronómicos mediante técnicas de aprendizaje computacional y minería de datos.**
- **Evaluar el modelo de clasificación mediante métricas que se ajusten al problema**

OBJETIVOS ESPECÍFICOS

- **Análisis y preparación del conjunto de datos para su estudio**
- **Transformar los datos en series temporales y extraer nuevos atributos de ellas**
- **Diseñar un modelo de clasificación sobre el conjunto de datos resultante**
- **Racionalizar los recursos de cómputo**
- **Implementar un *framewok* de machine learning que cumpla estos objetivos**

Universitat Oberta de Catalunya

uoc.edu

MOTIVACIÓN

- **¿Qué es el telescopio LSST?**
- **Primera misión científica: 10 años de observación**
- **Clasificación automática de 3.5 M de objetos**

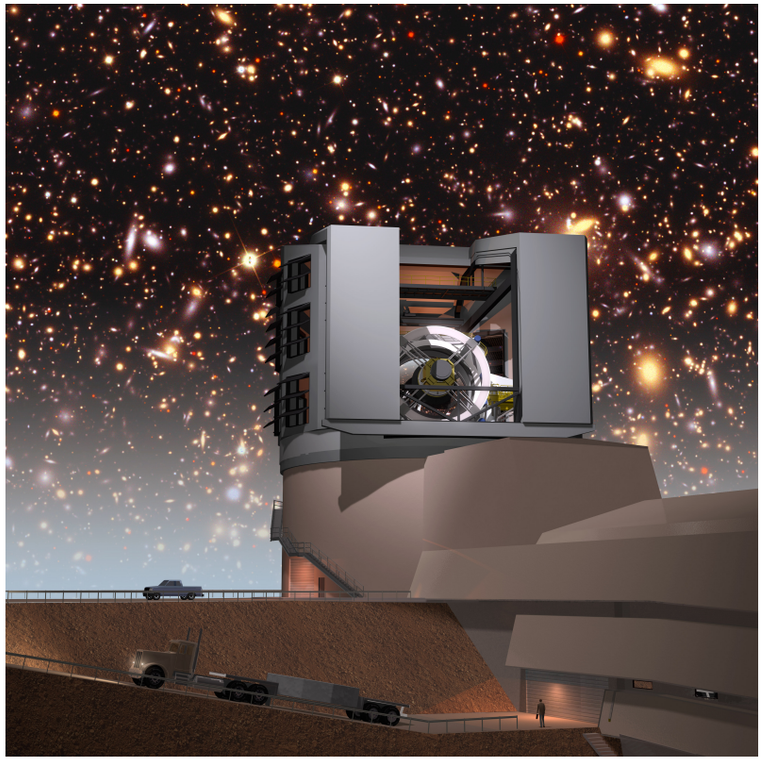


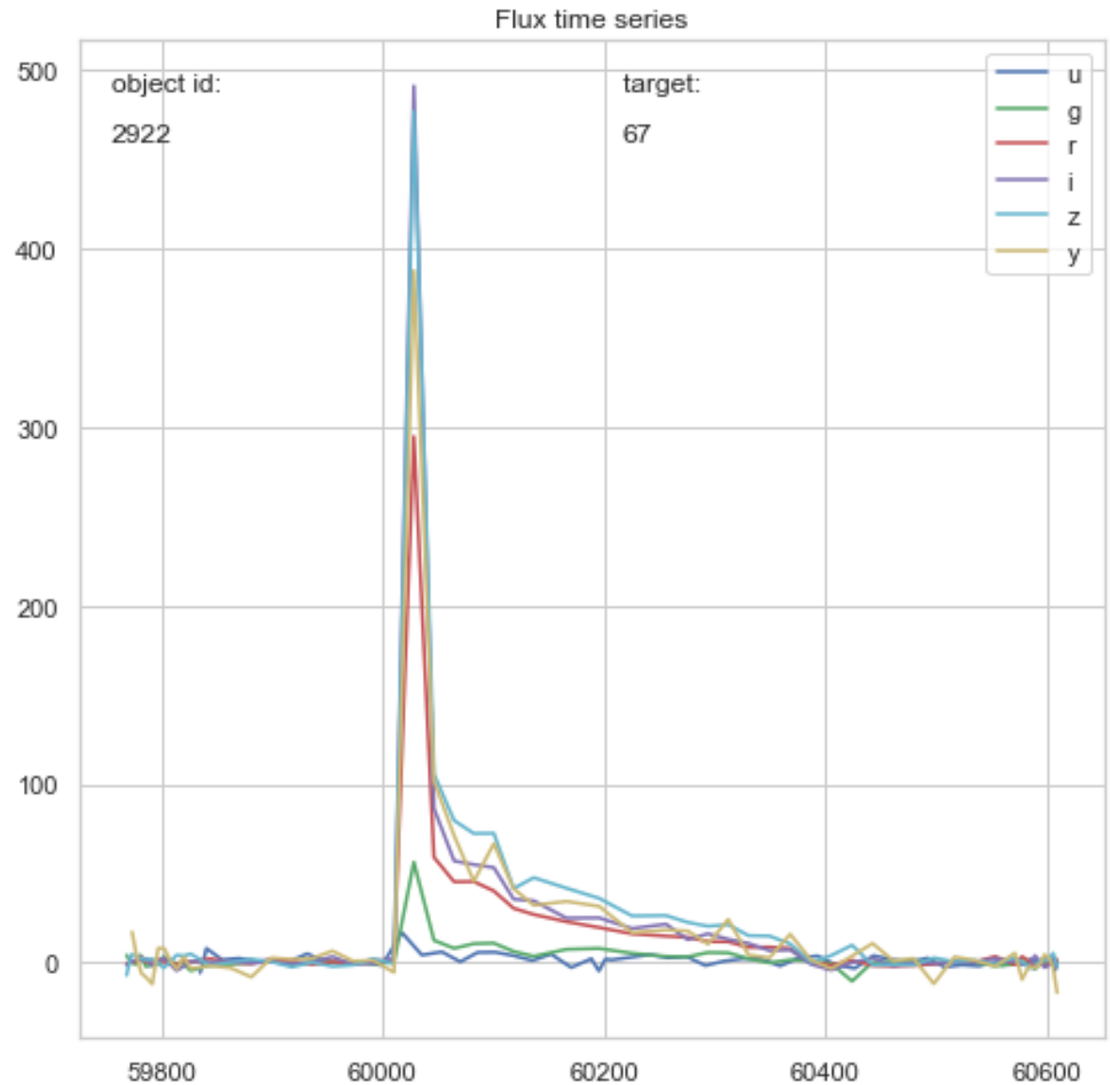
Imagen:https://www.lsst.org/

PLaSTiCC open challenge

kaggle

LOS DATOS

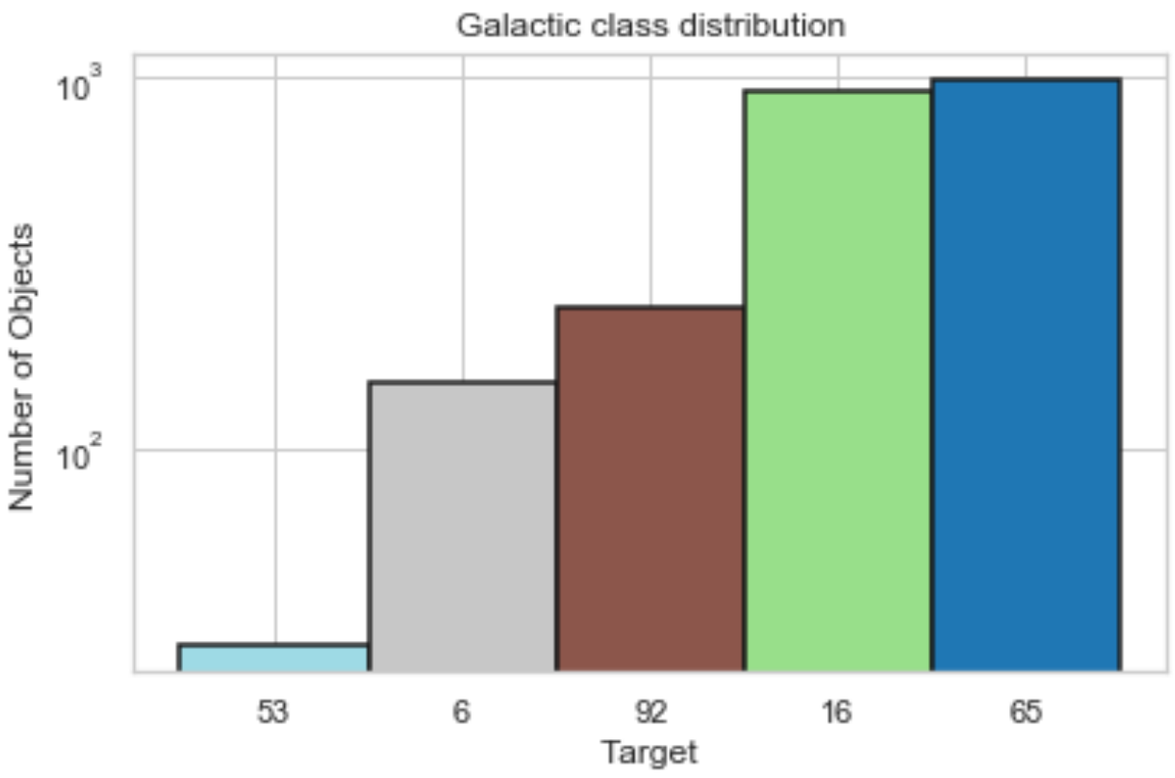
- **Los datos disponibles son una simulación de 8000 series de flujo, tiempo y error en 6 bandas de frecuencia**
- **Los metadatos son características de estos objetos no relacionadas con el tiempo**

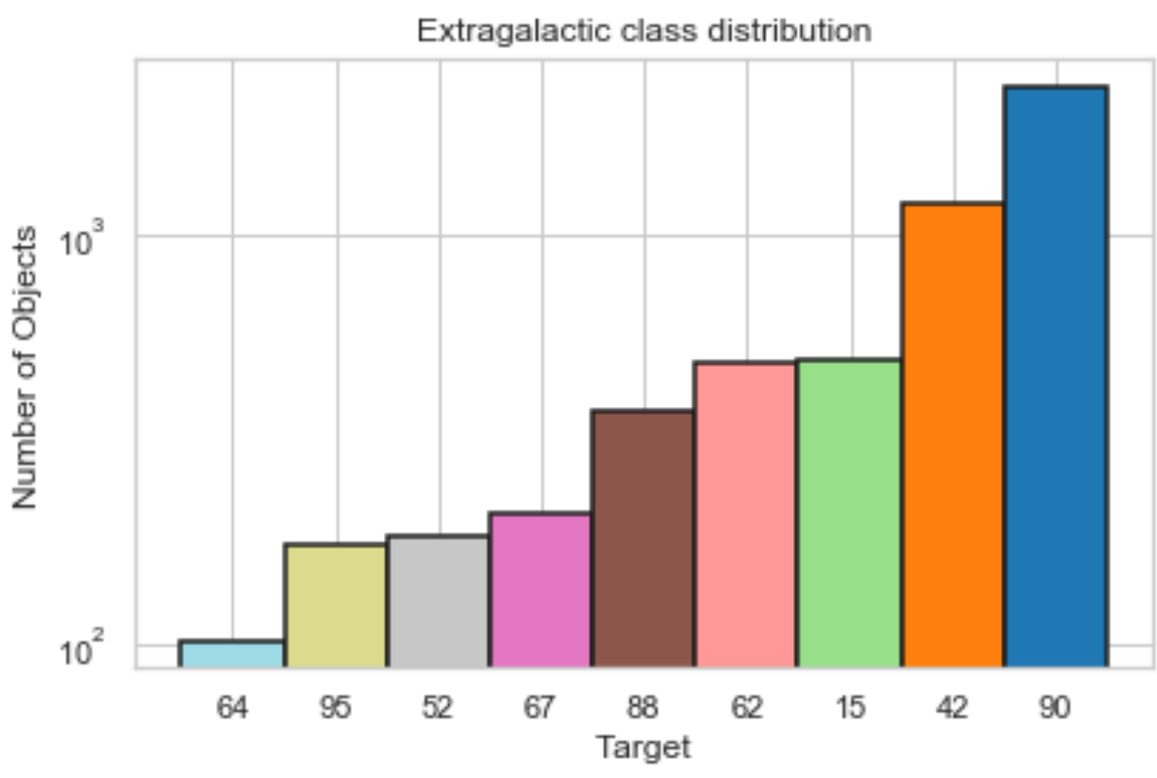


CONTEXTO ASTRONÓMICO

- **Desplazamiento al rojo (*Z*)**
- **Espectrometría vs fotometría**
- **Estrellas variables y curvas de luz**
- **Clases presentes en los datos**

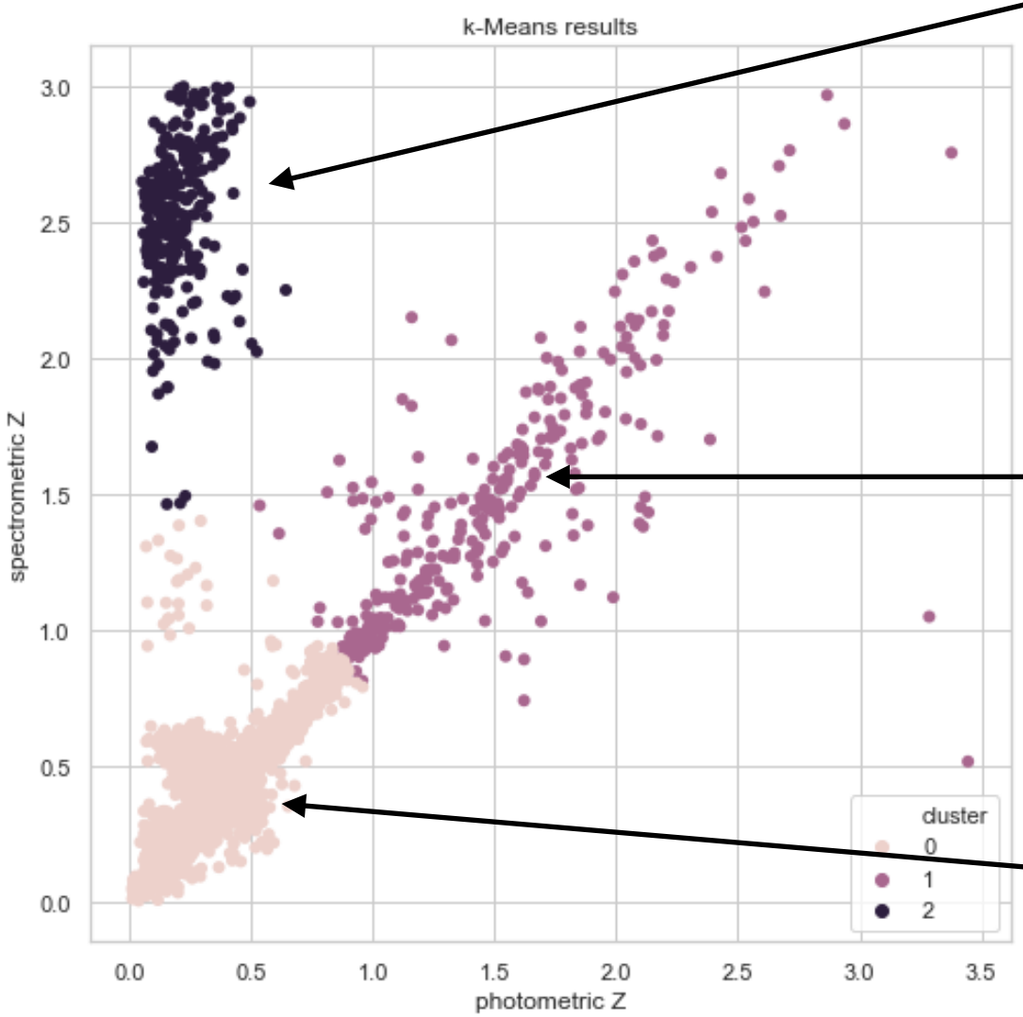
La distribución de clases en los subconjuntos de objetos galácticos y extragalácticos es completamente diferente

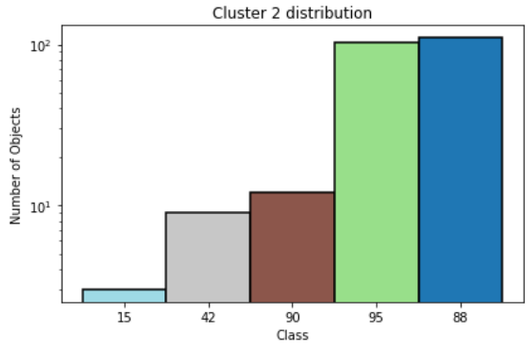


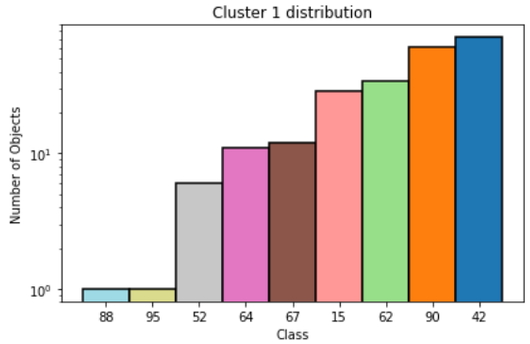


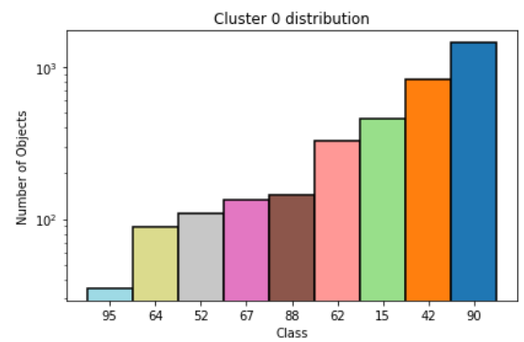
ANÁLISIS DE LOS DATOS II

Clustering sobre los datos extragalácticos





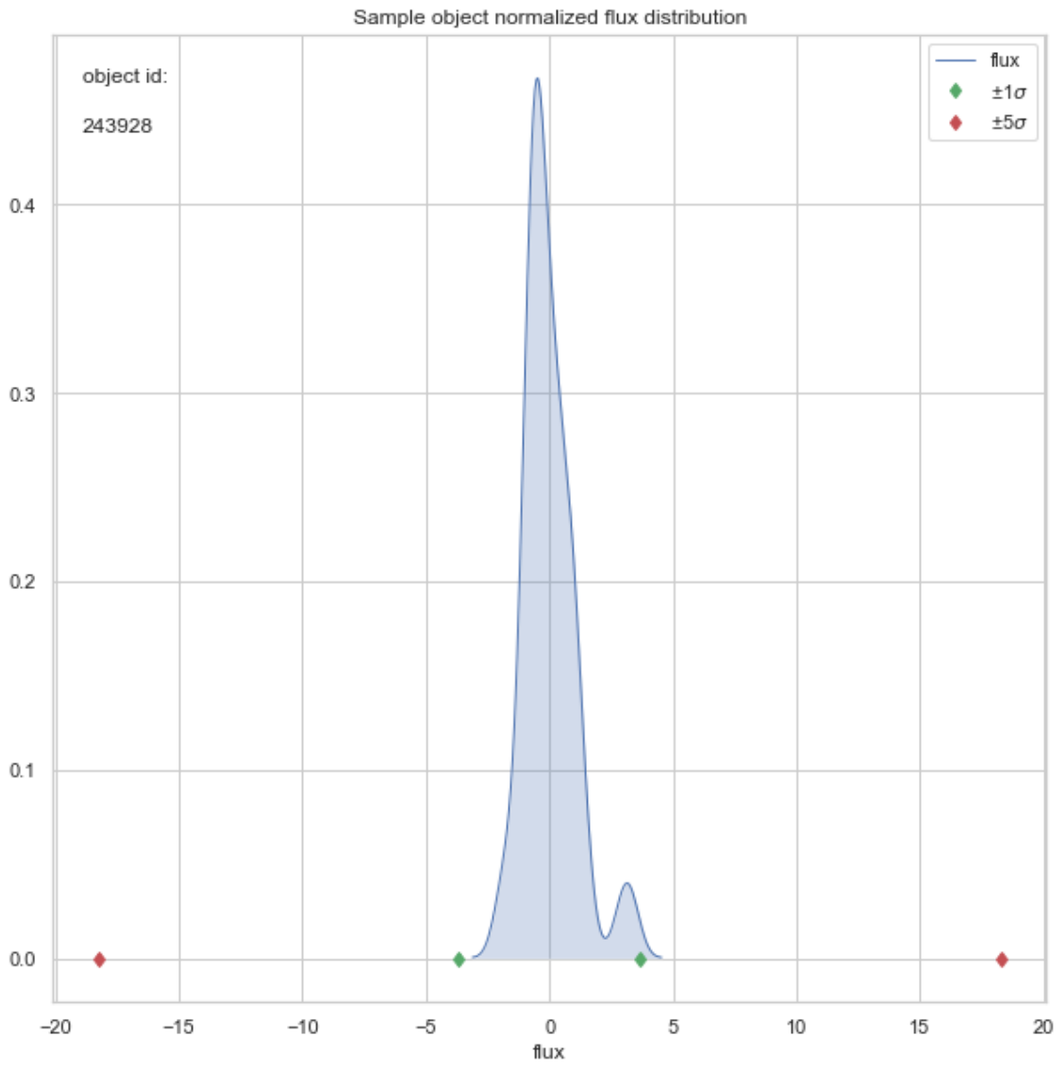




Clasificación independiente de los tres clusters

PROCESADO DE LOS DATOS

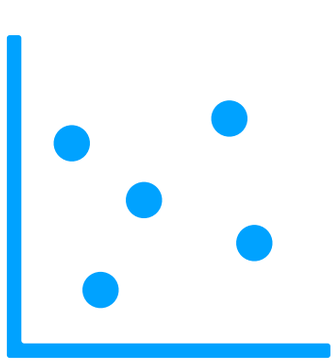
- **Magnitud**
- **Color**
- **Inferencia bayesiana**
- **Eliminación del ruido**



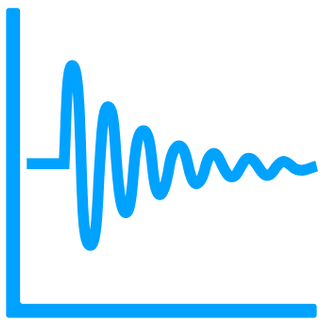
Universitat Oberta de Catalunya

uoc.edu

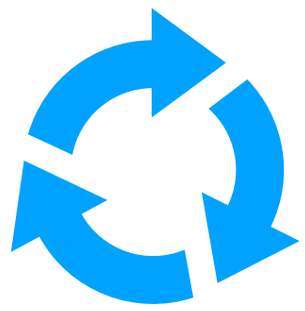
EXTRACCIÓN DE ATRIBUTOS



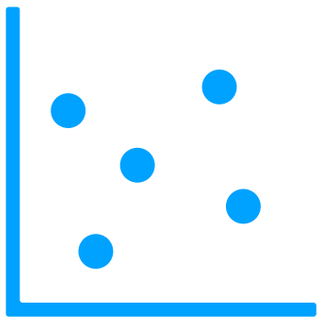
DATOS



SERIES TEMPORALES



EXTRACCIÓN ATRIBUTOS

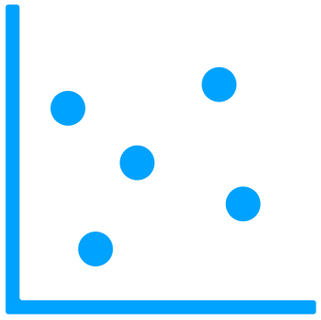


NUEVOS DATOS

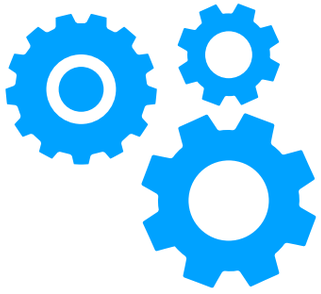
Universitat Oberta de Catalunya

uoc.edu

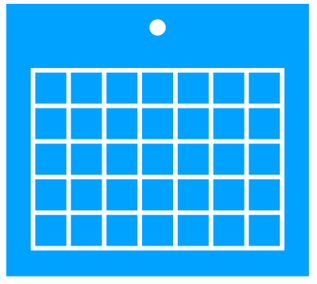
REDUCCIÓN DE ATRIBUTOS



NUEVOS DATOS



REDUCCIÓN DE ATRIBUTOS

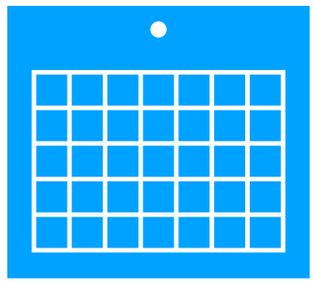


RANKING ATRIBUTOS

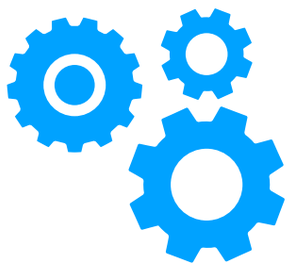
Universitat Oberta de Catalunya

uoc.edu

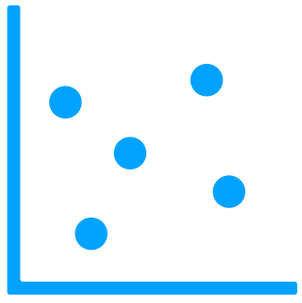
CONSTRUCCIÓN DE LOS CLASIFICADORES



RANKING ATRIBUTOS

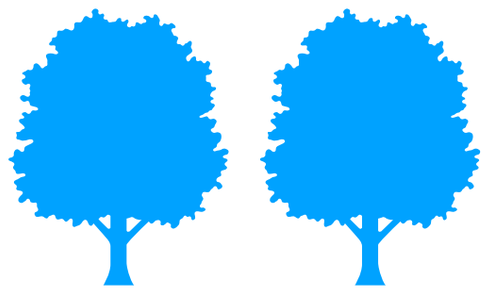


CONSTRUCCIÓN CLASIFICADORES

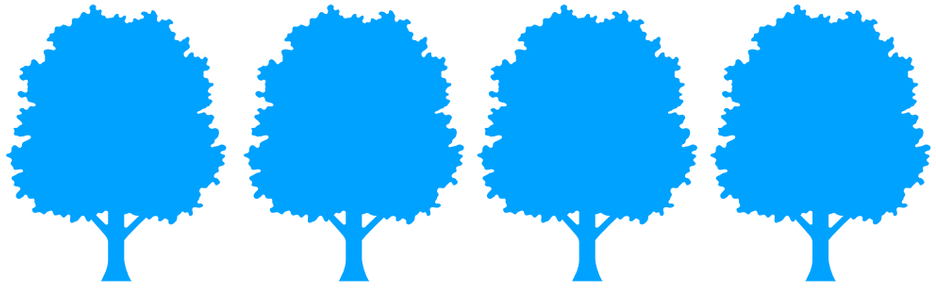


OTROS DATOS

1 BOSQUE ALEATORIO



2 BOSQUES ALEATORIOS



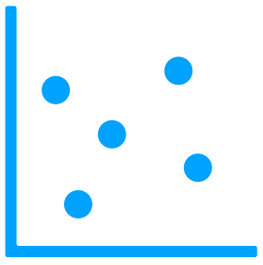
4 BOSQUES ALEATORIOS

Universitat Oberta de Catalunya

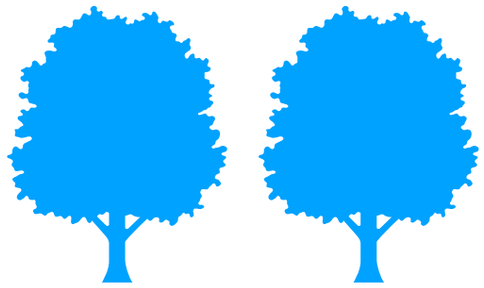
uoc.edu

VALIDACIÓN DE LOS CLASIFICADORES

1 BOSQUE ALEATORIO

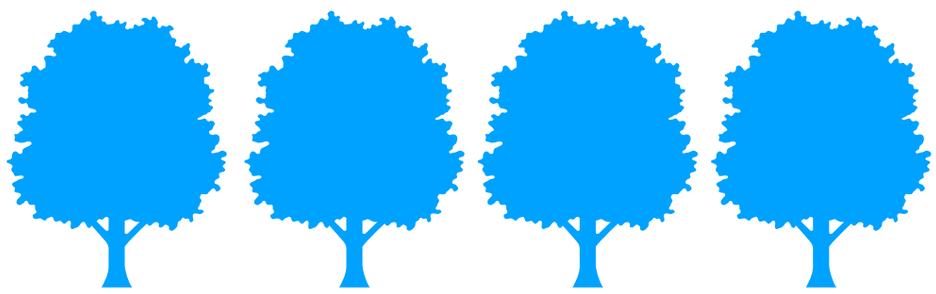


OTROS DATOS



2 BOSQUES ALEATORIOS

5 FOLD CROSS VALIDATION



4 BOSQUES ALEATORIOS

EVALUACIÓN DE RESULTADOS

5 FOLD CROSS VALIDATION

Tabla 4.1 Resultados del modelo

Test n	ne	bnr	1RF accuracy	2RF accuracy	gal. accuracy	2RFex gal acc	4RF accuracy
1	no	no	0.80 (+/- 0.04)	0.80 (+/- 0.04)	0.98 (+/- 0.04)	0.74 (+/- 0.04)	0.81 (+/- 0.07)
2	no	si	0.80 (+/- 0.03)	0.80 (+/- 0.04)	0.98 (+/- 0.03)	0.74 (+/- 0.05)	0.80 (+/- 0.06)
3	si	no	0.80 (+/- 0.03)	0.80 (+/- 0.03)	0.98 (+/- 0.02)	0.73 (+/- 0.03)	0.80 (+/- 0.06)
4	si	si	**0.81 (+/- 0.03)**	0.80 (+/- 0.04)	0.98 (+/- 0.02)	0.74 (+/- 0.05)	0.79 (+/- 0.05)

Test n	ne	bnr	4RF gal accuracy	k=0 accuracy	k=1 accuracy	k=2 accuracy	feats
1	no	no	0.98 (+/- 0.04)	0.75 (+/- 0.07)	0.84 (+/- 0.12)	**0.52 (+/- 0.18)**	53
2	no	si	0.98 (+/- 0.03)	0.75 (+/- 0.06)	0.84 (+/- 0.12)	**0.50 (+/- 0.17)**	54
3	si	no	0.98 (+/- 0.02)	0.75 (+/- 0.06)	0.83 (+/- 0.10)	**0.52 (+/- 0.21)**	57
4	si	si	0.98 (+/- 0.02)	0.74 (+/- 0.05)	0.82 (+/- 0.11)	**0.45 (+/- 0.14)**	100

DISCUSIÓN DE RESULTADOS

- **La mejor clasificación se obtiene con un único bosque aleatorio**
- **Z es el atributo de mayor importancia del *ranking***
- **El atributo calculado para la magnitud ocupa en segundo lugar del *ranking***
- **El proceso de eliminación de ruido e inferencia bayesiana no mejoran el *accuracy* pero reducen el error. Es posible que los datos hayan sido tratado previamente**
- **La división y clasificación separada de los datos no aporta la mejora esperada, probablemente porque el subconjunto es pequeño y *Z* es inexacto**
- **El color calculado queda en el puesto 44, ya que no se ha corregido en función de *Z***
- **El modelo detecta mejor los objetos de tipo periódico**

FUTURAS LÍNEAS DE TRABAJO

- **Sustituir algunos bosques aleatorios de los clasificadores múltiples por otros tipos de clasificadores desde otros paradigmas**
- **Mejorar la estimación de Z mediante procesos de regresión**
- **Incorporar a la extracción de atributos funciones que relacionen las bandas de frecuencia**
- **Corregir el color en función de Z**

CONCLUSIONES

- **Cumplimiento satisfactorio de los objetivos mayoritariamente**
- **Mala planificación temporal. No se ha planificado suficiente tiempo para experimentación y depuración del modelo**
- **Metodología mejorable. Hubiese sido más acertado tener un prototipo del clasificador primero y haber hecho experimentos conjuntamente con los análisis.**

Gracias por su atención