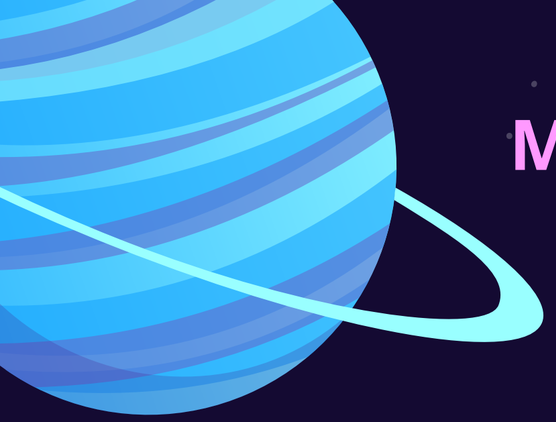
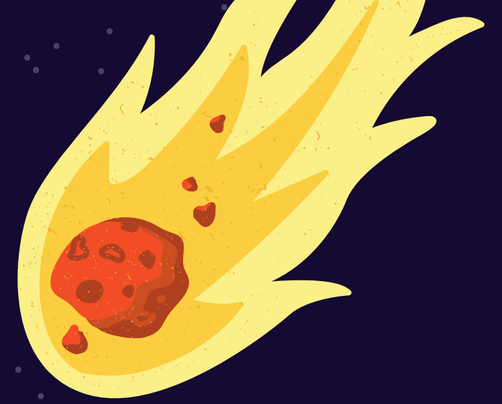
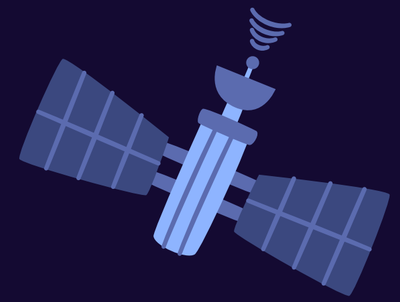
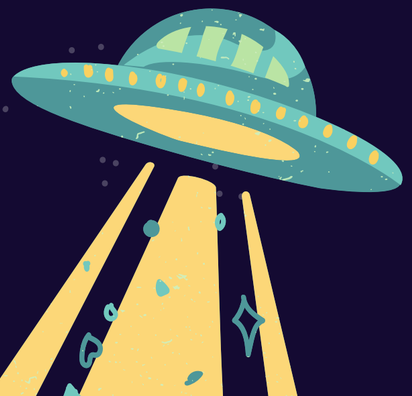
MINERIA DE DATOS

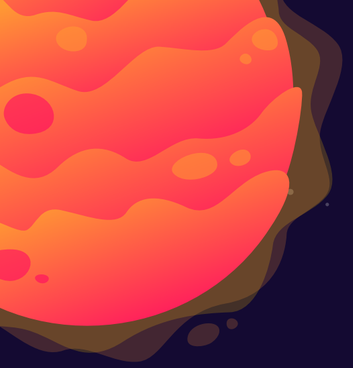
HITO Nº1

GRUPO 18


MOTIVACION:
OUR GALAXY



01 **Astronomía:**

- Tecnología
- Ciencias computacionales
- Gran poder de procesamiento → Gran volumen de datos
- Minería de datos → Herramienta fundamental

02 **Conceptos fundamentales:**

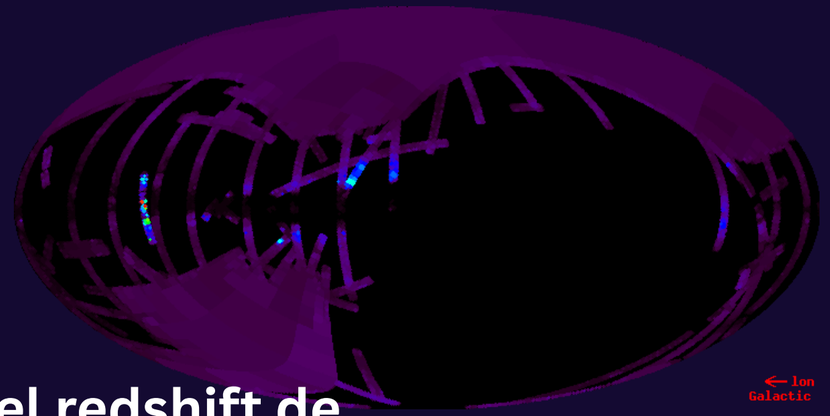
- Estrellas, agujeros negros, galaxias, AGN’s, QSO’s
- Redshift
- Magnitudes ugriz

03 ***Proyecto basado en: Brescia et. al 2015***

EXPLORACION DE DATOS:
OUR DATA

01

Datos usados: Sloan Digital Sky Survey (SDSS) Data Release 10. Alrededor de 3,6M de fuentes.



Pregunta de investigación: ¿Podemos clasificar y predecir el redshift de objetos usando fotometría?

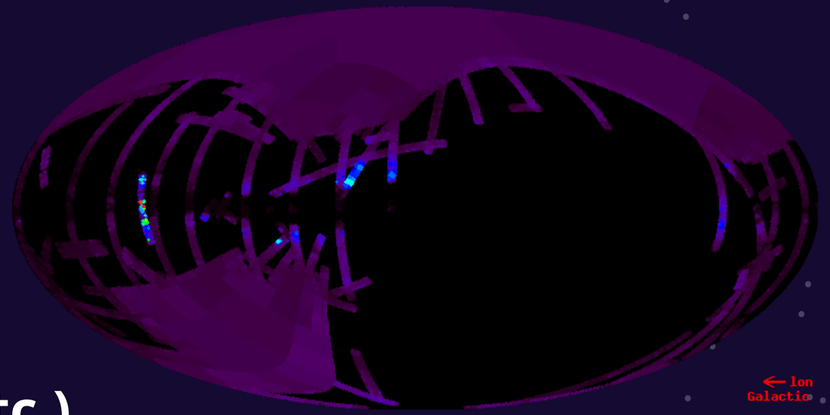
EXPLORACION DE DATOS:

Datos SDSS DR 10:

	objID	RAdeg	Dedeg	umag	gmag	rmag	imag	zmag	u'mag	g'mag	r'mag	i'mag	z'mag	specID	subclass	z	Qf
0	1237669697834516732	332.193566	6.213506	20.694180	20.070652	19.902487	19.833360	19.617641	20.702007	20.066208	19.903557	19.825838	19.612368	2615518058759874560	BROADLINE	0.000461	0.0
1	1237668331500929566	242.615071	9.127335	20.184590	20.125793	20.083216	20.264364	19.954805	20.170850	20.102804	20.053980	20.253532	20.000866	2844050179592579072	BROADLINE	0.000461	1.0
2	1237678859536302437	338.916923	11.483463	21.678028	21.292742	21.197329	21.170238	21.538815	21.655334	21.272352	21.143225	21.130280	21.298502	5684755593458876416	4.6062307E-4	0.000000	NaN
3	1237667730193581111	120.828351	9.728672	22.114391	21.625710	21.480595	21.595968	21.215242	22.183746	21.623096	21.470402	21.597070	21.242352	6185898695620820992	5.767762E-4	0.000000	NaN
4	1237654626786738691	116.210005	31.646465	21.893473	21.455797	21.365667	21.404415	21.234035	21.910515	21.379683	21.280743	21.330507	21.222286	5002396468775485440	5.7690655E-4	0.000000	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3602205	1237671686933971079	160.202834	79.700491	20.314724	20.115656	20.120022	19.846950	19.761147	20.332558	20.112700	20.103870	19.819887	19.801836	1	NaN	NaN	NaN
3602206	1237671686933971158	160.428454	79.625446	18.853277	18.651420	18.550852	18.575981	18.617682	18.875570	18.652070	18.556755	18.574717	18.629831	1	NaN	NaN	NaN
3602207	1237671686934037235	160.882926	79.442167	22.051360	21.963966	21.302395	20.819840	20.185960	21.975136	21.741077	21.046968	20.553333	20.002817	0	NaN	NaN	NaN
3602208	1237671686934037238	160.816211	79.432448	22.432056	22.067238	21.406551	21.063618	21.624708	22.473427	22.052916	21.424982	21.057661	21.563576	0	NaN	NaN	NaN
3602209	1237671686934036581	160.850917	79.513393	24.374573	20.937517	20.529442	20.632936	20.399504	24.458103	20.931923	20.535520	20.626488	20.409494	0	NaN	NaN	NaN

3602210 rows × 17 columns

EXPLORACION DE DATOS:
OUR DATA

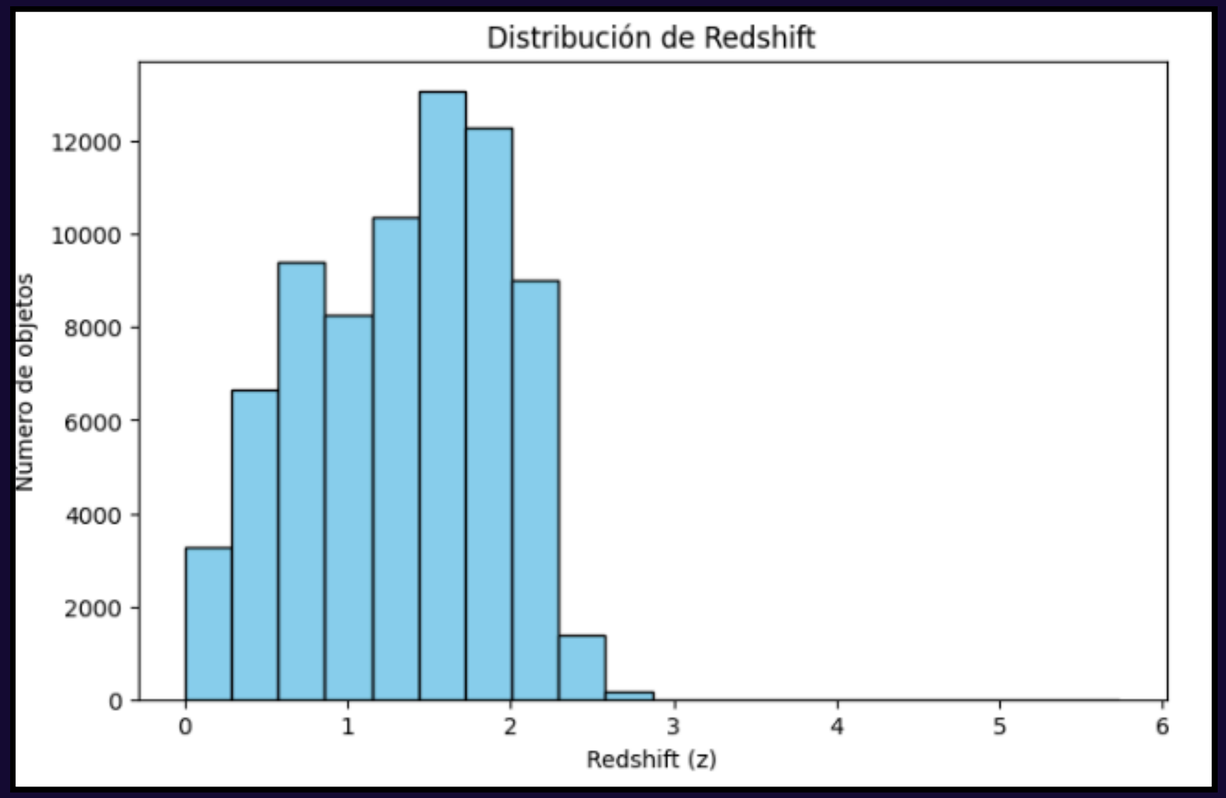


02

Filtro realizado:

- **Subclase válida (AGN, STARBURST, BROADLINE, etc.)**
- **Qf = 1 → 74k fuentes clasificadas de alta calidad**
- **Quedamos con cerca de 70mil datos**

EXPLORACION DE DATOS:

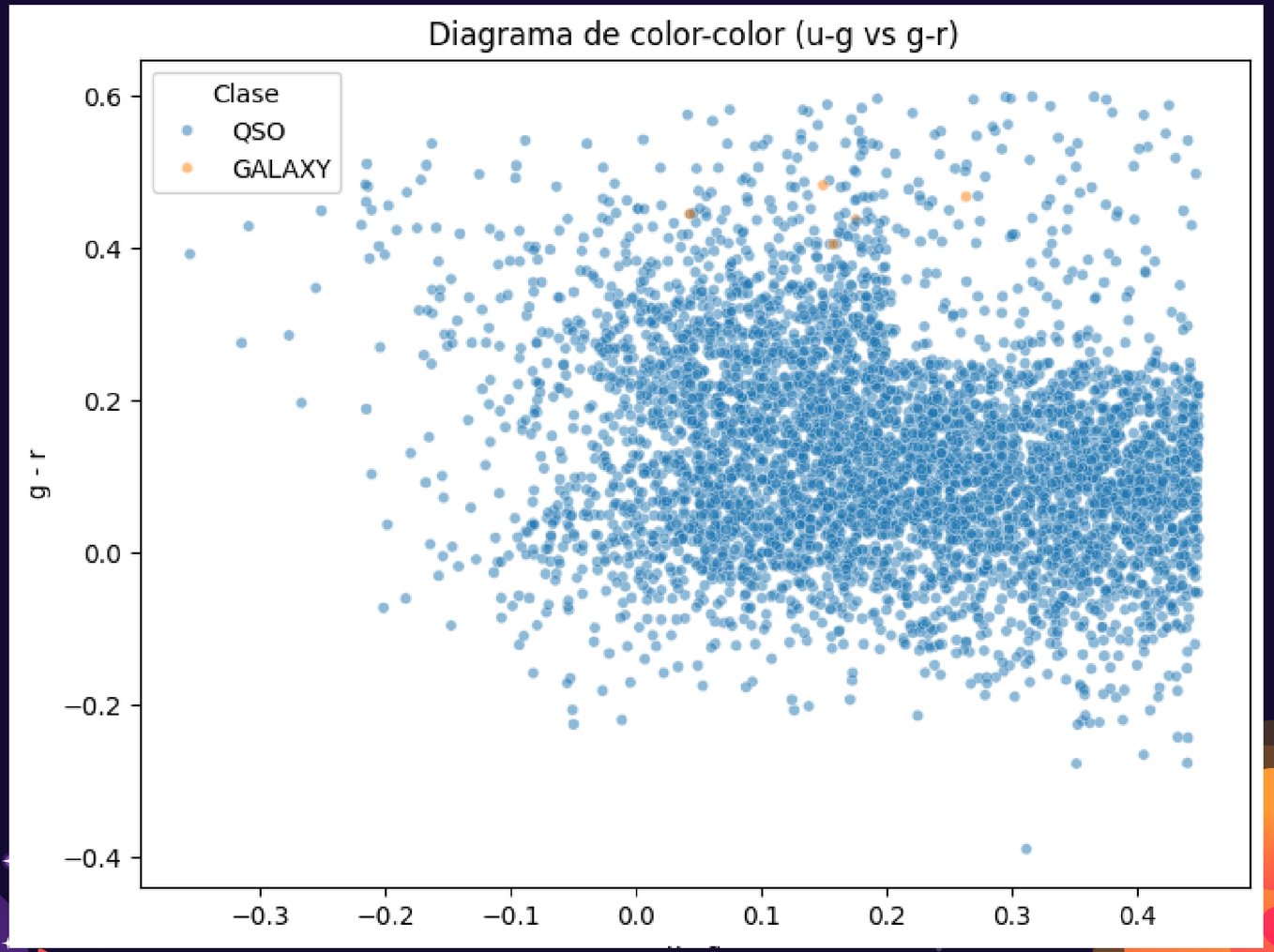


EXPLORACION DE DATOS:

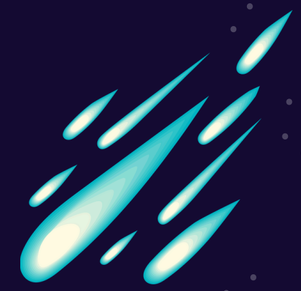
```
1  df["subclass"].value_counts()
```

```
subclass
BROADLINE                  71002
STARBURST_BROADLINE         2675
STARFORMING_BROADLINE         88
AGN_BROADLINE                 44
STARBURST                     36
AGN                           20
STARFORMING                   10
Name: count, dtype: int64
```

EXPLORACIÓN DE DATOS

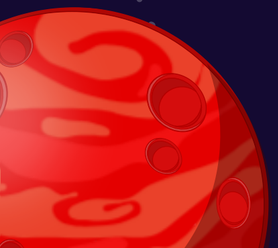


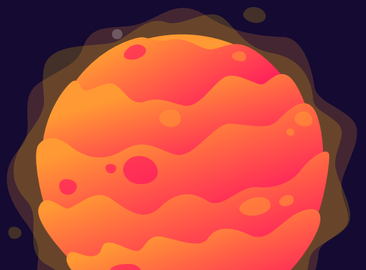
PREGUNTAS Y PROBLEMAS: OUTER SPACE



Problemas:

- **Desbalance de clases**
- **¿Over o under-sampling?**
- **Inyección de datos externos**





PROPUESTA METODOLÓGICA

01

MÉTODO DEL PAPER:

- MULTI LAYER PERCEPTRON WITH QUASI-NEWTON ALGORITHM (MLPQNA): 500K DATOS CONDENSADOS

02

PROPONEMOS MEJORAR LOS RESULTADOS:

- ACTUALIZAR DATOS (DR19 O AL MENOS SUPERIOR A DR10)
- DIFERENTES FORMAS DE CLASIFICACIÓN
- MANIPULACIÓN DE LOS DATOS
- OVER Y UNDERSAMPLING
- INYECCIÓN DE DATOS


MUCHAS
GRACIAS