Técnicas tradicionales y de Machine Learning para la identificación fotométrica de estrellas jóvenes (YSO)

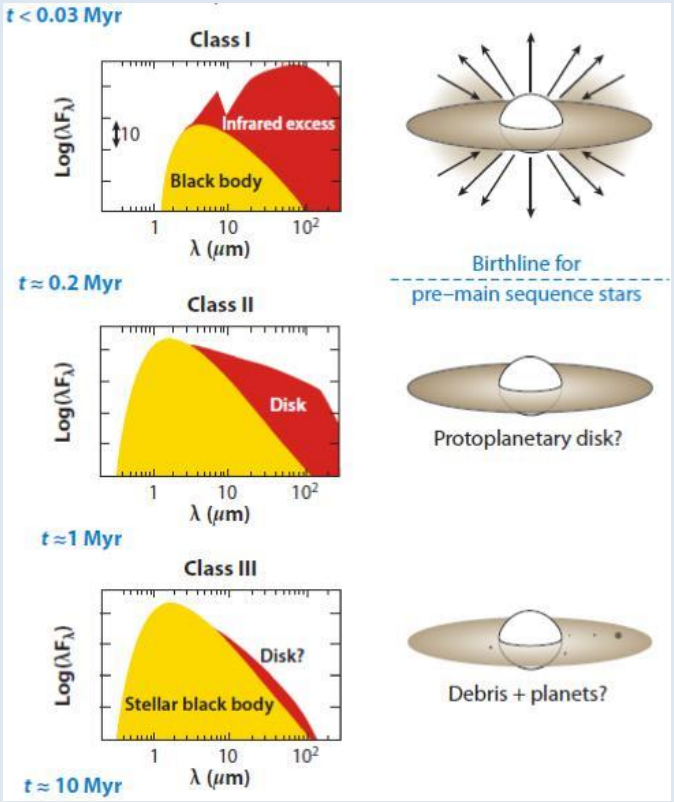
Nestor Sanchez(1), Elisa Nespoli(1), Marta González(1), Benjamín Arroquia(2)

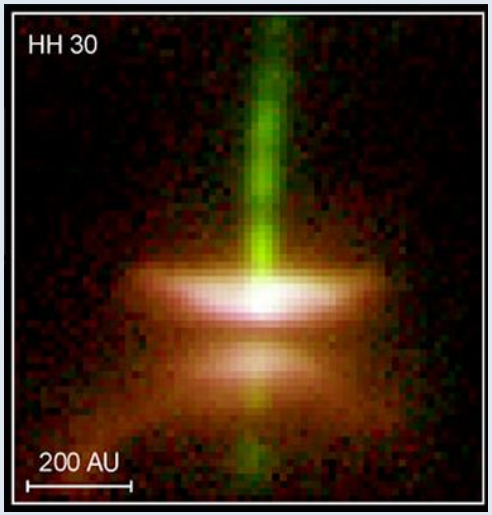
Esencia: Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

(1) ASGARD: AStronomy Group for Academic Research and Dissemination
(2) GRID: Grupo de investigación en Ciencia de Datos

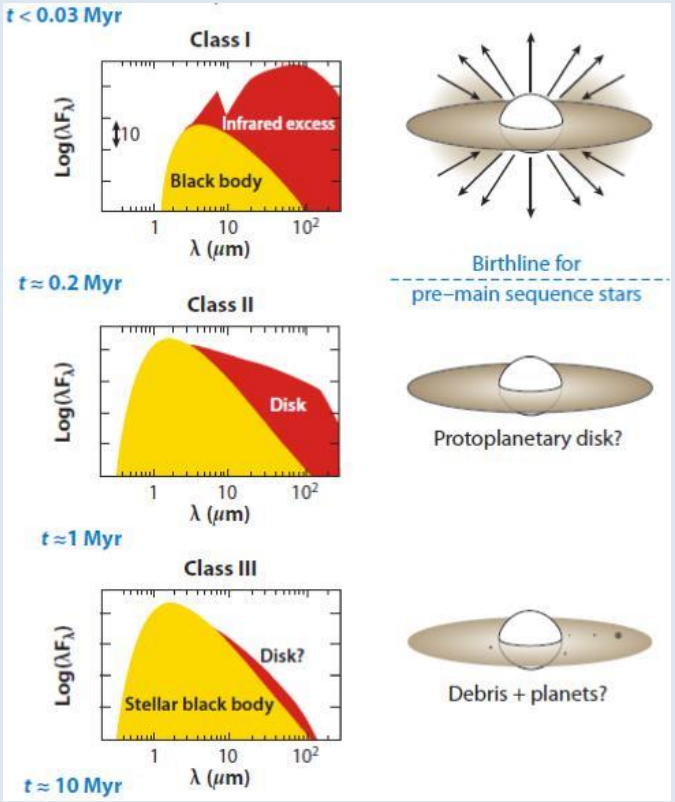
Universidad Internacional de Valencia (VIU)

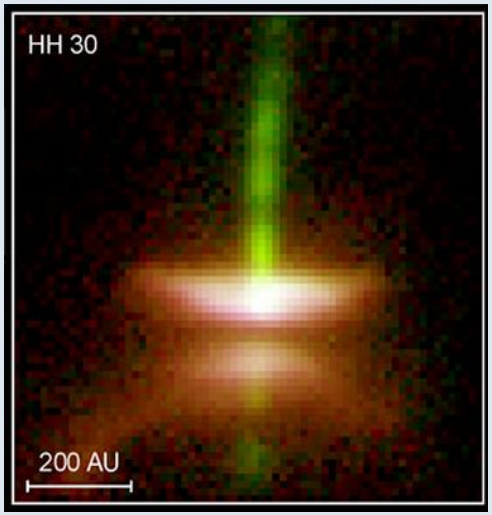
¿Qué es un YSO? ¿Cómo lo identifico?

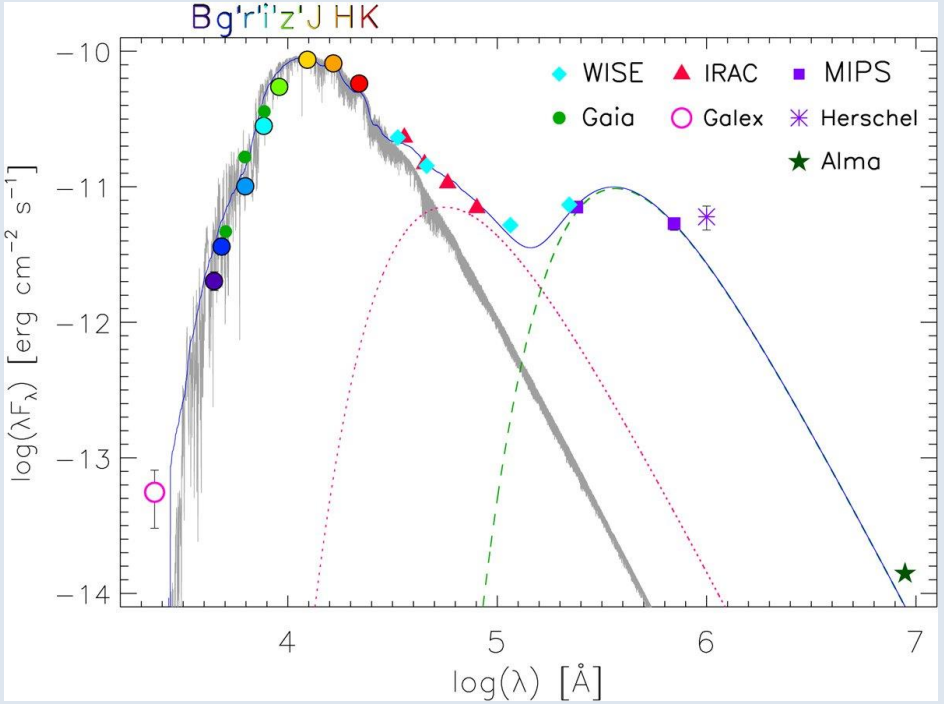




¿Qué es un YSO? ¿Cómo lo identifico?

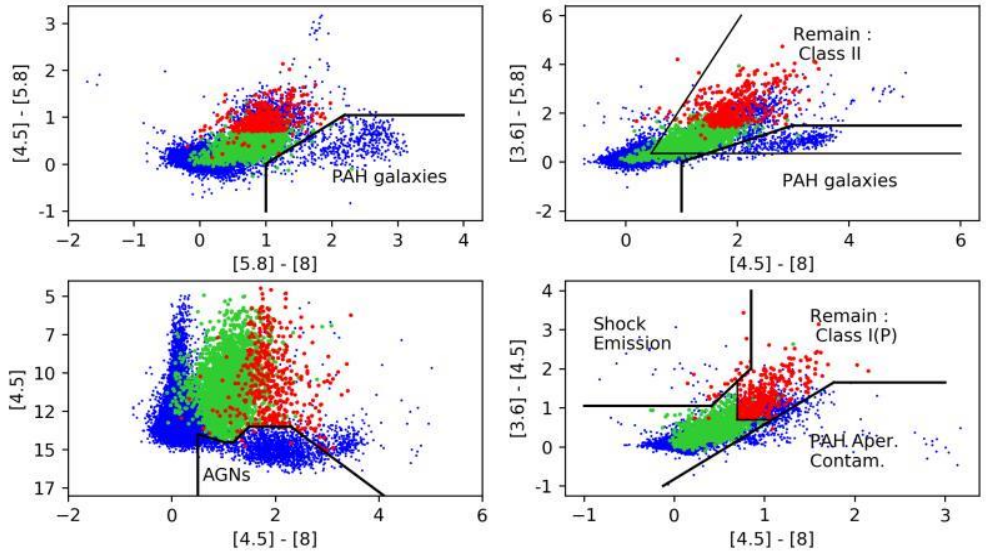






Frasca et al. (2020)

¿Qué es un YSO? ¿Cómo lo identifico?



Adapted from Gutermuth et al. (2009) method (G09) using IRAC at 3.6, 4.5, 5.8, 8.0 μm and MIPS at 24 μm. Class I in red and Class II in green, and Other in blue.

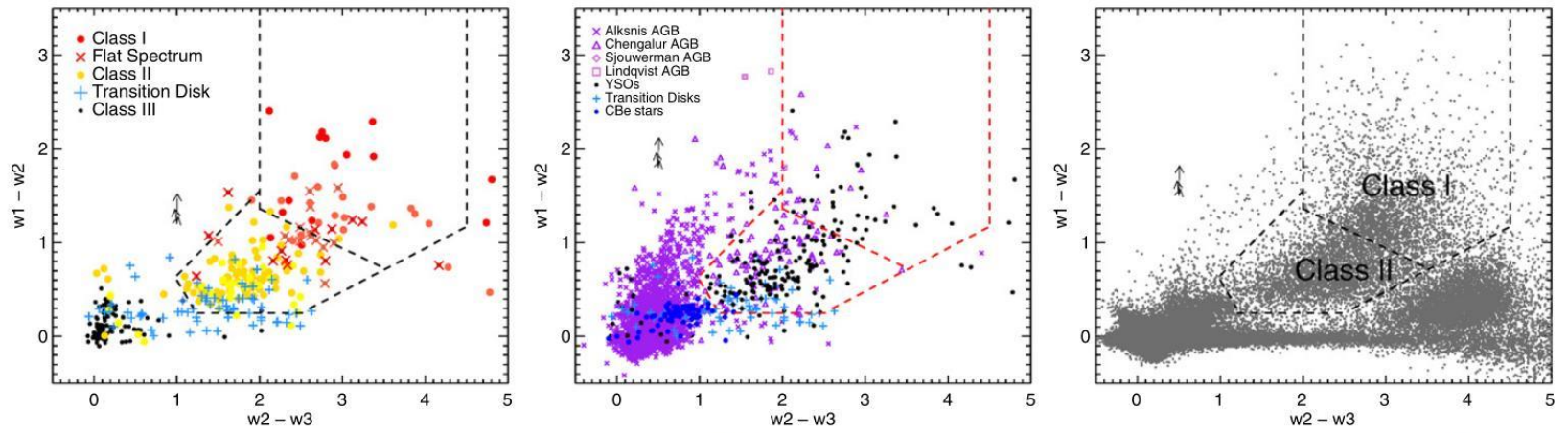
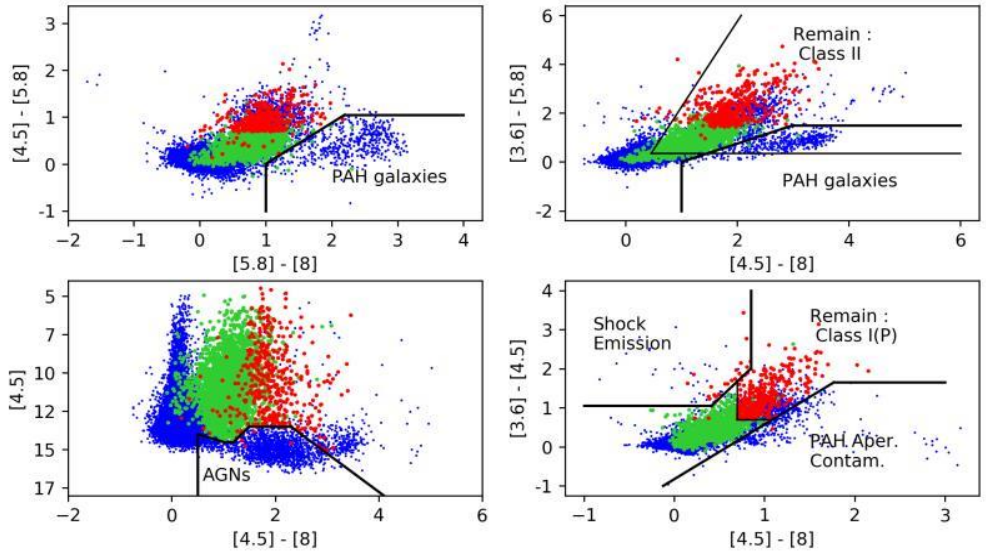


Figure 5. *WISE* color–color diagrams. Left: Taurus YSOs from Rebull et al. (2010) found in AllWISE and blue crosses, transition disks from Cieza et al. (2012). Center: AGB stars, CBe stars, YSOs, and transition disks, see the text. Right: outer Galaxy point sources from AllWISE found in the region defined by $105 < l < 153$, $|b| < 5$. All objects shown meet the $w1$, 2, and 3 requirements of Section 3.2. Dashed lines show our YSO class divisions. Arrows show extinction vectors of $A_{KS} = 0.4$, 0.8, and 2.

¿Qué es un YSO? ¿Cómo lo identifico?

Koenig & Leisawitz (2014)
---> KL14



Adapted from Gutermuth et al. (2009) method (G09) using IRAC at 3.6, 4.5, 5.8, 8.0 μm and MIPS at 24 μm. Class I in red and Class II in green, and Other in blue.

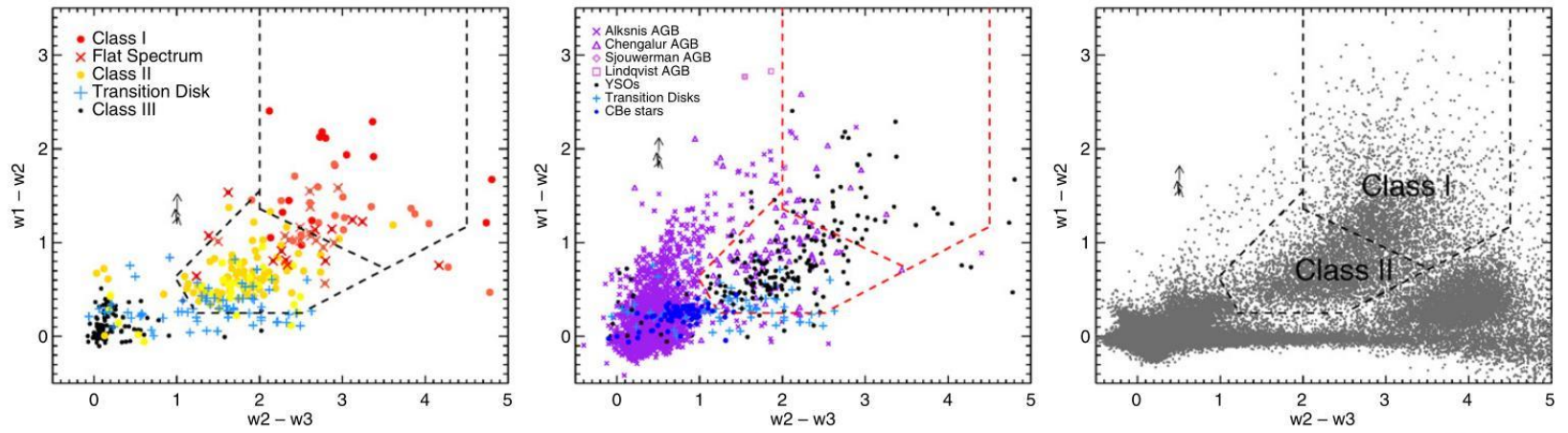


Figure 5. *WISE* color–color diagrams. Left: Taurus YSOs from Rebull et al. (2010) found in AllWISE and blue crosses, transition disks from Cieza et al. (2012). Center: AGB stars, CBe stars, YSOs, and transition disks, see the text. Right: outer Galaxy point sources from AllWISE found in the region defined by $105 < l < 153$, $|b| < 5$. All objects shown meet the $w1$, 2, and 3 requirements of Section 3.2. Dashed lines show our YSO class divisions. Arrows show extinction vectors of $A_{KS} = 0.4$, 0.8, and 2.

Datasets

- Dataset: YSO
- KYSO Catalogue: 11671 sources
- Crossmatch with WISE catalogue + non-null data in all bands: 8030 sources

- Dataset: non-YSO
- Crossmatch WISE-Simbad and non-null data: 9.4 million sources
- Cleaning (no relevant objects) and keeping -30<=b<=+30 deg: 680155 sources
  - MS stars (355212 sources)
  - Evolved stars (87103)
  - Normal galaxies (217420)
  - Active galaxies (20420)

Datasets

Final samples:

- Full imbalanced sample: 688185 sources (1.2% YSO)
- Training and testing sample: 16,040 sources (balanced classes/subclasses)

	AllWISE	RAJ2000	DEJ2000	W1mag	W2mag	W3mag	W4mag	Jmag	Hmag	Kmag	Class_KYSO	Class_RF	Class_KL14	grp_type	Name
1	J000000.08+320811.7	0,00036	32,13661	13,953	13,699	9,883	7,525	15,945	15,256	14,795	0	0	0	gal_normal	LEDA 1982072
2	J000000.31+481323.5	0,00132	48,22321	14,115	13,796	12,799	9,13	16,073	15,354	14,708	0	0	0	gal_normal	WISE J000000.31+481323.6
3	J000000.54+324211.0	0,00227	32,70308	13,889	13,644	12,5	9,005	15,674	14,916	14,366	0	0	0	gal_normal	SDSS J000000.55+324211.1
4	J000000.56+321233.2	0,00236	32,20925	12,519	12,467	12,583	8,777	14,403	13,684	13,295	0	0	0	gal_normal	LEDA 1985872
5	J000000.66+314836.8	0,00276	31,81023	15,177	15,188	12,349	9,118	16,505	15,723	15,113	0	0	0	gal_normal	SDSS J000000.67+314836.9
6	J000001.20+324236.9	0,00503	32,71026	14,505	14,323	10,944	8,752	16,269	15,149	15,08	0	0	0	gal_normal	LEDA 2009149
7	J000001.98+314603.7	0,00827	31,76772	13,844	13,741	12,384	8,994	15,391	14,868	14,318	0	0	0	gal_normal	SDSS J000002.00+314603.4
8	J000002.07+340728.5	0,00866	34,12459	14,921	14,709	12,152	8,878	16,606	15,546	15,064	0	0	0	gal_normal	SDSS J000002.07+340728.5
9	J000002.10+632747.0	0,00877	63,46306	10,706	10,796	10,66	9,071	11,88	11,107	10,861	0	0	0	star_evolved	2MASS J00000211+6327470
10	J000002.11+733946.2	0,00882	73,66285	8,924	8,942	8,946	9,203	9,779	9,261	9,094	0	0	0	star_ms	TYC 4306-1190-1
11	J000002.22+562535.9	0,00926	56,42666	11,281	11,332	11,167	9,439	11,982	11,433	11,398	0	0	0	star_evolved	2MASS J00000222+5625359
12	J000002.43+320249.2	0,01015	32,04702	13,928	13,754	11,026	8,819	15,872	15,011	14,504	0	0	0	gal_normal	2MASX J00000241+3202491
13	J000002.72+314800.9	0,01137	31,80025	14,215	14,123	12,449	9,103	15,709	15,005	14,705	0	0	0	gal_normal	SDSS J000002.73+314800.8
14	J000002.86+532411.0	0,01196	53,40307	13,714	13,708	12,064	9,029	15,383	14,515	14,178	0	0	0	gal_normal	2MASX J00000288+5324115
15	J000002.87+340618.8	0,012	34,10522	13,861	13,67	12,259	8,976	15,555	14,851	14,355	0	0	0	gal_normal	2MASX J00000287+3406192
16	J000003.02+442514.5	0,01261	44,4207	8,662	8,671	8,636	8,294	8,901	8,776	8,694	0	0	0	star_ms	BD+43 4599
17	J000003.40+605036.1	0,0142	60,84338	9,667	9,673	9,65	9,007	10,005	9,789	9,706	0	0	0	star_ms	TYC 4014-2752-1
18	J000003.46+334949.7	0,01442	33,8305	14,625	14,421	12,514	9,069	16,064	15,333	15,207	0	0	0	gal_normal	SDSS J000003.46+334949.6
19	J000003.52+314708.4	0,01469	31,78567	12,574	12,519	11,556	8,688	14,329	13,588	13,195	0	0	0	gal_normal	LEDA 1964345

Datasets

Final samples:

- Full imbalanced sample: 688185 sources (1.2% YSO)
- Training and testing sample: 16,040 sources (balanced classes/subclasses)

Features: magnitudes + colours (28 features).
Métodos: KL14 + Random Forest (RF)

	AllWISE	RAJ2000	DEJ2000	W1mag	W2mag	W3mag	W4mag	Jmag	Hmag	Kmag	Class_KYSO	Class_RF	Class_KL14	grp_type	Name
1	J000000.08+320811.7	0,00036	32,13661	13,953	13,699	9,883	7,525	15,945	15,256	14,795	0	0	0	gal_normal	LEDA 1982072
2	J000000.31+481323.5	0,00132	48,22321	14,115	13,796	12,799	9,13	16,073	15,354	14,708	0	0	0	gal_normal	WISE J000000.31+481323.6
3	J000000.54+324211.0	0,00227	32,70308	13,889	13,644	12,5	9,005	15,674	14,916	14,366	0	0	0	gal_normal	SDSS J000000.55+324211.1
4	J000000.56+321233.2	0,00236	32,20925	12,519	12,467	12,583	8,777	14,403	13,684	13,295	0	0	0	gal_normal	LEDA 1985872
5	J000000.66+314836.8	0,00276	31,81023	15,177	15,188	12,349	9,118	16,505	15,723	15,113	0	0	0	gal_normal	SDSS J000000.67+314836.9
6	J000001.20+324236.9	0,00503	32,71026	14,505	14,323	10,944	8,752	16,269	15,149	15,08	0	0	0	gal_normal	LEDA 2009149
7	J000001.98+314603.7	0,00827	31,76772	13,844	13,741	12,384	8,994	15,391	14,868	14,318	0	0	0	gal_normal	SDSS J000002.00+314603.4
8	J000002.07+340728.5	0,00866	34,12459	14,921	14,709	12,152	8,878	16,606	15,546	15,064	0	0	0	gal_normal	SDSS J000002.07+340728.5
9	J000002.10+632747.0	0,00877	63,46306	10,706	10,796	10,66	9,071	11,88	11,107	10,861	0	0	0	star_evolved	2MASS J00000211+6327470
10	J000002.11+733946.2	0,00882	73,66285	8,924	8,942	8,946	9,203	9,779	9,261	9,094	0	0	0	star_ms	TYC 4306-1190-1
11	J000002.22+562535.9	0,00926	56,42666	11,281	11,332	11,167	9,439	11,982	11,433	11,398	0	0	0	star_evolved	2MASS J00000222+5625359
12	J000002.43+320249.2	0,01015	32,04702	13,928	13,754	11,026	8,819	15,872	15,011	14,504	0	0	0	gal_normal	2MASX J00000241+3202491
13	J000002.72+314800.9	0,01137	31,80025	14,215	14,123	12,449	9,103	15,709	15,005	14,705	0	0	0	gal_normal	SDSS J000002.73+314800.8
14	J000002.86+532411.0	0,01196	53,40307	13,714	13,708	12,064	9,029	15,383	14,515	14,178	0	0	0	gal_normal	2MASX J00000288+5324115
15	J000002.87+340618.8	0,012	34,10522	13,861	13,67	12,259	8,976	15,555	14,851	14,355	0	0	0	gal_normal	2MASX J00000287+3406192
16	J000003.02+442514.5	0,01261	44,4207	8,662	8,671	8,636	8,294	8,901	8,776	8,694	0	0	0	star_ms	BD+43 4599
17	J000003.40+605036.1	0,0142	60,84338	9,667	9,673	9,65	9,007	10,005	9,789	9,706	0	0	0	star_ms	TYC 4014-2752-1
18	J000003.46+334949.7	0,01442	33,8305	14,625	14,421	12,514	9,069	16,064	15,333	15,207	0	0	0	gal_normal	SDSS J000003.46+334949.6
19	J000003.52+314708.4	0,01469	31,78567	12,574	12,519	11,556	8,688	14,329	13,588	13,195	0	0	0	gal_normal	LEDA 1964345

Resultados: KL14 & RF

KL14 versus RF:

- Similar YSO precisions (PRE).
- REC(KL14) << REC(RF).

PRE/REC---> YSO (err ~ 0.02)

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

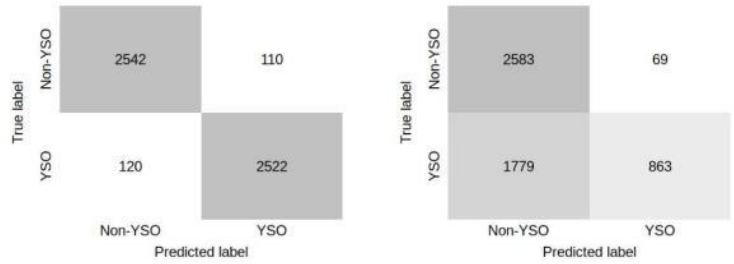


Fig. 1. Resulting confusion matrices of predicted and true classes for the execution of the RF using magnitudes and colours as features (left panel) and for the KL14 classification scheme (right panel).

Resultados: KL14 & RF

PRE/REC---> YSO (err ~ 0.02)

KL14 versus RF:

- Similar YSO precisions (PRE).
- REC(KL14) << REC(RF).

KL14 scheme is very restrictive because the authors prioritised purity over completeness and therefore RF is much more efficient at detecting YSOs within the sample.

If the goal is to produce a highly reliable catalogue of YSO for a given region while minimising contamination, using a relatively simple and direct method like KL14 may be as effective as training an RF.

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

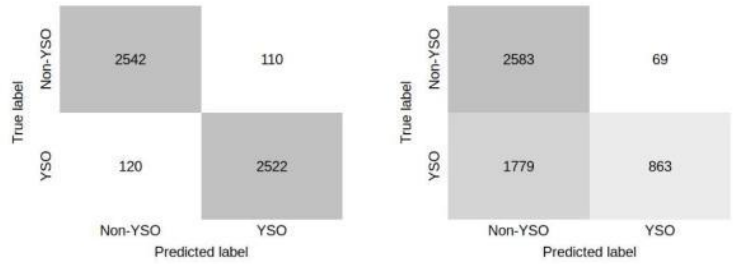


Fig. 1. Resulting confusion matrices of predicted and true classes for the execution of the RF using magnitudes and colours as features (left panel) and for the KL14 classification scheme (right panel).

Resultados: KL14 & RF

KL14 main colours:
- W1-W2
- W2-W3
- H-K

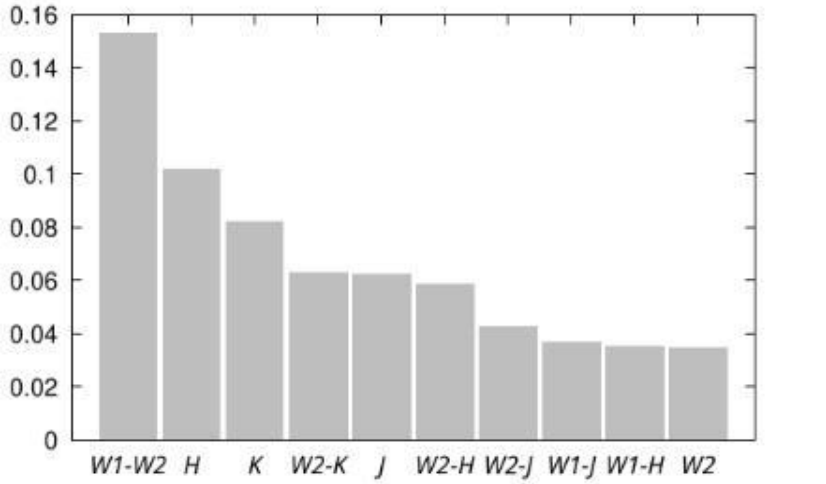


Fig. 2. Relative importance of the features for the same execution of the RF shown in Fig. 3.1. For clarity, only the ten most important features are presented.

Resultados: KL14 & RF

KL14 main colours:
- W1-W2
- W2-W3
- H-K

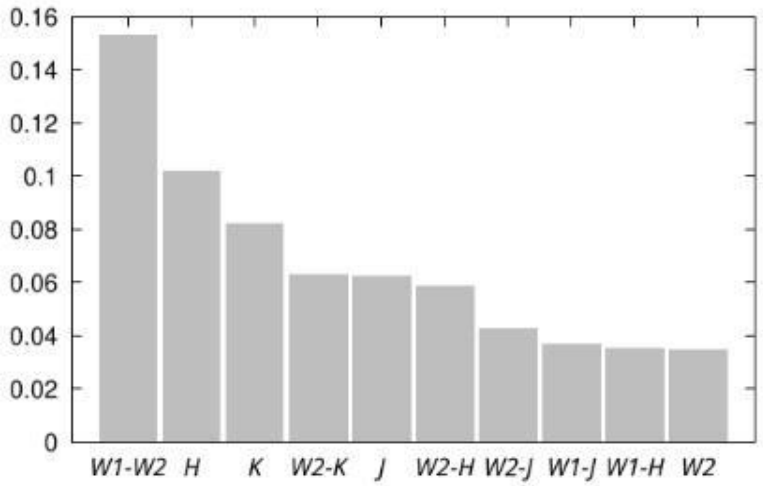


Fig. 2. Relative importance of the features for the same execution of the RF shown in Fig. 3.1. For clarity, only the ten most important features are presented.

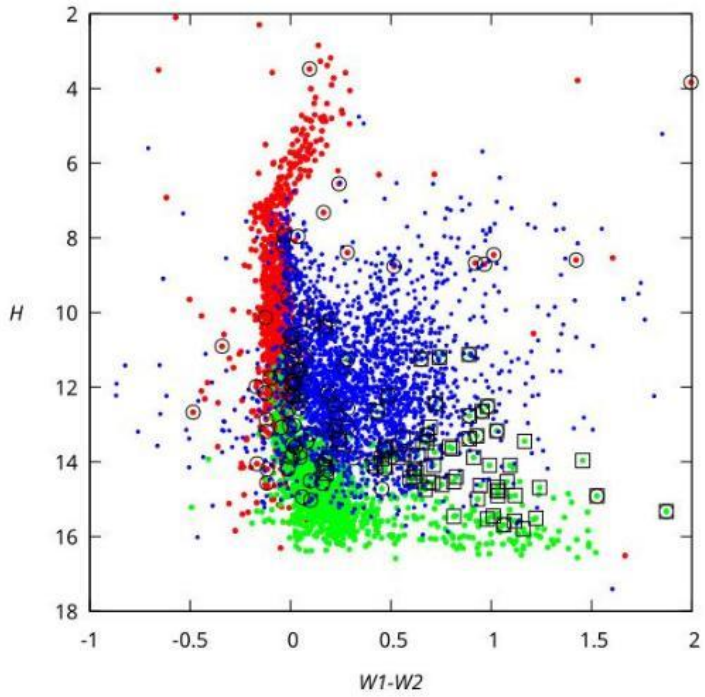


Fig. 3. Colour-magnitude diagram, H versus $W1-W2$, for the sources in the working sample: blue symbols refer to objects labelled as YSOs, red symbols are stars (main sequence and evolved stars), and green symbols are galaxies (both normal and actives). Sources erroneously classified as YSOs (false positives) are shown as open symbols: open circles represent misclassifications by the RF and open squares misclassifications by KL14.

Sources misclassified as YSO by KL14 tend to be concentrated near the region occupied by active galaxies: 94% of these false positives correspond to active galaxies.

Resultados: other ML algorithms

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

In this context of photometric classification of sources, having a high-quality training set is the key ingredient for any machine learning technique, whereas the classification model itself or the exact values of its free parameters do not seem to play an important role.

Resultados: full (imbalanced) sample

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

Optimised case al final

---> PRE disminuye significativamente mientras que REC permanece similar.

Resultados: full (imbalanced) sample

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

Optimised case al final

---> PRE disminuye significativamente mientras que REC permanece similar.

It is known that a critical point for classification algorithms, in particular RFs, is the transition from balanced training sets to highly imbalanced sets. This is because a low rate of false positives that is negligible in a balanced context results in a significant number of contaminants when the prior probability of the YSO class is several orders of magnitude lower than the rest of the sample.

Resultados: full (imbalanced) sample

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

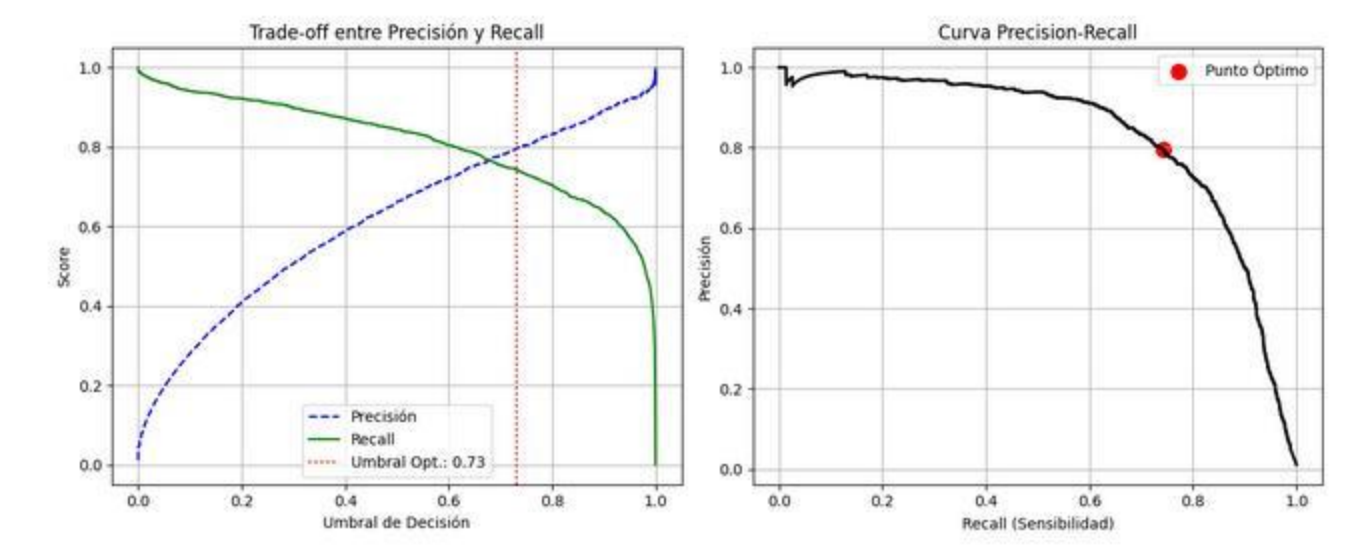
Optimised case al final

---> PRE disminuye significativamente mientras que REC permanece similar.

A combined RF+KL14 strategy yields a significantly higher precision (64%) than what would be achieved by applying RF or KL14 alone

---> It may be useful for extremely imbalanced samples where the true class proportions are unknown.

Resultados: optimised RF
---> work in progress



El umbral de predicción está 0.5 por defecto, donde según el valor de la predicción normalizado se clasifica cada objeto. En muestras desbalanceadas, se puede equilibrar el desempeño entre la precisión y recall del modelo cambiando el parámetro de probabilidad para la clase positiva como se muestra en la figura. Primero, se ha calculado los mejores parámetros del modelo de RF con una búsqueda aleatoria (RandomizedSearchCV) teniendo en cuenta que las clases están desbalanceadas, y se opta por optimizar la métrica de F1.

Resultados: optimised RF
---> work in progress

Table 1. Global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 metrics obtained for the classification of YSOs with different methods and data samples (see text for details).

Classifier	Sample	*ACC*	*PRE*	*REC*	*F*1
KL14	Test	0.65	0.93	0.33	0.48
RF	Test	0.96	0.96	0.95	0.96
KNN	Test	0.95	0.96	0.94	0.95
SVC	Test	0.95	0.96	0.95	0.95
GBoost	Test	0.95	0.96	0.95	0.95
KL14	Full sky	0.99	0.39	0.33	0.36
RF	Full sky	0.96	0.23	0.97	0.38
RF (optimised)	Full sky	0.00	0.00	0.00	0.00
RF+KL14	Full sky	0.99	0.64	0.33	0.43

PRE_opt = 0.80
REC_opt = 0.74
F1_opt = 0.77

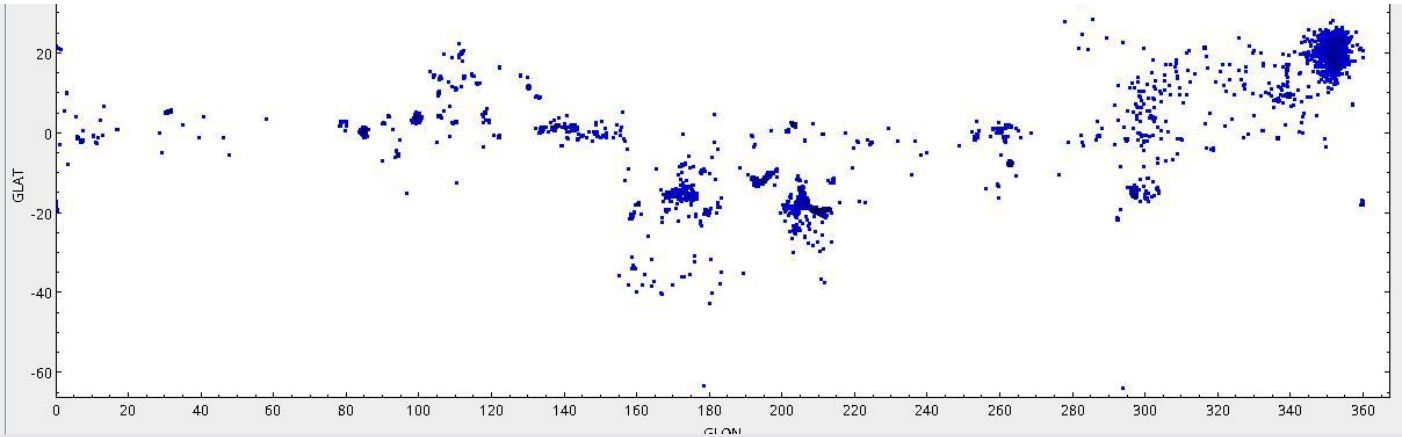
Resultados: individual star-forming regions ---> work in progress

Table 2. Defined regiones for comparing the resulting metrics. Also shown are the total number of sources (N_{dat}) and the actual number of YSOs and non-YSOs in each region.

Star forming Region	Galactic longitude (deg)	Galactic latitude (deg)	N_{dat}	N_{YSO}	$N_{non-YSO}$
Upper Scorpius / Ophichus	$343.36 \leq l \leq 360.28$	$+12.00 \leq b \leq +27.90$	6372	1317	5055
Taurus Dark Cloud	$166.97 \leq l \leq 183.54$	$-23.10 \leq b \leq -9.40$	6155	322	5833
Orion A / Orion Nebula Cluster	$207.68 \leq l \leq 213.22$	$-21.18 \leq b \leq -18.58$	2369	1782	587
Cepheus OB2	$98.81 \leq l \leq 100.80$	$+2.26 \leq b \leq +5.14$	977	357	620
NGC 7000 / IC 5070	$83.65 \leq l \leq 85.59$	$-0.56 \leq b \leq +1.54$	597	454	143
NGC 2264	$202.82 \leq l \leq 203.78$	$+1.25 \leq b \leq +2.40$	554	364	190

Table 3. Results obtained for global accuracy (*ACC*), precision (*PRE*), recall (*REC*), and *F*1 for different star-forming regions using the Koenig & Leisawitz (2014) scheme, Random Forest, and a combination of both classifiers (details in the text).

Star forming Region	Koenig & Leisawitz (2014)				Random Forest				Combined strategy			
	ACC	*PRE*	*REC*	*F*1	*ACC*	*PRE*	*REC*	*F*1	*ACC*	*PRE*	*REC*	*F*1
Upper Scorpius / Ophichus	0.84	0.84	0.26	0.40	0.88	0.63	0.99	0.77	0.84	0.90	0.26	0.40
Taurus Dark Cloud	0.95	0.56	0.35	0.43	0.87	0.29	1.00	0.45	0.96	0.73	0.35	0.47
Orion A / Orion Nebula Cluster	0.52	0.99	0.37	0.54	0.89	0.88	0.98	0.93	0.52	0.99	0.37	0.54
Cepheus OB2	0.74	0.99	0.30	0.46	0.87	0.76	0.96	0.85	0.74	0.99	0.30	0.46
NGC 7000 / IC 5070	0.62	1.00	0.51	0.67	0.95	0.97	0.98	0.97	0.62	1.00	0.50	0.67
NGC 2264	0.55	1.00	0.31	0.47	0.86	0.84	0.97	0.90	0.54	1.00	0.30	0.46



Conclusiones:

- Diferentes modelos de ML funcionan de manera similar (mismo datasets).
- KL14 y RF dan precisiones de YSOs similares (en datos balanceados).
- Con datos (muy) desbalanceados RF empeora en PRE pero la estrategia combinada (RF+KL14) mejora PRE de manera significativa.

Conclusiones:

- Diferentes modelos de ML funcionan de manera similar (mismo datasets).
- KL14 y RF dan precisiones de YSOs similares (en datos balanceados).
- Con datos (muy) desbalanceados RF empeora en PRE pero la estrategia combinada (RF+KL14) mejora PRE de manera significativa.

A tener en cuenta:

A common strategy is cost-sensitive learning in which different weights are assigned to classes during training by applying a higher penalty to minority class misclassification.

The issue with this strategy is that, for optimal performance, class weights should reflect prior probabilities of the actual population but, in practice, the weights are adjusted in proportion to class frequencies in the input data that in many cases differ from reality (for example: our working sample).