Análisis y diseño de procesos de minería de datos astrofísicos sobre catálogos fotométricos múltiple época.

Autor: Juan B Cabral
Directores: Pablo Granitto, Sebastián Gurovich.

FCEIA, Universidad Nacional de Rosario. IATE-OAC-CONICET.

Marzo 2019

Introducción: Astroestadística y astroinformática

La astronomía como ciencia de datos

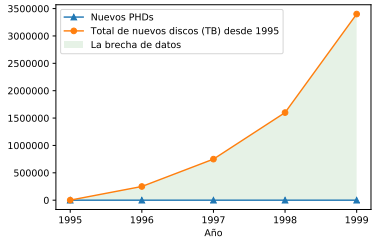
El campo de la astronomía se encuentra en una carrera para construir mejores telescopios hacen posible extraer cada vez más y mejores datos.

La astronomía como ciencia de datos

- **Paradigma científico clásico:**
  Hipótesis $\rightarrow$ *Diseño experimental* $\rightarrow$ *Recolección de datos*
- **Cuarto paradigma:** *Recolección de datos* $\rightarrow$ *Hipótesis* $\rightarrow$ *Diseño experimental*

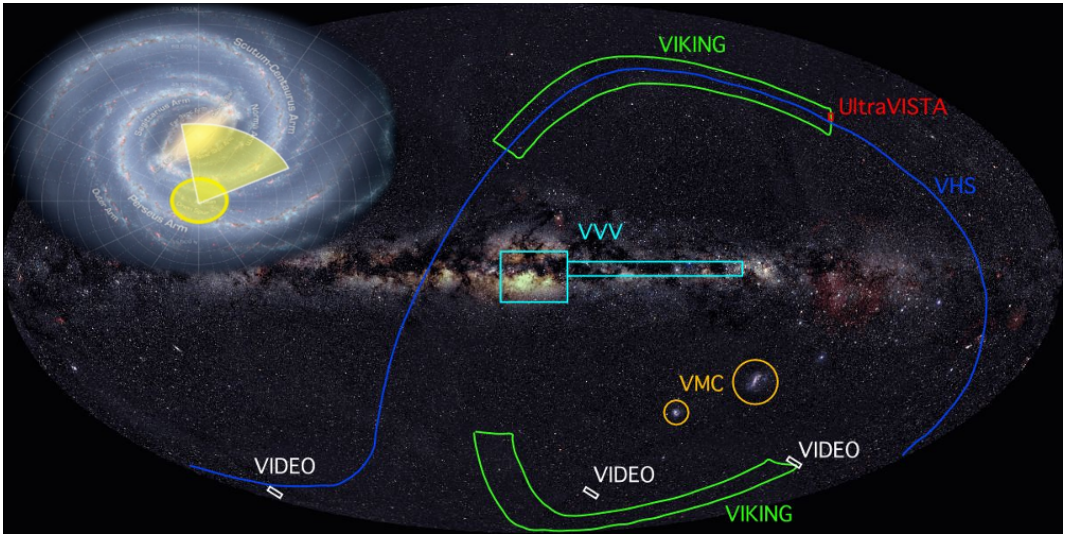
En otras palabras ahora el **éxito** radica en la capacidad de minar conocimiento desde los datos.

Así los astrónomos están siendo obligados a familiarizarse con términos como: aprendizaje automático, reconocimiento de patrones, minería de datos o reducción de dimensionalidad.
Todas estas técnicas han sido agrupadas en el campo de la **Astro-estadística** y la **Astro-informática** (*Borne, 2010*).



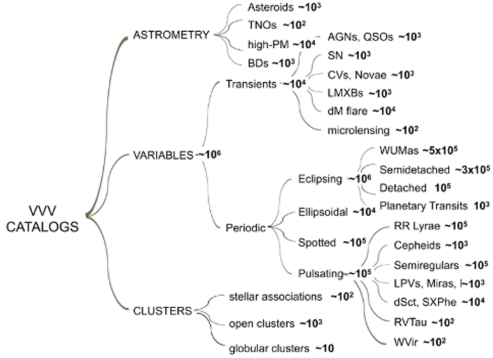
Crecimiento de datos vs. número de analistas.
Adaptado del trabajo de Grossman et al. (2013).

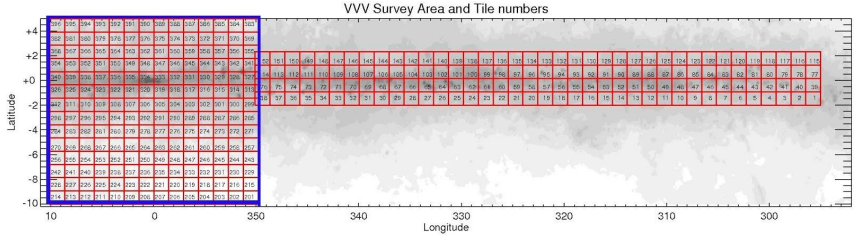
El relevamiento "*Vista Variables in the Via Lactea*"



El relevamiento "*Vista Variables in the Via Lactea*"

- Los datos del VVV se presentan en una unidad llamada "baldosa" (tile).
- Cada baldosa se compone de varias imágenes para diferentes frecuencias lumínicas en el infrarrojo cercano (cinco en total).
- Por cada imagen existe una base de datos numérica con las posiciones, magnitud y color de las fuentes de la imagen (catálogo fotométrico).





Esquema de la presentación

1. Aprendizaje automático.
2. Extracción de características.
3. Generación de catálogos de *RR-Lyrae*.
4. Detección de error instrumental.

Aprendizaje automático

¿Cómo aprenden las máquinas?

La rama que nos interesa de la inteligencia artificial postula que la forma de imitar la inteligencia humana por parte de una máquina se puede lograr a través del aprendizaje.

Formalmente Tom Mitchell en su libro "**Machine Learning**" define al aprendizaje automático como:

> *Se dice que un programa de computadora aprende de la experiencia E respecto a una tarea T y una medida de desempeño P, si el desempeño medido con P en una tarea T, mejora con la experiencia E.*

Importante

- El ML es un mecanismo inductivo de búsqueda de conocimiento (**Sesgo inductivo**).
- En la práctica, muchas veces el entendimiento es mucho más importante que la exactitud (**Navaja de Ockham**).

Pluralitas non est ponenda sine necessitate

¿Cómo aprenden las máquinas?

En una clasificación de dos clases los errores posibles son:

- Clasificar una clase positiva como negativa.
- Clasificar una clase negativa como positiva.

Matriz de confusión

	Clase Predicha	
Clase Real	**Positiva**	**Negativa**
Positiva	TP	FN
Negativa	FP	TN

Medición de errores

Precision

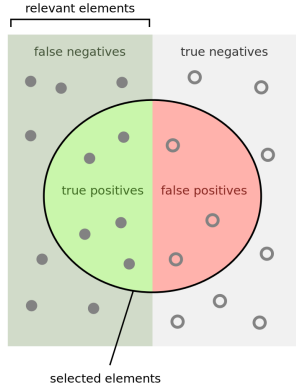
$$Precision = \frac{TP}{TP + FP}$$

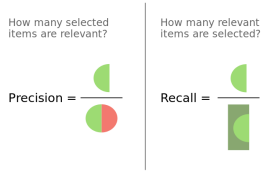
Recall

$$Recall = \frac{TP}{TP + FN}$$

Tasa de Falsos–Positivos (FPR)

$$FPR = \frac{FP}{FP + TN}$$

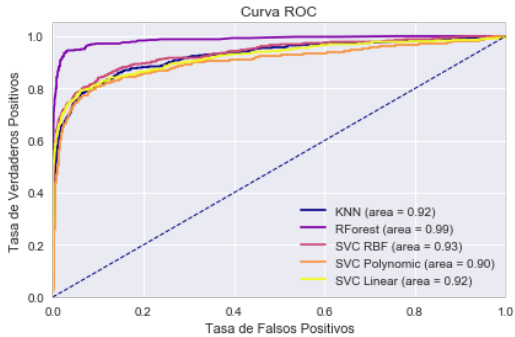




Medición de errores

Curva ROC

- Algunos clasificadores entregan un puntaje que representa con qué probabilidad cada ejemplo es miembro de una clase.
- Puede ser utilizado para producir muchos clasificadores variando el umbral de pertenencia de una clase a la otra.
- Si se grafican el *Recall* y el *FPR* correspondientes para cada variación, lo que se obtiene es una potente herramienta de evaluación llamada "Curva Característica Operativa del Receptor" (*ROC*)



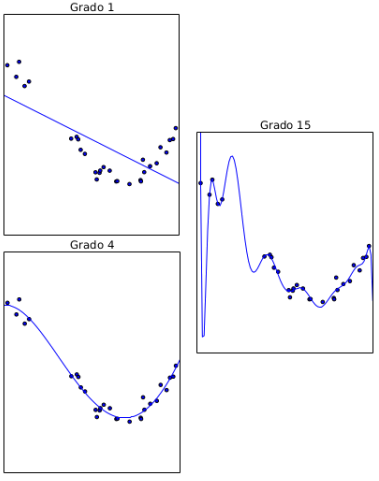
Consideraciones Importantes

Sobreajuste

Los patrones explicados en el entrenamiento disminuya la predicción de datos nunca vistos. Se percibe como una estimación optimista.
Se soluciona dividiendo el conjunto de entrenamiento en: **Entrenamiento, Prueba, Validación**.

Desbalance de clases

Sucede cuando la distribución de etiquetas de entrenamiento no es uniforme.
Soluciones: - Sobre-muestro aleatorio. - Sub-muestro aleatorio



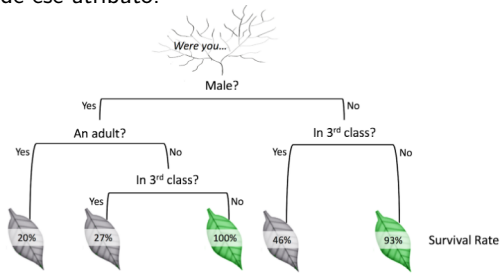
Selección de características

- Su utilidad es la de obtener un subconjunto de características más relevantes del conjunto completo de características, según una función de criterio determinada.
- Dentro de estos métodos se destaca el llamado "*Recursive Feature Elimination*" o "Eliminación recursiva de características" (*RFE*)
- Opera de la siguiente manera:
  1. Se calcula la importancia de todas las características utilizando un clasificador.
  2. Se eliminan las características con menor puntaje.
  3. Se repiten los pasos anteriores hasta que el número de características sea el deseado o se cumpla alguna condición de parada dada.
- Las condiciones de corte pueden ser varias, como por ejemplo que una métrica dada (*precision*, *recall* o *AUC*) disminuya o se les asigne una cota mínima o máxima.

Métodos de aprendizaje automático

Árboles de desición (DT)

Los **DT** clasifican instancias ordenándolas en un árbol partiendo desde la raíz hasta las hojas. Cada nodo especifica un atributo, y cada rama descendente de ese nodo corresponde a un valor de ese atributo.



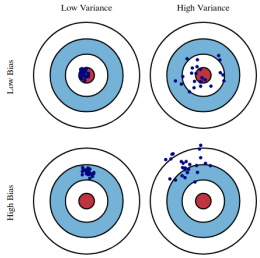
- Son fáciles de entrenar.
- Son fáciles de interpretar.
- Suele no alcanzar para capturar complejidad.

Random Forest (RF)

Utiliza un conjunto de árboles de decisión, entrenados sobre un sub-conjunto aleatorio de datos lo cual los hace especializarse y tener una muy buena predicción cuando trabajan en conjunto. Estos árboles votan la clase de un ejemplo dado.

Hiper-parámetros

- Cantidad de árboles.
- Cantidad de variables por predictor.



Métodos de aprendizaje automático

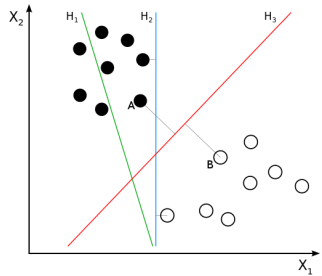
Máquinas de vectores de soporte (SVM)

El modelo representa a los ejemplos como puntos en un espacio vectorial, donde clases están divididas por una superficie con un claro margen.

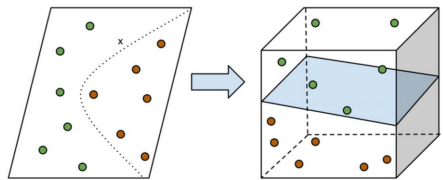
El ejemplo:

- H_1 no separa linealmente las clases.
- H_3 es el hiper-plano con mayor distancia a los puntos más cercanos y por lo tanto el que mejor generaliza.

En la práctica se utiliza el "margen débil", que soporta una tolerancia a puntos que quedan del lado incorrecto del hiper-plano.



Kernel

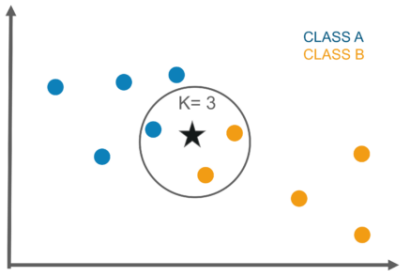


Métodos de aprendizaje automático

K-vecinos más cercanos (KNN)

La clase asignada a una observación es la que con más frecuencia se encuentre dentro de los K vecinos más próximos del conjunto de entrenamiento.

Es un tipo de algoritmo conocido como "basado en instancias" donde la función sólo es aproximada localmente y toda la computación del resultado se retrasa al momento de la clasificación.



Extracción de características

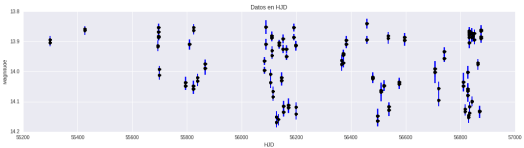
Datos

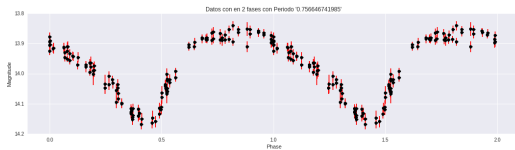
Las fuentes en nuestros catálogos están definidas como Curva de Luz. Esto es una colección de puntos con un época y una magnitud dada.

La época es el momento en el cual fue observada la fuente, y la magnitud es el brillo con el que fue observada.

Buscamos fuentes con algunas condiciones:

- Más de 30 observaciones (epocas).
- No saturadas (magnitud media > 12).
- No difusas (magnitud media < 16.5).



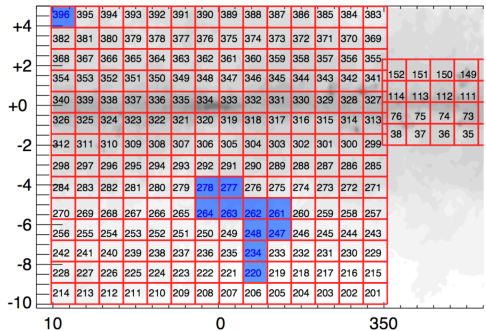


Generación del conjunto de datos

La clase positiva: Estrellas RR-Lyrae

Características: Estrellas muy brillanes, con variaciones de magnitud características con periodos muy bien definidos.
Por su estabilidad son utilizadas para la medición de distancias.

El conjunto de estrellas variables, se extrajo de los catálogos de OGLE-III, OGLE-IV y VizieR.



Nombre	Tamaño	Variables
b220	691713	149
b234	820452	1352
b247	939320	2601
b248	1081662	3485
b261	955119	4677
b262	1087417	5678
b263	920350	5503
b264	999947	4747
b277	979349	8686
b278	1018874	10179
b396	1075212	172
Total	10569415	47229

Características

La extracción de características que se llevó adelante en este trabajo fue llevada adelante en dos partes

1. **feATURE eXTRACTOR FOR tIME sERIES - *feets***

   Librería para extracción de características de series temporales estacionarias.

2. **Características basadas en color**

   Las estrellas variables se definen por su período, variación de magnitudes, morfología de su curva de luz; y su color.

   El problema:

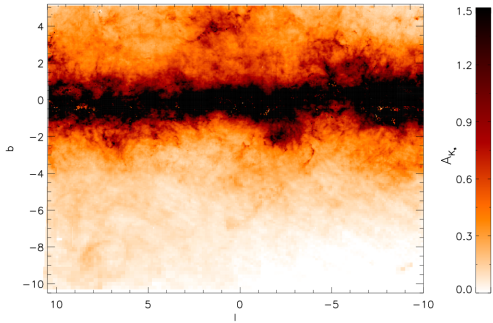
   - Solo hay observaciones multi-banda en la primer época.
   - Enrojecimiento

   Para eliminar este fenómeno se recurrieron a mapas de extinción.

Características

2. Características basadas en color

Utilizando el mapa de extinción brindado por el proyecto *"A VVV and 2MASS Bulge Extinction And Metallicity Calculator"* (Gonzales et al, 2012) Se calculo, pseudo-magnitudes, y pseudo-colores propuestas en el trabajo de Catelan et al, 2011; así como relaciones para convertir la `Amplitude` a las bandas J y H, además de calcular sus diferencias.



Características

56 características finales:

Características			
Amplitude	AmplitudeH	AmplitudeJ	AmplitudeJH
AmplitudeJK	Autocor_length	Beyond1Std	CAR_mean
CAR_sigma	CAR_tau	Con	Eta_e
FluxPercentileRatioMid20	FluxPercentileRatioMid35	FluxPercentileRatioMid50	FluxPercentileRatioMid65
FluxPercentileRatioMid80	Freq1_harmonics_amplitude_0	Freq1_harmonics_amplitude_1	Freq1_harmonics_amplitude_2
Freq1_harmonics_amplitude_3	Freq1_harmonics_rel_phase_0	Freq1_harmonics_rel_phase_1	Freq1_harmonics_rel_phase_2
Freq1_harmonics_rel_phase_3	LinearTrend	MaxSlope	Mean
Meanvariance	MedianAbsDev	MedianBRP	PairSlopeTrend
PercentAmplitude	PercentDifferenceFluxPercentile	PeriodLS	Period_fit
Psi_CS	Psi_eta	Q31	Rcs
Skew	SmallKurtosis	Std	c89_c3
c89_hk_color	c89_jh_color	c89_jk_color	c89_m2
c89_m4	cnt	n09_c3	
n09_hk_color	n09_jh_color	n09_jk_color	
n09_m2	n09_m4	ppmb	

Generación de catálogos de RR-Lyrae

Retomando las RR-Lyrae

- Las *RR-Lyrae* útiles para medir la distancia han adquirido un papel prominente en la determinación de la estructura de nuestra galaxia
- La búsqueda de este tipo de estrellas es uno de los objetivos principales del *VVV*.
- La identificación de este tipo de estrellas se obtuvo al realizar un *cross-matching* con los catálogos de los relevamientos *OGLE-III*, *OGLE-IV* y *VizieR*
- Para el comienzo de este experimento sólo se utilizó *OGLE-III*.

Detalles

- Cada Tile en total tiene $\sim 1 millon$ de fuentes de las cuales $\sim 500 mil$ son útiles y de estas ~ 300 son RR-Lyrae.
- De cada tile se tomaron muestras de estrellas desconocidas de tamaño 20 mil (Grande), 5 mil (Mediana) y 2500 (pequeña).

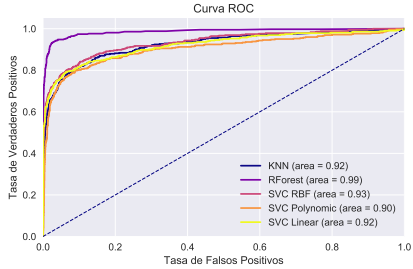
Preprocesado

Se realizó una limpieza de valores inválidos (nulos e infinito) quedándonos con 50 caracteísticas y con una distribución de clases representada en la tabla.

Tile	Pequeña	Mediana	Grande	RR-Lyrae
b261	2718	5212	20193	221
b262	2791	5288	20247	296
b263	2805	5302	20293	305
b264	2792	5292	20289	294
b278	2912	5406	20354	423

Comparación de clasificadores

	M. Pequeña			M. Mediana			M. Grande		
Clasificador	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**
SVM-L	0.86	0.68	0.92	0.91	0.55	0.92	0.91	0.43	0.91
SVM-P	0.88	0.53	0.91	0.85	0.43	0.88	0.86	0.40	0.88
SVM-R	0.91	0.66	0.93	0.91	0.53	0.93	0.95	0.43	0.90
RF	0.94	0.85	0.99	0.93	0.79	0.99	0.93	0.65	0.99
KNN	0.89	0.60	0.92	0.85	0.46	0.91	0.87	0.33	0.89



Es importante notar que para realizar un catálogo lo que se busca es tener una **alta detección de la clase positiva** (alto *recall* en *RR-Lyrae*), de modo que además se trata de conseguir la **menor cantidad posible de falsos positivos** (una *precision* alta en la clase *RR-Lyrae*)

Generalización

Surge la inquietud de analizar la bondad de predicción los datos de un *tile* del *VVV*, partiendo de otros *tiles*.

Experimento

1. Evaluar la detección en *b278*, al dividir los datos en 10 *K-Folds* estratificados.
2. Usar como conjunto de entrenamiento *b278* y probar con *b261*, *b262*, *b263* y *b264*, para evaluar si realmente los clasificadores son aplicables a otros *tiles*.
3. Usar como conjunto de entrenamiento *b261* y probar el modelo entrenado con *b278*, *b262*, *b263* y *b264*, para apreciar si los resultados dependen fuertemente del conjunto con que se entrena.
4. Usar como conjunto de entrenamiento *b278* $\cup$ *b261* y probar el modelo entrenado con *b262*, *b263* y *b264*, para evaluar si el resultado mejora al agrandar el conjunto de entrenamiento.
5. Usar como conjunto de entrenamiento *b278* $\cup$ *b261* $\cup$ *b264* y probar el modelo entrenado con *b262* y *b263*, para continuar el punto anterior.

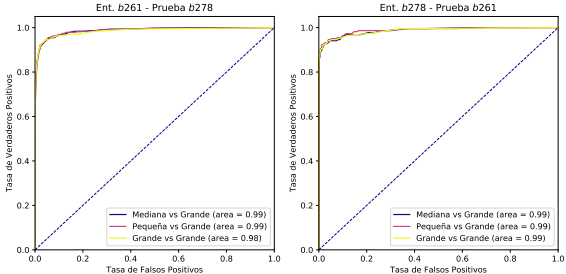
Generalización

		M. Pequeña			M. Mediana			M. Grande		
Ent.	**Prueba**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**	**Prec**	**Rec**	**AUC**
b278	K-fold	0.94	0.85	0.99	0.93	0.79	0.99	0.93	0.65	0.99
	b261	0.94	0.88	0.99	0.93	0.84	0.99	0.96	0.74	0.98
	b262	0.97	0.85	0.99	0.97	0.78	0.99	0.97	0.67	0.99
	b263	0.96	0.86	0.99	0.97	0.77	0.99	0.92	0.67	0.99
	b264	0.95	0.92	0.99	0.96	0.85	0.99	0.93	0.75	0.99
b261	b278	0.97	0.69	0.99	0.97	0.67	0.99	0.97	0.57	0.99
	b262	0.98	0.77	0.99	0.98	0.73	0.99	0.98	0.63	0.99
	b263	0.98	0.80	0.99	0.99	0.73	0.99	0.96	0.64	0.99
	b264	0.98	0.85	0.99	0.99	0.76	0.99	0.96	0.69	0.99
b261 $\cup$	b262	0.98	0.85	0.99	0.97	0.79	0.99	0.97	0.69	0.99
b278	b263	0.97	0.85	0.99	0.98	0.79	0.99	0.93	0.70	0.99
	b264	0.97	0.92	0.99	0.97	0.84	0.99	0.94	0.74	0.99
b261 $\cup$	b262	0.98	0.86	0.99	0.97	0.81	0.99	0.97	0.70	0.99
b278 $\cup$	b263	0.97	0.87	0.99	0.97	0.80	0.99	0.94	0.70	0.99
b264										

Balance de clases

Como diferirán los resultados si se altera el balance entre clases durante el entrenamiento.

Muestra	Entrenamiento	Prueba (Grande)	Prec	Rec	AUC
Pequeña	b278	b261	0.889	0.697	0.986
	b261	b278	0.735	0.878	0.988
Mediana	b278	b261	0.921	0.693	0.987
	b261	b278	0.845	0.842	0.987
Grande	b278	b261	0.969	0.582	0.984
	b261	b278	0.971	0.747	0.986



Balance de clases

Llegado este punto, la comparación se torna dificultosa.

- Todos los resultados anteriores no tienen el tamaño, ni el desbalance que tiene un *tile* completo.
- Puede sostenerse que la cantidad de FP va a crecer a medida que aumenta el conjunto de prueba.
- Los valores obtenidos de las muestras pueden ser utilizados para estimar los de un *tile* completo.
- Entonces es posible estimar el número de FP (FP^*) y la *Precision* (P^*) en un *tile* completo como:

$$FP^* = FP \times \frac{TR}{TM} \qquad P^* = \frac{TP}{TP + FP^*}$$

Agregado *OGLE-IV* y *VizieR*

			RR-Lyrae	
Tile	**T.Total**	**T.Útil**	**T. Catálogos**	**OGLE-III**
b220	691713	281150	65	0
b234	820452	297302	126	0
b247	939320	414497	192	0
b248	1081662	426369	222	0
b261	955119	575075	253	221
b262	1087417	591770	318	296
b263	920350	585661	319	305
b264	999947	614967	312	294
b277	979349	753146	434	0
b278	1018874	781612	441	423
b396	1075212	494646	15	0
Total	10569415	5746895	2697	1553
Promedio	960855.9	522445	245.18	139.91

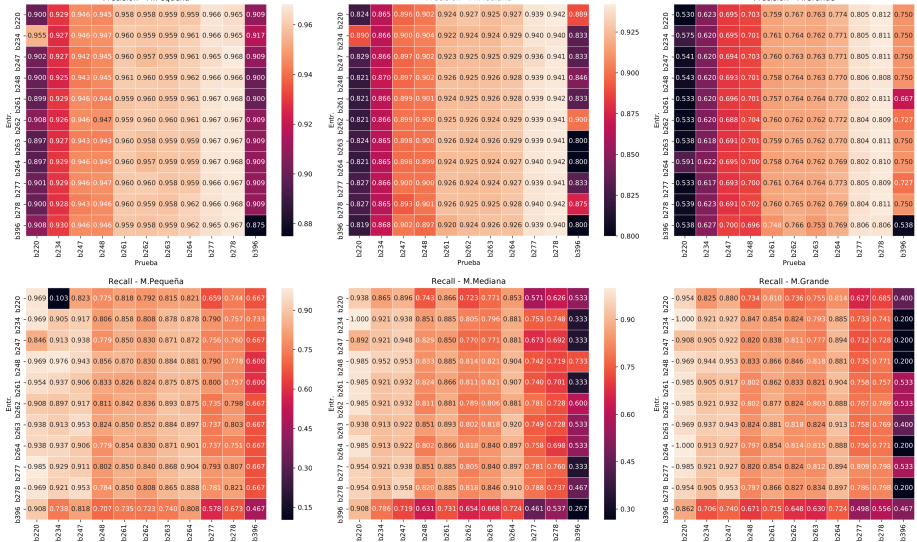
Predicción de nuevos *tiles*

Para comparar entonces los entrenamientos realizados con muestras de distinto tamaño sobre la base de datos ampliada se fijó un valor de $P^* = 0.1$

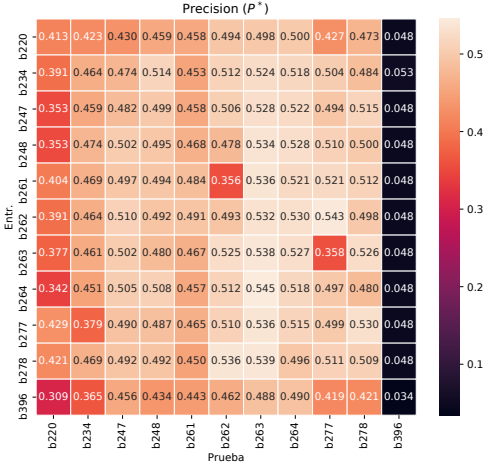
Así, el procedimiento para encontrar *precision* y *recall* de este punto de trabajo consiste en:

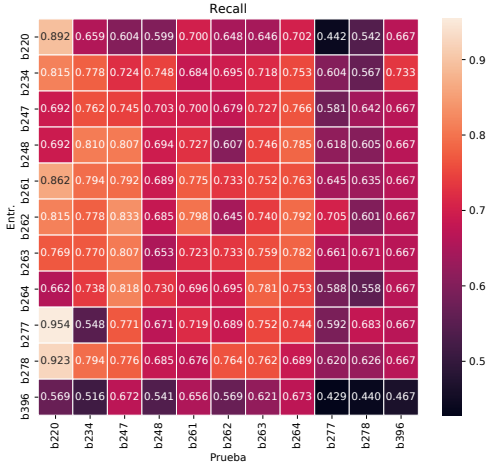
1. Entrenar un clasificador por cada tile en cada tamaño de muestra.
2. Por cada clasificador extraer las métricas de clasificación realizando pruebas en *k-folds* sobre sí mismo, y con cada uno de los demás *tiles*.
3. De cada una de estas métricas extraer todos los FP y TP de para todos los umbrales de decisión.
4. Calcular los FP^*.
5. Calcular la P^* para cada umbral de decisión.
6. Buscar el valor P^{**} el cual es el valor más cercano a 0.1 entre todos los P^*.
7. Extraer los valores de *precision* y *recall* correspondiente al valor de P^{**}.

Predicción de nuevos *tiles*



Predicción de nuevos *tiles*





Detección de error instrumental

Susceptibilidades al ruido observacional

- La información contenida en los catálogos está sujeta a ciertos errores experimentales que permean todos los datos recolectados.
- Por ende características derivadas son también proclives, en diferentes medidas, a estos errores.
- Es interesante preguntarse si se puede detectar qué características de las extraídas son más sensibles al ruido experimental.
- Es decir, interesa detectar cuáles son las característicasmás afectadas por efectos propios de la observación y que no reflejan propiedades del objeto observado.
- Un *tile* siempre es observado en conjunto, por lo cual sus fuentes comparten estados atmosféricos e instrumentales.

Hipótesis

Se propone un experimento que consiste en determinar si hay un subconjunto de las características medidas que permitan identificar si una fuente pertenece a un *tile*.

Si se puede identificar el origen de la fuente, se pone como hipótesis que lo que permite la identificación es ruido experimental, por lo que las características que más aporten al identificar el *tile* deberían ser las más ruidosas.

Muestreo

- No es necesario en este caso utilizar ningún tipo particular de fuente más allá de que éstas deben ser confiables en su fotometría.
- Se extrajo uniformemente 1000 fuentes no saturadas (magnitud media > 12) y no difusas (magnitud media < 16.5), de cada uno de los *tiles*.
- De estas selecciones se eliminaron fuentes con valores inválidos y se obtuvo la muestra final.

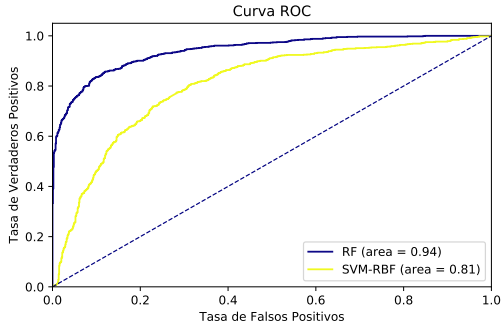
Tile	**Tamaño**
b220	1000
b234	1000
b247	1000
b248	998
b261	998
b262	996
b263	1000
b264	1000
b277	996
b278	994
b396	999

Selección de modelo

- Los modelos de clasificación evaluados fueron *SVM* y *RF*.
- Para determinar sus mejores hiper-parámetros se utilizó un análisis con *k-folds* sobre un conjunto de datos con las fuentes de los *tiles* *b*278 y *b*261.
- Con los hiper-parámetros determinados se evaluaron los modelos sobre los mismos set de datos utilizando *10 k-folds*.

Resultados

Modelo	Prec.	Recall	AUC
SVM-RBF	0.736	0.767	0.807
RF	0.888	0.839	0.941



Selección de características

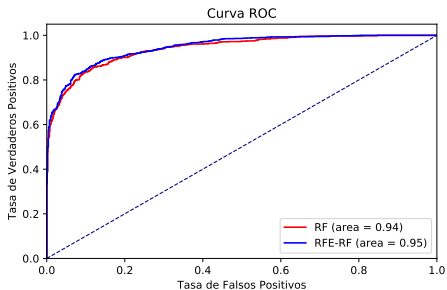
Se realizo un *RFE* eliminando de 1 característica por vez sin límite a la cantidad de features mínima seleccionada.

Características seleccionadas

Beyond1Std, Eta_e, Freq1_harmonics_amplitude_0, LinearTrend, MaxSlope, Mean, Meanvariance, Psi_eta, Rcs, Skew, **c89_m2**, **cnt**, **n09_c3**, **n09_hk_color**, **n09_m2**.

Rendimiento

Modelo	Prec	Recall	AUC
RF	0.888	0.839	0.941
RFE-RF	0.891	0.850	0.947

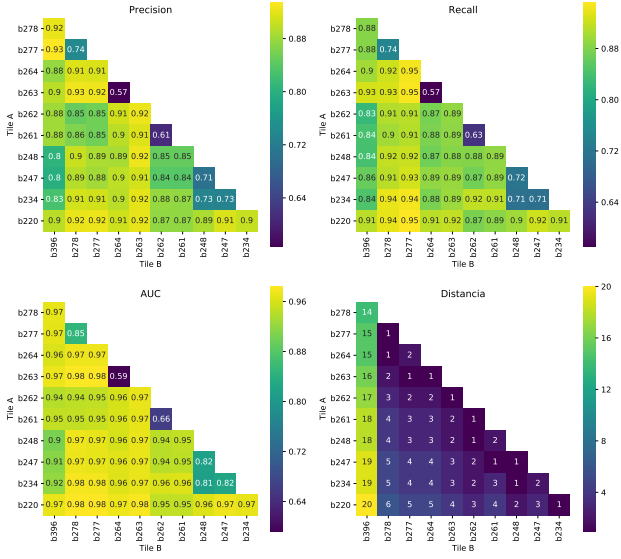


Generalización

Dado los resultados obtenidos en el caso de prueba del subconjunto de datos, es razonable preguntarse si estas mediciones se mantienen con todas las combinaciones de *tiles*. Entonces:

- Se procedió a crear 55 conjuntos de datos combinando de *2-en-2* todos los *tiles* disponibles en el trabajo
- Se entrenaron clasificadores y se extrajeron las mismas 3 métricas de clasificación (*Precision*, *Recall* y *AUC*) utilizando 10 *K-Folds*.

Generalización



Generalización

Distancia	Frecuencia	Precision	Recall	AUC
1.0	11	0.837	0.882	0.944
2.0	11	0.889	0.888	0.962
3.0	10	0.887	0.901	0.962
4.0	8	0.897	0.908	0.967
5.0	4	0.908	0.926	0.977

Conclusiones y trabajo a futuro

Sobre las características extraidas.

- Se presentó el relevamiento astronómico *VVV* y sus particularidades observacionales; haciendo hincapié en la forma en la cual se utiliza el telescopio *VISTA* y como utilizar sus curvas de luz de las fuentes observadas para luego extraer características útiles en métodos automáticos.
- **A futuro:**
  - Queda pendiente el diseño de características extraídas directamente de la morfología de la curva puesta en fase.

Sobre el software utilizado

- ***feets***, una herramienta extracción de características de curvas de luz, siendo esta una extensión de una herramienta existente, preparada para procesar con un gran volumen de datos en menos tiempo.
- **A futuro:**
  - Agregar herramientas para el análisis interactivo de las características extraídas.
  - Caracteristicas referidas a la morfología.

Sobre el software utilizado

- ***Corral***, una herramienta alternativa para el procesamiento de un gran volumen de datos, que brinda indicadores de calidad del pipile creado.
- **A futuro:** Extender o reescribir su funcionalidad sobre una plataforma real de cómputo distribuido como puede ser *Apache Spark* o *Dask*.

- **Carpyncho**, pipeline construido sobre *Corral* que brinda un ambiente confiable para la extracción de características del *VVV*.
- **A futuro:** Publicar el conjunto completo de características extraídas, y generar una herramienta web para su administración.

Sobre la generación de catálogos

- Se determinó el modo más eficiente para generar catálogos de estrellas variables tipo *RR-Lyrae* de forma automática de los *tiles* del *VVV*, teniendo especial cuidado en el balance de los datos utilizados para entrenamiento y prueba,
- También se estimó como será el funcionamiento de clasificadores entrenados con muestras de menor en tamaño que las que serán utilizadas para generar los catálogos.
- **A futuro:**
  - Se generarán y publicarán catálogos de estrellas *RR-Lyrae* candidatas para todos los *tiles* comprendidos en este estudio, haciendo principal énfasis en *b*396, por ser el menos estudiado.

Sobre la detección de error instrumental

- Se presentó una forma de evaluar la dependencia de las características extraídas de los catálogos, con la metodología de observación y reducción de datos.
- Así el procedimiento consiste en intentar determinar cuáles son las características más importantes de "**dónde**" se está observando obviando el "**qué**" se está observando.
- Se logró extraer un conjunto de 15 características, que poseen un *precision* promedio del 86% para determinar cuál es el *tile* al que pertenece una fuente, y que podrían estar afectadas en mayor medida por el ruido, el cual se incrementa cuando los *tiles* son más lejanos entre si.
- **A futuro:**
  - Queda pendiente el análisis de estas características, desde el punto de vista astronómico-observacional, así como la comparación de las medidas de calidad con más *tiles* para lograr una mejor apreciación de las mismas.

Gracias

¿Preguntas?