# Detección de exoplanetas a través de técnicas de aprendizaje automático

**Margarita Bugueño Pérez*** **and Francisco Mena Toro***
*Departamento de Informática, Universidad Técnica Federico Santa María

**ABSTRACT** Descubiertos los primeros planetas fuera de nuestro Sistema Solar en 1992 (en torno a un púlsar) y en 1995 (en torno a una estrella), la curiosidad sobre los diversos exoplanetas ha experimentado un considerable aumento y, con ello, el desarrollo de diferentes métodos para detectarlos. El procesamiento y análisis de curvas de luz emitidas por los cuerpos celestes es una de las técnicas de procesamiento más populares hasta el día de hoy. Por ello, haciendo uso del contenido y la funcionalidad de *Exoplanet Archive* de la NASA, una base de datos y conjunto de herramientas para apoyar a los astrónomos en la comunidad de exoplanetas, se propone la aplicación de diversas técnicas de aprendizaje automático enfocadas en la manipulación de datos brutos (secuencia de mediciones de la intensidad de la luz) con el objetivo de imitar el análisis y trabajo especializado de los astrónomos, permitiendo agilizar y simplificar el proceso de detección de exoplanetas. Para esto se utilizaron técnicas de extracción de características especializadas en secuencias, para así obtener una representación en la que un modelo detectase automáticamente aquellos elementos cruciales para lograr llevar a cabo la tarea encomendada.

**KEYWORDS** Visualización; curva de luz; exoplaneta; aprendizaje automático

*"Somewhere, something incredible is waiting to be known..."*

*Carl Sagan*

## I. Introducción

Un planeta extrasolar o exoplaneta es un planeta que orbita una estrella (o remanente de una estrella) fuera del Sistema Solar. La primera detección de un planeta extrasolar fue en 1992, cuando los astrofísicos Aleksander Wolszczan y Dale Frail descubrieron tres planetas extrasolares orbitando el púlsar PDR1257+12. Mayor and Queloz (1995) del Observatorio de Ginebra, detectaron el primer planeta extrasolar alrededor de la estrella 51 Pegasi. El planeta, denominado 51 Pegasi b, tiene alrededor de la mitad de la masa de Júpiter, gira a toda velocidad alrededor de su estrella en tan sólo cuatro días terrestres, y se encuentra ocho veces más cerca de ella que Mercurio del Sol.
Desde 1995 este campo se ha convertido en un área de investigación muy dinámica en la cual los astrónomos han encontrado, a la fecha, más de 3500 [1] exoplanetas valiéndose de diferentes técnicas. Sin embargo, el descubrimiento de planetas extrasolares no se trata de una tarea fácil y se requiere de mucho tiempo de investigación y análisis para lograr concluir sobre la posibilidad de tratarse o no de un exoplaneta. Lamentablemente los observatorios actuales, ya sean terrestres o espaciales, en la mayoría de los casos deben aplicar métodos indirectos para determinar la existencia de un posible exoplaneta, por ejemplo estudiando el leve movimiento que se produce en la estrella por la órbita del planeta alrededor de ésta (detección de velocidad radial) o la variación que se produce en la luz emitida por la estrella cuando el planeta pasa por delante de ésta (detección por curvas de luz o tránsito).
La gran cantidad de datos que generan los observatorios en la actualidad indica la necesidad del uso de técnicas automatizadas para los procesos que hasta el día de hoy deben llevarse a cabo extensamente de manera manual. Sin embargo, los grandes avances en computación científica y análisis de datos han permitido la aplicación de técnicas y algoritmos a problemas computacionalmente costosos, pudiendo hacer predicciones de factores cruciales, en este caso la validación en el caso de que una curva de luz corresponde o no a un planeta orbitando alrededor de una estrella. Cabe destacar que la detección mediante curvas de luz debiera ser la más precisa, puesto que el 70% de los planetas han sido detectados usando dicho método [2].
Para automatizar el proceso de detección de exoplanetas medi-

Introducción a la Informática y la Astronomía ILI-136
Documento generado: 7 de Diciembre, 2017
[1] exoplanets.nasa.gov

[2] https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

ante el análisis de las curvas de luz, el presente trabajo se enfoca en la aplicación de métodos de aprendizaje automático basado en la manipulación de las curvas de luz recolectadas por el observatorio Kepler (actualizado hasta el mes de septiembre del año 2017). Cabe destacar que con el propósito de lograr buenos resultados, se puso especial cuidado en lo que se conoce como *feature engineering*, el cual consta en el tratamiento de los datos brutos, su preprocesamiento y generación de nuevas características relevantes así como la selección de aquellos atributos más significativos para la tarea del algoritmo, pudiendo realizar una buena detección.
Para generar características a partir de la curva de luz se utilizó la librería FATS para Python, con la cual se pueden calcular una gran cantidad de métricas estadísticas enfocadas en secuencias de tiempo. Además de esto se trabajó con redes neuronales recurrentes, PCA e ICA, para generar nuevas características de forma no supervisada, es decir, reconocer patrones intrínsecos de todas las curvas de luz, independiente el tipo de problema al que se le enfrente (detección de exoplanetas, detección de anomalías, entre otros).
El documento se estructura de la siguiente manera: A continuación (sección II) se describe en mayor detalle que es un exoplaneta y la importancia de su descubrimiento. Luego (sección III) se presenta en extenso detalle las diversas técnicas de detección de las cuales se valen los astrónomos para la determinación de la existencia de un planeta extrasolar. Se presentan los datos con los cuales se trabajó (sección IV) así como toda la base teórica del proceder experimental (sección V), presentando los resultados y experimentaciones (sección VI). Finalmente se concluye y detallan los futuros procederes de la investigación (sección VII).

## II. Planetas Extrasolares, qué y por qué

Un planeta es un objeto que orbita alrededor de una estrella y que es lo suficientemente masivo como para despejar de polvo y otros desechos el disco protoplanetario del cual nació. Esto los diferencia de los planetas enanos (como Plutón), los cuales no tienen masa suficiente para limpiar el área de dicho disco.
Los planetas extrasolares se convirtieron en objeto de investigación científica a mediados del siglo XIX y aunque hubo algunas afirmaciones sin fundamento en cuanto a su descubrimiento, no se sabía cuán comunes eran, cuán similares eran a los planetas del Sistema Solar, o incluso cuán típica era la composición de nuestro Sistema Solar en comparación con sistemas planetarios alrededor de otras estrellas. Sin embargo, el estudio de éstos puede entregar respuestas a tales interrogantes.
Los discos protoplanetarios, donde se forman los planetas, son regiones de gas y polvo orbitando alrededor de estrellas muy jóvenes. Teorías actuales sugieren que las partículas de polvo empiezan a colapsar por la gravedad formando granos cada vez mayores. Si estos discos sobreviven a la radiación estelar y cometas o meteoritos, la materia continúa compactándose dando paso a un planetoide (objetos mayores que los meteoritos y cometas, pero menores que un planeta). Debido a las limitaciones en los métodos de detección, la mayoría de los planetas descubiertos han sido bastante grandes mientras que sólo unos pocos pueden ser comparados con las dimensiones de la Tierra. El actual foco del estudio de exoplanetas se centra en el desarrollo de teorías y conocimiento sobre la formación planetaria, cómo se formó el Sistema Solar y su futuro así como la habitabilidad de tales planetas. Es decir, la posible existencia de planetas similares a la Tierra y, de ser así, conocer cuáles son las condiciones necesarias para sustentar alguna forma de vida.

| | |
|---|---|
| Astrometría | 1 |
| Detección visual directa | 44 |
| Velocidad Radial | 658 |
| Tránsito | 2771 |
| Microlentes gravitacionales | 53 |
| Pulsos de radio de un púlsar | 5 |

**Tabla 1** Número de exoplanetas confirmados según método de detección

Por lo que el descubrir más exoplanetas es crucial para poder realizar este análisis.

## III. Técnicas de detección

Los planetas son fuentes de luz reflejada muy tenue en comparación con sus estrellas madre. Por ello, es sumamente difícil detectar este tipo de luz lográndose, a la fecha, haber fotografiado sólo un par de decenas de exoplanetas. Por el momento, la gran mayoría de los exoplanetas conocidos han sido detectados a través de métodos indirectos, mas detalle en la Tabla 1.
Algunos mecanismos de detección son:

1. **Velocidades radiales:** El planeta, al orbitar su estrella madre, ejerce una fuerza gravitacional sobre ésta tal que la estrella gira sobre el centro de masa del sistema. Las oscilaciones de la estrella pueden detectarse a través de pequeños cambios en las líneas espectrales de ella.

2. **Astrometría:** Dado que la estrella gira sobre el centro de masa del sistema, se registran sus variaciones de posición.

3. ***Tránsitos:** Observación fotométrica de la estrella y detección de variaciones en la intensidad de su luz cuando un planeta orbitante, pasa por delante de ella, bloqueando una fracción de la luz de la tal estrella.

4. **Medida de pulsos de radio de un púlsar:** Un pulsar emite ondas de radio regularmente a medida que gira. Leves anomalías en el momento de sus pulsos permiten rastrear los cambios en el movimiento del pulsar debido a la presencia de planetas.

5. **Microlentes gravitacionales:** Los campos de gravedad del planeta y su estrella actúan para focalizar la luz de una estrella distante. Se requiere que los tres objetos se encuentren alineados.

6. **Detección visual directa:** Método basado en la obtención de imágenes de los planetas extrasolares. Sin embargo, esto es casi imposible debido a la diferencia entre el brillo de las estrellas y el de los planetas.

Afortunadamente, los avances tecnológicos en fotometría han permitido que la sonda Kepler tenga sensibilidad suficiente para detectar una mayor gama de exoplanetas. Hecho que resulta evidente al observar la gran cantidad de planetas extrasolares que se han detectado hasta el día de hoy haciendo uso del método de tránsito [3].

[3] https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

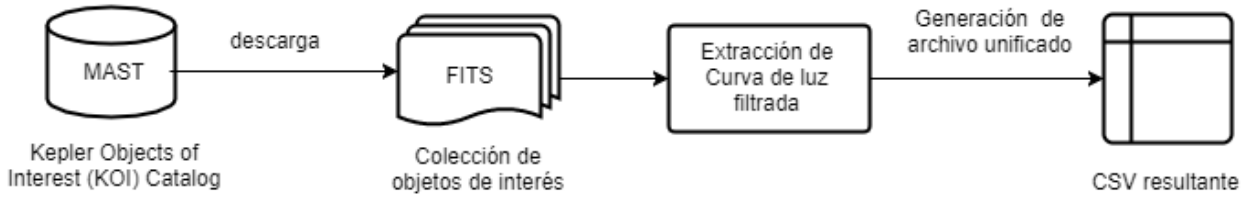



**Figura 1** Proceso de obtención de curvas de luz y creación de datos con los que se trabajó

***Misión de Kepler***

Kepler es un observatorio espacial lanzada por la NASA el 6 de marzo de 2009, con el objetivo de buscar planetas del tamaño de la Tierra en nuestro vecindario de la galaxia. Kepler mide la variación de la luz de miles de estrellas distantes, en busca de tránsitos planetarios. La medición de tránsitos repetidos, todos con un período regular, duración y cambio en el brillo, proporciona un método para descubrir y confirmar planetas y sus órbitas.
Kepler monitorea continuamente más de 100.000 estrellas similares al Sol para cambios de brillo producidos por tránsitos planetarios [4].

## IV. Datos

Hasta el día de hoy, NASA Exoplanet Science Institute[5] a mostrado que más del 65% de los planetas extrasolares descubiertos (3564) han sido detectados gracias a la misión Kepler. En vista entonces de que la mayoría de los exoplanetas descubiertos a la fecha han sido detectados haciendo uso del método de tránsito, y sacando provecho de los avances fotométricos de Kepler, se propone trabajar con el dataset *Kepler Objects of Interest* (KOI[6]). Los datos fueron provistos por MAST (*Mikulski Archive for Space Telescopes*), un proyecto fundado por la NASA para proveer datos astronómicos, compuesto por 9564 registros donde cada uno expone 44 atributos de la observación misma (metadatos), incluyendo un link hacia un FITS con la curva de luz.
El proceso en la Fig.1 muestra cómo se adquirieron los datos. En primer lugar se descargaron todos los FITS y se almacenaron de manera local, luego se extrajeron las curvas de luz filtradas de cada FITS y se generó un archivo unificado en formato CSV (*comma-separated values*). Por otro lado, todos los metadatos fueron directamente descargados desde MAST en formato *txt*.
Así entonces, se recolectaron 8054 FITS, donde algunos de ellos contenían múltiples extensiones de tablas binarias (BinTableHDU) debido a que se trataban de distintas observaciones en un mismo sistema, distintos *KOI* cada uno orbitando la misma estrella, dentro de estas tablas se encuentran los datos asociados a la medición de la intensidad de luz y los procesos realizados para la observación. Entre estos, el momento en que fue medido, la curva de luz, el error en la medición de ésta, la curva de luz bajo el filtro *whitened* y un modelo ajustado (de Mandel-Agol).

Cada registro en este dataset corresponde a un exoplaneta etiquetado como *CONFIRMED*, *FALSE POSITIVE* o *CANDIDATE* según el Instituto de la NASA de exoplanetas[7].

- 2281 *CONFIRMED*: aquellos que tras extensos análisis han sido catalogados como exoplanetas.
- 3976 *FALSE POSITIVE*: aquellos que se creían exoplanetas pero que, tras estudios, se ha determinado que no presentan las características de un planeta extrasolar.
- 1798 *CANDIDATE*: aquellos que continúan en análisis.

Los objetos de estudio etiquetados como *FALSE POSITIVE* se deben, posiblemente, a que las observaciones no coinciden con la posición de la estrella en estudio, indicando que el tránsito esta en torno a un objeto en segundo plano. Otra posibilidad es que la profundidad de los tránsitos pares sean estadísticamente diferentes a la profundidad de los tránsitos impares, indicando la ocurrencia de un eclipse binario, es decir, en lugar de tratarse de un planeta, lo que se estuvo observando se trataba de una estrella.
Así entonces, el largo de cada curva de luz (Figura 2) almacenada constaba originalmente de unas 70.000 mediciones. Sin embargo, considerando algunas de las propiedades que tiene una serie de datos (Falk *et al.* (2012)), cada dato no es generado de manera independiente y su dispersión varía en el tiempo. Por lo general, una serie está gobernada por una tendencia y/o posee ciclo. Esto es importante debido a que las mediciones del observatorio Kepler no son realizadas en tiempos uniformes obteniéndose entonces curvas de luz con datos faltantes (momentos en que no se realizó la medición propuesta) traduciéndose en la ausencia de observaciones.
En promedio, los datos faltantes correspondieron al 22,98% del largo de la serie original de todas las curva de luz, mostrado en las curvas de luz del Anexo. Es decir, cada curva se posee aproximadamente unas 55.000 mediciones efectivas. Dada esta situación, se propuso una técnica simple que consta de completar los valores faltantes de la serie con datos nulos (ceros). Dada la posibilidad de pérdida de información de valor de esta simple técnica, se propuso (de manera alternativa) completar dichos vacíos con un ajuste lineal.

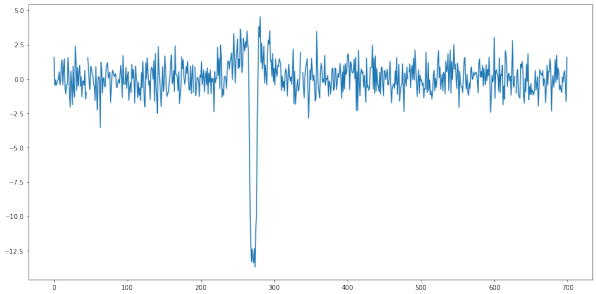


**Figura 2** Muestreo de una curva de luz en su forma bruta, con filtro de blanqueo aplicado.

[4] https://exoplanetarchive.ipac.caltech.edu/docs/KeplerMission.html
[5] https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html
[6] http://archive.stsci.edu/search_fields.php?mission=kepler_koi
[7] http://nexsci.caltech.edu/

Finalmente, de todos los metadatos con los cuales se disponía, solo 10 fueron aplicados en el presente estudio para entrenar los modelos, estos fueron:

- Period
- Transit Depth
- Planet Radius
- (*Planet*) Teq
- (*Stellar*) Teff
- (*Stellar*) log(g)
- (*Stellar*) Metallicity
- Stellar Radius
- Stellar Mass
- KOI count

Donde, *Period* indica el intervalo promedio entre tránsitos basado en un ajuste lineal a todos los tránsitos observados, *Transit Depth* indica la fracción del flujo estelar perdida en el mínimo del tránsito planetario, *Planet Radius* informa el radio de la región de interés (KOI), *Teq* la temperatura de equilibrio en la superficie del planeta, *Teff* la temperatura estelar efectiva (fotosférica) en grados Kelvin, *log(g)* el logaritmo de la gravedad superficial estelar, *metallicity* la metalicidad estelar (logaritmo de la relación *Fe* a *H* en la superficie de la estrella normalizado por la relación solar *Fe* a *H*), *Stellar Radius* radio estelar respecto al Sol (Sol = 1), *Stellar Mass* la masa estelar y *KOI count* indica el número de candidatos identificados en el sistema (el cual varía entre 1 y 7), mayor detalle puede hallarse en [6].

Definidos los atributos con los cuales se trabajaría, se procedió con la división de los datos para el entrenamiento, validación y prueba de los algoritmos a entrenar. Así entonces, el 64% de los datos se fijaron para la etapa de entrenamiento, un 18% para validar y otro 18% como conjunto de pruebas para evaluar y comparar los modelos entrenados, además de estudiar el comportamiento de los futuros datos (desconocidos).

#### *Filtro de blanqueo o whitened filtering*

Los datos con los que se trabajó correspondieron a la curva de luz bruta con *whitened filtering* aplicado, ya que el objetivo de este filtro es obtener una curva de luz con el menor ruido (ruido blanco) y donde las señales relevantes (mayores al ruido) se amplifiquen.
*Whitened filtering*, es una transformación lineal que transforma la secuencia de variables aleatorias (con matriz de covarianza conocida) en un conjunto de nuevas variables cuya covarianza es la matriz identidad, es decir, no existe correlación entre las variables y la varianza está normalizada.
La transformación se denomina "blanqueamiento" porque cambia la secuencia de entrada en una secuencia con ruido blanco. Este ruido blanco corresponde a una señal aleatoria que tiene la misma intensidad a diferentes frecuencias, lo que le da una densidad espectral de potencia constante. La operación que se realiza sobre la curva de luz es la de dividir la señal por su propia función de densidad de potencia espectral.

#### *Modelo de Mandel-Agol*

Dentro de los archivos FITS también se presenta el modelo Mandel-Agol ajustado, el cual modela el tránsito de un planeta esférico alrededor de una estrella esférica, como un eclipse, asumiendo una fuente uniforme. Se requiere conocer tanto la distancia desde el centro del planeta hasta el centro de su estrella madre, así como el radio de cada uno de los cuerpos.
La opacidad observada en la intensidad de la luz cuando el planeta eclipsa la estrella es máxima, cuando el planeta orbita sin eclipsar la estrella la opacidad es mínima y uniforme (modelada como cero o nula), mientras que, cuando el planeta está próximo a eclipsar a su estrella, la intensidad se modela como un polinomio cuadrático según indica Mandel and Agol (2002).

## V. Métodos

Sobre los métodos aplicados para la detección automática de exoplanetas, se presentan dos sub-secciones. La primera se centra en la aplicación de técnicas manuales para la extracción y obtención de características mientras que la segunda se enfoca en la aplicación de técnicas automáticas para esto. Ambos métodos son aplicados a las curvas de luz preprocesadas (mencionadas en la sección anterior).
Las técnicas a aplicar tienen por objeto entregar la clasificación correcta a los objetos actualmente investigados por NexScI (*CANDIDATE*) asignando las etiquetas *CONFIRMED* o *FALSE POSITIVE* según indique el modelo entrenado.

#### *Técnica Manual de extracción de características*

El trabajo realizado para la creación manual de características especializadas para secuencias de tiempo (en este caso secuencias de mediciones de intensidad de luz), *feature generation*, se inspiró en la librería *Feature Analysis for Time Series* (*FATS*) para Python (Hinners *et al.* 2017). Esta librería fue creada con el propósito de extraer características de datos astronómicos (curvas de luz) y ha sido utilizada anteriormente sobre el mismo dataset con el que se trabaja en la presente investigación (Hinners *et al.* 2017), además que este tipo de extracción de características ha sido utilizado en curvas de luz, Richards *et al.* (2011), Donalek *et al.* (2013) y Mahabal *et al.* (2017).
Se decidió no hacer uso de la librería propiamente tal puesto que, al parecer, no ha sido actualizada durante los últimos años[8]. Además, los tiempos de ejecución demostraron ser bastante altos para la cantidad de datos con los se disponía. Dada dicha situación, se implementaron algunas de las características que extraía la librería originalmente, las cuales fueron:

- *Amplitude*, definida como la diferencia entre el valor máximo y el valor mínimo de los datos dividido en dos.
- *Slope*, definida como la pendiente de un ajuste lineal a la curva de luz.
- *Max*, el valor máximo de la secuencia.
- *Mean*, la media aritmética de la secuencia.
- *Median*, la mediana de la secuencia.
- *Median Abs Dev*, definida como la mediana de la diferencia entre cada punto a la mediana de la secuencia.
- *Min*, el valor mínimo de la secuencia.
- *Q1*, el primer cuartil de los datos en la secuencia.
- *Q2*, el segundo cuartil de los datos en la secuencia.
- *Q31*, siendo la diferencia entre el tercer y el primer cuartil.
- *Residual bright faint ratio*, es la tasa entre el residuo de las intensidades mas tenues sobre las intensidades mas brillantes, con la media aritmética de toda la secuencia como umbral.
- *Skew*, definido como una medición de la asimetría de la secuencia (tercer momento).
- *Kurtosis*, definido como el cuarto momento de la secuencia.
- *Std*, desviación estándar de la secuencia.

Debido a las pocas características que pueden resumir a la curva de luz en su totalidad (55 mil mediciones de intensidad de la luz efectivas aproximadamente), se decidió agregar algunos de los metadatos presentados en la sección anterior, los cuales aportan información adicional a cada una de las observaciones de los exoplanetas y sus curvas de luz.

[8] https://github.com/isadoranun/FATS

### *Técnicas Automáticas de extracción de características*

Para la creación de características de manera automática se utilizaron técnicas no supervisadas, las cuales buscan detectar patrones intrínsecos de todos los datos independientemente de la tarea asignada, que en este caso la de detectar un exoplaneta.
El primer método corresponde a Principal Component Analysis (**PCA**) Pearson (1901). Este algoritmo es un método lineal que propone proyectar los datos a un espacio de dimensionalidad menor, es decir, lleva los datos desde todas sus dimensiones originales (en este caso el largo de la secuencia) a un nuevo espacio de menor dimensionalidad definido por las componentes (vectores) de mayor varianza. PCA es conocido en la literatura como uno de los mejores algoritmos para la reducción de dimensionalidad y ha sido aplicado en diversas ocasiones obteniendo muy buenos resultados en lo que se refiere a secuencias de tiempo (Gamit *et al.* (2015), Verleysen and François (2005), Cao *et al.* (2003)). Además de su gran eficiencia a la hora de enfrentarse con datos de gran escala, PCA se ve beneficiado de la optimización de los métodos lineales presentes en las distintas librerías de las cuales este método hace uso.
Como segundo algoritmo, se utilizó **FastICA** (*Fast algorithm for Independent Component Analysis*) Hyvärinen and Oja (2000), un algoritmo iterativo más eficiente para encontrar las componentes estadísticamente independientes de los datos, comparado con las no correlacionadas que busca PCA. Este algoritmo está enfocado en señales, puesto que intenta detectar las fuentes independientes que, mezcladas, emiten la señal que se observa.
Como variante, se probaron estos tres métodos automáticos sobre la representación de las curvas de luz aplicadas a una transformada de Fourier Discreta (Harris (1978)), es decir, pasar del dominio de tiempo en que fueron realizadas las mediciones, al dominio de frecuencias que generan la señal. Este método es generalmente aplicado en el procesamiento de señales periódicas, las cuales justamente cumplen las curvas de luz de objetos transientes. El potencial de Fourier radica en que permite descomponer una señal compleja en un conjunto de componentes de frecuencias únicas sin indicar en qué momento se emitieron.

### *Modelos a entrenar*

Dentro de los algoritmos entrenados sobre las características generadas con las técnicas ya mencionadas, el primero corresponde a un modelo bastante simple que considera los $k$ vecinos más cercanos, conocido como $k$**-NN**, Dasarathy (1991). Éste algoritmo busca clasificar un punto en base la clase mayoritaria de sus $k$ puntos (vecinos) más cercanos. La métrica de distancia asignada por defecto para este algoritmo corresponde a Minkowski [9].
$k$**-NN** opera en base a memoria, puesto que lo que hace es recordar todos los datos con los que se entrenó, para así, en base a la cercanía, clasificar futuros datos.
El segundo algoritmo corresponde a una **Regresión Logística Regularizada**, una variante de la regresión logística propuesta por Cox (1958), este es un método lineal de clasificación basado en un modelo probabilístico binario, en él se define una frontera lineal que separa las clases en base a la probabilidad de asignación a cada una de ellas y que es modelada utilizando una función logística.
La regularización del método es utilizada para penalizar el modelo y así evitar que se sobreajuste a los datos de entrenamiento, es decir, que no aprenda patrones específicos de los datos con los que se entrenó la máquina, sino que generalice de mejor manera para así predecir la data futura.
Otro algoritmo lineal entrenado corresponde a la **Máquina de Soporte Vectorial** lineal (SVM por sus siglas en inglés), Vapnik (1998), la cual intenta encontrar aquella frontera que mejor separe los puntos de las clases. SVM no corresponde a un modelo probabilístico como lo es la regresión logística. A diferencia de éste, SVM no considera todos los puntos para realizar su ajuste sino que sólo considera aquellos puntos que estén a una distancia menor o equivalente a 1 (llamados vectores de soporte). Como variación, se utiliza el modelo regularizado con el mismo objetivo que en Regresión Logística, evitar el sobreajuste sobre los datos de entrenamiento. Para ambos modelos, las predicciones se penalizan haciendo uso de la norma $l_1$ o $l_2$ según se especifique en los parámetros del algoritmo.
Otra variación a la SVM lineal lo es la SVM no lineal utilizando *Radial Basis Function* (RBF), kernel gaussiano. El uso de un kernel permite realizar una separación lineal en un espacio de mayor dimensiones a las que se trabaja, lo que (de manera visual) equivaldría a realizar una separación no lineal. En esta configuración, SVM se vale del parámetro *gamma* para definir cuán influyente es un dato de entrenamiento, en otras palabras, corresponde al inverso del radio de influencia de los datos de entrenamiento seleccionados por el algoritmo (vectores de soporte). De esta manera, se trabaja con una frontera mucho más potente que la frontera definida por una SVM lineal debido a su capacidad de ajuste cuando la máquina se enfrenta a datos no separables linealmente. Nuevamente, se utiliza la variante regularizada del modelo.
Como último modelo se utilizó **Random Forest Regressor**, Ho (1995), un método de ensamblado de máquinas, en el cual se entrenan múltiples árboles de decisión a la vez sobre distintas muestras *bootstrap* de los datos. Bajo este modelo se seleccionan atributos, de manera aleatoria, para particionar la muestra. En cuanto a la predicción del ensamblado, éste entrega como resultado la clase mayoritaria entre todos los árboles entrenados.

Como variante al modelo a entrenar y como técnica de extracción de características de manera automática se utilizó lo que son las redes neuronales recurrentes, en específico de las **LSTM** (*Long Short Term Memory*) Hochreiter and Schmidhuber (1997) y **GRU** (Gated Recurrent Unit Cho *et al.* (2014)), las cuales están especializadas en secuencias de datos puesto que cada registro es dependiente del siguiente dato en la secuencia. LSTM y GRU son cruciales debido al tamaño de la secuencia, puesto que éstas logran detectar patrones entre señales bastante extensas ya que tienen la capacidad de olvidar puntos que no son útiles y mantener aquellos que sí lo son.
Dada la dificultad para entrenar estas redes con lo extensa de las secuencias con las cuales se disponía, se transformó la curva de luz en secuencias de estadísticos por ventana. En este caso: máximo, mínimo, media, desviación estándar y tercer momento estadístico.

### *Métricas*

Debido a que el problema al que se enfrenta corresponde a un problema de clasificación binaria desbalanceado, se utilizaron las métricas de calidad *precision*, *recall* (por clase) y *f1 score*, donde este último resume *precision* y *recall* en una sola métrica sobre todos los datos (ambas clases), según lo indicado en [10].

[9] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html

[10] http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

- **Precision**
  Tasa entre aquellos objetos etiquetados como pertenecientes a una cierta clase sobre la suma de los objetos pertenecientes efectivamente a dicha clase y aquellos objetos mal etiquetados. En otras palabras, esto corresponde a la habilidad del clasificador para etiquetar de una clase A únicamente a los objetos que pertenezcan a esta clase A.

$$P = \frac{T_p}{T_p + F_p} \tag{1}$$

  Con $T_p$ como los *true positive* y $F_p$ como los *false positive*.

- **Recall**
  Tasa entre aquellos objetos etiquetados como pertenecientes a una cierta clase sobre la suma de los objetos pertenecientes efectivamente a dicha clase y aquellos objetos etiquetados como erróneamente pertenecientes a la clase contraria. En otras palabras, corresponde a la habilidad del clasificador en incluir a todos los ejemplos de una cierta clase A.

$$R = \frac{T_p}{T_p + F_n} \tag{2}$$

  Con $F_n$ como los *false negative*.

- **F1-score**
  Es definido como la media armónica entre las dos métricas mencionadas anteriormente, siendo alto cuando ambos, *precision* y *recall*, son altos, por lo que es una buena medida de calidad del clasificador en base a estas dos métricas.

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

  Estas métricas alcanzan su mejor valor en 1 y peor en 0.

## VI. Experimentos y resultados

Debido a la gran cantidad de datos que fueron procesados, se hizo uso del cluster proporcionado por ChiVO[11] (Chilean Virtual Observatory), en el cual se descargaron los 6257 datos etiquetados, correspondientes a 121 GB, además de los 1797 datos no etiquetados o candidatos, correspondientes a otros 33 GB. Así entonces, se agruparon 4000, 1000 y 1000 datos aproximadamente como conjuntos de entrenamiento, validación y pruebas respectivamente, tal que, para sintonizar los hiperparámetros estructurales de los distintos algoritmos se utilizó el conjunto de validación y luego se enfrentaron los mejores modelos con el conjunto de prueba para simular cómo se comportarán los modelos frente a datos futuros.
En cuanto a los híperparámetros de los algoritmos entrenados, fue necesario definir:

- $k$-NN: número de vecinos $k$.
- Regresión Logística: parámetro de regularización $C$.
- SVM: parámetro de regularización $C$.
- *Random Forest*: profundidad máxima *max depth*.

Cabe destacar que la selección de dichos hiperparámetros no fue costosa en términos computacionales debido a que se realizó sobre las representaciones de características extraídas de las técnicas ya comentadas, las cuales son mucho menores que la cantidad de datos ($d \ll n$, donde $n$ indica el número de datos y $d$ la dimensionalidad de éstos).

[11] www.chivo.cl

| | **5** | **10** | **15** | **20** | **25** | **50** |
|---|---|---|---|---|---|---|
| ICA | 0.711 | 0.709 | 0.709 | 0.686 | 0.679 | 0.675 |

| | **5** | **10** | **25** | **55** | **100** | **255** |
|---|---|---|---|---|---|---|
| PCA | 0.713 | 0.701 | 0.701 | 0.699 | 0.702 | 0.689 |

**Tabla 2** *F1 score* del mejor modelo, *Random Forest*, en función de las dimensiones que se experimentaron.

Para las técnicas automáticas de extracción de características se experimenó con dimensiones fijas (**5**, 10, 15, 20, 25, 50 para ICA y **5**, 10, 25, 55, 100 para PCA), pudiendo ver en la Tabla 2 cómo varía el desempeño del mejor modelo en función de la dimensionalidad. Es evidente que al aumentar las dimensiones (características) extraídas de los datos, el error aumenta. Fenómeno que indica un posible exceso de información bastando sólo con unos pocos atributos para definir un buen modelo predictivo.
Dentro de la experimentación, hacer uso de una transformada de Fourier discreta para pasar al dominio de frecuencias y luego extraer características en ésta nueva representación, demostró ser un procedimiento crucial (y de una clara mejora) al momento de extraer características automáticamente. Al contrario, al aplicar dichas técnicas de extracción de características a los datos brutos (en el dominio del tiempo), el error resultó ser aleatorio (todos los ejemplos de la clase *false positive* 0.486 o de la clase *candidate* 0.200 de *f1 score*).

Entre las técnicas para completar los datos faltantes que se nombraron (rellenar con ceros o completar con linea de tendencia) la que arrojó mejores resultados fue la de completar con ceros (mejora en $\sim 0.1$). Otra técnica que se llevó a cabo fue la de realizar un *sampling* de la secuencia tomando el dato con mayor magnitud cada 3 instantes de tiempo (considerando los datos faltantes como ceros) para luego completar los datos que quedaron faltantes haciendo uso de la línea de tendencia. Sin embargo, esto reflejó un mayor error, por lo que se decidió no informar tales resultados en el presente.
Al realizar el *sampling*, el total de datos faltantes (22.98% de la secuencia original), se reduce a un 20%, indicando que hay aproximadamente un 3% de datos no informados contiguos.
Si bien la extracción de características manual tiene la ventaja de no necesitar que las secuencias de datos tengan el mismo largo, ya que se extraen estadísticos de ésta independientemente de las otras, resulta ser peor que las técnicas automáticas debido a que éstas se adaptan a los datos (en el proceso de entrenamiento) y logran extraer información de valor para la predicción.
Además de extraer características de la curva de luz manualmente, se trabajó directamente con los metadatos de MAST, resultados que se pueden ver y comparar en la Tabla 3. En primer lugar, se hizo uso de los metadatos correspondientes al supuesto exoplaneta en estudio (KOI - *kepler object of interest*), los cuales son el período de la órbita, la profundidad de tránsito, el radio del planeta, la temperatura de equilibrio del planeta y el número de *KOI* en estudio en tal sistema. Este enfoque demostró obtener un rendimiento bastante alto comparado con las técnicas que se enfrentaron a los datos brutos. Otro enfoque correspondió al hacer uso de los metadatos de la estrella madre del exoplaneta siendo la temperatura efectiva, la metalicidad, la

| | Learners | | | |
|---|---|---|---|---|
| | *k-NN* | *Logistic Regression* | *SVM RBF* | *Random Forest* |
| Fourier + PCA | 0.679 | 0.493 | 0.486 | **0.713** |
| Fourier + ICA | 0.679 | 0.493 | 0.486 | **0.711** |
| OwnFATS | **0.666** | 0.583 | 0.575 | 0.658 |
| Planet metadata | 0.825 | 0.848 | 0.848 | **0.870** |
| Stellar metadata | 0.766 | 0.718 | 0.751 | **0.766** |
| OwnFATS + stellar & planet metadata | 0.844 | 0.864 | 0.876 | **0.883** |

**Tabla 3** F1 score en la clasificación de los distintos modelos (*learners*) sobre el conjunto de pruebas sobre las distintas representaciones generadas.

gravedad, el radio y la masa de la estrella. Al trabajar con estas características se pudo notar la gran mejora que se obtuvo, por lo que efectivamente se pudo extraer información de mayor valor para la clasificación de los ejemplos respecto a las técnicas utilizadas para extraer características netamente de la curva de luz. Como resultado alternativo se mezclaron estos procesos manuales, donde se utilizan los metadatos en conjunto con las características manuales extraídas de la curva de luz.

Los resultados son presentados en la Tabla 3, en donde se puede presenta la métrica *f1 score* para las mejores representaciones de cada técnica, con 5 características en PCA e ICA, rellenando con ceros los datos faltantes y asignando pesos balanceados a las clases (sin subsamplear la clase mayoritaria). Así entonces, luego de las diversas variantes en el proceso de experimentación (Tabla 3), el mejor resultado se obtuvo haciendo uso de las técnicas manuales, en la que se extraen estadísticos y características fijas de la curva de luz además de otras características a partir de los metadatos refentes tanto al planeta como la estrella, los cuales son de gran importancia y que están fuertemente relacionados con la observación. El rendimiento obtenido fue del 88,3% para la futura clasificación, según indica la métrica *f1 score*, significando que en un 88,3% de las veces la predicción será la correcta. Para esta configuración, el mejor modelo entrenado fue el de *Random Forest*, del cual en la Figura 3 se presenta la importancia de atributos que éste emplea en su proceso de selección. En dicha imagen, es posible notar que los atributos referentes al planeta resultan ser aquellos más relevantes, valor esperado en base a los resultados obtenidos, donde el radio del objeto de interés el más influyente para las clasificación en conjunto con el período y el número de objetos en estudio en tal sistema. Las características menos importantes pasan a ser las que fueron extraídas de la curva de luz, donde la pendiente (*slope*) y el cuartil 2 son los menos relevantes.

Para trabajar el desbalanceo de los datos se experimentó con la técnica de *undersampling* (subsamplear, (Drummond *et al.* 2003)) sobre la clase mayoritaria de manera aleatoria para que, de este modo, se equiparase la cantidad de ejemplos de ambas clases en el conjunto de entrenamiento. Por otro lado, se trabajó con los datos desbalanceados y al momento de entrenar los modelos Regresión Logística, SVM y *Random Forest* se utilizó la técnica de ponderación (mediante la signación de peso) a las clases de manera tal que éstas resultasen balanceadas. De este modo, los datos mayoritarios no tendrían un impacto mayor sobre los minoritarios al afectar más fuertemente en la función objetivo del modelo entrenado (King and Zeng 2001). Cabe destacar que esta última configuración fue aquella que logró mejores resultados (mejorando en un valor de $\sim 0.1\%$).

Ahora bien, respecto a la experimentación con redes neuronales recurrentes, se llevaron a cabo diversas experimentaciones con distintas representaciones de entrada, variando el tamaño de la ventana entre 300 y 500, con un *stride* fijo de 100. Cabe mencionar que, igualmente, se varió la architectura de la red modificando la profundidad, el número de neuronas, el optimizador (RMSprop y Adam), numero de *epochs* así como el tamaño del batch durante el proceso de entrenamiento, todo esto sin obtener resultados interesantes. Desafortunadamente, no se logró que la red aprendiera a partir de los estadísticos por ventana de las curvas de luz, siendo una pequeña red con GRU aquella de mejor rendimiento, con 0.567% según indicó *f1 score*. Esta situación pudo deberse a que las secuencias fueron demasiado largas y, por ello, los estadísticos por ventana no lograsen resumir la información necesaria para el correcto aprendizaje de la red.

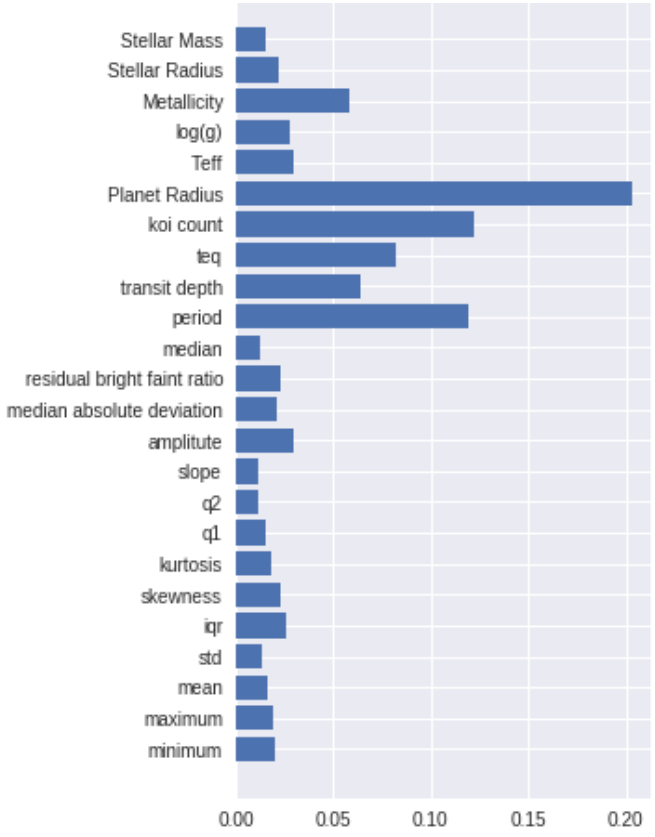



**Figura 3** Importancia de atributos en modelo *Random Forest* para la repressentación con mejor rendimiento (OwnFATS con metadatos)

El detalle de las métricas de *precision* y *recall* sobre la clasificación para ambas clases en el conjunto de pruebas, se puede apreciar en el Anexo. De la Tabla 5, la clasificación para los falsos positivos, se puede ver que el mejor modelo que logra identificar correctamente esta clase en términos de éstas métricas, es decir, aquel que incluye a la gran mayoría de los ejemplos de esa clase y lo hace de manera minuciosa (sin incluir ejemplos de la otra clase), es la SVM RBF, siendo la que mejor identifica en la mayoría de los casos esta clase. Esto se pudo deber al uso de un kernel tipo RBF, el cual ayuda a ajustar las fronteras de manera flexible y muy estricta a la clase en cuestión. Igualmente, en esta tabla se puede apreciar que aplicar la técnica ICA sobre el dominio de frecuencias logra un *precision* más alto que todas las técnicas, inclusive los metadatos, pero al tener un bajo *recall* esto se traduce en la inclusión de sólo una porción de todos los elementos de la clase. Por otro lado en la Tabla 6, la clasificación para los confirmados, se puede ver que tienen *scores* más bajos que la otra clase, indicando la gran dificultad que presenta la clasificación sobre los que efectivamente resultan ser exoplanetas. Esto, posiblemente porque no cuentan con características similares respecto a la curva de luz, dificultando al modelo el clasificar los ejemplos en todos los casos. Sin embargo, se puede ver que el modelo que logra identificar esta clase de mejor manera (en la mayoría de las representaciones) fue *Random Forest*.

**Resultados finales:** Luego de haber realizado las debidas pruebas y de haber identificado aquel método de mejor comportamiento frente a datos futuros (conjunto de pruebas) en base a las métricas *precision* y *recall*, se presentan las predicciones entregadas mediante la representación -**OwnFATS + stellar & planet metadata**- con el modelo *Random Forest* para la identificación de aquellos objetos *CONFIRMED*, con máxima profundidad de 15, y el modelo SVM RBF, con parámetro de regularización 100, para la identificación de aquellos *FALSE POSITIVE*.
Así entonces, se entrega la clasificación sobre los objetos de interés (*Kepler Objects of Interest*) que aún están siendo estudiados por el personal científico de NexScI a septiembre del pasado año 2017, es decir, aquellos objetos etiquetados como *CANDIDATE*. En la Tabla 7 (Anexo) se presenta una muestra de las etiquetas informadas por los algoritmos en cuestión, siendo éstas *CONFIRMED* o *FALSE POSITIVE* según corresponda. Cabe destacar que para aquellos objetos que fueron identificados tanto como exoplaneta como falso exoplaneta, fueron etiquetados como *UNCLASSIFIED* puesto que no se era posible llegar a un consenso entre ambos algoritmos. En esta tabla se puede ver por ejemplo que el sistema de la estrella Kepler 279, habiendo 2 exoplanetas confirmados ya por NexScI (Kepler 279 b y Kepler 279 c), nuestras técnicas indican que el tercer objeto en estudio **K01236.04** pasa a ser un exoplaneta válido como los otros orbitando la estrella, caso contraro se ve en el sistema de la estrella Kepler 619, en donde también orbitan 2 exoplanetas (Kepler 619 b y Kepler 619 c), nuestra técnica indica que el tercer objeto en estudio **K00601.02** es un falso positivo. También se puede ver que nuestra técnica clasifica a todo un sistema completo (KOI K01358), en la cual sus 4 regiones de interés (**K01358.01**, **K01358.02**, **K01358.03** y **K01358.04**) son clasificadas como exoplanetas válidos en ese sistema. Esto último también se ve en el caso en que se clasifica las regiones de estudios de la estrella Kepler 763, con el exoplaneta Kepler 763 b orbitando, indicando que existen 2 exoplanetas hermanos orbitando esta estrella y que el cuarto objeto de estudio **K01082.02** es un falso positivo.

A continuación se presenta un cuadro resumen de los resultados acusados por las técnicas presentadas en el documento:

| **Total *CANDIDATE*** | **1791** |
|---|---|
| Subtotal *CONFIRMED* | 975 |
| Subtotal *FALSE POSITIVE* | 434 |
| Sin clasificar | 382 |

**Tabla 4** Subtotales de exoplanetas candidatos, según método correspondiente.

## VII. Conclusiones

Se introduce una nueva forma de entregar la decisión sobre un objeto en estudio (KOI) de manera automática, haciendo uso de diversas técnicas de aprendizaje automático enfocadas en la manipulación de datos brutos (secuencia de mediciones de la intensidad de la luz), con el objetivo de imitar el extenso trabajo que hacen los expertos al identificar si el objeto es efectivamente un exoplaneta o resulta ser algún otro fenómeno.
En base a los resultados se pudo ver que las técnicas automáticas aquí presentadas para extraer información de la curva de luz no fueron lo suficientemente buenas en comparación con los metadatos, los cuales superan dicho enfoque en cuanto a *score*. Siendo, tal vez, los métodos no adecuados para la extracción de características o bien, muy simples para el problema enfrentado. Problema que resultó ser bastante complejo, por una parte el costo computacional debido a lo extenso que son las mediciones de la curva de luz y en cuanto a las características del dominio del problema, puesto que las curvas de luz son muy diversas en cuanto a la "morfología".
La variación en cuanto a los metadatos trabajados puede tener gran impacto en los resultados informados en el presente documento puesto que la elección de éstos se apoyó netamente en la descripción de los campos informados por MAST en su apartado KEPLER_KOI Field Descriptions. Dicha elección pudo verse modificada, logrando incluso mejores resultados, al haberse apoyado de personal experto en el área.
Respecto al trabajo futuro, se propone la modificación de las técnicas de autocompletado utilizadas para la eliminación de datos no informantes. De manera similar, se recomienda la aplicación de nuevas técnica para la extracción de nuevas características de la curva de luz.

## Agradecimientos

Este trabajo pudo ser realizado gracias a Chilean Virtual Observatory, ChiVO. Se agradece igualmente a los académicos Mauricio Araya y Ricardo Ñanculef, ambos del Departamento de Informática de la Universidad Técnica Federico Santa María por su contribución al presente trabajo.

## Referencias

Cao, L., K. S. Chua, W. Chong, H. Lee, and Q. Gu, 2003 A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. Neurocomputing **55**: 321–336.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, *et al.*, 2014 Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 .

Cox, D. R., 1958 The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) pp. 215–242.

Dasarathy, B., 1991 Nearest neighbor norms: Nn pattern classification techniques.

Donalek, C., S. G. Djorgovski, A. A. Mahabal, M. J. Graham, A. J. Drake, *et al.*, 2013 Feature selection strategies for classifying high dimensional astronomical data sets. In *Big Data, 2013 IEEE International Conference on*, pp. 35–41, IEEE.

Drummond, C., R. C. Holte, *et al.*, 2003 C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pp. 1–8, Citeseer.

Falk, M., F. Marohn, R. Michel, D. Hofmann, M. Macke, *et al.*, 2012 A first course on time series analysis: Examples with sas [version 2012. august. 01] .

Gamit, M. R., P. Dhameliya, and N. S. Bhatt, 2015 Classification techniques for speech recognition: A review. vol **5**: 58–63.

Harris, F. J., 1978 On the use of windows for harmonic analysis with the discrete fourier transform. Proceedings of the IEEE **66**: 51–83.

Hinners, T., K. Tat, and R. Thorp, 2017 Machine learning techniques for stellar light curve classification. arXiv preprint arXiv:1710.06804 .

Ho, T. K., 1995 Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pp. 278–282, IEEE.

Hochreiter, S. and J. Schmidhuber, 1997 Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pp. 473–479.

Hyvärinen, A. and E. Oja, 2000 Independent component analysis: algorithms and applications. Neural networks **13**: 411–430.

King, G. and L. Zeng, 2001 Logistic regression in rare events data. Political analysis **9**: 137–163.

Mahabal, A., K. Sheth, F. Gieseke, A. Pai, S. G. Djorgovski, *et al.*, 2017 Deep-learnt classification of light curves. arXiv preprint arXiv:1709.06257 .

Mandel, K. and E. Agol, 2002 Analytic light curves for planetary transit searches. The Astrophysical Journal Letters **580**: L171.

Mayor, M. and D. Queloz, 1995 A jupiter-mass companion to a solar-type star.

Pearson, K., 1901 Principal components analysis. The London, Edinburgh and Dublin Philosophical Magazine and Journal **6**: 566.

Richards, J. W., D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, *et al.*, 2011 On machine-learned classification of variable stars with sparse and noisy time-series data. The Astrophysical Journal **733**: 10.

Vapnik, V., 1998 *Statistical learning theory. 1998*. Wiley, New York.

Verleysen, M. and D. François, 2005 The curse of dimensionality in data mining and time series prediction. In *IWANN*, volume 5, pp. 758–770, Springer.

## Anexo

**Métricas:**

| | Learners | | | |
|---|---|---|---|---|
| | *k-NN* | *Logistic Regression* | *SVM RBF* | *Random Forest* |
| Fourier + PCA | P: 0.726<br>R: 0.809 | P: 0.902<br>R: 0.283 | P: 0.629<br>R: 1.000 | **P: 0.789**<br>**R: 0.736** |
| Fourier + ICA | P: 0.752<br>R: 0.728 | P: 0.899<br>R: 0.285 | **P: 0.933**<br>**R: 0.332** | P: 0.788<br>R: 0.730 |
| OwnFATS | P: 0.743<br>R: 0.695 | P: 0.821<br>R: 0.441 | **P: 0.890**<br>**R: 0.395** | P: 0.827<br>R: 0.569 |
| Planet metadata | P: 0.863<br>R: 0.857 | P: 0.917<br>R: 0.830 | **P: 0.927**<br>**R: 0.817** | P: 0.914<br>R: 0.871 |
| Stellar metadata | P: 0.781<br>R: 0.886 | P: 0.806<br>R: 0.714 | P: 0.818<br>R: 0.768 | **P: 0.819**<br>**R: 0.803** |
| OwnFATS + stellar & planet metadata | P: 0.860<br>R: 0.899 | P: 0.919<br>R: 0.857 | **P: 0.934**<br>**R: 0.861** | P: 0.924<br>R: 0.884 |

**Tabla 5** scores para Falsos positivos en conjunto de pruebas, donde (P) es Precision y (R) es Recall.

| | Learners | | | |
|---|---|---|---|---|
| | *k-NN* | *Logistic Regression* | *SVM RBF* | *Random Forest* |
| Fourier + PCA | P: 0.597<br>R: 0.481 | P: 0.438<br>R: 0.948 | P: 0.000<br>R: 0.000 | **P: 0.597**<br>**R: 0.666** |
| Fourier + ICA | P: 0.562<br>R: 0.592 | P: 0.438<br>R: 0.945 | P: 0.458<br>R: 0.960 | **P: 0.592**<br>**R: 0.665** |
| OwnFATS | **P: 0.533**<br>**R: 0.592** | P: 0.468<br>R: 0.836 | P: 0.471<br>R: 0.917 | P: 0.522<br>R: 0.798 |
| Planet metadata | P: 0.763<br>R: 0.773 | P: 0.755<br>R: 0.874 | P: 0.745<br>R: 0.893 | **P: 0.801**<br>**R: 0.864** |
| Stellar metadata | **P: 0.755**<br>**R: 0.585** | P: 0.600<br>R: 0.713 | P: 0.649<br>R: 0.716 | P: 0.681<br>R: 0.703 |
| OwnFATS + stellar & planet metadata | P: 0.818<br>R: 0.756 | P: 0.785<br>R: 0.874 | P: 0.795<br>R: 0.898 | **P: 0.820**<br>**R: 0.879** |

**Tabla 6** scores para Confirmados en conjunto de pruebas, donde (P) es Precision y (R) es Recall.

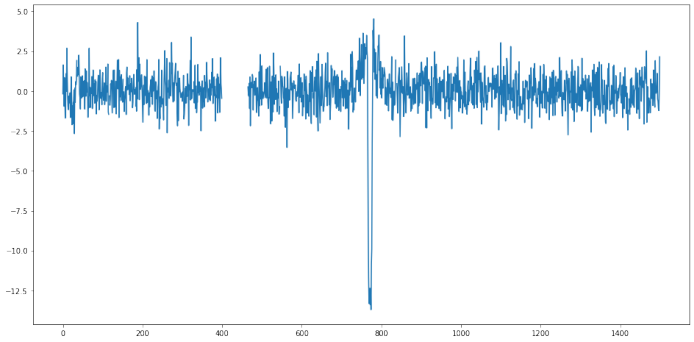
**Resultados:**

| *KOI name* | Disposición | Confirmados en sistema | Estrella |
|---|---|---|---|
| K00601.02 | *FALSE POSITIVE* | 2/3 | Kepler 619 |
| K00750.02 | *UNCLASSIFIED* | 1/3 | Kepler 662 |
| K01082.01<br>K01082.02<br>K01082.04 | *CONFIRMED*<br>*FALSE POSITIVE*<br>*CONFIRMED* | 1/4 | Kepler 763 |
| K01236.04 | *CONFIRMED* | 2/3 | Kepler 279 |
| K01358.01<br>K01358.02<br>K01358.03<br>K01358.04 | *CONFIRMED*<br>*CONFIRMED*<br>*CONFIRMED*<br>*CONFIRMED* | 0/4 | - |
| K01750.02 | *CONFIRMED* | 1/2 | Kepler 948 |
| K02064.01 | *UNCLASSIFIED* | 0/1 | - |
| K02420.02 | *CONFIRMED* | 1/2 | Kepler 1231 |
| K02578.01 | *FALSE POSITIVE* | 0/1 | - |
| K02828.02 | *FALSE POSITIVE* | 1/2 | Kepler 1259 |
| K03444.03 | *UNCLASSIFIED* | 0/4 | - |
| K03451.01 | *UNCLASSIFIED* | 0/1 | - |
| K04591.01 | *FALSE POSITIVE* | 0/1 | - |
| K05353.01 | *FALSE POSITIVE*E | 0/1 | - |
| K06267.01 | *CONFIRMED* | 0/1 | - |
| K06983.01 | *CONFIRMED* | 0/1 | - |
| K07279.01 | *CONFIRMED* | 0/1 | - |
| K07378.01<br>K07378.02 | *CONFIRMED*<br>*CONFIRMED* | 0/2 | - |
| K07434.01 | *FALSE POSITIVE* | 0/1 | - |
| K08082.01 | *CONFIRMED* | 0/1 | - |

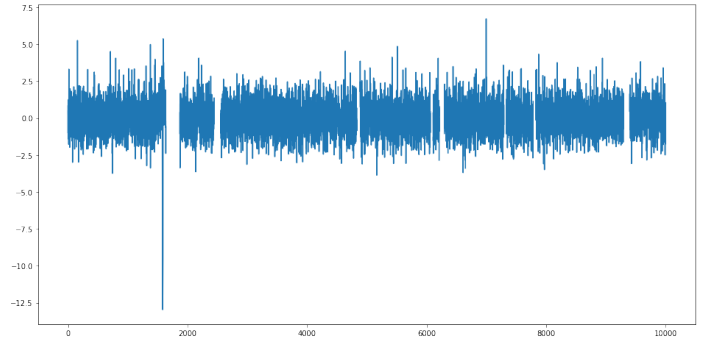
**Tabla 7** En esta tabla se muestran algunas de las predicciones asignadas tanto por **OwnFATS + stellar & planet metadata** con el modelo *Random Forest* para la detección de lo que es un exoplaneta (*CONFIRMED*), como por **OwnFATS + stellar & planet metadata** con *SVM RBF* para la detección de lo que no es un exoplaneta (*FALSE POSITIVE*). Cabe mencionar que la columna **Confirmados en sistema** señala la cantidad de exoplanetas confirmados a la fecha del estudio (Septiembre 2017), mientras que **Estrella** indica el nombre de la estrella madre del sistema en cuestión; Aquellos de los cuales no se tiene mayor información, no presentan este campo.

## Muestreo de los datos de entrada

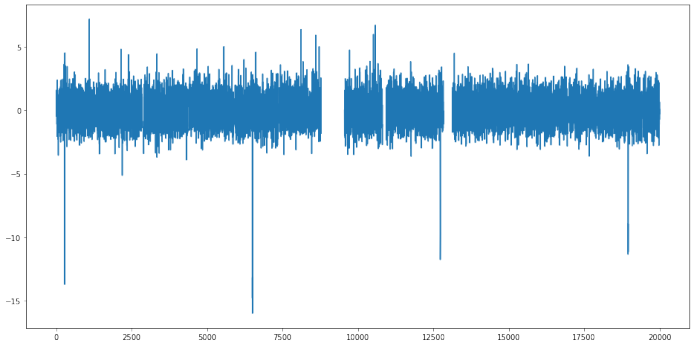
Curvas de luz en su forma bruta



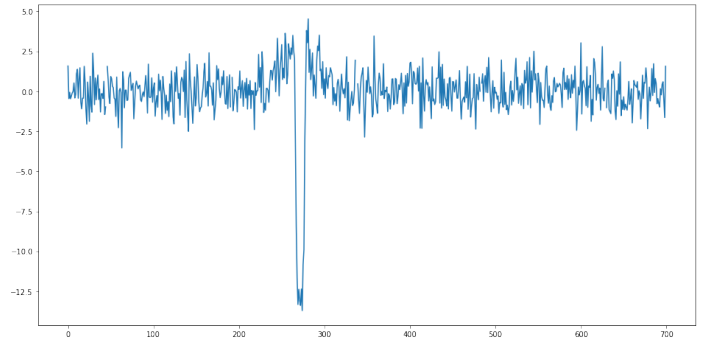
Curva 1



Curva 2

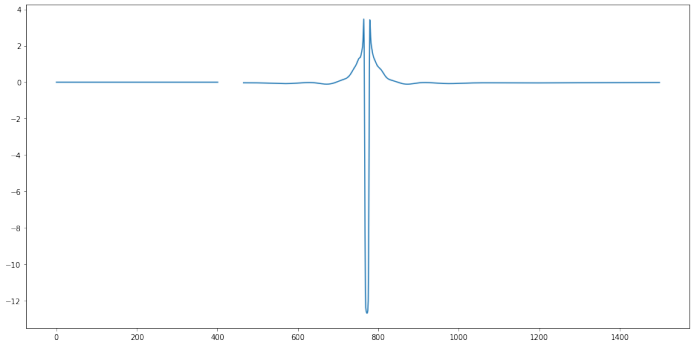


Curva 3

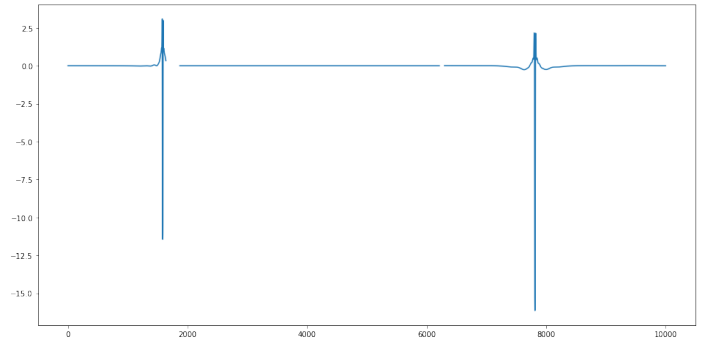


Curva 4

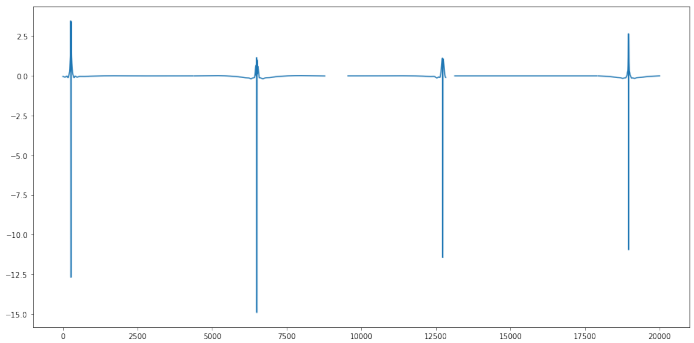
Curvas de luz modelo Mandel-Agol ajustado



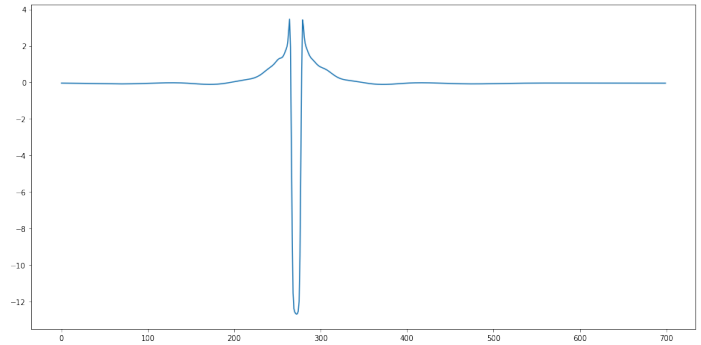
Modelo 1



Modelo 2



Modelo 3



Modelo 4