



PROGRAMA INCENTIVO A LA INICIACIÓN CIENTÍFICA MULTIDISCIPLINARIO PRIMER SEMESTRE 2019

1.1. TÍTULO DEL PROYECTO

Caracterización y comprensión de los procesos en la detección de exoplanetas a través de la validación de modelos de aprendizaje automáticos.

1.2. DEPARTAMENTOS

Departamento de Informática
Departamento de Electrónica


1.3. DURACIÓN

10 meses

1.4. RESPONSABLES

PROFESOR RESPONSABLE (USM)

Mauricio Araya López

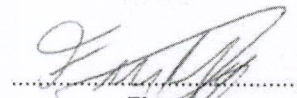

Firma

ALUMNOS INTEGRANTES DEL GRUPO

Margarita Buguño Pérez.....


Firma

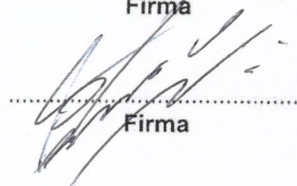
Francisco Mena Toro.....


Firma

Gabriel Molina Barra.....


Firma

Patricio Olivares Roncagliolo.....


Firma



1.5. RESUMEN

La detección y caracterización de exoplanetas es un área de la astronomía que propone el estudio de la formación e identificación de planetas similares a la Tierra fuera del Sistema Solar. Si bien esta área se ha visto fuertemente beneficiada por el incremento en sensibilidad y resolución de los detectores, el procesamiento de estos datos suele ser relativamente manual, exhaustivo y caso a caso, utilizando el análisis estadístico para manejar los grandes volúmenes de datos, pero sin automatizar el proceso de inferencia en base a todos estos datos. Esta propuesta, que inició con el proyecto FONDEF IT15I10041 (Chilean virtual observatory, ChiVO) y es apoyada en su continuidad por el proyecto BASAL FB-0008 (AC3E), tiene como objetivo el automatizar distintos procesos que realizan los astrónomos hoy en día en la detección y correcta categorización de los distintos objetos transientes (exoplanetas en efecto o no). A través de la teoría y las técnicas de las ciencias de la computación, planteamos detectar correctamente exoplanetas en distintos datasets a través del análisis automatizado de curvas de luz. Además, proponemos reproducir la extracción de características que realiza la pipeline de Kepler mediante aprendizaje, y así poder entender y mejorar los procesos de ingeniería de características (i.e., feature engineering). Estos modelos y resultados pueden ser extendidos para obtener objetos celestes sintéticos similares a los transientes reales, así poder generar objetos no medidos ni estudiados al día de hoy.



1.6. PROYECTOS DE INVESTIGACIÓN

Alumno	Programa
Margarita Bugeño Pérez	Magíster en Ciencias de la Ingeniería en Informática

Año	Nro. y Nombre del Proyecto
2017, 2018	IT15I10041 Implementación de Herramientas y Procesos del Observatorio Virtual Chileno (UTFSM en conjunto con UChile)
2018-presente	Instituto Milenio Fundamentos de los Datos, Inteligencia artificial con explicación (UTFSM en conjunto con PUC)

Alumno	Programa
Francisco Mena Toro	Magíster en Ciencias de la Ingeniería en Informática

Año	Nro. y Nombre del Proyecto
2017, 2018	IT15I10041 Implementación de Herramientas y Procesos del Observatorio Virtual Chileno (UTFSM en conjunto con UChile)

Alumno	Programa
Gabriel Molina Barra	Magíster en Ciencias de la Ingeniería en Informática

Año	Nro. y Nombre del Proyecto
2017	IT15I10041 Implementación de Herramientas y Procesos del Observatorio Virtual Chileno (UTFSM en conjunto con UChile)



Alumno	Programa
Patricio Olivares Roncagliolo	Doctorado en Ingeniería Electrónica

Año	Nro. y Nombre del Proyecto
2018-presente	Proyecto basal FB0821 (CCTVAL)-Investigador en temas de predicción con herramientas de Machine Learning

2. PROPUESTAS PREVIAS PIIC O PUBLICACIONES

Si ha tenido financiamiento previo de PIIC favor indicar la siguiente información: nombre proyecto, académico patrocinador, año de adjudicación, estado de artículo (publicado, en prensa, etc.). Si ha tenido otras publicaciones, indicar:

Alumno	Año y Título trabajo	Profesor Responsable	Estado o link Publicación
Francisco Mena Toro	(PIIC) 2018 - Diseño y validación de un modelo predictivo de ozono troposférico basado en la distribución espacial y temporal de emisión contaminantes	Juan Ricardo Ñanculef Alegría	En proceso de escritura para conferencia (CIARP 2019)
Margarita Bugueño Pérez y Francisco Mena Toro	2018 - Refining Exoplanet Detection Using Supervised Learning and Feature Engineering	Mauricio Alejandro Araya López	http://cleilaclo2018.mackenzie.br/docs/SLIOIA/182772.pdf Esperando indexación IEEEExplorer
Margarita Bugueño Pérez	2019 - An Empirical Analysis of Rumor Detection on Microblogs with Recurrent Neural Networks	Marcelo Mendoza Rocha	Aceptado en conferencia HCI 2019. Futura indexación en Lecture Notes in Computer Science, Springer.



3. PROPUESTA DE PROYECTO

3.1 Descripción de propuesta

El estudio de exoplanetas, i.e. planetas que orbitan estrellas fuera de nuestro sistema solar, es un campo de la astronomía que comenzó desde la primera detección confirmada del planeta gigante 51 Pegasi b [1]. Gracias a los avances en instrumentación y técnicas de análisis se ha logrado el descubrimiento de miles de exoplanetas hasta el día de hoy. La NASA reporta que más de 3500 exoplanetas han sido detectados¹ utilizando diferentes técnicas. La dificultad de detectar exoplanetas yace en sus diminutos tamaños en comparación a la distancia de observación, reflejando (o emitiendo) poca luz en comparación a otros objetos cercanos como sus estrellas albergantes. Consecuentemente, se requiere de un análisis bastante detallado y cuidadoso en su estudio.

Al día de hoy, solo se ha fotografiado una cantidad muy limitada de exoplanetas, mientras que la mayoría se han detectado a través de métodos indirectos [3]. El método de detección más exitoso es el de **velocidad radial**, el cual estudia las variaciones de velocidad de una estrella producto de sus planetas orbitantes, analizando las variaciones en frecuencia de las líneas espectrales debido al efecto Doppler. Otro método es el **tránsito fotométrico**, que corresponde a la observación fotométrica sistemática para detectar variaciones en la intensidad de la luz cuando un planeta en órbita eclipsa a la estrella, bloqueando así una fracción de su luz. A este registro de intensidad *versus* tiempo se le conoce como curva de luz. Cabe destacar que este método detecta eficientemente planetas de gran volumen independientemente de la proximidad del planeta a su estrella.

Ahora bien, al momento de trabajar con datos en astronomía, uno de los varios tipos de recolección de información es denominado *survey* astronómico, que para nuestro contexto, corresponde a la recolección de imágenes o espectros de una región de interés en el cielo o un cuerpo celeste en específico [2]. Un claro ejemplo del cómo es que se recolectan estos datos es representado por la misión Kepler².

Kepler fue lanzado por la NASA en 2009 y tiene por objetivo buscar planetas similares a la Tierra, demostrando ser bastante efectivo gracias a los avances tecnológicos en la instrumentación con la cual éste se vale. La serie de procesos, i.e. **pipeline**, que realiza para la identificación de los píxeles que albergan una estrella y las variaciones (eventos) en intensidad de luz que sobrepasan un cierto criterio [4], se realiza midiendo qué tanto se ajustan estas variaciones a un tránsito, requiriendo el conocimiento de expertos en el dominio.

Actualmente se ha estudiado en gran detalle el problema de detectar estrellas variables, sin embargo éstas no describen el mismo problema que se plantea en el presente proyecto, puesto que las estrellas variables varían su intensidad luminosa en base a su composición y no a causa de un objeto eclipsante. Por ejemplo, [5] y [6] clasifican estrellas variables mediante la extracción de características simples (estadísticos y análisis de período/frecuencia) los cuales resultan efectivos para ese problema, pero requieren cierto conocimiento específico del dominio.

Se han propuesto previamente modelos de aprendizaje para automatizar la identificación de objetos candidatos, donde las mediciones de la curva de luz se asemejan a objetos transientes, *versus* objetos Falsos Positivos (mediciones que no corresponden a objetos transientes). En particular se ha trabajado sobre los datos TCE (*Thresholding Crossing Event*) de Kepler, el cual es la versión de datos previa a la que trabajaremos nosotros. El proyecto de *Autovetter* [7] utiliza un modelo de aprendizaje Random Forest que realiza la clasificación de objetos utilizando las características de la *pipeline* de Kepler mostrando gran efectividad. En [8] se utiliza aprendizaje no supervisado para extraer información directamente desde la curva de luz (en su representación bruta) a través de la medición de los *flux* (potencia luminosa) capturados por Kepler.

¹ <http://exoplanets.nasa.gov>

² <https://www.nasa.gov/kepler/overview/historybyborucki>



Las redes neuronales artificiales [9] son un modelo computacional inspirado en el comportamiento básico de las neuronas biológicas del cerebro. Estas redes están compuestas por capas de neuronas conectadas entre sí, donde cada conexión representa la composición de una serie de funciones no lineales, cuyos parámetros se ajustan a partir de un proceso de entrenamiento. Esto permite a la red especializarse en una tarea determinada.

El proceso de entrenamiento de una red neuronal puede ser de tipo *supervisado*, *no supervisado* o *transfer learning*. Mientras en el aprendizaje supervisado existe una relación conocida entre las entradas del sistema y las salidas esperadas, en el aprendizaje no supervisado estas relaciones son inferidas a partir de patrones identificados en los datos entregados. En el caso de transfer learning [10], el entrenamiento parte con otra red pre entrenada sobre algún dominio similar (*source domain*), la cual se usa como base para el aprendizaje de una nueva tarea (*target domain*).

En base a la amplia gama de herramientas y modelos que ofrecen las técnicas de computación e inteligencia artificial a la hora de enfrentarse al dominio de la astronomía, proponemos la automatización de los diferentes procesos que realizan los astrónomos hoy en día en la correcta detección y categorización de los distintos objetos de interés que puedan existir (exoplanetas en efecto o no). En específico, planteamos detectar correctamente exoplanetas en **distintos ambientes**, a través del manejo **directo** de las curvas de luz, en su representación bruta de serie de tiempo, y el refinamiento de parámetros de los modelos aplicados; Reproducir la extracción de características que realiza la *pipeline* de Kepler, la cual ha demostrado ser bastante efectiva, además de utilizar estas dos herramientas para validar y simular cuerpos celestes similares a los transientes que se observan en las curvas de luz de exoplanetas orbitantes.

Los datos con los que se trabajará en esta propuesta corresponden a Objetos de Interés de Kepler (*Kepler Object of Interest*, KOI³), los cuales son un subconjunto de los datos de TCE nombrados anteriormente. El *dataset* corresponde a mediciones de intensidad de luz (curvas de luz) que se sospechan que tienen ciertos patrones periódicos que se ajustan a ser tránsitos fotométricos pudiendo evidenciar la presencia de un exoplaneta o no.

3.2 Hipótesis

La hipótesis que guiará el trabajo de investigación es que el proceso de detección e inferencia de parámetros de objetos transientes en curvas de luz puede automatizarse mediante el uso de aprendizaje de máquinas, procesos de simulación y validación matemática del comportamiento de objetos celeste, obteniendo tasas de error similares a los métodos semi-manuales acusados en el estado del arte.

3.3 Objetivos

Objetivo general:

- Utilizando series de tiempo, en específico curvas de luz, proponemos automatizar la detección exoplanetas así como su caracterización y modelado a través de herramientas de aprendizaje automático.

Objetivos específicos:

- Diseñar e implementar un modelo supervisado que aprenda directamente de curvas de luz extensas en el dominio temporal.
- Diseñar e implementar un modelo no supervisado que aprenda a imitar los procesos de la *pipeline* de Kepler directamente de curvas de luz.
- Encontrar un modelo base, derivado de los propuestos, para realizar *transfer learning* sobre otros *surveys* de cuerpos transientes.

³ https://exoplanetarchive.ipac.caltech.edu/docs/Kepler_KOI_docs.html



- Generar datos sintéticos a través de un modelo probabilista no supervisado, validado con datos reales, de objetos transientes que no hayan podido ser estudiados en profundidad debido a limitaciones tecnológicas.

3.4 Metodología

Para cumplir con los objetivos específicos se deberá, en primer lugar, generar una representación de los datos que sea válida para distintos *datasets/surveys*. Esto permitirá generar modelos válidos para diferentes tipos de datos sin necesidad de especializarse en un conjunto específico para resolver la tarea encomendada. Para ésto se requiere un paso de estudio de la estructura de los datos en diferentes *surveys* y ver cómo se pueden representar de manera normalizada, enfrentando el problema de mediciones a distintos instantes de tiempo y con vacíos (mediciones no realizadas) en ciertos intervalos.

Para trabajar con el diseño de una red neuronal profunda que aprenda directamente de los datos, se deberá experimentar con un modelo que tenga la capacidad de aprender de secuencias largas de datos, ya que las mediciones de objetos transientes por lo general corresponden a años. Para esto se propone aplicar redes convolucionales [12], que son independientes del largo del objeto en sí, ya que procesan la secuencia a través de ventanas de datos (largo fijo) deslizantes, de manera que aprende la transformación a realizar sobre los datos en esa ventana. Otra forma es procesar a través de una red recurrente [13], modelo propuesto para series de tiempo, pero adaptada a procesar secuencias largas, realizando una estructura jerárquica de niveles de procesamiento, por ejemplo procesar cada 3 meses la secuencia y con estos datos clasificar. El entrenamiento y validación de ese tarea se realizará sobre los datos etiquetados de Kepler (KOI).

Para trabajar con el diseño de una red neuronal no supervisada que aprenda la extracción de características que realiza la *pipeline* de Kepler sin la tarea de detección de exoplanetas asignada, se utilizará la técnica de *Autoencoder* (AE) [14]. Para esto, se requiere codificar la información de la curva de luz (*encoder*) para luego reconstruir la misma curva de luz (*decoder*): esta codificación debiera reflejar las características de la *pipeline* de Kepler. Se forzará ésto asignando una función que premie la imitación de los datos de Kepler, además de reconstruir el dato en sí. Como forma de validar este objetivo se medirá qué tan cercana es la codificación de la curva de luz a los datos reales de la *pipeline* de Kepler, además de ver qué tan efectiva es esta nueva representación aprendida (*encoder*) para la clasificación de exoplanetas comparada con las técnicas del estado del arte.

Para replicar sobre otros *datasets* lo aprendido por los modelos propuestos en los datos de Kepler a (*transfer learning*), se utilizará la información extraída por los modelos propuestos en los objetivos específicos 1 y 2, trabajando con *surveys* de curvas de luz, como *Transiting Exoplanet Survey Satellite* (TESS), *SuperWASP Survey*, *CoRoT mission*, *Kilodegree Extremely Little Telescope* (KELT), *HATNet Exoplanet Survey*. Algunos de estos datos son supervisados (etiquetados) y otros no, por lo que se experimentará con qué tan buena es la clasificación de un modelo o red base (una fracción del modelo) para los datos etiquetados. Mientras que para los *datasets* no etiquetados se visualizará la codificación que genera la red base para verificar si encuentra ciertos comportamiento de grupos de datos, además de verificar de manera cuantitativa con algoritmos de *clustering*, donde se pueda ver que la representación sirve para identificar **tipos** de objetos.

En pos de tener un modelo/algoritmo que genere datos artificiales de las curvas de luz (tránsitos) de cuerpos celestes se utilizará un modelo no supervisado probabilista, similar al comentado anteriormente, conocido como *Variational Autoencoder* (VAE) [15], debido a que es un modelo generativo que resulta invariante a datos similares. Éste tiene el beneficio de que el *decoder* puede ser utilizado como reconstrucción de una variable latente que modelaremos como ciertas características de los cuerpos celeste, y así poder generar curvas de luz de objetos que no se han



logrado estudiar al día de hoy, a través de lo que el modelo de aprendizaje aprendió y logró inferir. La validación se realizará de manera cualitativa en base al comportamiento periódico de los objetos generados *versus* los datos reales con similares características, además de corroborar cuantitativamente esto con una métrica de error. También se propone verificar si la *pipeline* de Kepler genera las mismas características con las que se simuló/generó el objeto (por ejemplo radio del planeta y período de la órbita).

3.5 Trabajo adelantado

Se cuenta con la publicación de “*Refining exoplanet detection using supervised learning and feature engineering*” el cual fue invitado a un *journal* (CLEI-EJ⁴). Ya contamos con los datos de Kepler descargados, además de tener experimentaciones previas sobre los datos. Respecto al *journal*, se está experimentando con ciertas contribuciones de lo formulado en esta propuesta.

3.6 Recursos disponibles

- China-Chile Astronomical Data Center (Chi2AD), afiliado a ChiVO: 860 TB de almacenamiento, 192 núcleos y 1 TB de RAM en total (distribuido en 8 nodos de cómputo).
- Computador Intel i7, 8700K, 6 Núcleos, 32 GB RAM, GPU GTX 1080 TI.
- Laboratorio de Postgrado Departamento de Informática.
- Laboratorio de Postgrado Departamento de Electrónica.

3.7 Plan de trabajo

Categorías	Actividad	2019								2020	
		MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC	ENE	FEB
Implementación de modelo supervisado a cargo de: Margarita Buguño	Revisar estado del arte	X	X	X	X						
	Manejo de Datos		X	X	X						
	Experimentar distintas arquitecturas				X	X	X				
	Validar arquitectura con datos Kepler						X	X	X		
	Preparar paper								X	X	X
Implementación de modelo no supervisado a cargo de: Francisco Mena	Revisar estado del arte	X	X	X	X	X					
	Manejo de Datos		X	X	X						
	Experimentar distintas arquitecturas				X	X	X				
	Derivar nueva función objetivo					X	X				
	Validar arquitectura con datos Kepler							X	X		
	Preparar paper								X	X	X

⁴ <http://www.clei.org/cleij/index.php/cleij>



Investigar modelo base a cargo de: Gabriel Molina	Revisar estado del arte	x	x	x	x	x	x				
	Investigar la fracción a extraer de los modelos supervisado y no-supervisados propuestos					x	x	x	x		
	Validación supervisada de modelos base en otros <i>datasets</i> (K2-TESS-HATNet)							x	x	x	
	Validación no supervisada de modelos base en otros <i>datasets</i> (SuperWasp-KELT-CoRoT)							x	x	x	
	Preparar paper								x	x	x
Generación de datos de cuerpos celestes a cargo de: Patricio Olivares	Revisar estado del arte	x	x	x	x	x					
	Implementar generación de cuerpos celestes para curvas de luz				x	x	x	x			
	Validar de manera cuantitativa y cualitativa a través de datos y la <i>pipeline</i> de Kepler						x	x	x		
	Preparar paper								x	x	x

Referencias

- [1] Mayor, M., & Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature*, 378(6555), 355.
- [2] American Astronomical Society meeting in January 2018. 'The TESS Science Writer's Guide. Online: <https://www.nasa.gov/sites/default/files/atoms/files/tesssciencewritersguidedraft23.pdf>
- [3] Rice, K. (2014). The detection and characterization of extrasolar planets. *Challenges*, 5(2), 296-323.
- [4] Twicken, J. D., Catanzarite, J. H., Clarke, B. D., Girouard, F., Jenkins, J. M., Klaus, T. C., ... & Wohler, B. (2018). Kepler Data Validation I—Architecture, Diagnostic Tests, and Data Products for Vetting Transiting Planet Candidates. *Publications of the Astronomical Society of the Pacific*, 130(988), 064502.
- [5] Donalek, C., Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Drake, A. J., Kumar, A. A., ... & Longo, G. (2013, October). Feature selection strategies for classifying high dimensional astronomical data sets. In *2013 IEEE International Conference on Big Data* (pp. 35-41). IEEE.
- [6] Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., ... & Rischard, M. (2011). On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1),.
- [7] McCauliff, S. D., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., ... & Cote, M. (2015). Automatic classification of Kepler planetary transit candidates. *The Astrophysical Journal*, 806(1), 6.
- [8] Thompson, S. E., Mullally, F., Coughlin, J., Christiansen, J. L., Henze, C. E., Haas, M. R., & Burke, C. J. (2015). A machine learning technique to identify transit shaped signals. *The Astrophysical Journal*, 812(1), 46.
- [9] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436.
- [10] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.
- [11] Margon, B., Ford, H. C., Grandi, S., & Stone, R. P. S. (1979). Enormous periodic Doppler shifts in SS 433. *The Astrophysical Journal*, 233, L63-L68.
- [12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- [13] Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.
- [14] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec).
- [15] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.