

# Clasificación de exoplanetas: desarrollo de una estrategia didáctica para abordar la construcción de modelos observacionales en Física Educativa



ISSN 1870-9095

**Leandro Chinellato<sup>1</sup>, Santiago H. Luna<sup>3,4</sup>, María Sol Pera<sup>1,2</sup>, Gabriel I. Perren<sup>1,5</sup>, Rodrigo Menchón<sup>1,2</sup>, Mauro G. Spagnuolo<sup>3</sup>, Hugo D. Navone<sup>1,2</sup>**

<sup>1</sup>Universidad Nacional de Rosario, Av. Pellegrini 250, Rosario, C.P. 2000, Argentina.

<sup>2</sup>Instituto de Física de Rosario (CONICET-UNR), C.P. 2000, Argentina.

<sup>3</sup>Instituto de Estudios Andinos "Don Pablo Groeber" (UBA-CONICET), C.P. C1428EGA, Ciudad Autónoma de Buenos Aires, Argentina.

<sup>4</sup>Universidad Nacional de Hurlingham, Hurlingham, C.P. 1688, Argentina.

<sup>5</sup>Instituto de Astrofísica de La Plata (UNLP-CONICET), C.P. B1900FWA, La Plata, Argentina.

**E-mail:** hnavone@fceia.unr.edu.ar

(Recibido el 15 de octubre de 2022, aceptado el 30 de noviembre de 2022)

## Resumen

En este trabajo, se presenta una estrategia educativa cuyo propósito central es introducir y explorar la construcción de modelos observacionales en la formación de grado en Física y disciplinas afines. La propuesta recurre al aprendizaje activo basado en el desarrollo de proyectos, asume el carácter distribuido de los procesos cognitivos e integra algunos aspectos transversales de relevancia social. Los tópicos generadores que articulan su desarrollo se basan en la utilización de técnicas de minería de datos de amplio espectro de aplicación, cuya ubicuidad las hace especialmente significativas. Se utiliza como motivación la extracción de conocimientos a partir de un catálogo de exoplanetas, proponiendo la categorización de estos objetos mediante el desarrollo de algoritmos basados en técnicas de aprendizaje automatizado de carácter no supervisado y supervisado. Se concluye que la estrategia es viable y suficientemente accesible y flexible como para ser recreada en distintos escenarios educativos, y que contribuye a promover el desarrollo de competencias requeridas en los procesos de formación académica y en el mundo laboral.

**Palabras clave:** física educativa, física computacional, aprendizaje automático, exoplanetas.

## Abstract

In this work, we present an educational strategy whose main purpose is to present and explore the creation of observational models in the context of a degree in Physics and related fields. This proposal relies on project-based active learning strategies, assumes the distributed character of cognitive processes and incorporates some transversal aspects of social relevance. The topics that articulate this work's development are based on the use of data mining techniques with a wide range of applications, whose ubiquity makes them especially significant. The acquisition of knowledge from an exoplanet catalog is used as motivation, proposing the categorization of these objects through the development of algorithms based on automated learning techniques of both an unsupervised and supervised nature. In conclusion we find that the educational strategy is feasible, and accessible and flexible enough to be recreated in different educational settings, and that it contributes to promote the development of skills required in academic training processes and in the world of work.

**Keywords:** physics education, computational physics, machine learning, exoplanets.

## I. INTRODUCCIÓN

La producción de contenidos relacionados con el diario quehacer del campo científico y tecnológico, y que luego son difundidos por medios masivos de comunicación y redes sociales, ha registrado en los últimos años un importante crecimiento con un notable impacto sociocultural. En nuestro ámbito de trabajo, y a partir de consultas y relevamientos informales, es posible observar que estos contenidos han tenido o tienen alguna influencia

en los intereses vocacionales de los estudiantes que ingresan y cursan las carreras de Profesorado en Física y de Licenciatura en Física de la Universidad Nacional de Rosario (Argentina).

Medios, plataformas y canales traducen, transforman y difunden hallazgos, observaciones, experimentos, métodos, conjeturas y teorías provenientes de múltiples campos disciplinares, científicos y tecnológicos. Este estado de situación genera una multiplicidad de oportunidades que pueden ayudar a enriquecer el trabajo docente, puesto que

permite darle ingreso a temáticas que ya están presentes en las vivencias del colectivo estudiantil, repercutiendo y resonando en intereses, inquietudes y vocaciones.

En estrecha relación con todo esto, los estudiantes perciben que el mundo laboral asociado a las carreras que estudian se expande y complejiza cada vez más, emergiendo áreas de frontera en donde varias disciplinas convergen para poder abordar diversas problemáticas, dando lugar a respuestas y construcciones de carácter multidisciplinar e interdisciplinar. Resulta interesante observar que no se trata de casos aislados, puesto que es muy frecuente encontrar que el desempeño de quienes se gradúan en ciencias se da en campos de trabajo que se alejan cada vez más respecto de la formación inicial de grado, demandando conocimientos y habilidades no previstas en los correspondientes trayectos educativos [1].

En el caso particular de este trabajo, las grandes áreas temáticas que están presentes en los medios de comunicación y en las redes sociales, que nos interesan porque despiertan inquietudes y vocaciones en el estudiantado, y que proponen desafíos en el borde de las disciplinas, son: Astrofísica, Informática y Ciencia de Datos. En nuestras carreras, los saberes de estas áreas temáticas se articulan en el campo problemático de trabajo educativo y disciplinar que se configura en torno a la Física Computacional. Se trata de grandes áreas del conocimiento que contienen en potencia temas generadores y significativos para el desarrollo de actividades de comprensión [2] de interés estratégico para la formación inicial en Física y que, por supuesto, se proyectan hacia el mundo del trabajo.

En los últimos años, debido al auge de las diversas plataformas digitales de intercambio de información y en respuesta a distintos intereses, en donde los de mercado juegan un rol muy importante, proliferan denominaciones como *big data*, *data analytics*, inteligencia artificial, análisis inteligente de datos y *machine learning*, por mencionar sólo algunas. Informática, Ciencias de la Computación, Estadística y Ciencia de Datos, así como sus proyecciones en el seno de otras disciplinas y actividades, serían las grandes áreas del saber desde donde provienen o se ubican estas temáticas. La difusión de sus conocimientos y productos, el uso de sus tecnologías y aplicaciones –ya sea implícito o explícito–, así como su ingreso al lenguaje de todos los días, ofrecen múltiples oportunidades para el trabajo docente en diversas temáticas.

Desde esta perspectiva, el trazado de relaciones, el agrupamiento, la categorización y, en definitiva, la posible extracción de información y de conocimiento a partir de bases de datos, constituye hoy un objetivo de carácter ubicuo que responde a variados intereses. Se trata de cuestiones y problemáticas de gran potencial educativo, puesto que también permiten dar ingreso a reflexiones de carácter ético que pueden estar presentes en el tratamiento de los datos y de la información. Según Mayer-Schönberger y Cukier (2013) [3]:

*Ver el mundo como información, como océanos de datos que pueden explorarse cada vez más lejos y más hondo, nos ofrece un nuevo panorama de la realidad. Es una perspectiva mental que puede penetrar todas las áreas de la vida.*

Mientras que los datos son símbolos que representan, de alguna manera, propiedades de objetos, eventos o personas, la información se constituye en torno a la utilidad de los datos ya procesados, y el conocimiento se traduce en un conjunto organizado de informaciones que pretenden explicar fenómenos, situaciones y problemas. En este contexto, la sabiduría implicaría inevitables reflexiones de carácter ético en torno a diversas opciones de valor [4]. No cabe duda entonces que resulta imprescindible estar advertidos al respecto para poder comprender y reflexionar críticamente sobre estas cuestiones, así como para emprender un trabajo formativo en torno a ellas. Justamente, el carácter ubicuo de las tecnologías de análisis de datos hace que su uso en Física Computacional y Astrofísica no diste mucho de las aplicaciones que involucran otro tipo de intereses y que, por lo tanto, resulte imprescindible no perder de vista todas estas oportunidades de trabajo educativo, aunque se trate de relaciones que en una primera instancia parecieran ser de largo alcance.

Situados ahora en el campo de la Astrofísica: el interés natural que despierta la posibilidad de encontrar otros mundos parecidos al nuestro, junto a los descubrimientos de planetas extrasolares o exoplanetas que se realizan periódicamente y la libre disponibilidad de registros observacionales, son aspectos que configuran una interesante oportunidad para el diseño de estrategias de enseñanza basadas en el análisis de datos.

Desde el descubrimiento del primer exoplaneta, realizado por Michel Mayor y Didier Queloz en 1995 [5], se han confirmado alrededor de 4000 nuevos exoplanetas, la mayoría de ellos gracias al telescopio espacial Kepler. Además, en los próximos años, nuevos dispositivos de observación terrestres y misiones espaciales prometen incrementar significativamente este número.

El espectro de masas, tamaños y parámetros orbitales de los exoplanetas detectados es muy amplio. Abarca desde objetos más grandes que Júpiter orbitando extremadamente cerca de su estrella anfitriona (más próximos que Mercurio del Sol) a planetas del tamaño de la Tierra y posiblemente habitables. Por lo tanto, y a los efectos de poseer un lenguaje común a la hora de estudiarlos, surge la necesidad de clasificarlos y caracterizarlos en grupos y subgrupos. Objetos que también se comparten, como ya mencionamos, con otras áreas del conocimiento e intereses diversos.

En síntesis, resulta posible sumar a la oportunidad educativa que nos brinda el acceso libre a la información relacionada con exoplanetas, la posibilidad de aplicar metodologías de análisis de datos de amplio espectro de aplicación para intentar el agrupamiento y caracterización de estos objetos celestes; al mismo tiempo que se analizan críticamente las cuestiones mencionadas con anterioridad.

Desde el punto de vista de la Física Computacional, tomado como un campo problemático de convergencia de todas estas temáticas y métodos, que se extiende hacia los bordes de nuestra disciplina, y que comparte intereses con todas las áreas del saber mencionadas, es de fundamental importancia trabajar en el diseño de algoritmos y su codificación en lenguajes de programación. Todo esto posibilita el desarrollo de competencias que son requeridas

para transitar la formación inicial en nuestras carreras y, posteriormente, para el trabajo profesional en los nuevos desafíos que plantea el mundo laboral.

En base a todo lo expuesto, en este trabajo se presenta una estrategia didáctica que acerca a los estudiantes al mundo del aprendizaje automático (*machine learning*) proponiéndoles la tarea de clasificar y caracterizar exoplanetas utilizando esta metodología. Se trata de algoritmos que posibilitan la construcción de modelos observacionales y la extracción de conocimientos a partir de los propios datos disponibles. Al respecto, es importante destacar que la propuesta es de carácter educativo y que, por lo tanto, se encuentra a una cierta distancia del trabajo científico en este campo del conocimiento, lo que nos permitirá realizar adecuaciones didácticas y trabajar con flexibilidad sin perder de vista los aspectos disciplinares específicos.

El proceso de construcción de la estrategia didáctica se puso en práctica en el contexto de un programa de investigación-acción participativa desarrollado por la cátedra de Física Computacional, unidad curricular electiva de la carrera de Licenciatura en Física de la Universidad Nacional de Rosario. La construcción metodológica se centró en el desarrollo de un proyecto computacional integrador focalizado en explorar la posibilidad de aplicar técnicas de *machine learning* de carácter general sobre una base de datos de exoplanetas observados. Todo esto teniendo en cuenta que se trata de una temática que naturalmente suscita el interés de los posibles destinatarios debido al inevitable misterio que rodea a la existencia de otros mundos y la eventual posibilidad de explorarlos. A partir de la identificación del problema se seleccionaron los recursos necesarios para abordarlo: una base de datos de exoplanetas y dos algoritmos de aprendizaje automático, uno de clasificación no supervisada (*K-means*) y otro de clasificación supervisada (*árbol de decisión*). De esta manera, se iniciaron los ciclos de acción y reflexión necesarios en todo programa de investigación de este tipo, concluyendo en el diseño de la estrategia didáctica que compartimos en este trabajo. En las próximas secciones presentamos los referentes teóricos utilizados en la planificación del trabajo, las etapas de la estrategia diseñada y los resultados obtenidos durante los ciclos de investigación-acción educativa. La presentación de este estudio constituye una síntesis abierta de todo el proceso realizada en términos de una hipótesis de trabajo para la acción, que se materializa en las etapas de la secuencia didáctica que presentamos. Se trata, entonces, de compartir una experiencia con la intención de brindar orientaciones prácticas para la acción educativa, algo que usualmente le es demandado al campo de la didáctica [6].

Si bien el proceso de investigación-acción educativa y las etapas exploratorias de diseño se implementaron en el contexto de la asignatura Física Computacional del último año de la carrera de Licenciatura en Física de la Universidad Nacional de Rosario, los resultados de la experiencia sugieren que la estrategia didáctica elaborada puede ser utilizada con adecuaciones en diversos escenarios educativos.

## II. REFERENTES TEÓRICOS

La Física Computacional es un campo problemático de trabajo educativo de naturaleza complementaria, y al mismo tiempo transversal, a los tradicionales enfoques teóricos y experimentales de la Física. Está basado en el diseño de algoritmos que se expresan en lenguajes de programación con el propósito de construir modelos computacionales que permitan simular y estudiar objetos, sistemas, procesos y fenómenos de diversa naturaleza y complejidad. En esta tarea, el ordenador desempeña el rol de una *máquina-laboratorio* en donde es posible llevar a cabo experimentos numéricos considerando diversas situaciones de estudio.

En términos generales, en este trabajo asumiremos que un modelo es un constructo de carácter abstracto y provisional, que incorpora en términos de variables sólo los aspectos relevantes de aquello que pretende describir, estableciendo supuestos y haciendo simplificaciones, y utilizando, a veces, entidades que no son directamente observables [7].

A los efectos de facilitar el trabajo didáctico y teniendo en cuenta la metodología empleada en el diseño de los modelos podemos agrupar a estos constructos en tres grandes categorías: modelos analíticos, modelos numéricos y modelos observacionales. Los modelos analíticos se obtienen a partir de primeros principios y su expresión final se establece en términos de operaciones y funciones explícitas y conocidas. Los datos experimentales, en este caso, son utilizados para verificar si el modelo analítico elaborado subyace en ellos y, también, para estimar, aquellos parámetros que no pueden medirse directamente [8]. Cuando no es posible obtener una expresión explícita y cerrada para los modelos analíticos se recurre al cálculo computacional, incorporando el uso de métodos numéricos en su resolución. De esta manera, se construyen modelos numéricos de los sistemas en estudio a los efectos de poder abarcar su complejidad [9].

Muchas veces, sin embargo, sucede que no se conocen o no existen primeros principios a los que se pueda recurrir para construir un modelo analítico y, de ser necesario, su posterior derivación numérica. También puede ocurrir que resulte demasiado engorroso proponer una formulación basada en primeros principios, y que las suposiciones y simplificaciones necesarias en su construcción lleven a obtener resultados poco realistas. Ante estas situaciones, complejas y diversas, nos encontramos con el problema de estimar dependencias estadísticas a partir de los datos; esto es, nos enfrentamos a la necesidad de inferir modelos a partir de los propios registros disponibles [8]. En estos casos es cuando se propone la construcción de modelos observacionales [9], utilizando para ello las tecnologías de análisis de datos disponibles y sus innumerables variantes. Sin perder de vista este marco teórico de referencia, hoy resulta claro que la construcción de modelos observacionales derivados de tecnologías de análisis de datos, mediadas por sistemas y plataformas provistas por las Tecnologías de la Información y de la Comunicación, también impactan en diversos aspectos de nuestras vivencias cotidianas. Su influencia sociocultural, lejos de

ser neutral, afecta tanto a los individuos –en su propia subjetividad y construcciones simbólicas– como al sistema educativo en su conjunto. Las redes sociales, y el procesamiento de datos subyacente en ellas, se nutren constantemente del trazado de relaciones a partir de la construcción de modelos computacionales de carácter observacional. Por lo tanto, la naturaleza sociocultural ubicua de estos constructos hace que resulte imprescindible asumir el carácter situado y, a la vez, distribuido de los procesos de enseñanza, aprendizaje y construcción del conocimiento relacionados con ellos [10, 11, 12]. Estos procesos son situados y distribuidos porque forman parte, y son producto a la vez, de las actividades que se desarrollan, del contexto en donde se encuentran inmersos y, además, de la cultura en la que se despliegan y utilizan [11].

De esta manera, y como ya adelantáramos, actualmente no resulta para nada extraño escuchar menciones a *big data*, *data analytics* y a sistemas que aprenden a partir de los datos (*machine learning*) en diversos ámbitos y situaciones. En consecuencia, es posible aprovechar la presencia significativa de estas temáticas y asumir que se trata de un conocimiento socialmente distribuido para utilizarlo en el desarrollo educativo de actividades de comprensión hacia el interior de nuestras disciplinas. Todo esto sin dejar de advertir la falsa creencia en que la sola presencia de nuevas oportunidades y medios generarán en nuestros estudiantes los procesos de aprendizaje que imaginamos, fenómeno educativo que Perkins identifica y describe como *efecto oportunista* [2]. También, es muy importante observar que el tratamiento de estas temáticas no sólo se dirige hacia el interior de nuestra disciplina, sino que también se proyecta, por su propia naturaleza y en un mismo movimiento, hacia los bordes y hacia el exterior de la misma.

Desde esta perspectiva de trabajo educativo, centrada en una pedagogía y didáctica de la comprensión, en nuestra propuesta asumimos que comprendemos algo cuando no sólo lo sabemos y podemos repetirlo en términos informativos, cuando no sólo podemos ejecutar habilidades de carácter rutinario relacionadas con ese saber, sino cuando realmente podemos pensar y actuar con flexibilidad frente a distintas situaciones y contextos problemáticos a partir del conocimiento que hemos construido [2, 13].

Los referentes teóricos expuestos nos permitieron ir contextualizando adecuadamente cada ciclo del proceso de investigación-acción educativa de carácter exploratorio en el que participamos, para arribar, finalmente, a la propuesta que aquí compartimos en términos de una hipótesis-acción.

### III. ETAPAS DE LA ESTRATEGIA DIDÁCTICA

La estrategia didáctica que aquí proponemos se divide en una serie de etapas o episodios que, a modo de segmentos programados, se van articulando en una narrativa de aula. Partimos realizando una breve introducción histórica en donde se relatan los antecedentes y las motivaciones en torno a nuestro problema. Luego, continuamos describiendo los métodos de detección y observación de exoplanetas, la recopilación de la información en bases de datos y la necesidad de su pre-procesamiento. En la siguiente etapa,

se introducen las tecnologías de análisis de datos y se propone la construcción de los algoritmos necesarios para obtener modelos observacionales del problema en estudio. Finalmente, arribamos a la etapa de reconstrucción, evaluación y síntesis de todo el trabajo realizado.

En las siguientes sub-secciones describimos brevemente cada una de las etapas que logramos construir en términos de hipótesis para la acción, siendo éste el principal resultado de todo el proceso de investigación-acción educativa de carácter exploratorio que hemos desarrollado. Nuestro propósito es brindar los elementos mínimos necesarios para hacer posible la implementación de la propuesta en distintos escenarios educativos. Sin ser taxativos ni exhaustivos al respecto, las indicaciones que se presentan tienen la intención de facilitar las modificaciones y recortes de la estrategia didáctica que se consideren más adecuados a la hora de planificar y de llevar a la práctica intervenciones educativas en situaciones y contextos específicos.

#### A. Perspectiva histórica, antecedentes y motivaciones

En esta etapa de la secuencia, se propone una aproximación a la temática que nos ocupa desde el campo de la Naturaleza de la Ciencia, destacando no sólo los aspectos históricos, sino también la importancia que adquiere en la construcción del conocimiento científico la mera posibilidad de filosofar, hipotetizar y conjeturar, y de hacerlo, en este caso, acerca de la posible existencia de otros mundos. Ya Jenófones (570-475 a. C.) pensaba que existían innumerables astros además del Sol y de la Luna, y Epicuro (341-270 a. C.), en su *Carta a Heródoto*, afirmaba, sin dejar lugar a dudas, que:

*Existe una infinidad de mundos, tanto parecidos como diferentes al nuestro* [14],

para llegar a Giordano Bruno (1548-1600), quien muchas veces cae injustamente en el olvido frente a otras figuras no menos importantes. Bruno tuvo la valentía de desafiar el canon de la época afirmando en voz alta que el hecho de vivir no era privilegio de la humanidad, que se oponía a la conquista de América y que la religión Cristiana no era la única manera de acercarse a Dios; insistiendo a su vez, en la voz de *Elpino*, personaje de su obra *Sobre el infinito universo y los mundos*, que:

*Hay entonces un sinnúmero de soles, y un número infinito de tierras giran alrededor de éstos, como las siete que podemos observar lo hacen alrededor de este Sol cercano a nosotros.* [14]

Cuestiones, todas estas, que le costarían la vida un 17 de febrero de 1600 en Roma por orden de la Santa Inquisición. Desde aquellos tiempos, es posible retomar los aportes de Copérnico, Brahe, Kepler, Galileo y sus vicisitudes, Newton y los “universos-islas” de Wright, Kant y Lambert, para arribar, aunque no finalmente, con Leavitt, Hubble y tantos otros, a nuestra concepción actual del universo, en donde el descubrimiento de otros mundos ya es cosa de todos los días.

En esta etapa, que también podría ser de carácter positivo, consideramos que es conveniente trabajar generando un diálogo participativo con todo el grupo-clase.

Para ello, se pueden utilizar interrogantes que movilicen inquietudes en los participantes y que funcionen como promotores reflexivos: *¿Qué consecuencias tuvo a lo largo de la historia la posible existencia de otros mundos? ¿Qué implicaba en la época de Bruno imaginar otros mundos en el contexto de la disputa entre heliocentrismo y geocentrismo? ¿Qué significa todo esto hoy en esta nueva era exploratoria?*, entre muchos otros posibles. Por supuesto, es esperable que otros interrogantes y derivaciones temáticas vayan surgiendo a partir del diálogo colectivo.

## B. Métodos de observación y obtención de datos

Nuevamente, en términos didácticos y pedagógicos, es muy conveniente partir haciéndonos algunas preguntas. *¿Qué noticias nos llegan hoy acerca de la existencia de otros mundos? ¿Qué sabemos acerca de ellos? y ¿qué entendemos por exoplaneta?*, podrían ser algunos interrogantes que ofician de disparadores temáticos. De esta manera, compartiendo ideas previas y relevando conocimientos socialmente distribuidos, podremos arribar a una definición para aquellos objetos celestes que pueden considerarse exoplanetas o planetas extrasolares en términos más significativos. Un exoplaneta es un cuerpo en órbita alrededor de una estrella anfitriona que no es nuestro Sol y que cumple con los requisitos necesarios para ser considerado como un planeta: en su evolución debe haber “limpiado” sus alrededores de cuerpos menores, ser suficientemente masivo como para que su gravedad supere a las fuerzas del sólido rígido adquiriendo una forma aproximadamente esférica, y orbitar alrededor de una estrella y no de otro cuerpo [5, 15].

*¿Cómo es posible detectar estos objetos?* sería el próximo interrogante en esta secuencia didáctica. Sin entrar en mayores detalles aquí, y sólo a los efectos de brindar breves indicaciones para la posible recreación y ampliación de esta etapa, mencionaremos solo dos de las técnicas existentes para detectar la presencia de estos cuerpos: (1) método de las velocidades radiales y (2) método de los tránsitos.

El método de las velocidades radiales se basa en la medición de la componente de la velocidad en la dirección de la visual del observador y asume que, si una estrella presenta variaciones en esta componente, esto puede deberse a la existencia de un cuerpo que la hace girar en torno al centro de masa del sistema. Estas variaciones se miden a partir del desplazamiento que se registra en los espectros estelares por *efecto Doppler* debido al movimiento de la fuente luminosa. El método de los tránsitos se fundamenta en que, si un planeta pasa frente a una estrella, el astro se ve tenuemente “eclipsado” y su brillo debe sufrir pequeños cambios.

Ambas metodologías son particularmente importantes en el contexto de este trabajo, no sólo por su probado éxito en la detección de exoplanetas, sino porque, además, el método de las velocidades radiales posibilita la estimación de la masa de estos objetos, y la técnica de los tránsitos permite la evaluación aproximada de sus radios. De esta

manera, mediante la combinación de ambas técnicas, es posible inferir la densidad media del cuerpo en cuestión.

## C. Pre-procesamiento de los datos

Para el desarrollo de esta propuesta educativa, los datos se extrajeron del catálogo de exoplanetas que provee el *NASA Exoplanet Science Institute* [16]. Teniendo en cuenta el carácter amplio e introductorio de la estrategia didáctica que presentamos, y a los efectos de simplificar el análisis de los datos, se sugiere no considerar los errores asociados a los valores de los atributos. Mencionamos esto porque no descartamos que esta alternativa pueda ser específicamente incluida en posibles implementaciones de la estrategia, en correspondencia con los contextos de aplicación y con los propósitos y objetivos de aprendizaje planificados por el equipo docente.

Una vez obtenidos los datos observacionales, se implementan las siguientes operaciones de pre-procesamiento o filtrado de los mismos:

1. Sólo se consideran como atributos de cada exoplaneta la identificación o denominación del mismo, su densidad estimada y su radio.
2. Los planetas extrasolares que no poseen valores numéricos en los atributos seleccionados no se tienen en cuenta.
3. También son descartados los cuerpos con valores de densidad mayores a la del hierro ( $7,87 \text{ g/cm}^3$ ). Este criterio nos permite simplificar el análisis de los datos, evitando la complicación innecesaria que, desde un punto de vista didáctico, introduce su consideración en el proceso de clasificación.
4. Finalmente, se normalizan los datos y se les aplica la función logaritmo a ambos atributos numéricos. Estas operaciones de pre-procesamiento se realizan para que los algoritmos de agrupamiento y clasificación trabajen más eficientemente.

Luego de esta etapa de pre-procesamiento, el número de exoplanetas se reduce de aproximadamente unos 4000 a 436, siendo ésta la base de datos o catálogo que se propone utilizar de ahora en más en el desarrollo de la estrategia educativa.

La implementación de esta etapa de la secuencia didáctica implica la elaboración de un algoritmo que realice las operaciones de pre-procesamiento sobre los datos originales. Esta actividad puede ser propuesta como parte de la estrategia educativa o bien puede ser saltada para pasar directamente a trabajar sobre un catálogo ya procesado.

Nuevamente, todas las decisiones relacionadas con la implementación o puesta en práctica de esta propuesta estarán condicionadas por los propósitos y objetivos del equipo docente. En este sentido, consideramos que los conocimientos previos de los posibles destinatarios y el tiempo disponible para el trabajo en taller pueden llegar a ser factores decisivos, entre otras cuestiones asociadas al contexto y a la situación particular de enseñanza y aprendizaje que se requieran tener en cuenta. A partir de estas últimas consideraciones de carácter general, seleccionamos la técnica *K-means* [17, 18] para generar

grupos de exoplanetas de manera no supervisada, por ser éste un algoritmo relativamente sencillo de comprender e implementar. Luego, con el mismo criterio de carácter didáctico, utilizamos árboles de decisión (*Decision Trees* o *DT*) a los efectos de introducir una técnica de clasificación supervisada de uso muy difundido en el campo de la minería de datos [19]. Las etapas que abordan el trabajo con los datos (pre-procesamiento) y la aplicación de los métodos no supervisados y supervisados de clasificación constituyeron el núcleo central sobre el que se desarrollaron los ciclos de investigación-acción educativa que dieron origen a la construcción de la estrategia didáctica. En las subsecciones D y E se describen los algoritmos mencionados y en la sección IV se presentan los resultados de su aplicación obtenidos durante el proceso de investigación-acción participativa implementado.

#### D. K-means

En este episodio de la estrategia didáctica, y a partir de la base de datos construida en la etapa previa, se propone el agrupamiento no supervisado de sus registros en conglomerados o *clusters* mediante la aplicación del algoritmo *K-means* [17, 18].

Para abordar este problema, es necesario especificar en primer lugar la cantidad de conglomerados o *clusters*  $k$  que se desean construir a partir de los  $n$  registros o patrones  $x \in \mathcal{X}$  disponibles en la base de datos. Cada registro representa a un exoplaneta en el espacio de los atributos  $\mathcal{X}$ , donde  $\mathcal{X} \subset R^d$ , siendo  $d$  la dimensión del espacio o cantidad de atributos considerados.

El algoritmo, entonces, tratará de construir  $k$  centros o *centroides*  $c \in \zeta$ , que representan a  $k$  *clusters* o conglomerados, mediante la minimización de una función  $\phi$  que evalúa la distancia entre patrones y centros. Esta función viene dada por:

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \zeta} \|x - c\|^2 \quad (1)$$

A partir de estos *centroides*, es posible construir los agrupamientos o conglomerados asignando a cada registro del catálogo el centro más cercano en el espacio de los atributos.

Es interesante notar que resolver exactamente el problema planteado es extremadamente difícil, puesto que existen  $k^n$  formas de dividir las  $n$  observaciones en  $k$  grupos. No obstante, gracias a la técnica de agrupamiento *K-means* contamos con una metodología simple y rápida para la resolución aproximada de este problema [18].

En el contexto de esta propuesta educativa, el algoritmo *K-means* se introduce formalizando y describiendo la siguiente secuencia de acciones que lo definen:

1. Tomando puntos aleatorios en el espacio de los datos  $\mathcal{X}$  selecciona arbitrariamente  $k$  *centroides*  $\zeta = \{c_1, c_2, \dots, c_k\}$ .

2. Para cada  $i \in \{1, \dots, k\}$ , genera el conglomerado o *cluster*  $C_i$ , que es el conjunto de patrones  $x$  en  $\mathcal{X}$  que están más cerca de  $c_j$  para todo  $j \neq i$ .

3. Para cada  $i \in \{1, \dots, k\}$ , establece un nuevo centroide  $c_i$  como el “centro de masa” de todos los puntos  $x$  en cada *cluster*  $C_i$ ,

$$c_i = \frac{1}{\#(C_i)} \sum_{x \in C_i} x$$

Siendo  $\#(C_i)$  el número de elementos o patrones asociados al conglomerado  $C_i$ .

4. Se repiten los pasos 2 y 3 de esta secuencia hasta que el cambio en  $\zeta$  resulte menor a un valor de tolerancia.

Como el proceso de *clustering* implementado mediante el *K-means* es no supervisado, se hace necesario realizar una validación interna de los grupos o conglomerados construidos. Para hacer esto, existen diversos coeficientes o *scores*; en este trabajo, se optó por utilizar el *coeficiente de silueta* (*silhouette*) [19]. Este indicador se define para cada objeto y se calcula de la siguiente manera:

$$s = \frac{b - a}{\max(a, b)} \quad (2)$$

donde  $a$  es la distancia media entre un objeto y todos los demás puntos o patrones en la misma clase (conglomerado, *cluster* o grupo) y  $b$  es la distancia media entre un objeto y todos los demás puntos en el siguiente grupo más cercano.

El puntaje asignado a un conjunto de objetos estará dado por la media del coeficiente  $s$  de cada objeto. Un valor más alto del *coeficiente de silueta* se relaciona con un modelo que posee agrupamientos mejor definidos.

Para aplicar el método *K-means* a la base de datos de exoplanetas ya pre-procesada se propone el diseño de un algoritmo codificado en lenguaje *Python* que invoca a la librería *Scikit-learn*, cuya descripción puede consultarse en [19] o en diversos foros y textos de acceso libre. En este trabajo no entraremos en mayores detalles al respecto, puesto que la comunidad de práctica constituida en torno a *Python* y sus aplicaciones en *machine learning* es muy amplia, prolífica y abierta a la resolución de consultas; cuestión que se desea promover en este tipo de abordajes, dando cuenta, a su vez, del carácter distribuido del conocimiento.

El programa que se propone desarrollar debe tomar como datos de entrada el catálogo de exoplanetas y el número máximo de *clusters*  $k$  a considerar, para proceder a la optimización del número de agrupamientos haciendo uso del *coeficiente de silueta*. Como resultado, se espera que el código genere un gráfico de dispersión con los conglomerados que optimizan el mencionado coeficiente, presentando a los datos de entrada etiquetados con su pertenencia –o membresía– a un dado agrupamiento. Se debe consignar, además, el valor promedio y el desvío

estándar de los atributos asociados al conjunto de miembros de cada una de las clases construidas.

### E. Árboles de decisión

Los árboles de decisión (*DT*) son un método de aprendizaje supervisado, no paramétrico, utilizado tanto en problemas de clasificación como en problemas de regresión. El objetivo de esta metodología de análisis de datos es la construcción de un modelo computacional que permita predecir el valor de una variable objetivo mediante la aplicación sucesiva de reglas de decisión simples que se infieren a partir de un conjunto de observaciones.

Cada etapa en donde se toma una decisión se denomina *nodo*. En cada nodo se produce una separación o no del conjunto de datos, de acuerdo a la aplicación de una regla de decisión sobre variables predictoras. Una de las formas más usuales de medir la *impureza* de un árbol de decisión consiste en calcular la entropía  $H$  de cada nodo [18]. El propósito del algoritmo es tratar de homogeneizar cada nodo intentando anular la *entropía* del mismo, cantidad que se define como:

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk}), \quad (3)$$

donde  $X_m$  simboliza al conjunto de objetos en cada nodo y  $p_{mk}$  es la probabilidad de los posibles valores de las clases correspondientes a los objetos en el nodo.

Para poder cuantificar la *exactitud* del modelo, es decir, conocer qué tan bien se predice el valor de la variable objetivo mediante la aplicación de las reglas previamente inferidas, es necesario validarlo con datos no utilizados en su construcción. Para hacer esto, se seleccionan aleatoriamente un cierto número de objetos para realizar el proceso de entrenamiento o aprendizaje del modelo, y otro para su posterior validación. Por supuesto, es necesario contar con una función matemática o indicador que compare los valores predichos por las reglas de decisión con los valores reales o etiquetas asociadas a los mismos. En esta propuesta, se utiliza el *Accuracy Classification Score (ACS)*.

Sea  $\hat{y}_i$  el  $i$ -ésimo valor predicho e  $y_i$  el correspondiente valor real o conocido, entonces la fracción de predicciones correctas sobre las  $n_s$  predicciones hechas, o *exactitud*, se define como:

$$ACS(y_i, \hat{y}_i) = \frac{1}{n_s} \sum_{i=0}^{n_s-1} I(\hat{y}_i = y_i), \quad (4)$$

donde  $I(x)$  es la *función indicatriz*, dada por:

$$I(x) = \begin{cases} 1 & \text{si } \hat{y}_i = y_i \\ 0 & \text{si } \hat{y}_i \neq y_i \end{cases}. \quad (5)$$

Si el modelo predice correctamente todos los datos de validación, el *score* tomará el valor uno; mientras que, en el

extremo opuesto, un valor cero indicará una *exactitud* nula. Análogamente, si definimos la *función indicatriz* de manera que evalúe los desaciertos, se obtiene el *error promedio* de clasificación, que es una medida complementaria a la anterior [18].

La librería *scikit-learn* provee un algoritmo de clasificación basado en árboles de decisión denominado *Decision Tree Classifier (DTC)* [19]. Al igual que con otros clasificadores, el *DTC* toma como entrada dos matrices: una matriz  $X$  de tamaño  $[n_s, n_f]$  que contiene los objetos de entrenamiento, y una matriz  $Y$  de tamaño  $[n_s]$ , que contiene las etiquetas de clase para los objetos de entrenamiento.

En esta propuesta didáctica, se utilizan como atributos o *features* la densidad y el radio de los exoplanetas, de acuerdo con los datos provistos en el catálogo luego de que el mismo fuera pre-procesado en la segunda etapa de la secuencia de trabajo.

### F. Evaluación y síntesis

Como último episodio de la secuencia, se propone la reconstrucción, evaluación y síntesis de todo lo desarrollado. Hay diversas estrategias educativas que permiten abordar esta etapa, siendo algunas de ellas: 1) la elaboración de un informe en formato de artículo para presentar en un congreso o enviar a una revista; 2) la presentación de los resultados en formato póster para su exposición en una jornada o reunión académica; 3) la redacción de una monografía como informe de investigación, en donde se profundizan un poco más los distintos aspectos de la temática; 4) el desarrollo de un seminario [20] destinado a estudiantes, promoviendo la discusión y la evaluación entre pares, en donde se presenta toda la temática, desde los aspectos históricos, antecedentes y motivaciones, hasta los resultados obtenidos, alcances y posibles derivaciones o, 5) la planificación y puesta en práctica de la estrategia educativa –o de algunos aspectos de la misma– en formato de micro clase [21], cuyos destinatarios sean los alumnos de alguna unidad curricular previamente seleccionada o de la misma asignatura en donde se desarrolla la experiencia. Estas últimas dos modalidades son particularmente importantes y requeridas en los procesos de formación inicial docente.

En el contexto de esta propuesta, si bien consideramos que todas las estrategias de integración y de síntesis expuestas son muy adecuadas, pensamos que aquellas que promueven la discusión entre pares constituyen el dispositivo de evaluación que más puede aportar en términos de desarrollo de competencias en los futuros profesionales. Por supuesto, todo esto teniendo en cuenta las características específicas de los escenarios educativos en donde se intente recrear esta secuencia didáctica.

## IV. RESULTADOS DE LA PUESTA PRÁCTICA

La estrategia educativa que hemos diseñado, cuyos principales aspectos presentamos en este artículo, se

construyó a partir de un proceso de investigación-acción educativa desarrollado en el contexto del espacio curricular *Física Computacional*, cuya modalidad pedagógica es taller. Esta unidad es de carácter electivo y pertenece al quinto año de la carrera de Licenciatura en Física de la Universidad Nacional de Rosario (Argentina).

En sucesivos ciclos de planificación, acción y reflexión, se abordaron las etapas C, D y E presentadas anteriormente, que involucraron la obtención y pre-procesamiento de la base de datos, la aplicación del algoritmo *K-means* (aprendizaje no supervisado) y la utilización de la técnica de árboles de decisión (aprendizaje supervisado). El propósito central de este abordaje fue explorar la viabilidad de una propuesta educativa que aborde estas temáticas, y la identificación de sus posibles obstáculos y derivaciones, así como el dimensionamiento de sus alcances generales en términos pedagógicos y didácticos, tratando de visualizar su proyección hacia distintas problemáticas y escenarios de actuación.

Las tareas de diseño, implementación y evaluación de los algoritmos fueron llevadas adelante por un estudiante en el contexto de un proceso de aprendizaje activo basado en el desarrollo de un proyecto computacional focalizado en la problemática planteada; acompañado por pares, docentes y especialistas en la temática. Desde el punto de vista del proceso de indagación implementado, el trabajo realizado puede pensarse como un estudio de caso único desarrollado en el marco de un esquema de investigación-acción educativa, de carácter exploratorio y cualitativo, que produjo como resultado una estrategia didáctica que se constituye en una hipótesis para la acción [22].

Con la intención de facilitar la posible recreación de la secuencia didáctica en distintos contextos educativos, en las siguientes sub-secciones se presentan los resultados obtenidos por el estudiante durante la puesta en práctica del proceso de investigación-acción participativa.

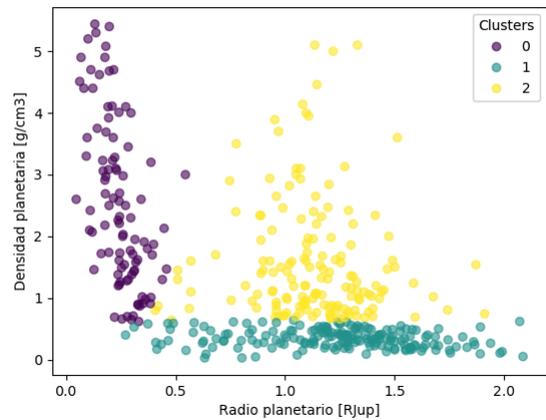
### A. Resultados obtenidos con *K-means*

El procesamiento de la base de datos de exoplanetas mediante el algoritmo *K-means* y el *score silhouette*, tal como fuera descrito anteriormente, produce como resultado tres agrupamientos o conglomerados (*clusters*).

En la Figura 1 se presentan los miembros que componen cada uno de los tres grupos construidos por el algoritmo en el espacio de los atributos e identificados por color.

En base a lo hallado y tomando en cuenta otras clasificaciones y nombres propuestos en la bibliografía especializada sobre la temática, en la Tabla I se especifican las categorías de clasificación asignadas a los exoplanetas de la base de datos utilizada. Al respecto, es importante no perder de vista que estamos presentando una propuesta de carácter educativo y que, por lo tanto, los conocimientos científicos específicos y al día sólo entran en juego como referentes disciplinares últimos a ser tenidos en cuenta en términos de cultura académica.

En la tabla I se consignan los valores promedios y las desviaciones estándares de los atributos que definen a cada una de las categorías.



**FIGURA 1.** Densidad vs. radio de los exoplanetas clasificados por el algoritmo *K-means*. Cada exoplaneta se identifica con un punto en el espacio de atributos y los colores indican la pertenencia o membresía a un dado grupo o clase.

**TABLA I.** Parámetros de los tres conglomerados encontrados.

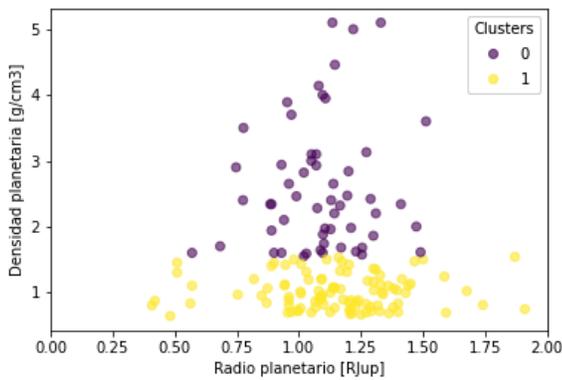
Nombres propuestos (cluster)	Cantidad de miembros	Densidad promedio [ $g/cm^3$ ]	Radio promedio [ $R_{Jup}$ ]
Exotierra (cluster 0)	98	$2 \pm 1$	$0,25 \pm 0,09$
Exojúpiter (cluster 1)	182	$0,4 \pm 0,2$	$1,2 \pm 0,4$
Planeta de Transición (cluster 2)	156	$1,6 \pm 0,9$	$1,1 \pm 0,3$

Los planetas del agrupamiento denominado «Exotiemras» (*cluster 0*) presentan un radio y una densidad promedio cercanos a los de nuestro planeta u otros planetas rocosos conocidos en nuestro sistema solar. De la misma manera, los «Exojúpiter» (*cluster 1*) comparten densidades similares a las de Júpiter y Saturno, con radios en el intervalo de  $0,25 R_{Jup}$  a  $2 R_{Jup}$ . Mientras que los denominados «Planetas de Transición» (*cluster 2*), son los que presentan radios en el orden del de Júpiter, pero caracterizados por una alta dispersión en las densidades.

Esta particularidad que se presenta en los *Planetas de Transición* sugiere la necesidad de una sub-clasificación dentro de este grupo. Motivados por lo anteriormente expuesto, se propone una nueva ejecución del algoritmo de *clustering*, pero ahora sólo sobre este último conglomerado. Los resultados obtenidos se muestran en la Figura 2.

Como se observa en la figura, se obtienen dos agrupamientos claramente diferenciados. Esto nos permite proponer nuevas categorías de clasificación teniendo en cuenta ahora las características de los dos subgrupos que conforman el conglomerado de origen que habíamos denominado «Planetas de Transición».

En la Tabla II se presentan los resultados obtenidos, consignando la cantidad de objetos en cada grupo y los promedios y desviaciones que los caracterizan.



**FIGURA 2.** Densidad vs. radio de los exoplanetas del grupo «Planetas de Transición» clasificados por el algoritmo *K-means*. Los colores representan los subgrupos obtenidos.

**TABLA II.** Parámetros de los subgrupos encontrados.

Nombres propuestos (cluster)	Cantidad de miembros	Densidad promedio [ $g/cm^3$ ]	Radio promedio [ $R_{Jup}$ ]
Júpiter Denso (cluster 0)	56	$3 \pm 1$	$1,0 \pm 0,2$
Gigante de Hielo (cluster 1)	100	$1,1 \pm 0,2$	$1,1 \pm 0,3$

Se observa que el grupo denominado «Júpiter Denso» (cluster 0) presenta densidades acordes a planetas rocosos pero con radios del orden del radio de Júpiter, por lo que también podrían llamarse «Supertierras», término que suele usarse en la literatura. En el caso de los denominados «Gigantes de Hielo» (cluster 1), las densidades resultan ser similares a las de Neptuno o Urano –nombrados en la literatura como *Gigantes de Hielo*–, mientras que los radios caen en un intervalo de aproximadamente  $0,4 R_{Jup}$  a  $2 R_{Jup}$ , siendo el radio de Neptuno y Urano de aproximadamente  $0,35 R_{Jup}$ .

**B. Resultados obtenidos con árboles de decisión**

El desarrollo de un árbol de decisión con el propósito de extraer reglas explícitas que permitan la clasificación de los exoplanetas de nuestro catálogo requiere contar con un archivo de datos etiquetados según los grupos de pertenencia, ya que se trata de un método de aprendizaje supervisado.

A su vez, y tal como adelantáramos previamente, es necesario disponer de un conjunto de datos de entrenamiento para el aprendizaje de las reglas y otro de validación a los efectos de poder constatar que efectivamente se produce tal “aprendizaje”.

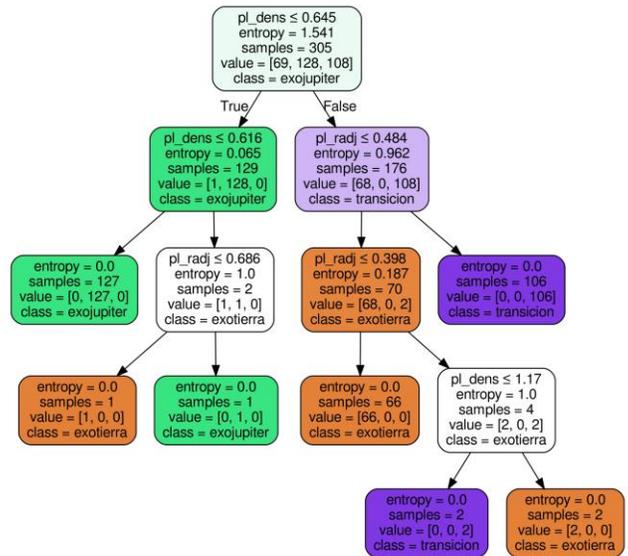
En nuestro caso, se toma como entrada el archivo de datos en donde cada exoplaneta está etiquetado con su grupo de pertenencia de acuerdo con *K-means*. El 70% de los registros se utiliza en el proceso de entrenamiento y el 30% restante se usa para validación.

Una vez generado el árbol de decisión (DT), se calcula la entropía de cada nodo utilizando los datos de validación.

El código entrega un gráfico del DT con el valor de la entropía y el número de objetos correspondientes a cada nodo.

Utilizando la librería *Graphviz* [19], es posible representar en forma gráfica el árbol de decisión generado por el *Decision Tree Classifier*, explicitando, al mismo tiempo, las reglas de clasificación correspondientes a cada nodo, el valor de la entropía y la cantidad de patrones asociados a los mismos.

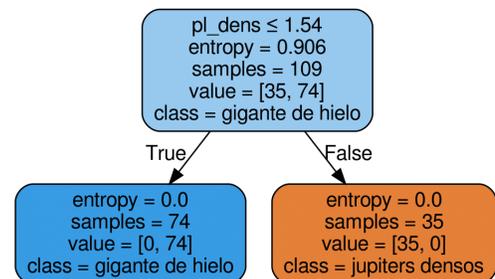
Los resultados obtenidos en esta etapa se muestran en la Figura 3.



**FIGURA 3.** Árbol de decisión obtenido con el algoritmo *Decision Tree Classifier*, graficado usando la librería *Graphviz*. En cada nodo se especifica la regla de decisión, el valor de la entropía y la cantidad de objetos involucrados.

A partir del gráfico es posible deducir que las reglas obtenidas con el *Decision Tree Classifier* discriminan correctamente a cada grupo, logrando hacer nula la entropía de los nodos de la última capa.

Finalmente, se aplica el mismo algoritmo para generar reglas de decisión correspondientes a los dos subgrupos pertenecientes al conglomerado denominado «Planetas de Transición», obteniéndose el árbol que se presenta en la Figura 4.



**FIGURA 4.** Árbol de decisión correspondiente al grupo «Planetas de Transición».

## V. SÍNTESIS FINAL Y CONCLUSIONES

En este trabajo, presentamos una estrategia educativa cuyo propósito central es introducir en la formación de grado en Física la construcción de modelos observacionales a partir de los propios datos que produce un sistema, fenómeno o problemática en estudio. En correspondencia con el sentido que establece Perkins [2], consideramos que se trata de un tema generador por: (1) las implicancias que tiene; (2) su importancia o centralidad en la formación de los futuros profesionales; (3) su accesibilidad, ya que promueve el desarrollo de procesos comprensivos de amplio espectro, y (4) su ubicuidad y riqueza, ya que permite la realización de conexiones y extrapolaciones hacia diversos campos del conocimiento y del trabajo.

La propuesta se basa en una pedagogía orientada a la comprensión, que reconoce el carácter distribuido de los procesos cognitivos y, desde este marco, hace uso de mediaciones: de carácter físico, a través del trabajo con ordenadores; de carácter social, a partir de intercambios de todo tipo en el contexto de comunidades de pares, de aprendizaje, de enseñanza y de práctica; y de carácter simbólico, apelando a distintos lenguajes formales, comunicaciones, constructos, gráficos, tablas y representaciones [2].

También, apela al uso socialmente responsable de los conocimientos [23] al advertir acerca de la necesidad de reflexionar críticamente sobre cuestiones de carácter ético relacionadas con el manejo de la información. Y, en este sentido, para lograr mayores niveles de responsabilidad social, articula contenidos cognitivos desde una perspectiva de carácter ético sin perder de vista lo emocional [24], reconociendo conocimientos previos, motivaciones, intereses, inquietudes, dudas y vocaciones, entre otras cuestiones. Por supuesto, algunos de los contenidos que se trabajan son de naturaleza transversal, ya que no pertenecen a un área curricular específica y, además, son demandados socialmente. En estos casos, lo que se pone en juego es la *relación transversal-disciplinar* [25], cuestión teórico-práctica que la mayoría de las veces es desatendida.

La puesta en práctica de la propuesta de enseñanza bajo el formato de aprendizaje activo basado en el desarrollo de un proyecto-problema en el contexto de un programa de investigación-acción participativa, nos permitió no sólo evaluar su riqueza formativa, sus alcances y viabilidad, sino también dimensionarla adecuadamente para su comunicación en términos de estrategia, prototipo o experiencia a considerar para ser recreada en diversos escenarios de actuación educativa, cuestión que intentamos abordar en este artículo.

La construcción de modelos observacionales a partir de los datos se realizó utilizando técnicas del área de aprendizaje automatizado o *machine learning*, tomando un problema de interés actual, y que ha inquietado a la humanidad durante toda su existencia. Se trata de la posibilidad cierta de encontrar otros mundos, con las connotaciones que esto adquiere al considerar la presencia de otras vidas. Específicamente, se trabajó con una base de datos de exoplanetas de acceso libre y, a partir de ella, se construyó un catálogo propio respondiendo a criterios

pedagógicos y didácticos. Luego, se implementó un algoritmo de agrupamiento o *clustering* no supervisado, basado en la técnica *K-means*, a los efectos de organizar los datos, construyendo categorías y extrayendo conocimiento a partir de los mismos. Esto permitió distinguir, en primera instancia, tres grupos: *Exotierras* (planetas similares a la Tierra), *Exojúpiters* (planetas similares a Júpiter) y *Planetas de Transición* (Tablas I); y luego, dividir a esta última categoría en dos sub-categorías: *Júpiters Densos* (planetas densos de gran tamaño o Supertierras) y *Gigantes de Hielo*, similares a Neptuno o Urano (Tabla II).

Como última etapa de procesamiento, y con el propósito de hacer explícitas las reglas de clasificación, se desarrolló un algoritmo que hace uso de un método de aprendizaje supervisado basado en árboles de decisión y se lo aplicó al catálogo de exoplanetas, ya agrupados en categorías mediante *K-means*.

Es muy importante destacar que el sistema de clasificación generado no es definitivo, puesto que, por propia naturaleza, puede actualizarse aprendiendo a partir de la experiencia utilizando nuevos descubrimientos y datos, que seguramente se irán obteniendo gracias a la puesta en servicio de nuevos telescopios espaciales y otras tecnologías de última generación.

La experiencia desarrollada en todo el proceso de investigación-acción educativa ha posibilitado la construcción de competencias en el diseño de algoritmos y su codificación en lenguajes formales; la incorporación de conceptos y habilidades relacionadas con técnicas de aprendizaje automatizado y minería de datos; el trazado de posibles relaciones con otras áreas del conocimiento y del mundo del trabajo, así como con problemáticas vinculadas al uso responsable del conocimiento, y ha favorecido el intercambio entre pares, docentes y profesionales, todo esto en el contexto de comunidades de aprendizaje, de enseñanza y de práctica.

El diseño de la estrategia didáctica basado en la articulación de segmentos programados (etapas) facilita su adecuación a distintas situaciones y escenarios educativos. En este sentido, a partir de la experiencia recabada nos es posible brindar una serie de orientaciones para la acción educativa que pueden ayudar en diversas implementaciones de acuerdo con los objetivos educativos que fije el equipo docente: 1) en relación al pre-procesamiento de la base de datos, esta etapa puede no incluirse y utilizarse directamente la base de datos ya depurada por el equipo docente; 2) respecto de los algoritmos de *machine learning*, es posible recurrir a una versión mínima de la estrategia didáctica utilizando solamente la técnica de *K-means*, si así se lo considera conveniente; 3) por supuesto, también es posible reemplazar o complementar las técnicas utilizadas con otras de igual disponibilidad; 4) la estrategia educativa puede plantearse como un proyecto computacional a ser desarrollado con cierta autonomía, tanto en forma individual como grupal, ya sea durante los encuentros en formato taller o bien como un trabajo práctico a resolver para ser presentado oportunamente; 5) si bien nos alejamos de nuestros propósitos iniciales, cumplimos en indicar que también puede implementarse en forma expositiva y demostrativa, en donde el equipo docente desarrolla todos

los contenidos para el grupo-clase; 6) la primera y segunda etapa de la secuencia, en donde se trabajan contenidos de Naturaleza de la Ciencia y se describen los métodos de observación, pueden trabajarse de manera introductoria, o bien, profundizar aún más en sus diversos aspectos. En relación al tiempo de desarrollo requerido, suponiendo que se cuenta con un laboratorio de informática con 15 puestos de trabajo y que la estrategia se implementa completamente en tiempo de taller en donde los participantes desarrollan los algoritmos con la asistencia del equipo docente, se estima que se necesitarían dos encuentros de 4 horas reloj cada uno. Durante el primer encuentro podrían abordarse las etapas A, B y C de la secuencia, y durante el segundo encuentro la etapa D, en donde aquí incluimos la preparación de los datos de entrada categorizados en la etapa previa para generar los árboles de decisión. La etapa final de síntesis, en este tipo de implementación puede consistir en un diálogo entre todos los participantes evaluando la experiencia desarrollada o bien puede concluir con la elaboración de un informe final -individual o grupal- a ser presentado como trabajo práctico de síntesis. Si la propuesta se plantea como proyecto computacional individual a ser desarrollado autónomamente, aquí los tiempos son mucho más flexibles y dependen directamente del trabajo de los estudiantes involucrados. La estrategia educativa elaborada también puede ser utilizada como parte del proceso de acreditación de un espacio curricular; en este último caso, el equipo docente debe establecer claramente la consigna de trabajo.

En términos de posibles escenarios educativos de aplicación, consideramos que esta propuesta puede ser recreada y adecuada para su integración en distintos espacios curriculares vinculados al área de Física Computacional, Astrofísica, Computación Científica e Informática, ya sea en el contexto de las carreras de formación de grado en Física (Licenciatura en Física y Profesorado en Física), o en otros campos del conocimiento en donde esta estrategia adquiera algún sentido o sea de utilidad práctica.

## AGRADECIMIENTOS

Este trabajo fue realizado en el contexto de un proyecto de investigación en Física Educativa radicado en la Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la Universidad Nacional de Rosario (Argentina).

## REFERENCIAS

- [1] Landau, R., *Computational Physics: A Better Model for Physics Education?* Computing in Science & Engineering, 22-30 (2006).
- [2] Perkins, D., *La escuela inteligente* (Gedisa, Barcelona, 2003).
- [3] Minniti, D. y J. J. Clariá, *Nuevos Mundos: Hacia la búsqueda de otras Tierras* (Comunicarte, Córdoba, 2014).
- [4] Mayer-Schönberger, V. y Cukier, K., *Big data: la revolución de los datos masivos* (Turner Publicaciones S.L., Madrid, 2013).
- [5] Pérez Gómez, A., *Pedagogías para tiempos de perplejidad: De la información a la sabiduría* (Homo Sapiens, Rosario, 2017).
- [6] Steinman, J., *Más didáctica en la educación superior* (Miño y Dávila, Buenos Aires, 2012).
- [7] Lombardi, O., *Los modelos como mediadores entre teoría y realidad*, en Galagovsky (coord.), *Didáctica de las Ciencias Naturales: el caso de los modelos científicos* (Lugar, Buenos Aires, 2010).
- [8] Cherkassky, V. y Mulier, F., *Learning from data: Concepts, Theory, and Methods* (John Wiley & Sons, NY, 1998).
- [9] Gershenfeld, N., *The Nature of Mathematical Modeling*, (Cambridge University Press, Cambridge, 1999).
- [10] Brown, J.S., Collins, A. y Duguid, P., *Situated cognition and the culture of learning*, Educational Researcher, 32-42 (1989).
- [11] Díaz Barriga, F., *Enseñanza situada: vínculo entre la escuela y la vida* (McGraw-Hill, México, 2006).
- [12] Litwin, E., *Las configuraciones didácticas: una nueva agenda para la enseñanza superior* (Paidós, Buenos Aires, 2008).
- [13] Perkins, D., *El aprendizaje pleno* (Paidós, Buenos Aires, 2016).
- [14] Gangui, A., *El big bang: la génesis de nuestra cosmología actual* (Eudeba, Buenos Aires, 2005).
- [15] IAU (B5), *Definition of a Planet in the Solar System*. <[https://www.iau.org/static/resolutions/Resolution\\_GA26-5-6.pdf](https://www.iau.org/static/resolutions/Resolution_GA26-5-6.pdf)>, consultado el 14 de julio de 2022.
- [16] NASA Exoplanet Science Institute (NExSci, IPAC), <<https://exoplanetarchive.ipac.caltech.edu/docs/data.html>>, consultado el 23 de julio de 2020.
- [17] National Research Council (NRC), *Frontiers in massive data analysis* (The National Academies Press, Washington D.C., 2013).
- [18] Gareth, J., Witten, D., Hastie, T. y Tibshirani, R., *An introduction to statistical learning with applications in R* (Nueva York, Springer, 2014).
- [19] Albon, C., *Machine learning with Python cookbook: practical solutions from preprocessing to deep learning* (Sebastopol, O'Reilly, 2018).
- [20] Davini, M.C., *La formación en la práctica docente* (Paidós, Ciudad Autónoma de Buenos Aires, 2015).
- [21] Anijovich, R., Cappelletti, G., Mora, S. y Sabelli, J.J., *Transitar la formación pedagógica: dispositivos y estrategias* (Buenos Aires, Paidós, 2014).
- [22] Hernández Sampieri, R. y Mendoza Torres, C. P., *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta* (México, Mc Graw Hill, 2018).
- [23] Tedesco, J.C., Aberbuj, C. y Zacarias, I., *Pedagogía y democratización de la universidad* (Aique, Ciudad Autónoma de Buenos Aires, 2014).
- [24] Tedesco, J. C., *Debates de política educativa* (Paidós, Ciudad Autónoma de Buenos Aires, 2016).
- [25] Cullen, A., *Crítica de las razones de educar: temas de filosofía de la educación* (Paidós, Buenos Aires, 2008).