GRADO EN ECONOMÍA

CURSO ACADÉMICO 2023/2024

TRABAJO FIN DE GRADO

Análisis y aplicación de técnicas de Machine Learning: un enfoque comparativo y práctico

Analysis and application of Machine Learning techniques: a comparative and practical approach

AUTOR

MIGUEL KOUVCHINOVE VORONINE

DIRECTOR

PEDRO SOLANA GONZALEZ

10 de junio de 2024

Resumen:

El presente trabajo de fin de grado se centra en aportar una visión general de las diferentes técnicas de análisis que existen en el campo del Machine Learning, destacando la relevancia y aplicabilidad en diversos contextos, y haciendo énfasis en las técnicas de aprendizaje supervisado (regresión y clasificación) y no supervisado (clustering y reducción de la dimensionalidad), aunque mencionando brevemente otras como el aprendizaje semi-supervisado, por refuerzo, así como las redes neuronales y el Deep Learning, explicando brevemente sus fundamentos y aplicaciones potenciales. Se hace un análisis de los diferentes tipos de algoritmos que existen en las distintas técnicas de aprendizaje, empleando algunas de estas para un análisis empírico utilizando los programas como R Studio y Weka con el fin de ilustrar la aplicación práctica de estas técnicas.

Palabras clave: Machine Learning, economía, aprendizaje supervisado, aprendizaje no supervisado, algoritmos.

Abstract:

The tesis focuses on providing an overview of different analysis techniques in the Machine Learning field, highlighting their relevance and appropriateness in various contexts, giving a special emphasis in supervised learning techniques (regression and classification) and unsupervised learning techniques (clustering and dimensionality reduction), while briefly mentioning other techniques such as semi-supervised learning, reinforcement learning, as well as neural networks and deep learning, briefly explaining their fundamentals and potential applications. An analysis of the different types of algorithms, belonging to different types of learning, is carried out using programs such as R Studio and Weka in order to demonstrate the practical application of these techniques.

Keywords: Machine Learning, economics, supervised learning, non-supervised learning, algorithms.

ÍNDICE

1. INTRODUCCIÓN

La inteligencia artificial está teniendo un gran auge en los últimos tiempos (Pulido, 2023) y una rápida evolución, que hace que muchas veces se utilicen conceptos como el de Machine Learning, Deep Learning y Data Mining como si fueran sinónimos, aunque se refieran a cosas diferentes.

La primera definición del Machine Learning (ML) se introdujo en 1959, estaba aplicado a una partida de damas en la que jugaba una máquina, y básicamente se programaba a un ordenador con el objetivo de que aprendiese a jugar a un juego, mejor que la persona por la que había sido programado (Samuel, 1959). Este concepto se fue desarrollando a posteriori, surgiendo diversas definiciones de ML donde se puede ver definido como "*mecanismo o conjunto de técnicas, dentro del ámbito de la inteligencia artificial, que utiliza métodos estadísticos para la búsqueda de patrones a partir de los cuales construimos maquinas inteligentes, capaces de aprender y tomar decisiones en base a datos empíricos obtenidos desde diversas fuentes de datos, como sensores, bases de datos, …"* (Álogos, 2018), simplemente como métodos computacionales que utilizan la experiencia para mejorar el rendimiento o para hacer predicciones precisas (Mohri, et al., 2018).

Aun así los conceptos relacionados con nuevas tecnologías siempre generan confusión, y por tanto es importante distinguir bien que significan los conceptos, como puede ser el de Inteligencia Artificial (IA) definida como "*la habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana*" (Rouhiainen, 2018), el de Deep Learning, una de las ramas del ML que consiste en utilizar un método jerárquico donde en cada paso se transforma la información del paso anterior en representaciones más complejas de datos (Storm, et al., 2020), u otros como el de Data Mining (DM), que se trata del proceso de extracción de correlaciones o patrones dentro de grandes bases de datos (Nájera & Calleja, 2018).

Uno de los problemas más grandes relacionados con el uso de las herramientas de IA es la preocupación creciente por la privacidad de los datos (Sharifani & Amini, 2023), ya que los datos de la gente se encuentran en una situación de debilidad (Albornoz, 2021) debido a que muchos de los modelos de IA son vulnerables a ataques informáticos (Oseni, et al., 2020).

Los problemas del planeta tienen una gran complejidad, y por eso utilizar ML puede facilitar la tarea de encontrar una predicción a estos. El ML ayuda a solucionar problemas específicos de clasificación, asociación, agrupamiento y selección de rasgos (Sammut & Webb, 2010; Verona Pérez & Arco García, 2016).En nuestro día a día, estamos rodeados de diversos procesos que involucran el uso de ML, como puede ser el filtrado de los correos de spam, tecnologías de reconocimiento facial, previsiones de tráfico, etc. (Alzubi, et al., 2018). Como se puede ver, los usos que tiene el ML son muy numerosos, y en el futuro cada vez más tareas estarán desarrolladas con ayuda de procesos de tecnología, ya que cada día disponemos de más datos que es necesario procesar (Storm, et al., 2020).

El objetivo de este trabajo es dar una visión general acerca de las diferentes técnicas de análisis que existen en ML, centrando el análisis sobre todo en las técnicas de aprendizaje supervisado y no supervisado, y realizando aplicaciones empíricas sencillas de algunas de las técnicas explicadas para explicar de forma práctica el funcionamiento y la interpretación de estas técnicas a través de los programas R Studio y Weka.

2. FUNDAMENTOS TEÓRICOS DEL MACHINE LEARNING

2.1. TÉCNICAS DE ANÁLISIS EXISTENTES EN MACHINE LEARNING

2.1.1. Aprendizaje supervisado

Consiste en que un algoritmo aprenda a través de ejemplos, introduciendo datos y los resultados esperados (Nasteski, 2017), para posteriormente, predecir la respuesta cuando se le dé un conjunto de datos diferente (Mueller & Massaron, 2021). Es el método preferido por la comunidad del ML (Japkowicz, 2001).

2.1.1.1. Regresión

Si el resultado esperado de la técnica consiste en una o más variables continuas, entonces tenemos una regresión (Bishop, 2006). Es una de las bases y procedimientos más sencillos dentro del ML, ya que a través de unos datos existentes genera una predicción que se corresponde con el objeto a estudiar que teníamos (Huang, et al., 2020). Existen distintos tipos de regresiones, una de las más utilizadas en análisis realizados por artículos es la logística, aunque también se utilizan otras como la regresión lineal. Con este método se realiza un análisis de regresión a través de la suma de cuadrados que evalúe los coeficientes de las variables independientes que explican la variable dependiente (Rong & Bao-wen, 2018). En otros estudios utilizando las regresiones se pudo analizar que el uso de la herramienta educativa Moodle y los teléfonos inteligentes influyen positivamente en la motivación de los alumnos (Salas-Rueda, et al., 2023). También se han utilizado las regresiones para evaluar políticas, así como predecir el clima y los desastres naturales, aunque encontrar una tendencia sea difícil (Lazo Pilatuña & Moreano Moncayo, 2021).

2.1.1.2. Clasificación

Otro de los tipos de aprendizaje supervisado más comunes es la clasificación, encargado de clasificar los datos mediante etiquetas (Lazo Pilatuña & Moreano Moncayo, 2021), donde una de las técnicas más destacadas son los árboles de decisión (Rivero Suguiura, 2022) por su alta precisión (Nájera & Calleja, 2018), y su fácil comprensión (Ville, 2013).

Este método se desarrolla de forma escalonada, partiendo de todo el conjunto de datos al que conoceremos como un nodo raíz, este, normalmente se parte en otros nodos conocidos como nodos internos o de prueba, que finalmente acaban terminando en los nodos hoja o nodos de decisión (Rokach & Maimon, 2005), es decir, se van realizando divisiones del conjunto de datos en subconjuntos, uno para cada valor del atributo, este proceso se repite hasta que el proceso de bifurcación se detiene obteniendo finalmente el árbol de decisión ya que los datos no pueden seguir ramificándose (Aggarwal, 2015; Nájera & Calleja, 2018; Arana, 2021).

Algunos estudios realizados utilizando árboles de decisión han sido para identificar las razones del abandono universitario en los estudiantes de ingeniería informática (Bello, et al., 2020), otros en general para predecir las tasas de abandono de los estudiantes (Kemper, et al., 2020). No solo tiene aplicaciones educativas, ya que también se utiliza diariamente en consultas médicas para encaminarse a un problema médico o a otro según las pruebas que se realicen al paciente (Navada, et al., 2011).

Para la implementación de árboles de decisión existen diversos algoritmos como Random Tree, Reduced Error Pruning (REP Tree) y J48.

Random Tree es un algoritmo que aleatoriamente genera un árbol de decisión a partir de un set de posibles árboles (Zhao & Zhang, 2008), estos pueden ser generados eficientemente de una combinación de otros árboles de decisión y aun así llevarnos a

modelos precisos. (Ali, et al., 2012). Para realizarlo, el algoritmo selecciona una prueba basándose en un numero especifico de características aleatorias de cada nodo (Hamoud, et al., 2018).

Reduced Error Pruning (REP Tree) es uno de los algoritmos más sencillos para crear un árbol de decisión (Gupta, et al., 2012), y se caracteriza por ser de rápida decisión y tener un error reducido (Gokilam & Shanthi, 2016), está basado en los principios de entropía y minimización del error derivado de la varianza (Hamoud, et al., 2018). Los valores ausentes, no impiden el uso de este algoritmo, ya que a través del algoritmo C4.5 se puede utilizar solo la información presente para generar el árbol de decisión (Mohamed, et al., 2012).

J48, es uno de los más utilizados debido a la estabilidad, la precisión, la velocidad y la interpretabilidad de los resultados (Hamoud, et al., 2018). El objetivo es generalizar progresivamente un árbol de decisión hasta conseguir un equilibrio entre flexibilidad y precisión (Kaur & Chhabra, 2014). Este algoritmo fue creado para el programa Weka por su equipo, siendo una adaptación del algoritmo C4.5 (Cruz & Tumibay, 2019).

En algunos estudios, se ha visto como entre estos tres algoritmos, J48 ofrece una mejor precisión (Ai Munandar & Sumiati, 2017), seguido de REP Tree, y en último lugar Random Tree (Hamsagayathri & Sampath, 2016), considerado entre estos tres como el método más impreciso. Una clara ventaja de los árboles de decisión es que pueden ser visualizados y entendidos por gente no experta en la materia, y una clara desventaja es la gran generalización que se hace con los datos (Müller & Guido, 2016).

2.1.2. Aprendizaje no supervisado

El ML tiene una forma de aprendizaje no supervisado, esta consiste en introducir al sistema unos datos, con la diferencia de que en este caso no se le añaden unos resultados esperados (Nasteski, 2017) y teniendo por objetivo aprender a producir los resultados correctos sin ningún estimulo (Ghahramani, 2003). Se recomienda su uso sobre todo en conjuntos de datos sin procesar con la finalidad de obtener un análisis partiendo de datos no etiquetados (Usama, et al., 2019). Existen diversos algoritmos según la finalidad para la que se vaya a utilizar.

2.1.2.1. Clustering

El clustering es una técnica cuyo objetivo es encontrar patrones ocultos en el conjunto de datos separándolos en grupos de elementos, o también conocidos como clusters (Grira, et al., 2005). Esta técnica cada vez tiene más relevancia, ya que hay una gran demanda para descubrir métodos que reconozcan grupos distintos dentro de unos datos (Gentleman, et al., 2006). Dentro del clustering se han desarrollado diversos tipos (Jain, et al., 1999), algunos de ellos son k-means, clustering basado en modelos, clustering espectral, clustering jerárquico, clustering bayesiano y clustering particional (Shutaywi & Kachouie, 2021).

El clustering jerárquico descompone el conjunto de datos jerárquicamente (Usama, et al., 2019) y crea un árbol de soluciones (Wade, 2023) que puede ser aglomerativo y divisivo (Jain & Dubes, 1998; Kaufman & Rousseeuw, 1990).

El clustering bayesiano crea un modelo probabilístico de los datos con el que se decide el destino de un nuevo punto de prueba de forma probabilística (Usama, et al., 2019).

El clustering particional genera diversas particiones y las evalúa teniendo en cuenta un criterio o una característica como puede ser la distancia euclídea (Usama, et al., 2019). Entre las distintas formas de realizar el clustering particional, cabe destacar el algoritmo K-medias, el algoritmo K-medoids o Partitioning Around Medoids (PAM) (Kaufman &

Rousseeuw, 1990)) y el algoritmo Clustering Large Applications (CLARA) (Kassambara, 2017).

El método de clustering tiene diversas aplicaciones, ya que se ha utilizado para segmentar un beneficio (Saunders, 1980) o diversos mercados de sectores como el de bebidas alcohólicas o el sector inmobiliario (Mariño Santos, 2023; Jhon Rios, 2023). También se han realizado aplicaciones relacionadas con el comportamiento de los consumidores (Gough & Sozou, 2005), así como la creación de clústers para el sector del retail (Holý, et al., 2017).

2.1.2.2. Estimación de la densidad

Cuando el algoritmo trata de determinar la distribución de los datos, entonces estaríamos hablando de un algoritmo de estimación de la densidad (Bishop, 2006). Se pueden clasificar este tipo de algoritmos en paramétricos, semi-paramétricos y no paramétricos (Wang & Scott, 2019).

2.1.2.3. Reducción de la dimensionalidad

Los algoritmos de reducción de la dimensionalidad se basan en describir con exactitud los valores de p variables, utilizando un pequeño subconjunto de las variables, con esto se consigue reducir la dimensión del problema, con la desventaja de una ligera perdida de información (Peña, 2002). Hay diversas formas de reducir la dimensionalidad, encontrandose entre ellas el Análisis de Componentes Principales (ACP), la regresión de componentes principales y el análisis lineal discriminante entre otros métodos (Babenko, et al., 2021).

El ACP transforma un grupo de variables correlacionadas en un grupo de variables sin correlación (Reddy, et al., 2020), que además representan adecuadamente la información con un número menor de variables que han surgido como combinaciones lineales de las originales (Peña, 2002). Estas nuevas variables (componentes principales) estarán ordenadas según su varianza, y se podrán seleccionar para continuar realizando los análisis.

2.1.3. Aprendizaje semi-supervisado

Se trata de una técnica que combina el uso del aprendizaje supervisado con el aprendizaje no supervisado, su funcionamiento se basa en un conjunto de técnicas de aprendizaje supervisado que añade una parte del proceso sin supervisar, normalmente la parte del proceso sin supervisar es predominante (Mouriño García, 2018). Con este método mixto se pueden preprocesar los datos antes de realizar los análisis (Usama, et al., 2019).

2.1.4. Aprendizaje por refuerzo

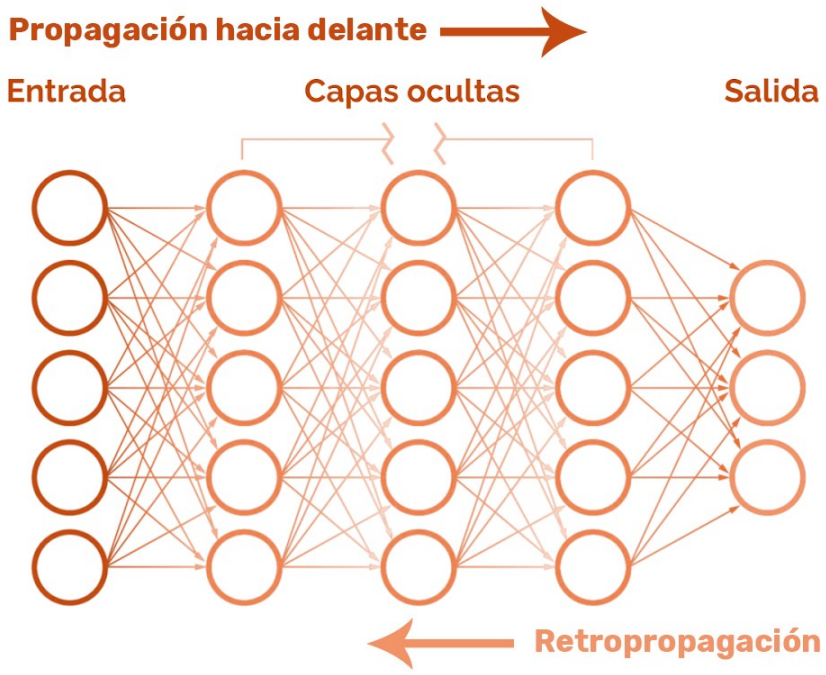
Consiste en la prueba y error, que es la característica más distintiva de este tipo de aprendizaje (Sutton, 1992). El funcionamiento se basa en introducir unos datos, sin dar detalles acerca del resultado esperado, y a través de acciones, el sistema va probando (Ghahramani, 2003) hasta obtener el mejor resultado, que puede no solo influir en la situación de prueba actual, sino que también puede influir en el resultado obtenido de una prueba futura, repitiéndose este proceso una infinidad de veces (Sutton & Barto, 1998; Kaelbling, et al., 1996; Boada, et al., 2005) La técnica de aprendizaje por refuerzo está avanzando rápidamente dentro de la comunidad del ML (Sugiyama, et al., 2012).

2.1.5. Redes neuronales y deep learning

Se trata de un modelo inspirado en las conexiones neuronales biológicas (Choi, et al., 2020), que se crea y entrena con una gran cantidad de datos para resolver un problema (De Luca, et al., 2021). El proceso de aprendizaje consiste en dos partes, la propagación hacia delante, donde los datos se envían desde la capa inicial, hacia una o varias capas

ocultas donde se procesan, para después acabar en la capa final donde se obtiene un resultado. En el caso de que este resultado no se corresponda con los datos reales que se tienen, se pasaría a utilizar la segunda parte, la retropropagación, con el que se ajustarán los pesos que se le otorgan a cada conexión entre las capas con el objetivo de reducir el error en el resultado. Repitiendo estos pasos, se llega a un punto donde el error que se obtiene en el resultado es aceptable. (Huang, et al., 2020).

Ilustración 1: Red neuronal



Fuente: (Datademia) modificada

Cuando estamos ante un caso de una red neuronal muy extensa, podemos hablar de DL, donde a través de numerosas capas de procesamiento, se crea una representación jerárquica que ordena los datos siendo las capas más cercanas al input simples, a diferencia de aquellas que están más procesadas que tienen altos niveles de complejidad. (Shinde & Shah, 2018).

2.1.6. Ensamble de modelos

Generando submuestras de los datos originales, y la estimación para cada una de esas submuestras de un modelo, las estimaciones se agregan y se produce una estimación final que tiene una capacidad predictiva superior a la de un solo clasificador (Rosati, 2021).

Muchas de las aplicaciones del ML en la economía están centradas en la estimación de parámetros (Mullainathan & Spiess, 2017), pero hay que tener en consideración la existencia de otros métodos, y que cada uno de ellos tiene su contexto de aplicación, ya que como hemos visto, no todos los métodos realizan lo mismo, sino que se utilizan para diferentes funciones. En cuanto a los distintos algoritmos, estamos ante un campo que día a día va creciendo rápidamente, por lo que cada cierto tiempo, algoritmos anteriores son mejorados y sustituidos por nuevas opciones, siendo más eficientes y precisos que los anteriores, pero también más complejos (Sharifani & Amini, 2023).

3. DESARROLLO EMPÍRICO

En esta sección se realiza el desarrollo empírico enfocado en la aplicación de las técnicas de ML supervisadas y no supervisadas. De esta forma se podrá comprender de forma básica qué información es necesaria para utilizar los algoritmos y cómo interpretar el resultado que nos ofrece cada tipo de algoritmo.
Se hará uso de diferentes bases de datos, para de esta forma, aportar explicaciones con diferentes datos, y no ceñirse simplemente a un solo modelo.

3.1. APRENDIZAJE SUPERVISADO

3.1.1. Regresión

En este caso se realiza una regresión lineal múltiple $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$ donde y es la variable dependiente (crime), β_0 es la constante, y cada x pertenece a una de las variables independientes (enroll, priv, pólice) siendo el β que acompaña a cada variable independiente su coeficiente, por tanto obteniendo la siguiente regresión $crime = \beta_0 + \beta_1 police + \beta_2 enrol + \beta_3 priv + e$.

Los datos para realizar la regresión han sido extraídos de la base de datos “campus” creada por Wooldridge, y se van a utilizar dos softwares diferentes, siendo R Studio y Weka, para comparar las ejecuciones de los métodos por ambos programas.

Con R Studio ejecutando la regresión especificada anteriormente, obtenemos un primer resultado, donde vemos que la variable priv no es significativa, por lo que tenemos que eliminar esa variable de la regresión para obtener una regresión donde todas las variables sean significativas. En el caso de Weka, esto no ocurre, ya que directamente ejecuta el algoritmo que nos muestra la regresión con las variables significativas, lo que nos ahorra tiempo y permite elegir el mejor modelo de forma más eficiente, obteniendo de ambos programas el mismo resultado final, pero utilizando menos tiempo en Weka, con la que vemos que el número de crímenes aumenta en 7,57 crímenes por cada policía adicional en el campus, manteniendo el resto de los factores constantes.

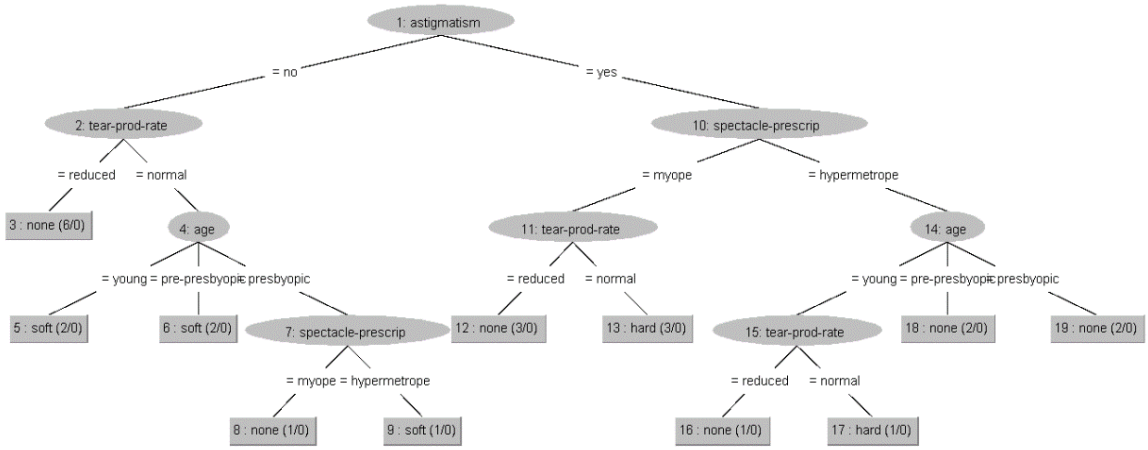
$$\widehat{crime} = -153{,}6529 + 7{,}5702\ police + 0{,}0244\ enrol$$

Este resultado es llamativo, y puede haber sido influenciado por el sesgo de variable omitida, es decir, se ha podido dejar fuera de la regresión una variable independiente importante que hace que las conclusiones extraídas del modelo puedan ser incorrectas.

3.1.2. Clasificación

Utilizando la base de datos “contact lenses” incluida en el programa Weka, sin elegir ningún tipo de semilla de entrenamiento, y dejando las opciones del programa por defecto, realizamos un árbol de decisión del tipo Random Tree para comprender si una persona necesita lentillas blandas, duras o directamente no necesita tener lentillas. Se obtiene un árbol de decisión algo complejo, donde se tienen en cuenta todas las variables de la base de datos y según la respuesta que se tenga o el grupo al que se pertenezca se llega al resultado, por ejemplo, si una persona no tiene astigmatismo, posteriormente a través del índice de lagrimeo si este es reducido no tendrá que usar lentillas, sin embargo, si su índice de lagrimeo es normal, se crea una bifurcación más, donde según la edad del ojo, si es joven y no tiene presbicia, se usarán lentillas blandas, mientras que si el ojo ya tiene presbicia se vuelve a bifurcar según si tiene prescripción de gafas, donde si se es miope no se usará ninguna lentilla, mientras que si se es hipermétrope se tendrá que usar una lentilla blanda.

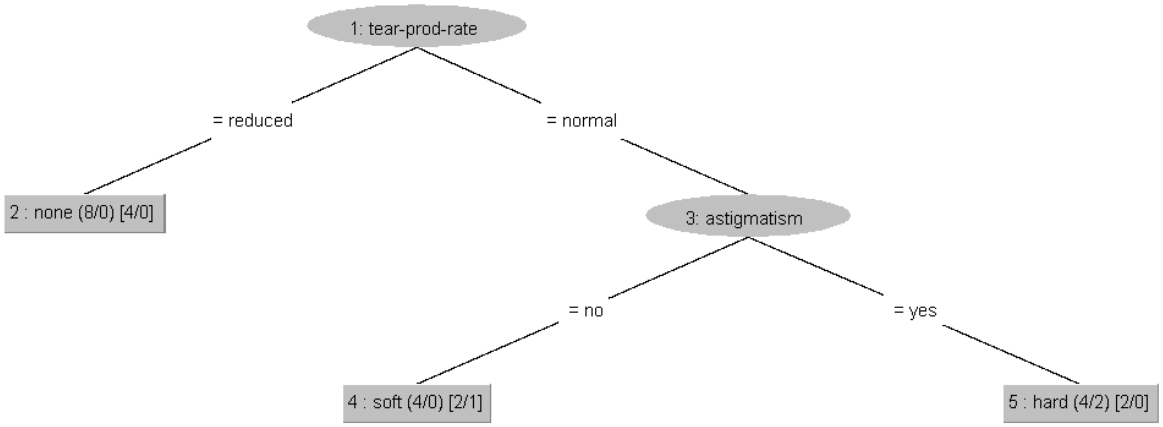
Ilustración 2: Árbol de decisión (Random Tree)



Fuente: elaboración propia a partir de la base de datos contact lenses de Weka

El algoritmo Random Tree, ofrece un resultado muy completo, pero tenemos otros algoritmos como se ha visto anteriormente que generan arboles de decisión distintos, como puede ser el REP Tree. En este caso se obtiene una solución más sencilla de comprender, ya que, según el índice de lagrimeo, si este es reducido, sabemos que no harán falta lentillas, mientras que en el caso de que el índice de lagrimeo sea normal, el tener o no astigmatismo marcara la diferencia entre tener que utilizar unas lentillas blandas o duras.

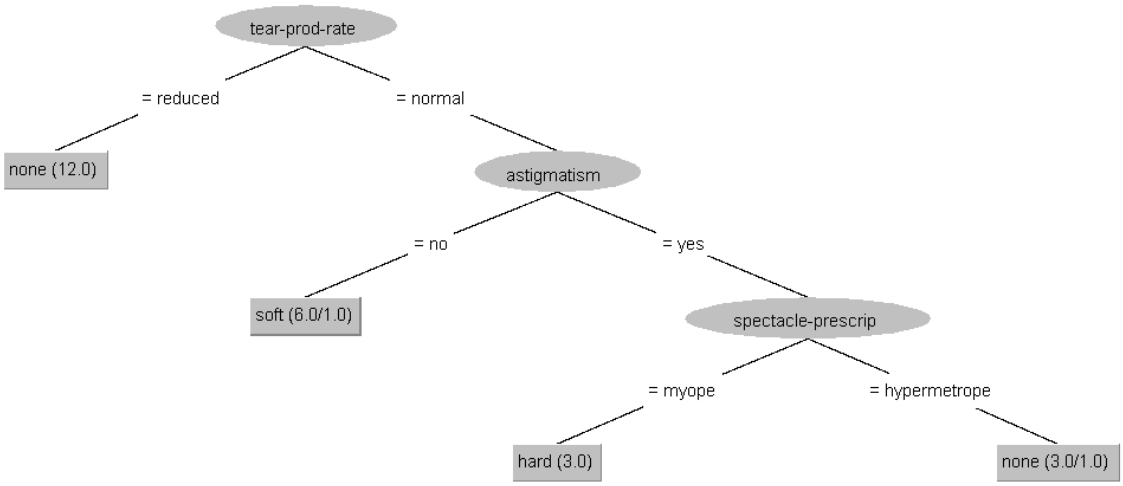
Ilustración 3: Árbol de decisión (REP Tree)



Fuente: elaboración propia a partir de la base de datos contact lenses de WEKA

Utilizando el algoritmo J48, se obtiene un árbol extremadamente similar al anterior generado por REP Tree, solo que esta vez desglosa en el caso afirmativo del astigmatismo, una categoría adicional, que es la prescripción óptica, en este caso al ser miope debería utilizar una lentilla dura, mientras que si es hipermétrope no debería utilizar ninguna.

Ilustración 4: Árbol de decisión (J48)



Fuente: elaboración propia a partir de la base de datos contact lenses de WEKA

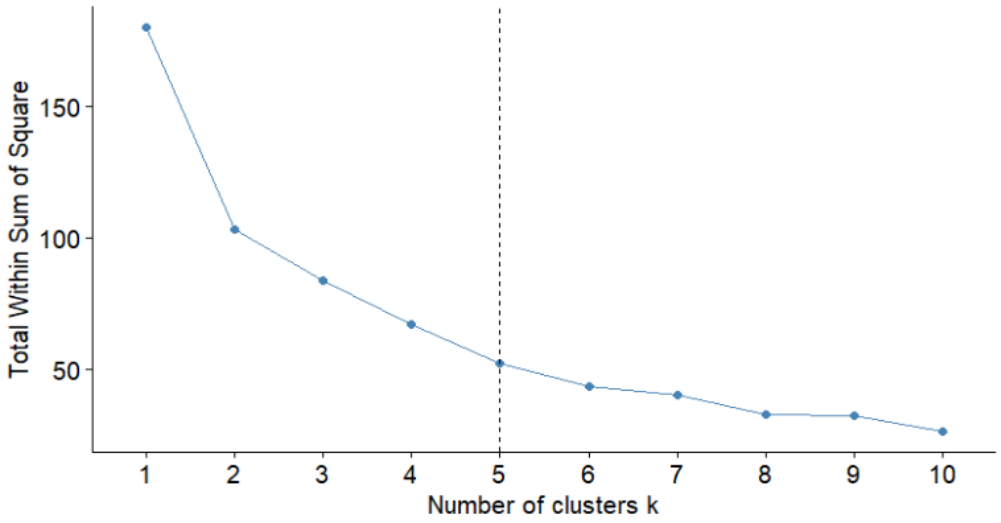
3.2. APRENDIZAJE NO SUPERVISADO

3.2.1. Clustering

Para poner en práctica el método de clustering, se realizará una demostración con el programa R Studio, utilizando la base de datos "corn" procedente de Wooldridge y el paquete adicional factoextra para poder realizar el procedimiento. Es muy importante que la base de datos no tenga ningún valor ausente, y los datos estén estandarizados, es decir hacer que tengan una media de 0, y que su desviación típica sea 1, lo que ayudará a las variables sean comparables (Kassambara, 2017).

Una parte importante del procedimiento es saber cuántos clústers o grupos de elementos es recomendable utilizar para los datos que se tienen, en este caso se emplea el análisis de codo (Elbow Method), que utiliza las k-medias y la suma de los cuadrados (Syakur, et al., 2018) para determinar el número óptimo de clústeres que deberíamos utilizar para hacer el análisis.

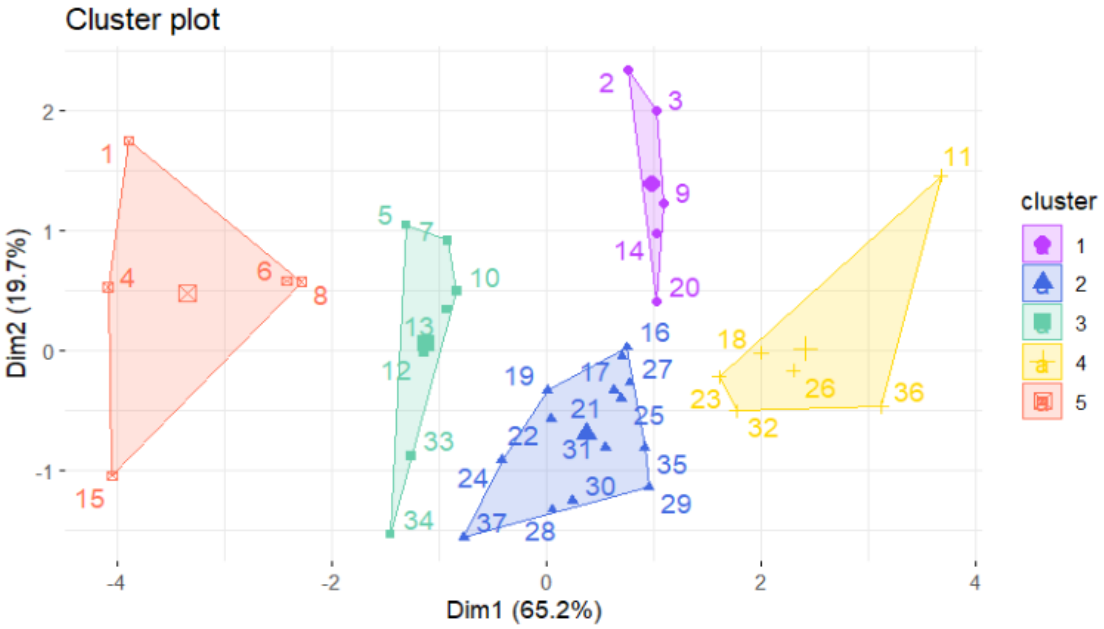
Ilustración 5: Número óptimo de clústers



Fuente: elaboración propia a partir de la base de datos corn de Wooldridge

Como se puede apreciar en la figura, a partir del clúster 5, se produce un ligero cambio de pendiente, donde la tasa de disminución de la suma de los cuadrados dentro de los clústeres se desacelera, por lo que se puede utilizar esta referencia para establecer que en este análisis usaremos 5 clústeres. Para poder reproducir los resultados a posteriori, es necesario fijar una semilla, que en este caso se ha definido en 2002 y se han definido las iteraciones en un máximo de 10 para la realización de las k-medias, generando un gráfico para poder apreciar visualmente la distribución de los clúster utilizando como referencia la primera y la segunda dimensión, que representan aproximadamente un 85% de la varianza conjuntamente.

Ilustración 6: Representación de los clústers



Fuente: elaboración propia a partir de la base de datos corn de Wooldridge

Cabe destacar como el agrupamiento es distintivo y se separan bien los clústeres, por lo que las características que definen estos grupos de elementos son diferentes entre ellos. Por ejemplo, en el caso del clúster 1, podemos observar claramente como en la primera dimensión las observaciones son muy homogéneas respecto a las características que predominan en esta dimensión, mientras que, en el caso de la segunda dimensión, estas se ven algo más dispersas, sobre todo si comparamos la observación 2 con la observación 20.

3.2.2. Reducción de la dimensionalidad

El Análisis de Componentes Principales (ACP) se va a realizar con el programa R Studio para la base de datos “401K” de pensiones perteneciente a Wooldridge con el paquete adicional “psych”. Se utilizan todas las variables, estandarizándolas para eliminar la diferencia de escalas y evitar que se vean influenciados los resultados del análisis.

El primer paso para realizar una reducción de la dimensionalidad es conocer si este método es adecuado, es decir, observando si las variables originales tienen correlación, para ello se utilizan dos herramientas, el Test de Esfericidad de Bartlett y el Índice KMO. Comenzando con el Test de Esfericidad, tenemos un contraste de hipótesis:

$$H_0: |\boldsymbol{R}| = 1$$

$$H_1: |\boldsymbol{R}| \neq 1$$

Al ser el p-valor inferior a 0.5, podemos rechazar la hipótesis nula, lo que significa que tiene sentido aplicar un análisis de componentes principales a esta base de datos. Por otro lado, también tenemos el Índice de Kaiser, Meyer y Olkin, en el que los valores oscilan entre 0 y 1, teniendo una mayor adecuación cuanto más cercano a 1 sea este indicador. En el caso de este análisis, el resultado es 0,71, por lo tanto, tendríamos una alta adecuación de ACP. Como se ha visto con anterioridad el método de la ACP transformaba las variables originales. Así, la transformación por componentes principales se define de la siguiente manera:

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{T}$$

Siendo $\boldsymbol{Y}$ una combinación lineal de $\boldsymbol{X}$, de forma que $Y_1 = \boldsymbol{X}\boldsymbol{t}_1$, $Y_2 = \boldsymbol{X}\boldsymbol{t}_2, \dots$, $Y_k = \boldsymbol{X}\boldsymbol{t}_k$, donde:

$$\boldsymbol{X} = (X_1,\, X_2, \dots,\, X_k)$$

$$\boldsymbol{Y} = (Y_1,\, Y_2\,, \dots,\, Y_k)$$

$$\boldsymbol{T} = (t_1,\, t_2\,, \dots,\, t_k)$$

X se trata de las variables originales, Y las componentes principales, y T es la matriz de vectores normalizados que definen a dichas componentes principales (autovectores). Definiendo k en este caso como 8, ya que es el número de variables que utilizaremos para el análisis. Estas nuevas variables obtenidas (componentes principales) se ordenan en función de su varianza y contendrán la mayor parte de la información de X.

Normalmente un paso importante es seleccionar el número de componentes principales para realizar el análisis, para ello se pueden utilizar varios criterios, como el grafico de sedimentación, o con un análisis de la variabilidad explicada, con el que al 80% de variabilidad explicada, en este caso se cumpliría para las cuatro primeras componentes principales, explicando estas un 83% de la variabilidad, pero debido a la extensión del trabajo, solo comentaremos una de las componentes para que se pueda comprender el análisis que habría que seguir con el resto.

Tabla 1: Autovectores de la matriz de correlaciones

	Componente 1	**Componente 2**	**Componente 3**
Prate	0.076	- 0.615	0.087
Mrate	0.028	- 0.570	- 0.017
Totpart	- 0.501	- 0.140	- 0.178
Totelg	- 0.509	- 0.091	- 0.187
Age	- 0.141	- 0.365	0.667
Totemp	- 0.495	- 0.050	- 0.167
Sole	0.190	- 0.328	- 0.660
Ltotemp	- 0.426	0.161	0.125

Fuente: elaboración propia a partir de la base de datos 401k de Wooldridge

En relación a la primera componente se puede observar cómo destacan las variables de total de participantes, total de participantes aptos, total de trabajadores de la empresa, y el logaritmo del total de trabajadores de la empresa, todas estas variables, tienen valores similares con signo negativo, por lo que las pensiones que más destacan

son aquellas que tienen valores más elevados de estas variables, en este caso destaca la persona número 999 con los valores más elevados de estas variables anteriormente comentadas, seguida de la persona 935. En el caso del análisis de otras componentes, estas darán un peso diferente a cada variable, por ejemplo, en la segunda componente se le da un gran peso a la tasa de participación en el plan 401k, a diferencia de la primera componente donde el peso que tenía esta variable es ínfimo.

4. CONCLUSIONES

En este trabajo se han estudiado las diversas técnicas de aprendizaje existentes relacionadas con Machile Learning, destacando como cada método tiene su propio contexto de aplicación y funciones específicas, haciendo énfasis en el hecho de que el ML es un campo en constante evolución, lo que implica que van surgiendo nuevos algoritmos que mejoran y reemplazan a los anteriores aportando resultados más eficientes y precisos, aunque también más complejos.

En el desarrollo empírico, a través del uso de varias bases de datos y los programas Weka y R Studio, se han aplicado distintas técnicas de aprendizaje supervisado y no supervisado, obteniendo una clara visión de la información necesaria para implementarlas y la forma en la que se interpretan los resultados obtenidos de estos algoritmos.

Entre las técnicas de aprendizaje supervisado, se ha hablado de regresiones, una de las aplicaciones más utilizadas en la economía, donde se ha conocido que usar el programa Weka para este tipo de aplicaciones simplifica el trabajo debido a que las variables no significativas son automáticamente excluidas del análisis, y que es necesario tener en cuenta otros factores para las regresiones como el sesgo de variable omitida. También se ha utilizado la clasificación a través de diversos arboles de decisión, destacando el algoritmo J48 debido a la mejor precisión.

Entre las técnicas de aprendizaje no supervisado, se analizó el clustering, donde se puede apreciar la creación de grupos a partir de las observaciones con características similares, y, por otro lado, la reducción de la dimensionalidad con el Análisis de Componentes Principales, con la que se ha descrito las herramientas para comprobar la idoneidad de una base de datos para el uso con esta herramienta, y la facilidad que aportan para realizar un análisis sin perder mucha información.

Por lo que, aunque los métodos y los algoritmos de ML sean distintos y por ende tengan distintas aplicaciones, la selección del método adecuado es determinante para obtener resultados pertinentes eficientes y precisos. El trabajo demuestra que utilizando diversos enfoques y combinando las herramientas se pueden mejorar los procesos de análisis y predicción de diversas áreas.

5. REFERENCIAS

Aggarwal, C. C., 2015. *Data Mining: The Textbook.* Nueva York: Springer.

Ai Munandar, T. & Sumiati, 2017. The Classification of Cropping Patterns Based on Regional Climate Classification Using Decision Tree Approach. *Journal of Computer Science,* 13(9), pp. 408-415.

Albornoz, M. M., 2021. El titular de datos personales, parte débil en tiempos de auge de la Inteligencia Artificial. ¿Cómo fortalecer su posición?. *Revista Ius,* 15(48), pp. 209-242.

Ali, J., Khan, R., Ahmad, N. & Maqsood, I., 2012. Random Forests and Decision Trees. *International Journal of Computer Science Issues (IJCSI),* 9(5).

Álogos, 2018. *Introducción a Machine Learning.* [Online] Available at: https://web.archive.org/web/20190225135018/http://alogos.es:80/introduccion-machine-learning
[Accessed 18 febrero 2024].

Alzubi, J., Nayyar, A. & Kumar, A., 2018. Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series,* Volume 1142.

Arana, C., 2021. *Modelos de aprendizaje automoático mediante árboles de decisión.* [Online]
Available at: https://www.econstor.eu/bitstream/10419/238403/1/778.pdf
[Accessed 15 abril 2024].

Babenko, V. et al., 2021. Classical Machine Learning Methods in Economics Research: Macro and Micro Level Examples. *WSEAS Transactions on Business and Economics,* Volume 18, pp. 209-217.

Bello, F. et al., 2020. *Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout.* s.l., IEEE Computer Society.

Bishop, C. M., 2006. *Pattern recognition and machine learning.* 1 ed. Nueva York: Springer.

Boada, M. J. L., Boada, B. L. & López, V. D., 2005. Algoritmo de aprendizaje por refuero continuo para el control de un sistema de suspensión semi-activa. *Revista Iberoamericana de Ingeniería Mecánica,* 9(2), pp. 77-91.

Choi, R. Y. et al., 2020. Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology,* 9(2), p. 14.

Cruz, A. P. D. & Tumibay, G. M., 2019. Predicting Tuberculosis Treatment Relapse: A Decision Tree Analysis of J48 for Data Mining. *Journal of Computer and Communications,* Volume 7, pp. 243-251.

Datademia, n.d. *¿Qué es Deep Learning y qué es una red neuronal?.* s.l.:s.n.

De Luca, A. M., Irigoitia, M. E., Pérez, G. A. & Pons, C. F., 2021. *Uso de la técnica de Transfer Learning en Machine Learning para la clasificación de productos en el Banco Alimentario de La Plata.* Mendoza, s.n.

Gentleman, R. et al., 2006. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Nueva York: Springer.

Ghahramani, Z., 2003. Unsupervised learning. In: *Advanced Lectures on Machine Learning.* Heidelberg: Springer Berlin Heidelberg, pp. 72-112.

Gokilam, G. & Shanthi, D. K., 2016. Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset. *COMPUSOFT, An international journal of advanced computer technology,* 5(3).

Gough, O. & Sozou, P., 2005. Pensions and retirement savings: cluster analysis of consumer behaviour and attitudes. *International Journal of Bank Marketing,* 23(7), pp. 558-570.

Grira, N., Crucianu, M. & Boujemaa, N., 2005. Unsupervised and Semi-supervised Clustering: a brief survey. *Review of Machine Learning Techniques for Processing Multimedia Content.*

Gupta, D. L., Malviya, A. K. & Singh, S., 2012. Performance Analysis of Classification Tree Learning Algorithms. *International Journal of Computer Applications,* 55(6).

Hamoud, A. K., Hashim, A. S. & Awadh, W. A., 2018. Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence,* 5(2), pp. 26-31.

Hamsagayathri, P. & Sampath, P., 2016. Decision Tree Classifiers For Classification Of Breast Cancer. *International Journal of Current Pharmaceutical Research,* 9(2), pp. 21-36.

Holý, V., Sokol, O. & Černý, M., 2017. Clustering retail products based on customer behaviour. *Applied Soft Computing,* Volume 60, pp. 752-762.

Huang, J.-C., Ko, K.-M., Shu, M.-H. & Hsu, B.-M., 2020. Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Comput & Applications,* Volume 32, pp. 5461-5469.

Jain, A. K. & Dubes, R. C., 1998. *Algorithms for clustering data.* Prentice Hall ed. Englewood Cliffs: s.n.

Jain, A., Murty, M. & Flynn, P., 1999. Data clustering: a review. *ACM Computing Surveys,* 31(3), pp. 264-323.

Japkowicz, N., 2001. Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks. *Machine Learning,* Volume 42, pp. 97-122.

Jhon Rios, L. P., 2023. *Aplicación de técnicas de Data Analytics: Clustering y Regresión Lineal Múltiple, para la segmentación de la oferta y proyección de ciclos inmobiliarios en el mercado de oficinas prime.* [Online].

Kaelbling, L. P., Littman, M. L. & Moore, A. W., 1996. Reinforcement Learning : A Survey. *Journal of Artificial Intelligence ,* Issue 4, pp. 237-285.

Kassambara, A., 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning.* 1 ed. s.l.:Sthda.

Kaufman, L. & Rousseeuw, P., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* 1 ed. Nueva York: Wiley.

Kaur, G. & Chhabra, A., 2014. Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications,* 98(22).

Kemper, L., Vorhoff, G. & Wigger, B., 2020. Predicting student dropout: A machine learning approach. *European Journal of Higher Education,* pp. 1-20.

Lazo Pilatuña, J. R. & Moreano Moncayo, A. V., 2021. *Desarrollo de un sistema inteligente para predecir los consumos de medicamentos genéricos de mayor demanda en el distrito de salud 06d05 guano-penipe, aplicando técnicas de regresión de machine learning..* [Online]
Available at: http://dspace.espoch.edu.ec/handle/123456789/19266

Mariño Santos, C., 2023. *Análisis de clustering para la segmentación del mercado: un caso de estudio de una aplicación de una bebida alcohólica en las principales ciudades de Colombia.* [Online].

Mohamed, W. N. H. W., Salleh, M. N. M. & Omar, A. H., 2012. *A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms.* s.l., s.n., pp. 392-397.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2018. *Foundations of machine learning.* s.l.:MIT press.

Mouriño García, M. A., 2018. *Clasificación multilingüe de documentos utilizando machine learning y la Wikipedia.* [Online]
Available at: http://hdl.handle.net/11093/928

Mueller, J. P. & Massaron, L., 2021. *Machine Learning For Dummies.* 2 ed. s.l.:Wiley.

Mullainathan, S. & Spiess, J., 2017. Machine Learning: An Applied Econometric Approac. *Journal of Economic Perspectives,* 31(2), pp. 87-106.

Müller, A. C. & Guido, S., 2016. *Introduction to Machine Learning with Python.* 1 ed. s.l.:O'Reilly Media, Inc..

Nájera, A. B. U. & Calleja, J. d. l., 2018. Selection of academic tutors in higher education using decisiontrees. *Revista Española de Orientación y Psicopedagogía,* 29(1), pp. 108-124.

Nasteski, V., 2017. An overview of the supervised machine learning methods. *Horizons.B,* Volume 4, pp. 51-62.

Navada, A., Ansari, A. N., Patil, S. & Sonkamble, B. A., 2011. *Overview of use of decision tree algorithms in machine learning.* s.l., s.n., pp. 37-42.

Oseni, A. et al., 2020. Security and privacy for artificial intelligence: Opportunities and challenges. *ACM,* 37(4).

Peña, D., 2002. *Análisis de datos multivariantes.* Madrid: McGraw-Hill Interamericana de España.

Pulido, I. G., 2023. El uso de la inteligencia artificial generativa en la investigación de la ciberdelincuencia de género: ante el auge de los deepfakes. *IUS ET SCIENTIA,* 9(2), pp. 157-180.

Reddy, G. T. et al., 2020. Analysis of dimensionality reduction techniques on big data. *Ieee Access,* Volume 8, pp. 54776-54788.

Rivero Suguiura, F. O., 2022. Árbol de Decisión en Aprendizaje Automático. *Revista Varianza,* Issue 19, pp. 39-46.

Rokach, L. & Maimon, O., 2005. Decision trees. In: L. Rokach & O. Maimon, eds. *Data Mining and Knowledge Discovery Handbook.* s.l.:Springer, pp. 165-192.

Rong, S. & Bao-wen, Z., 2018. *The research of regression model in machine learning field.* Chuan, MATEC Web Conf 176.

Rosati, G., 2021. Métodos de Machine Learningcomo alternativa para la imputación de datos perdidos. Un ejercicio en base a la Encuesta Permanente de Hogares. *Revista De La Asociación Argentina De Especialistas En Estudios Del Trabajo (ASET),* Volume 61.

Rouhiainen, L., 2018. *Inteligencia artificial 101 cosas que debes saber hoy sobre nuestro futuro.* Barcelona: Alienta editorial.

Salas-Rueda, R.-A., Ramírez-Ortega, J., Martínez-Ramírez, S.-M. & Alvarado-Zamorano, C., 2023. Uso de los algoritmos Machine Learning para analizar Moodle y los teléfonos inteligentes en el proceso educativo de la Física. *Texto Livre,* Volume 16, pp. 1-20.

Sammut, C. & Webb, G. I., 2010. *Encyclopedia of machine learning.* s.l.:Springer Science & Business Media.

Samuel, A. L., 1959. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development,* 3(3).

Saunders, J., 1980. Cluster Analysis for Market Segmentation. *European Journal of Marketing,* 14(7), pp. 422-435.

Sharifani, K. & Amini, M., 2023. Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal,* 10(07), pp. 3897-3904.

Shinde, P. P. & Shah, S., 2018. A Review of Machine Learning and Deep Learning Applications. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),* pp. 1-6.

Shutaywi, M. & Kachouie, N. N., 2021. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy,* 23(6).

Storm, H., Baylis, K. & Heckelei, T., 2020. Machine learning in agricultural and applied economics. *European Review of Agricultural Economics,* 47(3), pp. 849-892.

Sugiyama, M., Suzuki, T. & Kanamori, T., 2012. *Density Ratio Estimation in Machine Learning.* 1 ed. s.l.:Cambridge University Press.

Sutton, R. S., 1992. Introduction: The challenge of reinforcement learning. In: R. S. Sutton, ed. *Reinforcement learning.* s.l.:Springer, pp. 1-3.

Sutton, R. S. & Barto, A. G., 1998. *Reinforcement Learning: An introduction.* 1 ed. Cambridge: The MIT Press.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & Satoto1, B. D., 2018. *Integration k-means clustering method and elbow method for identification of the best customer profile cluster.* s.l., IOP Publishing.

Usama, M. et al., 2019. Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access,* Volume 7, pp. 65579-65615.

Verona Pérez, I. C. & Arco García, L., 2016. Una revisión sobre aprendizaje no supervisado de métricas de distancia. *Revista Cubana de Ciencias Informáticas,* 10(4), pp. 43-67.

Ville, B. d., 2013. Decision trees. *WIREs Computational Statistics,* Volume 5, pp. 448-455.

Wade, S., 2023. Bayesian cluster analysis. *Philosophical Transactions of the Royal Society,* 281(2247).

Wang, Z. & Scott, D. W., 2019. Nonparametric density estimation for high-dimensional data—Algorithms and applications. *WIREs Computational Statistics,* 11(4).

Zhao, Y. & Zhang, Y., 2008. Comparison of decision tree methods for finding active objects. *Advances in Space Research,* pp. 1955-1959.