Ejercicio 1

Estas son algunas posibles definiciones de Machine Learning extraidas de la web:

1. El machine learning logra el aprendizaje de los ordenadores a partir de los datos que se le introducen, así como de la ejecución de algoritmos.[1]

2. El aprendizaje automático es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.[2]

3. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.[3]

La definición 2 explica bien la diferencia entre Machine Learning e Inteligencia Artificial. Machine Learning es una rama de la Inteligencia Artificial, es decir que es un subconjunto de dicha disciplina.

En cuanto al Análisis Estadístico, este se utiliza para derivar información de los datos, pero se diferencia con Machine Learning dado que ML implica un procesamiento adicional de los datos.

Data Mining se utiliza en un conjunto de datos existente para encontrar patrones. Con Machine Learning, por otro lado, se entrena en un conjunto de datos de "entrenamiento", que le enseña a la computadora cómo dar sentido a los datos y luego hacer predicciones sobre nuevos conjuntos de datos.

Machine Learning tiene diversas aplicaciones, algunas de ellas son: motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, y robótica.

---

[1] https://planetachatbot.com/cu%C3%A1les-son-las-principales-diferencias-entre-inteligencia-artificial-y-machine-learning-3ffa6db9e43

[2] https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

[3] https://expertsystem.com/machine-learning-definition/

Ejercicio 2

Estas son algunas de las herramientas más populares para Machine Learning:

- RapidMiner: Es un software para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación, educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales. Proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización.[4]

- Weka: Weka contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y el modelado predictivo, junto con interfaces gráficas de usuario para un fácil acceso a estas funciones. Admite varias tareas estándar de minería de datos, más específicamente, preprocesamiento de datos, agrupamiento, clasificación, regresión, visualización y selección de características. Weka también proporciona acceso a bases de datos SQL utilizando Java Database Connectivity y puede procesar el resultado devuelto por una consulta de base de datos.[5]

- KNIME: Es una plataforma de análisis, informes e integración de datos. KNIME integra varios componentes para el aprendizaje automático y la minería de datos a través de su concepto de canalización de datos modular. Una interfaz gráfica de usuario y el uso de JDBC permiten el ensamblaje de nodos que combinan diferentes fuentes de datos, incluido el preprocesamiento, modelado, análisis y visualización de datos. KNIME integra varios otros proyectos open-source, por ejemplo: algoritmos de aprendizaje automático de Weka, H2O.ai, Keras, Spark, R project y LIBSVM.[6]

- Shogun: Es una libreria open-source de Machine Learning escrita en C ++. Ofrece numerosos algoritmos y estructuras de datos para problemas de aprendizaje automático. Ofrece interfaces para Octave, Python, R, Java, Lua, Ruby y C#. Estos son algunos de los tipos de algoritmos que Shogun soporta:[7]
    - Máquinas de vectores de soporte
    - Algoritmos de reducción de dimensionalidad
    - Algoritmos de aprendizaje online
    - Algoritmos de clustering
    - Análisis discriminante lineal

[4] https://es.wikipedia.org/wiki/RapidMiner
[5] https://en.wikipedia.org/wiki/Weka_(machine_learning)
[6] https://docs.knime.com/
[7] https://www.shogun-toolbox.org/

Ejercicio 3

CRISP-DM significa “cross-industry process for data mining”. La metodología CRISP-DM proporciona un enfoque estructurado para planificar un proyecto de minería de datos. Consta de 6 pasos y pueden tener iteraciones cíclicas según las necesidades de los desarrolladores. Estos pasos son Comprensión empresarial, Comprensión de datos, Preparación de datos, Modelado, Evaluación y Despliegue.

El primer paso es la comprensión empresarial y su objetivo es dar contexto a los objetivos y a los datos para que el desarrollador tenga una noción de la relevancia de los datos en ese modelo de negocio en particular.
Se compone de reuniones, lectura de documentación, aprendizaje de campo específico y otras formas que ayudan al equipo de desarrollo a hacer preguntas sobre el contexto relevante.

El segundo paso es la comprensión de datos y su objetivo es saber qué se puede esperar y lograr a partir de los datos. Comprueba la calidad de los datos, en varios términos, como la integridad y la distribución de valores.
Esta es una parte crucial del proyecto porque define cuán viables y confiables pueden ser los resultados finales.

El tercer paso es la preparación de datos e involucra los procesos que convierten los datos en algo útil para los algoritmos. Algunos algoritmos funcionan mejor bajo ciertos parámetros, algunos no aceptan valores no numéricos, otros no funcionan bien con una gran variación en los valores. Le corresponde al equipo de desarrollo normalizar la información.

El cuarto paso es el modelado y es el núcleo de cualquier proyecto de aprendizaje automático. Este paso es responsable de los resultados que deben satisfacer o ayudar a satisfacer los objetivos del proyecto.
Algunos algoritmos, como agrupamiento jerárquico, series de tiempo, regresión lineal, k vecinos más cercanos, y muchos otros, son utilizados en este paso en la metodología.

El quinto paso es la evaluación, donde se debe verificar que los resultados sean válidos y correctos. En caso de que los resultados sean incorrectos, la metodología permite volver a revisar el primer paso, para comprender por qué los resultados están equivocados.
Por lo general, en un proyecto de ciencia de datos, el científico de datos divide los datos en entrenamiento y pruebas. En este paso se utilizan los datos de prueba, su objetivo es verificar que el modelo sea aproximado a la realidad.

El sexto y último paso es el despliegue y consiste en presentar los resultados de una manera útil y comprensible, la cual varía según el usuario final.[8]

[8] https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781

Otra metodología es TDSP. Similar a CRISP-DM, TDSP proporciona un ciclo de vida para estructurar el desarrollo de proyectos de ciencia de datos, describiendo todos los pasos que generalmente se toman al ejecutar un proyecto. El ciclo de vida de TDSP se compone de 5 etapas:

1. Comprensión empresarial
2. Adquisición y comprensión de datos
3. Modelado
4. Despliegue
5. Aceptación del cliente

TDSP es una metodología de ciencia de datos más detallada y actualizada, adaptada a enfoques más ágiles. También incluye un paso de comprensión empresarial más detallado al comienzo de un proyecto.[9]

[9] https://deeperinsights.com/how-to-run-a-data-science-team/

Ejercicio 5

“Forest Fires Data Set”:[10]

- Intenta predecir el área impactada por incendios forestales, principalmente en base a datos meteorológicos.
- 13 atributos.
- La mejor configuración utiliza un SVM (Support Vector Machine) y cuatro entradas meteorológicas (temperatura, humedad relativa, lluvia y viento) y es capaz de predecir el área quemada de pequeños incendios, que son más frecuentes. Este conocimiento es particularmente útil para mejorar la gestión de recursos de extinción de incendios.[11]

[10] https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[11] https://www.researchgate.net/publication/238767143_A_Data_Mining_Approach_to_Predict_Forest_Fires_using_Meteorological_Data#:~:text=A%20Data%20Mining%20Approach%20to%20Predict%20Forest%20Fires%20using%20Meteorological%20Data,-Article%20(PDF%20Available&text=Forest%20fires%20are%20a%20major,damage%20while%20endangering%20human%20lives.&text=In%20this%20work%2C%20we%20explore,burned%20area%20of%20forest%20fires.