Centro Universitario de Ciencias Exactas e Ingenierías

Universidad de Guadalajara

Actividad 2: Tipos de aprendizaje, validación y métricas

Aprendizaje Máquina

Alumno: Samuel David Pérez Brambila

Código: 222966286

Profesora: Karla Ávila Cárdenas

Sección: D01

Fecha de Entrega: 01 de Septiembre de 2024

Introducción

El aprendizaje máquina, machine learning o también llamado aprendizaje automático es “una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo).” (Iberdrola, s.f.)

Hoy en día, encontramos diversas aplicaciones del machine learning en nuestra vida cotidiana, donde podemos destacar los algoritmos utilizados en redes sociales, procesamiento de lenguaje natural (PLN) que podemos encontrarlo en asistentes que utilizamos día a día como lo son Siri, Alexa, entre otros; sistemas de recomendaciones de compra, entre muchos otros.

Estas aplicaciones se basan en la implementación de algoritmos de machine learning, que, según Iberdrola (s.f.) se dividen en 3 principales categorías:

- **Aprendizaje supervisado:** Estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a datos que les permiten tomar decisiones o hacer predicciones, por ejemplo, un detector de spam que etiqueta un e-mail como spam o no dependiendo de los patrones que ha aprendido del histórico de correos.
- **Aprendizaje no supervisado:** Estos algoritmos no cuentan con un conocimiento previo, se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de cierta manera. Por ejemplo, en marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y de esa forma crear campañas publicitarias altamente segmentadas.
- **Aprendizaje por refuerzo:** Su objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto se refiere a que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error. En la actualidad se utiliza para posibilitar el reconocimiento facial, clasificar secuencias de ADN y para diagnósticos médicos.

Además, algunos expertos consideran una cuarta categoría: el aprendizaje semi-supervisado, el cual “es una rama del aprendizaje automático que combina el aprendizaje supervisado y no supervisado, utilizando datos etiquetados y sin etiquetar para entrenar modelos de inteligencia artificial (IA) para tareas de clasificación y regresión.” (Bergmann, 2023). Estas categorías, así como algunas adicionales, se explorarán en mayor detalle en este trabajo.

Un aspecto crucial al desarrollar modelos de machine learning, es la medición de la calidad de los mismos, esto se define como “medidas utilizadas para evaluar el rendimiento de los modelos de aprendizaje automático. Estas métricas se utilizan para comparar diferentes modelos y seleccionar el que tenga un mejor rendimiento

en base a las necesidades y objetivos del problema a resolver.” (DataBitAI, 2023), estas medidas o métricas son explicadas posteriormente.

Así como medimos la calidad de los modelos también contamos con métodos de mejora que nos permiten optimizar el rendimiento, los cuales “son técnicas y estrategias utilizadas para optimizar el rendimiento de los modelos, mejorando su capacidad para hacer predicciones precisas y generalizar a nuevos datos.” (Yildirim, 2020)

En resumen, el presente trabajo explorará en profundidad los diferentes tipos de machine learning, el tratamiento de datos, las métricas o medidas de calidad y métodos de mejora de los modelos de aprendizaje máquina. A través de esta exploración, se busca proporcionar una comprensión completa y aplicada del machine learning.

Contenido de la Actividad

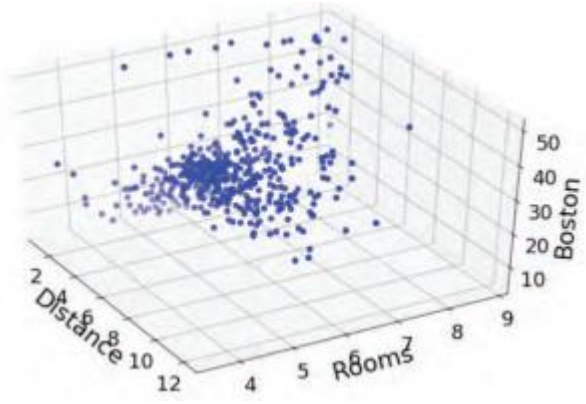
Tipos de machine learning

Con el objetivo de poder abordar cualquier tarea específica, el ingeniero de datos debe conocer algunos conceptos importantes de machine learning. Los conocimientos básicos incluyen la identificación de las tareas, empezando por la clasificación de los problemas de machine learning en alguno de los siguientes tipos:

- Aprendizaje supervisado
  - Regresión
  - Clasificación
- Aprendizaje no supervisado
  - Clustering (Agrupamiento)
  - Reducción de dimensiones
- Aprendizaje semi-supervisado
- Aprendizaje por refuerzo

El **aprendizaje supervisado** en machine learning se aplica cuando cada dato, o conjunto de datos de entrada (muestra) tiene asociada una etiqueta, partiendo del conjunto de datos se pueden usar diferentes algoritmos de clasificación con el objetivo de "entrenar" un modelo y poder, al acabar el entrenamiento, predecir la etiqueta correspondiente, éste es un problema de clasificación. De igual manera, podemos hacer uso de un conjunto de datos que contenga muestras con valores numéricos asociados, esto es conocido como un problema de regresión.

Por ejemplo, en el siguiente gráfico tridimensional se pueden observar los datos correspondientes a un problema de regresión, el objetivo es predecir el precio de una propiedad inmobiliaria en Boston atendiendo a diferentes tipos de información. En dicho gráfico solo se muestra el número de habitaciones y la distancia a la autopista como tipos de información, sin embargo, el término utilizado en machine learning para los tipos de datos es *característica o feature*.



Es así, que tanto número de habitaciones (rooms) como distancia a la autopista (distance) son características, que se representan en los ejes horizontales x e y. El eje vertical z muestra los valores objetivos para cada precio de venta de las propiedades inmobiliarias. En este caso, los valores objetivos no son etiquetas o nombres de categorías, si no valores numéricos. De esa forma,

afrontamos un problema de regresión, donde al suministrar un nuevo dato lo que se obtiene es la predicción esperada, el precio de la propiedad inmobiliaria.

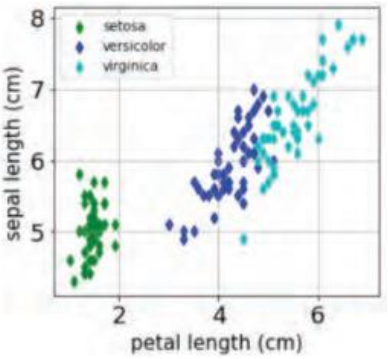
De los párrafos anteriores se puede destacar un concepto importante: el modelo de machine learning. Este es un elemento clave, ya que la mayoría de los algoritmos de machine learning crean un modelo a partir de los datos, dicho modelo puede ser tan simple como la solución lineal que mejor ajuste las muestras de origen a los valores objetivo, o mucho más complejo, como la búsqueda de factores ocultos que representen la información más importante que esta contenida en los datos.

La importancia del aprendizaje supervisado en machine learning está aumentando rápidamente debido a las siguientes razones:

1. Las nuevas oportunidades brindadas por el Internet de las Cosas (Internet of Things o IoT), de donde se pueden obtener cantidades masivas de datos etiquetados de manera automática.
2. Las redes sociales, en cuyos servidores se almacena una enorme cantidad de interacciones y cuyo número de aplicaciones, no para de crecer.
3. Los nuevos algoritmos, destinados a resolver diferentes tipos de aprendizaje supervisado, que hacen posible obtener resultados comerciales significativos, como lo son la conducción automática, reconocimiento facial, sistemas de recomendación, etc.
4. Las crecientes capacidades de procesamiento, particularmente las supercomputadoras paralelas y las unidades de procesamiento gráfico (GPUs).
5. La democratización del machine learning, por la que todos podemos trabajar con recursos altamente tecnológicos como Tensorflow o granjas de GPUs, así como con potentes APIs, entornos, IDEs, etc.

El **aprendizaje no supervisado** utiliza información no etiquetada. La aplicación más conocida del aprendizaje no supervisado es la de clustering o agrupamiento. El objetivo de la técnica de clustering es agrupar muestras.

El siguiente gráfico muestra un esquema típico de clustering, el cual contiene tres clusters (grupos o clases) correspondientes a 3 tipos diferentes de lirios, además podemos observar que es fácil diferenciar el tipo “setosa” de los otros dos, mientras que se podrían presentar dificultades para precisar los grupos de versicolor y virginica.



Un modelo de clustering podría proporcionar varios hiperplanos lineales de separación, mientras que un modelo de clustering diferente

podría proporcionar algunos elementos representativos (centroides), cuya área de influencia determina a qué cluster pertenece cada una de las muestras. Hay más tipos de modelos de clustering, pero los que se han indicado pueden ayudar a entender el concepto de modelo y el hecho de que diferentes algoritmos de machine learning pueden estar basados en diferentes tipos de modelos.

La reducción de dimensionalidad se usa habitualmente como una etapa de pre-procesamiento en algún otro tipo de labores de machine learning, principalmente en clasificación o regresión. Muchos escenarios reales aportan datos dispersos o datos que en su mayoría proporcionan muy poca información, un ejemplo de datos dispersos es la información que se maneja en un sistema de recomendación, donde los usuarios solamente compran, hacen clic, consumen o votan una proporción muy pequeña de los productos. Siendo capaces de comprimir los datos y donde los datos comprimidos contendrían casi toda la información, pero de una manera "condensada". Trabajar con esta información comprimida es mucho más eficiente y produce resultados más precisos.

El **aprendizaje semi-supervisado** trata con conjuntos de datos en los que una porción de los datos está etiquetada y el resto no. Normalmente, la cantidad de muestras etiquetadas es mucho más pequeña que las no etiquetadas. La mayoría de los algoritmos de aprendizaje semi-supervisado son una mezcla de métodos supervisados y no supervisados.

El **aprendizaje por refuerzo** es un área innovadora y con un gran futuro, ya que está inspirada en mecanismos naturales. En este caso, el algoritmo de aprendizaje recibe información de un entorno real o simulado, cuando el sistema realiza una acción es recompensado o penalizado, tal y como pasa con los seres vivos. Tales algoritmos de aprendizaje se denominan agentes y pueden aprender siguiendo los principios de la evolución natural. Los agentes aprenden estrategias, denominadas políticas, que maximizan las recompensas y minimizan las penalizaciones. La mayoría de los sistemas de inteligencia artificial actuales, están basados en el enfoque de aprendizaje por refuerzo.

Los métodos de machine learning también pueden ser clasificados como:

- Basado en modelos o basados en memoria
- Aprendizaje incremental o aprendizaje por lotes
- Aprendizaje superficial (shallow learning) o aprendizaje profundo (deep learning)

Los **algoritmos basados en memoria** (basados en instancias) toman las muestras de datos como entrada y procesan directamente la predicción o la clasificación, si se necesita una nueva predicción se procesa de nuevo, partiendo de las muestras de datos. Por otro lado, los **algoritmos basados en modelos** necesitan actualizar el modelo periódicamente, aunque el proceso de predicción es mucho más rápido.

Los **algoritmos de aprendizaje por lotes** (batch learning) siempre calculan el modelo desde el principio. Si disponemos de 2000 muestras, el proceso por lotes las usa todas para crear el modelo, cuando se aporten 300 nuevas muestras, el proceso por lotes crea el modelo desde el principio, usando las 2300 muestras y así sucesivamente.

Los **algoritmos de aprendizaje incremental** no crean sucesivos modelos desde el principio, si no que actualizan el modelo existente. Como se mencionó en el ejemplo anterior, los algoritmos de proceso por lotes obtendrían el modelo procesando desde cero las 2300 muestras, sin embargo, los algoritmos de aprendizaje incremental usarían las 300 nuevas muestras para cambiar los valores de la pendiente existente y el punto de corte. Los algoritmos incrementales presentan una importante ventaja: pueden ser usados como el núcleo de sistemas de machine learning escalables.

En el **aprendizaje superficial** (shallow learning), los parámetros (pendiente, punto de corte, etc.) se aprenden directamente de las características de las muestras de datos. En el **aprendizaje profundo** (deep learning) siempre existe una arquitectura con más de un nivel (capa), en el segundo nivel (y sucesivos) los parámetros "aprenden" de los resultados de las capas precedentes. Las arquitecturas deep learning pueden conformarse a base de varias capas con métodos de machine learning iguales o diferentes.

Tratando con datos

En machine learning, los datos son la base de todo; no habrá aprendizaje si no hay suficientes datos o éstos no son representativos o presentan información sesgada. Cuando la cantidad de datos es insuficiente, los algoritmos de machine learning no pueden generalizar los resultados, simplemente aprenden los patrones específicos de las muestras existentes. Este concepto es muy importante, el cual se denomina *sobreajuste (overfitting)* y se debe prevenir. Incluso si disponemos de suficiente cantidad de datos, éstos podrían no ser aceptables para algunos propósitos específicos de machine learning si no son representativos o están sesgados.

Incluso si vamos a usar algún conjunto de datos que contiene información representativa y no sesgada, podría presentar fallos si la información no es de calidad. Algunos ejemplos de información de mala calidad son:

- Cuando hay muchas muestras con valores vacíos en alguna característica.
- Valores atípicos (outliers): datos incorrectos provenientes de errores humanos, sensores de IoT que funcionan incorrectamente, errores en los programas.
- Datos incorrectos e inconsistentes.

También las características irrelevantes pueden estropear un proceso de machine learning, para poder obtener resultados adecuados necesitamos datos relevantes.

Cuando se trabaja en el campo de machine learning debemos diferenciar entre características de tipo continuo y características de tipo categórico (características discretas). Habitualmente resulta equivalente a la división previamente presentada de clasificación vs. regresión, pero aplicado a los datos de entrada.

Para representar valores categóricos se usa normalmente el proceso denominado *codificación one-hot (one-hot encoding)*. Básicamente, la codificación one-hot representa una variable categórica mediante el uso de varias características binarias nuevas (una nueva característica por cada valor categórico existente). La codificación one-hot hace más fácil para los métodos de machine learning la extracción de patrones y la relación entre características, con el objeto de predecir objetivos. El proceso denominado *binning (bucketing)* convierte una característica numérica en varias características binarias.

Muchos de los algoritmos de machine learning funcionan mejor cuando todas las características y el valor numérico objetivo están en el mismo rango. Además, evitando que los valores de entrada sean muy grandes se ayuda a los algoritmos a encontrar la solución y de la manera más rápida. Por esta razón, de manera habitual es necesario llevar a cabo un proceso de *normalización* sobre las muestras. La normalización más simple consiste en dividir cada muestra por el valor máximo de dichas muestras. La ecuación de normalización es:

$$x_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

La normalización puede fallar cuando hay información sesgada o existen valores atípicos (outliers). Si un valor atípico se cuela entre los datos de entrada, el efecto de la normalización se deteriorará. Por este motivo, la *estandarización (normalización z-socre o standardization)* es, muy a menudo, la mejor opción. La estandarización centra los datos en su media y los distribuye de acuerdo con el valor de la desviación típico. Un valor estandarizado puede ser interpretado como el número de desviaciones típicas que lo separada de la media.

$$x_i = \frac{x_i - \propto(X)}{\sigma(X)}$$

Para tratar con los valores vacíos en las características, los enfoques típicos son:

- Eliminar las muestras que contengan valores vacíos.
- Insertar los valores correctos cuando sea posible.
- Utilización de la técnica de *data imputation*.

La técnica de *data imputation* consiste en reemplazar los valores vacíos en una característica por alguna predicción.

Medición de la calidad

La regla de oro en metodología de la validación es: no medir la calidad con el mismo conjunto de datos empleado para obtener el modelo. Queremos evitar situaciones en las que el modelo aprende cada una de las muestras de datos, pero es incapaz de tratar datos nuevos. Esta es la situación de sobreajuste (overfitting) que se mencionó anteriormente. Es necesario dividir las muestras de datos y los valores objetivo en varios conjuntos:

- Entrenamiento
- Validación
- Test

El conjunto de datos de entrenamiento es el mayor, un valor típico es que sea el 80% del conjunto total de muestras, se usa para entrenar el modelo. El conjunto de validación contiene las muestras usadas para mejorar el modelo a base de realizar un ajuste fino en los híper-parámetros (valores que controlan diferentes variaciones en el funcionamiento de los algoritmos). Por último, el conjunto de test (o pruebas) nos permitirá medir la calidad del modelo utilizando las muestras que no hayan sido usadas para entrenar el modelo o para mejorarlo. Debemos asegurarnos de que los tres conjuntos tienen una distribución similar de los datos, que contienen datos representativos y que su intersección es nula. Una vez que hayamos entrenado el modelo con el conjunto de muestras de entrenamiento y lo hayamos mejorado usando el conjunto de muestras de validación, podremos usar las muestras de test para hacer predicciones con estos datos que no han sido previamente procesados.

Dos medidas de calidad bien conocidas son el *Error Medio Absoluto (Mean Absolute Error o MAE)* y el *Error Cuadrático Medio (Mean Squared Differences o MSD)*. Ambas medidas de calidad penalizan la distancia que hay entre cada valor de una predicción y el valor objetivo. Sus ecuaciones son:

$$MSD(X, y) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - p(X_i))^2$$

$$MAE(X, y) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - p(X_i)|$$

Ambas ecuaciones devuelven el error medio cometido en las N muestras de test. El error en ambas ecuaciones es la diferencia entre el valor objetivo real y el valor predicho para la muestra X. MSD penaliza los errores con valores grandes (los eleva al cuadrado) mucho más de lo que lo hace el MAE.

Mientras que las medidas anteriores devuelven valores absolutos, la ecuación del coeficiente R^2 (R^2 score) devuelve valores relativos. Esta medida de calidad proporciona un valor entre 0 y 1, donde el 1 significa una predicción perfecta y 0 se

corresponde al modelo de regresión más simple, predecir la media del conjunto de muestras de test. Al coeficiente R^2 se le denomina también *coeficiente de determinación*. La ecuación de R^2 es:

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1}(y_i - f(x_i))^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}$$

Donde $f(x_i)$ es la predicción del regresor para la muestra x_i

Los algoritmos de clasificación también hacen predicciones, las medidas de calidad de regresión podrían utilizarse también en clasificación asignando ceros a los errores de las muestras clasificadas correctamente y asignando unos a los errores de las muestras clasificadas erróneamente. Existe un enfoque diferente que hace posible diseñar medidas de calidad más específicas para la clasificación. Si centramos nuestra atención en la clasificación binaria, en la matriz de confusión solamente pueden existir dos clases (grupos):

Predicción	clase 0 (positivo)	clase 1 (negativo)
clase 0	verdadero positivo (TP)	falso negativo (FN)
clase1	falso positivo (FP)	verdadero negativo (TN)

La medida de calidad denominada *Precision* nos muestra la proporción de aciertos en la predicción. La medida de calidad *Recall* centra su atención en el número relativo de muestras positivas. Mientras que *Precision* nos proporciona un valor de calidad relativo al número total de predicciones realizadas, *Recall* nos proporciona un valor de calidad relativo al número total de muestras positivas. Merece la pena mantener niveles altos de *Precision* cuando el número de predicciones crece y merece la pena mantener niveles altos de *Recall* cuando el número de muestras positivas es bajo.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Existe una medida de calidad de clasificación que une Precision y Recall, es la F1:

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

En el caso de que todas las clases sean igualmente importantes se puede usar la siguiente medida de calidad de exactitud (accuracy):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Para medir la calidad de la clasificación también se usa la curva ROC, el cual combina los valores verdaderos positivos con los falsos positivos, definida como:

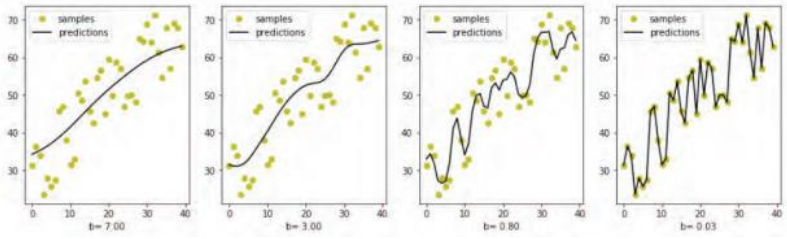
$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

Cuanto mayor sea el área bajo la curva ROC, mejor será el modelo. Los valores obtenidos deben ser mayor de 0.5, ya que 0.5 se corresponde con un clasificador aleatorio.

Mejora del modelo

Para mejorar un modelo de machine learning se debe medir la calidad que obtiene al actuar sobre el conjunto de muestras de test (pruebas). Primero se usa el conjunto de datos de entrenamiento para crear el modelo, y podemos entonces obtener sus medidas de calidad (aplicadas a las muestras de entrenamiento). Si las predicciones del modelo presentan muchos errores, las medidas de calidad devolverán valores muy bajos. Esto significa que los datos no son suficientemente numerosos o presentan muchos valores atípicos o las características no son relevantes o que el algoritmo de aprendizaje utilizado no está indicado para la presente tarea (normalmente por ser muy simple) o los híper-parámetros de aprendizaje elegidos no son correctos. En estos casos, el modelo sufre *subajuste (underfitting)*.

La situación opuesta al subajuste es la del sobreajuste (overfitting). Un modelo presenta sobreajuste cuando predice con suficiente exactitud el conjunto de datos de entrenamiento y, sin embargo, no puede predecir adecuadamente las muestras de test, no generaliza. En la siguiente figura se visualizan varios modelos obtenidos a partir de entrenamientos que utilizan el mismo conjunto de muestras, los dos gráficos de la izquierda representan modelos más simples, que son capaces de generalizar para predecir de manera adecuada nuevas muestras. Los dos gráficos de la derecha representan modelos más complejos, que no son capaces de generalizar, presentan situaciones de sobreajuste.



Para evitar el sobreajuste podemos entrenar el modelo usando más datos, en este caso será más difícil para el algoritmo ajustar todos los datos. Otra posibilidad es usar un modelo más simple (quizás un modelo lineal). Reducir la dimensionalidad de las muestras también puede funcionar, puede ser llevado a cabo usando técnicas de reducción de dimensiones.

Por último, regularizar el modelo es, normalmente, la solución más simple y más eficiente. La *regularización* se puede llevar a cabo de diferentes maneras, dependiendo de cada algoritmo de machine learning que se esté usando, pero su concepto subyacente es siempre simplificar el modelo. A menudo, la regularización significa mantener los pesos de aprendizaje tan pequeños como sea posible, limita la importancia de algunas características, simplificando el modelo.

Otra técnica específica de regularización es la denominada *data augmentation* (enriquecimiento de datos). Ya que incrementar el número de muestras de entrenamiento reduce el sobreajuste. Aunque la técnica de data augmentation puede ser usada teóricamente en cualquier ámbito, se aplica especialmente en clasificación de imágenes.

Conclusiones

En conclusión, es importante reconocer la importancia y el impacto del machine learning en diversos ámbitos de la tecnología y la ciencia de datos. Conocer los diferentes tipos de machine learning, como lo son el aprendizaje supervisado, semi-supervisado, no supervisado y por refuerzo, permite seleccionar la mejor aproximación o categoría para resolver problemas específicos de la vida cotidiana o vida laboral. Además, se debe destacar que el tratamiento adecuado de los datos es crucial para evitar el sobreajuste o subajuste, los cuales son problemas que pueden comprometer la capacidad de un modelo para generalizar y ofrecer predicciones precisas con datos nuevos. La medición de la calidad de los modelos, a través de métricas y técnicas de validación, es esencial para garantizar que el rendimiento del modelo esté alineado con los objetivos establecidos. Finalmente, los métodos de mejora de los modelos, como la regularización, son herramientas indispensables para alcanzar un desempeño óptimo en aplicaciones del mundo real. En conjunto, estos aspectos destacan la importancia de una comprensión profunda y matizada del machine learning, no solo como una tecnología que ha cobrado gran relevancia actualmente, sino como un campo en constante evolución que requiere un enfoque riguroso y meticuloso para su implementación exitosa.

Referencias

- Bergmann, D. (2023, 12 de diciembre). *¿Qué es el aprendizaje semisupervisado?*. IBM. Recuperado el 27 de Agosto de 2024 de: https://www.ibm.com/mx-es/topics/semi-supervised-learning
- Bobadilla, J. (2020). *Machine Learning y Deep Learning Usando Python, Scikit y Keras* (1ª ed.). Ra-Ma Editorial; Ediciones de la U.
- DataBitAI (2023, 17 de abril). *Métricas de Evaluación en Machine Learning*. DataBitAI. Recuperado el 27 de Agosto de 2024 de: https://databitai.com/machine-learning/metricas-de-evaluacion-en-machine-learning/
- Iberdrola (s.f.). *Descubre los principales beneficios del 'Machine Learning'*. Iberdrola. Recuperado el 27 de Agosto de 2024 de: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico
- Yildirim, S. (2020, 21 de julio). *Mejorar El Rendimiento De Un Modelo De Aprendizaje Automático*. DataSource.ai. Recuperado el 27 de Agosto de 2024 de: https://www.datasource.ai/es/data-science-articles/mejorar-el-rendimiento-de-un-modelo-de-aprendizaje-automatico