Métricas de evaluación para machine learning: precisión y robustez

Breve descripción:

Este componente ofrece una guía práctica sobre las métricas de evaluación en machine learning, enfocándose en la precisión y robustez de los modelos. Cubre desde conceptos fundamentales y técnicas de ensamblado como Bagging y Boosting, hasta métricas avanzadas y estrategias de ajuste, incluyendo comunicación efectiva de resultados para garantizar modelos confiables y aplicables.

Diciembre 2024

Tabla de contenido

Introducción

La evaluación de modelos de machine learning es esencial para garantizar su efectividad y aplicabilidad en escenarios reales. Medir la precisión y la robustez de un modelo valida su desempeño y asegura su capacidad para generalizar ante datos nuevos y variados.

Este componente aborda de manera sistemática las métricas de evaluación fundamentales para machine learning, enfocándose en la precisión y la robustez. Se exploran desde conceptos básicos y técnicas de ensamblado como Bagging, Random forest y Boosting, hasta métricas avanzadas como la matriz de confusión, las curvas ROC y el manejo de datos desbalanceados. Además, se incluyen estrategias para ajustar y optimizar modelos, así como métodos efectivos para comunicar y documentar los resultados obtenidos.

A lo largo de este componente, se combinan conceptos teóricos con ejemplos prácticos, proporcionando las herramientas necesarias para que los lectores puedan aplicar las métricas de evaluación de manera efectiva en sus propios proyectos. Se enfatiza la importancia de una evaluación exhaustiva para desarrollar modelos confiables y robustos, capaces de enfrentar los desafíos del mundo real.

La integración de técnicas de evaluación con prácticas de comunicación y documentación garantiza que los resultados no solo sean precisos, sino también comprensibles y utilizables por diferentes audiencias. Como dice un principio fundamental en machine learning: "Un modelo es tan bueno como su evaluación".

¡Le invitamos a descubrir las métricas y metodologías clave para evaluar y mejorar sus modelos de machine learning, asegurando su precisión y robustez en cada aplicación!

Video 1. Métricas de evaluación para machine learning: precisión y robustez

Enlace de reproducción del video

Síntesis del video: Métricas de evaluación para machine learning: precisión y robustez
En el componente formativo «Métricas de evaluación para machine learning: precisión y robustez» se abordan las herramientas esenciales para medir y garantizar el rendimiento de los modelos de inteligencia artificial. Comprender y aplicar adecuadamente estas métricas es fundamental para desarrollar modelos

precisos y robustos, capaces de generalizar eficazmente en distintos contextos y datos.

Durante el desarrollo de este componente, se busca una comprensión profunda de las métricas clave que permiten evaluar la efectividad de los modelos de machine learning. Se inicia con una introducción a la importancia de las métricas de evaluación, destacando cómo estas influyen en la selección y mejora de los modelos.

El componente explora técnicas de ensamblado de modelos, como Bagging, Random Forest y Boosting, que son fundamentales para aumentar la precisión y reducir la variabilidad de las predicciones. A continuación, se profundiza en métricas avanzadas como la matriz de confusión, las curvas ROC y el área bajo la curva (AUC), además del manejo de datos desbalanceados, aspectos determinantes para una evaluación detallada y equilibrada.

Asimismo, se abordan estrategias para probar y ajustar los modelos, incluyendo la optimización de hiperparámetros y la ingeniería de características, garantizando así un rendimiento óptimo. Finalmente, se enfatiza la importancia de la comunicación y documentación de resultados, utilizando técnicas de storytelling y la elaboración de informes técnicos que faciliten la comprensión y aplicación de los hallazgos por parte de diferentes audiencias.

¡Bienvenido al mundo de las métricas de evaluación en machine learning! Le invitamos a descubrir cómo medir, analizar y mejorar tus modelos para alcanzar niveles superiores de precisión y robustez!

1. Introducción a la evaluación de modelos de machine learning

Este tema proporciona una base para comprender la importancia y los conceptos clave de las métricas de evaluación en machine learning. Se explora cómo la precisión y la robustez son pilares fundamentales en la evaluación de modelos y se discuten los desafíos comunes que pueden surgir durante este proceso. Al abordar estos desafíos con estrategias y soluciones adecuadas, podemos garantizar que nuestros modelos sean más precisos, confiables y aplicables en situaciones del mundo real.

1.1. La importancia de las métricas de evaluación

En el mundo del machine learning, construir modelos capaces de realizar predicciones es solo una parte del proceso. Es esencial evaluar su rendimiento de manera objetiva y sistemática para garantizar que cumplan con los objetivos planteados y sean útiles en aplicaciones prácticas. Las métricas de evaluación son herramientas fundamentales que nos permiten cuantificar el desempeño de un modelo, identificar áreas de mejora y compararlo con otros modelos o enfoques.

Sin una evaluación adecuada, es imposible determinar si un modelo es realmente efectivo o si simplemente parece funcionar bien debido a coincidencias en los datos de entrenamiento. Además, las métricas de evaluación facilitan la comunicación de resultados a equipos multidisciplinarios, permitiendo que todos comprendan el valor y las limitaciones del modelo desarrollado.

Importancia en el ciclo de vida del modelo

La evaluación es determinante en varias etapas del ciclo de vida de un modelo:

- **Durante el desarrollo**: para seleccionar el mejor modelo entre múltiples candidatos.
- **Antes de la implementación**: para asegurar que el modelo cumple con los requisitos de precisión y robustez necesarios.
- **Después de la implementación**: para monitorear el rendimiento del modelo en producción y detectar degradaciones a lo largo del tiempo.

1.2. Conceptos clave: precisión y robustez

Antes de profundizar en las métricas específicas, es fundamental entender dos conceptos esenciales que guían la evaluación de modelos:

- **Precisión**: se refiere a la exactitud con la que el modelo realiza predicciones correctas. Un modelo preciso tiene un alto porcentaje de aciertos, lo que indica un buen desempeño en términos de exactitud. Sin embargo, la precisión por sí sola puede ser engañosa en conjuntos de datos desbalanceados.
- **Robustez**: indica la capacidad del modelo para mantener un rendimiento consistente ante variaciones en los datos de entrada, como ruido, valores atípicos o cambios en la distribución de los datos (`concept drift`). Un modelo robusto es resistente a perturbaciones y generaliza bien a datos no vistos, lo cual es fundamental para aplicaciones en entornos dinámicos.

Relación entre precisión y robustez

Aunque un modelo puede ser muy preciso en un conjunto de datos específico, si carece de robustez, su rendimiento puede degradarse significativamente cuando se enfrenta a datos nuevos o cambiantes. Por lo tanto, es esencial equilibrar ambos aspectos para desarrollar modelos confiables y duraderos.

1.3. Visión general de las métricas comunes

Existen diversas métricas para evaluar modelos de Machine learning, cada una adecuada para diferentes tipos de problemas y contextos. A continuación, se presenta una visión general de las métricas más utilizadas:

Tabla 1. Resumen de métricas de evaluación comunes

Tipo de problema	Métricas comunes	Descripción
Clasificación.	Precisión (Accuracy), Precision, Recall, F1-Score.	Miden el desempeño en la clasificación de categorías discretas, evaluando verdaderos positivos, falsos positivos, etc.
Regresión.	MSE, RMSE, MAE, R^2.	Evalúan la diferencia entre los valores predichos y los valores reales continuos, proporcionando una medida del error promedio y la calidad del ajuste.
Datos Desbalanceados.	AUC-ROC, Curva Precision-Recall.	Analizan el rendimiento en situaciones donde las clases están desbalanceadas, enfocándose en la capacidad del modelo para distinguir entre clases minoritarias y mayoritarias.
Clustering.	Índice de Silueta, SSE.	Miden la calidad de los grupos formados en problemas de agrupamiento sin supervisión, evaluando la cohesión interna y la separación entre clusters.

Fuente. OIT, 2024.

Ejemplos de Aplicación

Ejemplo: diseñar un programa que verifique si un número es primo. El pensamiento algorítmico te guiará a:

- **Clasificación**: en la detección de spam, podríamos usar la precisión para saber qué porcentaje de correos electrónicos fueron clasificados correctamente. Sin embargo, también es importante considerar la **precisión (Precision) y el Recall** para entender cuántos correos legítimos fueron marcados incorrectamente como spam y cuántos correos spam fueron pasados por alto.
- **Regresión**: al predecir los precios de viviendas, métricas como el **RMSE** proporcionan una idea de cuánto, en promedio, se desvían las predicciones del modelo respecto a los valores reales, lo que es muy importante para entender la fiabilidad de las estimaciones.

1.4. Desafíos en la evaluación de modelos

La evaluación de modelos de machine learning presenta varios desafíos que deben ser abordados para asegurar resultados confiables y útiles.

a) Datos desbalanceados

Cuando una clase predomina significativamente sobre otras, algunas métricas pueden ser engañosas. Por ejemplo, en un problema de detección de fraude donde solo el 1 % de las transacciones son fraudulentas, un modelo que siempre predice "no fraude" tendría una precisión del 99 %, pero sería inútil para detectar fraudes reales.

Soluciones:

- Utilizar métricas como el **Recall, Precisión (Precision) y F1-Score** que ofrecen una mejor perspectiva en conjuntos de datos desbalanceados.
- Aplicar técnicas de muestreo como submuestreo de la clase mayoritaria o **sobremuestreo** de la clase minoritaria.
- Emplear algoritmos diseñados para manejar desequilibrios de clases, como **Balanced Random forest o Adaptive Boosting**.

b) Sobreajuste y subajuste

Sobreajuste (Overfitting): ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento, capturando ruido y patrones irrelevantes. Esto conduce a un pobre rendimiento en datos nuevos.

Subajuste (Underfitting): sucede cuando el modelo es demasiado simple para capturar la estructura subyacente de los datos, resultando en un rendimiento deficiente tanto en el conjunto de entrenamiento como en nuevos datos.

Soluciones:

- **Validación cruzada**: dividir el conjunto de datos en múltiples subconjuntos para validar el rendimiento del modelo de manera más robusta.
- **Regularización**: aplicar técnicas como Lasso o Ridge para penalizar la complejidad del modelo.
- **Ajuste de hiperparámetros**: optimizar parámetros clave del modelo para encontrar el equilibrio adecuado entre sesgo y varianza.

c) **Interpretabilidad vs. Complejidad**

Modelos más complejos, como las redes neuronales profundas, pueden ofrecer mayor precisión pero son menos interpretables. En ciertos dominios, como la medicina o las finanzas, la interpretabilidad es importante para cumplir con regulaciones y generar confianza en los usuarios.

Soluciones:

- Optar por modelos más interpretables como árboles de decisión o regresiones lineales cuando sea apropiado.
- Utilizar técnicas de interpretabilidad como **LIME** o **SHAP** para explicar las predicciones de modelos complejos.
- Equilibrar la necesidad de precisión con la interpretabilidad según los requisitos del proyecto.

d) **Costos y consecuencias de los errores**

No todos los errores tienen el mismo impacto. Por ejemplo, en una aplicación médica, un falso negativo (no detectar una enfermedad cuando está presente) puede ser más grave que un falso positivo (diagnosticar una enfermedad cuando no existe).

Soluciones:

- Definir una **función de costo** personalizada que refleje las consecuencias reales de los diferentes tipos de errores.
- Ajustar los umbrales de decisión del modelo para minimizar los errores más críticos.
- Priorizar métricas que capturen la gravedad de los errores, como el **Recall** en casos donde los falsos negativos son más costosos.

e) **Variabilidad en los datos**

Los datos pueden cambiar con el tiempo debido a tendencias, estacionalidades o cambios en el comportamiento. Esto puede causar que el rendimiento del modelo disminuya si no se actualiza regularmente.

Soluciones:

- Implementar un sistema de monitoreo para detectar cambios en el rendimiento del modelo.
- Programar reentrenamientos periódicos del modelo con datos actualizados.
- Utilizar técnicas de aprendizaje en línea que permiten al modelo adaptarse continuamente a nuevos datos.

f) **Limitaciones de las métricas**

Cada métrica tiene sus propias limitaciones y puede no reflejar completamente el rendimiento del modelo en todos los aspectos.

Soluciones:

- Utilizar múltiples métricas para obtener una visión más completa del rendimiento.
- Analizar las curvas ROC y `Precision-Recall` para entender mejor el comportamiento del modelo en diferentes umbrales.
- Realizar análisis cualitativos adicionales, como inspeccionar casos mal clasificados para identificar patrones o problemas subyacentes.

Conclusión

Enfrentar estos desafíos requiere un enfoque crítico y cuidadoso en la selección y aplicación de las métricas de evaluación. Es esencial comprender no solo qué miden las

métricas, sino también cómo interpretarlas en el contexto específico del problema que se está abordando.

En los siguientes temas, profundizaremos en técnicas avanzadas y metodologías específicas para validar y mejorar modelos, incluyendo el uso de técnicas de ensamblado y la comunicación efectiva de resultados a través de informes y storytelling.

2. Técnicas de ensamblado de modelos de inteligencia artificial

Los métodos de ensamblado representan un enfoque poderoso en el campo del machine learning para mejorar el rendimiento de los modelos predictivos. Al combinar múltiples modelos base, es posible superar las limitaciones individuales y obtener predicciones más precisas y robustas. Este tema introduce los fundamentos de las técnicas de ensamblado, explorando cómo algoritmos como Bagging, Random forest y Boosting pueden utilizarse para reducir la variabilidad, disminuir el sesgo y manejar datos complejos. A través de esta exploración, proporcionaremos una comprensión profunda de cómo y cuándo aplicar estos métodos para optimizar modelos de inteligencia artificial en diversas aplicaciones.

2.1. Fundamentos de los métodos de ensamblado

Los métodos de ensamblado, o ensemble methods, son técnicas que combinan múltiples modelos de aprendizaje automático para mejorar el rendimiento predictivo en comparación con modelos individuales. la idea central es que al combinar varios modelos, se puede reducir la variancia, el sesgo o mejorar las predicciones generales.

Algunas ventajas de los métodos de ensamblado son:

- **Mejora de la precisión**: al combinar modelos, se suelen obtener predicciones más precisas.
- **Reducción de la variabilidad**: se mitiga el efecto de modelos que podrían haber sobreajustado los datos.
- **Robustez**: los ensamblados son generalmente más resistentes al ruido y a los datos atípicos.

Tipos principales de métodos de ensamblado:

- **Promediación (averaging)**: los modelos se combinan promediando sus predicciones. ejemplos incluyen Bagging y random forest.
- **Boosting**: los modelos se construyen secuencialmente, y cada modelo intenta corregir los errores del anterior.
- **Stacking**: combina predicciones de múltiples modelos a través de un modelo meta.

2.2. Bagging (bootstrap aggregating)

El Bagging es una técnica que mejora la estabilidad y precisión de los algoritmos de machine learning al reducir la variancia y ayudar a evitar el sobreajuste consiste en:

- Muestreo aleatorio con reemplazo: se crean múltiples subconjuntos de datos a partir del conjunto de entrenamiento original.
- Entrenamiento de modelos independientes: se entrena un modelo separado en cada subconjunto.
- Agregación de predicciones: las predicciones se combinan promediando (para regresión) o mediante votación mayoritaria (para clasificación).

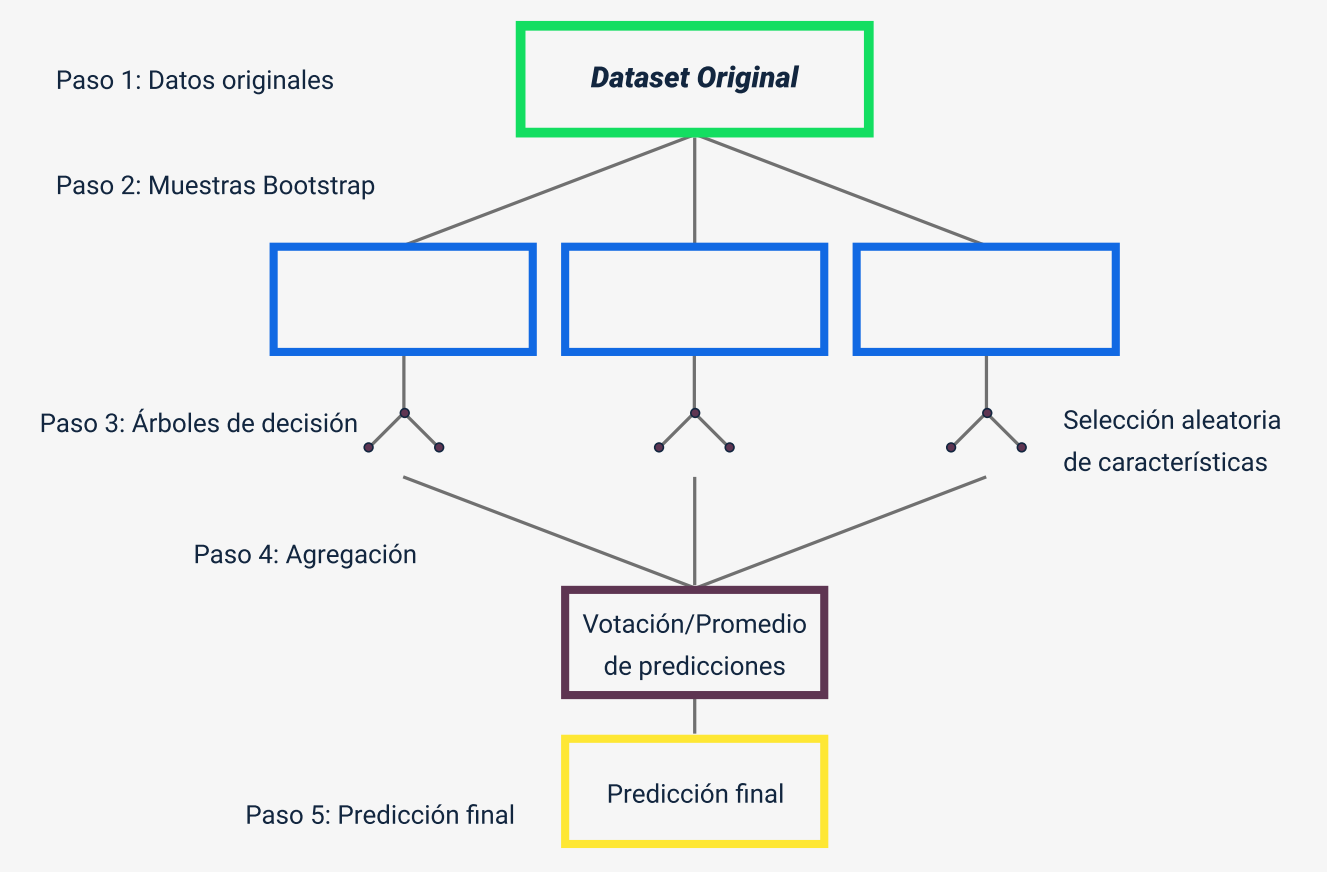
Ventajas del Bagging:

- Reducción de variancia: al promediar múltiples modelos, se reduce la variabilidad de las predicciones.
- Simplicidad: fácil de implementar y aplicar a diversos algoritmos.

2.3. Random forest

El random forest es una extensión del Bagging que utiliza árboles de decisión como modelos base y agrega aleatoriedad adicional.

Figura 1. Arquitectura Random forest



Fuente. OIT, 2024.

El método Random forest se caracteriza por lo siguiente:

- **Selección aleatoria de características**: en cada división del árbol, se considera un subconjunto aleatorio de características en lugar de todas.
- **Construcción de múltiples árboles de decisión**: se construyen numerosos árboles utilizando diferentes subconjuntos de datos y características.

- **Agregación de resultados**: las predicciones de los árboles se combinan por promedio (regresión) o votación mayoritaria (clasificación).

Ventajas del random forest:

- **Mejora de la precisión**: suele superar a los árboles de decisión individuales.
- **Reducción de la correlación entre modelos**: la selección aleatoria de características disminuye la correlación entre los árboles.
- **Manejo de datos de alta dimensionalidad**: funciona bien con muchos atributos y puede estimar variables importantes.

2.4. Métodos de Boosting

El Boosting es otra familia de métodos de ensamblado que construye modelos secuencialmente, enfocándose en corregir los errores de los modelos anteriores. a diferencia del Bagging, que construye modelos independientes, el Boosting da más peso a las observaciones que fueron mal predichas en iteraciones anteriores. Los principales algoritmos de Boosting son los siguientes:

a) Adaboost (adaptive Boosting):

- **Funcionamiento**: inicializa pesos iguales para todas las observaciones. en cada iteración, se ajusta un modelo y se actualizan los pesos aumentando aquellos de las observaciones mal clasificadas.
- **Agregación**: los modelos se combinan en una suma ponderada donde los pesos se determinan según la precisión de cada modelo.

b) Gradient Boosting:

- **Funcionamiento**: los modelos se construyen secuencialmente, y cada uno intenta minimizar el error residual del modelo anterior.
- **Agregación**: se suman las predicciones de todos los modelos anteriores para mejorar gradualmente el rendimiento.

c) XGBoost (eXtreme Gradient Boosting):

- **Mejoras sobre gradient Boosting**: optimización de velocidad y rendimiento mediante técnicas como paralelización, regularización y manejo eficiente de memoria.
- **Uso común**: muy popular en competiciones de machine learning por su alta precisión y eficiencia.

Ventajas del Boosting:

- Alta precisión: tiende a producir modelos con alto rendimiento predictivo.
- Flexibilidad: puede utilizarse con diferentes tipos de modelos base.
- Manejo de datos desbalanceados: al enfocarse en observaciones difíciles, puede mejorar la predicción de clases minoritarias.

Consideraciones al usar Boosting:

- Propensión al sobreajuste: debido a su naturaleza, es importante regularizar y ajustar correctamente los hiperparámetros.
- Mayor tiempo de entrenamiento: los modelos se construyen secuencialmente, lo que puede aumentar el tiempo de cómputo.

2.5. Evaluación de modelos ensamblados

La evaluación de modelos ensamblados sigue los mismos principios que la de modelos individuales, pero con algunas consideraciones adicionales.

a) Validación cruzada

Es fundamental utilizar técnicas de validación cruzada para evaluar el rendimiento real del modelo y evitar el sobreajuste.

- **k-fold cross-validation**: se divide el conjunto de datos en k pliegues y se realiza entrenamiento y validación k veces, cambiando el pliegue de validación cada vez.
- **stratified k-fold**: similar al anterior, pero mantiene la proporción de clases en cada pliegue, útil en problemas de clasificación desbalanceados.

b) Importancia de variables

Los métodos de ensamblado como `random forest` permiten evaluar la importancia de las variables, lo cual es útil para:

- **Interpretación del modelo**: identificar qué variables contribuyen más a las predicciones.
- **Reducción de dimensionalidad**: eliminar variables irrelevantes para simplificar el modelo.

c) Métricas específicas

Además de las métricas comunes, es posible que desee evaluar:

- **Out-of-bag error (oob error)**: en `Bagging` y `random forest`, es una estimación del error de generalización calculado utilizando las muestras no incluidas en cada `bootstrap`.

- **Curvas de aprendizaje**: gráficas que muestran cómo cambia el rendimiento del modelo con respecto al número de árboles o iteraciones, útil para determinar cuándo se alcanza un rendimiento óptimo.

Tabla 2. Comparación de métodos de ensamblado

Características	Bagging	Random forest	Boosting
Construcción de modelos	Paralelo	Paralelo	Secuencial
Base learner	Cualquier modelo	Árboles de decisión	Cualquier modelo
Reducción de variancia	Sí	Sí	Sí
Reducción de sesgo	No	No	Sí
Selección aleatoria de features	No	Sí	No
Propenso al sobreajuste	Menos	Menos	Más si no se regula

Fuente. OIT, 2024.

2.6. Casos prácticos y aplicaciones

Finalmente, es útil discutir casos prácticos donde los métodos de ensamblado han demostrado ser efectivos:

- **Detección de fraude**: Boosting es utilizado para mejorar la detección de transacciones fraudulentas.

- **Predicción en medicina**: random forest ha sido aplicado en la predicción de enfermedades debido a su capacidad para manejar datos complejos y variables.
- **Sistemas de recomendación**: los ensamblados pueden mejorar la precisión de recomendaciones al combinar diferentes modelos.

Conclusiones

Los métodos de ensamblado son herramientas poderosas en el arsenal de machine learning que permiten mejorar la precisión y robustez de los modelos, al combinar múltiples modelos, es posible superar las limitaciones de modelos individuales y obtener mejores resultados en una amplia gama de aplicaciones, sin embargo, es importante comprender las diferencias entre las distintas técnicas de ensamblado y saber cuándo aplicarlas adecuadamente. En el próximo tema, profundizaremos en las métricas avanzadas de evaluación y en cómo ajustar los modelos para optimizar su rendimiento.

3. Métricas avanzadas de evaluación y ajuste de modelos

3.1. Matriz de confusión y análisis de errores

La matriz de confusión es una herramienta esencial en la evaluación de modelos de clasificación. Permite visualizar el desempeño del modelo al mostrar las predicciones correctas e incorrectas en cada categoría. Esta matriz es fundamental para comprender no solo cuántas predicciones fueron acertadas, sino también para identificar dónde y por qué se producen los errores.

Imaginemos un modelo que clasifica imágenes de gatos y perros. La matriz de confusión nos indicaría cuántas imágenes de gatos fueron correctamente clasificadas como gatos (verdaderos positivos), cuántas fueron incorrectamente clasificadas como perros (falsos negativos), y lo mismo para las imágenes de perros. Este análisis detallado ayuda a identificar patrones en los errores, como si el modelo confunde más a menudo un tipo específico de perro con un gato, lo que podría indicar la necesidad de más datos de entrenamiento o ajustes en el preprocesamiento.

El análisis de errores va más allá de contar las predicciones incorrectas; implica entender las causas subyacentes. Por ejemplo, si el modelo confunde consistentemente gatos negros con perros, podría deberse a condiciones de iluminación en las imágenes o a características visuales similares. Al profundizar en estos errores, podemos realizar mejoras específicas, como ajustar el balance de clases en el conjunto de datos, aplicar técnicas de aumento de datos o refinar la arquitectura del modelo.

3.2. Curvas ROC y área bajo la curva (AUC)

Las curvas ROC (Receiver Operating Characteristic) son una representación gráfica que permite evaluar el rendimiento de un modelo de clasificación binaria a

diferentes umbrales de discriminación. En el eje vertical se representa la tasa de verdaderos positivos (sensibilidad) y en el eje horizontal la tasa de falsos positivos (1 - especificidad). Al trazar la curva ROC, podemos observar cómo varía el desempeño del modelo al modificar el umbral de decisión.

El área bajo la curva (AUC) es una medida que resume el rendimiento del modelo en una sola cifra. Un AUC cercano a 1 indica un modelo con excelente capacidad de discriminación entre las clases, mientras que un AUC de 0.5 sugiere que el modelo no es mejor que una decisión al azar. La ventaja del AUC es que es independiente del umbral elegido, proporcionando una evaluación general del modelo.

Por ejemplo, en el diagnóstico médico, donde es tan sensible detectar enfermedades graves, las curvas ROC y el AUC permiten seleccionar el umbral que equilibra adecuadamente la sensibilidad y la especificidad según las necesidades clínicas. Si priorizamos minimizar los falsos negativos, podríamos elegir un umbral que aumente la sensibilidad, aceptando un mayor número de falsos positivos.

3.3. Manejo de datos desbalanceados

En muchos problemas reales, las clases en los datos no están equilibradas. Esto significa que una clase puede estar representada por una cantidad significativamente menor de ejemplos que la otra. Un ejemplo común es la detección de fraudes financieros, donde las transacciones fraudulentas son una fracción muy pequeña del total. Este desbalance puede causar que los modelos aprendan a predecir siempre la clase mayoritaria, ignorando la minoritaria.

Para abordar este desafío, es importante aplicar estrategias que compensen el desbalance. Una de ellas es el muestreo, que puede ser de dos tipos: submuestreo de

la clase mayoritaria o sobremuestreo de la clase minoritaria. El submuestreo implica reducir el número de ejemplos de la clase dominante para equilibrar el conjunto de datos, mientras que el sobremuestreo consiste en aumentar el número de ejemplos de la clase minoritaria, ya sea replicando datos existentes o generando nuevos ejemplos sintéticos mediante técnicas como SMOTE (Synthetic Minority Over-sampling Technique).

Otra estrategia es utilizar algoritmos que sean intrínsecamente robustos a los datos desbalanceados. Los métodos de ensamblado, como el Boosting, pueden dar mayor peso a las observaciones de la clase minoritaria, mejorando su detección. Además, es clave seleccionar métricas de evaluación adecuadas. En lugar de confiar únicamente en la precisión global, es preferible utilizar métricas como el recall, la precisión (Precision) y el F1-score, que ofrecen una visión más equilibrada del rendimiento en ambas clases.

3.4. Probar y ajustar el modelo

El proceso de probar y ajustar el modelo es una etapa iterativa que busca optimizar su rendimiento. Después de entrenar el modelo inicial, es fundamental evaluarlo en un conjunto de validación para medir su capacidad de generalización. Esta evaluación permite identificar problemas como el sobreajuste, donde el modelo aprende demasiado bien los detalles del conjunto de entrenamiento y no generaliza bien a datos nuevos.

El ajuste del modelo puede involucrar varias acciones. Una de ellas es la optimización de hiperparámetros, que son parámetros del modelo que no se aprenden directamente durante el entrenamiento, como la profundidad máxima de un árbol de decisión o la tasa de aprendizaje en un algoritmo de Boosting. La búsqueda de los

valores óptimos para estos hiperparámetros puede realizarse mediante métodos sistemáticos como la búsqueda en cuadrícula (`grid search`) o algoritmos más avanzados como la optimización bayesiana.

Además, es importante considerar la ingeniería de características. Esto implica seleccionar, transformar o crear nuevas variables que puedan mejorar el rendimiento del modelo. Por ejemplo, en un problema de predicción de ventas, podríamos crear una nueva característica que represente la estacionalidad, como el mes del año o si es un día festivo.

La validación cruzada es otra técnica clave en esta etapa. Al dividir el conjunto de datos en múltiples subconjuntos y entrenar el modelo varias veces, podemos obtener una estimación más confiable de su rendimiento y reducir la variabilidad asociada a una sola división de entrenamiento y prueba.

Finalmente, es esencial monitorear continuamente el desempeño del modelo una vez que se implementa en un entorno real. Los datos pueden cambiar con el tiempo, un fenómeno conocido como deriva de datos, lo que puede degradar el rendimiento del modelo. Establecer un proceso de reentrenamiento periódico y ajustar el modelo en función de nuevos datos garantiza que siga siendo preciso y robusto.

Conclusión

En este tema, hemos profundizado en métricas avanzadas y técnicas esenciales para la evaluación y ajuste de modelos de aprendizaje automático. La matriz de confusión y el análisis de errores nos permiten entender detalladamente el desempeño del modelo y las razones detrás de sus errores. Las curvas ROC y el AUC ofrecen una

perspectiva completa de la capacidad de discriminación del modelo, especialmente útil en situaciones donde los umbrales de decisión pueden variar.

Abordar el problema de los datos desbalanceados es determinante para asegurar que el modelo sea efectivo en todas las clases, especialmente en aquellas que son de mayor interés pero menos representadas. Las estrategias discutidas proporcionan herramientas prácticas para manejar este desafío.

El proceso de probar y ajustar el modelo es continuo y vital para garantizar su relevancia y eficacia. Al aplicar técnicas de optimización de hiperparámetros, ingeniería de características y validación adecuada, podemos mejorar significativamente el rendimiento del modelo. Además, al mantener un monitoreo constante y adaptativo, nos aseguramos de que el modelo siga siendo útil en un entorno cambiante.

Este enfoque integral en la evaluación y ajuste de modelos fortalece nuestra capacidad para desarrollar soluciones de aprendizaje automático precisas, robustas y adaptables a las necesidades reales, cumpliendo con los objetivos de precisión y robustez que son fundamentales en el campo de la inteligencia artificial.

4. Comunicación y documentación de resultados

En el ámbito de la inteligencia artificial y el machine learning, no es suficiente con desarrollar modelos precisos y robustos; es también clave comunicar eficazmente los hallazgos y resultados obtenidos. La capacidad para presentar información técnica de manera clara y accesible es esencial para que las partes interesadas comprendan el valor y las implicaciones de los modelos desarrollados. En este tema, exploraremos las mejores prácticas para la comunicación y documentación de resultados, incluyendo el arte del storytelling en ciencia de datos, la elaboración de informes técnicos, el desarrollo de manuales técnicos y la socialización y presentación de resultados.

4.1. El arte del storytelling en ciencia de datos

El storytelling, o narración de historias, es una técnica poderosa que permite transmitir información compleja de manera comprensible y atractiva. En ciencia de datos, el storytelling ayuda a contextualizar los análisis y resultados, conectándolos con las necesidades y objetivos del negocio o proyecto.

Una buena historia en ciencia de datos debe tener una estructura clara, que generalmente incluye:

- **Introducción**: presentación del problema o pregunta de investigación.
- **Desarrollo**: descripción del enfoque metodológico y los análisis realizados.
- **Clímax**: presentación de los hallazgos más significativos.
- **Desenlace**: interpretación de los resultados y sus implicaciones.
- **Conclusión**: recomendaciones y pasos a seguir.

Por ejemplo, al presentar un modelo de predicción de abandono de clientes, podríamos comenzar describiendo el impacto que el abandono tiene en la empresa,

cómo se abordó el problema mediante el análisis de datos y qué resultados clave se obtuvieron. Al final, se ofrecerían recomendaciones basadas en los hallazgos para retener a los clientes.

El uso de visualizaciones es fundamental en el storytelling, ya que facilita la comprensión de datos complejos. Gráficos, diagramas y tablas ayudan a ilustrar tendencias, patrones y relaciones que respaldan la narrativa.

Figura 2. El arte del storytelling

Fuente. OIT, 2024.

4.2. Elaboración de informes técnicos

Los informes técnicos son documentos que detallan de manera exhaustiva los procesos, análisis y resultados de un proyecto técnico. Su objetivo es proporcionar información precisa y detallada que permita a otros profesionales comprender y reproducir el trabajo realizado. Al elaborar un informe técnico, es importante considerar los siguientes elementos:

- **Resumen ejecutivo**: visión general concisa del proyecto, sus objetivos, métodos y conclusiones principales.

- **Introducción**: contextualización del problema y objetivos específicos del proyecto.
- **Metodología**: descripción detallada de los métodos y técnicas utilizados, incluyendo la selección y preparación de datos, algoritmos implementados y parámetros configurados.
- **Resultados**: presentación de los hallazgos, acompañados de tablas, gráficos y análisis estadísticos.
- **Discusión**: interpretación de los resultados, análisis de limitaciones y consideraciones sobre la validez y confiabilidad de los hallazgos.
- **Conclusiones y recomendaciones**: síntesis de las principales conclusiones y sugerencias para futuras acciones o investigaciones.
- **Anexos**: información adicional, como código fuente, detalles técnicos específicos o datos complementarios.

Es esencial que el informe sea claro y esté bien estructurado, facilitando su lectura y comprensión. El uso adecuado del lenguaje técnico y la precisión en la terminología son determinantes para mantener la rigurosidad del documento.

4.3. Desarrollo de manuales técnicos

Un manual técnico es un documento que proporciona instrucciones detalladas sobre cómo utilizar, mantener o desarrollar un sistema o producto. En el contexto de modelos de inteligencia artificial, un manual técnico puede incluir instrucciones para la implementación, configuración y mantenimiento del modelo. Al desarrollar un manual técnico, se deben considerar los siguientes aspectos:

- **Claridad y precisión**: las instrucciones deben ser claras, concisas y libres de ambigüedades.
- **Estructura lógica**: organizar el contenido de manera que siga un flujo lógico, facilitando al usuario encontrar la información que necesita.
- **Instrucciones paso a paso**: proporcionar guías detalladas para tareas específicas, apoyadas por capturas de pantalla o diagramas cuando sea pertinente.
- **Glosario de términos**: incluir definiciones de términos técnicos para ayudar a los usuarios menos familiarizados con el tema.
- **Soporte y mantenimiento**: instrucciones sobre cómo actualizar el modelo, resolver problemas comunes y contactar al equipo de soporte.

Por ejemplo, un manual técnico para un modelo de clasificación de imágenes podría incluir instrucciones sobre cómo instalar las dependencias necesarias, cómo preparar los datos de entrada, cómo ejecutar el modelo y cómo interpretar los resultados.

4.4. Socialización y presentación de resultados

La socialización de los resultados implica compartir los hallazgos con diferentes audiencias, que pueden incluir equipos técnicos, gerentes, clientes o el público en general. Es importante adaptar la presentación según el público objetivo, ajustando el nivel de detalle y el lenguaje utilizado. Al preparar una presentación de resultados, se deben considerar los siguientes puntos:

- **Conocer a la audiencia**: comprender el nivel de conocimiento técnico y los intereses de la audiencia para ajustar el contenido y la terminología.

- **Estructurar la presentación**: organizar la información de manera lógica, comenzando con una introducción clara y avanzando hacia los detalles más específicos.
- **Utilizar recursos visuales**: incorporar gráficos, diagramas y tablas que apoyen la narrativa y faciliten la comprensión.
- **Destacar los puntos clave**: enfatizar los hallazgos más importantes y sus implicaciones prácticas.
- **Fomentar la interacción**: incluir espacios para preguntas y discusiones, lo que permite aclarar dudas y obtener retroalimentación.

Algunas viñetas para tener en cuenta al presentar resultados:

- **Sé conciso**: evita sobrecargar a la audiencia con información excesiva.
- **Usa ejemplos prácticos**: ilustra los conceptos con casos reales o simulaciones.
- **Anticipa preguntas**: prepárate para abordar inquietudes comunes o puntos de confusión.
- Además, es recomendable practicar la presentación con anticipación y considerar el uso de herramientas interactivas o demostraciones en vivo para hacer la sesión más dinámica.

Conclusiones

La comunicación y documentación de resultados son componentes esenciales en cualquier proyecto de inteligencia artificial y machine learning. A través del storytelling, podemos conectar los hallazgos técnicos con narrativas que resuenen con las necesidades y objetivos de las partes interesadas. Los informes técnicos y manuales proporcionan documentación detallada y estructurada que respalda la

reproducibilidad y el mantenimiento del trabajo realizado. Finalmente, la socialización efectiva de los resultados asegura que el valor generado por los modelos desarrollados sea comprendido y aprovechado al máximo por todos los involucrados.

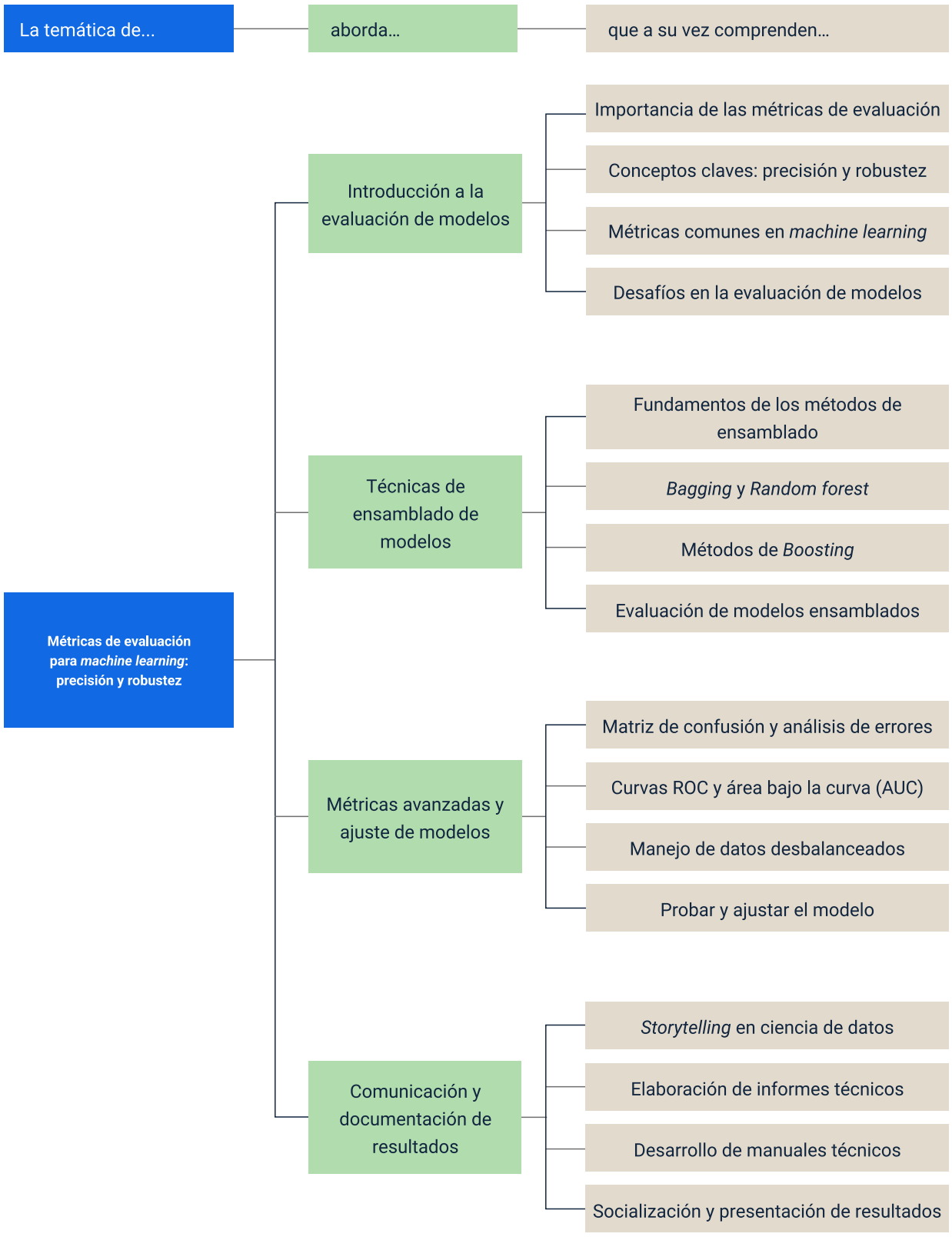
Síntesis

A continuación, se muestra un mapa conceptual con los elementos más importantes desarrollados en este componente.

El diagrama representa la estructura integral del componente sobre métricas de evaluación para machine learning, centrado en la precisión y robustez de los modelos. Partiendo del concepto central de las métricas de evaluación, se ramifica en cuatro áreas esenciales: introducción a la evaluación de modelos, técnicas de ensamblado de modelos, métricas avanzadas y ajuste de modelos, y comunicación y documentación de resultados. Cada una de estas áreas incorpora subtemas específicos que conforman los elementos fundamentales para comprender y aplicar eficazmente las métricas en el desarrollo y mejora de modelos de inteligencia artificial.

Esta organización ilustra el flujo lógico del proceso de evaluación y optimización de modelos de machine learning. Comienza con la comprensión de la importancia de las métricas y los conceptos clave de precisión y robustez, proporcionando una base sólida. Luego, profundiza en las técnicas de ensamblado como Bagging, Random forest y Boosting, que son esenciales para mejorar el rendimiento y reducir la variabilidad de los modelos. A continuación, aborda las métricas avanzadas y estrategias de ajuste, incluyendo el uso de la matriz de confusión, las curvas ROC y el manejo de datos desbalanceados, herramientas estratégicas para una evaluación detallada y ajuste fino de los modelos. Finalmente, se enfoca en la comunicación y documentación de resultados, destacando el papel del storytelling, la elaboración de informes técnicos y la socialización efectiva de los hallazgos.

El diagrama funciona como una hoja de ruta visual para comprender la estructura y el alcance del componente, permitiendo al lector visualizar rápidamente la progresión del aprendizaje y las conexiones entre los diferentes temas. Se sugiere utilizarlo como referencia para organizar el estudio y entender cómo se integran los diversos aspectos en la evaluación y mejora de modelos de machine learning, garantizando tanto su precisión como su robustez en aplicaciones prácticas.


La temática de...
aborda...
que a su vez comprenden...
Métricas de evaluación para *machine learning*: precisión y robustez
Introducción a la evaluación de modelos
Importancia de las métricas de evaluación
Conceptos claves: precisión y robustez
Métricas comunes en *machine learning*
Desafíos en la evaluación de modelos
Técnicas de ensamblado de modelos
Fundamentos de los métodos de ensamblado
Bagging y *Random forest*
Métodos de *Boosting*
Evaluación de modelos ensamblados
Métricas avanzadas y ajuste de modelos
Matriz de confusión y análisis de errores
Curvas ROC y área bajo la curva (AUC)
Manejo de datos desbalanceados
Probar y ajustar el modelo
Comunicación y documentación de resultados
Storytelling en ciencia de datos
Elaboración de informes técnicos
Desarrollo de manuales técnicos
Socialización y presentación de resultados

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
1. Introducción a la evaluación de modelos de machine learning	Ecosistema de Recursos Educativos Digitales SENA. (2023, septiembre 15). *Introducción al Machine learning*.	Video	https://www.youtube.com/watch?v=xwjQmGJ3q0I
1. Introducción a la evaluación de modelos de machine learning	Ecosistema de Recursos Educativos Digitales SENA. (2020, 13 septiembre). *¿Qué es Machine learning?*	Video	https://www.youtube.com/watch?v=J9w6KquPKbE
2. Técnicas de ensamblado de modelos de inteligencia artificial	Ecosistema de Recursos Educativos Digitales SENA. (2023, agosto 18). *Aprendizaje no supervisado K-means.*	Video	https://www.youtube.com/watch?v=yVIf4MtNg_c
2. Técnicas de ensamblado de modelos de inteligencia artificial	Ecosistema de Recursos Educativos Digitales SENA. (2023, marzo 27). *Algoritmos usados en aprendizaje no supervisado*.	Video	https://www.youtube.com/watch?v=iZ6soC3Nx9M
3. Métricas avanzadas de evaluación y ajuste de modelos	Ecosistema de Recursos Educativos Digitales SENA. (2023, octubre 10). *Machine learning con Python.*	Video	https://www.youtube.com/watch?v=noMy4-zjR9Q

Tema	Referencia	Tipo de material	Enlace del recurso
3. Métricas avanzadas de evaluación y ajuste de modelos	Ecosistema de Recursos Educativos Digitales SENA. (2023, marzo 17). *Optimización de datos previo a la creación de modelos de machine learning.*	Video	https://www.youtube.com/watch?v=tkeCwwPOVIU
4. Comunicación y documentación de resultados	Ecosistema de Recursos Educativos Digitales SENA. (2022, 26 diciembre). *Introducción a la visualización de datos.*	Video	https://www.youtube.com/watch?v=-7nn2bm07Dw

Glosario

Área bajo la curva (AUC): valor numérico que resume el desempeño de un modelo de clasificación binaria, calculado como el área bajo la curva ROC; un AUC más alto indica mejor discriminación.

Bagging (Bootstrap Aggregating): técnica de ensamblado que mejora la precisión y estabilidad de los modelos al entrenar múltiples modelos en subconjuntos aleatorios y promediar sus predicciones.

Boosting: técnica de ensamblado que combina secuencialmente modelos débiles, enfocándose en corregir los errores de los modelos anteriores para crear un modelo final más preciso.

Curva ROC: gráfico que representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos a diferentes umbrales de decisión.

Datos desbalanceados: conjuntos de datos donde las clases no están representadas de manera equitativa, lo que puede conducir a modelos que ignoran las clases minoritarias.

F1-Score: medida que combina la precisión y el recall en una sola métrica armonizada, proporcionando una evaluación equilibrada del rendimiento del modelo.

Matriz de confusión: tabla que permite visualizar el desempeño de un modelo de clasificación, mostrando las predicciones correctas e incorrectas y distinguiendo entre verdaderos y falsos.

Precisión (Accuracy): medida que indica el porcentaje de predicciones correctas realizadas por un modelo en comparación con el total de predicciones realizadas.

Precisión (Precision): métrica que indica la proporción de verdaderos positivos entre todos los casos que el modelo ha predicho como positivos.

Random forest: algoritmo de ensamblado que construye múltiples árboles de decisión utilizando muestras aleatorias y combina sus predicciones para mejorar el rendimiento.

Recall (Sensibilidad): métrica que mide la proporción de verdaderos positivos identificados correctamente por el modelo en relación con el total de casos reales positivos.

Robustez: capacidad de un modelo para mantener un rendimiento consistente ante variaciones en los datos de entrada, como ruido, valores atípicos o cambios en la distribución.

Sobreajuste (**Overfitting**): situación en la que un modelo aprende demasiado bien los detalles y el ruido del conjunto de entrenamiento, resultando en un rendimiento deficiente en nuevos datos.

Storytelling en ciencia de datos: técnica de comunicación que utiliza narrativas y visualizaciones para presentar resultados y hallazgos de manera comprensible y atractiva.

Subajuste (Underfitting): ocurre cuando un modelo es demasiado simple para capturar la estructura subyacente de los datos, resultando en un rendimiento deficiente en entrenamiento y prueba.

Referencias bibliográficas

AWS. (n.d.). ¿Qué es el aprendizaje automático? Recuperado de https://aws.amazon.com/es/what-is/machine-learning/

Barbara, J. (2023). Practical C++ Backend Programming: Crafting Databases, APIs, and Web Servers for High-Performance Backend. GitforGits.

Brandao, M., Iago, & da Costa, C. (2022). Fault diagnosis of rotary machines using machine learning. Eletrônica de Potência, 27(03), 1–8. http://dx.doi.org/10.18618/rep.2022.3.0013

García, S., Luengo, J., & Herrera, F. (2021). Data Preprocessing in Data Mining and Machine learning: New Frameworks, Algorithms and Applications. Springer.

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (3rd ed.). O'Reilly Media.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Knaflic, C. N. (2023). Storytelling with Data: A Data Visualization Guide for Business Professionals (2nd ed.). Wiley.

Molnar, C. (2022). Interpretable Machine learning: A Guide for Making Black Box Models Explainable (2nd ed.). Leanpub. https://christophm.github.io/interpretable-ml-book/

Moreno-Jiménez, J. M., & Escobar Urmeneta, M. T. (2020). El proceso analítico jerárquico (AHP). Fundamentos, metodología y aplicaciones. RECTA Monográficos, 1, 21-53.

Mudunuru, M. K., Ahmmed, B., Rau, E., Vesselinov, V. V., & Karra, S. (2023). Machine Learning for Geothermal Resource Exploration in the Tularosa Basin, New Mexico. Energies, 16(7), 3098. https://doi.org/10.3390/en16073098

Nelli, F. (2023). Python Data Analytics: With Pandas, NumPy, and Matplotlib (3rd ed.). Apress.

Polanía Arias, L. A. (2021). Evaluación de modelos de Machine learning para Sistemas de Detección de Intrusos en Redes IoT. Universidad de los Andes. Recuperado de https://repositorio.uniandes.edu.co/server/api/core/bitstreams/c6f1ea83-58d7-4eb6-a1ba-c3457306c054/content

Raschka, S., Patterson, J., & Nolet, C. (2022). Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133

Sarker, I. H. (2021, May 1). Machine learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. Springer. https://doi.org/10.1007/s42979-021-00592-x

Topuz, K., Bajaj, A., & Abdulrashid, I. (2023). Interpretable Machine learning. In Proceedings of the Annual Hawaii International Conference on System Sciences (Vol. 2023-January, pp. 1236–1237). IEEE Computer Society. https://doi.org/10.1201/9780367816377-16

Créditos

Elaborado por:

Organización Internacional del Trabajo