**ACTIVIDAD SESIÓN NIVEL DE AJUSTE DE UN MODELO Y VALIDACIÓN CRUZADA**

En esta actividad, evaluaremos la comprensión de los conceptos relacionados con el ajuste de modelos, el trade-off entre sesgo y varianza, y las técnicas de validación cruzada, incluyendo su implementación con la librería Scikit-Learn. La actividad consta de preguntas de reflexión y un caso práctico.

**INSTRUCCIONES:**

**PARTE 1: PREGUNTAS DE REFLEXIÓN (6 Puntos)**

Responde de manera argumentada y con ejemplos cuando sea necesario.

1. Niveles de ajuste de un modelo: Explica con tus propias palabras qué es el sobreajuste, el subajuste y el ajuste apropiado de un modelo. Proporciona un ejemplo práctico para cada caso.

Sobreajuste: El modelo se ajusta mucho a los datos de entrenamiento, toma en consideración los ruidos y anomalías por lo que predice aprendiendo los casos particulares de los datos de entrenamiento y no es bueno para generalizar. Por ejemplo, estoy buscando con mi modelo identificar si la imagen corresponde a un gato o no, le paso a mi modelo un Dataset de entrenamiento de gatos donde todos son de diferentes razas, pero de color blanco. Entreno a la maquina y luego le entrego una imagen de un gato negro, dando como resultado que no es un gato. Esto es porque el modelo aprendió las particularidades del Dataset.

Subajuste: El modelo es muy simple por lo que no puede captar la relación entre variable de entrada y salida, generando mucha tasa de error. Un ejemplo, siguiendo con el modelo de identificar un gato, podría ser entrenar el modelo con 5 muestras de una sola raza de gato y pasarle una imagen de un gato de otra raza, este modelo no podrá identificar que es un gato por la falta de suficientes muestras en el entrenamiento.

Para ajustar apropiadamente un modelo debemos encontrar un punto medio entre el sobreajuste y el subajuste, para esto es importante dividir el Dataset en datos de entrenamiento y datos de validación o test, donde estos últimos no serán conocidos por el modelo antes de utilizar los datos de entrenamiento. Luego debemos ir entrenando el modelo y evaluar como está funcionando el modelo a través de métricas, para esto es importante utilizar la validación cruzada, el error de entrenamiento, el error de validación, el error de prueba, etc.

1. Trade-off entre sesgo y varianza:

* ¿Qué es el sesgo (bias)y la varianza en el contexto de modelos de aprendizaje automático?

Sesgo: error debido a las suposiciones simplificadas que hace el modelo.

Varianza: sensibilidad del modelo a las pequeñas fluctuaciones en los datos de entrenamiento.

* ¿Cómo se relacionan con el sobreajuste y el subajuste?

Un modelo con alto sesgo es demasiado simple y no captura los patrones subyacentes en los datos, resultando en subajuste

Un modelo con alta varianza se ajusta demasiado a los datos de entrenamiento y tiene un rendimiento pobre en datos nuevos, resultando en sobreajuste

* Proporciona un ejemplo de cómo equilibrar el sesgo y la varianza en un modelo.

Random forest puede ser una buena opción para el problema de predicción de precios de viviendas. Este modelo utiliza múltiples árboles de decisión, lo que le permite capturar relaciones complejas (baja varianza) y evitar suposiciones simplificadas (bajo sesgo), proporcionando un rendimiento equilibrado en datos nuevos.

1. Validación cruzada:

* Explica qué es la validación cruzada y por qué es importante en la evaluación de modelos.

Técnica utilizada para evaluar el rendimiento de un modelo de aprendizaje automático. Consiste en dividir los datos disponibles en varios subconjuntos y entrenar el modelo en algunos mientras se valida en otros. Esto se hace demanera iterativa, asegurando que cada subconjunto se use al menos una vez como conjunto de validación. La validación cruzada es una herramienta crucial para medir cómo de bien puede generalizar un modelo a datos no vistos, previniendo el sobreajuste.

* Describe brevemente las siguientes técnicas de validación cruzada:
  + Método de retención (Hold-Out).

Consiste en dividir el conjunto de datos en dos partes: un conjunto de entrenamiento(usado para ajustar el modelo) y un conjunto de prueba(usado para evaluar su rendimiento). Proporción típica: 80%-20%. Es rápido y fácil de implementar, pero su estimación del rendimiento puede variar según cómo se dividan los datos.

* + Validación cruzada de k-iteraciones (k-Fold).

técnica robusta y ampliamente utilizada en aprendizaje automático para evaluar modelos predictivos. Consiste en dividir el conjunto de datos en k subconjuntos (o "folds") de tamaño similar. El modelo se entrena y evalúa k veces, utilizando en cada iteraciónk−1folds para entrenamiento y el fold restante para evaluación. Finalmente, se calcula el promedio de las métricas de rendimiento obtenidas en cada iteración.

* + Validación cruzada aleatoria (Random Subsampling).

Se divide repetidamente el conjunto de datos en conjuntos de entrenamiento y prueba de manera aleatoria, sinuna estructura fija. En cada iteración, se selecciona un subconjunto aleatorio de los datos para entrenamiento y el resto para prueba, repitiendo el proceso varias veces

* + Validación cruzada dejando uno afuera (Leave-One-Out).

Es un caso extremo de la validación cruzada k-Fold, dondekes igual al número total de muestras en el conjunto de datos. En cada iteración, se deja fuera una única muestra para evaluación y se entrena el modelo con el resto de los datos. Este proceso se repite hasta que cada muestra ha sido utilizada exactamente una vez como conjunto de prueba.

1. Implementación con Scikit-Learn:

* ¿Cómo se implementa la validación cruzada utilizando la librería Scikit-Learn?

Se importa la librería en conjunto con los tipos de validación cruzada que se quiera utilizar, luego se carga el Dataset, y luego se llama a cada tipo de validación junto con los parámetros a modificar

* Menciona al menos dos funciones o clases de Scikit-Learn que se utilizan para este propósito.

KFold, LeaveOneOut, StratifiedKFold, ShuffleSplit, cross\_val\_score

**PARTE 2: ANÁLISIS DE CASO (4 Puntos)**

Lee el siguiente caso y responde las preguntas que se plantean.

**Caso:**

Un equipo de científicos de datos está desarrollando un modelo de predicción para estimar el precio de viviendas en una ciudad. Han notado que el modelo tiene un rendimiento inconsistente: en algunos casos predice muy bien, pero en otros falla significativamente. Sospechan que el modelo puede estar sufriendo de sobreajuste o subajuste.

**Preguntas:**

1. ¿Cómo podrían diagnosticar si el modelo está sobreajustado o subajustado? Describe los pasos que seguirías para identificar el problema.

Evaluar rendimiento en datos de entrenamiento y datos de validación, error muy alto en ambos podría indicar subajuste y error en entrenamiento bajo y en prueba muy alto podría indicar sobreajuste

1. ¿Qué técnica de validación cruzada recomendarías para evaluar el rendimiento del modelo y por qué?

K-Fold Cross Validation, a pesar de que se necesitan recursos computacionales para hacerlo funcionar, permite generar un buen equilibrio entre sesgo y varianza en la estimación del rendimiento y proporciona una estimación más confiable del rendimiento al evaluar el modelo en múltiples particiones.

1. ¿Cómo podrían utilizar la validación cruzada para ajustar los hiperparámetros del modelo y mejorar su rendimiento?

Podría utilizarse diferente combinaciones de hiperparametros, ir evaluando cada combinación y luego comparar los rendimientos.

1. ¿Qué beneficios y desafíos podrían enfrentar al implementar la validación cruzada en este caso?

Como beneficio permite medir que tan bien generaliza el modelo a datos no vistos, previniendo el sobrajuste. Y como desafio, podríamos encontrarnos la dificultad de elegir correctamente la técnica de validación cruzada, ya que si no se elige correctamente podríamos no estar obteniendo datos verídicos y que hay un mayor uso de recursos computacionales, sobre todo en datasets grandes o modelos m8y complejos.