Definiciones de términos

# Datos de Train (Entrenamiento)

Son el conjunto de datos utilizados para entrenar un modelo de aprendizaje automático. Estos datos permiten al modelo aprender patrones y relaciones que puede utilizar para realizar predicciones.

# Datos de Validation (Validación)

Son el conjunto de datos utilizados para validar el modelo durante el proceso de entrenamiento. Este conjunto ayuda a ajustar los hiperparámetros del modelo y a evitar el sobreajuste, evaluando el rendimiento del modelo en datos no vistos.

# Datos de Test (Prueba)

Son el conjunto de datos utilizados para evaluar el rendimiento final del modelo después de haber sido entrenado y validado. Proporcionan una medida imparcial de la efectividad del modelo en datos completamente nuevos.

# GridSearchCV

Es una técnica de búsqueda de hiperparámetros en modelos de aprendizaje automático que permite encontrar la combinación óptima de parámetros mediante la evaluación exhaustiva de todas las combinaciones posibles en un conjunto de parámetros especificado. 'CV' significa 'Cross-Validation' (validación cruzada), lo que implica que se utilizan múltiples divisiones de los datos para obtener una estimación más robusta del rendimiento del modelo.

# One Hot Encoding

Es una técnica de preprocesamiento de datos que convierte variables categóricas en un formato numérico adecuado para modelos de aprendizaje automático. Cada categoría se transforma en una columna binaria (0 o 1), donde 1 indica la presencia de la categoría y 0 indica su ausencia.

# Matriz de Confusión

Es una herramienta que permite visualizar el rendimiento de un modelo de clasificación al comparar las predicciones del modelo con los valores reales. La matriz muestra verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, lo que facilita la evaluación del rendimiento del modelo.

# Precision (Precisión)

Es la proporción de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo. Se calcula como: Precisión = VP / (VP + FP), donde VP son los verdaderos positivos y FP son los falsos positivos. La precisión mide la exactitud de las predicciones positivas del modelo.

# Accuracy (Exactitud)

Es la proporción de todas las predicciones correctas (tanto positivas como negativas) respecto al total de predicciones realizadas. Se calcula como: Exactitud = (VP + VN) / Total, donde VN son los verdaderos negativos.

# Specificity (Especificidad)

Es la proporción de verdaderos negativos sobre el total de casos negativos. Se calcula como: Especificidad = VN / (VN + FP). La especificidad mide la capacidad del modelo para identificar correctamente los negativos.

# Recall (Sensibilidad o Tasa de Verdaderos Positivos)

Es la proporción de verdaderos positivos sobre el total de casos positivos reales. Se calcula como: Recall = VP / (VP + FN), donde FN son los falsos negativos. El recall mide la capacidad del modelo para identificar todos los casos positivos.

# F1 Score

Es la media armónica entre la precisión y el recall. Se utiliza como una métrica que combina ambas para obtener un equilibrio entre ellas. Se calcula como: F1 = 2 \* (Precisión \* Recall) / (Precisión + Recall).

# Curva ROC (Receiver Operating Characteristic)

Es una representación gráfica que ilustra el rendimiento de un modelo de clasificación en función de su umbral de decisión. Muestra la tasa de verdaderos positivos (recall) frente a la tasa de falsos positivos. Una curva ROC más cercana a la esquina superior izquierda indica un mejor rendimiento del modelo.

# R cuadrado (R²)

Es una medida estadística que indica el porcentaje de la variabilidad en la variable dependiente que es explicada por el modelo. Se utiliza principalmente en regresión. Un valor de R² de 1 indica que el modelo explica completamente la variabilidad de los datos, mientras que un valor de 0 indica que no explica nada.