Padrón: _____ Nombre y Apellido:_____

Parcial 23-05-2023

Ejercicio 1

Explique la diferencia entre variables **cuantitativas** y **cualitativas**. ¿Cómo se relacionan con los tipos de problemas a resolver (clasificación, regresión, agrupamiento)? Ejemplifique.

Variables Cualitativas: variables categóricas, sus valores representan categorías y no se puede establecer un ordenamiento entre ellas. Ejemplo: color de pelo, tipo de propiedad, modelo de auto, etc

Variables Cuantitativas: son variables que toman valores numéricos pueden ser discretos o continuos. Ejemplo: edad, altura, distancia recorrida, etc

En los problemas de clasificación la variable dependiente es cualitativa y las independientes pueden ser de cualquier tipo. Ejemplo: clasificar la especie de una flor a partir del largo de su pétalo, ancho del sépalo y color de las hojas.
En los problemas de regresión la variable dependiente es cuantitativa (de rango continuo) y las independientes pueden ser de cualquier tipo. Ejemplo: predecir el precio de una propiedad a partir de los metros cuadrados cubiertos, el barrio y la cantidad de ambientes.
En los problemas de agrupamiento no existe una variable dependiente que queramos predecir, sino que se trata de agrupar los datos en N conjuntos distinguibles, en base a características similares.

Ejercicio 2

Explique brevemente el concepto de **Outliers** e indique cuál es la diferencia entre valores atípicos **univariados y multivariados**. Mencione métodos/técnicas para detectar estos valores en cada caso. Dé un ejemplo concreto de algún método utilizado en el TP1.

Los outliers o valores atípicos son observaciones distantes del resto de los datos, pueden deberse a un error de medición, aleatoriedad,etc. Siempre son subjetivos al problema que estamos estudiando y deben ser cuidadosamente inspeccionados porque pueden estar alertando anomalías.
Los outliers univariados son valores atípicos que podemos encontrar en una simple variable por ejemplo: en un conjunto de datos correspondientes a niños en edad escolar, encontramos una observación donde la edad es de 45 años. Métodos para detectarlos: IQR, Z-score, Z-score modificado.
Los outliers multivariados son valores atípicos que se pueden encontrar en un espacio n-dimensional, es decir analizando más de una variable o atributo. Por ejemplo en un dataset de frutas una manzana de color naranja. Métodos para detectarlos: Mahalanobis, LOF, Isolation Forest, Clustering.

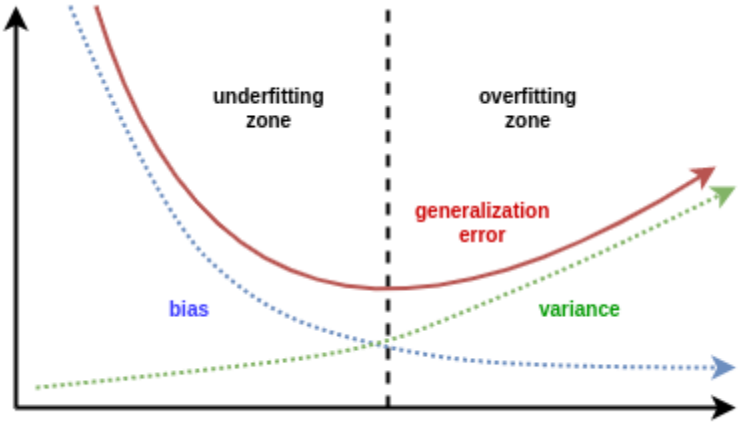
EJEMPLO CONCRETO DE APLICACIÓN AL TP1

Padrón: _____ Nombre y Apellido:_____

Ejercicio 3

Explique brevemente en qué consiste y para qué se utiliza el método **Early Stopping** en Redes Neuronales.

Es una técnica de regularización para Redes Neuronales que consiste en detener el entrenamiento de la red cuándo el error sobre el set de validación comienza a aumentar. Este método busca quedarse con los pesos en la instancia óptima.
Se utiliza para evitar el overfitting (sobreajuste) ya que ayudan a una mejor generalización, es decir, que el modelo funcione adecuadamente en datos que nunca vió.



Ejercicio 4

Seleccione una **técnica de preprocesamiento** aplicada en el **Checkpoint 1 del TP1** y describa detalladamente cómo se implementó dando ejemplos concretos del dataset utilizado.

EJEMPLO CONCRETO DE APLICACIÓN AL TP1
Detección y corrección de valores faltantes, imputación, encoding, generación de features,etc

Ejercicio 5

Explique por qué el set de datos suele particionarse en un conjunto de **entrenamiento** y uno de **test** respectivamente. ¿Qué consideraciones se deben tener en el preprocesamiento de datos respecto a estos conjuntos?

La partición del conjunto en train y test es un procedimiento que se utiliza para la validación de modelo que permite simular cómo se comportaría el mismo con datos nuevos/no vistos. El modelo se entrenará sobre el conjunto de train y luego se evaluará su performance sobre el conjunto de test. Las proporciones utilizadas comúnmente son 70/30 y 80/20. Es importante realizar el mismo preprocesamiento de los datos en ambos conjuntos evitando el Data Leakage.

Padrón: _____ Nombre y Apellido:_____

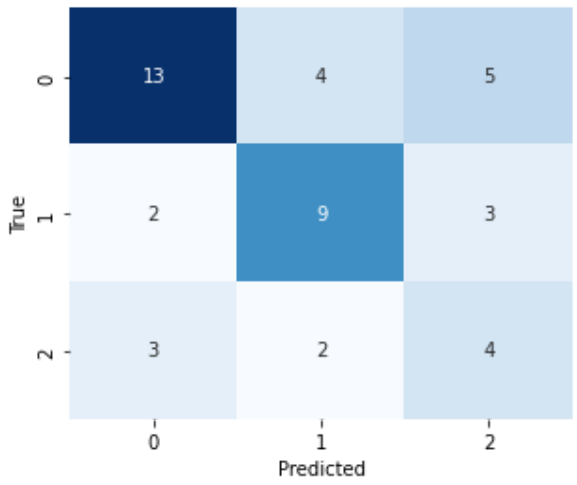
Ejercicio 6

Se tiene un conjunto de datos de 400 filas y 10 columnas y se quiere entrenar un árbol de decisión para resolver un problema de clasificación. Se genera un conjunto de entrenamiento (70%) y un conjunto de test (30%) y se quiere evaluar la performance del modelo en entrenamiento utilizando k-fold cross validation con k=4. Determinar si las siguientes afirmaciones son verdaderas o falsas (V / F):

a) Habrá 280 registros que estarán 4 veces en un conjunto de entrenamiento y 1 vez en un conjunto de validación. **FALSO**
b) Se partirá el conjunto de entrenamiento en 4 subconjuntos, luego cada subconjunto se partirá con proporción 70/30 para entrenar y validar el árbol .**FALSO**
c) El árbol de decisión se entrenará sólo una vez.**FALSO**
d) No se puede utilizar k-fold cross validation para evaluar el entrenamiento, es un método para optimizar hiperparámetros. **FALSO**

Ejercicio 7

Se entrenó un modelo de clasificación para detectar la cepa de un virus : cepa 0, cepa 1 y cepa 2. Luego se evaluó el modelo en los datos de test y se obtuvo la siguiente matriz de confusión:



a) ¿Qué cepa tiene el mayor recall? CEPA **1** RECALL **9/14 = 0.64**

b) ¿Qué cepa se detecta con menor precisión? CEPA **2** PRECISION **4/12 = 0.33**

c) ¿Cuál es el **porcentaje** total de aciertos del modelo? **57.77% = (26 / 45)*100**

Padrón: _____ Nombre y Apellido:_____

Ejercicio 8

Para un proyecto de ciencia de datos, se entrena un modelo predictivo y se evalúa su performance con la métrica F1-Score. Se obtienen los siguientes resultados:

-F1-Score en entrenamiento: 0.25
-F1-Score en test: 0.4

Indique si las siguientes afirmaciones son verdaderas o falsas (V / F):

a) El modelo generaliza muy bien para datos nuevos.**FALSO**
b) El modelo podría estar sobreajustando a los datos de entrenamiento (overfitting).**FALSO**
c) Hay un problema en los datos de test, se debería evaluar en otro conjunto.**FALSO**
d) El modelo podría estar subajustando a los datos de entrenamiento (underfitting) **VERDADERO**