Tema 1

1. ¿Este es un problema de clasificación binaria o un problema de clasificación multiclase? ¿Cuál es la diferencia entre ambos?

Este es un problema de clasificación binaria, ya que la variable objetivo es si la combinación de ingredientes y cantidades conduce o no a la cura de la princesa. La diferencia entre la clasificación binaria y multiclase es que la clasificación binaria tiene solo dos opciones posibles para la variable objetivo (por ejemplo, "sí" o "no", "verdadero" o "falso"), mientras que la clasificación multiclase tiene más de dos opciones posibles (por ejemplo, "rojo", "verde" o "azul").

2. En sus palabras, ¿por qué es importante y primordial efectuar el proceso del feature engineering para un problema de clasificación?

Es importante y primordial realizar el proceso de feature engineering en un problema de clasificación porque el conjunto de características o variables que se utilizan para predecir la variable objetivo tienen un impacto significativo en la capacidad del modelo para generalizar y hacer predicciones precisas. El feature engineering implica seleccionar, transformar y crear características que sean relevantes para la variable objetivo y que puedan mejorar la capacidad del modelo para hacer predicciones precisas.

3. En un dataset con tantas características (features) diferentes, ¿cómo puedo elegir de manera objetiva (numérica) qué features probar en el modelo y cuáles no? Explique su razonamiento.

Para elegir de manera objetiva qué características probar en el modelo, se pueden utilizar técnicas de selección de características. Estas técnicas pueden clasificarse en tres categorías: basadas en filtro, basadas en envoltorio y basadas en incrustación.

Las técnicas basadas en filtro evalúan cada característica de manera independiente y seleccionan aquellas que están más correlacionadas con la variable objetivo. Estas técnicas son rápidas y eficientes, pero no tienen en cuenta la interacción entre características.

Las técnicas basadas en envoltorio prueban diferentes combinaciones de características y evalúan su rendimiento utilizando un modelo específico. Estas técnicas pueden capturar la interacción entre características, pero son más costosas computacionalmente.

Las técnicas basadas en incrustación ajustan el modelo directamente en el conjunto de datos y seleccionan las características más relevantes durante el proceso de ajuste del modelo. Estas técnicas pueden ser más precisas y eficientes, pero pueden ser menos interpretables.

En este caso, dado que tenemos una cantidad limitada de ingredientes disponibles, es posible que sea útil utilizar técnicas basadas en filtro para seleccionar las características más relevantes. También es importante tener en cuenta la correlación entre características y

seleccionar solo aquellas que no están altamente correlacionadas entre sí, para evitar la multicolinealidad y mejorar la interpretación del modelo.

4. ¿Cómo se puede prevenir el overfitting o el underfitting en este caso?

Para prevenir el overfitting o el underfitting en este caso, es importante seleccionar características relevantes y tener un conjunto de datos de entrenamiento variado y suficientemente grande. Para prevenir el overfitting, se pueden utilizar técnicas como la regularización y la validación cruzada. Por otro lado, para prevenir el underfitting, es importante tener un modelo lo suficientemente complejo y ajustar los hiperparámetros para mejorar su rendimiento. También se puede aumentar el tamaño del conjunto de datos o utilizar técnicas de aumento de datos.

5. A lo largo de la última parte del curso, se expusieron varias metricas para medir el éxito del modelo de clasificación (RMSE, Accuracy, precision, recall, f1 score), y teniendo en mente que el objetivo es intentar encontrar la combinación de ingredientes que nos brinden la mejor probabilidad de encontrar una combinación de ingredientes que salven a la princesa, ¿cuál de estas cuatro métricas sería la más adecuada para poder medir el modelo? EXPLIQUE SU RAZONAMIENTO.

Se está trabajando en un problema de clasificación binaria para determinar si una princesa puede ser curada o no. Para evaluar el modelo, se deben utilizar las métricas adecuadas: precisión, recall y F1 score. La precisión mide la fracción de verdaderos positivos entre todas las instancias clasificadas como positivas, lo que indica la confianza del modelo.

El recall mide la fracción de verdaderos positivos entre todas las instancias que son realmente positivas, lo que indica la capacidad del modelo para identificar combinaciones efectivas de ingredientes. Finalmente, el F1 score combina la precisión y el recall en una sola métrica y es una buena medida general del rendimiento del modelo.

Tema 2

¿Cual fue el mejor modelo?

El modelo “gam” fue el que mas estuvo acertando, por lo que la princesa tiene un 90.38% de sobrevivir con las mezclas del alquimista.

Experimentos

- EXPERIMENTO 1:
  - El modelo tomado fue una regresión lineal multiple simple (LM), remover datos ni columnas. Por lo que esta regresión es la mas simple.
- EXPERIMENTO 2:
  - El modelo usado fue un GLM, este ya que es general dio un accurracy similar al lineal.
- EXPERIMENTO 3:
  - El ultimo fue un modelo GAM , lo que permiteajustar modelos lineales con términos suaves, lo que permite modelar relaciones no lineales entre las variables.