UNIVERSIDAD RAFAEL LANDÍVAR

FACULTAD DE INGENIERÍA

ANALISIS DE DATOS

SECCIÓN 1 VESPERTINA

EXAMEN FINAL

Julio Anthony Engels Ruiz Coto 1284719

GUATEMALA DE LA ASUNCIÓN, MAYO 3 DE 2023

CAMPUS CENTRAL

Tema I – 50 puntos
Utilizando R, cargue el dataset y observe los datos antes de responder. (sugerencia: verifique los datos con la función summary)

1. ¿Este es un problema de clasificación binaria o un problema de clasificación multiclase? Cuál es la diferencia entre ambos. 5 puntos

El problema que se nos presenta es de una clasificación binaria, porque la variable objetivo solo tiene dos posibles resultado si esta Curado, si dicha combinación de ingredientes y cantidad conduce a una cura o por el contrario No curado, (si este no conduce a una cura). La diferencia entre ambos es que la clasificación binaria implica categorizar estos datos en dos clases diferentes, por otro lado, la clasificación multiclase lo que implica es categorizar estos datos en más de dos clases.

2. ¿En sus palabras porque es importante y primordial efectuar el proceso del feature engineering, para un problema de clasificación? 5 puntos

Feature engineering es fundamental para un problema de clasificación en específico ya que este permite mejorar la calidad de las características en este problema los ingredientes en este caso, estos que se utilizan para alimentar el modelo de aprendizaje automático. Cuando se aplican las feature engineering, tiende a poder crear nuevas características a partir de las existentes en el dataset, conlleva en eliminar características irrelevantes, transformar estas características para una mejor representatividad, lo que conlleva en un modelo más preciso y eficiente, le da la oportunidad de mejorar así dicha capacidad de clasificación y en último caso, la probabilidad de encontrar la combinación correcta que se nos pide de ingredientes para curar a dicha princesa.

3. En un dataset con tantas características (features) diferentes, como puedo elegir de una manera objetiva (numérica) que features probar en el modelo y cuáles no. Explique su razonamiento. 5 puntos

Cuando se cuenta con un dataset con tantas características diferentes lo que se puede utilizar es un enfoque objetivo, como la selección de características basada en correlación, esta para poder elegir qué características se puede probar en el modelo y cuales no, la importancia de dichas características puede calcularse utilizando algoritmos como lo es Random Forest. También se puede calcular la correlación entre cada característica y nuestra variable objetivo, y con esto poder seleccionar aquellas características que tengan una correlación más alta, ya sea tanto positiva o negativa con la variable objetivo.

4. Como puede prevenir el overfitting o el underfitting en este caso? 15 puntos

Para prevenir el overfitting o el underfitting en este caso, se pueden utilizar diversas estrategias, como:

- *Seleccionar un número apropiado de características para reducir la dimensionalidad y evitar el overfitting.*

- *Se puede dividir el dataset en conjuntos de entrenamiento y prueba para validar el rendimiento del modelo en datos no vistos.*

5. A lo largo de la última parte del curso, se expusieron varias métricas 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. 20 puntos

La métrica más adecuada sería el recall, este mide la proporción de los casos positivos verdaderos para este problema las combinaciones que la princesa necesita identificados correctamente por el modelo. Cuando tiende a un alto recall lo que quiere decir es que el modelo es capaz de identificar la mayoría de las combinaciones exitosas para este caso los ingredientes ya que queremos minimizar dicha probabilidad de pasar por alto una combinación efectiva, con esto es preferible tener un mayor recall incluso si se experimenta una precisión baja, ya que asegurarnos de la cura de la princesa es el objetivo primordial.

Tema II - 50 puntos
Teniendo en cuenta el tiempo, desarrolle al menos tres experimentos de su modelo, aplicando lo visto en clase y tomando en cuenta lo consultado en el Tema I. Deje constancia de cada uno de los tres experimentos en su notebook de R, mejor si lo deja bien documentado.

En base a los experimentos ejecutados y basándose en la métrica elegida en la pregunta 5 del tema I, efectué la recomendación del modelo que mejor podría ayudarnos a predecir un éxito para poder curar a la princesa