Nombre: Edy German Perez Calcina

Materia: Inteligencia-Artificial DAT-245

0. Selección del clasificador:

Para abordar el problema planteado en el código, se utilizaron dos algoritmos: Random Forest como clasificador supervisado y KMeans como método no supervisado. Estas estrategias fueron seleccionadas porque se complementan entre sí, ajustándose tanto a las características del conjunto de datos como a los objetivos del análisis.

Clasificador supervisado: Random Forest

El Random Forest fue elegido como el modelo principal para el análisis supervisado debido a sus ventajas técnicas y prácticas:

1. Capacidad para manejar datos multiclase: La variable objetivo, "Discovery method", incluye varias clases que representan diferentes enfoques para descubrir exoplanetas. Random Forest es particularmente eficaz para manejar estos escenarios multiclase porque utiliza un conjunto de árboles de decisión. Al combinar las predicciones de múltiples árboles, este modelo ofrece un análisis robusto y confiable.

2. Resistencia a datos ruidosos: Este algoritmo minimiza el impacto de valores atípicos o ruido en los datos, ya que promedia las decisiones de múltiples árboles independientes. Esto reduce significativamente el riesgo de que un solo árbol domine el resultado, haciéndolo más fiable para conjuntos de datos complejos y variables.

3. Importancia de las características: Una de las ventajas clave de Random Forest es su capacidad para identificar las variables que tienen un mayor impacto en las predicciones. Esto no solo permite mejorar el modelo al centrarse en las características más relevantes, sino que también ayuda a entender mejor las relaciones entre las variables y los métodos de descubrimiento.

Para maximizar el rendimiento del modelo, se empleó GridSearchCV, una técnica que explora combinaciones óptimas de hiperparámetros como el número de estimadores (n_estimators), la profundidad de los árboles (max_depth) y el mínimo de muestras requeridas para dividir un nodo (min_samples_split). Gracias a este proceso de optimización, el modelo alcanzó un rendimiento excepcional en el conjunto de prueba.

Algoritmo no supervisado: KMeans

En el análisis no supervisado, se utilizó KMeans para identificar patrones subyacentes en los datos, sin depender de una variable objetivo. Este algoritmo clasificó los exoplanetas en tres clusters, considerando características escaladas como masa, radio y temperatura.

La elección de KMeans se basa en varias razones:

1. Simplicidad y eficiencia computacional: KMeans es un algoritmo rápido y fácil de implementar, lo que lo convierte en una opción ideal para analizar datos de gran escala y alta dimensionalidad.

2. Revelación de estructuras ocultas: Aunque asume que los datos forman clusters esferoidales, su aplicación sigue siendo válida en este contexto, ya que ayuda a descubrir

conexiones y patrones ocultos en los datos. Esto proporciona información valiosa para complementar el análisis supervisado.

3. Complemento al enfoque supervisado: Los resultados de KMeans no solo ofrecen una perspectiva alternativa sobre las estructuras del conjunto de datos, sino que también pueden servir como base para futuras iteraciones del modelo, ayudando a identificar áreas que requieren mayor exploración o refinamiento.

Justificación del Clasificador

La elección de Random Forest como el principal clasificador supervisado está fundamentada en las siguientes ventajas:

1. Interpretabilidad: Random Forest permite inspeccionar de manera directa la importancia de las características, facilitando la comprensión de los resultados tanto para audiencias técnicas como no técnicas. Esto lo hace más accesible en comparación con modelos más complejos como las redes neuronales.

2. Flexibilidad: Este algoritmo puede manejar datos heterogéneos, incluyendo variables categóricas y continuas. Por ejemplo, en este caso, procesó con éxito variables como "Mass (Tierra)" y "Temp. (K)", adaptándose a las necesidades específicas del conjunto de datos.

3. Mitigación de sobreajuste: Random Forest reduce el riesgo de sobreajustar los datos de entrenamiento al combinar múltiples árboles de decisión. Esto asegura que el modelo generalice bien a nuevos datos, mejorando su rendimiento en escenarios del mundo real.

Referencias técnicas

- Fuente ISBN: 978-1-59327-868-4. Este libro describe en detalle la teoría y práctica de Random Forest, con ejemplos aplicados a problemas similares.
- Fuente DOI: 10.1007/978-1-4471-7448-4_6. Este capítulo destaca las ventajas y limitaciones de Random Forest en el aprendizaje supervisado.

Preprocesamiento y su relación con el clasificador

El éxito de los clasificadores en este análisis depende en gran medida de las técnicas de preprocesamiento aplicadas, que garantizan que los datos estén listos para su uso:

1. Imputación de valores faltantes: La imputación con la mediana asegura que no se pierda información valiosa debido a datos incompletos. Esto permite que Random Forest y KMeans trabajen con conjuntos de datos completos y confiables.

2. Escalado de características: La normalización mediante MinMaxScaler asegura que todas las características tengan la misma escala, evitando que una variable domine el modelo. Esto es especialmente crucial para KMeans, pero también beneficia a Random Forest al equilibrar la contribución de cada característica.

3. Balanceo de datos: Dado que la variable objetivo puede estar desbalanceada, el uso de SMOTE crea un conjunto de datos balanceado, mejorando significativamente el rendimiento del clasificador supervisado.

Estas técnicas no solo mejoran la eficiencia de los clasificadores, sino que también aseguran que las conclusiones obtenidas sean más confiables y precisas.

La combinación de Random Forest y KMeans representa un enfoque robusto y complementario para analizar datos complejos como los exoplanetas. Random Forest sobresale en tareas supervisadas, ofreciendo predicciones precisas y explicaciones claras, mientras que KMeans aporta valor al identificar patrones implícitos en los datos. Además, un proceso de preprocesamiento bien diseñado asegura que ambos algoritmos trabajen al máximo de su potencial. Este análisis no solo mejora nuestra comprensión del conjunto de datos, sino que también establece una base sólida para futuras investigaciones en el campo.