UNIVERSIDAD EAN

ESPECIALIZACIÓN EN MACHINE LEARNING

PRUEBA FINAL DEL PROYECTO: PREDICCIÓN DE OBESIDAD A PARTIR DE HÁBITOS DE VIDA

TUTOR:

CARLOS ISAAC ZAINEA MAYA

INTEGRANTE:

YEISON ANDRES BARON LOPEZ

Tabla de Contenido

RESUMEN EJECUTIVO

Este proyecto tiene como objetivo mejorar un modelo de predicción de obesidad utilizando técnicas de análisis de clústeres y optimización de hiperparámetros. Inicialmente, se implementó un árbol de decisión para clasificar los datos, pero para mejorar su rendimiento, se introdujeron clústeres basados en características clave. Posteriormente, se optimizaron los hiperparámetros mediante técnicas de *grid search*.

Los resultados muestran que la adición de clústeres mejora significativamente la precisión del modelo, alcanzando una precisión del 100%. Se concluye que los clústeres pueden ser una herramienta útil para mejorar la clasificación en modelos de predicción.

INTRODUCCIÓN

En este proyecto, se aborda la tarea de mejorar un modelo de predicción sobre la obesidad, utilizando un conjunto de datos con características como la presión arterial, edad, entre otros.

El objetivo es predecir las categorías de obesidad de los individuos basándose en estas características. Para ello, se exploran técnicas de análisis no supervisado, como los clústeres, para identificar patrones ocultos en los datos. Además, se optimizan los hiperparámetros del modelo de árbol de decisión para maximizar su rendimiento. Este trabajo tiene como propósito demostrar cómo la combinación de métodos no supervisados y supervisados puede mejorar la precisión de un modelo de predicción.

METODOLOGÍA

Exploración No Supervisada (Clústeres)

1. **Preprocesamiento de Datos:** Se realizaron las siguientes tareas de limpieza de datos:
    - Eliminación de valores nulos o inconsistentes.
    - Normalización de los datos usando un escalador estándar para asegurar que las características tuvieran la misma escala.
2. **Análisis de Componentes Principales (PCA):** Se utilizó PCA para reducir la dimensionalidad de los datos, seleccionando las primeras 2 componentes principales para facilitar el análisis de clústeres.
3. **Clustering (K-means):** Se aplicó el algoritmo K-means con un número de 4 clústeres, basándose en la información proporcionada por PCA. Los clústeres obtenidos se añadieron como una nueva característica en los datos para alimentar el modelo supervisado.

Entrenamiento Supervisado (Árbol de Decisión)

1. **Entrenamiento Inicial:** Se entrenó un modelo de árbol de decisión sin utilizar los clústeres obtenidos. El modelo inicial logró una precisión de 98% en los datos de prueba.
2. **Entrenamiento con Clústeres:** Posteriormente, se añadieron los clústeres como una nueva característica al modelo. El rendimiento mejoró, alcanzando una precisión de 100%.
3. **Optimización de Hiperparámetros:** Se utilizó *Grid Search* para optimizar los hiperparámetros del modelo. Los mejores parámetros encontrados fueron:
    - criterion: gini
    - max_depth: None
    - min_samples_leaf: 1
    - min_samples_split: 2

Tras la optimización, el modelo mostró un desempeño similar al modelo con clústeres, con una precisión del 100%.

Validación y Evaluación

Se realizó validación cruzada para evaluar la robustez del modelo optimizado. La precisión media en 5 pliegues fue de 97.16%. Se compararon los resultados del modelo original, el modelo con clústeres añadidos y el modelo optimizado, con todos los modelos alcanzando una precisión del 100% en los datos de prueba.

RESULTADOS

Análisis de Modelos

- **Modelo Original:** Un árbol de decisión entrenado sin clústeres logró una precisión de 98%.
- **Modelo con Clústeres Añadidos:** Al añadir los clústeres, la precisión aumentó al 100%, indicando que los clústeres ayudan a mejorar la capacidad del modelo para hacer predicciones precisas.
- **Modelo Optimizado:** Después de la optimización de hiperparámetros, el modelo también alcanzó una precisión del 100%, lo que sugiere que los parámetros adecuados fueron seleccionados correctamente para maximizar el rendimiento.

Visualizaciones de Clústeres

- Los clústeres obtenidos fueron visualizados utilizando gráficos de dispersión. Se observó que los clústeres estaban bien definidos, con separación clara entre las diferentes clases de obesidad. Esto refuerza la idea de que los clústeres pueden capturar patrones subyacentes en los datos, mejorando así la predicción.

Comparación de Modelos

- La inclusión de los clústeres resultó en un modelo significativamente mejorado.
- La optimización de hiperparámetros permitió afinar aún más el modelo, pero los resultados fueron similares a los obtenidos con los clústeres.

Conclusiones y Recomendaciones

1. **Conclusiones:**
    - El uso de clústeres ha demostrado ser eficaz para mejorar el modelo de árbol de decisión, proporcionando una predicción más precisa.
    - La optimización de hiperparámetros también contribuyó a la mejora del rendimiento, aunque no hubo una mejora significativa al compararlo con el modelo con clústeres.

- Los resultados sugieren que los clústeres pueden ser una herramienta valiosa para mejorar modelos supervisados, especialmente cuando los datos son complejos o no lineales.

2. **Recomendaciones:**
    - Para futuros trabajos, se recomienda explorar otras técnicas de reducción de dimensionalidad como t-SNE o UMAP para ver si los clústeres mejoran aún más.
    - También es importante realizar validaciones adicionales con conjuntos de datos más grandes o con diferentes características para confirmar la robustez del modelo.
    - Además, se recomienda implementar un análisis más detallado de las características que más influyen en las predicciones para mejorar la interpretabilidad del modelo.

Referencias

- "Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido.
- Scikit-learn documentation: https://scikit-learn.org/stable/
- SHAP documentation: https://shap.readthedocs.io/en/latest/