

Descripción del Proceso Realizado

El proceso realizado para la carga, análisis, preprocesamiento, construcción de modelos y evaluación de un conjunto de datos relacionado con las fuentes de energía renovables. El objetivo es comparar el desempeño de dos modelos de clasificación: Support Vector Machine (SVM) y Random Forest (RF).

Se utilizaron las bibliotecas de pandas ,numpy ,seaborn y matplotlib para visualización de datos y sklearn para la construcción de modelos, preprocesamiento, selección de hiperparámetros y evaluación.

Se cargó el conjunto de datos desde un archivo CSV utilizando pandas y se usaron múltiples graficas para analizar los datos. Desde la visualización de las primeras filas del conjunto de datos, el resumen estadístico de las variables, distribución de variable objetiva y numérica y la correlación entre las variables. En el Preprocesamiento no se encontró ninguna variable nula y no se tuvieron que eliminar ni columnas, instancias o imputar datos.

Se dividió el conjunto de datos y los datos se dividieron en conjuntos de entrenamiento y prueba, con una reducción inicial del tamaño del conjunto de datos para pruebas. Además de la construcción de los modelos SVM Y Random Forest

Se evaluaron ambos modelos utilizando el conjunto de prueba y se imprimieron los reportes de clasificación, Curvas de Aprendizaje Y Matrices de Confusión.

La matriz de confusión fue el método de comparación en este caso la cual permitió comparar el desempeño de dos modelos de clasificación (SVM y Random Forest) en un conjunto de datos relacionado con fuentes de energía renovables.

La matriz de confusión del SVM solo predice una única clase (clase 4). Esto indica que con el modelo SVM, no es capaz de distinguir entre las diferentes clases de manera adecuadas .

La matriz de confusión del Random Forest muestra un comportamiento más disperso que el modelo SVM haciendo predicciones en múltiples clases.

En resumen, la matriz de confusión del SVM indica un modelo que no está funcionando correctamente para este conjunto de datos, mientras que la matriz de confusión del Random Forest muestra un rendimiento mejor, demostrando que soporta mejor carga de datos grandes.