Memoria del Trabajo: Modelo Predictivo de Calidad del Agua

Introducción

En este proyecto, se desarrolló un modelo predictivo para estimar la clasificación de la calidad del agua en función de diversas características del agua y el entorno. El objetivo principal es ayudar en la toma de decisiones relacionadas con el uso del agua en la agricultura y otros sectores. Se utilizó un conjunto de datos que contiene información de calidad del agua de varios años y se aplicaron técnicas de preprocesamiento de datos y aprendizaje automático para construir y evaluar modelos de clasificación.

Conjunto de Datos

El conjunto de datos utilizado para este proyecto consiste en mediciones de calidad del agua recopiladas en diferentes distritos y años. Las características incluyen valores químicos, físicos y geográficos, como la concentración de iones, el pH, la temperatura, la latitud y la longitud, entre otros. La variable objetivo es la "Clasificación de la Calidad del Agua", que describe la idoneidad del agua para el riego y su impacto en el suelo y los cultivos.

Preprocesamiento de Datos

Se llevaron a cabo las siguientes etapas de preprocesamiento de datos:

- Eliminación de columnas innecesarias, como "season" y "Unnamed: 8".
- Renombrado de columnas para cumplir con un estándar de nomenclatura.
- Relleno de valores nulos en la columna "CO3" utilizando las medias por distrito en el año 2019.
- Concatenación de los datos de diferentes años en un solo conjunto de datos.
- Eliminación de categorías de clasificación raras que causaban ruido en el modelo.

- Modelos de Aprendizaje Automático

Se construyeron varios modelos de aprendizaje automático para predecir la clasificación de la calidad del agua. Las etapas clave incluyeron:

- División del conjunto de datos en datos de entrenamiento y prueba.
- Normalización de los datos de entrada utilizando StandardScaler.
- Reducción de la dimensionalidad con PCA (Análisis de Componentes Principales).
- Entrenamiento de los siguientes modelos:
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier (SVC)
- K-Nearest Neighbors (KNN)
- Evaluación de los modelos mediante la precisión (accuracy) en el conjunto de prueba.

Resultados:

Los resultados de los modelos mostraron diferentes niveles de precisión en la clasificación de la calidad del agua. Se observó que el modelo de Random Forest Classifier obtuvo la precisión más alta, seguido por Support Vector Classifier, Decision Tree Classifier y K-Nearest Neighbors. La elección del modelo final depende de las necesidades específicas y del equilibrio entre precisión y otros factores.

Importancia de las Características:

Se analizó la importancia de las características para el modelo Random Forest Classifier. Las características más importantes para el modelo incluyeron, entre otras, la concentración de iones, el pH y la temperatura. Este análisis puede proporcionar información valiosa sobre qué variables son más relevantes para predecir la calidad del agua.

Uso de Modelos Entrenados

Se proporcionó la capacidad de cargar los modelos previamente entrenados para realizar predicciones en nuevos datos. Los modelos de Decision Tree, Random Forest, SVC y KNN se guardaron en archivos .pkl y se pueden utilizar según sea necesario.

Conclusión

Este proyecto demuestra la aplicación exitosa de técnicas de aprendizaje automático para predecir la calidad del agua a partir de datos químicos, físicos y geográficos. Los modelos construidos pueden ser valiosos para la toma de decisiones relacionadas con el uso del agua en la agricultura y otros sectores. Se pueden realizar mejoras adicionales en el proyecto, como la optimización de hiperparámetros y la inclusión de más datos históricos para mejorar la precisión y la robustez de los modelos.