أكاديمية سدايا
SDAIA Academy

1443/3/15

(Smart Agriculture)

Final Project

Eng.Nouf Rashed Alswuida

T5 BOOTCAMP

Abstract

The goal of this project was to use classification models to predict the operating condition of water points (Irrigation sensors) in smart farm in order to help improve solution will help farmers prepare their irrigation schedules more efficiently.
I worked with data provided by real collected data by soil moisture sensor during multiple days and , leveraging geographic and categorical feature engineering along with KNN Regression model to achieve promising results for this multiclass problem.

Design

Classifying statuses accurately via machine learning models would enable the Farmers to take action to improve operations and increase performance of these lands, allocate resources more quickly to needed areas, ensuring that the fields receive enough water without an increase or decrease, and this is different for each type of planted product and the appropriate soil moisture.

Data

The dataset contains (28049) soil humidity of fields & sensor status with 15 features. A few feature highlights include Air temperature (C), Air humidity (%), Pressure (KPa), Wind speed (Km/h), Wind gust (Km/h), Wind direction (Deg).

Algorithms

Feature Engineering

1. Mapping latitude and longitude to 3-dimensional coordinates so nearby continuous values would also be close in reality.
2. Visualizing the missing values and fill it.
3. Combining particular dummies and ranges of numeric features to highlight strong signals and illogical values for irrigation sensor status identified during EDA.
4. Selecting subsets of the total unique values for categorical features, according to the number of samples they were associated with and their contribution to certain statuses

Models

Logistic regression, k-nearest neighbors, and random forest, Decision Tree classifiers were used. Before settling on KNN as the model with strongest ROC AUC curve.

Model Evaluation and Selection

The entire training dataset of 28049 records was split into 75/25 train vs. Predictions on the 25% K Nearest Neighbors were limited to the very end, so this split was only used and scores seen just once. and accuracy of model is calculated also with The ROC curve. The official metric for Driven Data was classification rate (accuracy)

logistic regression: 2 features

Accuracy 0.8796185935637664

ROC AUC score = 0.8242375096358965

K Nearest Neighbors: 2 features

Accuracy 0.882797

ROC AUC score of KNN = 0.8692753165165139

Random Forest: 2 features

Accuracy 0.861740

Decision Tree: 2 features

Accuracy 0.860946

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting