



Detection of nitrogen concentration in oilseed rape

23-12-2021

Noam Jow

Introduction

The project involves the hyperspectral imaging for rapid detection of nitrogen concentration in oilseed rape leaf. Notably, the oilseed rape is one of the most important oil crops globally. Better fertilizer management is required to improve the quality and the productivity of the oilseed rape. So, the detection of nitrogen concentration is really important to properly manage the nitrogen fertilizer applications which can greatly benefit in improving the quality and the productivity of the oilseed rape.

Objective

The aim of the project is the detection of the nitrogen concentration in the oilseed rape. In other words, the detection of this chemical property is mainly required to solve the fertilization problem of oilseed rape.

Dataset

Before doing any machine learning project, it is necessary to collect enough and consistent sample data for comprehensive and exploratory data analysis. For successful detection of nitrogen concentration in oilseed rape, the collected csv-filed dataset constitutes 512 number of input features containing 192 number of observations. The input feature columns range from 'Band1' to Band512 thus, containing 512 input feature columns with a predictive output column 'N Values'. There is also a column named 'Dataset' containing binary values. So, after some brief description of the dataset, it's the right time to perform dataset splitting which is done in the next section.

Data splitting

The data splitting is performed on the basis of the column 'Dataset'. The observations or rows where the 'Dataset' column contains the value '1' are used for the training set while the remaining rows containing the value '0' are used for the test set. Now, the most important step is to perform the feature selection considering the number of feature columns in the dataset. The feature selection is explained in the next section.

Feature selection

There are several techniques for feature selection provided by the scikit-learn library. For the given case, the following feature selection technique is used.

SelectKBest

The technique involves selecting the k number of important feature columns having great impact on the output column to predict. The k value is passed to the 'SelectKBest' method to pick the k number of feature columns. The selection is all done on the basis of the score of each feature column. Greater the score, the more important or dominant the feature is. For the given case, 350/512 feature columns are selected depending upon the complexity of the dataset while the other features having low impact on the output column are eliminated from the dataset. After the feature selection step, the most significant step is now to be performed which is explained in the succeeding part.

Model training

After complete data preprocessing, the model selection and training is performed using the classical machine learning algorithms provided by the scikit-learn library. Considering the complexity and the format of the dataset, the following three machine learning regressors are chosen:

- Decision tree regressor
- Support vector regressor (SVR)
- Random forest regressor

Each of the above regressors is used as a model for training phenomena. The training is performed using the x_train and y_train dataset constituting 280 number of observations. After training is done, the trained regressor or model is evaluated in the next section.

Model evaluation

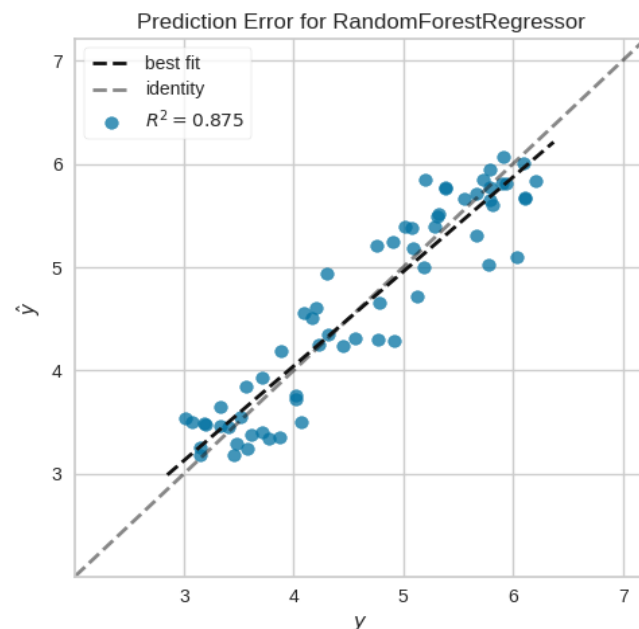
The regression models are usually evaluated by calculating the prediction errors and scores. All three trained regressors or models are evaluated using the prediction errors and scores provided by the scikit-learn library. All three models and their relative evaluation statistics can be shown in the following table.

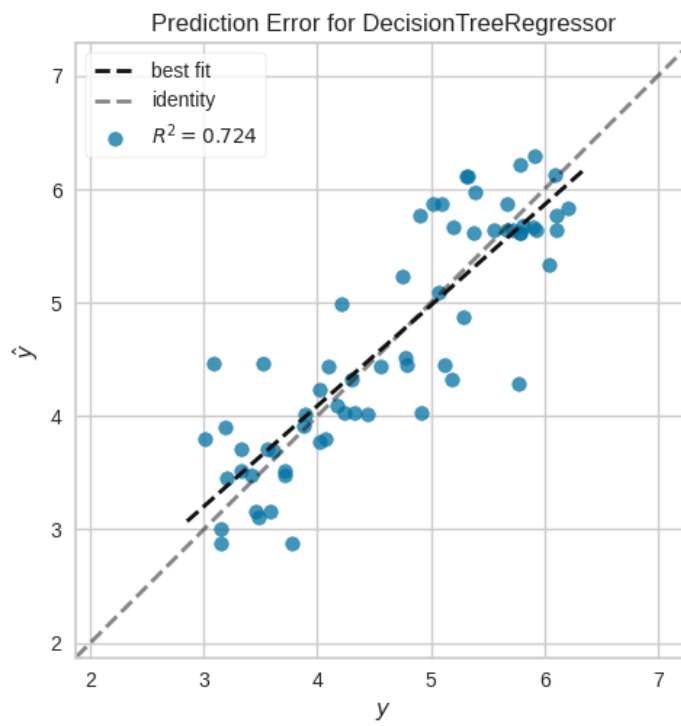
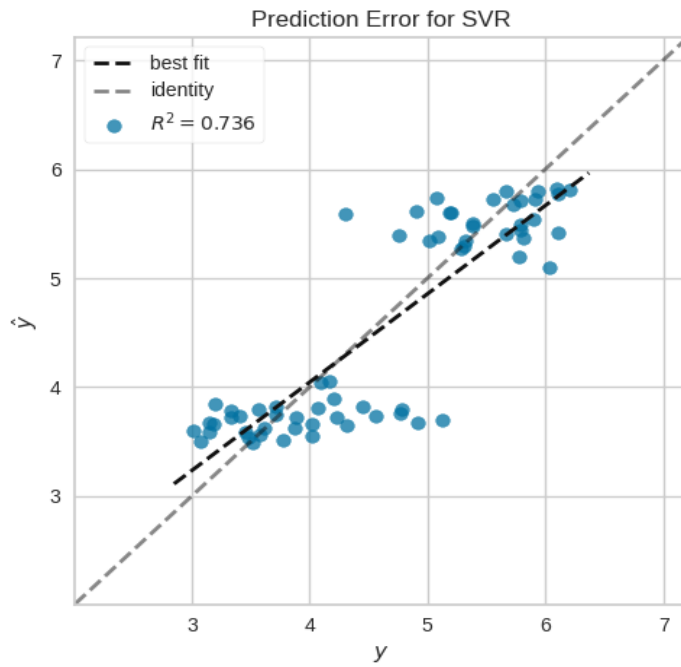
Model	Accuracy	Variance score	R2 score	Mean absolute error	Mean squared error
Random forest	87-89	0.8752	0.8751	0.290	0.122
Decision Tree	72-74	0.7238	0.7235	0.405	0.269
Support vector	73-75	0.7414	0.7356	0.391	0.257


Considering the above table, the random forest regressor or model possesses the best performance for the given case statistically. Due to the high number of features, the decision tree could not perform because the decision tree algorithm is mostly used for the short datasets containing a low number of features. The support vector regressor also does not have good predictive analysis for the given dataset. The 3 regression models are represented pictorially in the next section showing the plots for model performance representation.

Prediction plots

The plots giving the insight for the prediction errors can be shown below.







The above plots involve the identity line and the best fit. The identity line indicates the ground truth line and the best fit line indicates the prediction line evaluated by the trained models.

Tools and technologies

The following tools and technologies are used at various steps of the project. The tool or technology and for what it is used is described briefly below.

- Programming language: python
- Scikit-learn: feature selection and training
- Pandas profiling: exploratory data analysis
- Matplotlib: plotting the prediction errors
- Pickle: saving and loading the model files.

Conclusion

The random forest regressor is the best performer in quantifying the nitrogen concentration in the oilseed rape.