1. Data scrapying

First, use data scrapying packages to collect data on the official website of agricultural product statistics, select a specific product in a specific area - wild mushrooms, and mainly crawl the wild mushroom production in various time periods (weeks and months).

Then, crawl the weather conditions in the corresponding time period in the area.

The indicators are shown as follows:

{"Average low temperature/℃": 6, "Average high temperature/℃": 21, "Extreme low temperature/℃": 0, "Extreme high temperature/℃": 29, "Total number of rainy/snowy days/d": 19, "Total rainfall/mm": 18.4, "Average air quality": 34.0, "Average daily maximum wind speed/km/h": 5.9, "Average visibility/km": 26.2, "Air humidity/%": 42.2, "CO2 concentration in the lower atmosphere/ppm": 368.0, "Oxygen concentration in the lower atmosphere/%": 53.8, "Light intensity/W/m²": 272.3, "Wild morel dry yield/Kg": 106.8}

Save the data in a json file.

The others are predicted independent variables, and the wild morel dry yield is the dependent variable.

2. Data cleaning and preprocessing

Remove data including null values ​​and outliers, and then preprocess each parameter, retain the integer part of the temperature, and retain one decimal for other parameters, then connect and combine the data sets, combine them through the same date, and finally build the initial data set.

3. Task objectives

It is hoped that the output of agricultural products can be predicted based on future weather conditions, so as to help enterprises better purchase and control inventory and ensure the stability of product supply. Similarly, the economic supply and demand model can be used to effectively predict product prices.

4. Model selection

Random forest, SVM, GBM, MLP, LSTM

5. Evaluation indicator selection

Mean squared error (MSE)

The mean squared error measures the average of the squares of the difference between the predicted value and the true value. The smaller it is, the smaller the prediction error of the model.

R² score (R-squared, R²)

R² is an indicator of the goodness of fit of the regression model, ranging from 0 to 1. R² indicates the degree of explanation of the independent variable to the dependent variable. The closer the value is to 1, the better the prediction effect of the model.

Output visualization icons:

Regression curve: true value vs predicted value

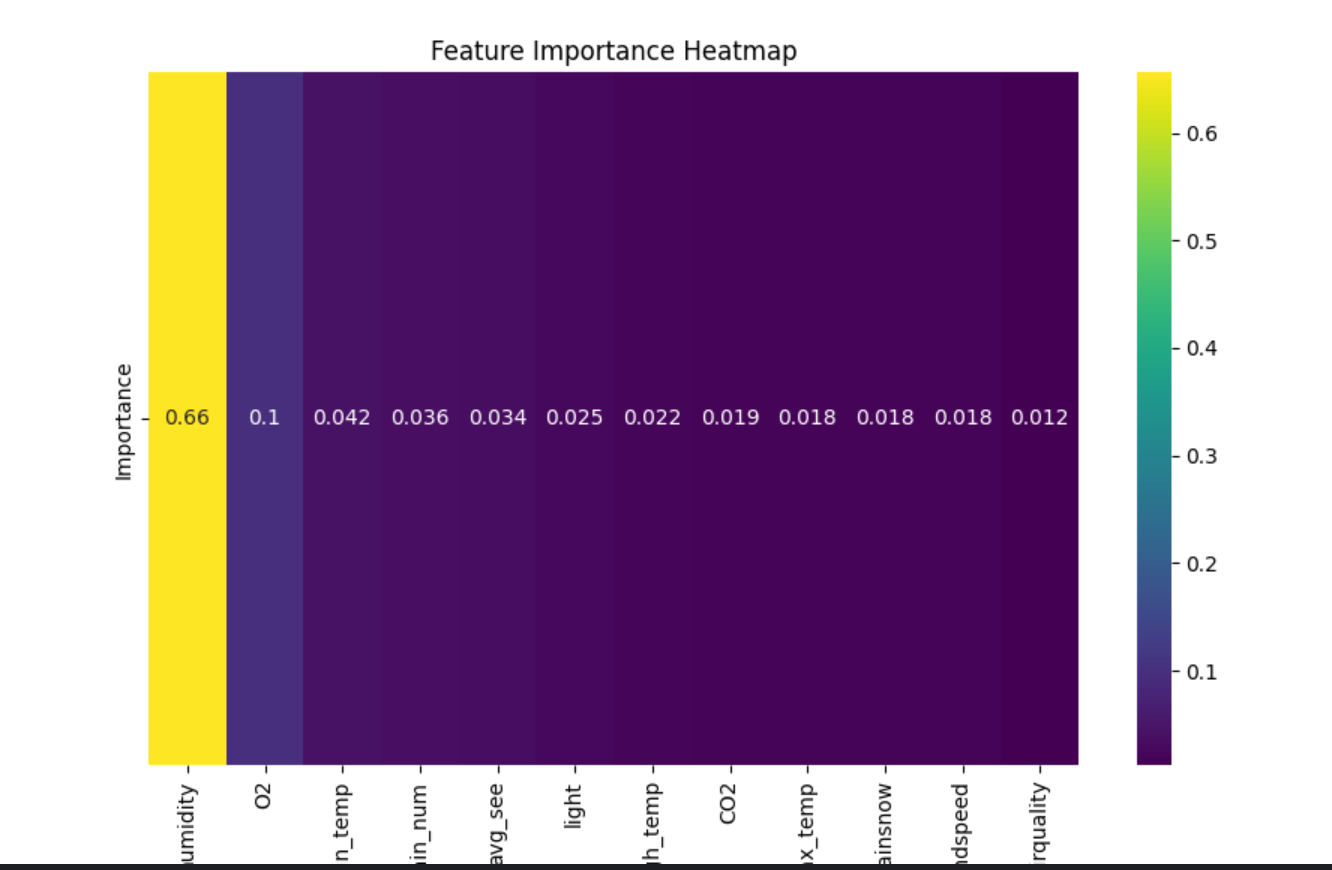
This graph shows the comparison between the model's prediction results and the true value. Through this curve, you can intuitively see the model's prediction effect.

Learning curve: training score vs cross-validation score

The learning curve shows the performance of the model under different numbers of training samples, and is compared with the training score and cross-validation score (i.e. the score on the test set).

The training score reflects the performance of the model on the training set, and the cross-validation score reflects the performance of the model on the validation set. This graph can help identify whether the model is overfitting or underfitting

Heatmap, a data visualization technique, uses the depth of color (usually from light to dark) to represent the numerical size of the data, and intuitively shows the relationship or pattern between the numerical values. Use heatmaps to determine the correlation of factors affecting production.



1. Results

According to the results of the heat map, humidity, oxygen concentration and average high temperature are the most important factors affecting yield.

And the model prediction LSTM has the highest accuracy of 78%, followed by random forest, GBM, MLP, and SVM at 70%, and SVM at 68%.

Finally, LSTM is selected for yield prediction in model selection.

2. Wechat Mini program settings

Using flask to design lightweight programs, users can enter the future weather forecast queried on the front end, and then the model will return the prediction results.