Crop Health Classification

Progress Report
MALIS - Advanced Project

LABIDI Souha
ZAPATA Juan David

Introduction

Monitoring crop health is crucial for making timely decisions. It helps farmers take quick actions to protect crops and increase yields, ensuring food security and promoting sustainable farming. Traditional methods, like field inspections, are often costly and slow, highlighting the need for innovative solutions. Machine learning and remote sensing provide powerful tools to address these challenges. Using real-time satellite images, algorithms can analyze crop health faster and more accurately. The Telangana Crop Health Challenge focuses on classifying crop types using satellite data. This helps improve farming practices, reduce waste, and support better decisions. Finally, the project advances precision agriculture and sustainable farming.

Methods

Exploratory Data Analysis (EDA)

The very first step towards an appropriate development of a machine learning model is the understanding of the dataset with which it will be trained. First, a quick overview over the data is done through basic DataFrame methods such *describe()* and *info()*.

Afterwards, a univariate analysis is performed for each of the variables. The categorical variables are studied through pie and bar charts, while the numerical ones are studied through histograms. Then, bivariate analysis were performed between the target feature and the rest of the variables. For categorical variables, contingency tables and stacked bar charts were used while numerical variables were grouped by their category. Finally, we applied the Interquartile Range (IQR) technique to identify potential outliers in the numerical variables.

Feature Engineering

Following the results of the performed EDA some of the variables are removed from the dataset. The categorical ones remaining are encoded through a traditional label-encode approach while the numerical ones are standardize through scikit-learn *StandardScaler*. Additionally the date data is transformed to only keep the month information.

Model Training

Initially a Support Vector Machine model is built with a C parameter of 1. Then, a KNeighbors model with n equals to 3. Although, for this last model only two variables were chosen for training in order to avoid the curse of dimensionality (*CropCoveredArea* and *WaterCov*).

Results

Exploratory Data Analysis (EDA)

The first overview done through the data shows some of its most basic characteristics, such as:
1. There are 8775 data points, each representing a farm and associated to 19 features from which 6 are numerical, 12 are categorical and one represents the farm's geographic position.
2. The target feature has 4 possible values: Healthy, Diseased, Pests and Stressed.
3. There are not missing values present in the provided dataset.

The most important insight from the univariate analysis is the remarkable imbalance present in the target variable:

Table 1: Category Counts

Category	Count
Healthy	7214
Diseased	537
Pests	536
Stressed	488

From the bivariate analysis it is possible to observe that none of the variables has a strong correlation with the target feature behavior. However, some patterns can be spotted by sight.
After outlier detection analysis, we observed that only the variable CHeight contains values flagged as outliers. To better understand these values, we examined the rows with the minimum and maximum values:

- Groundnut crops typically have a height of around 20.
- Maize crops can grow up to 250.

We determined that these data points are not anomalies but legitimate variations based on the crop type. As a result, we decided to retain these outliers.

Feature Engineering

Through this process the dataset was reduced to only keep 10 attributes.

Model Trainning

The Support Vector Machine fails to separate the test data properly as it assigns the *Healthty* class to each one of the test samples.

On the other hand, the KNeighbors classifies the test data as follows: 2819 Healthy samples, 137 Diseased, 37 Stressed and 23 Pests. This first results were uploaded to the challenge platform, and a F1 Multi Class error metric of 0.7121 was returned.

Conclusions

1. The class imbalance remarked in the provided dataset must somehow be dealt with if better results are expected.
2. As none of the variables present in provided tabular dataset shows to have a strong relationship with the studied phenomenon occurrence, the need to collect satellite data through the locations given for the farms is enforced.
3. The categorical variables studied are not ordinal data, therefore using a label-encode approach could bias the learning process as it is better suited for ordinal data. Other approaches must be explored.
4. The model training phase is still in an early stage as very few options have been tested. Different models with different hyper-parameters must be tried out in order to find a better solution.

Contributions

Souha did the outliers identification while Juan the univariate and bivariate analysis. We both worked together for the feature engineering and model training steps.