**Data Science Project: winter catch crop classification using sentinel data as alternative to on site control**

*Author(s): Erez Kauffman, MSc horticultur**e*

Introduction:

The European legislative guidelines on the Common Agricultural Policy (CAP) lead to an obligation of area-wide information about cross-compliant Greening measures in EU countries. These guidelines require by farmers, cooperatives, and authorities to actively measures, but by on-site controlling, agricultural authorities can monitor only a small portion of all registered parcels. Using remote sensing-based agricultural monitoring on a large-scale it is possible to overcome this problem.

Catch crops, often referred to as cover crops, are crop mix usually grown in summer or winter between the main cropping seasons and are subsidized as Greening measures for farmland diversification and soil protection (Magdoff and Van Es, 2009; OSCAR, 2016).Cover crops, have couple of benefits to the soil, in which: soil micro-organism prosperity, soil protection due to erosion, water percolation, Nitrogen & organic compounds enrichment, prevention of nitrogen contamination of the groundwater caused by over-fertilization in winter time and more. following farmers success in the past year, Germany government provides differential subsidies to farmers based on the rate of success.

Nowadays, fields tagging for cover crops is implemented by field technicians which inspect cover crop plantation few times a year between July-April and by summing the data, gives their decision based on the cover crops coverage. These controls require large numbers of inspectors to survey several hundred thousand parcels within a short time frame of each catch crop season, in addition to other expenses. Consequently, a very low percentage of approximately 5% of all registered crop parcels can be controlled each year.

in this work I will present a novel approach to classify between cover crops and non-cover crops plantation using agriculture index (NDVI) and climate parameters on a parcel-level.

The data for the crop parcels obtained from different years (2016–2019) and four federal states in Germany, while the NDVI index obtained using satellite image (Sentinel 1 and Sentinel 2) and depicted using different time series calculations.

The outcome for the classification problem will be provided as 'yes' (1) for cover crop plantation and 'No' (0), for other. This outcome is influenced by the level of 'chlorophyl a' molecule in plants leaves, which is represented by the 'NDVI' index.The predictive model was validated with the test data in a model-internal accuracy assessment.

The objective of this study is the development of a new, fully automated method for large-scale winter catch crop detection based on the example of Germany.

The used approach allows authorities large-scale winter catch crops monitoring and helps them to understand if there are specific crop parcels which require special attention and different subsidies.

By merging the training datasets from different federal states and years, I could overcome the typical spatial and temporal overfitting problem in machine learning. Therefore, the study’s final classifier can be reliably transferred to new datasets in Germany and other regions with similar bio-geographical conditions.

Methodology (Project design):

The data that I will use is online-free download database from an academic paper which published on 8 May 2021. The data arrived consisted of 12 different tables, which correlates to different zones and years in Germany. In addition to the available data, I connected to data from different meteorological stations in accordance to the different zones and years in Germany and I pulled data regards: minimum monthly temperature, maximum monthly temperature, average relative humidity per month, average precipitation per month, average monthly soil temperature and average monthly sunshine duration. Using feature engineering in SQL I manipulate the different features and created new ones according to their importance.

The project will be based on the past 3 years data (Inclusion criteria) and will try to use it for the data classification of the current year (after the year is finished).

As explained earlier the outcome variable in which I would like to classify is 'Class' (1 for cover crop success, 0 for other-no success). The main confounder variables which can have some bias with the target variable, could be related to the climate features, which can have some dependencies within each other. I will have to identify it in the exploratory data analysis process and in some case, I might remove them.

The data exploration process will be subjected to the current use protocol in this field: descriptive statistics of all the variables and of the target variable include suspicious variables vs the target variable, correlation analysis. Outliers detection and missing values, included data cleansing. The distribution of the variables will be analyzed with and without missing values and in case distribution is changed and correlation is changed, it won’t be possible to remove the feature. Feature with high percentages of outliers will be also analyzed and in case the distribution and correlation doesn’t change, it will be possible to remove the variable. At the end of this part, features with very high correlation with each other will be examined, and in case they promote bias or act the same, one of them will be removed. Finally, a new exploratory data analysis process will be held after all the changes and as a preparation for the last part of the machine learning models.

## Models:

The data will be divided to different parts to examine it using machine learning models.

1st the data which is categorical will be converted using one-hot encoding to prepare it for the machine learning process (the old labels will be dropped), then the data will be randomly shuffled to prepare it for the next part of the data division and to prevent selecting same row values.

The data will be split using 'sklearn.model\_selection, train\_test\_split' to 'Test' (20%), 'Train' (without the target variable, 80%) and validation (which will be examined using cross validation in 10 parts for 'test score'). The observed models for the classification of the outcome variables will be: 'Logistic regression', 'Ridge classifier', 'K-nearest-neighbors', 'Support vector machine', 'Bagged decision trees', 'Random Forest classifier' and 'Gradient boosting machine (stochastic boosting machine)'.

Each model will be fitted using Train dataset (80%) and cross-validation in addition to the evaluation using 'Root mean squared logarithmic error' (lower value best) and 'Score' which presents the 'Accuracy' (higher value best), I found these two-evaluation metrics as the easiest to decide for the best model. At the end of this part each model will be improved using 'Hyper-parameters fine-tuning' using grid search for the main parameters. Best model score will be possibly use for future production.

## Deployment of your model

The QA of the project should be done by the data scientist which take care for the new data classification, some features could be a bit different and they may resemble the same similarity. The main units to be assessed are the NDVI features within the same time-frame.

The final user of the predictions could be companies related to local authorities which assess the final decision in digital platform for them. Prediction should be presented as a % accuracy in regard to specific plot, then a threshold for the minimum cover crop should be define. Plots with different coverage (>90%,80%-60%,<60%) should be subsidized differentially. In order to maintain best results the model should be incremented each year with thee relevant data, according to model success. Best result model (from all the observed models which mentioned previously is random forest classifier (85% accuracy) as same as Gradient boosting machine. Due to faster running time as compared to Gradient boosting machine, this model is with high priority for the production.

It is important to emphasize that due to climate features addition and feature engineering compared to the article researchers, accuracy % is higher in my project (85%) compared to them (82%).

Results:

As explained before, Train data was divided to 80% and test to 20% (154 features, 14390 rows).

Feature with Outliers were examined with and without, and result wasn't significant. Missing values were tested in case there are more than 60%, the variable was deleted. Imputation executed using 'MICE' imputation. Some feature distributed normally and some not, as a rule of thumb, features were treated using methods for non-normally distributed.

variable features in regards to the NDVI index, were treated using time-frames from early-July to beginning of April next year, accordingly climate parameters were treated the same.

Pre exploratory data analysis, I used SQL to transform the data and generate new features.

Conclusion:

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how you got the final prediction. You have to discuss the limitations of the model, when it can be used and when not.

Project started finding the relevant topic for my career, this part was not easy, since the use of agricultural data in machine learning is quite a new topic, not all the datasets were fits with the requested amount of data for machine learning.

I tried to ask researchers for the data share, which wasn't easy, but finally I found this article, which enabled free dataset. Than, I wanted to enrich the data, so according to my knowledge the most important features were climate features (I pull them from meteorological stations in Germany for the specific zones), this part evolves extracting the relevant data for the time frame of the project in the relevant years (July-April).

Luckily, the article researchers also added a column with the manual tagging for the block and their final prediction column for their prediction, I removed those columns in order to use this dataset for my Train set.

The model has some limitations. Since the data was gain for Germany, it will be possible to use it for other countries with similar climate in Europe (Austria, North Italy, Belgium etc.) but precautions should take in consideration. It will be best to collect the data from one year from different country and first compare same features for similarities. In case the % deviation is high, and distributions are not equal, it will be better to collect new data and not aggregates the data for analysis.