Architectural Decisions Document Template

1 Method for data quality assessment

The information has been revised to verify that the crop codes are correct and that all codes are present in all years.

The distribution of the data has been revised to verify that there is not a year in which the distribution of the data is out of the ordinary.

2 Method for feature engineering

The first feature extraction technique applied was the conversion of the crop codes so the can be used in the LSTM model. Crop codes is a categorical variable and must be converted to numerical values in order to be used as a deep learning model input. This conversion has been done using the feature embedding technique, to convert each crop code to a vector of fixed dimensions.

3 Algorithm decision and model performance indicator

The initial idea was to use LSTM models, these models take advantage of the contextual information of a series, so they are perfect for modeling data with a temporal component. In our case, the problem is that the time series is short (9 years) and there is no access to a previous series. Different models have been made combining LSTM networks with 1-dimensional convolution networks to extract new characteristics, the result is relatively satisfactory. As a performance metric, the **f1-score on the test set** has been used, this metric balances between precision and recall and makes it more robust in unbalanced datasets.

Because LSTM networks have a longer training time than other deep learning models, especially if regularization parameters that disable the use of the cuDNN implementation are used, a reduced dataset has been extracted for model training to speed up the construction of the different iterations of the model. The frequencies of each crop have been maintained to ensure the representativeness of this dataset. The sample dataset has been divided into three blocks to have separate data for training, validation and testing with percentages 70% / 15% / 15%. Train and test sets are used in the keras callback to measures performance during the training and the test set is used for final evaluation.

4 Framework

Se ha utilizado keras como librería para el entrenamiento del modelo por ser una librería ampliamene soportada por tensorflow, que permite entrenamiento basado en GPU y soporta modelos LSTM.