**Loop Q Prize**

**Challenge B: Crop Yield Prediction**

Alexandros P. Pouroullis

1. Overview of the Solution
   1. A screenshot of a computer

      Description automatically generatedData flow
   2. Key components
      1. Pandas dataframe for storing the data after extracting it from a .csv file
      2. sklearn MinMaxScaler for scaling features using the min-max formula
      3. Numpy arrays for storing scaled features
      4. Sklearn implementation of the K-Nearest Neighbours algorithm for regression
   3. Model used

The algorithm that I settled on after extensive experimentation and comparison was the K-Nearest Neighbours algorithm for regression, using the sklearn implementation. In its default state it performed exceptionally well, but after hyperparameter tuning I managed to squeeze a little bit more performance. The configuration of the tuned model is as follows:

* + - * Distance metric: Manhattan distance
      * Weight function: distance (places greater emphasis on closer neighbours)
      * Number of neighbours: 2

1. Graphical user interface, application

   Description automatically generatedStrengths of the solution  
   The final model generalizes quite well and obtained a root-mean-squared error of ~8000 hg/ha after 10-fold cross-validation, which can give us surety that it’s likely to perform similarly on unseen data. As was seen in the graphs of the predictions and errors of the predictions made by the model, it, for the most part, came extremely close to the truth value.

Figure : Graphs of predictions and error of predictions made by the final KNN model

Moreover, the model was fitted to data that is quite readily available to most farmers and doesn’t require advanced systems of sensors to capture the data, and thus would make implementing such a predictive model feasible for many.

1. Limits and bias of the dataset/solution  
     
   The Dataset  
   As was pointed out in the data exploration phase in the notebook, the dataset has a gross imbalance in the representation of countries, and to a lesser extent crops. It also doesn’t contain data for all countries and is limited in the number of crop categories that it accommodates. Thus, the fitted model can only make predictions for countries and crop items that are present in the training data set, restricting the generalizability of the model.

The solution

The model is, of course, only as robust as the training data allows it to be. And because the training data contains imbalances and doesn’t have data for all countries and crop items, the model is only capable of making predictions on data that it’s familiar with, i.e., with countries and crop items it’s seen before.