**Team Name**: Green Thumb Coders

**Objective/Problem Statement:** build ML model that most accurately predict corn phenotype values:

1. Predict Yield (y\_YLD\_BE)
2. Moisture Content (y\_MST)
3. Test Weight (y\_TWT) of the corn crops in 2008 using prior year’s genomic, environmental, phenomic data

And then Utilize contextual information and data science techniques to develop a model that minimizes the average RMSE

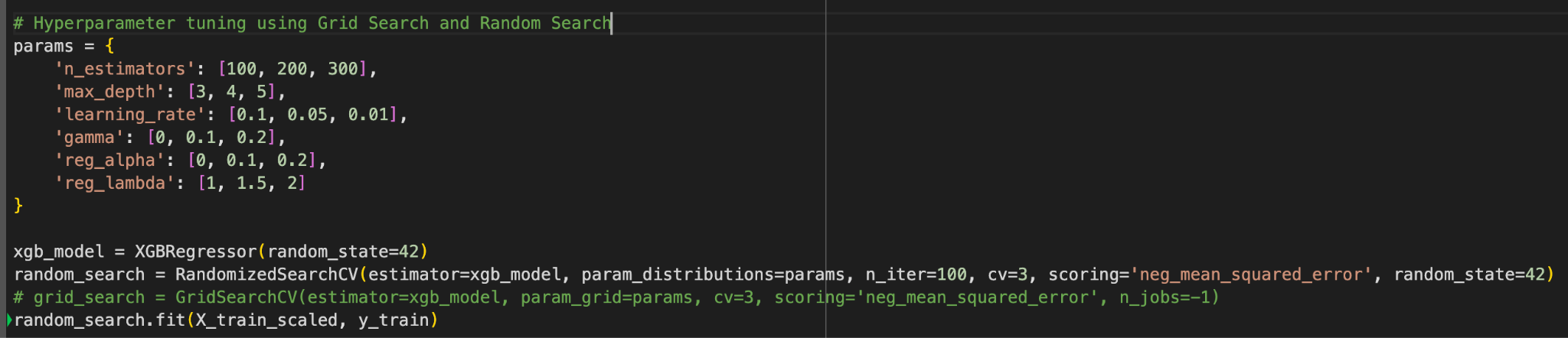
Our code: [**Colab Link**](https://colab.research.google.com/drive/1fETzv4ortt1YqhYVM8K-D1sLEswre0ce#scrollTo=LrCoPEcrfU6m)**,** [**codebase**](https://colab.research.google.com/drive/1fe2xi6lOVvlW_x415QIZ3NEK6G4OlLzs?usp=sharing)

**Method used**: Random Forest Regression

**Data Preprocessing**: checking for missing values(NaN), Impute missing values using techniques like mean imputation and KNN imputation, standardized features using StandardScaler (by removing the mean and scaling to unit variance).

**Model fitting and optimization:**

* ***Random Forest Regression*:** combines *multiple decision trees* to make predictions, resulting in a robust and accurate model. It is an *ensemble learning technique* used for regression tasks, which involve *predicting continuous target variables*.
* The reason we chose this tool is because it is highly robust to overfitting because:
  + It combines multiple trees, which collectively generalize better.
  + It can handle large datasets with many features.
  + It provides feature importance scores, indicating the contribution of each feature to the model's predictions.
* We used techniques like Grid Search (performs better for this dataset) and Random Search(which is faster than grid search and is used when we have a large parameter space) to find the best hyperparameters for our model and implement cross-validation (e.g., k-fold cross-validation) to ensure the model's performance is consistent across different subsets of the data.

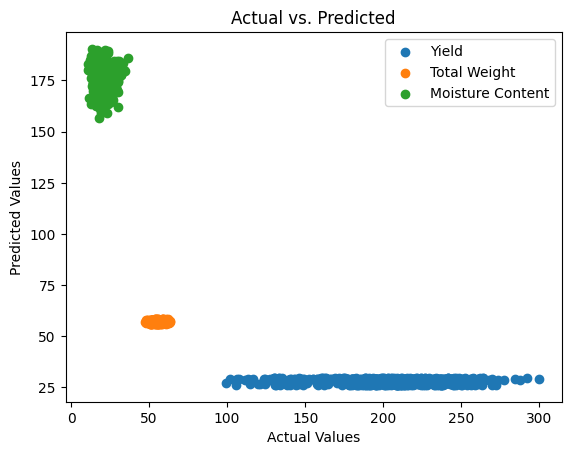
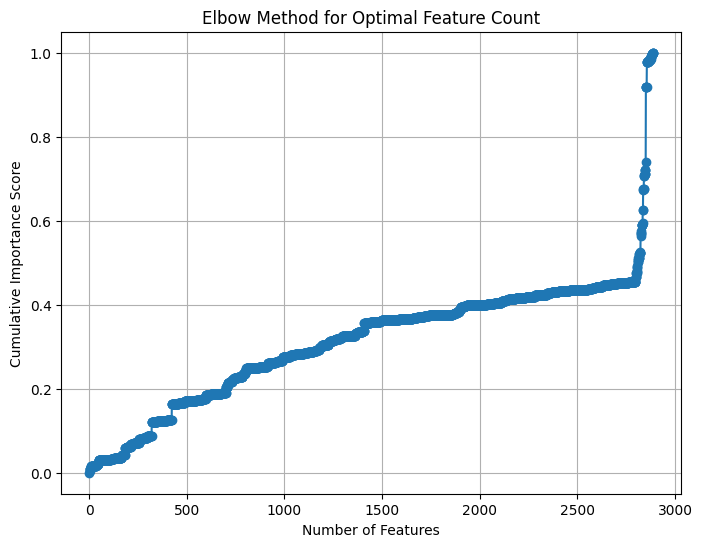


* We also applied regularization techniques in case our model tends to overfit. For XGBoost, we can control overfitting by adjusting parameters like gamma, reg\_alpha, and reg\_lambda.

**Evaluation, predictions, visualization:**

* Lastly, we evaluated the models using Mean Squared Error (MSE). Lower MSE(and in turn RMSE) values indicated better model performance and the best MSE. We observed amongst all our iterations, the best RMSE to be 8.9060.

A graph with blue bars

Description automatically generated with medium confidence

**Data Interpretation:**

Environmental variables, majorly precipitation, temperature and number of days it rains in a month has the greatest influence on the corn yield, total weights of corn and moisture content. In addition, nitrogen at soil depth 15-30 cm also has influence on the prediction. Two of the genomic variables are also in the top 10 important features. Lastly, it is important to pay attention to the location of the corn farm considering that each location has its unique characteristic as indicated in the prediction.

**Conclusion:** Given the limited time frame, we have optimized our model with several transformations and hyperparameter tuning on our base model. However, we see good potential in the feature importance from our interpretation and we can possibly improve our model and results further if we implement the feature importance scaling in our data as future research and study.