*\*\*This was a school project done in a group of 5 people, and this document is adapted from a presentation on the project.*

Main focus of this model

How might we leverage machine learning to recommend crop types to farmers based on the surrounding temperature, amount of rainfall, quality of the surrounding air, and quantity of pesticides used to maximise yield?

Data preparation

A screenshot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated

The separate csv files were loaded into dataframes and null values were replaced to prevent errors when handling the data later on.

A screen shot of a computer program

Description automatically generated

As the range of years, countries, and crops are different for each dataset, we only took data from the common countries across the datasets, chose 6 staple crops, and limited the year to be between 1990 and 2020.

A computer screen shot of text

Description automatically generated

For each crop, we matched the data of the independent variables from the separate datasets, which are the four agricultural production factors, to the dependent variable, which is the yield, based on their corresponding country and year and stored them in a dictionary.

A computer screen with white text

Description automatically generated

This dictionary was then converted into a dataframe and rows with missing values were removed to prevent them from skewing the results.

Data analysis

By plotting the data, we observed a few trends.

A graph of different types of yield

Description automatically generated with medium confidence

Firstly, in general, crop yield tends to decrease with respect to all of the independent variables except for temperature.

A graph of blue dots

Description automatically generated with medium confidence

We decided to further investigate the crop yield with respect to temperature for each crop and realized that yield is not linearly related to temperature. Instead, it tends to peak at differing temperatures for different crops. However, due to the importance of temperature in affecting yield, we decided not to leave it out of the model.

A group of blue graphs

Description automatically generatedvs A graph of blue dots

Description automatically generated with medium confidence

Lastly, an interesting observation was that despite the vastly different relationships that temperature and precipitation have with respect to yield,

A graph of blue circles

Description automatically generated

Temperature and precipitation seem to be related as precipitation levels tend to rise as temperature rises.

Model

A black screen with white text and colorful lines

Description automatically generated

To ensure consistency in the test results, we fixed the random state, distribution between test and feature data sets, number of iterations of gradient descent the model goes through, and the rate at which gradient descent is implemented.

A computer screen with white and blue text

Description automatically generated



With those variables set, we move on to the actual training of the model. While yield is dependent on crop types, it is a variable that we are not able to quantify. Hence, we trained a separate multiple linear regression model for each crop. We start by retrieving the relevant columns of data for the specified crop, then splitting that data into test and training sets.

A computer screen with white and blue text

Description automatically generated



A computer screen shot of a program code

Description automatically generated

A computer screen shot of a program code

Description automatically generated

After that, we scaled the training data. This step was needed as when we initially tried to train the model using the raw data, it led to an error as the values in the columns were too large for the program to compute. Scaling the values using a common integer like 1000 or using the log function were also not possible as the range of values between columns were too large and the temperature variable contained negative values. Therefore, we settled on scaling the values with the column means.

A computer screen with white and blue text

Description automatically generatedA screen shot of a computer

Description automatically generated



With that, we trained the training dataset using gradient descent and obtained the beta values, before using those beta values to conduct predictions on the test data set.

A computer screen shot of a program

Description automatically generated

With those results, we then evaluated each crop’s model accuracy by calculating the root mean square value as well as the ratio of the root mean square value to the mean yield of the crop.

A screenshot of a computer

Description automatically generated

Overall, while individual RMSE values seem quite low, it is actually relatively large in comparison to the yield values, which leads us to believe that the model is not the most accurate, likely due to the fact that there are other qualitative independent variables (e.g. skill level of farmers, farmers' familiarity with the crops) which affect the dependent variable that were not taken into account.

By comparing the individual ratio of RMSE to mean yield for each crop, we can see that it is higher for some which shows the model’s varying accuracy between crops.

A screenshot of a computer program

Description automatically generated

With this, we then decided to improve on the model by changing the data processing technique used and scale the data using z-normalization.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

Lastly, we calculated the new RMSE values and compared the differences. From this, we can see that the RMSE value decreased for all crops, which means that the average deviation of the predicted versus actual yield has decreased across the board, indicating that scaling the values with Z Normalization has increased the accuracy of all models.

Individually, the RMSE value and ratio decreased by differing magnitudes for the crops, which shows that z normalization has varying effectiveness on the different models, likely because some datasets had more outliers than others since z normalization reduces the impact of outliers.

Conclusion

In conclusion, our model can roughly estimate the yield of crops with varying accuracy based on our selected independent variables, which should be good enough to give a recommendation on the crop since the difference between yields of different crops tend to differ more greatly than the RMSE value. However, it is important to note that the actual predicted yields are not very accurate. To boost its accuracy, more independent variables can be explored for the various crops to find variables with a stronger relationship to the yield of the specific crops.