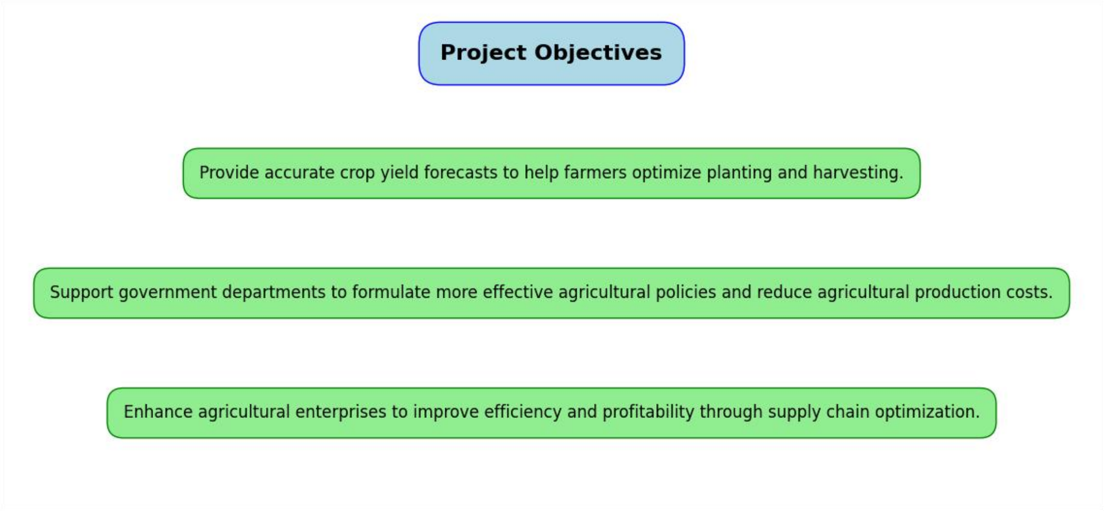
The topic I chose for Assignment 3 is the same as Assignment 1

How the Victorian government uses agricultural data to predict crop yields？

Project Description:

This project predicts crop yields in Victoria, aiding agriculture and stakeholders in planning and resource allocation. By integrating historical data like crop yields, weather patterns, soil characteristics, and agricultural practices, accurate predictions for various crops and seasons are generated. These insights assist farmers, policymakers, and stakeholders in decision-making for crop planning, resource allocation, risk management, and market strategies, fostering sustainable growth in Victoria's agriculture.

Project Objectives:



Project Objectives

Provide accurate crop yield forecasts to help farmers optimize planting and harvesting.

Support government departments to formulate more effective agricultural policies and reduce agricultural production costs.

Enhance agricultural enterprises to improve efficiency and profitability through supply chain optimization.

Data Science Roles and Responsibilities:

1. Data Engineer

- Responsible for collecting, cleaning and processing agricultural data from various sources, ensuring data quality and optimizing the data retrieval process.

2. Data Scientist

- Design and develop machine learning models to analyze historical agricultural data and identify patterns in future crop yields. Work with agricultural experts to communicate factors affecting crop yields to stakeholders.

3. Machine Learning Engineer

- Develop algorithms, optimize model performance, and adjust algorithms based on feedback, working closely with data scientists.

4. Agricultural experts

- Provide domain knowledge, guide data science projects, validate model predictions, and interpret analysis results for practical agricultural applications.

5. System Architect

-Responsibilities: Design and implement data storage and processing infrastructure to ensure system scalability and reliability.

Business model：

Application areas：

The project is suitable for the agricultural sector, farmers, agricultural enterprises and all aspects of the agricultural product supply chain.

Benefit：

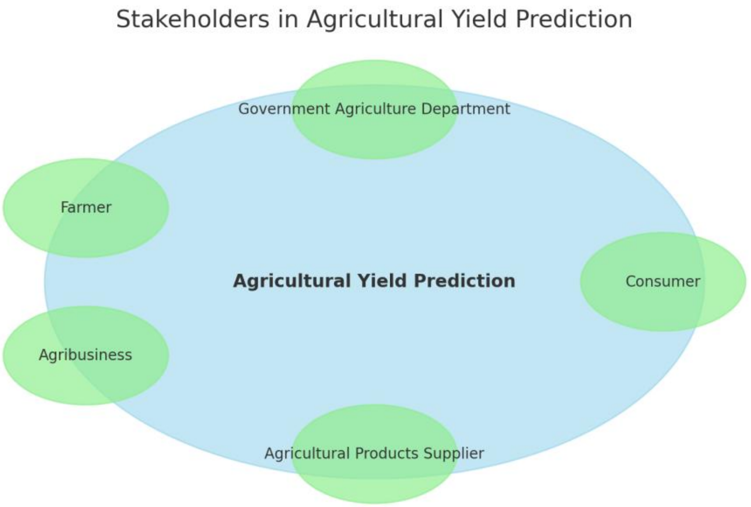
1. Government departments: Reduce agricultural production costs by 5% to 10% through more effective policy formulation.

2. Farmers: Increase yields by 10% to 15% by optimizing planting and harvesting.

3. Agribusiness: Improve efficiency and profitability through supply chain optimization.

4. Agricultural product suppliers and consumers: Stable market supply and fair prices promote market stability and development.

Stakeholders：


Stakeholders in Agricultural Yield Prediction
Government Agriculture Department
Farmer
Agricultural Yield Prediction
Consumer
Agribusiness
Agricultural Products Supplier

Challenge:

1. The unstructured or semi-structured nature of data and the difficulty of storing it in different formats.

2. Data quality issues, such as sensor errors and human recording errors, require data cleaning, verification, and correction.

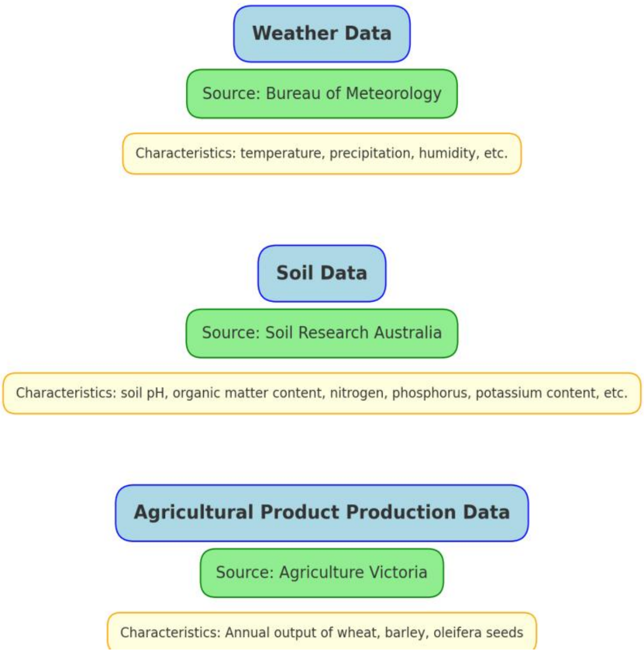
3. Data that exist in paper or non-digital form need to be digitized.

4. Uncertainty in the agricultural production environment and data privacy security issues.

To overcome these challenges, the project will take a variety of measures, including using data cleaning and standardization technologies to improve data quality and availability, adopting advanced predictive models and algorithms to deal with uncertainties, and implementing strict data security and privacy protection measures to ensure that the project sustainable operation and development.

Data collection sources and analysis：

Data Sources：



Weather Data

Source: Bureau of Meteorology

Characteristics: temperature, precipitation, humidity, etc.

Soil Data

Source: Soil Research Australia

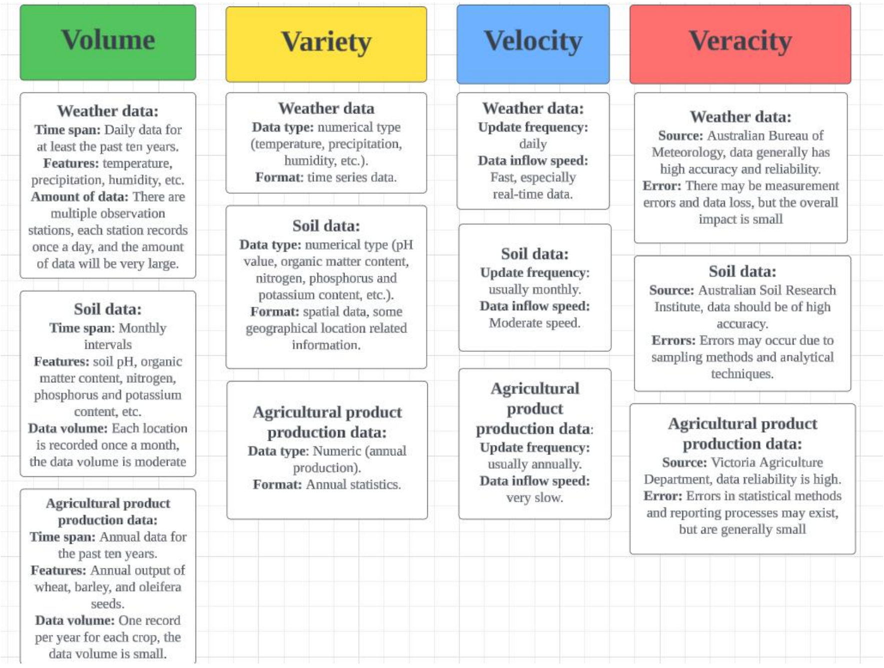
Characteristics: soil pH, organic matter content, nitrogen, phosphorus, potassium content, etc.

Agricultural Product Production Data

Source: Agriculture Victoria

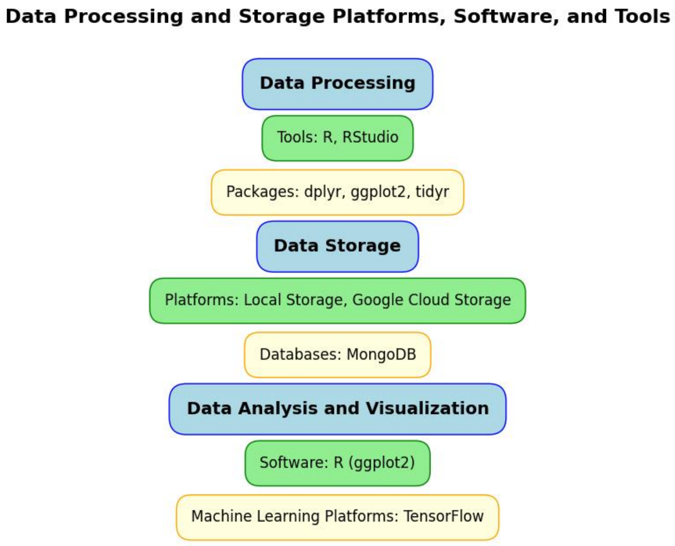
Characteristics: Annual output of wheat, barley, oleifera seeds

Data characteristics (4V)：



Volume	Variety	Velocity	Veracity
Weather data: **Time span:** Daily data for at least the past ten years. **Features:** temperature, precipitation, humidity, etc. **Amount of data:** There are multiple observation stations, each station records once a day, and the amount of data will be very large.	**Weather data** **Data type:** numerical type (temperature, precipitation, humidity, etc.). **Format:** time series data.	**Weather data:** **Update frequency:** daily **Data inflow speed:** Fast, especially real-time data.	**Weather data:** **Source:** Australian Bureau of Meteorology, data generally has high accuracy and reliability. **Error:** There may be measurement errors and data loss, but the overall impact is small
Soil data: **Time span:** Monthly intervals **Features:** soil pH, organic matter content, nitrogen, phosphorus and potassium content, etc. **Data volume:** Each location is recorded once a month, the data volume is moderate	**Soil data:** **Data type:** numerical type (pH value, organic matter content, nitrogen, phosphorus and potassium content, etc.). **Format:** spatial data, some geographical location related information.	**Soil data:** **Update frequency:** usually monthly. **Data inflow speed:** Moderate speed.	**Soil data:** **Source:** Australian Soil Research Institute, data should be of high accuracy. **Errors:** Errors may occur due to sampling methods and analytical techniques.
Agricultural product production data: **Time span:** Annual data for the past ten years. **Features:** Annual output of wheat, barley, and oleifera seeds. **Data volume:** One record per year for each crop, the data volume is small.	**Agricultural product production data:** **Data type:** Numeric (annual production). **Format:** Annual statistics.	**Agricultural product production data:** **Update frequency:** usually annually. **Data inflow speed:** very slow.	**Agricultural product production data:** **Source:** Victoria Agriculture Department, data reliability is high. **Error:** Errors in statistical methods and reporting processes may exist, but are generally small

Data processing and storage platforms, software and tools:

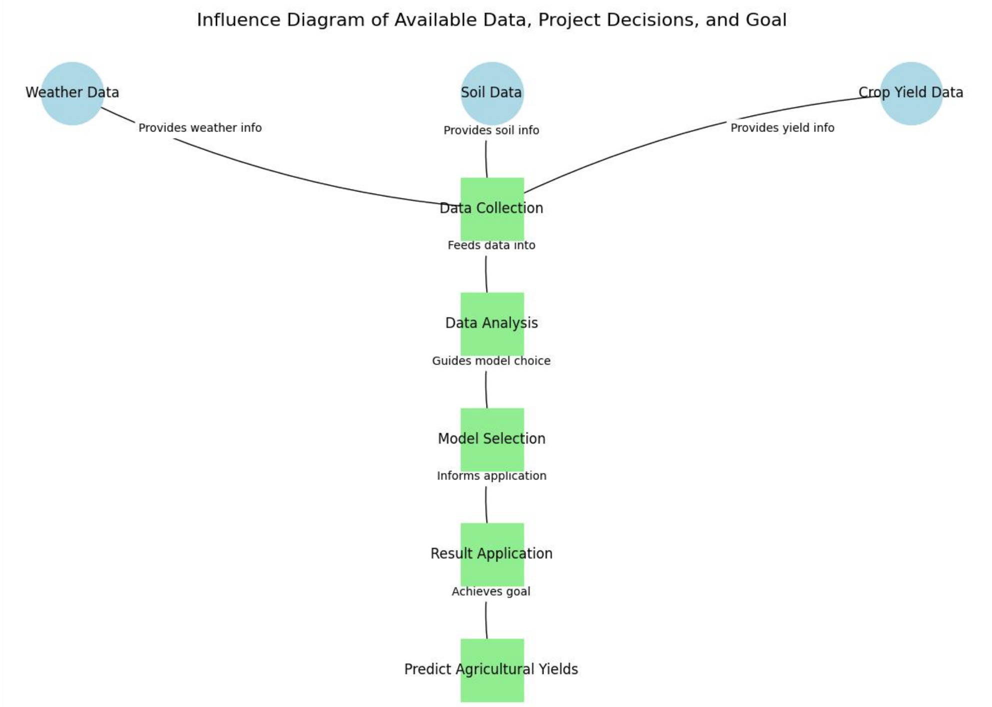

Data Processing and Storage Platforms, Software, and Tools
Data Processing
Tools: R, RStudio
Packages: dplyr, ggplot2, tidyr
Data Storage
Platforms: Local Storage, Google Cloud Storage
Databases: MongoDB
Data Analysis and Visualization
Software: R (ggplot2)
Machine Learning Platforms: TensorFlow

R and R Studio are used for data analysis and visualization, and they have excellent statistical and data analysis packages.

Local storage and Google Cloud Storage are used for the storage and management of small-scale and large-scale data respectively.

MongoDB is suitable for storing and processing diverse, unstructured agricultural data. TensorFlow is used to build and train complex machine learning models to predict agricultural yields

Impact diagram:



Data analysis and statistical methods：

1. Time series analysis(ARIMA)

Purpose:

Analyze time series trends of weather data and annual production data.

Reason:

（1）Identify seasonality and trends:
Agricultural yields are closely related to seasonal changes, and time series analysis can capture these changes and identify long-term trends.

（2）Predicting the future:
By analyzing historical data, models can be use to predict future weather conditions and yields, helping farmers and policymakers make more informed decisions.

Advanced output:

（1）Seasonal component: Break down time series data and identify cyclical changes.

（2）Trend Component: Identifying long-term growth or decline trends.

（3）Prediction intervals: Provides future predictions and their confidence intervals to help quantify the uncertainty of the predictions.

2.Multiple linear regression：

Purpose: Establish the relationship between weather, soil data and annual crop yields.

Reason:

（1）The model is simple and easy to interpret:
Multiple linear regression provides a direct explanation of each independent variable, helping to understand which factors have a significant impact on yield.

（2）Quickly verify the relationship in the data:
As a baseline model, multiple linear regression can quickly verify the linear relationship in the data and evaluate the importance of each variable.

Advanced output:

（1）Regression coefficient: Provides the magnitude and direction of the impact of each independent variable on yield.

（2）Significance test: Determine which variables contribute significantly to the model.

（3）Model fit: Evaluate the overall fitting effect of the model, such as R^2 value and adjusted R^2 value.

3. Random Forest Regression：

Purpose: Improve prediction accuracy and reduce overfitting.

Reason:

（1）Handle a large number of features and complex non-linear relationships:
Random forests can automatically select important features and handle complex interactions between features.

（2）Strong robustness: It has good robustness to missing values and noisy data, reducing the risk of over-fitting.

Advanced output:

（1）Feature importance: Provides a ranking of the importance of each feature to predictions to help understand which features are the most important.

（2）Prediction error analysis: Evaluate the performance of the model on the training set and test set, and detect potential overfitting or underfitting problems.

（3）Robustness of the ensemble method: Improve the robustness and generalization ability of the model through the integration of multiple trees

4. Support Vector Regression (SVR)：

Purpose: Processing high-dimensional data and non-linear relationships.

Reason:

（1）Good generalization ability: SVR finds the best hyperplane in high-dimensional space, has good generalization ability, and is suitable for complex nonlinear relationships.

（2）Robustness to noisy data: SVR can handle noisy data in high-dimensional space, reducing the risk of over-fitting.

Advanced output:

（1）Support vectors: Identify support vectors that have an important impact on the model.

（2）Kernel function selection: Capture different types of nonlinear relationships by selecting different kernel functions (such as linear kernel, polynomial kernel, RBF kernel).

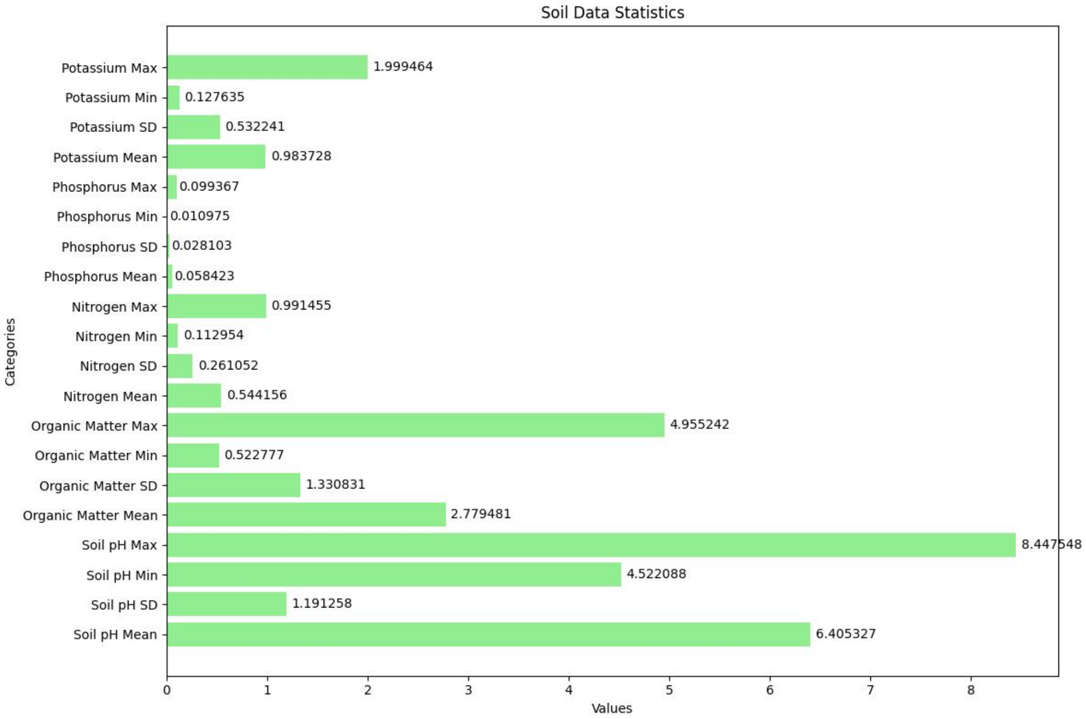
（3）Model performance evaluation: Evaluate model performance through cross-validation and select optimal parameter settings.

Data analysis demonstration:

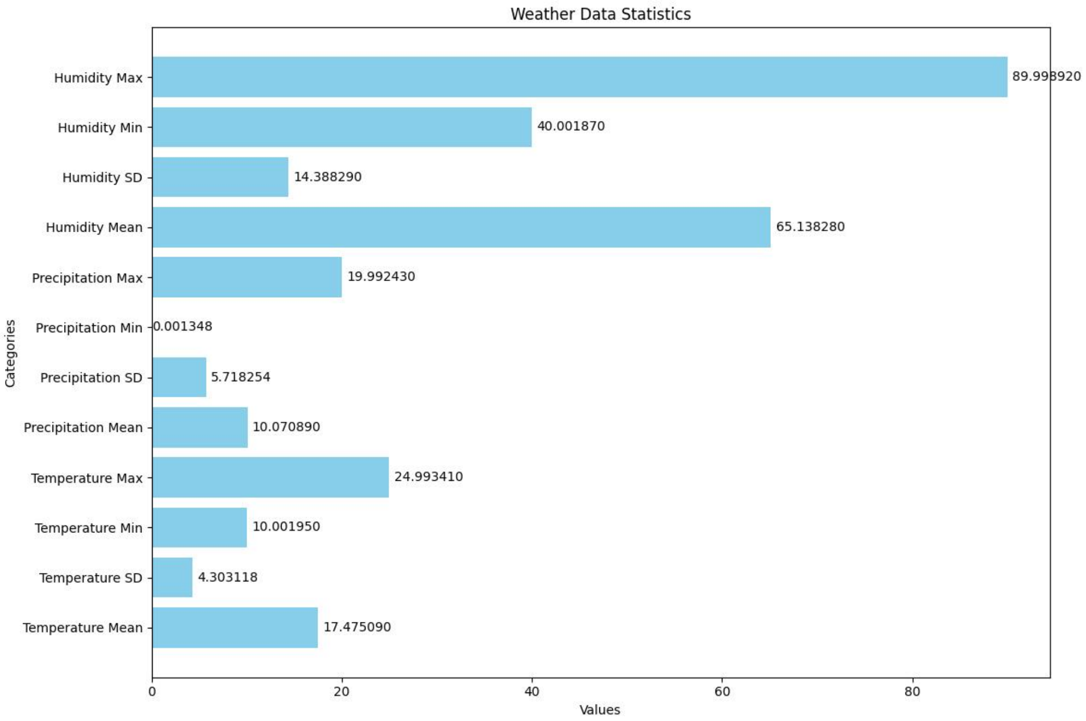
Dataset description:

Agricultural dataset contains wheat, barley and rapeseed production for different regions on different dates. Melbourne weather dataset records temperature, precipitation, and humidity on different dates. Soil dataset provides soil pH, organic matter, nitrogen, phosphorus, potassium content, and longitude and latitude information in different areas on different dates. These data can be used to analyze the relationship between agricultural yields and weather and soil conditions, thereby helping to develop more effective agricultural management and decision-making strategies.

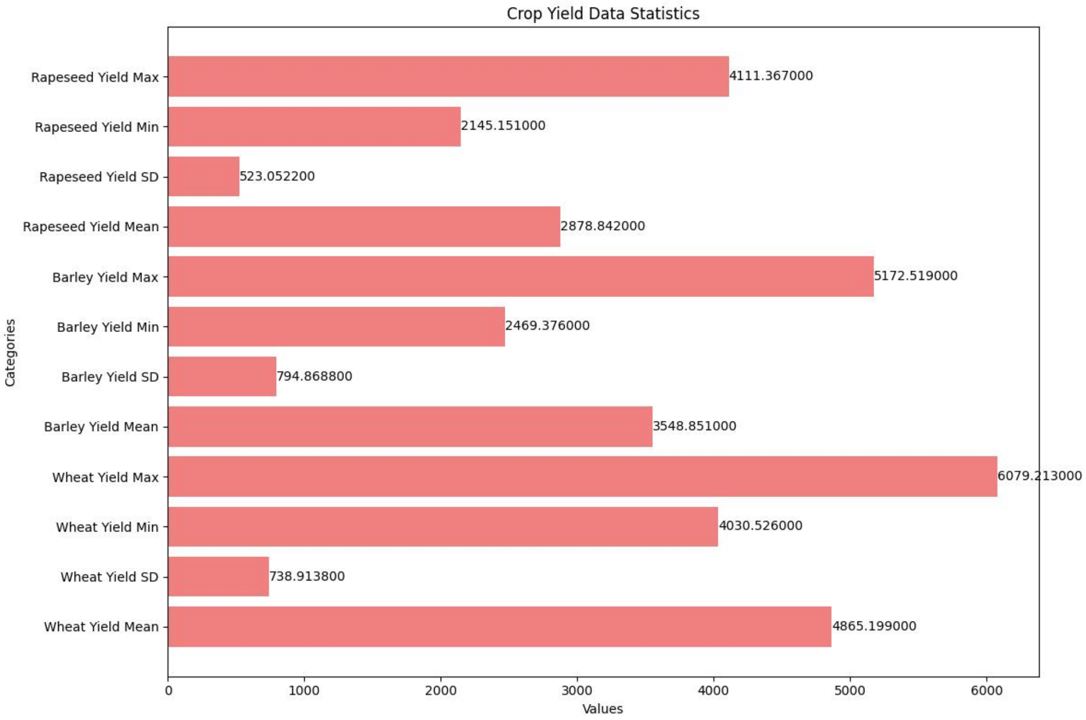
Basic analysis of data set data:



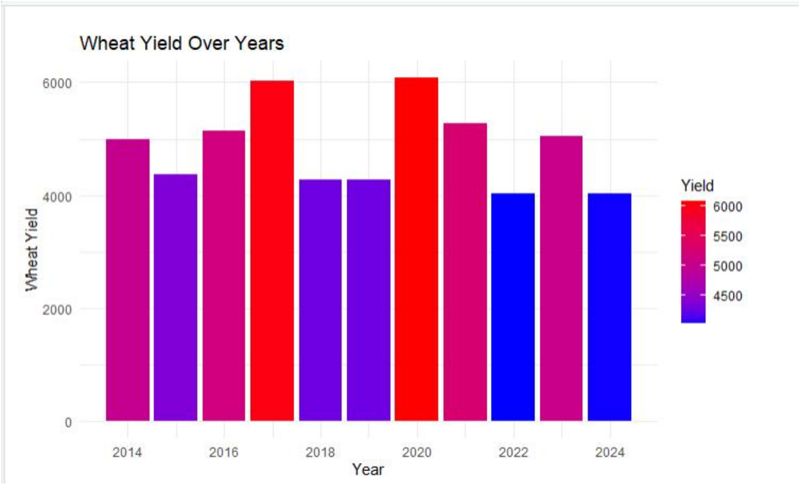
First, summary statistics of the soil data are calculated to fully understand the various properties of the soil. These statistics include the average, standard deviation, minimum and maximum values of soil pH, organic matter content, nitrogen, phosphorus and potassium.



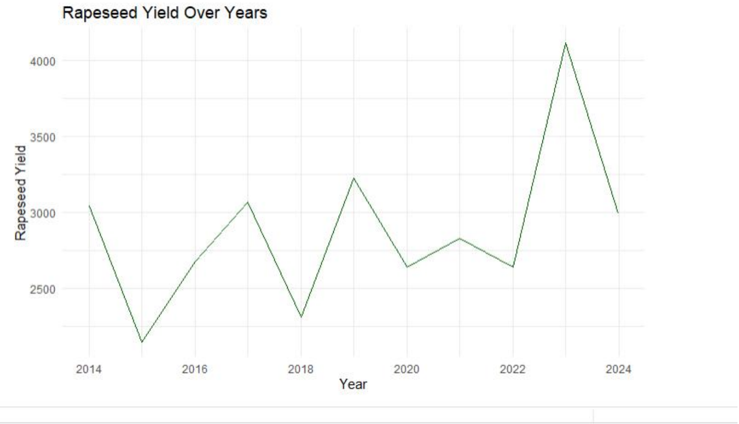
Learn about temperature, precipitation, and humidity by calculating summary statistics for weather data. These statistics include mean, standard deviation, minimum, and maximum values for temperature, precipitation, and humidity.



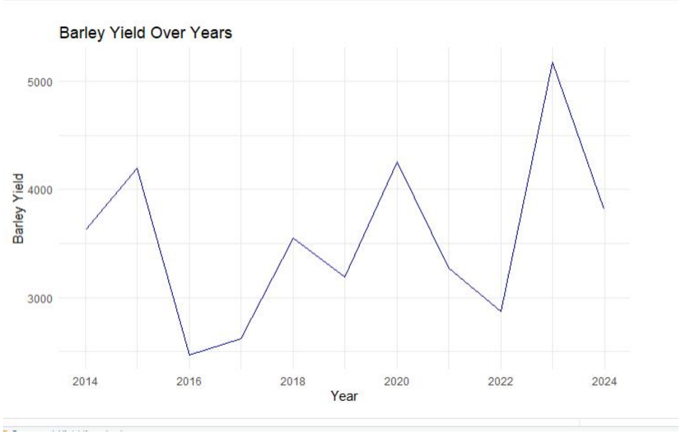
Calculate summary statistics for agricultural production data to get a complete picture of wheat, barley, and rapeseed production. These statistics include the mean, standard deviation, minimum and maximum values of each crop yield.



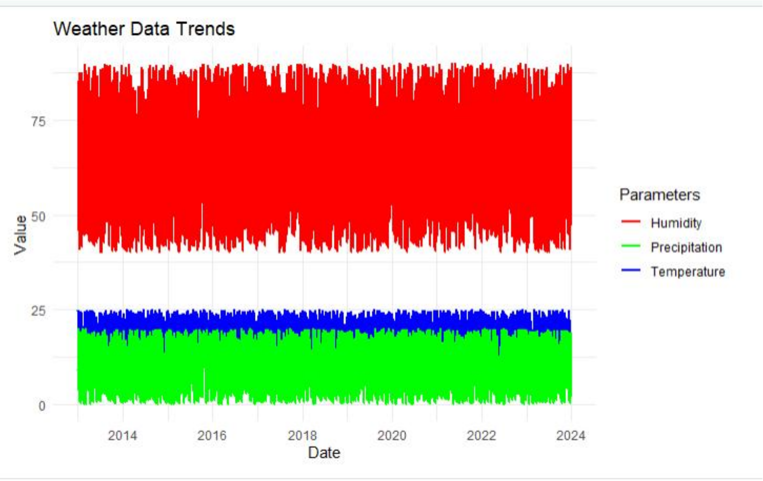
Bar chart of wheat production over the years



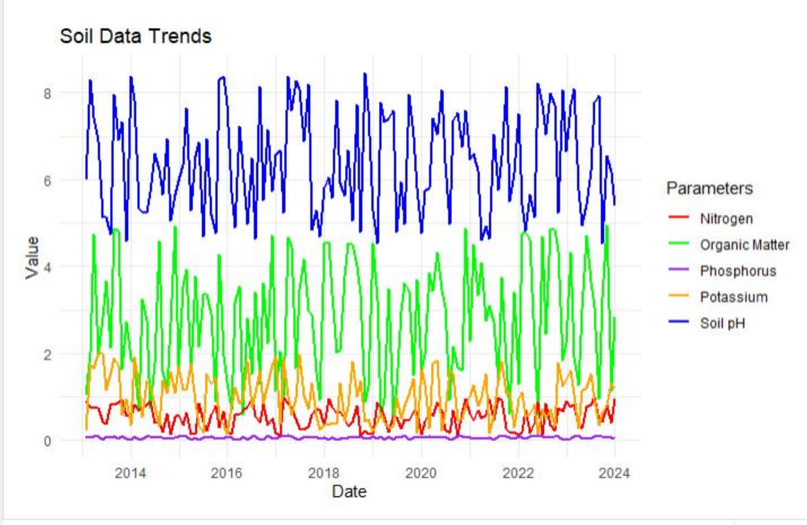
Rapeseed Yield line chart over the years



Barley Yield line chart over the years



Weather Trends chart

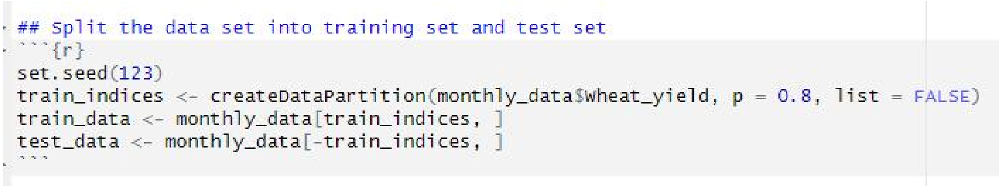


Soil Data Trends chart

Build a random forest regression model and predict future agricultural output:

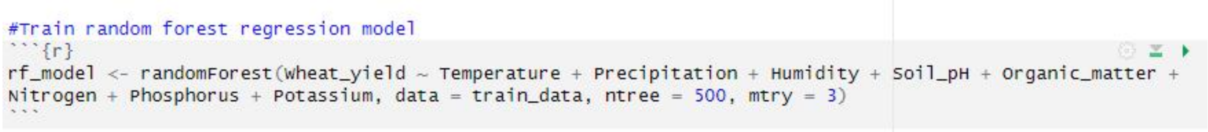
After analyzing the basic features of the dataset, I selected some features that impact the prediction target for further processing. This involved converting the data format, extracting the required year and month information, and summarizing it.

First, I divided the monthly_data dataset into a training set and a test set, where 80% of the data is used for training and 20% for testing. This ensures that model training and evaluation use different data subsets to reduce the risk of overfitting and improve the model's generalization ability.



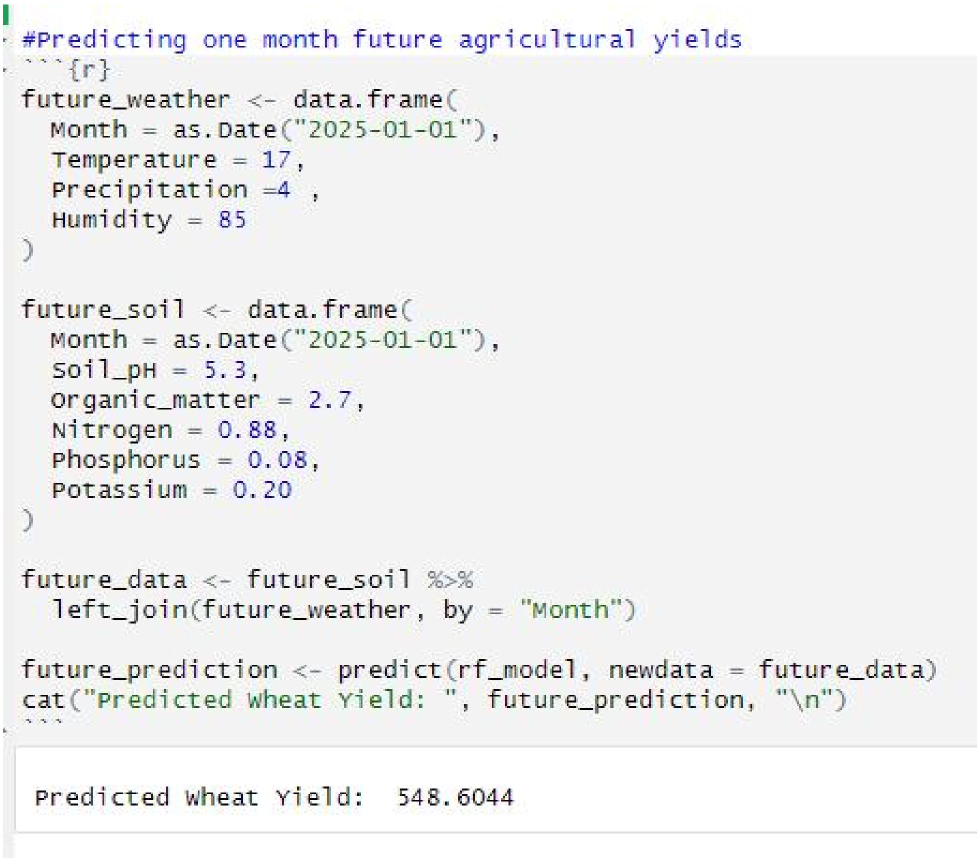
~~~
## Split the data set into training set and test set
```{r}
set.seed(123)
train_indices <- createDataPartition(monthly_data$Wheat_yield, p = 0.8, list = FALSE)
train_data <- monthly_data[train_indices, ]
test_data <- monthly_data[-train_indices, ]
```
~~~

Then, I used train_data to train a random forest model so that it can learn the relationship between input features (such as temperature, precipitation, etc.) and the target variable (wheat yield).



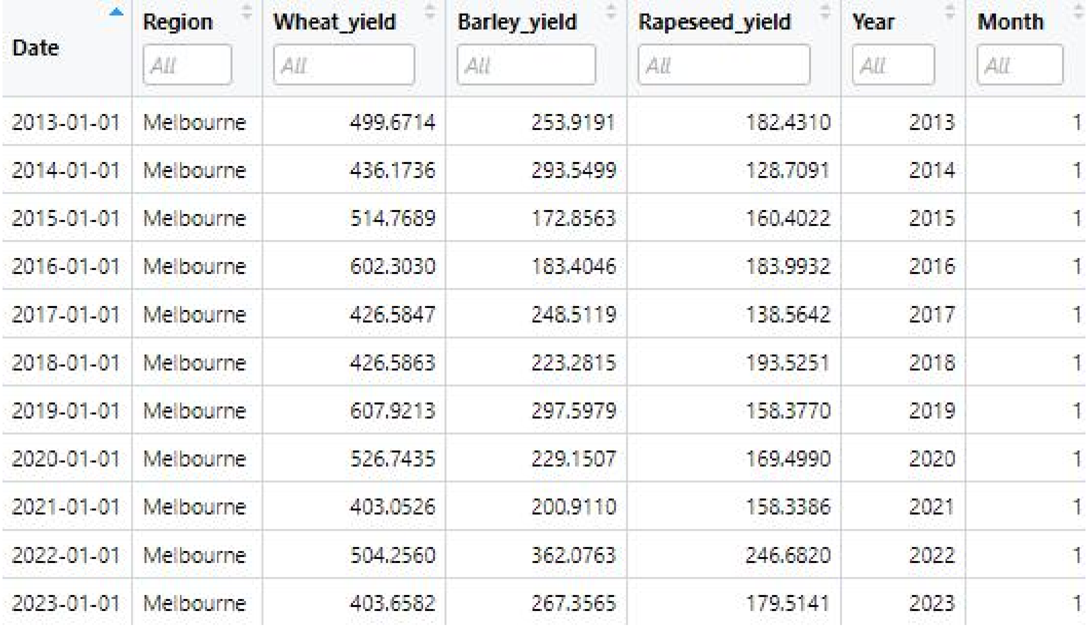
~~~
#Train random forest regression model
```{r}
rf_model <- randomForest(Wheat_yield ~ Temperature + Precipitation + Humidity + Soil_pH + Organic_matter +
Nitrogen + Phosphorus + Potassium, data = train_data, ntree = 500, mtry = 3)
```
~~~

Then, I used the model to predict future agricultural output based on set data.



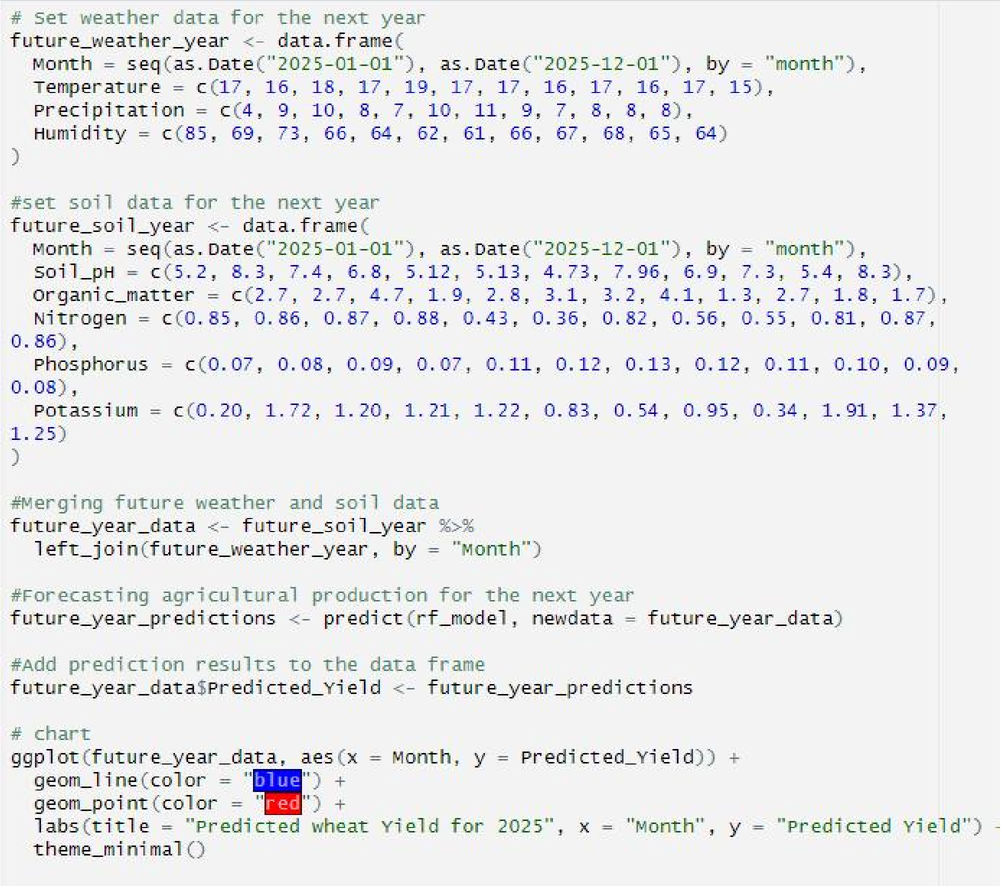
~~~
#Predicting one month future agricultural yields
```{r}
future_weather <- data.frame(
  Month = as.Date("2025-01-01"),
  Temperature = 17,
  Precipitation =4 ,
  Humidity = 85
)

future_soil <- data.frame(
  Month = as.Date("2025-01-01"),
  Soil_pH = 5.3,
  Organic_matter = 2.7,
  Nitrogen = 0.88,
  Phosphorus = 0.08,
  Potassium = 0.20
)

future_data <- future_soil %>%
  left_join(future_weather, by = "Month")

future_prediction <- predict(rf_model, newdata = future_data)
cat("Predicted Wheat Yield: ", future_prediction, "\n")
```
~~~

```
Predicted Wheat Yield:  548.6044
```

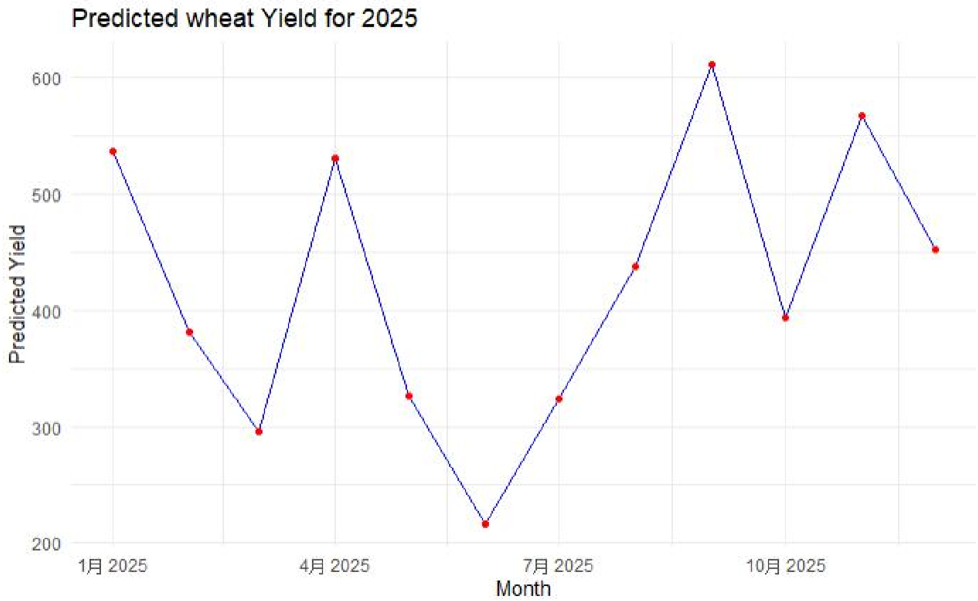


Date	Region	Wheat_yield	Barley_yield	Rapeseed_yield	Year	Month
	All	All	All	All	All	All
2013-01-01	Melbourne	499.6714	253.9191	182.4310	2013	1
2014-01-01	Melbourne	436.1736	293.5499	128.7091	2014	1
2015-01-01	Melbourne	514.7689	172.8563	160.4022	2015	1
2016-01-01	Melbourne	602.3030	183.4046	183.9932	2016	1
2017-01-01	Melbourne	426.5847	248.5119	138.5642	2017	1
2018-01-01	Melbourne	426.5863	223.2815	193.5251	2018	1
2019-01-01	Melbourne	607.9213	297.5979	158.3770	2019	1
2020-01-01	Melbourne	526.7435	229.1507	169.4990	2020	1
2021-01-01	Melbourne	403.0526	200.9110	158.3386	2021	1
2022-01-01	Melbourne	504.2560	362.0763	246.6820	2022	1
2023-01-01	Melbourne	403.6582	267.3565	179.5141	2023	1

Finally, I use the model to predict agricultural yields for the next year based on given data. By comparing with the data from previous years, we can see that the difference is small, which shows that the prediction results are relatively accurate.



```r
# Set weather data for the next year
future_weather_year <- data.frame(
  Month = seq(as.Date("2025-01-01"), as.Date("2025-12-01"), by = "month"),
  Temperature = c(17, 16, 18, 17, 19, 17, 17, 16, 17, 16, 17, 15),
  Precipitation = c(4, 9, 10, 8, 7, 10, 11, 9, 7, 8, 8, 8),
  Humidity = c(85, 69, 73, 66, 64, 62, 61, 66, 67, 68, 65, 64)
)

#set soil data for the next year
future_soil_year <- data.frame(
  Month = seq(as.Date("2025-01-01"), as.Date("2025-12-01"), by = "month"),
  Soil_pH = c(5.2, 8.3, 7.4, 6.8, 5.12, 5.13, 4.73, 7.96, 6.9, 7.3, 5.4, 8.3),
  Organic_matter = c(2.7, 2.7, 4.7, 1.9, 2.8, 3.1, 3.2, 4.1, 1.3, 2.7, 1.8, 1.7),
  Nitrogen = c(0.85, 0.86, 0.87, 0.88, 0.43, 0.36, 0.82, 0.56, 0.55, 0.81, 0.87,
0.86),
  Phosphorus = c(0.07, 0.08, 0.09, 0.07, 0.11, 0.12, 0.13, 0.12, 0.11, 0.10, 0.09,
0.08),
  Potassium = c(0.20, 1.72, 1.20, 1.21, 1.22, 0.83, 0.54, 0.95, 0.34, 1.91, 1.37,
1.25)
)

#Merging future weather and soil data
future_year_data <- future_soil_year %>%
  left_join(future_weather_year, by = "Month")

#Forecasting agricultural production for the next year
future_year_predictions <- predict(rf_model, newdata = future_year_data)

#Add prediction results to the data frame
future_year_data$Predicted_Yield <- future_year_predictions

# chart
ggplot(future_year_data, aes(x = Month, y = Predicted_Yield)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Predicted wheat Yield for 2025", x = "Month", y = "Predicted Yield") +
  theme_minimal()
```



Data usage standards:

(1) Data quality check: We cleaned the original data, including handling missing values, etc. Ensure data accuracy and completeness.

(2) Data splitting criteria: We used 80% of the data as the training set and 20% of the data as the test set. This segmentation method helps reduce the risk of overfitting and ensures that the model has good generalization capabilities.

Data accessibility, security and confidentiality:

(1) Ensure data accessibility to those involved in agricultural yield forecasting projects while preventing unauthorized access.

(2) Protect weather, agriculture, and land data from unauthorized access and leakage by implementing encryption and access control measures.

(3) Ensure the confidentiality of sensitive data and comply with relevant data privacy regulations and policies.

Potential ethical issues in data use:

(1) Data Privacy: Ensure that data collected and used does not violate the privacy rights of farmers and other stakeholders.

(2) Data fairness: Avoid possible biases in data analysis and models to ensure the fairness of agricultural forecast results.

Think critically and creatively：

Potential limitations：

（1）The training data may not adequately represent future conditions or extreme situations, causing the predictive model to perform poorly in these situations.

（2）Agricultural data are often subject to historical conditions, and models may struggle to accurately predict future environmental conditions if extreme climate changes occur

Future direction：

（1）Further combine multiple models to improve the prediction model and improve the accuracy of predictions.

（2）Expand the data source and add more dimensions of data to enhance the robustness of the model.

Conclusion：

By forecasting crop yields in Victoria, this project will provide valuable insights to agriculture and related stakeholders to help them make effective decisions and resource allocations, thereby promoting the sustainable development of Victorian agriculture.

Dataset download link

[1]http://www.bom.gov.au/climate/averages/tables/ca_vic_names.shtml#name_s
[2]https://www.abs.gov.au/statistics/industry/agriculture/agricultural-commodities-australia/2021-2022
[3]https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/metadata/70105
[4]https://github.com/chriszou710/5145

Reference List

[1] Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and electronics in agriculture*, *121*, 57-65.

[2] Ahmed, M. U., & Hussain, I. (2022). Prediction of wheat production using machine learning algorithms in northern areas of Pakistan. *Telecommunications policy*, *46*(6), 102370.

[3] Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., ... & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and forest meteorology*, *274*, 144-159.

[4] Ruan, G., Li, X., Yuan, F., Cammarano, D., Ata-Ul-Karim, S. T., Liu, X., ... & Cao, Q. (2022). Improving wheat yield prediction integrating proximal sensing and weather data with machine learning. *Computers and Electronics in Agriculture*, *195*, 106852.

[5] Li, L., Wang, B., Feng, P., Li Liu, D., He, Q., Zhang, Y., ... & Yu, Q. (2022). Developing machine learning models with multi-source environmental data to predict wheat yield in China. *Computers and Electronics in Agriculture*, *194*, 106790.

[6] Murakami, K., Shimoda, S., Kominami, Y., Nemoto, M., & Inoue, S. (2021). Prediction of municipality-level winter wheat yield based on meteorological data using machine learning in Hokkaido, Japan. *Plos one*, *16*(10), e0258677

[7] CSIRO. (n.d.). Soil and Landscape Grid of Australia. Retrieved May 18, 2024, from https://www.clw.csiro.au/aclep/soilandlandscapegrid/GetData-DAP.html

[8] Australian Bureau of Meteorology. (n.d.). Climate data for Site 086147. Retrieved May 18, 2024, from http://www.bom.gov.au/climate/averages/tables/cw_086147.shtml

[9] Department of Agriculture, Water and the Environment. (n.d.). Victorian crop report. Retrieved May 18, 2024, from https://www.agriculture.gov.au/abares/research-topics/agricultural-outlook/australian-crop-report/victoria