Climate Change and Environmental Influences on Wheat Growth

Arpon Sarker

47454130

Contents

[Executive Summary 3](#_Toc213234846)

[Introduction 4](#_Toc213234847)

[Background 4](#_Toc213234848)

[Related Works 6](#_Toc213234849)

[Problems 9](#_Toc213234850)

[Objectives 10](#_Toc213234851)

[Solution 11](#_Toc213234852)

[Problem Formulation 11](#_Toc213234853)

[Data Cleaning and Merging 13](#_Toc213234854)

[Exploratory Data Analysis 15](#_Toc213234855)

[Model Selection and Training 21](#_Toc213234856)

[Discovering the Most Climate-Resistant Cultivar 21](#_Toc213234857)

[Modelling Frost Damage 26](#_Toc213234858)

[Identifying the Most Productive Environment for Wheat Production 30](#_Toc213234859)

[Image Segmentation 35](#_Toc213234860)

[Costs 44](#_Toc213234861)

[Ethics and Privacy 45](#_Toc213234862)

[Conclusion 45](#_Toc213234863)

[Appendix 47](#_Toc213234864)

[Reference List 54](#_Toc213234865)

# Executive Summary

This report investigates the multifaceted impacts of climate change on wheat production and proposes actionable solutions to enhance yield, identify frost damage, optimise environmental conditions, and segment images of wheat using machine learning and data science techniques. By analysing plot-level, simulation and weather data from the National Variety Trials, the study identifies cultivars with superior climate resilience using clustering and regression, models frost damage using neural networks, quantifies the effects of soil and geographical factors on yield using random forest regression and lasso regression and uses computer vision techniques and deep learning for wheat image segmentation. Firstly, the most climate-resistant cultivars found exhibited a mean yield of 837.29 kg/m^2 and the lowest coefficient of variation which highlights this particular breed had a strong performance despite the changing environmental conditions and highlights this and the other top 10 breeds for future research in breeding programs. Secondly, using a feedforward neural network to model the frost damage, an f1-score of 0.93 was achieved which indicated a robust distinction between crops suffering from no frost damage and at least level 1 of frost damage. This provides a reliable tool to be used for early warning and risk assessment for crops. Thirdly, the analysis investigates environmental factors that optimise yield revealed that longitude, latitude along with trace elements of copper, magnesium and organic carbon exert a substantial influence on yield amount. Lastly, three models were compared to test on a global wheat image dataset and were Deeplabv3, Fully Convolutional Network, and Segformer. The final Segformer model which was chosen to represent the solution to this project incorporated data augmentation techniques, normalisation, early stopping, and focal loss to achieve a mean Intersection over Union of 0.6 which is below the threshold for reliability. All of these outputs of these projects were outputted into a PowerBI dashboard to enable farmers and agronomists to visualise and interpret complex data for informed decision-making. This report also considers the costs, ethics and privacy in implementing these projects, especially when regarding the National Variety Trial dataset. This report hopes to provide actionable insights and recommendations for enhancing wheat yield and supporting sustainable food supply chains under a more uncertain future of climate change.

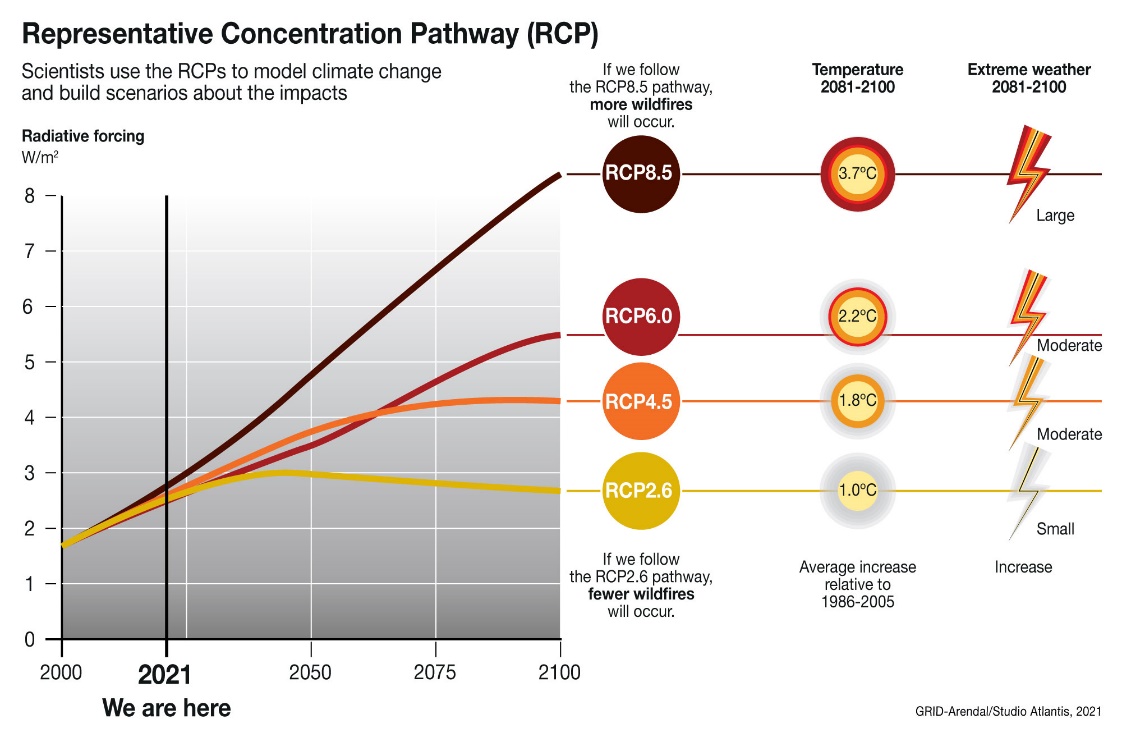
# Introduction

By 2030, wheat production must grow to 60% to keep up with demand which will be severely impacted by climate change providing disastrous consequences to farmers, livestock, and Australian food supply chains. (FOLUR, n.d.) This report argues for the investigation of the impacts of weather and climate change on wheat growth and seeks to address a key question: what variety of wheat could be analysed to lead to enhanced breeds resilient to climate change and erratic weather patterns? The significance of this analysis is to explore options that aim to reduce grain shortages and explore the efficacy of organic agriculture using real-world data. The analysis will be conducted on plot-level, simulation, and weather data from the National Variety Trials. Using the data science techniques of data fusion and machine learning, the datasets will be combined and analysed for insights into which cultivar or varieties of wheat to focus on and will provide models for farmers to mitigated damage from weather patterns. In the first section, I illustrate the background knowledge required to better inform this report’s applications to the industry. The second and third sections delve into the current literature and the gaps or issues that require adapting to this report’s thesis. The fourth section is the objectives which lists the projects and goals to be met. The fifth section is the proposed solution and covers the entire data science process in solving the goals and objectives using machine learning models and using computer vision techniques on an image dataset to be summarised in a dashboard. The sixth and seventh sections are the costs and privacy and ethical considerations. Lastly, the conclusion summarises the report, identifies the extent to which goals were achieved, and persuades the use of the report in legitimate use cases.

# Background

The report’s interpretations on the effects of wheat production and its use cases provide value to farmers and agronomists and require the understanding of the National Variety Trials (NVT), climate change, and APSIM. Firstly, the National Variety Trials is a multi-location research program for evaluating and comparing the performance of different varieties of wheat across different growing conditions. The significance of this program is for informed decision-making and enhanced crop productivity. The datasets that will be used for this report are from this program which includes plot level, simulation, and weather data. The core limitations of using these datasets are the unbalanced nature of the trials which can invalidate the statistical tests conducted since they require assumptions that the varieties and agronomical treatments have been allocated randomly to trials. (Eichi, Okamoto, Garnett, & Eckermann, 2020) This may even introduce bias where inconsistencies in data collection methods across different regions or times can reduce model reliability and generalisability. Moreover, many of the columns could not be used due to the high number of missing records which will be expounded in more detail in the Exploratory Data Analysis section. Furthermore, the value of the grain is based on protein content which is not measured. This lack of data renders deciding on breeding programs based on profitability infeasible. Another factor in understanding which breeds to be focused on is climate variability.

Secondly, this report focuses on climate change and the effects of weather to better understand the significant impacts on yield. The projected climate change modelled for Southwest Australia predicts median wheat yield based on Representative Concentration Pathways (RCP). This model illustrates concentrations of greenhouse gases leading to radiative forcing amongst the years 2000 to 2100. This can be seen in figure 1 which shows the detrimental effects predicted by these models. The median wheat yields decline by 26% to 38% under RCP 4.5 by 2090 which is the ‘most-likely’ scenario. Another more severe scenario projects a decline of 41% to 49% for RCP 8.5. (Department of Primary Industries and Regional Development, 2020) The Western Australian government’s modelling determined that rising temperatures and a decrease of rainfall by 28% accounted for 27% reduction in water-limited yield potential. (Department of Primary Industries and Regional Development, 2020) Furthermore, these two factors will reduce the flowering period which is the period where the yield of the potential crop is established. (Cotton Seed Distributors, n.d.) Hence, climate variability is a nationwide issue which needs to be investigated for more informed decision-making on climate-resistant crops using the National Variety Trials datasets. Dealing with climate change requires urgency where integrating crop simulation tools give a quantitative understanding of this important issue.

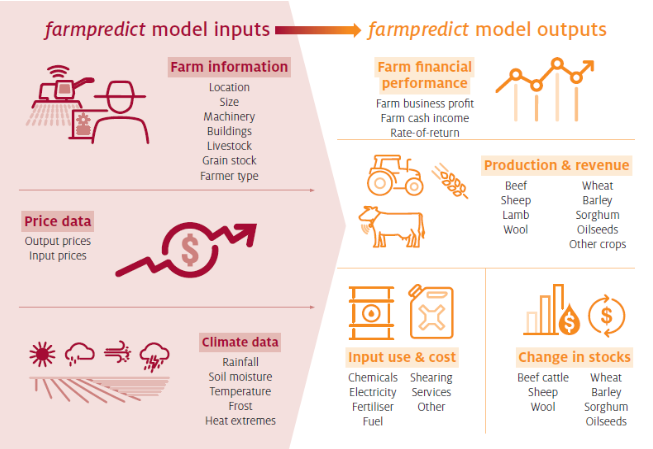


*Figure 1: RCP Graph from 2000 to 2100* (Grida, 2022)

# Lastly, the current literature and the NVT datasets given use the Agricultural Productions Systems Simulator (APSIM) to simulate crop data aiding the understanding of wheat growth against numerous environmental factors. APSIM is a modelling framework and can simulate wheat alongside more than 20 other crops and is flexible in attempting user inputs on crop, soil or utility modules. (CSIRO, n.d.) The advantages of using APSIM is improved management in farm resources and vital in experimentation and research for experts. The current literature utilises APSIM for research endeavours that are aligned very closely with the objectives of this report.

# Related Works

This section reviews the current literature on climate change and its effects on wheat production and profitability, erosion risk and frost damage. Firstly, researchers investigating the effects of climate change on the profitability of Australian farms – dependent on yield – utilised the statistical model *farmspredict* which outputs revenue, costs, farm inventories and profit based on weather conditions under a projected 2050 climate. Figure 2 showcases the model’s inputs and outputs which better inform how changes in climate data and farm information affect production and revenue. The study focused on Australian broadacre farms which accounts for 60% of Australia’s land mass. (Hughes et al., 2021) The study farms sampled are between 2015 to 2019 excluding 2017 and consists of 2712 observations out of 6312 total observations that are related to crops. The observations were inputted into the *farmspredict* model which was trained on 40269 farms and corresponding special climate data and used the *xgboost* regression algorithm to demonstrate the performance of the model with out-of-sample validation tests. (Hughes et al., 2021) The climate projection data that was inputted into the model is from CSIRO and the Bureau of Meteorology which was daily time step data centred on 2050. The scenarios included different environments across RCP4.5 and RCP8.5. The results from this study demonstrated the changes of rainfall and temperature across different cropping regions and the loss of profits. From the most severe future scenario RCP8.5, the simulated profits declined by 50%. The solutions or adaptations that are recommended based on the loss of profits are quite simplistic and doesn’t consider breeding programs as a lasting and permanent solution. (Hughes et al., 2021) The next study also attempts to find solutions by looking into the connections between climate change, erosion risk and wheat production.



*Figure 2: farmspredict model inputs and outputs*

Secondly, the effects of climate change on wheat production and erosion risk in South Australia’s cropping zones supports the report’s purpose in choosing varieties of wheat resistant to climate change and erosion. This study focuses on soil as a significant resource in wheat growth and the effects of wind and water erosion on soil fertility, carbon content, and water holding capacity. The researchers analysed these effects using the APSIM crop simulation model, South Australia’s soil and land attribute special datasets and time series climate datasets. The assumptions for the modelling are the climate change projection will increase by 0.4C to 1.8C by 2030 and 0.8C to 5.5C by 2070 and average annual rainfall can decrease by 15% at 2030 or by 45% at 2070. The APSIM model takes these assumptions and uses the inputs of soil characterisation, starting conditions, and climate information to output the dataset. The results for 5% rainfall reduction average around a 6% loss in post-harvest biomass production compared to a 10% rainfall reduction average to a 11% loss and a 20% reduction in rainfall which leads to a 24% reduction in biomass. Another factor that requires minimising the risk is frost damage which is examined to minimise crops unable to contribute their yield. (Liddicoat et al., 2012)

Thirdly, a study conducted on frost damage trends and their impact on yield is related to this report’s analysis of mitigating the risk of further damage to crops during the developmental stages. Frost damage was chosen as a focal point in damage reduction due to higher average nighttime temperatures increasing the occurrence of frost based on climatic conditions of the last six decades. This is reinforced by the idea that the simulation study conducted demonstrated a 20% increase in mean national yield just from a 10% improvement in frost tolerance and earlier sowing dates. (Zheng et al., 2015) The study consists of using the APSIM 7.6 model to simulate wheat crops for 60 sites.

The assumptions made were that the vegetative stage of wheat is unaffected by frost due to being more frost tolerant. The datasets that were used spanned from 1957 to 2013 whereas, this report is focused on a more frequent timeframe due to the focus on climate change. Moreover, the frequency of frost events increased from 1957 in 20% of the Australian wheatbelt region. Another result is the majority of the wheat yield advantage derives from reducing the frost damage temperature threshold from 0C to 1C. Figure 3 graphed below showcases FT1 to FT5 meaning a frost tolerance up to -1C to -5C and Ftot refers to completely no frost damage where the higher the frost tolerance the greater the simulated yield increase. (Zheng et al., 2015) All of these studies contribute to the report in different ways but their methods and focuses leave a gap in reliably acknowledging the effects of climate change and environment on yield.



*Figure 3: Frost Tolerance vs. Simulated Yield Increase*

# Problems

This section informs the issues in analysing the impacts of climate change on wheat production to reduce assumptions and make clearer decisions in breeding crops with higher resistance to weather.

Firstly, the study on climate change’s effects on Australian farms is limited by the large uncertainties faced by the wide range of potential outcomes over projected rainfall. For instance, the projections for winter rainfall are disagreed upon by multiple climate models used in the study. For southeastern Australia, the data from the last 25 years demonstrates a negative trend strongly where the observed rainfall is at the dry end of the projected range which has the least amount of variability compared to the other regions. In appendix 1, 6 different climate models are visually represented as illustrating different results. The report’s solutions will identify the uncertainty in future rainfall to support farmers adapt earlier and more assuredly. Moreover, the study used both farm-scale analysis using the *farmspredict* micro-simulation model and broad regional coverage with the given datasets. These models and statistical approaches allow updating the farm data to the current date to identify impacts based on regions and farm types but cannot account for future adaptation or carbon dioxide effects which leads to larger estimated impacts that bio-physical model studies. The report focuses on the future of wheat crop production and thus this further analysis cannot be used for the project. Furthermore, the study also takes into account the sheep and beef industries which may influence the profits of the farms, leading to a weaker correlation between the profits of the farm and the crop yields. Another issue is the emphasis on average farm profits but not analysing the effects on risk such as frequency of drought years. The study also emphasises broadacre farms and only investigates the Australian ‘Wheat-Sheep zone’ whereas, the National Variety Trial datasets are much more widespread and offer comparably more varied locations. In addition, the issues pertaining to this study can be extended to the issues of climate change affecting wheat production and erosion risk which also does not focus nor solve this report’s contention. (Hollaway & Knights, 2025)

Secondly, the study on the effects of climate change on wheat production and erosion risk includes issues that this report will take into account when considering choosing the best breeds with the highest resistance. The study used ‘modelling soil profiles’ as APSIM soil files based on landscape mapping but due to resource constraints, considerable simplifications have been made rather than considering all local-scale soil data. This did not account for water flow across the land which may have affected the results in identifying erosion risk. Moreover, the generalisation of soil parameters is also used to reduce confusion due to the large scale the study is conducting. This report will not take into account erosion but will look at soil data corresponding to weather changes. Moreover, the climate change scenarios used are based on historic records of daily temperature and rainfall and hence the future scenarios have the same variability as the historic datasets which introduces uncertainty. In addition, the crop variety was only limited to one medium maturing wheat variety which is not useful for this project since no crop can be chosen to further use climate-resistant traits for breeding. Furthermore, the study’s model overestimates the post-harvest biomass due to APSIM unable to account for weed or disease impacts. Another set of methods used, based on simulation data, are pervasive in the study of frost damage’s impact on crop yields and requires an extension to align itself with the report’s aim in modelling frost damage.

Thirdly, the study on frost damage’s effects on simulated yields has issues that need to be extended by this report to be able to reliably determine which environmental factors and to what extent can frost damage affect wheat production. The findings based on frost trends is derived from simulation yield outcomes where the dataset ends in 2013 which lacks recency, especially in a climate change context which is dependent on recent data for reliable forecasting. Moreover, my solution will not be completely dependent on simulation data and post-heading frost damage is seldom accurately recorded given the difficulty of estimating the yield if the crop was not affected by frost. For my solution, it will use real-world data for further accuracy and reducing any assumptions made. The last limitation that can be expanded upon is the topography of the farm allowing for different areas of the farm to be afflicted by different levels of frost damage, but this report uses the NVT dataset which has plot-level data to avoid this. (Zheng et al., 2015) These issues will now be solved in the report’s objectives which align itself with understanding crop yield against the effects of climate change.

# Objectives

The project’s objectives all pertain to satisfying farmer’s needs for informed decision-making, enhancing crop productivity, and creating preliminary alerts based on severe weather events. These objectives can be split into multiple parts which include finding the best cultivars or genotypes of wheat to further breed climate resistance, identifying frost damage using the crop’s attributes, and identifying the weather and soil conditions that are best suited to wheat growth.

The overall objectives for this project utilise data science techniques for analysis to best identify environmental factors that influence wheat production and optimise the yield through informing breeding programs and the environmental conditions of farms. The intended outcomes are as follows:

* An integrated dataset that combines plot-level, simulation, and weather data from the National Variety Trials, providing a comprehensive foundation for subsequent analysis
* A cultivar performance analysis that identifies top-performing wheat varieties based on yield stability and climate resilience
* A frost damage prediction model that uses machine learning to detect and forecast frost damage risk for early intervention and management
* An environmental suitability guideline that quantifies the influence of key weather and soil factors
* An image segmentation model capable of automatically identifying and classifying images of wheat which differentiates its physical characteristics
* Actionable recommendations for breeding programs and farm management practices to optimise yield and mitigate the effects of adverse weather patterns

These objectives will be considered in communicating a solution that will be useful in many scenarios, ensuring practical relevance in decision making regarding wheat production.

# Solution

The report’s solution addresses the identified problems and intends to achieve all the objectives. The implementation will go through the data science process which is the problem formulation, data cleaning and merging, exploratory data analysis (EDA) and model selection and training.

## Problem Formulation

The project’s design is to infer the effects of climate variability and extreme weather patterns that threaten future wheat production in Australia. To accomplish this core objective, the formulation for the problem is to develop a framework that identifies which wheat varieties demonstrate larger yield performance under various weather conditions. This will be implemented using machine learning techniques where the algorithms used in current literature are neural networks, random forests, multiple linear regression model, and clustering based on temperature, soil type, rainfall, crop information and soil maps as the top 5 features used in over 50 primary publications. (van Klompenburg, Kassahun, & Catal, 2020) In addition, the group features most used relates to soil information, solar information, humidity, and nutrients. To answer the question of identifying the cultivar with the highest return on yield against the effects of weather, the problem shifts to regression of yield based on traits and classification to distinguish between climate resilient and non-resilient crops. Moreover, the assumptions required to create a reliable and accurate result is based on the historical data being a good proxy for future climate projections, the trial data is reliable and trustworthy in its usage and is identical and independently distributed as random across similar climate environments.

The next project being undertaken is identifying the risks or conditions of a crop’s attributes inputted into a machine learning model for frost damage. The problem will be formulated as a supervised classification task where the inputs may include weather and soil metadata, growth stage, cultivar, and attributes related to the crop’s trial. The possible models for this task could be tree-based, logistic regression, linear discriminant analysis (LDA) or support vector machines. The output of the chosen model is an ordinal label consisting of the levels of frost damage from the least severe, 1, to the most severe, 4. However, this implementation assumes that the frost damage records are reliable, that the crop and environmental data are consistent with each other, and the explanatory variables are able to explain most of the variability and that each of the four classes are balanced. The main goal is interpretability for any farmers or agronomists to be able to understand and ensure actionable steps are taken with the output of the 4 classes of frost damage.

The last project examines the best environmental attributes for wheat production. The project will be formulated as a supervised regression problem where the objective is to model wheat yield based on environmental predictors such as weather and soil attributes. Furthermore, data science techniques useful in ranking features based on importance in explaining the variance of the results the most will be used to identify the most significant features to consider when dealing with future climate change events. The possible techniques that are being considered are principal component analysis (PCA) and LASSO regression for feature ranking and selection respectively. The techniques are fairly similar with studies confirming a more than 50% similarity in selected features using both methods on socioeconomic, environmental, and image processing datasets against multiple linear regression and artificial neural network models. (Rahmat et al., 2023) Further advantages as to why these techniques are needed are demonstrated by the reduction in dimensionality for ease in processing and drilling down on significant features which can explain the variance and trends in the data which is clearer for interpretability. The assumptions required to use PCA is linearity between the yield and predictors, independence across observations, and a large variance for PCA. Moreover, the assumptions required for LASSO regression is linearity, low multicollinearity, normality of errors and sparsity of the true model to utilises the feature selection ability. (Huang & Deng, 2021)

The report aims for above an 80% success rate in determining the level of frost damage and using the key environmental attributes found to predict the amount of yield that results in the greatest return.

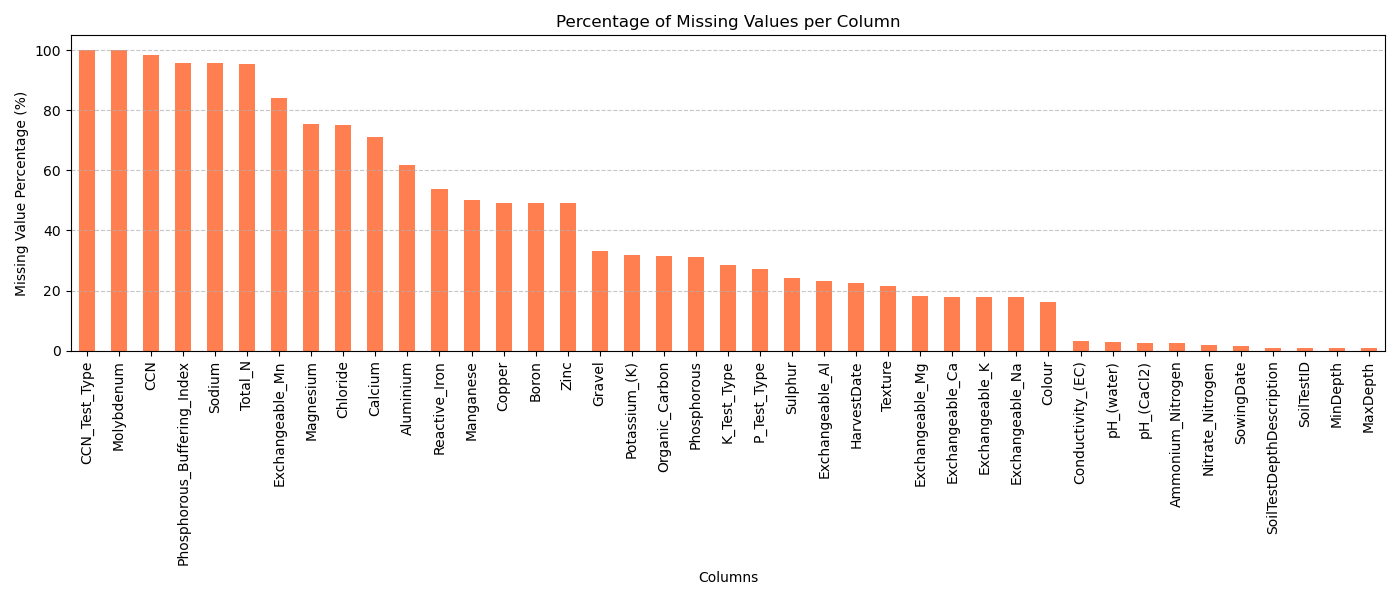
The reasoning behind all three projects answers the report’s question of which cultivar gives the greatest return in yield in a climate change scenario. In these scenarios, frost damage is frequent and can create a large deficit in yield if unaccounted for which directly conflicts with finding a greater yield possible with current breeds and delving into the most favourable conditions for further guidance in informing farmers and agronomists, the optimal conditions to be designed against climate change. These three projects provide a pipeline for identifying climate-resilient varieties, mitigating the risk involved in breeding them and ensuring an environment conducive to greater expected yield.

## Data Cleaning and Merging

The datasets involved include ‘soil and weather metadata’, ‘2015-2019 main plot data’, ‘2015-2020 main trial layout data’, ‘2020 main wheat data’, ‘2020 NVT Wheat Plot Level Measurements’, ‘weather data’, and ‘APSIM daily data’. Table 1 showcases the primary layout and details for each dataset:

|  |  |  |
| --- | --- | --- |
| Dataset Name | File Name | Format |
| Soil and Weather Metadata – Soil Data | *2015-2020 NVT Soil and Weather metadata sg 01.10.2021\_anonymized.xlsx* | 3031 rows x 52 columns |
| Soil and Weather Metadata – Rainfall Data | *2015-2020 NVT Soil and Weather metadata sg 01.10.2021\_anonymized.xlsx* | 2170 rows x 20 columns |
| 2015-2019 Main Plot Data | *2015-2019 Wheat Main Plot Data (All traits)\_anonymized.xlsx* | 99499 rows x 64 columns |
| 2015-2020 Main Trial Layout Data | *2015-2020 Wheat Main Trial Layout\_anonymized.xlsx* | 122323 rows x 16 columns |
| 2020 Main Wheat Data | *2020 Main Wheat - kg per plot, harvest length and width - cultivarID added\_anonymized.xlsx* | 19038 rows x 19 columns |
| 2020 NVT Wheat Plot Level Measurements | *2020 NVT Wheat Plot Level Measurements 26.03.2021\_anonymized.xlsx* | 26451 rows x 55 columns |
| Weather Datasets | *trial\_716\_SILOweather.csv* to *trial\_1684\_SILOweather.csv* | 517 rows x 17 columns |
| Simulated Dataset | *apsim\_out\_daily\_anonymized.csv* | 145387 rows x 41 columns |

*Table 1: Initial Format of Datasets*

Using the Pandas library, the soil data from the metadata datasets showcased the missing columns graphed in figure 4. 

*Figure 4: Percentage of missing values per column for the soil metadata dataset*

Using a missing value threshold of 50%, the attributes from ‘Manganese’ to ‘MaxDepth’ are able to be used where imputation based on a linear regression model will be used and all columns with missing values greater than the threshold will be dropped. The reasoning for using linear regression is that it utilises the relationship between the variables and preserves the multivariate structure which is much more reliable than using the mean for imputation. (Spanos, 1986) Moreover, mode imputations is used to impute categorical columns. Unfortunately, a major weakness of the NVT datasets given is the amount of data needing to be removed which may prove fruitful for the report’s objectives. The graphs for the rest of the missing columns are under the appendix.

The rest of the datasets are undergoing the same data cleaning process. Firstly, the rainfall dataset only had zero columns with over 50% missing values with the only column close being ‘Name’ at 47%. Secondly, the 2015-2019 main plot dataset which originally had 64 columns now has 23 columns and was the sparsest dataset. This dataset included frost damage as a variable which is crucial for the report’s objectives and 4 csv files were created by extracting the rows of frost damage of differing levels. Thirdly, the 2015-2020 main trial layout data had zero columns that exceeded the 50% threshold and the column with the highest percentage of missing values was ‘RowRep’ at 1.678%. Fourthly, the 2020 main wheat data had zero columns dropped with the largest number of missing values for a column being 30.6%. The fifth dataset is the 2020 NVT wheat dataset which now has 19 columns compared to the previous 55. The weather datasets will not be cleaned due to how numerous the number of files is, and it is inefficient to clean a trial-level dataset. The weather datasets will use summary statistics to merge to the corresponding trial code in the main dataset. Lastly, the simulated dataset had zero missing values.

After finishing the data cleaning process, the various cleaned datasets need to be merged into one main dataset to conduct exploratory data analysis (EDA) across different domains to ensure meeting the report’s purpose. The first step in the merging process is to merge all the plot level data together on ‘TrialCode’ and ‘CultivarID’ to create a dataset consisting of all plot level data. The next step is joining the rainfall data based on ‘RegionName’ and ‘Year’ and soil data based on ‘Year’ and ‘TrialCode’. The final dataset has a shape of 482834 rows with 65 columns and is seen in figure 4.

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) needs to be performed to gain an initial understand of the cleaned and merged datasets to visualise the trends of cultivar, frost damage, and climate variability. Firstly, the trends of cultivar and yield is a direct insight into this report’s purpose. Figure 6 represents the relationship between each cultivar and their respective yield in kg/plot.

A graph with blue squares

AI-generated content may be incorrect.

*Figure 6: Cultivar vs. Sum of Kg/Plot*

From figure 6, the top 5 cultivar varieties are 7ec49d23, 60934168, 78d8adec, 838e8df2, and 77b88288. These cultivar varieties need to be further analysed based on the environment and region to distinguish whether the yield’s performance is based on a specific ‘CultivarID’ or sharing the same region. Figure 7 showcases the top 10 cultivar varieties with the greatest yield with regional context.

A screenshot of a graph

AI-generated content may be incorrect.

*Figure 7: graphs of yield for top 10 cultivars against CultivarID, RegionName, and MET\_Analysis\_Mega\_Region*

The graphs reveal that most of the yield of the 10 cultivars are from Agzone2, Mallee, S/W, Agzone4, and S/E. This can be aggregated as the western region having a much greater impact on yield as seen in the pie chart. The top 10 varieties total 14.76% of the total yield with nearly half of the yield from the western analysis mega region. Furthermore, the total yield is also affected by the sowing date and harvest date as one of the few factors controlled by farmers which is represented in figure 8 which graphs ‘CultivarID’ against both dates.

A screenshot of a graph

AI-generated content may be incorrect.

*Figure 8: CultivarID, sowing date and harvest date vs yield of top 20 cultivars*

The first 20 cultivar varieties all are primarily sowed in May in the middle graph and the top graph showcases the harvest date which primarily in November and December. This is the usual practice in Australian agriculture and hence this can be controlled by the farmers and is not inherently based on uncontrollable environmental factors. However, the cultivar varieties that rank the highest in yield can be affected by soil as a key factor which is more damaging and can be changed based on soil type for a more productive environment. Figure 9 demonstrates 6 environmental factors against yield which are manganese, pH of CaCl2, copper, organic carbon, pH of water, and boron.

A screenshot of a graph

AI-generated content may be incorrect.

*Figure 9: soil attributes vs. yield for top 20 cultivars*

Focusing on the first 20 cultivars, the manganese is much more centred just above 0 which is where it peaks the most. The highlighted segments corresponding to the top 20 cultivars for manganese are situated near the middle in a highly dense part of the graph which may reveal the optimal manganese levels to be above 0 at around 18. Moreover, organic carbon, carbon and boron also have similar patterns where the highlighted sections corresponding to the top 20 cultivars are close to the middle of the peak in what can be assumed to be a normal distribution or skewed for both manganese and organic carbon. The optimal level for organic carbon is 0.98 based on the peak. In addition, the optimal amount of copper is 1.05 and the optimal amount of boron is 2.4. For the pH levels, they are bimodal graphs with the corresponding data points of the 20 cultivars following the same overall pattern which is close around the peaks and lower in the middle. Moreover, it seems likely that the peaks indicate a non-linear relationship where different pH levels are optimal. The last key factor being analysed is frost damage which is also affected by environmental and weather conditions. Figure 10 showcases frost damage across different regions.

A chart of different colored circles

AI-generated content may be incorrect.

*Figure 10: Frost Damage vs MET analysis mega region*

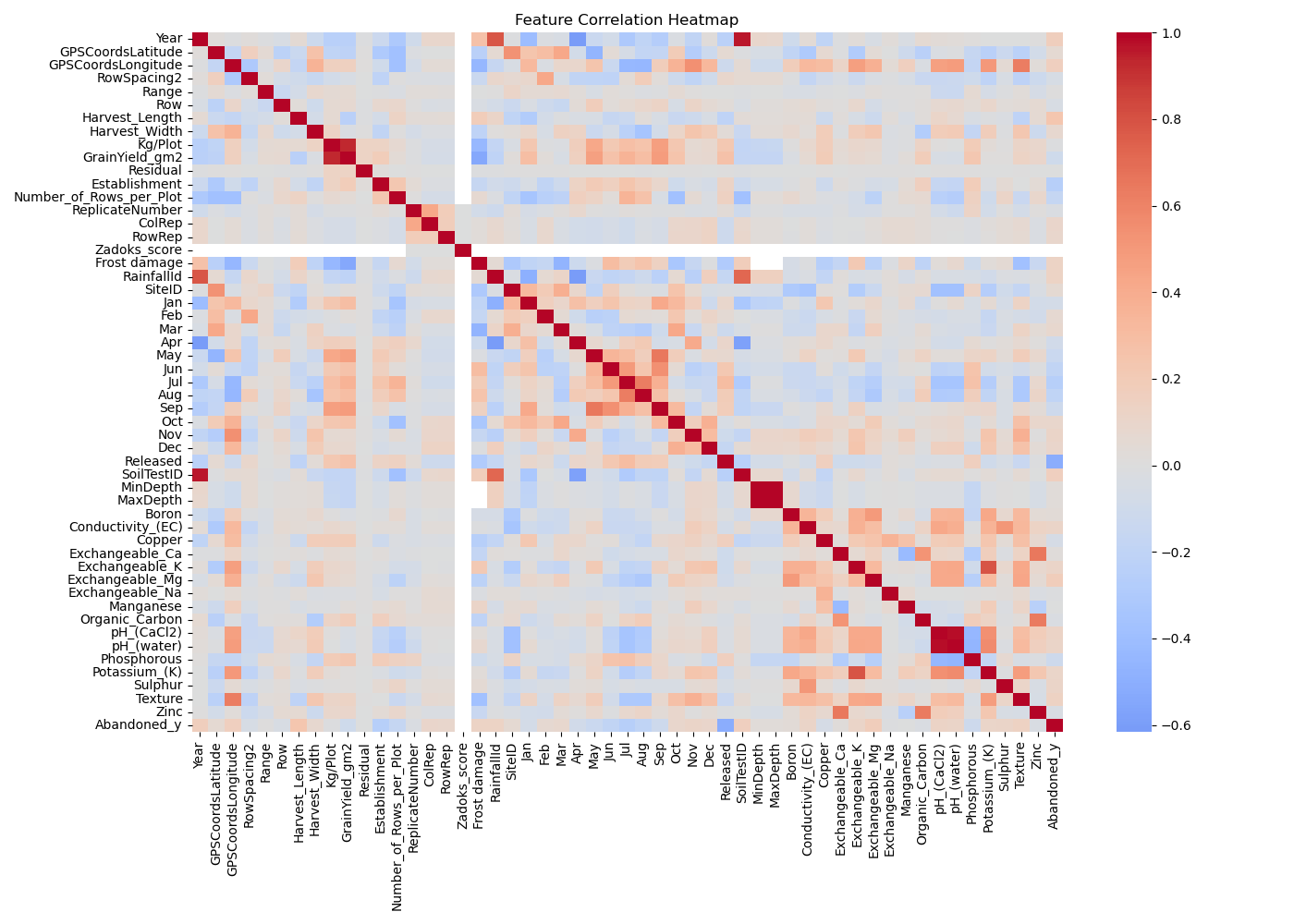
These graphs are not very consistent where frost damage level 1 and frost damage level 4 have southern regions as having the count of crops affected by frost. However, levels 2 and 3 of frost damage have Eastern as the largest count. This may be due to regional differences where southern regions experience more extreme cold wave and eastern regions are much more moderate which leads to frost damage of levels 2 and 3. Frost damage level 1 has all regions nearly the same which can mean frost damage of level 1 is frequent in all regions regardless of climate differences due to being the least severe case. On the other hand, the effects of frost damage and yield can be seen in figure 11 where all levels of frost damage has the shape of a normal distribution. This means that the peaks of the frost damage are near the lower values with lower yield being a result of low frost damage and higher yield also being a result of low frost damage with a certain middle value that is susceptible to frost damage. The peaks are at around 2 to 4 kg/plot of yield with yield that is higher being a result of lower frost damage as it tapers lower down to the end of the graph.

A graph of blue lines

AI-generated content may be incorrect.

*Figure 11: Yield vs levels of frost damage*

Figure 12 showcases the heatmap of the merged dataset which only has 8 correlations between two different features that are greater than 0.8. A more detailed description of the correlation between features are presented in table 2.



*Figure 12: Feature correlation heatmap of merged dataset*

|  |  |  |
| --- | --- | --- |
| Feature 1 | Feature 2 | Correlation |
| MinDepth | MaxDepth | 0.999490 |
| pH\_(CaCl2) | pH\_(water) | 0.975181 |
| Year | SoilTestID | 0.967899 |
| Kg/Plot | GrainYield\_gm2 | 0.926703 |
| Exchangeable\_K | Potassium\_(K) | 0.790980 |
| Year | RainfallID | 0.783657 |
| RainfallID | SoilTestID | 0.721524 |
| Exchangeable\_Ca | Zinc | 0.651306 |
| May | Sep | 0.646771 |
| Organic\_Carbon | Zinc | 0.637413 |

*Table 2: Top 10 strongest correlations between features in merged dataset*

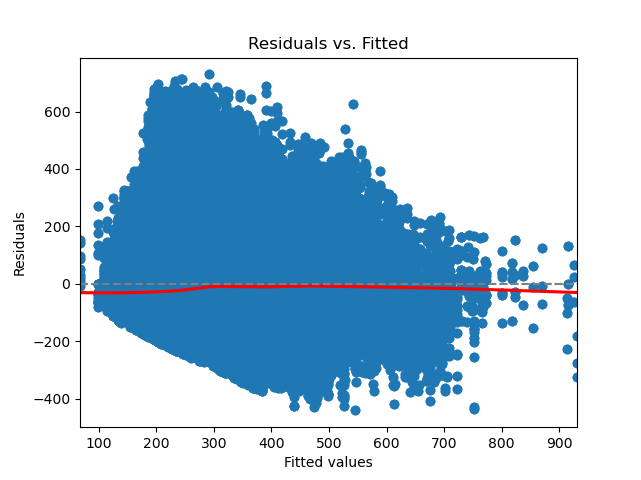
From table 2, many of the correlations between two features are redundant such as Kg/Plot with GrainYield\_gm2 since they are positively correlated with each other where one feature will, by definition, increase the other. Exchangeable\_K and Potassium\_(K) is also redundant by definition. The most interesting feature correlations are between pH\_(CaCl2) and pH\_(water) which entails a measure of soil acidity since they are both similar in pH levels. Another feature of the soil’s chemistry is calcium and zinc being correlated which shows more nutrient-rich soils have high levels of both alongside carbon and zinc which also supports the claims of having higher quality soils involving these elements. The other feature correlations can be explained by the experimental conditions such as MinDepth and MaxDepth being set for each trial as well as the IDs of different features being set by the experimenters.

## Model Selection and Training

The model selection and training section goes over each objective which includes testing any assumptions that were stated in the problem formulation section, going through the training process and evaluating the models for generalisability and functionality towards the chosen objective.

### Discovering the Most Climate-Resistant Cultivar

The first project in aiding understanding of climate change on wheat production is ranking the cultivars based on their respective yield with respect to environmental variables such as soil and rainfall. To achieve this, multiple models and techniques were considered such as multiple linear regression, random forest regression, and clustering methods. Firstly, the assumptions when using multiple linear regression include linearity which in figure 13 is proven valid from the graphed residual plot.

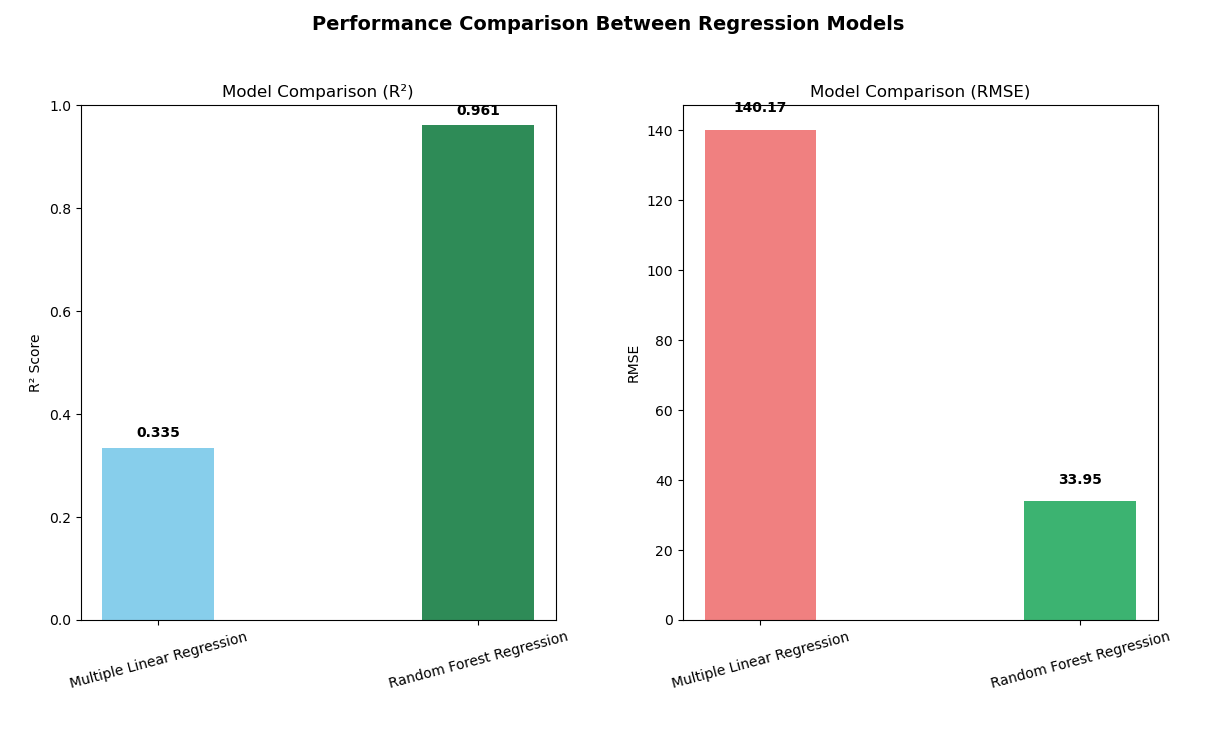


*Figure 13: Residual plot of multiple linear regression*

Furthermore, another assumption required is testing for multicollinearity which is significant since the regression model’s coefficients for one environmental variable increasing or decreasing could affect another variable which reduces interpretability in explaining which predictor directly affects the crop yield. (Qu, 2024) This can be examined using variance inflation factors (VIF) with the most common threshold being VIFs greater than 4 require investigation and anything above 10 require serious correction. (Pennsylvania State University, n.d.) Appendix 7 contains all the VIFs for each environmental variable with the major variables with values over the threshold of 4 being ‘MinDepth’ at 945.36, ‘MaxDepth’ at 945.30, ‘pH\_(CaCl2)’ at 24.02, and ‘pH\_(water)’ at 25.12. Hence, these variables will be dropped in the multiple linear regression model for the model to be able to be considered and deliver reliable results in the project. The preliminary model using only ‘TotalRainfall’ and ‘CultivarID’ only managed to get an R-squared value of 0.35 which is quite poor and jumps to 0.94 when adding the rest of the variables unaffected by multicollinearity. However, this may be skewed due to adding ‘TrialCode' which decreases to 0.43 and needs to be finetuned in the actual implementation. The second model considered is random forest regression. The assumptions needed are much more relaxed compared to the linear regression which needs identical and independently distributed data and informative features. The advantages relevant to the report’s objectives are capturing feature importance which is related to identifying which environmental factors the cultivar is most resistant to and captures complex non-linear interactions such as one variable positively affecting one cultivar and not the other. (AIML.com, 2025) The implementation of the random forest regression model uses some feature engineering such as calculating total rainfall and using one-hot encoding for ‘CultivarID’. The feature set consists of environmental variables which were similarly used for the previous linear regression model. I would then train this set on a random forest regressor model and evaluate the performance using R-squared and RMSE metrics. For assessing whether the cultivar is resistant to a wide range of variations in environmental factors, I would generate yield predictions for all the given data points and group the results by cultivar and calculate its coefficient of variation which identifies the stability to weather or dispersion to the mean under different conditions. (ScienceDirect, n.d.)​ Lastly, clustering is an unsupervised learning technique which will focus less on yield and more on grouping cultivars with similar environmental properties. I will use either K-Means or hierarchical clustering based on model evaluation and analyse the resulting clusters for cultivars that produce high-yield in diverse environments, showcasing stability or tolerance in different environments. For each cluster, the average yield and its variance can also be calculated for each cluster to identify entire clusters as climate resistant. Moreover, dimensionality reduction techniques such as PCA can be used for visualisation by plotting each cluster or an average of each cluster’s yields for comparison between different cluster’s yields and different environments. The evaluation metric to be used is Davies-Bouldin Index to quantify the similarities of cultivars within a cluster compared to other clusters. (GeeksforGeeks, 2023)

From the above preliminary exploration of the possible solutions to the first objective’s problem of finding the most climate-resistant cultivar, the next section details the implementation and experimentation section. This required creating a pipeline capable of not only predicting the yield accurately but also assessing the stability of the cultivar and the environmental adaptability. The final code in the appendix undergoes data exploration, model experimentation, and statistical validation. To ensure reproducibility and scalability when testing the merged dataset with the codebase’s machine learning models, a Pandas-based ingestion process was established with a test mode to sample the full dataset on the cluster for troubleshooting errors and ensuring performance metrics were behaving suitably for the full dataset.

The preprocessing pipeline automatically detected categorical and numerical columns and allowed different transformations each such as median imputation and standardisation for numeric data, and frequency-based imputation with one-hot encoding for categorical data. These steps ensured consistency across both linear and non-linear models while avoiding data leakage. To understand the relationships between environmental features and yield, the first approach employed was multiple linear regression. Linear regression was chosen for its interpretability and the ability to confirm the previous assumptions of linearity and multicollinearity. Using residual plots and variance inflation factors, several predictors like MinDepth, MaxDepth, and pH\_(water). These variables were removed to avoid unstable coefficients and improved the model’s interpretability. However, after all this, the explanatory power remained limited with an R-squared value of 0.335 and a RMSE of 140.167. From this poor performance, the next approach to transition to was towards Random Forest Regression. This relaxed the assumptions of linearity and independence. The random forest, unlike multiple linear regression, could handle complex, non-linear interactions between environmental features and the yield. Furthermore, the feature importance scores provided valuable insight into which environmental variables had the greatest influence on the variation of yield. Through fine-tuning hyperparameters such as tree depth and number of estimators, the model achieved a very high performance with an R-squared value of 0.961 and a RMSE of 33.951. Figure 14 shows the comparison between these two models.



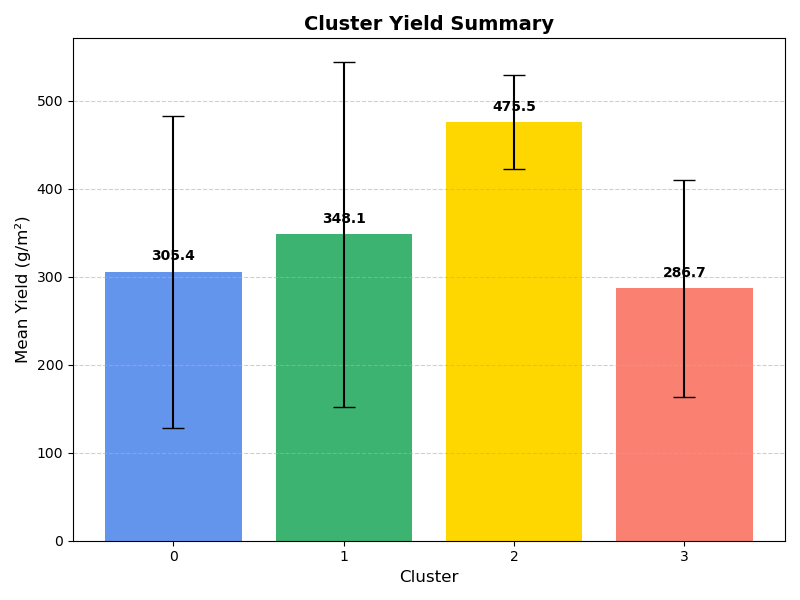
*Figure 14: Performance comparison between multiple linear regression and random forest regression*

After finding the model with the best performance, the focus can now be shifted towards interpretability of the model’s output and cultivar evaluation. The feature importance analysis derived from the random forest proved a ranked list of the most influential environmental variables. With these predictions, cultivar stability was assessed using the coefficient of variation (CV) which is a normalised measure that quantifies yield stability relative to the mean. Cultivars with low CV values were interpreted as more stable which suggested that their performance was less affected by environmental variability. Table 3 below shows the top 10 most stable (climate-resistant) cultivars.

|  |  |  |  |
| --- | --- | --- | --- |
| CultivarID | Mean yield | Standard Deviation of Yield | Coefficient of Variation |
| b2536bb8 | 837.2851 | 24.9747 | 0.0298 |
| 9426cdfe | 737.5918 | 44.7849 | 0.0607 |
| 2478f263 | 779.8819 | 53.6176 | 0.0687 |
| a45a1d12 | 753.2708 | 52.1044 | 0.0692 |
| 8ce6fc70 | 726.9733 | 51.5902 | 0.0710 |
| a80bd9bf | 389.5569 | 29.2164 | 0.0750 |
| 4f1730a9 | 796.7332 | 63.4267 | 0.0796 |
| 03271d33 | 701.6202 | 55.9298 | 0.0797 |
| 70876df6 | 722.3801 | 59.0005 | 0.0817 |
| 315b4df9 | 776.1679 | 65.7667 | 0.0847 |

*Table 3: Top 10 most stable (climate-resistant) cultivars*

To complement the prediction’s analysis, unsupervised learning techniques were introduced to investigate the environmental groupings that correspond to different soils or climates. K-means clustering was used to segment trials into four clusters based on similarity in environmental variables. Aggregating yield statistics within each cluster allowed comparison of mean and standard deviation of yields, helping identify which environmental variables foster the most stable yield production. Figure 15 shows the cluster yield summary.



*Figure 15: Cluster yield summary of dataset and their corresponding mean and std. yields*

The next project details modelling the effects of climate variability in modelling frost damage.

### Modelling Frost Damage

**Project 2: Modelling frost damage**

The second project in the report is discovering for a given inputted crop, the level of frost damage to mitigate risks against climate change for farmers and agronomists. Before implementing the project, the cleaned dataset was searched and for each class of frost damage – from 1 to 4 – a separate csv file was created. The models trained on these datasets considered are logistic regression, XGBoost, and neural networks. Firstly, logistic regression requires the assumption that there is a linear relationship between the logits and each predictor variable, no multicollinearity which was proven in project 1, and independence of observations. Figure 16 showcases the linearity of the traits of ‘Kg/Plot’ against ‘Frost damage’ where a higher amount of yield results in class 1 and 4 having a lower chance of frost damage as opposed to classes 2 and 3. In addition, class 4 has a much lower first predictor value than the other 3 classes which means crops with a frost damage of level 4 decreases in yield much more quickly for a lower amount of yield. This interpretation makes sense as crops with more frost damage implies lower yield. However, classes 2 and 3 are actually positive which requires further investigation. Furthermore, extensions to be made will incorporate more environmental variables such as rainfall and soil data which will then be constructed into a feature set if they satisfy the linearity assumption. To implement a logistic regression model, I will combine all 4 csv files and split the dataset into 70% for training and 30% for testing and evaluate the models using accuracy and visualise performance using a confusion matrix.

A graph of a graph

AI-generated content may be incorrect.

*Figure 16: logits vs predictor for Kg/Plot vs Frost Damage*

The second model to be considered is XGBoost which will be similarly implemented as the logistic regression model but will use a 60% training, 20% validation, and 20% testing split. The validation set will be used to finetune the learning rate, maximum depth and number of estimators. To evaluate this model I will use accuracy and a confusion matrix for visualisation of the performance. Furthermore, since XGBoost is a collection of trees, the model provides feature importance plots to further determine which environmental variables affect frost damage the most for interpretability to farmers and agronomists. The advantages of this model outweigh the logistic regression implementation due to the performance benefits, supporting regularisation and high accuracy. However, the features of the model to account for are the many hyperparameters which may complicate the process and requires evaluating many combinations of XGBoost models through cross-validation. Furthermore, there is a significant risk of overfitting based on the number of estimators and learning rate which also need to be carefully evaluated by monitoring learning curves and interpreting the training vs test and validation accuracy. Furthermore, plotting the ROC curve of the model can also determine overfitting and visualise the performance. (Chen & Guestrin, 2016)

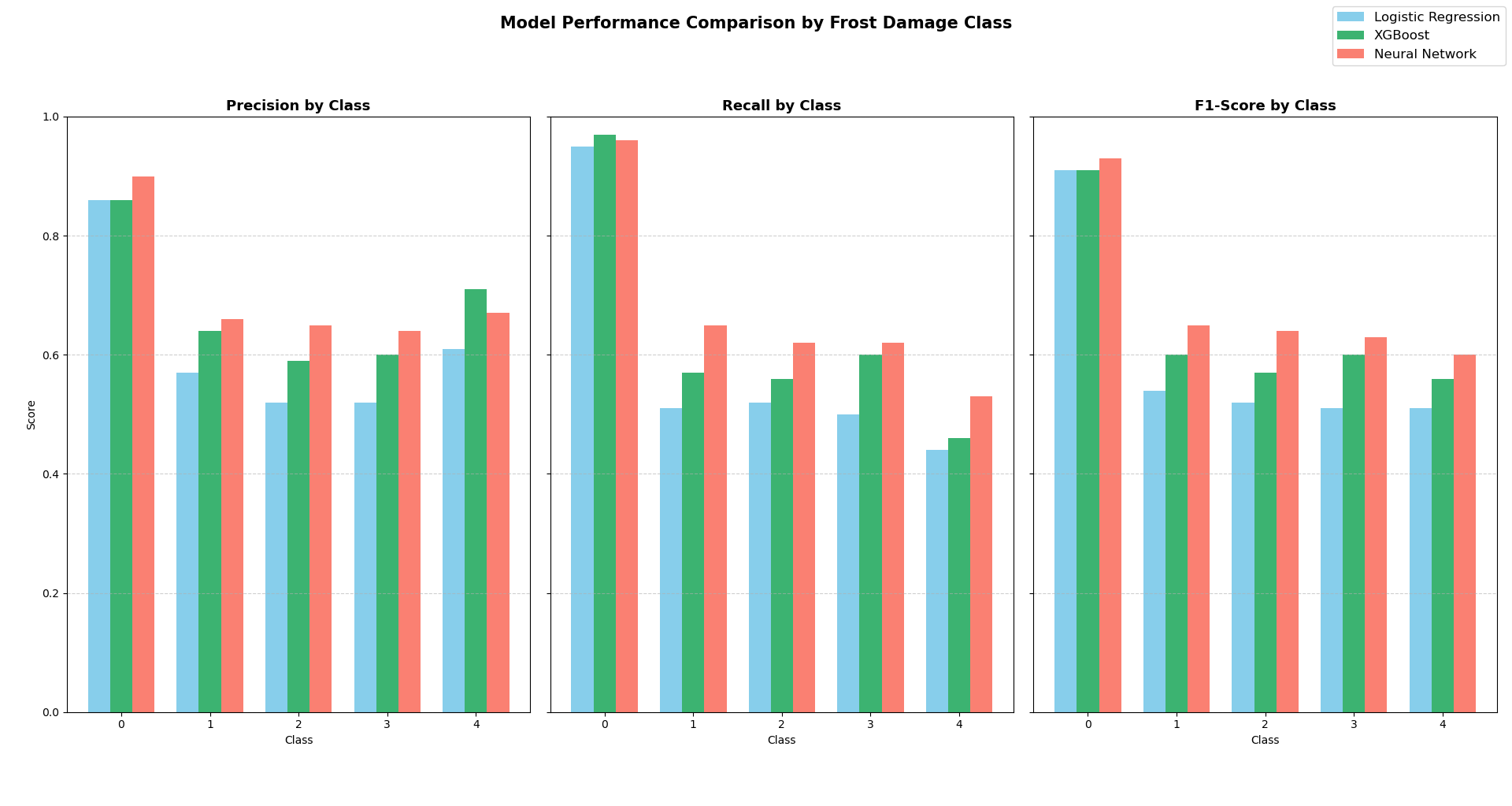
The third model in determining the level of frost damage on an inputted set of features is using a neural network. To implement this model, I would define a simple neural network architecture with one or two hidden layers. Grid search will be used in arriving at the most effective number of layers and neurons for performance by choosing the hyperparameters that result in the least loss. Moreover, a softmax activation function will be used at the end of the model to output the class probabilities for each level of frost damage. The model will also have to iteratively choose between different loss functions starting with categorical cross-entropy loss as the use case is specifically for classification problems with more than two classes. (Yathish, 2022)

From the above preliminary exploration of the possible solutions to the second objective’s problem of modelling different levels of severity of frost damage across multiple wheat trials, the next section details the implementation and experimentation section. This prediction from the machine learning models being implemented could aid in decision making regarding the treatment of crops. The earliest challenge was the massive amounts of data incompleteness for the frost damage column and so the first 4 levels of frost damage are considered due to the high prevalence rather than using frost damage level 9 which was encountered in 1% of the dataset’s rows. The feature encoding technique used was one-hot encoding which was applied to all categorical fields for numerical representation within the models. After the encoding and feature standardisation, the dataset was split into training, validation, and testing datasets. For models that relied on hyperparameter optimisation or early stopping such as XGBoost and the neural network, the split employed was 60% for training, 20% for validation, and 20% for testing. For logistic regression, a 70% training set and 30% testing set split was used. For both approaches, the splits were stratified by proportion of each frost damage severity label to ensure balanced representation among all classes since they were imbalanced as seen in figure 17.



*Figure 17: Proportion of frost damage levels*

To further mitigate this class imbalance, class weights were computed and incorporated into the model training ensuring that misclassifications of rare classes had higher penalties. For logistic regression, these weights were passed into the ‘class\_weight’ parameter. For the other models, a weighted cross-entropy loss was used, where each class’s contribution to the overall loss was scaled according to its proportion in the dataset. The first model was the multinomial logistic regression which acted as a baseline model due to its simplicity and interpretability. This was trained using the ‘saga’ solver and handles both L1 and L2 regularisation. The model was also configured to run up to 2000 iterations to ensure convergence given the high dimensionality after encoding. The results showed an overall accuracy of 0.69 and a weighted average of 0.68. The best class that was classified properly was class 0 or no frost damage which had an f1-score of 0.91. The next model was the XGBoost which needed tuning of the hyperparameters, including ‘eta’ and ‘max\_depth’. An eta or learning rate of 0.05 was used for convergence speed while a maximum depth of 6 was used to mitigate overfitting. The model also used early stopping as another deterrent for overfitting. This model did much better with an overall accuracy of 0.73 and a weighted average of 0.72. The class that had the greatest classification performance was class 0 or no frost damage which an f1-score of 0.91 which was the same as the logistic regression model. The final model evaluated was the feedforward neural network which three fully connected dense layers as its architecture. The input layer accepted the scaled feature matrix, followed by hidden layers of 256 neurons and then 128 for the rest. Each hidden layer had batch normalisation and ReLU activation functions to model the nonlinear relationship better. Dropout regularisation was set at 30% to mitigate overfitting. Moreover, the Adam optimiser was used and the learning rate was set to 0.001 and was trained for 30 epochs using mini-batch gradient descent with a batch size of 32. Each epoch displayed significant improvement with the loss decreasing steadily. The performance plateaued at epoch 20 and the training was terminated to prevent overfitting. This had the best performance with an accuracy of 0.76 and a weighted average of 0.76. The class that had the greatest performance was class 0 or no frost damage with an f1-score of 0.93. Figure 18 graphs the comparison of the model’s performance.



*Figure 18: Model performance comparison by frost damage class*

From the above figure, the neural network model had a higher f1-score for all the classes whereas, it was lower than the XGBoost’s performance on two occasions which were the precision of class 4 and the recall of class 0. Each of the model’s performance on class 0 was the highest which was completely reasonable due to the imbalanced nature of the dataset evidenced by figure 17. Regarding the frost damage classes, the f1-score showed that class 1 had the greatest performance overall but for precision and recall this was classes 4 and 3, respectively. The next project details determining the most productive environment for yield.

### Identifying the Most Productive Environment for Wheat Production

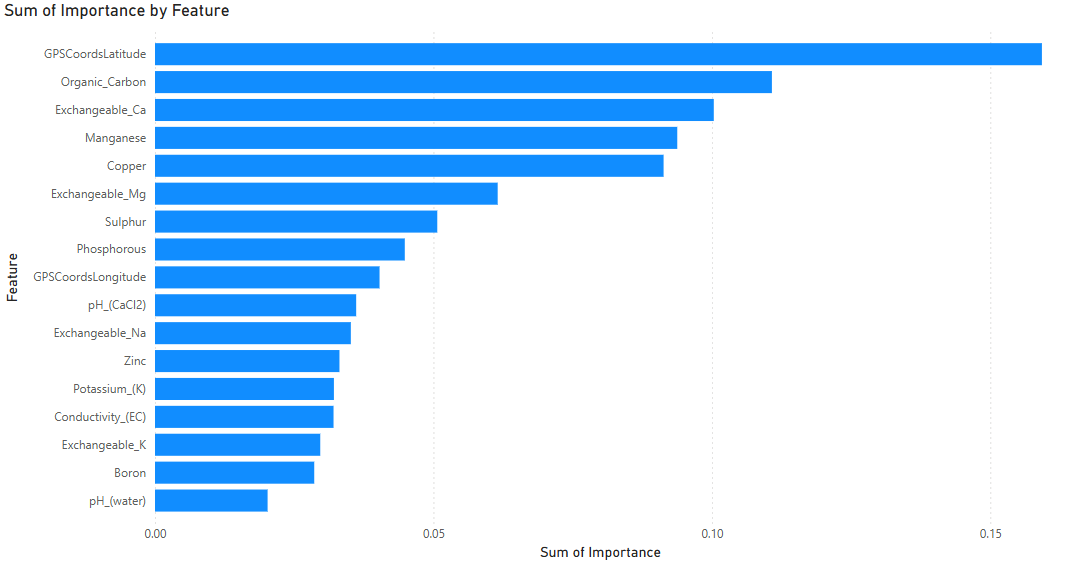
The third project entails identifying the features of the environment most conducive to wheat production. The models being considered for this feature selection and ranking task are random forest regression and lasso regression. Random forest regression is useful in this domain since it includes feature importance scores after training to identify the most important environmental features. The model’s training process will include only environmental variables as the explanatory variables of the model using an 80% training and 20% test split. Once the selected features are found, I will use a grid search using the selected features to iterate through every possible combination within the range to identify the combination that results in the highest yield. The range will be calculated by using the interquartile range and ensuring any values outside of the range are treated as outliers. In addition, constraint-based optimisation tools can be used to use the selected features as lists of constraints and find the optimal values based on a yield-based objective function.

The second model that can be evaluated is lasso regression which also prioritises feature selection of the environmental variables to find the optimal environment for wheat production. I would standardise the environmental variables and split the data into 60% training, 20% validation and 20% testing sets. This is to provide enough data for the regularisation parameter to balance model complexity. (Jain, 2025) Furthermore, lasso performs regression and simultaneously conducts feature selection by creating sparsity by reducing coefficients to zero. The coefficients not reduced are the selected features that have the greatest influence on the amount of yield a crop produces. To evaluate the model, I would use R-squared and mean squared error for interpretability and sensitivity to large errors respectively. After finding the significant features and evaluating the model, I would similarly use grid search and constraint-based optimisation tools to find the optimal environment for wheat production. The two models used both have their advantages and disadvantages, but both will be used to be able to combine interpretations.

The comparison between random forest regression and lasso regression is that lasso regression is much simpler in being able to interpret the model based on the non-zero coefficients and can be easily finetuned using cross-validation but assumes the model is linear which was proven for project 1. On the other hand, random forest regression also works with non-linear models and is less prone to overfitting, if using pruning techniques but the feature importance scores must be given through a function which loses its interpretability within the context of the model.

Lastly, each model will be compared to each other using the same evaluation metrics as every other model in their corresponding project. In the case of similar metrics for both models, ensemble methods will be used such as using hard voting, bagging, and boosting to identify the best cultivar varieties accurately and reliably, improve generalisations in having to predict frost damage, and to be able to identify the same features selected from different models which supports a greater influence on the environment.

From the above preliminary exploration of the possible solutions to the third objective’s problem of determining the most productive environment for yield, the next section details the implementation and experimentation section. The first model used was the Random Forest regression model which allowed modelling the yield from all the environmental variables simultaneously. This model was trained on 70% of the dataset and tested on the remaining 30%. The performance of this model regarding its R-squared value was quite high at 0.9304 with an RMSE of 45.4334. Furthermore, the Random Forest model provided an ordered ranking of features by importance which highlighted the variables that had the strongest influence on the yield. Figure 19 shows the top 10 features with the greatest importance computed by the Random Forest model.



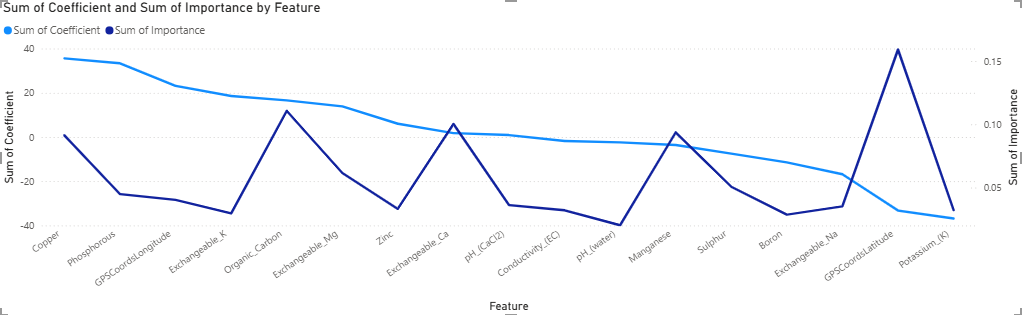
*Figure 19: top 10 features by importance from random forest regression*

Another model that was employed was Lasso regression to better refine the feature selection process. Lasso regression includes an L1 regularisation penalty that encourages sparsity in the coefficients as a form of feature selection to gather the most highly influential environmental variables on yield. Figure 20 shows the top 10 features with the greatest importance computed by Lasso regression and figure 21 shows the comparison between both feature importance lists of the two models.

A graph of a bar graph

AI-generated content may be incorrect.

*Figure 20: top 10 features by importance from lasso regression*



*Figure 21: comparison between feature importance of random forest and lasso regression*

Comparing both figures 19 and 20, the features chosen to be the most important are majorly different between the two approaches. The environmental features that are common in both are GPSCoordsLatitude, Organic\_Carbon, Copper, Exchangeable\_Mg, and GPSCoordsLongitude. From getting these most relevant features which seems to be based on geographic location and particular concentrations of elements found in the soil, the next step was to understand the natural variability of each feature for optimisation. Firstly, this was done by computing the interquartile ranges which provided a spread of the data for each selected feature. For instance, features with larger IQRs such as Phosphorous can indicate a higher potential for influence on yield, if adjusted within realistic bounds. This is crucial for grid search and constraint-based optimisation (inspired by linear programming) to search for the optimal values within these IQRs.

The first method used to identify optimal combinations of environmental features was a grid search over the IQRs of the selected features. The approach taken was to generate a fine grid of potential values for each feature and systematically evaluate all possible combinations by inputting them into the previously trained Random Forest model to predict yield. The final result is the grid search identified the combination of feature values that maximised the predicted yield. The second method implemented was the constraint-based method which used the L-BFGS-B algorithm. This method uses continuous optimisation rather than discrete points such as the previous method to adjust values iteratively to maximise the yield. This optimisation method complemented the grid search results by providing more precise values and confirming the general trends. Table 4 shows the optimal yield which is the most realistic and table 5 shows the best grid search results which is just greedily maximising all the features which is quite unrealistic.

*A screenshot of a graph

AI-generated content may be incorrect.*

*Table 4: Optimal Yield for top environmental features*

A screenshot of a graph

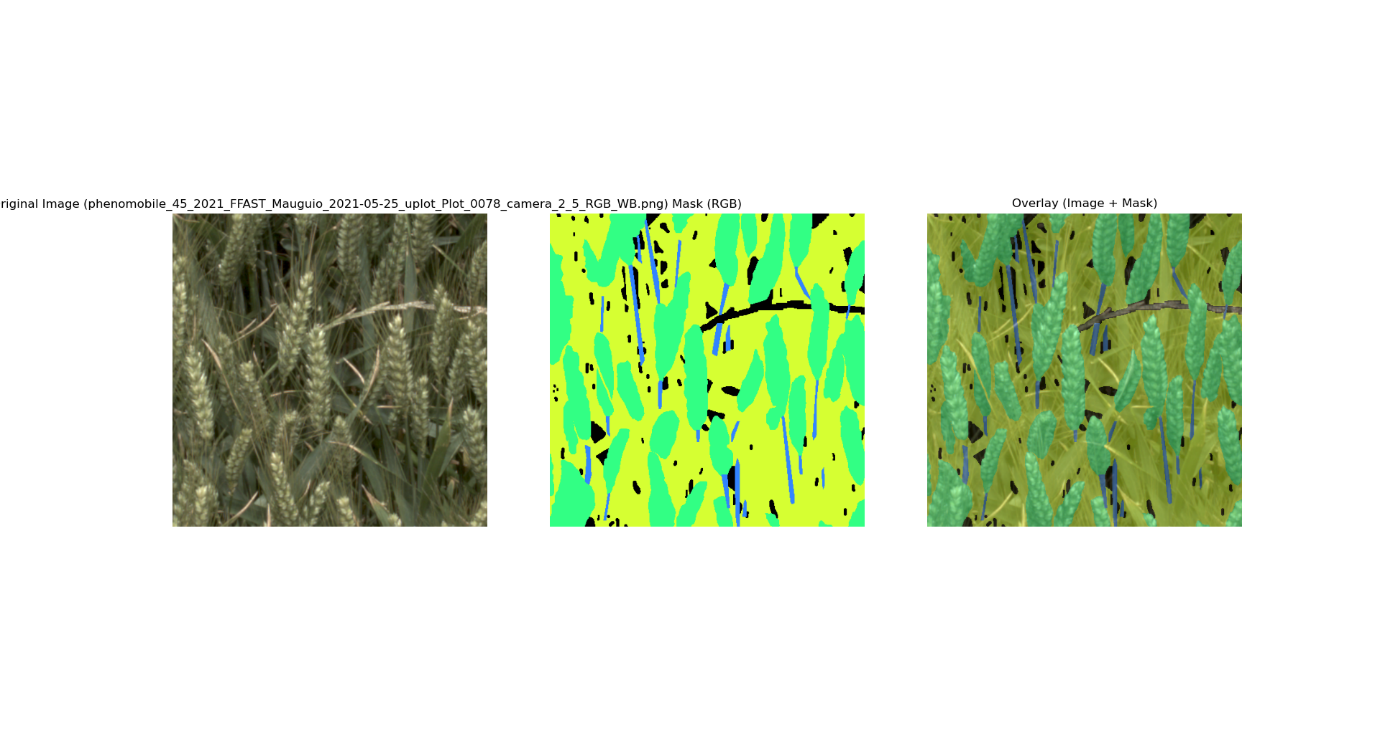
AI-generated content may be incorrect.

*Table 5: Grid Search Best Result for top environmental features*

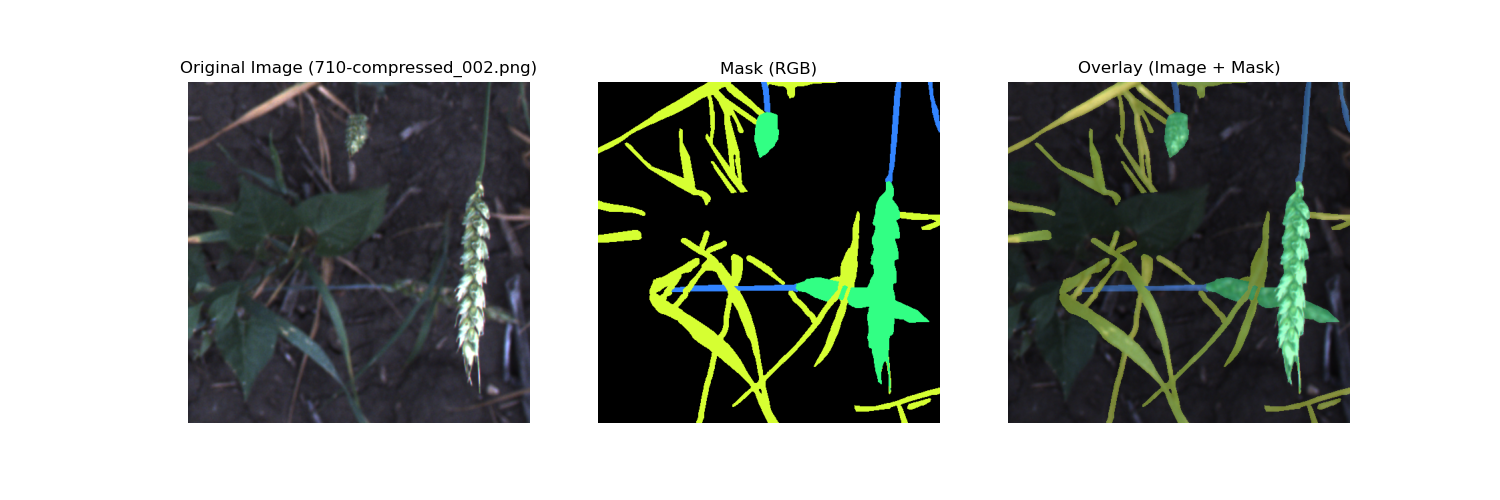
The next project details using computer vision techniques to conduct image segmentation on the wheat images dataset.

### Image Segmentation

This section details the solution to the image segmentation objective which was to segment the given wheat images dataset into four classes: background, head, stem and leaf. The exploratory data analysis conducted on this dataset shows image sizes of both 512x512 and 1024x1024 with all the images having an aspect ratio of 1.0. Figures 22 to 24 showcase examples of the image (stored in dataset/images) and the corresponding mask (stored in dataset/masks) with the corresponding overlay of the mask on the image.



*Figure 22: image, mask and overlay of phenomobile\_45\_2021\_FFAST\_...*

**

*Figure 23: image, mask and overlay of 710-compressed\_002.png*

*A close-up of a plant

AI-generated content may be incorrect.*

*Figure 24: image, mask and overlay of 0166\_17\_im1\_2021\_05\_11\_11\_41\_09\_L776.png*

The image and mask datasets were split into different folders comprising of the origin of the images which ranged from Japan (NARO & UTokyo), France (Arvalis), Canada, UK, Switzerland, China, and Australia. (David et al., 2020) To fully train all these images under a model, the dataset was chosen to remove the geographical and institutional subfolders and store all the images under one folder. The choice of analysing all the images into one dataset was to improve the overall generalisability of the model on an international level and to improve the model by utilising a larger dataset.

The models chosen were the DeepLabv3, Fully Convolutional Network (FCN), and Segformer models and was tested on the dataset without any data augmentation techniques, changing of the architecture, or manipulation of the loss function to serve as a baseline for evaluation. Firstly, DeepLabv3 loaded the images as PIL.Image and converted this to RGB format and used Torchvision’s transforms library to resize all the images to 256x256. The DeepLabv3 was chosen for its ability to use atrous convolutions and Atrous Spatial Pyramid Pooling to capture multi-scale contextual information. (Chen et al., 2017) The model was initialised with a pretrained ResNet-50 backbone and was modified to output only 4 classes. 5-fold cross validation was implemented with an Adam optimiser, cross entropy loss, a batch size of 4, and 20 epochs. This was all evaluated using Intersection over Union. Figure 25 shows the confusion matrix, figure 26 shows mIoU graph and figure 27 shows the predictions made.

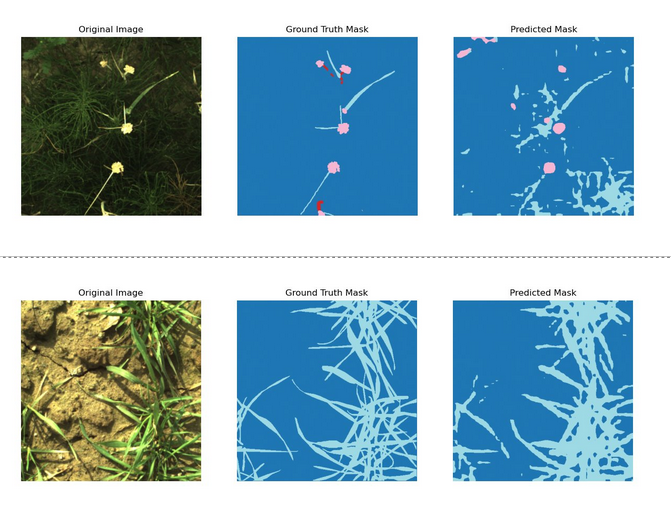


*Figure 25: DeepLabv3 Confusion Matrix*

A graph of blue rectangular bars

AI-generated content may be incorrect.

*Figure 26: DeepLabv3 IoU graph for each class*



*Figure 27: DeepLabv3 Image, Mask, and Predicted Mask*

The next baseline model tested was the FCN which loaded images and masks in the same way and converted the RGB to class IDs. Furthermore, the images and masks were both resized to 256x256 using the transforms library from Torchvision. The model chosen was the FCN-ResNet50 since it provided a fully convolutional architecture and was much more lightweight than the other models. Furthermore, this could be easily compared to DeepLabv3’s baseline model which used the same underlying architecture. Furthermore, I used 5-fold cross validation, Adam optimiser and cross entropy loss as before. Figure 28 shows the confusion matrix, figure 29 shows the IoU graph and figure 30 shows the predictions made.

A graph showing the difference between a number and a matrix

AI-generated content may be incorrect.

*Figure 28: FCN Confusion matrix*

A graph of a graph

AI-generated content may be incorrect.

*Figure 29: FCN IoU graph*

A collage of different images

AI-generated content may be incorrect.

*Figure 30: FCN Image, Mask, and Predicted Mask*

Lastly the baseline segformer model, unlike the two other models relying on convolutional layers uses Vision Transformers with hierarchical feature extraction. (Xie et al., 2021) This architecture make Segformer particularly suitable for high-resolution and class-imbalanced datasets which will be useful for the research question posed later in the section. The model chosen was the Segformer-B0 for a balance between efficiency and accuracy. The model was initialised using weights from the ADE20K dataset from NVIDIA for transfer learning on the wheat dataset. The metrics used here were the same as the other 2 baseline models and used class-wise IoU. Figure 31 shows the confusion matrix; figure 32 shows the IoU graph and figure 33 shows the predictions made.

A blue squares with numbers and numbers

AI-generated content may be incorrect.

*Figure 31: Segformer confusion matrix (classes: 0-background, 1-stem,2-head,3-leaf)*

A graph of a graph

AI-generated content may be incorrect.

*Figure 32: Segformer IoU graph*

*A collage of images of a mask

AI-generated content may be incorrect.*

*Figure 33: Segformer Image, Mask and Predicted Mask*

As can be seen in all three confusion matrices, the imbalanced dataset makes it difficult for all the models to learn the relationship for classifying the stem of the wheat since the number of pixels of the stem class are dominated by the number of pixels of the other three classes, especially background and leaf. Figure 34 shows the distribution of classes.

A graph of a number of blue rectangular objects

AI-generated content may be incorrect.

*Figure 34: Class-wise pixel frequency distribution*

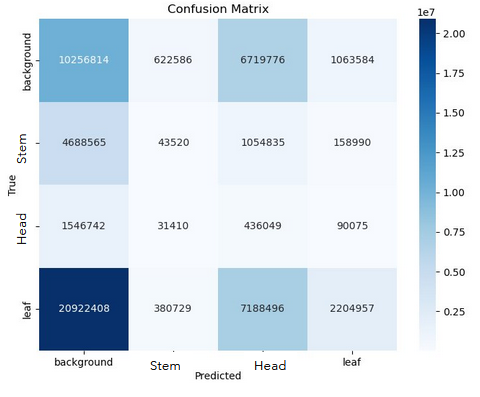
Thus, the research question posed is *what methods improve segmentation accuracy for small structures such as stems, wheat head tips or fine leaves*? For strategies to improve the classification of the stem class, two papers were used which required modifying the loss function and adding data augmentation techniques. The first paper, “calibrating imbalanced classifiers with focal loss: an empirical study” explains that models trained with focal loss tend to be better calibrated than those trained with just cross-entropy on imbalanced tasks. However, improving calibration can reduce some aspects of accuracy unless tuned well and there are precision/recall trade-offs with tuning hyperparameters and per-class calibration. (Wang et al., 2022). Moreover, the second paper, “unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation” goes through a medical image segmentation context with foreground elements such as organs much smaller and imbalanced relative to the background. Moreover, this approach unifies distribution-based loss (cross entropy) with region-based loss (dice) which measures region overlap and distribution losses which led to a more consistent improvement. (Yeung et al., 2021) The model chosen to answer the research question using these approaches is the Segformer model since it classified the highest amount of stem pixels at 433520 compared to 45112 for FCN and 66223 for DeepLabv3. To further improve the model against overfitting, normalisation and early stopping was also considered. Figure 35 shows the training loss curve for each of the five cross validation folds.

A graph of different colored lines

AI-generated content may be incorrect.

*Figure 35: Segformer training loss curves*

Firstly, the Segformer implementation of the dice loss failed in its classification performance for all classes as seen in figure 36 for the confusion matrix and figure 37 for the class-wise IoU graph.



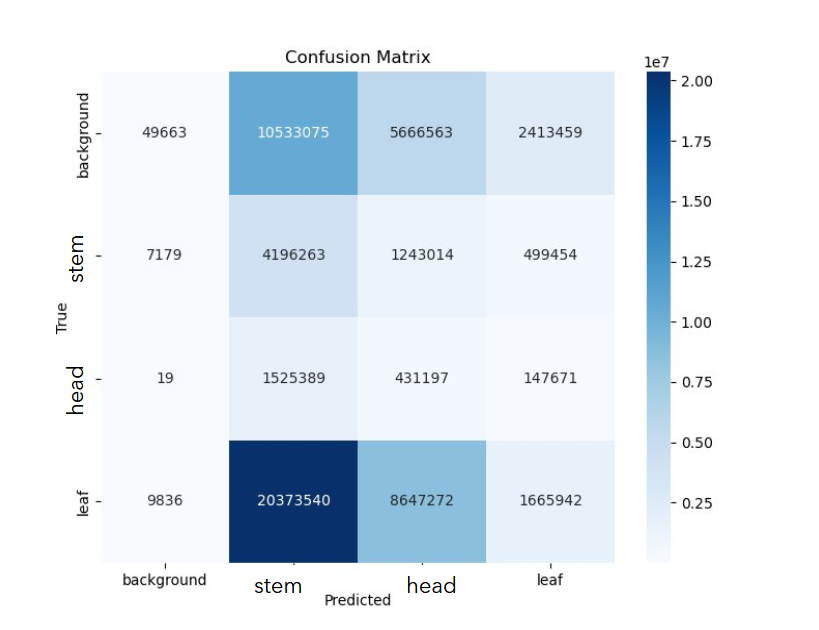
*Figure 36: Segformer dice loss confusion matrix*

A graph with blue squares

AI-generated content may be incorrect.

*Figure 37: Segformer dice loss IoU graph*

From the above two figures, it can be seen that the background was being predicted much higher which was not reasonable as this is the highest proportion class. However, the stem class was predicted much less which was reasonable as there are less pixels related to this class. The Segformer model with focal loss did much better but a weighting issue had to be overcome since different tunings of the weight parameters for each class led to worse performance as seen in figure 38 for the confusion matrix and figure 39 for IoU graph.



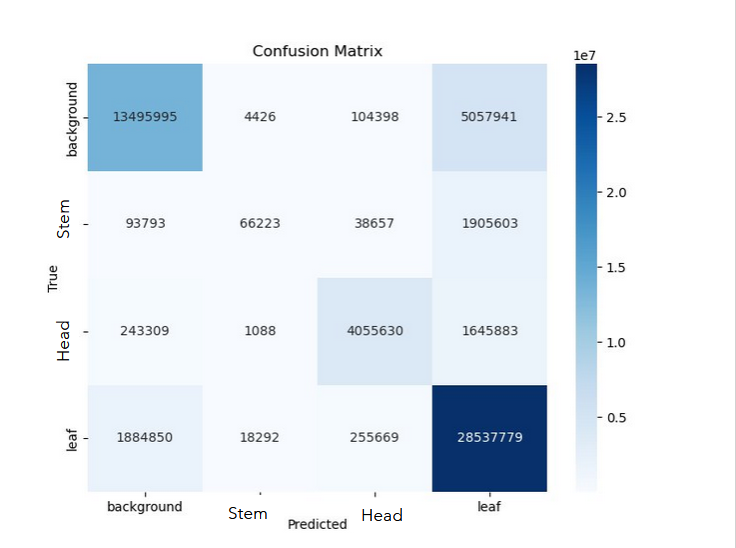
*Figure 38: Segformer focal loss with weighting issue confusion matrix*

A graph with blue squares

AI-generated content may be incorrect.

*Figure 39: Segformer focal loss with weighting issue IoU graph*

From figure 39, stem had a substantial increase in its IoU but the performance of all the other classes had to be sacrificed. This is seen especially for the background class which was previously the highest performing class. On the other hand, the weighting issue was fixed with the correct hyperparameter tuning to balance out the prioritisation of classifying stem while balancing the performance of all the other classes. Figures 40 and 41 are the confusion matrix and IoU graph for the final Segformer model as the proposed solution to this objective.



*Figure 40: Segformer focal loss with correct weighting confusion matrix*

A graph of blue rectangular bars

AI-generated content may be incorrect.

*Figure 41: Segformer focal loss with correct weighting IoU graph*

# Costs

This report has no financial costs associated with it and only considers time as its only constraint. The cost of time can be divided based on the sections of the report’s outline of the solution which is conceptualisation, development, implementation, and evaluation. Firstly, conceptualisation deals with an initial understanding of the data, data preprocessing such as standardisation, combining different datasets for specific models and removing outliers. This phase will take about a week for all the models. Secondly, the development phase consists of transforming the dataset to conform with each model’s assumptions, feature engineering for aggregating multiple variables into one variable and if necessary, creating an initial framework of the model. This second phase will take approximately two weeks to complete. Thirdly, the implementation phase deals with model training and finetuning hyperparameters. This phase will be completed in two weeks. Lastly, the evaluation phase will consist of choosing different evaluation metrics for each model by choosing the highest metric score. Moreover, the metric score will be compared against the threshold and other models to determine if an ensemble method is needed or whether the model should be discarded or finetuned. This phase will span for approximately three weeks. This project should then take 2 months to complete in total. During the phases of these projects, caution will be taken in handling the NVT datasets given and consider the greater implications of the results and the models.

# Ethics and Privacy

The report utilises data that conforms with ethical and privacy considerations through understanding the data sources given and the processing techniques used which were addressed to execute the report’s objectives. The NVT datasets given are already anonymised and thus no information pertaining to individual farmers were known or revealed during the data science process. However, the datasets do include longitude and latitude data which was not used or required to be stated during EDA, modelling or interpretation. Furthermore, the datasets used will not be shared with anyone in case bad faith actors utilise the dataset for negative intentions. Another consideration is the environmental impact based on a recommender system for frost damage which should still not be overshadowed by sustainable agricultural practices as well as not mandating for unsustainable practices to reach an optimal level of wheat yield. Further steps will have to be taken to ensure that the results of the projects do not pressure farmers or agronomists to forgo sustainable agricultural practices for higher production, especially affected by climate change. The last consideration is for the responsible use of AI which is just seen as tools for decision making and not replacing the guidance and expertise of areas in the field and can be prone to errors. Caution should be taken in interpreting these results and predictions by using a secondary source.

# Conclusion

In closing, climate change is a very serious threat to wheat production that this report aims in delivering results that can lead to actionable steps in breeding wheat varieties, mitigating the risk of damage to the crop, and favouring environmental conditions more optimal for yield. This report aims to accomplish all four machine learning objectives in unison towards the core goal of creating a safer and sustainable future food supply chains. This report intends on extending the works of current literature by addressing real-world data and using machine learning techniques to prioritise wheat yield production on a variety of cultivars, to primarily focus on yield, and use recent data for frost damage. The solution consists of four projects: determining the most climate resistant cultivar, determining frost damage on a chosen set of features, recognising the most yield-productive environment based on environmental features, and segmenting a global wheat image dataset. Firstly, the most climate-resistant cultivars found are listed in table 3 with a mean yield of 837.29kg/m^2 whilst being resistant to different environmental conditions which requires further research on these breeds to understand the impacts under climate change. For the second project, the neural network had the greatest performance and was able to model the overall dataset with an f1-score of 0.93 which was classified as a success. The third project used a combination of techniques such as feature importance, grid search, and optimisation to find the most important features and their predicted yields. The features which had the greatest importance was found to be geographical positioning based on the longitude and latitude and trace elements in the soil such as copper or carbon. Lastly, for segmenting the images using computer vision techniques, the final model found was a Segformer model using data augmentation techniques, normalisation, early stopping and focal loss with an mIoU of 0.6 which requires serious consideration and is not considered to have succeeded. Each of these projects were incorporated into a Microsoft PowerBI dashboard to further inform users of the effects of climate change on wheat. The main constraints of this report were time-based, and any ethical considerations were considered. In considering this report, the use cases are innumerable and will provide immense values to researchers and farmers alike.

# Appendix

1. Percentage change in farm profits for different climate change models

A map of australia with red and blue spots

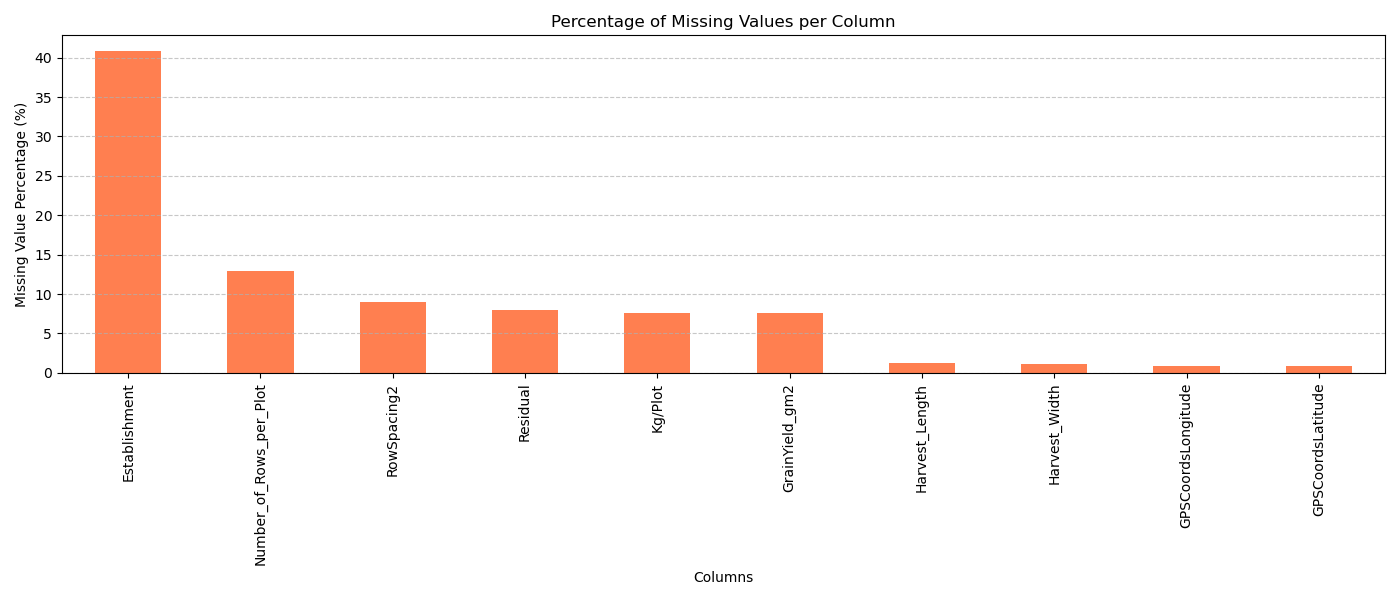
AI-generated content may be incorrect.

1. Percentage of missing values per column for rainfall metadata dataset

A graph with orange squares

AI-generated content may be incorrect.

1. Percentage of missing values per column for 2015\_2019 Main Plot Dataset



1. Percentage of missing values per column for 2015\_2020 Main Trial Layout Dataset

A graph with a rectangle and black text

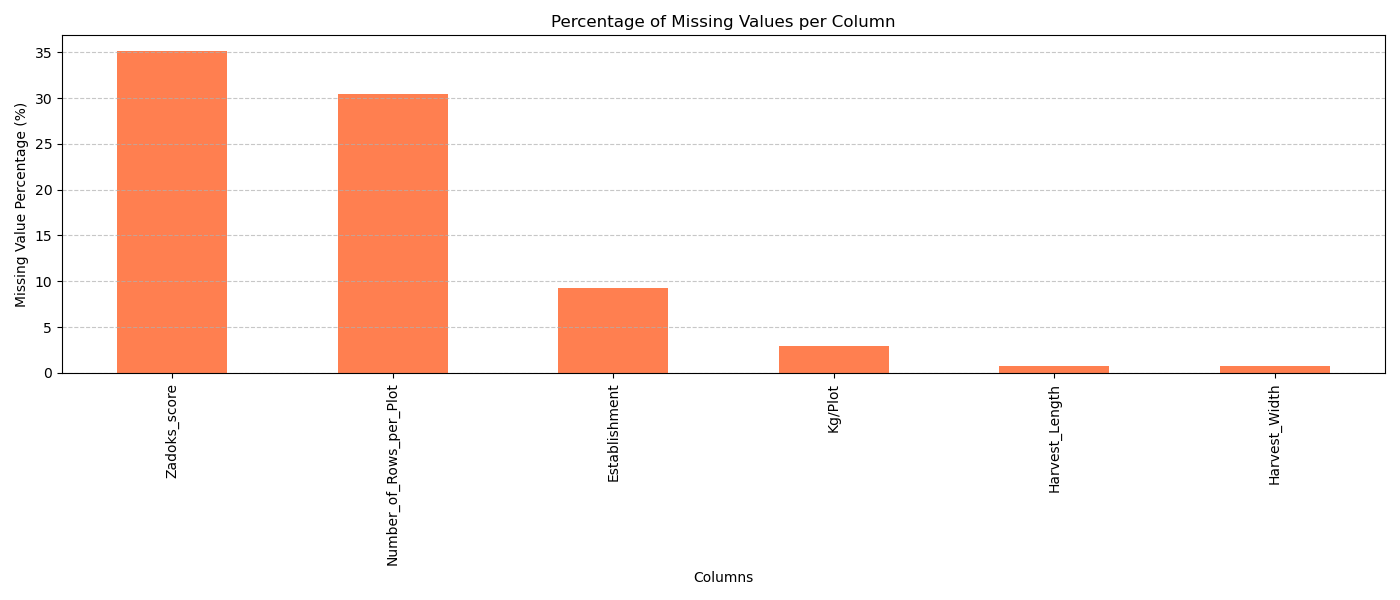
AI-generated content may be incorrect.

1. Percentage of missing values per column for 2020 Main Wheat Dataset

A graph with orange and white lines

AI-generated content may be incorrect.

1. Percentage of missing values per column for 2020 NVT Wheat Dataset



1. Missing values in merged dataset

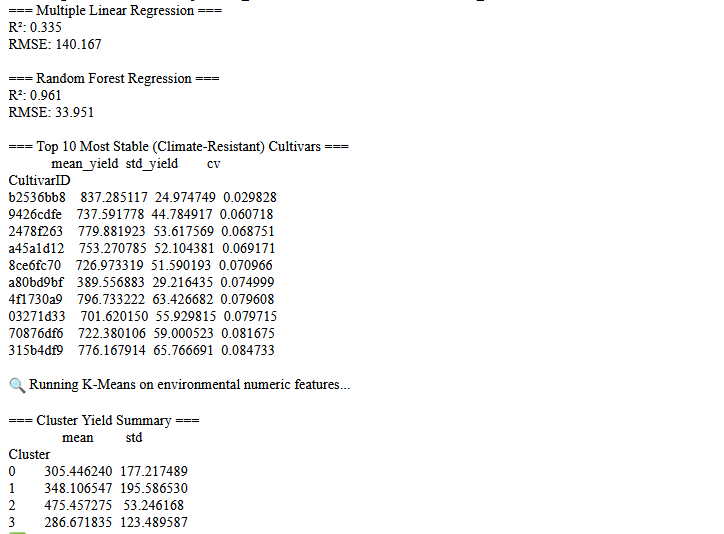
A close up of a screen

AI-generated content may be incorrect.

1. Variance Inflation Factors for Multiple Linear Regression

|  |  |
| --- | --- |
| **Variable** | **VIF** |
| TotalRainfall | 1.14 |
| MinDepth | 945.36 |
| MaxDepth | 945.30 |
| Boron | 1.57 |
| Conductivity\_(EC) | 2.21 |
| Copper | 1.49 |
| Exchangeable\_Ca | 3.40 |
| Exchangeable\_K | 3.21 |
| Exchangeable\_Mg | 1.71 |
| Exchangeable\_Na | 1.18 |
| Manganese | 1.84 |
| Organic\_Carbon | 2.24 |
| pH\_(CaCl2) | 24.02 |
| pH\_(water) | 25.12 |
| Phosphorous | 2.00 |
| Potassium\_(K) | 4.73 |
| Sulphur | 1.55 |
| Texture | 1.86 |
| Zinc | 2.83 |

1. Most Climate-resistant cultivars output



1. Modelling Frost Damage Output

A screenshot of a computer

AI-generated content may be incorrect.A black background with white text

AI-generated content may be incorrect.

A black background with white text

AI-generated content may be incorrect.

1. Most Productive Environment for Yield Output

A screenshot of a computer

AI-generated content may be incorrect.

A black background with white lines

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

# Reference List

1. AIML.com. (2025, April 10). *Advantages and disadvantages of Random Forest*. AIML.com. <https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/>​[AIML](https://aiml.com/category/ml-interview-questions/supervised-learning/classification/?utm_source=chatgpt.com)
2. Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking atrous convolution for semantic image segmentation.* arXiv preprint arXiv:1706.05587. <https://doi.org/10.48550/arXiv.1706.05587>
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
4. Cotton Seed Distributors. (n.d.). *Flowering is a critical period of crop development*. <https://csd.net.au/knowledge_library/flowering-is-a-critical-period-of-crop-development/>
5. CSIRO. (n.d.). *Agricultural Production Systems Simulator*. <https://www.csiro.au/en/about/corporate-governance/ensuring-our-impact/impact-case-studies/future-industries/agricultural-productions-systems-simulator>
6. David, E., Madec, S., Sadeghi-Tehran, P., Aasen, H., Zheng, B., Liu, S., Kirchgessner, N., Ishikawa, G., Nagasawa, K., Badhon, M. A., Pozniak, C., de Solan, B., Hund, A., Chapman, S. C., Baret, F., Stavness, I., & Guo, W. (2020). *Global Wheat Head Detection (GWHD) dataset: A large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods.* **Plant Phenomics, 2020**, 3521852. <https://doi.org/10.34133/2020/3521852>
7. Department of Primary Industries and Regional Development. (2020, June 15). *How wheat yields are influenced by climate change in Western Australia*. Government of Western Australia. <https://www.agric.wa.gov.au/climate-change/how-wheat-yields-are-influenced-climate-change-western-australia>
8. Eichi, V. R., Okamoto, M., Garnett, T. P., & Eckermann, P. (2020). *Strengths and weaknesses of National Variety Trial data for multi-environment analysis: A case study on grain yield and protein content.* *Agronomy, 10*(5), 753. <https://doi.org/10.3390/agronomy10050753>
9. FOLUR. (n.d.). *Wheat.* Food Systems, Land Use, and Restoration. <https://www.folur.org/wheat>
10. GeeksforGeeks. (2023, November 5). *Davies-Bouldin Index*. <https://www.geeksforgeeks.org/davies-bouldin-index/>
11. Hollaway, K., & Knights, S. (2025). *NVT harvest report interim version – Central Queensland* (GRDC Report No. 619408). Grains Research and Development Corporation. <https://nvt.grdc.com.au/__data/assets/pdf_file/0028/619408/nvt-harvest-report-central-queensland-interim-version-grdc-20250327.pdf>
12. Huang, S., & Deng, H. (2021). Chapter 8: Scalability: LASSO & PCA. In *Data Analytics: A Small Data Approach*. Retrieved May 1, 2025, from <https://dataanalyticsbook.info/chapter-8-scalability-lasso-pca.html>
13. Hughes, N., Lu, M., Soh, W., & Lawson, K. (2021). *Simulating the effects of climate change on the profitability of Australian farms* (ABARES Working Paper). Australian Bureau of Agricultural and Resource Economics and Sciences. <https://www.agriculture.gov.au/sites/default/files/documents/abares-climate-change-working-paper.pdf>
14. Jain, A. (2025, March 10). *Ridge and lasso regression in Python: A complete tutorial*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/>
15. Liddicoat, C., Hayman, P., Alexander, B., Rowland, J., Maschmedt, D., Young, M.-A., Hall, J., Herrmann, T., & Sweeney, S. (2012). *Climate change, wheat production and erosion risk in South Australia’s cropping zone: Linking crop simulation modelling to soil landscape mapping* (Technical Report 2012/05). Department of Environment, Water and Natural Resources. <https://cdn.environment.sa.gov.au/environment/docs/kb-gen-climate-change-wheat-production-and-erosion-risk.pdf>​
16. Pennsylvania State University. (n.d.). *10.7 – Detecting multicollinearity using variance inflation factors*. STAT 462: Applied Regression Analysis. Retrieved May 1, 2025, from <https://online.stat.psu.edu/stat462/node/180/>
17. Qu, K. (2024). Research on linear regression algorithm. *MATEC Web of Conferences, 395*, 01046. <https://doi.org/10.1051/matecconf/202439501046>
18. Rahimi Eichi, V., Okamoto, M., Garnett, T., Eckermann, P., Darrier, B., Riboni, M., & Langridge, P. (2020). Strengths and weaknesses of national variety trial data for multi-environment analysis: A case study on grain yield and protein content. *Agronomy, 10*(5), 753. <https://doi.org/10.3390/agronomy10050753>
19. Rahmat, M. F., Zulkafli, Z., Ishak, A. J., Abdul Rahman, R. Z., De Stercke, S., Buytaert, W., Tahir, W., Ab Rahman, J., Ibrahim, S., & Ismail, M. (2023). Supervised feature selection using principal component analysis. *Knowledge and Information Systems, 66*(3), 1955–1995. <https://doi.org/10.1007/s10115-023-01993-5>
20. ScienceDirect. (n.d.). *Coefficient of variation*. In *ScienceDirect Topics*. Retrieved May 1, 2025, from <https://www.sciencedirect.com/topics/engineering/coefficient-of-variation>
21. Spanos, A. (1986). The multivariate linear regression model. In *Statistical foundations of econometric modelling* (pp. 571–607). Cambridge University Press. <https://doi.org/10.1017/CBO9780511599293.024>
22. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709. <https://doi.org/10.1016/j.compag.2020.105709>
23. Wang, C., Balazs, J., Szarvas, G., Ernst, P., Poddar, L., & Danchenko, P. (2022). *Calibrating imbalanced classifiers with focal loss: An empirical study.* In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 145–153). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-industry.14>
24. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). *SegFormer: Simple and efficient design for semantic segmentation with transformers.* *arXiv preprint* arXiv:2105.15203. <https://doi.org/10.48550/arXiv.2105.15203>
25. Yathish, V. (2022, August 4). *Loss functions and their use in neural networks*. Towards Data Science. <https://towardsdatascience.com/loss-functions-and-their-use-in-neural-networks-a470e703f1e9>
26. Yeung, M., Sala, E., Schönlieb, C.-B., & Rundo, L. (2021). *Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation.* *Computerized Medical Imaging and Graphics, 95,* 102026. <https://doi.org/10.1016/j.compmedimag.2021.102026>
27. Zheng, B., Chapman, S. C., Christopher, J. T., Frederiks, T. M., & Chenu, K. (2015). Frost trends and their estimated impact on yield in the Australian wheatbelt. *Journal of Experimental Botany, 66*(12), 3611–3623. <https://doi.org/10.1093/jxb/erv163>