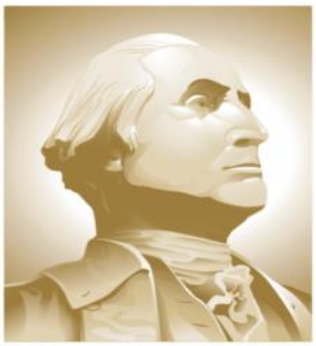


THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Enhance Sustainable Agriculture in South Africa

This Final Report is submitted in fulfillment of the requirement for the Capstone project of

Master of Science

in

GW Data Science Program

Columbian College of Arts and Sciences

By

MANOJKUMAR YERRAGUNTLA

Guided by

Prof. Michael Mann

Supervised by

Prof. Abdi Awl

Table of Contents:

Glossary of Terms:

1. **Class Imbalance:** An unequal distribution of classes within a dataset, which can lead to biased model predictions and reduced performance, often addressed through techniques such as oversampling, undersampling, or class weighting.
2. **Cross-Validation:** A model evaluation technique that partitions the dataset into multiple subsets, training the model on a subset and validating it on another, used to assess model performance and generalization.
3. **Ensemble Learning:** A machine learning technique that combines multiple models to improve predictive performance, often used to reduce variance, enhance accuracy, and improve generalization.
4. **Exploratory Data Analysis (EDA):** The process of analyzing and visualizing data to uncover patterns, trends, and insights, often used to understand the structure and characteristics of a dataset before modeling.
5. **Feature Engineering:** The process of transforming raw data into meaningful features that can improve the performance of machine learning models, including techniques such as scaling, normalization, and creation of derived features.
6. **Feature Importance:** A measure of the relative importance of each feature in predicting the target variable, often used to identify the most informative features for model training.
7. **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics and offering an overall measure of model performance.
8. **Geospatial Data:** Data that represents spatial features on the Earth's surface, typically captured through satellites, or other geographic information systems (GIS) technologies.
9. **Hyperparameters**: Parameters that are set before model training and remain constant during training, such as learning rate, number of trees, and maximum depth, which can significantly impact model performance and behavior.
10. **LightGBM:** Light Gradient Boosting Machine, a gradient boosting framework designed for efficiency and speed, utilizing histogram-based techniques to optimize training time and memory usage.
11. **Mode Prediction:** A prediction approach where the most frequently occurring class within a set of predictions is selected as the final prediction, often used to aggregate predictions at a higher level of granularity.
12. **Precision:** The proportion of correctly predicted positive cases out of all predicted positive cases, indicating the accuracy of positive predictions made by the model.
13. **Random Forest:** An ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.
14. **Recall:** The proportion of actual positive cases that were correctly identified by the model, representing the model's ability to capture all positive instances.
15. **Satellite Bands:** Specific spectral bands of electromagnetic radiation captured by satellite sensors, each representing a different range of wavelengths and providing unique information about surface features.
16. **XGBoost:** Extreme Gradient Boosting, a scalable and efficient gradient boosting algorithm known for its superior performance in classification and regression tasks.

1. Introduction

1.1. Introduction/Background

This project aims to address critical food security challenges in South Africa by leveraging data science techniques, particularly advanced crop classification methods. South Africa faces issues such as crop failures, fluctuating production, and managing imports, which threaten its food security.

1.2. Problem Statement

The primary challenge is to utilize data science tools to enhance food security by improving crop monitoring and resource allocation in South Africa's agricultural sector.

1.3. Problem Elaboration

The project will focus on developing advanced models for multi-crop classification to ensure sustainable agricultural practices. This involves analyzing various factors like vegetation health, land cover classification, and land surface temperatures.

1.4. Motivation

Recent advancements in data analytics and machine learning offer promising solutions for addressing food security challenges. We can enhance crop identification accuracy and optimize agricultural practices by harnessing these technologies.

1.5. Project Scope

This project's scope encompasses implementing advanced feature selection techniques, refining predictive models, and evaluating their effectiveness across diverse regions in South Africa. To achieve its objectives, the project will utilize a comprehensive dataset comprising remote sensing data, satellite imagery, and historical crop data.

2. Literature Review

2.1. Relevant Research

Satellites equipped with remote sensing instruments capture data by detecting electromagnetic radiation (light) reflected or emitted from the Earth's surface. When sunlight falls on the Earth's surface, it interacts with various surface features such as vegetation, water bodies, soil, and built-up areas. Different surface materials absorb, transmit, or reflect light at different wavelengths across the electromagnetic spectrum.

Band	Resolution	Central Wavelength	Description
B1	60 m	443 nm	Ultra Blue (Coastal and Aerosol)
B2	10 m	490 nm	Blue
B3	10 m	560 nm	Green
B4	10 m	665 nm	Red
B5	20 m	705 nm	Visible and Near Infrared (VNIR)
B6	20 m	740 nm	Visible and Near Infrared (VNIR)
B7	20 m	783 nm	Visible and Near Infrared (VNIR)
B8	10 m	842 nm	Visible and Near Infrared (VNIR)
B8a	20 m	865 nm	Visible and Near Infrared (VNIR)
B9	60 m	940 nm	Short Wave Infrared (SWIR)
B10	60 m	1375 nm	Short Wave Infrared (SWIR)
B11	20 m	1610 nm	Short Wave Infrared (SWIR)
B12	20 m	2190 nm	Short Wave Infrared (SWIR)

Remote sensing satellites typically carry sensors capable of capturing light in multiple spectral bands. These bands are strategically chosen to capture specific information about the Earth's surface. For instance, visible bands capture light in the range of wavelengths visible to the human eye, while near-infrared and shortwave infrared bands capture light just beyond the visible spectrum. Each band provides unique information about surface properties, allowing for a comprehensive understanding of the Earth's surface and its changes over time.

In the context of agricultural applications, remote sensing data, particularly from satellites, plays a crucial role in monitoring crop health, assessing vegetation dynamics, and optimizing agricultural practices. By analyzing data from different spectral bands, researchers and practitioners can derive valuable insights into various aspects of crops, including growth patterns, stress detection, yield estimation, and land use management.

For example:

- The blue band (B2) is sensitive to the presence of water and can help in mapping water bodies and assessing soil moisture levels.
- Near-infrared bands (such as B6 and B7) are highly sensitive to chlorophyll content in vegetation, allowing for the detection of vegetation health and biomass.
- Shortwave infrared bands (such as B11 and B12) provide information about moisture content in vegetation and soil, aiding in drought monitoring and irrigation management.

- The Enhanced Vegetation Index (EVI) is derived from satellite data, combining information from multiple bands, to provide quantitative measures of vegetation health, accounting for atmospheric effects and background noise.
- Additionally, the hue band, representing the color or hue of the surface as observed by satellite sensors, can aid in land cover classification and crop type mapping.

By integrating data from multiple spectral bands and employing advanced analysis techniques, remote sensing-based approaches offer a powerful tool for monitoring agricultural systems at various spatial and temporal scales. These approaches contribute to sustainable agriculture practices by enabling informed decision-making, resource management, and precision farming techniques.

Overall, the utilization of satellite remote sensing data, along with appropriate analysis methods, holds immense potential for advancing agricultural research, supporting agricultural policy-making, and enhancing food security on a global scale.

3. Methodology

3.1. Dataset Description

Objective: The dataset aims to facilitate the analysis and modeling of various agricultural parameters, including vegetation health, moisture content, and land cover classification, using satellite remote sensing data.

Data Source: Satellite imagery obtained from SENTINEL 2 satellites.

Format: Data from satellite images is extracted as Parquet files.

Spectral Bands: The dataset includes six Parquet files per location, each representing a different spectral band. The spectral bands included are:

- Blue (B2)
- Near-Infrared (B6)
- Shortwave Infrared (B11)
- Shortwave Infrared (B12)
- Enhanced Vegetation Index (EVI)
- Hue (HUE)

Geographical Locations: The dataset covers two geographical locations: 34S_19E_258N and 34S_19E_259N.

Dataset Structure: Each location file contains approximately 3.7 million rows and 231 variables.

Dataset Size: The total compressed dataset size is 3.7 GB.

3.2. Data Preprocessing

The data preprocessing phase of our data science capstone project involved several key steps to ensure the quality and suitability of the dataset for subsequent analysis and modeling tasks. Below is a summary of the preprocessing steps undertaken:

Data Integration:

The dataset comprised six Parquet files per geographical location, each representing a different spectral band from the SENTINEL 2 satellite. These files were combined into a single unified data frame, allowing for easier manipulation and analysis of the entire dataset.

Handling Missing Values:

Upon inspection, it was found that the variable "ts_complexity_cid_ce" contained no values. As a result, this variable was dropped from the dataset to avoid any potential bias or confusion in subsequent analyses.

Column Deduplication:

A total of 75 column duplicates were identified within the dataset. These duplicates were removed to eliminate redundancy and streamline the dataset, ensuring that each variable represented unique information.

Handling Duplicates:

No duplicate rows were detected within the dataset, eliminating the need for any deduplication procedures.

Outlier Detection and Treatment:

Outliers were identified based on location-specific thresholds. For Location (34S_19E_258N), 47476 outliers were identified, while for Location (34S_19E_259N), 118119 outliers were detected. Outliers were removed to mitigate their influence on subsequent analyses.

Feature Elimination:

Feature importance analysis was conducted to identify the least important features within the dataset. A total of 74 such features were identified and subsequently removed to reduce dimensionality and improve computational efficiency.

3.3. Feature Engineering

Feature engineering played a pivotal role in our data science capstone project, enabling the creation of meaningful and informative derived features from remote sensing data. The goal of feature engineering was to enhance the predictive power of our models by incorporating domain-specific knowledge and extracting relevant information from the raw data. Below is an overview of the feature engineering process:

Remote Sensing Derived Features:

Leveraging remote sensing data, we derived several new features using mathematical formulas tailored to capture important aspects of agricultural landscapes and environmental conditions. These derived features included:

- Vegetation Ratio Index (VRN)
- Normalized Burn Ratio 2 (NBR2)
- Enhanced Vegetation Index (EVI)
- Soil Adjusted Vegetation Index (SAVI)
- Normalized Difference Water Index (NDWI)
- Normalized Difference Vegetation Index (NDVI)
- Modified Improved Ratio Brightness Index (MIRBI)

By incorporating these derived features, we aimed to capture nuanced relationships between spectral bands and surface characteristics, providing deeper insights into vegetation health, water content, soil properties, and environmental changes.

Integration with Existing Features:

While existing geographical features may not have yielded satisfactory results, the inclusion of derived remote sensing features offered a more comprehensive representation of the underlying agricultural landscape dynamics. These features were seamlessly integrated with the existing dataset, enriching it with additional information relevant to our project objectives.

Mathematical Formulas and Transformations:

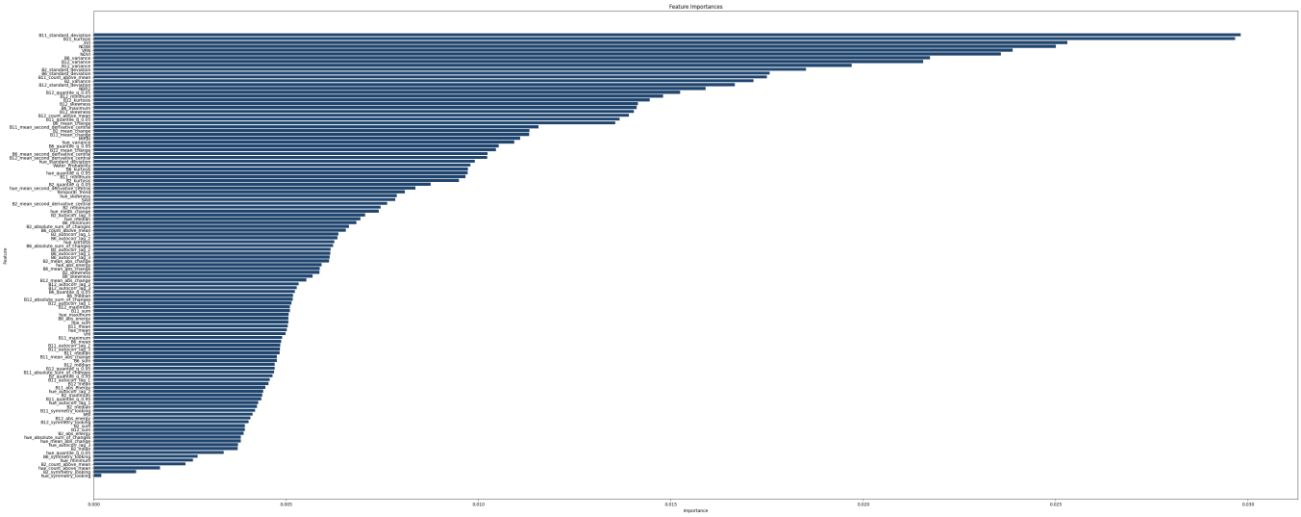
Each derived feature was calculated using specific mathematical formulas and transformations tailored to extract meaningful information from the raw satellite imagery data. These formulas were carefully selected based on domain knowledge and prior research to ensure their relevance and effectiveness in capturing relevant agricultural parameters.

Evaluation and Validation:

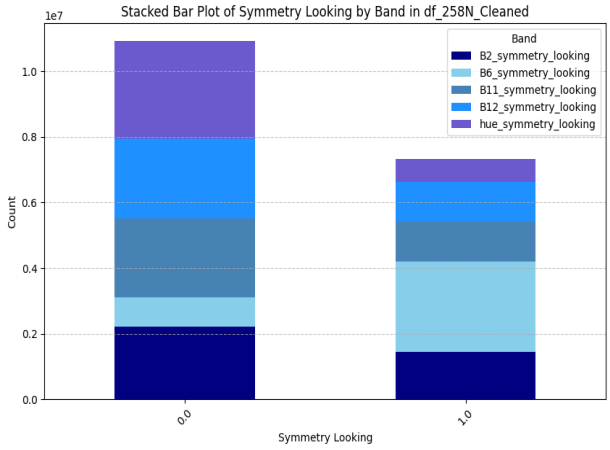
The efficacy of the derived features was assessed through rigorous evaluation and validation processes. This involved analyzing feature importance, conducting correlation analyses, and assessing their contribution to model performance through techniques such as cross-validation and model comparison.

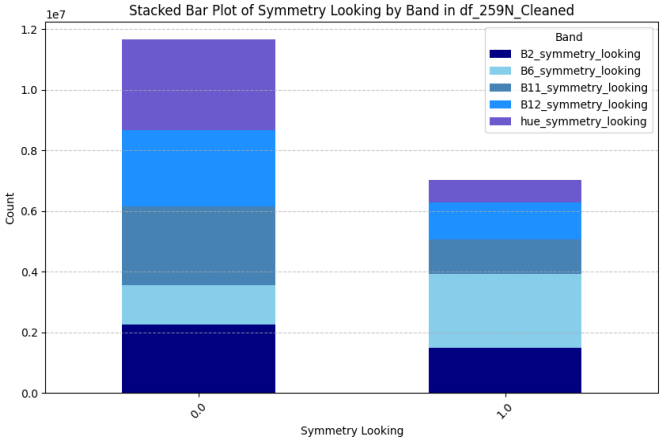
3.4. Exploratory Data Analysis

Histogram of Feature Importance:



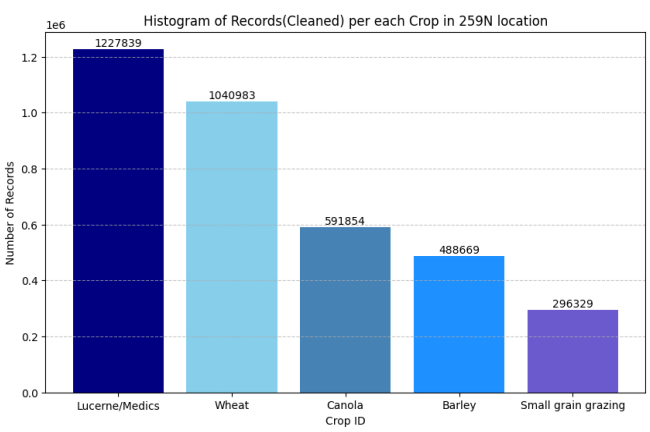
- The histograms provided a visual representation of the feature importance scores within the dataset.
- Features were ranked from most important to least important, with the top-ranked feature being "B11_standard Deviation" and the least important feature being "hue_symmetry_looking."
- This analysis helped identify the most influential features for predicting crop types, with an emphasis on those contributing to approximately 80% of the variability in crop classification.
- Notably, four out of the seven derived features were among the top six most important features, indicating their significance in enhancing the predictive performance of the model.

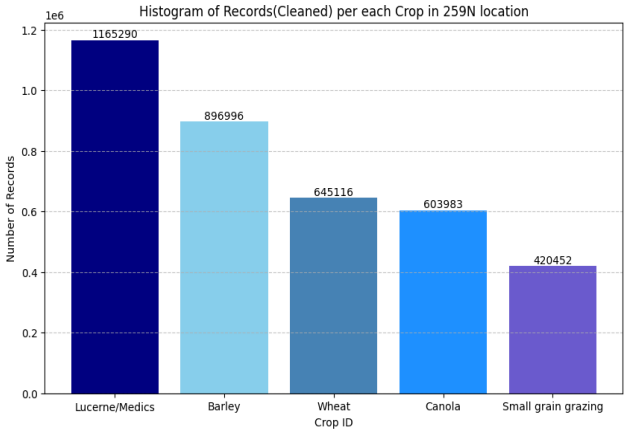




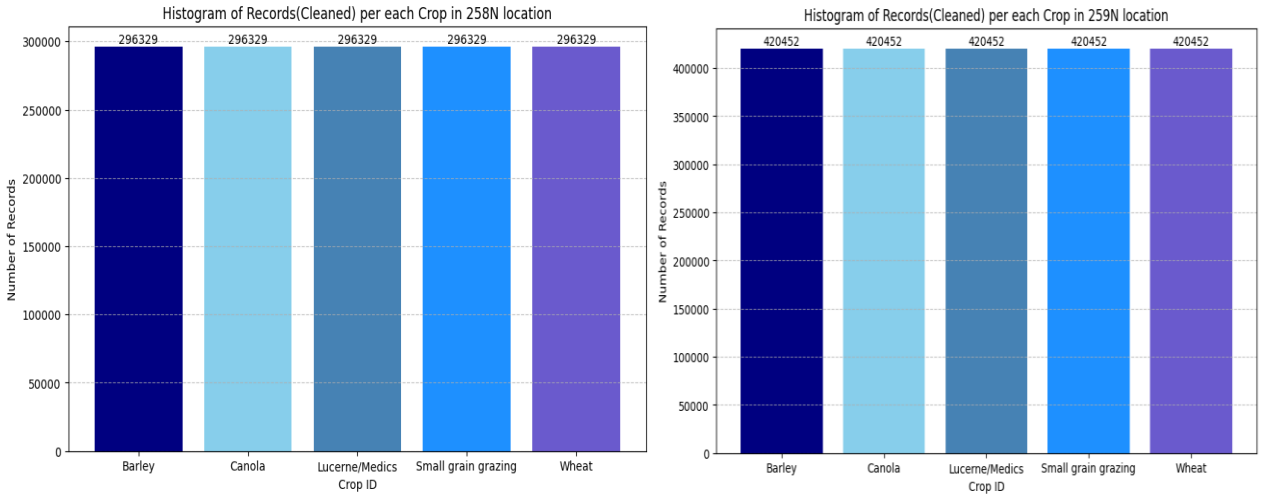
- The stacked bar charts visualized the symmetry looking of different crops across different bands of the UV spectrum (B2, B6, B11, B12, and Hue bands) for two distinct locations.
- Analysis revealed that, for both locations, all bands except B6 exhibited a higher proportion of non-symmetrical crops compared to symmetrical ones.

- This observation provided valuable insights into the relationship between spectral bands and crop characteristics, suggesting potential differences in crop health and development across different bands and locations.





- The histograms depicted the distribution of crop records across two distinct locations, showcasing the prevalence of each crop within each location.
- In both locations, lucerne/medics emerged as the most commonly grown crop, while small grain grazing was identified as the least grown crop.
- Notably, variations were observed in the rankings of other crops across the two locations, with wheat, canola, and barley occupying different positions in the crop hierarchy.
- These findings provided insights into regional agricultural practices and preferences, highlighting variations in crop cultivation patterns across different locations.
- Analysis of the histograms revealed a significant class imbalance between the most grown crop class (lucerne/medics) and the least grown crop class (small grain grazing).
- The presence of class imbalance poses challenges for modeling, as it can skew model predictions and affect the accuracy and reliability of classification results.
- Addressing class imbalance is essential for developing robust predictive models that effectively capture the variability in crop types and ensure a balanced representation of all classes.



- The above histograms displayed a balanced number of records in each location, achieved through under-sampling.
- Under-sampling was chosen to address the class imbalance, reduce computational power, and optimize storage space on AWS resources.
- Both under and oversampling techniques were considered, with under-sampling selected due to its effectiveness in improving model predictions while reducing computational overhead.

3.5. Data Modeling

Model Selection:

- We utilized three ensemble learning algorithms for our modeling tasks: Random Forest, XGBoost, and LightGBM.
- These models were chosen for their ability to handle complex datasets, including highly correlated geospatial data, and their robust performance in classification tasks.

Hyperparameter Tuning:

- Each model was fine-tuned using a set of hyperparameters to optimize performance and generalization.
- For Random Forest, hyperparameters included n_estimators, max_depth, criterion, and min_samples_split.
- XGBoost hyperparameters included n_estimators, max_depth, learning_rate, colsample_bytree, and gamma.
- LightGBM hyperparameters included learning_rate, max_depth, num_boost_round, reg_lambda, and reg_alpha.

Model Justification:

Random Forest: Random Forest is known for its robustness to overfitting, ability to handle large datasets, and interpretability. It aggregates multiple decision trees to reduce variance and improve predictive values, making it well-suited for handling highly correlated geospatial data.

XGBoost: XGBoost is a scalable and efficient gradient-boosting algorithm known for its superior performance in classification tasks. It utilizes an ensemble of weak learners to boost predictive accuracy and handle complex relationships within the data, making it a suitable choice for our modeling tasks.

LightGBM: LightGBM is a gradient-boosting framework designed for efficiency, speed, and accuracy. It employs a novel tree-growing algorithm and histogram-based techniques to optimize training time and memory usage while maintaining high predictive performance. Its ability to handle large datasets and capture complex patterns makes it an ideal candidate for modeling highly correlated geospatial data from different bands.

Handling Correlated Geospatial Data:

- Random Forest, XGBoost, and LightGBM are well-suited for handling highly correlated geospatial data due to their ensemble-based nature and ability to capture complex relationships within the data.
- Ensemble methods like Random Forest, XGBoost, and LightGBM are robust to multicollinearity and can effectively exploit the information captured by correlated features to improve predictive performance.
- Moreover, these models can capture nonlinear relationships and interactions between different bands, allowing us to leverage the rich information contained in geospatial data for accurate crop classification.

4. Results & Analysis

Model Performance:

Models	Precision	Recall	F1-Score
Random Forest	0.58	0.58	0.55
XGBoost	0.58	0.58	0.56
LightGBM	0.49	0.47	0.42

Analysis of Model Metrics:

- Precision measures the proportion of correctly predicted positive cases out of all predicted positive cases. Our models achieved a precision of approximately 0.58, indicating that around 58% of predicted crop types were accurate.
- Recall, or sensitivity, measures the proportion of actual positive cases that were correctly identified by the model. Our models achieved a recall of approximately 0.57 to 0.58, indicating that they correctly identified around 57% to 58% of actual positive cases.

- F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. Our models achieved an F1-Score of approximately 0.55 to 0.56, indicating overall good performance in capturing both precision and recall.

Across all three models, we observed consistent performance in terms of precision, recall, and F1-Score, with minor variations between models.

XGBoost does show slightly higher performance in terms of precision, recall, and F1-Score compared to Random Forest and LightGBM.

The models demonstrated comparable performance, with F1-Scores ranging from 0.55 to 0.56, indicating balanced performance in both precision and recall.

5. Conclusion

5.1. Conclusion

- Our project successfully applied Random Forest, XGBoost, and LightGBM models to predict crop types using geospatial data, achieving comparable performance in precision, recall, and F1-Score.
- Exploratory data analysis revealed insights into crop distribution patterns and dataset biases, highlighting the need for further research to address limitations.
- Future efforts should focus on incorporating additional bands, in-depth feature engineering, and evaluation at the plot level to enhance model accuracy and generalization.
- The implications of accurate crop classification extend to precision agriculture practices and optimized resource allocation, underscoring the importance of continued research in agricultural predictive modeling.

5.2. Project Limitation

- One significant limitation is the absence of certain crucial satellite bands, such as B3, B4, B7, and B8, which are essential for deriving specific remote sensing data.
- Incorporating additional satellite bands, such as B3, B4, B7, and B8, can enrich the dataset and provide more comprehensive information about vegetation health, soil properties, and environmental conditions.
- Exploring additional derived features and spectral indices based on these bands can further enhance the predictive power of our models and capture subtle variations in crop characteristics.
- Our analysis may be limited in scope, focusing primarily on predicting crop types using highly correlated geospatial data. Other relevant factors or variables that could impact agricultural dynamics, such as weather patterns, soil characteristics, or socio-economic factors, may not have been fully accounted for in our analysis.

5.3. Future Research

- Future research could focus on incorporating additional satellite bands, such as B3, B4, B7, and B8, to enrich the dataset and capture more nuanced information about vegetation health, soil properties, and environmental conditions.
- In-depth feature engineering techniques, including exact calculations and advanced spectral indices, can further enhance the predictive power of our models and improve the accuracy of crop classification.
- Extending the analysis to evaluate performance at the plot level can provide deeper insights into model accuracy and generalization.
- By aggregating predictions at the plot level and comparing them to ground truth labels, we can assess the model's ability to capture spatial variability and accurately classify crops within individual fields.
- Integrating additional external data sources, such as weather data, soil maps, or historical crop yield data, can enrich the predictive modeling framework and provide valuable context for understanding agricultural dynamics.

6. References

In our data science capstone project, we relied on various references to inform our methodology and deepen our understanding of remote sensing techniques and derived features. These references encompassed authoritative sources, research papers, and online resources pertinent to our project objectives. Below is a compilation of the references used:

1. "Sentinel-2 Bands Combinations" (https://gisgeography.com/sentinel-2-bands-combinations/): This online resource provided valuable insights into the optimal combinations of Sentinel-2 spectral bands for various remote sensing applications. It guided our selection of bands and informed our feature engineering process.
2. "Composites in Remote Sensing" (https://gsp.humboldt.edu/olm/Courses/GSP_216/lessons/composites.html): This resource from Humboldt State University offered a comprehensive overview of composite images and their utility in remote sensing analysis. It contributed to our understanding of image compositing techniques and their relevance to our project.
3. "Spectral Indices Adopted Equations with Sentinel-2 Multispectral Instrument (MSI) Bands" (https://www.researchgate.net/figure/Spectral-indices-adopted-equations-with-Sentinel-2-Multispectral-Instrument-MSI-bands_tbl3_342010458): This research paper provided a detailed compilation of spectral indices equations tailored for Sentinel-2 Multispectral Instrument (MSI) bands. It guided our formulation of derived features and spectral indices used in our analysis.
4. "Spectral-Spatial Classification of Sentinel-2 Imagery for Land Cover Mapping in Urban Areas" (https://isprs-archives.copernicus.org/articles/XLII-5/683/2018/isprs-archives-XLII-5-683-2018.pdf): This research article presented a methodology for spectral-spatial classification of Sentinel-2 imagery for land cover mapping in urban areas. It informed our approach to image classification and land cover analysis.
5. "Custom Scripts for Sentinel-2 Bands" (https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/bands/): This online repository provided custom scripts and code snippets for processing Sentinel-2 satellite imagery, including band combinations and spectral index

calculations. It served as a valuable reference for implementing custom processing algorithms in our project.

These references were instrumental in guiding our project's methodology and ensuring the credibility and rigor of our analysis.