Project Report Mining of Massive Data Spring 2024

Crop Yield Prediction Using Big Data Techniques

Manal Rizwan[1], Muhammad Ibrahim[2], and Aisha Muhammad Nawaz[3]

[1]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

May 6, 2024

Abstract

The increasing concern of Crop Yield prediction has encouraged the use of technical means to predict crop yield in advance. With the massive data, using Big Data Techniques has become a necessity. Our project focuses on taking advantage of PySpark and using methods like Linear Regression, Gradient Boosted, K-Means Clustering, and Random Forest Regressor to solve the problem of crop yield prediction. After pre-processing the data and applying the models, we fine tuned the models and managed to obtain an amazing accuracy of 99 percent with Gradient Boosted. We then combined our models and deployed the PySpark code to Cluster on Google DataProc. Overall, we managed to obtain a decent accuracy but improvements such as increasing the number of features can be made to increase performance.

1 Introduction

Agriculture, which has been the foundation of civilizations since antiquity, has changed into an industry where traditional methods and cutting-edge technology now coexist. The modern agricultural environment is defined by a symbiotic relationship between tradition and innovation, with data-driven insights guiding every choice toward sustainability and efficiency. This change, driven by the combination of agricultural knowledge and data science, has ushered in a new era of productivity. Leading this movement is PySpark, a powerful framework for data processing known for its ability to handle large datasets with ease. Our project demonstrates the revolutionary potential of this combination by first properly preprocessing the Crop Yield dataset and then applying carefully chosen PySpark's methods such as Linear Regression, Gradient Boosted Trees, K-Means Clustering, and Random Forest Regressor to predict agricultural yield with amazing accuracy and efficiency. After this, we deployed our PySpark code to a cluster on Google DataProc. We hope that this project will further agricultural analytics and add to the conversation on resource management and sustainable food production.

2 Problem Background

With the many moving parts of agricultural production, forecasting crop yield is still a difficult task full with unknowns. A wide range of interrelated factors, such as insect pressures, market dynamics, soil health, and climate variability, affect crop productivity. Conventional approaches to yield estimation are frequently insufficient, requiring more effective and scalable alternatives. Crop yield forecasting is made more difficult by the intrinsic complexity of agricultural production, the increasing unpredictability of weather patterns, and the growing need for sustainable farming methods. Particularly in light of climate change, which is increasing the frequency and severity of extreme weather occurrences, there are substantial concerns. These variations can throw off planting plans, have an impact on crop development and growth, and intensify pest and disease challenges already present. Yield prediction is further complicated by the important roles that market dynamics and customer preferences play in determining crop pricing and agricultural practices. Therefore, there is a pressing need for sophisticated analytical tools and methodologies that

can handle these complex issues and offer useful information for raising resilience and production in agriculture.

3 Literature Review

3.1 BIG DATA ANALYTICS IN AGRICULTURE [1]

The research paper discusses the application of Big Data Analytics in agriculture, and how structured and unstructured data can help farmers to have insights. This not only will allow them to make informed decisions but it will also reduce losses due to unforeseen disasters. It outlines various techniques such as predictive analysis, recommendation systems, and data mining, along with specific applications like intelligent crop recommendation systems and precision agriculture using MapReduce. Moreover, the paper also discusses different machine learning approaches including Grey wolf optimization and K-means clustering are explored for crop prediction, while Smart Farming services like the Internet of Things and Cloud Computing are detailed. The paper also delves into crop analysis using data mining techniques and proposes a Spark-based system for distributed data analytics in agriculture. It also mentions some of the problems faced in big data analysis while also addressing future scopes including product traceability and genetic engineering. Lastly, a comparison table is also provided that suggests MapReduce for weather data and K-Means Clustering for crop and vegetation data analysis.

3.2 CROP PREDICTION ON THE REGION BELTS OF INDIA: A NAIVE BAYES MapReduce PRECISION AGRICULTURAL MODEL [2]

This research paper suggests a data-driven utilized by big data analytics to optimize crop selection for agriculture in the city of Telangana, India. It starts off by analyzing the agricultural system in Telangana while also collecting relevant data from various sources. During this time, the development of a recommendation system based on the Naive Bayes classifier to suggest suitable crops like rice, cotton, maize, and chillies, was still ongoing. Moreover, after building upon previous research in precision agriculture, the paper introduces a methodology to predict crop suitability across various zones in Telangana. The methodology also takes into account other factors such as soil types, temperature, rainfall, and atmospheric pressure. The results from this methodology are then used to advise the optimal timing for sowing and harvesting crops. Lastly, the paper analyses the effectiveness of the proposed approach, with the goal of improving agricultural productivity in Telangana through future enhancements.

3.3 PREDICTION OF CROP YIELD USING BIG DATA [3]

The paper explores the use of big data analytics for crop yield prediction and food security in China. It emphasises the need for early and exact crop projections to ensure food security. The paper presents a strategy that uses MapReduce and closest neighbour analysis to accurately predict crop yield. The results show that the proposed strategy is excellent at predicting crop production with minimum deviation. The paper continues by pushing for additional research to improve data collection and computing efficiency, putting emphasis on the approach's scalability across diverse geographical contexts.

3.4 ENVIRONMENT CHANGE PREDICTION TO ADAPT CLIMATE-SMART AGRICULTURE USING BIG DATA ANALYTICS [4]

The paper focuses on using Big Data Analytics to forecast weather patterns and assist farmers in making informed agricultural decisions, which will solve global food insecurity aggravated by climate change. The authors use the Hadoop framework to collect data from a variety of sources, including social media, sensors, and weather forecasts, with an emphasis on precipitation, temperature, and cloud cover in Karnataka. They use Hive for data processing, taking use of its SQL capabilities to manage massive datasets stored on HDFS. MapReduce is used for data analysis, which improves query response times. The work uses Apache Mahout to create prediction functions using k-means clustering and logistic regression techniques, then evaluates accuracy and visualises

the findings using Flot. The authors plan to improve the model for generating alerts on disasters in the future, showing its potential for solving important global concerns using advanced data analytics approaches.

3.5 CROP YIELD PREDICTION USING AGRO ALGORITHM IN HADOOP [5]

The paper provides a new approach that uses the Agro Algorithm to predict crop production and select suitable crops, therefore improving farmers' profitability and supporting the agricultural sector. It combines soil, and crop disease datasets to make more precise predictions. The authors analyse existing crop prediction systems and their limitations, emphasising the importance of variables such as soil pH and seed selection. Implementation on the Hadoop platform allows for the processing of massive datasets, while normalisation and classification improve prediction accuracy. They present an architecture for data collecting, analysis, and prediction that addresses technical gaps and limited data availability. The model predicts soil quality and recommended crops using varied datasets, allowing farmers to make informed decisions. The paper shows that this technique improves agricultural output and production standards while providing enormous value to farmers by directing crop choices depending on weather and soil conditions. Future research intends to classify illnesses comprehensively, hence improving crop quality.

3.6 SURVEY ON WEATHER PREDICTION USING BIG DATA ANALYTICS [6]

The paper mentions several weather prediction systems and their effects, emphasising the importance of precise forecasts in minimising damage and improving agricultural planning. It describes the collecting and analysis of weather information using big data analytics, focusing on problems such as dataset size and parameter changes. Different prediction approaches, such as MapReduce and Linear Regression, are reviewed, with comparative analysis demonstrating their efficacy in weather forecasting. Overall, the paper recommends for the use of MapReduce for efficient processing of big weather datasets and proposes using K-means clustering to improve agricultural predictions, with the goal of increasing resilience and optimising farming techniques.

3.7 A SURVEY ON DATA MINING FOR CROP YIELD PREDICTION [7]

The paper examines various data mining strategies for crop yield prediction, showing their importance in dealing with the complex dynamics of agricultural systems. It emphasises the need of accurate yield forecasting for farmers and agricultural organisations, as well as the importance of new technology in increasing production efficiency. This paper examines the use of techniques such as K-Means, K-Nearest Neighbour, Artificial Neural Networks, and Support Vector Machines in agricultural data analysis using a thorough literature review. It explores the problems and opportunities of data mining in agriculture, demonstrating its ability to extract significant insights from vast datasets. The paper supports the integration of data mining techniques in agricultural decision-making processes by presenting a methodology for crop yield prediction and emphasising the role of inputs such as weather and soil conditions, ultimately leading to increased productivity and sustainability in agriculture.

3.8 PREDICTIVE ABILITY OF MACHINE LEARNING METHODS FOR MASSIVE CROP YIELD PREDICTION [8]

The paper compares the predictive accuracy of machine learning and linear regression algorithms for crop yield prediction across ten crop datasets, emphasising the necessity of precise yield estimation for agricultural planning. Several machine learning approaches, including multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression, and k-nearest neighbour, were tested on real data from an irrigation zone in Mexico. Four different accuracy metrics were used: root mean square error (RMSE), root relative square error (RRSE), normalised mean absolute error (MAE), and correlation factor (R). The results show that M5-Prime and k-nearest neighbour algorithms consistently outperform others, with the lowest average errors

across metrics. This shows their applicability for large-scale crop yield prediction in agricultural planning. M5-Prime stands out for its excellent ability in producing models with low errors, indicating its usefulness as a tool for predicting agricultural yields.

3.9 MACHINE LEARNING FOR LARGE SCALE CROP YIELD FORECASTING [9]

This paper offers a machine learning baseline for large-scale crop yield forecasting, with the goal of addressing previous approach constraints, particularly transferability and scalability. By merging agronomic concepts with machine learning approaches, the study creates a workflow that prioritises correctness, modularity, and reusability. Crop growth and development principles guide the creation of features, which rely on data from a variety of sources such as crop simulation outputs, weather, remote sensing, and soil. The workflow is modular and can be easily converted to different crops and nations with small setup modifications, allowing for repeatable tests and outcomes. Case studies including five crops from three countries show encouraging results, with early season estimates equivalent to operational forecasting systems in several crops and regions. The paper identifies areas for further improvement, such as incorporating other data sources and refining predictive features, with the ultimate goal of promoting the use of machine learning in large-scale crop output forecasting.

3.10 WB-CPI: WEATHER BASED CROP PREDICTION IN INDIA USING BIG DATA ANALYTICS [10]

The paper addresses the critical topic of crop yield prediction in India using big data analytics and innovative approaches. Recognising agriculture's critical significance in India's economy, as well as the problems faced by unexpected weather conditions, the paper points out the use of modern predictive technologies. The authors take a multi-step strategy, using data from several places on temperature, rainfall, soil, seed, and crop yield. They preprocess the data, use the MapReduce framework for analysis, then use k-means clustering to extract insights. They also create a recommender system that suggests crops depending on the analysed data. They use bar graphs, scatter plots, and a graphical user interface to visualise and explain the correlations between many elements that influence crop yield. The paper emphasises the importance of big data analytics in revolutionising agricultural practices while also proposing scalable crop prediction tools. By combining meteorological and agricultural data, the authors hope to deliver useful information for farmers, resulting in increased agricultural output .

4 Dataset Information

The dataset comprises agricultural data spanning from 1997 to 2020, encompassing multiple crops cultivated across various states in India. It includes essential features for crop yield prediction, such as crop types, crop years, cropping seasons (e.g., Kharif, Rabi, Whole Year), states, cultivated area, production quantities, annual rainfall, fertilizer usage, pesticide usage, and calculated yields. Each entry in the dataset provides specific information about the crop, including its name, the year of cultivation, the cropping season, the state where it was cultivated, the total land area under cultivation, the quantity of crop production, the annual rainfall received, the amount of fertilizer used, the amount of pesticide used, and the calculated yield (production per unit area). It contains 19,689 entries and 20 columns, providing comprehensive agricultural data for analysis and modeling.

5 Data Preprocessing

Before applying big data techniques on the dataset, the data was preprocessed and cleaned. The preprocessing steps were carried out in python.

5.1 Label Encoding

Three columns in the dataset were of object type. These were Crop, Season and State. The aforementioned columns were encoded to type int using the LabelEncoder of sklearn.

5.2 Outlier Detection and Removal

Upon visualization of the box plots of columns, it was observed that there were a considerable amount of outliers in the data. To amend this, winsorization was applied with an upper limit of 0.8 and a lower limit of 0.1. The box plots visualising the dataset before and after Winsorization are shown below.

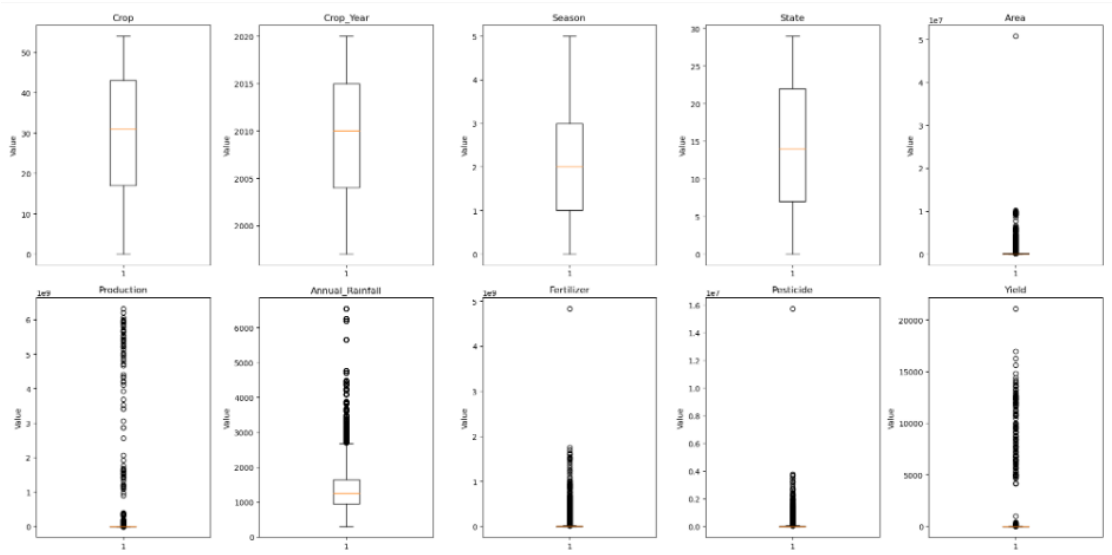

Figure 1: Box Plots Before Winsorization

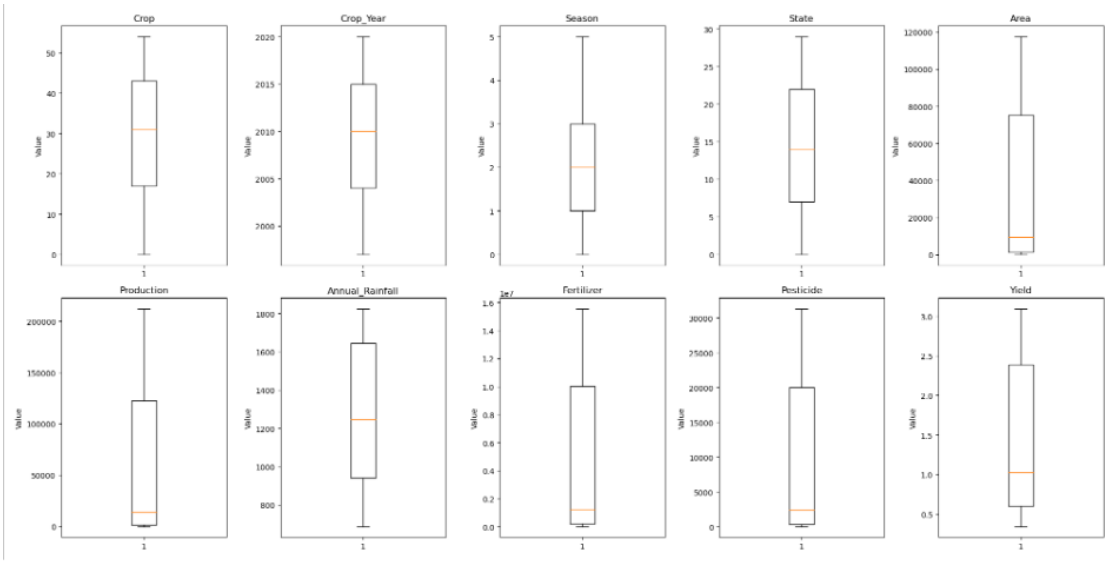

Figure 2: Box Plots After Winsorization

5.3 Standardization

Standardization was applied on the data using the StandardScaler of sklearn. As the range of values of the columns differed largely, this step was necessary so that the model didn't develop a bias towards columns having larger values. Due to this step, the distribution of data improved. The histograms visualising the dataset before and after standardization are shown below.

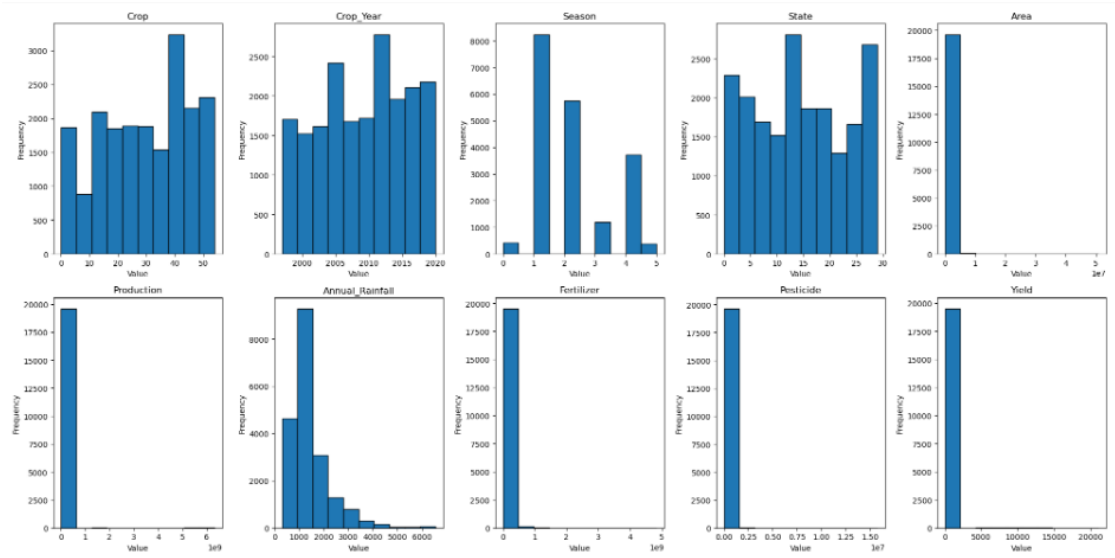

Figure 3: Histograms Before Standardization

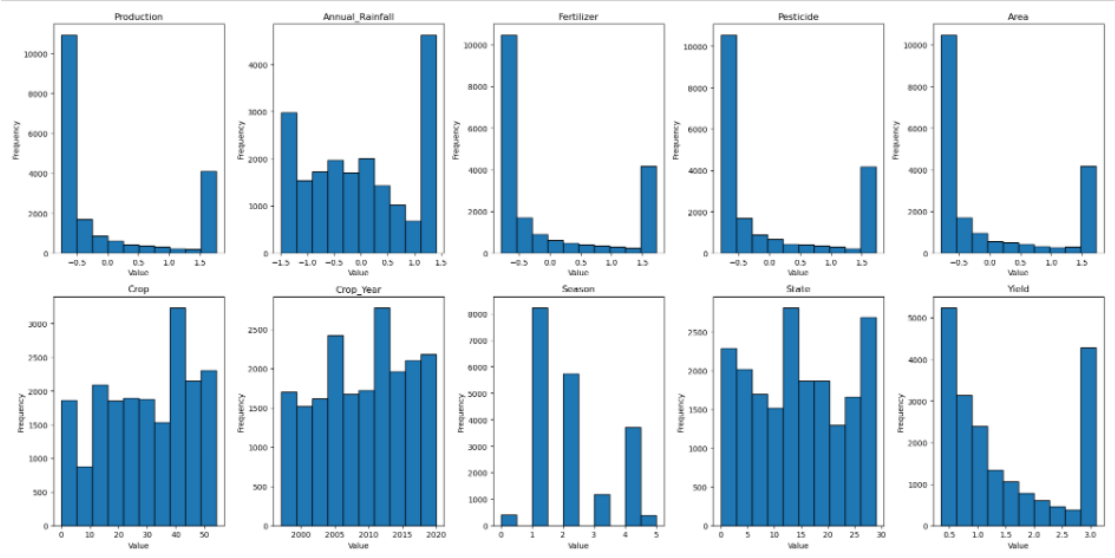

Figure 4: Histograms After Standardization

6 Methodology

Several big data techniques were applied on the dataset using PySpark's machine learning library. These include Linear Regression, Gradient Boosted Trees, K-Means Clustering, and Random Forest Regressor. After the individual assesments, we combined the best performing models of all the techniques and delpoyed the PySpark code to a cluster on Google DataProc. The detailed methodology followed and results obtained from each technique is explained below.

6.1 Linear Regression

Multiple linear regression was applied on the dataset as the first technique. Three models were developed. For model 1, 70 percent of the data was used for the training phase while the rest was used for the testing phase. Model 2 had a split of 80-20 whereas the third model had a split of 90-10. Model 1 outperformed the other models with a R2 score of 0.43, Root Mean Squared Error of 0.77, and Mean Absolute Error of 0.62. In addition to these evaluation metrics, Accuracy with a 1.5 tolerance was also calculated. The Accuracy of the best model was 95 percent.

6.1.1 Scatter Plot

The scatter plot of the best performing model is displayed below. As can be seen from the plot, the model only captures about 42 percent of the variability of the data.

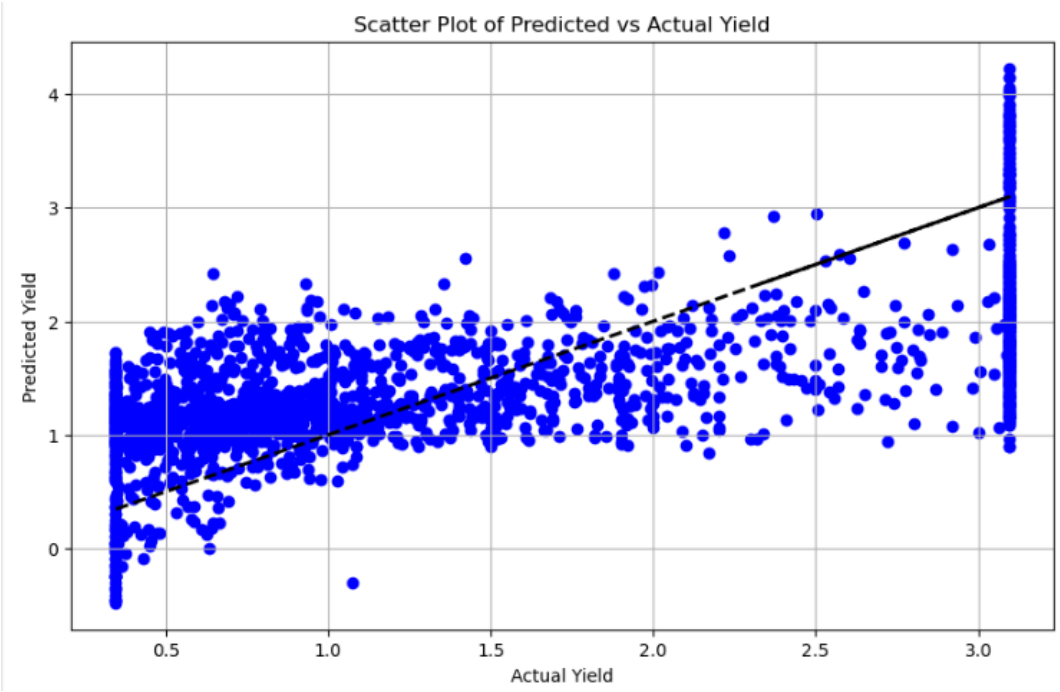

Figure 5: Scatter Plot of Predicted vs Actual Yield

6.1.2 Residual Plot

The shape of the plot suggests that a non-linear relationship exists between the independent and dependent variables. This explains the poor R2 score of the model.

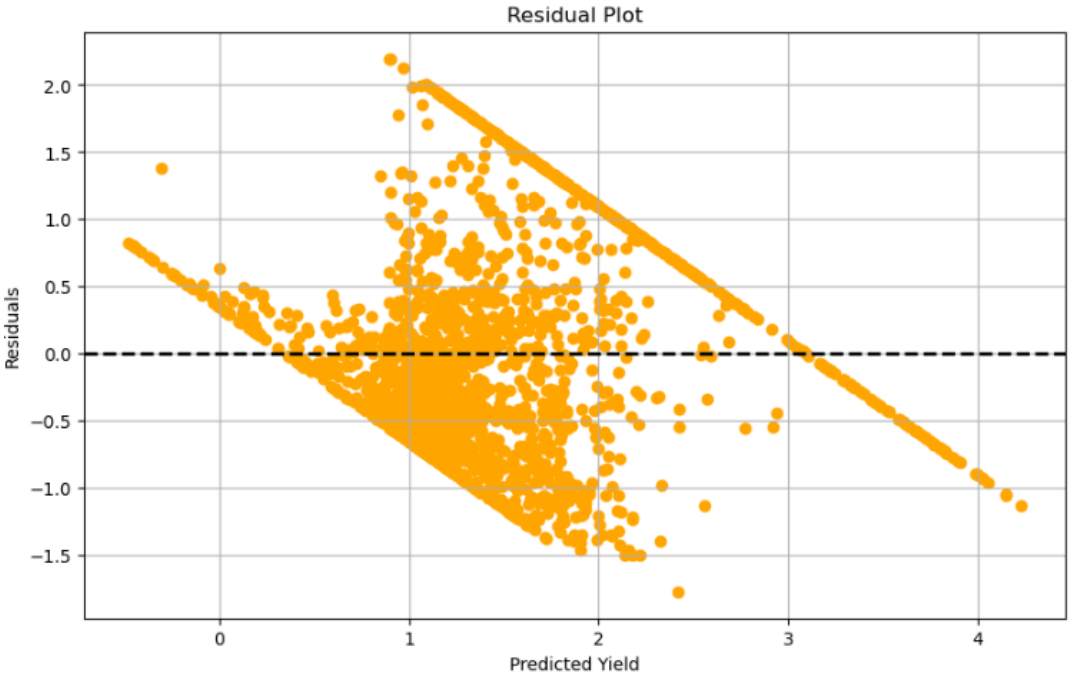

Figure 6: Residual Plot

6.1.3 Frequency Distribution of Residuals

The data distribution of the residuals is right-skewed. This suggests that the model is overestimating the target variable.

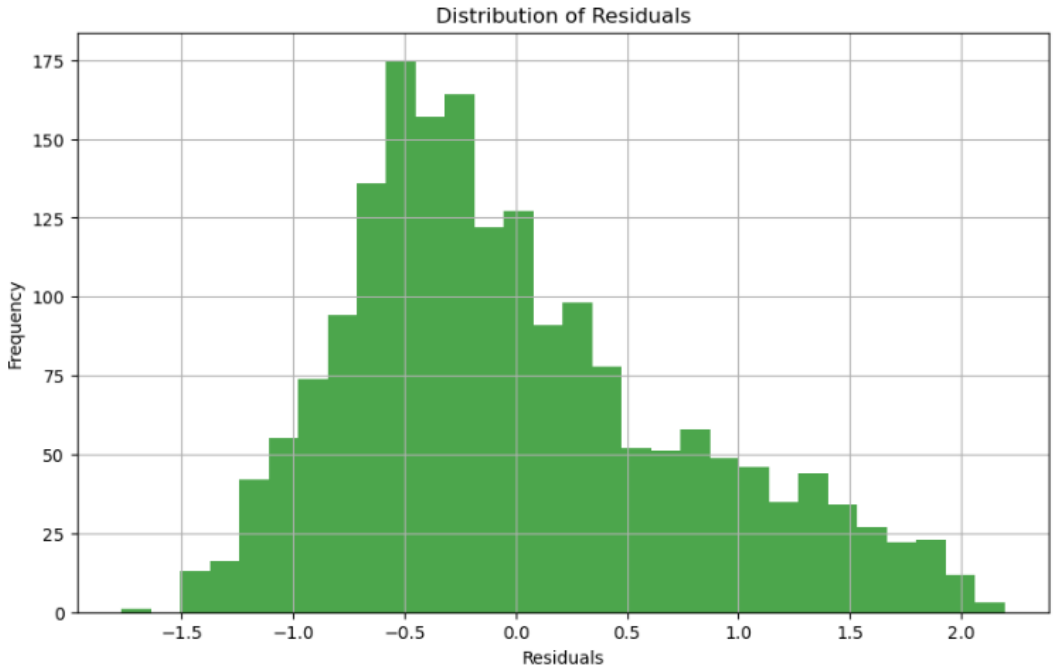

Figure 7: Frequency Distribution of Residuals

6.2 Gradient Boosted Trees

Three models of Gradient Boosted Trees were developed. The best performing model had a train-test split of 90-10. It had a R2 score of 0.85, Root Mean Squared Error of 0.39, and Mean Absolute Error of 0.26. The model's Accuracy with a 1.5 tolerance level was 99 percent.

6.2.1 Scatter Plot

The scatter plot of the best performing model of Gradient Boosted Trees is displayed below. As can be seen from the plot, the model seems a fairly good fit.

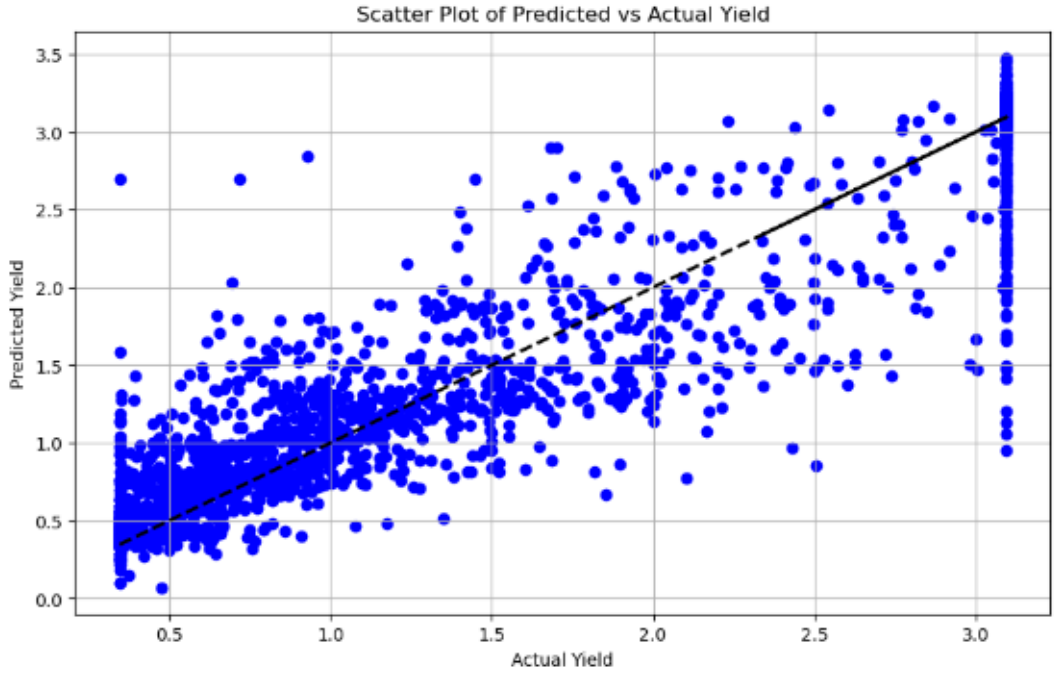

Figure 8: Scatter Plot of Predicted vs Actual Yield

6.2.2 Residual Plot

As the points in the plot are randomly scattered around the horizontal line, with no clear pattern, it indicates that the model's predictions were unbiased and have constant variance across the range of predicted values.

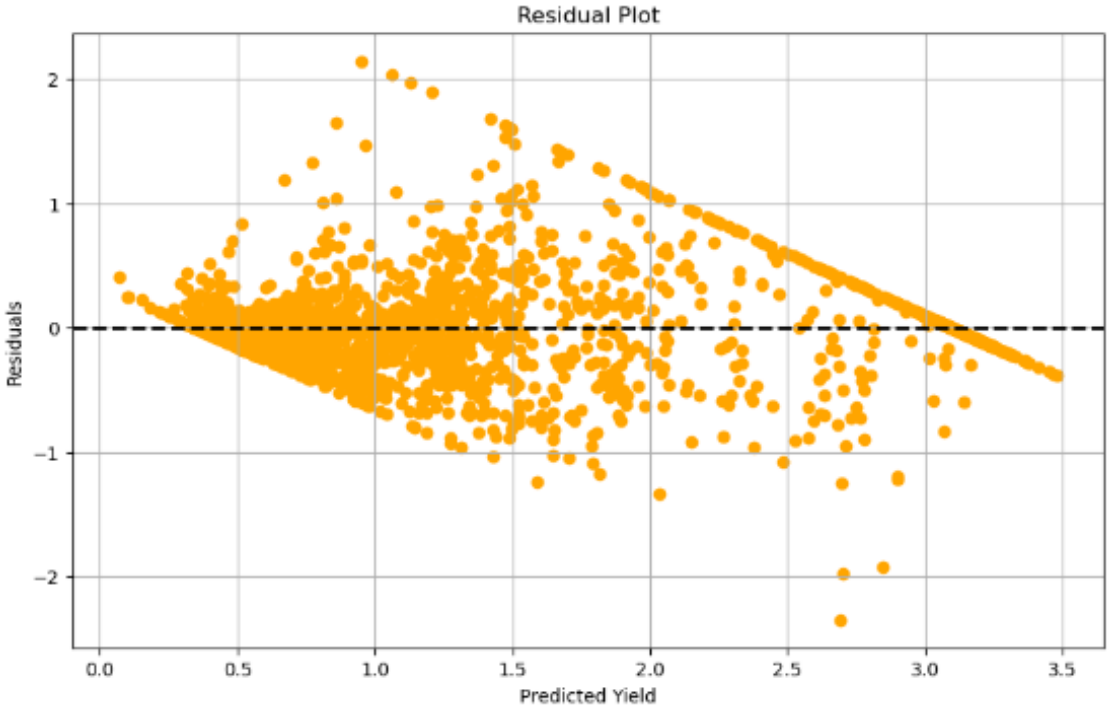

Figure 9: Residual Plot

6.2.3 Frequency Distribution of Residuals

The data distribution of the residuals is fairly normal. This suggests that on average, the model was making predictions that were fairly accurate.

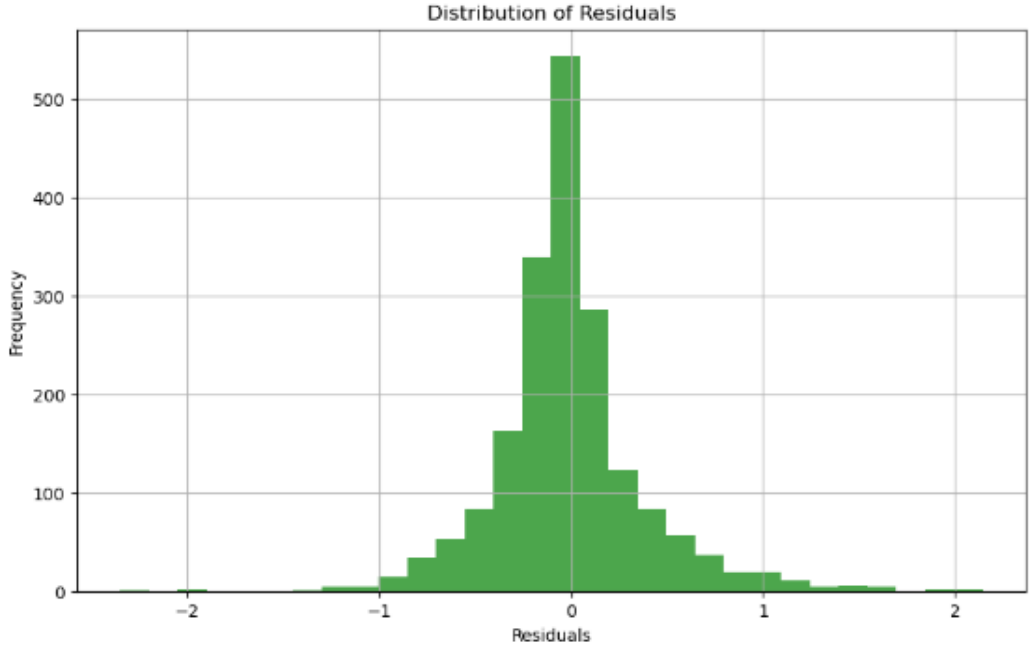

Figure 10: Frequency Distribution of Residuals

6.3 Random Forest Regressor

To find the best results of predicting yield through different methods in PySpark, we also explored the Random Forest Regressor technique. A model was built with number of trees set to 3, feature subset strategy set to auto, impurity set to variance, and max bins set to 32. This model was later fine tuned to obtain an amazing accuracy (with a 1.5 tolerance level) of 94.5 percent, R2 score of 0.51, Root Mean Root Mean Squared Error of 0.711, Mean Absolute Percentage Error of 59.13, Mean Absolute Error of 0.533, Mean Squared Error of 0.528. This model used a 80-20 split for training and testing datasets, with a maximum depth of trees set to 10. The Scatter, Residual, and Frequency Distribution of Residuals plots of the best performing model of Random Forest Regressor is visible in figure 11, 12 and 13 respectively.

Among the other models of Random Forest Regressor we experimented with, we obtained an accuracy (with a 1.5 tolerance level) of 92 percent and an R2 Score of 0.33 with model 1 which used a 70-30 split for training and testing datasets, with maximum depth of trees set to 6. With model 2 we obtained an accuracy (with a 1.5 tolerance level) of 93.5 percent and an R2 Score of 0.45 with the same 70-30 split for training and testing datasets but with maximum depth of trees set to 8. With model 3 we obtained an accuracy (with a 1.5 tolerance level) of 94.2 percent and an R2 Score of 0.48 with 70-30 split for training and testing datasets, with maximum depth of trees set to 9. This trend suggests that increasing the depth of trees increases the accuracy and R2 score values but this was observed only up to maximum depth of trees set to 10.

6.3.1 Scatter Plot

The scatter plot of the best performing model of Random Forest Regressor is shown in Figure 11. As visible, the model captures roughly 51 percent of the variability of the data.

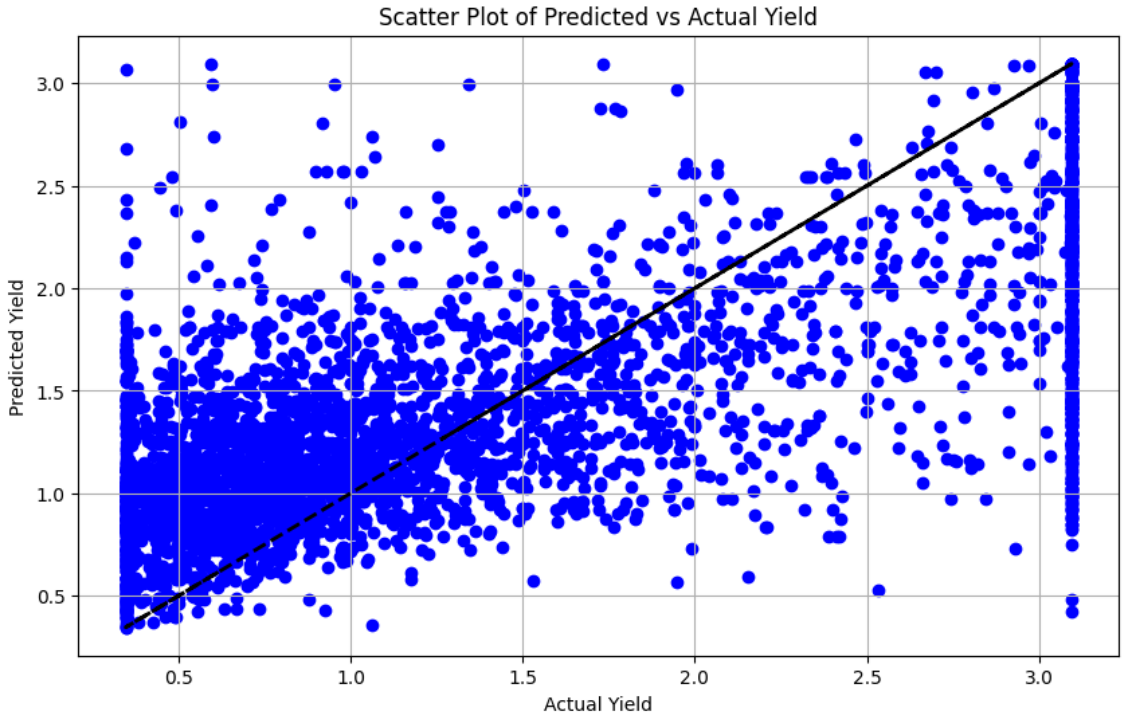

Figure 11: Scatter Plot of Predicted vs Actual Yield

6.3.2 Residual Plot

The residual plot of the best performing model of Random Forest Regressor is shown in Figure 12. The plot's shape implies a non-linear association between the independent and dependent variables, which likely contributes to the model's average R2 score as it may struggle to capture the complex relationship accurately.

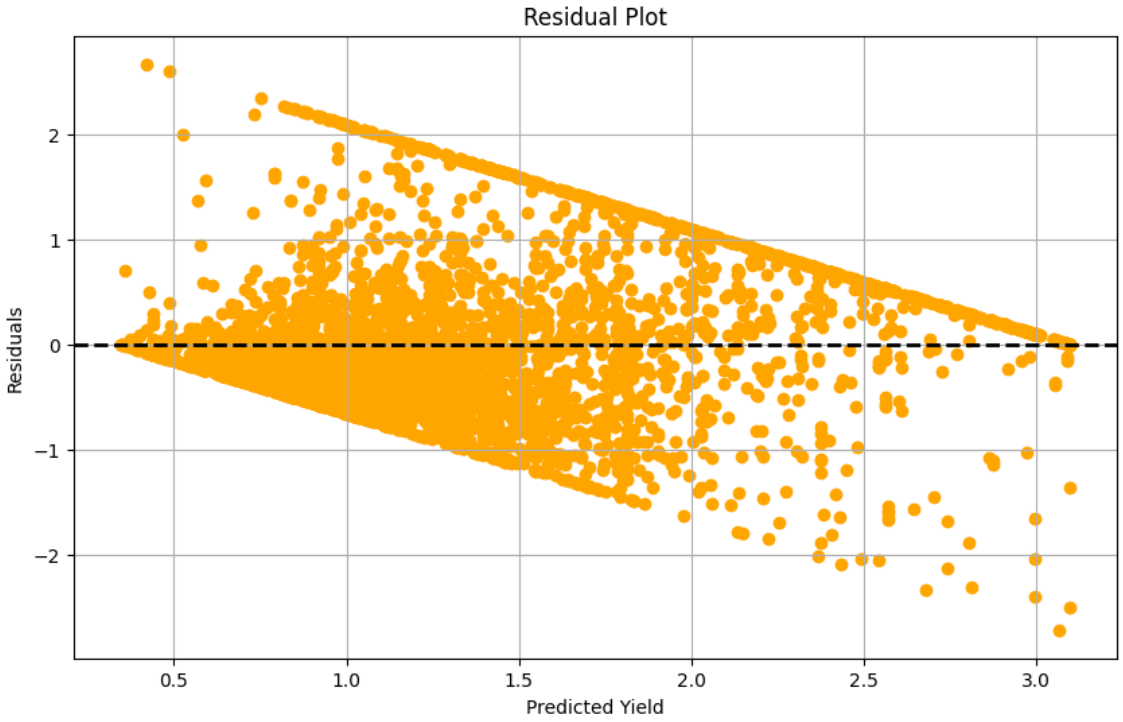

Figure 12: Residual Plot

6.3.3 Frequency Distribution of Residuals

The frequency distribution of residuals plot of the best performing model of Random Forest Regressor is shown in Figure 13. The data distribution of the residuals seems normally distributed, suggesting a symmetrical pattern around zero, implying the model's predictions exhibit minimal bias and effectively capture the variability in the data.

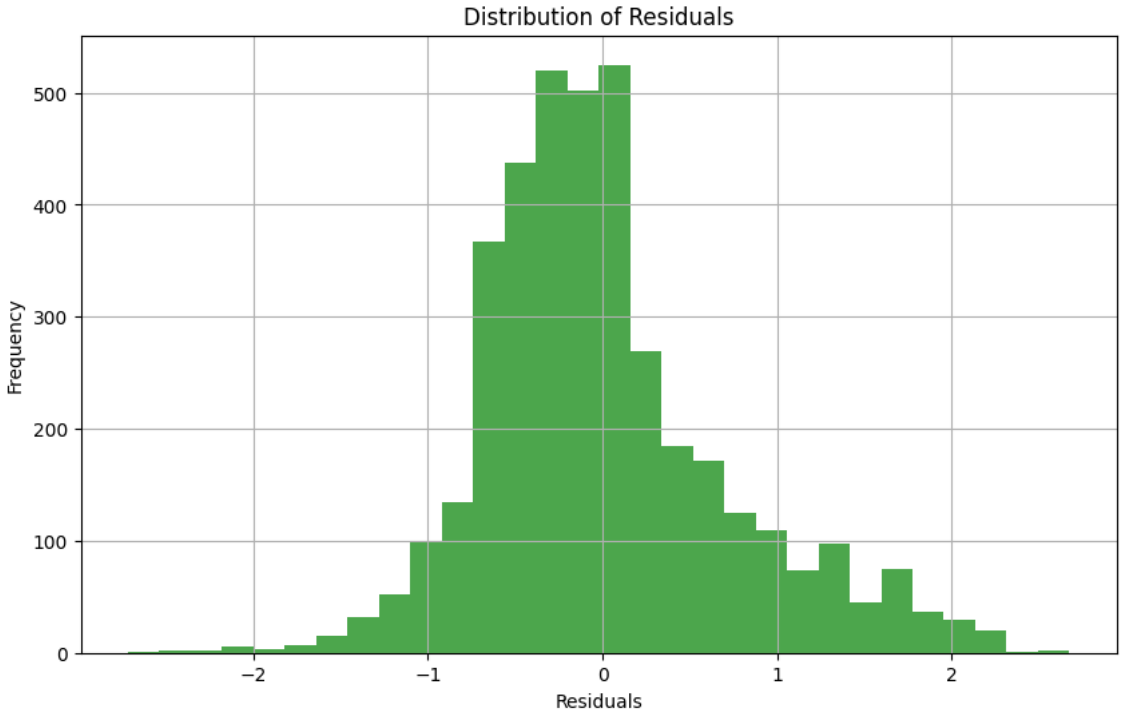

Figure 13: Frequency Distribution of Residuals

6.4 K-Means Clustering

Before applying clustering, correlation of Yield with other independent variables was calculated. This was done to select the top three variables which had the highest correlation with Yield. As there were around 10 variables, this step was necessary. Yield, Production and Season were picked as their correlation with the yield column was 1, 0.41, 0.31 respectively which was more than that of remaining columns. K-means clustering was run on the dataset for three values of k. The optimal value of k turned out to be 5. The Silhouette Score was around 0.7, the Within-Cluster Sum of Squares Error was 17459.67, and the Between-Cluster Sum of Squares Error was 161.64.

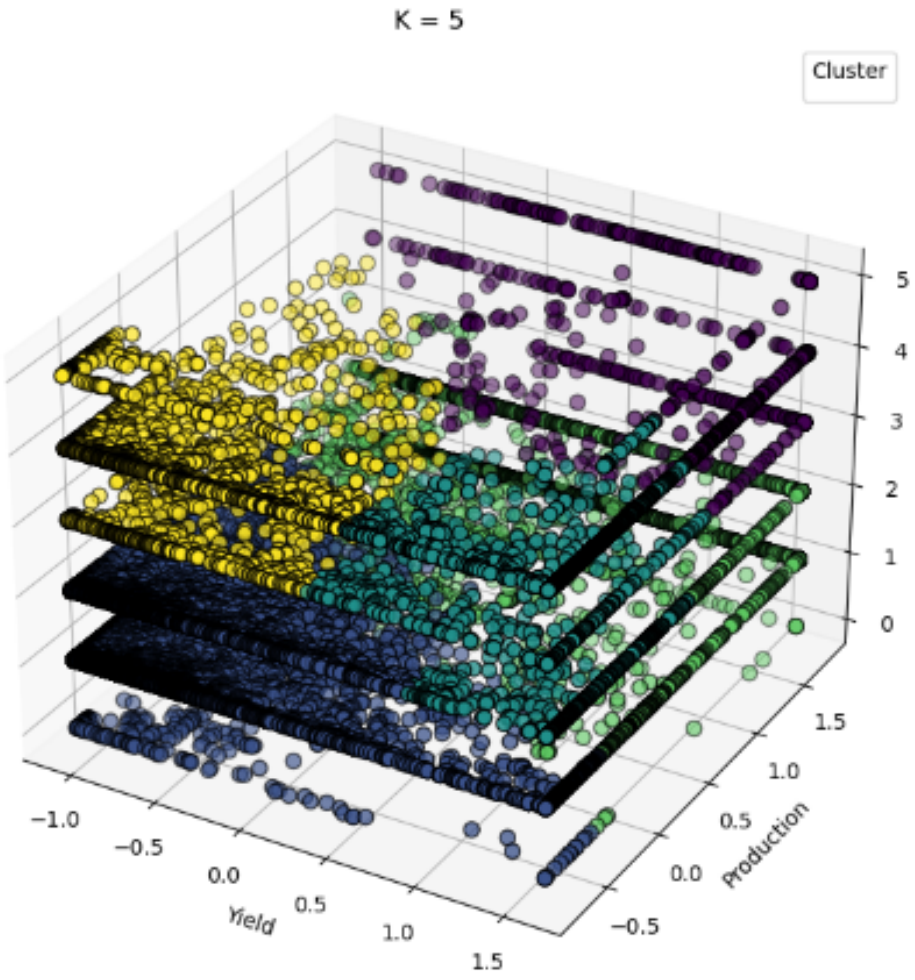

Figure 14: K-Means Clustering for K=5

6.4.1 Results of Clustering

The image below shows the cluster centroids for k=5.

```
Centroid for Cluster 1: [1.34893237 1.64358697 3.95892169]
Centroid for Cluster 2: [-0.50591819 -0.59375849  1.36450512]
Centroid for Cluster 3: [ 1.42497265 -0.41895242  3.33460986]
Centroid for Cluster 4: [0.41156477 1.54144321 1.30060841]
Centroid for Cluster 5: [-0.63492937 -0.53909821  3.81575311]
```

Figure 15: Cluster Centroids

- Cluster 1 has an average yield of 1.35 and average production of 1.64. The most frequent season in this cluster is 4 which is actually the 'Whole Year' season.
- Cluster 2 has an average yield of -0.51 and average production of -0.59. The most frequent season in this cluster is 1 which is actually the 'Kharif' season.
- Cluster 3 has an average yield of 1.42 and average production of -0.42. The most frequent season in this cluster is 3 which is actually the 'Summer' season.
- Cluster 4 has an average yield of 0.41 and average production of 1.54. The most frequent season in this cluster is 1 which is actually the 'Kharif' season.
- Cluster 5 has an average yield of -0.63 and average production of -0.54. The most frequent season in this cluster is 4 which is actually the 'Whole Year' season.

6.5 Google DataProc Deployment

After individually assessing each of the techniques and fine-tuning the models, we combined them and deployed the PySpark code to a cluster on Google DataProc. The successful results of running the job are visible in Figure 16.

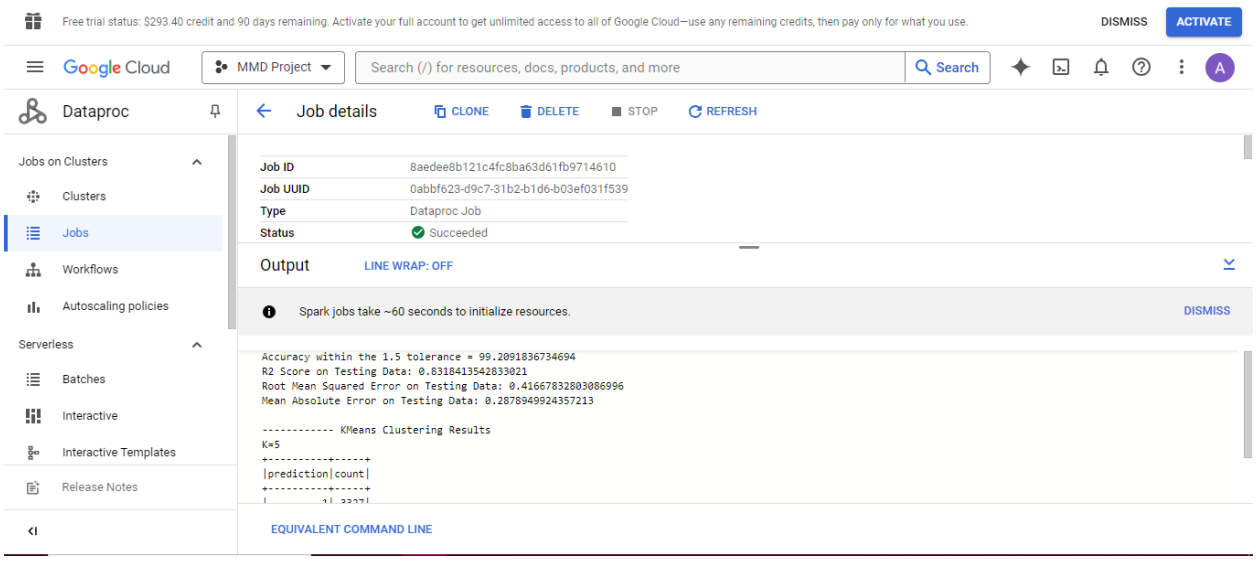

Figure 16: Deploying PySpark Code to Cluster on Google DataProc

7 Contributions

In this project we used a mixture of techniques and evaluated each one using different measures to ensure that each model's performance can be judged. Linear Regression, Gradient Boosted, and Random Forest Regressor code was written by ourselves using built in libraries. The K-Means clustering was implemented with the help of github codes. Three of the techniques were implemented by each member individually except K-Means Clustering was discussed and implemented together to ensure proper understanding.

8 Improvements and Future Work

There are a number of directions that future study and development in the field of agricultural production prediction can go. First, model performance can be enhanced by adding more agricultural variables to the feature engineering process, such as soil parameters, weather forecasts, and satellite imagery. Furthermore, investigating ensemble learning methodologies to merge forecasts from many machine learning models may result in improved prediction precision. Lastly, further insights can be gained by performing temporal and spatial studies to identify seasonal trends, cyclic patterns, and geographic factors impacting crop yield variability.

Conclusion

In conclusion, our research demonstrates the potential of PySpark and machine learning techniques in advancing agricultural analytics. By preprocessing data and applying various models like Linear Regression, Gradient Boosted Trees, Random Forest Regressor, and K-Means Clustering we've shown significant progress in predicting crop yields. However, while our findings are promising, there are still challenges to overcome, such as improving model interpretability and addressing data limitations.

References

[1] D. Bose *et al.*, "Big data analytics in agriculture," 2020.

[2] R. Priya, D. Ramesh, and E. Khosla, "Crop prediction on the region belts of india: a naïve bayes mapreduce precision agricultural model," in *2018 international conference on advances in computing, communications and informatics (ICACCI)*, pp. 99–104, IEEE, 2018.

[3] W. Fan, C. Chong, G. Xiaoling, Y. Hua, and W. Juyun, "Prediction of crop yield using big data," in *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, pp. 255–260, IEEE, 2015.

[4] M. Ramya, C. Balaji, and L. Girish, "Environment change prediction to adapt climate-smart agriculture using big data analytics," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 4, no. 5, pp. 1995–2000, 2015.

[5] A. K. Kushwaha and S. Bhattachrya, "Crop yield prediction using agro algorithm in hadoop," *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, vol. 5, no. 2, pp. 271–274, 2015.

[6] P. C. Reddy and A. S. Babu, "Survey on weather prediction using big data analystics," in *2017 Second international conference on electrical, computer and communication technologies (ICECCT)*, pp. 1–6, IEEE, 2017.

[7] R. A. Medar and V. S. Rajpurohit, "A survey on data mining techniques for crop yield prediction," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 2, no. 9, pp. 59–64, 2014.

[8] A. González Sánchez, J. Frausto Solís, W. Ojeda Bustamante, *et al.*, "Predictive ability of machine learning methods for massive crop yield prediction," 2014.

[9] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, and I. N. Athanasiadis, "Machine learning for large-scale crop yield forecasting," *Agricultural Systems*, vol. 187, p. 103016, 2021.

[10] R. Gupta, A. K. Sharma, O. Garg, K. Modi, S. Kasim, Z. Baharum, H. Mahdin, and S. A. Mostafa, "Wb-cpi: Weather based crop prediction in india using big data analytics," *IEEE access*, vol. 9, pp. 137869–137885, 2021.