**Crop Yield Prediction using Agricultural Data Clustering**

**Group Member Names:**

1. Manal Rizwan Qureshi (L200915) [Group Lead]
2. Muhammad Ibrahim (L202098)
3. Aisha Muhammad Nawaz (L200921)

**Mining of Massive Data Techniques:**

A range of data mining techniques will be utilized, with a primary emphasis on clustering algorithms like k-means and bisecting k-means. We want to find trends and correlations by classifying agricultural data points according to characteristics such as Crop\_Year, Fertilizer, Pesticide, and Annual\_Rainfall. This will help us forecast crop production more precisely. We'll also investigate dimensionality reduction techniques like PCA or UV Decomposition, whichever gives the best result to make the dataset simpler while maintaining essential information, which should help with visualization and possibly enhance clustering performance. We will use PySpark to integrate Apache Spark, allowing for efficient processing across distributed computing nodes, in order to handle the dataset's scale and complexity. In addition, we plan to employ Google Cloud Platform's Data Cluster to execute and deploy our PySpark code with ease. This will enable teamwork in development and provide scalable computing power to help us accomplish our research goals.

**Project Objective:**

We aim to properly pre-process, visualize and then use clustering techniques on agricultural data to transform the forecasting of crop yields. Our main goal is to create a prediction model that can correctly project agricultural output for future seasons and geographical areas. We aim to uncover important elements impacting crop output by evaluating extensive historical data covering crop types, seasons, states, cultivated area, production quantities, annual rainfall, fertilizer and pesticide consumption, and crop yield. By using big data approaches like clustering, we want to put similar agricultural data points together so that important insights may be extracted more easily. The expected result will be the development of an accurate crop yield prediction model that will enable farmers, agricultural officials, and stakeholders to make well-informed decisions about crop planning, risk management, and resource allocation. This will increase agricultural sustainability and productivity.

**Related Research Paper:**

**Title:** Analysis of Crop Yield Prediction using Data Mining Techniques

**Authors:** D Ramesh, B Vishnu Vardhan

**Summary:** This paper explores MLR and density-based clustering for crop yield prediction in East Godavari, AP, India. It collected historical agricultural data and evaluated MLR and clustering's effectiveness in predicting crop yields. Insights refine yield prediction models tailored to the region, aiding agricultural productivity.

**Relation to Project:**

This paper validates the project's focus on crop yield prediction using data mining techniques. It provides insights and methodologies applicable to the project.

**Dataset Details:**

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.

**Dataset Source:**

The agricultural dataset utilized in this research was sourced from Kaggle. Retrieved from: Gupta, A. (n.d.). Crop Yield in Indian States Dataset. Kaggle. Available at: [https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset.](https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset)

**References:**

[1] Vani, P. S., & Rathi, S. (2023). Improved data clustering methods and integrated A-FP algorithm for crop yield prediction. Distributed and Parallel Databases, 41(1), 117-131. <https://doi.org/10.1007/s10619-021-07350-1>

[2] Verma, A., Jatain, A., & Bajaj, S. (2018). Crop Yield Prediction of Wheat Using Fuzzy C Means Clustering and Neural Network. International Journal of Applied Engineering Research, 13(11), 9816-9821. Retrieved from<https://www.researchgate.net/profile/Aman-Jatain/publication/330555314_Crop_Yield_Prediction_of_Wheat_Using_Fuzzy_C_Means_Clustering_and_Neural_Network/links/5c480216a6fdccd6b5c193c4/Crop-Yield-Prediction-of-Wheat-Using-Fuzzy-C-Means-Clustering-and-Neural-Network.pdf>

[3] Vardhan, B. V., Ramesh, D., & Goud, O. S. C. (2014). Density Based Clustering Technique on Crop Yield Prediction. International Journal of Electronics and Electrical Engineering, 2(1), 56-59. Retrieved from <http://www.ijeee.net/uploadfile/2014/0214/20140214102624649.pdf>