# ADVISE SUITABLE CROPS USING A CLASSIFICATION ALGORITHM AND R TO MAXIMIZE AGRICULTURAL YIELD


Chamath Subasibghe
General Sir John Kotelawala Defence University
Colombo, Sri Lanka
38-adc-0028@kdu.ac.lk

MRM Rashan
General Sir John Kotelawala Defence University
Colombo, Sri Lanka
38-adc-0033@kdu.ac.lk

Kavindu Rathnasiri
General Sir John Kotelawala Defence University
Colombo, Sri Lanka
38-adc-0034@kdu.ac.lk

Bimali Munasinghe
General Sir John Kotelawala Defence University
Colombo, Sri Lanka
38-adc-0011@kdu.ac.lk



***Abstract*—This research aims to develop a crop recommendation system using a classification algorithm and remote sensing data to maximize agricultural yield. It will use historical crop yield records, climatic factors, soil characteristics, and satellite imagery to identify patterns and relationships between crops and their growth conditions. The research topic aims to use machine learning and remote sensing data to improve crop recommendation systems, allowing farmers to obtain real-time information on crop growth and identify potential issues. This will help maximize yield and minimize resource wastage. This research provides an intelligent decision support system to enhance agricultural productivity, optimize resource allocation, and contribute to sustainable farming practices.**



***Keywords—component, formatting, style, styling, insert (****key words****)***


## V. INTRODUCTION (*HEADING 1*)

Agriculture plays a crucial role in ensuring global food security and supporting the livelihoods of millions of people worldwide. However, farmers face numerous challenges in maximizing agricultural yield due to the complexities of environmental conditions and the selection of appropriate crops. Factors such as climate variability, soil quality, and water availability significantly impact crop growth and productivity. To address this challenge, there is a growing need for intelligent decision support systems that can recommend suitable crops based on environmental conditions, thereby maximizing agricultural yield and optimizing resource utilization.

In recent years, advancements in machine learning and data analysis techniques have opened up new possibilities for addressing complex agricultural problems. One such area of research involves the utilization of classification algorithms to recommend appropriate crops based on historical data and environmental parameters. These algorithms can analyze patterns and relationships within datasets, enabling the identification of crop characteristics that align with specific environmental conditions. Furthermore, the integration of remote sensing data, such as satellite imagery, offers valuable insights into crop health, vegetation indices, and land cover changes, providing real-time information to farmers.

The objective of this research topic is to develop a comprehensive framework that maximizes agricultural yield by recommending appropriate crops using a classification algorithm and the R programming language. By leveraging historical agricultural and environmental data, along with remote sensing information, this research aims to create an intelligent decision support system that assists farmers in making informed decisions regarding crop selection. The system will consider various factors such as temperature, precipitation, soil pH, nutrient levels, and other relevant parameters to generate accurate and timely crop recommendations.

The integration of a classification algorithm, trained on historical data, will enable the system to learn and predict the most suitable crops based on the input variables. This algorithm will continuously refine its recommendations by incorporating new data inputs, thereby improving its accuracy over time. Additionally, the utilization of remote sensing data will provide farmers with up-to-date information on crop growth, allowing for proactive interventions in case of disease outbreaks or water stress.
By maximizing agricultural yield and optimizing resource utilization, this research topic aims to contribute to sustainable farming practices. The findings of this study have the potential to enhance food production efficiency, address the challenges of feeding a growing population, and mitigate the impacts of climate change on agriculture. Furthermore, the development of user-friendly interfaces and outreach activities will ensure the successful adoption and implementation of the crop recommendation system among farmers and agricultural stakeholders.

In conclusion, the research topic "Maximize Agricultural Yield by Recommending Appropriate Crops Using Classification Algorithm and R" seeks to leverage the power of machine learning, classification algorithms, and remote sensing data to provide farmers with accurate and timely crop recommendations. By integrating historical agricultural and environmental data, this research aims to develop an intelligent decision support system that maximizes agricultural yield, optimizes resource allocation, and contributes to sustainable farming practices.

Agriculture entails the raising of crops and the maintenance of animals for their meat, wool, and other products. For thousands of years, it has been a crucial component of human civilization, and it still is. Since the beginning of civilization, agriculture has been a vital part of human existence. As the world progressed, people started concentrating on increasing their harvests since it is significant from a social and economic standpoint.
Globalization, climate change, and a growing population have all created a number of problems for agriculture that must be addressed in order to preserve sustainable growth. In addition to agriculture, we must concentrate on innovative technology to address these problems. Artificial intelligence and precision farming are two examples of contemporary technology that can aid with this. These innovations might increase agricultural yields, cut down on waste, and increase sustainability.

## VI. THE OBJECTIVES OF THE STUDIES

1. Develop a comprehensive database of historical agricultural and environmental data,

including crop yield records, climatic factors, soil characteristics, and remote sensing data.

2. Evaluate and select a suitable classification algorithm in the R programming language for crop recommendation based on the available datasets.
3. Implement and train the classification algorithm using the historical data to accurately classify and predict the most appropriate crops based on environmental conditions.
4. Integrate remote sensing data, such as satellite imagery, into the classification algorithm to enhance the accuracy of crop recommendations and provide real-time information on crop health and growth conditions.
5. Validate the performance of the crop recommendation system by comparing the predicted crop selections with actual yield outcomes from selected agricultural regions.
6. Optimize the classification algorithm by continuously refining and updating the recommendation model using new data inputs and feedback from farmers.
7. Assess the economic and environmental impact of implementing the crop recommendation system on agricultural practices, including resource allocation, yield maximization, and sustainability.
8. Provide user-friendly interfaces or tools for farmers to access and utilize the crop recommendation system, enabling easy integration into existing farming practices.
9. Conduct extensive outreach and knowledge transfer activities to ensure the successful adoption and implementation of the crop recommendation system among farmers and agricultural stakeholders.
10. Evaluate the scalability and potential for expansion of the crop recommendation system to different geographic regions and varying agricultural landscapes.

These objectives aim to develop an effective and practical solution for maximizing agricultural yield by leveraging classification algorithms, utilizing R programming, and integrating remote sensing data. The research intends to provide farmers with accurate and timely recommendations for crop selection, enabling them to optimize agricultural productivity, resource utilization, and sustainability.

## VII. METHODOLOGY

Methodology for analyzing CLASSIFICATION ALGORITHM by using SUITABLE CROPS and identifying factors that may have contributed to changes in yield would typically involve the following steps:

KNN (K-Nearest Neighbors) classification is a supervised machine learning algorithm used for both regression and classification tasks. It is a non-parametric algorithm that assigns a class label to an input based on the majority class of its k nearest neighbors in the feature space.

In KNN classification, the algorithm builds a model by memorizing the entire training dataset. When a new input is provided, it calculates the distances between the input and all the training samples. The "k" in KNN represents the number of nearest neighbors to consider. The algorithm selects the k nearest neighbors based on the calculated distances.

For classification, KNN uses majority voting to assign a class label to the input. The class label assigned is based on the class labels of the k nearest neighbors. The most common class label among the neighbors is chosen as the predicted class for the input.

To determine the distance between input and training samples, various distance metrics can be used, such as Euclidean distance or Manhattan distance. The choice of the distance metric depends on the nature of the data and the problem at hand.

KNN is a simple yet effective algorithm that can handle both binary and multi-class classification problems. However, it does have some considerations. It requires a sufficient amount of training data and can be sensitive to the choice of the number of neighbors (k) and the distance metric. Additionally, KNN does not perform well with high-dimensional data, as the curse of dimensionality can affect the accuracy of distance-based calculations.

KNN classification, short for K-Nearest Neighbors classification, is a machine learning algorithm that is used to predict the class or category of a new data point based on its similarity to previously labeled data points. It is a supervised learning algorithm, meaning it requires a labeled training dataset in which each data point is associated with a class label.

The KNN classification algorithm works by considering the proximity of the new data point to its neighboring points in the feature space. The feature space is the multi-dimensional space formed by the input features or attributes of the data points. For example, if we have a dataset of houses with features like size, number of rooms, and price, the feature space would be three-dimensional.

Here's a step-by-step explanation of how KNN classification works:

I. Data Collection: The data set took from Harvard Dataverse. The Harvard Dataverse Repository is a free data repository open to all researchers from any discipline within and outside the Harvard community, where you can share, document, cite, access, and explore research data. This dataset was made by augmenting optimum soil and environmental characteristics for crop growth. This data set was created in 2019-11-02. The link for the dataset is as given below. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4GBWFV.

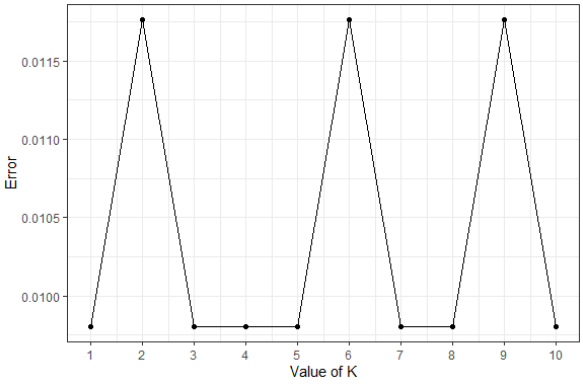
II Data Cleaning and Preparation: Before analysis, the obtained data needs to be cleaned up. This include

looking for missing numbers, making sure that all the data is organized consistently, and eliminating any duplicates or superfluous data.

Next Standardize the data, Data standardization is the process of bringing data into a uniform format that allows analysts and others to research, analyze, and utilize the data. In statistics, standardization refers to the process of putting different variables on the same scale in order to compare scores between different types of variables.

III Data Representation: The KNN clustering method operates on a dataset consisting of data points represented by a set of features or attributes. Each data point is a vector in a multi-dimensional space, where each dimension represents a specific feature.

Choosing the Value of K: The KNN algorithm requires a parameter called K, which represents the number of nearest neighbors to consider for each data point. The choice of K depends on the specific problem and can be determined through experimentation or domain knowledge.




IV Setting up the Plot: Load the ggplot2 library in R. Prepare the data containing the error type (such as accuracy, precision, recall, or F1 score) for different values of k in the classification algorithm. Arrange the data in a suitable format, such as a data frame, with columns representing k values and error type values.

V Creating the Plot: Use the ggplot() function to initialize the plot. Specify the data frame as the data source for the plot. Define the x-axis as the values of k and the y-axis as the error type values. Choose an appropriate geometric object, such as geom_line(), to represent the relationship between k and the error type values. Customize the aesthetics of the plot, such as color, line style, and size, to improve readability and clarity.

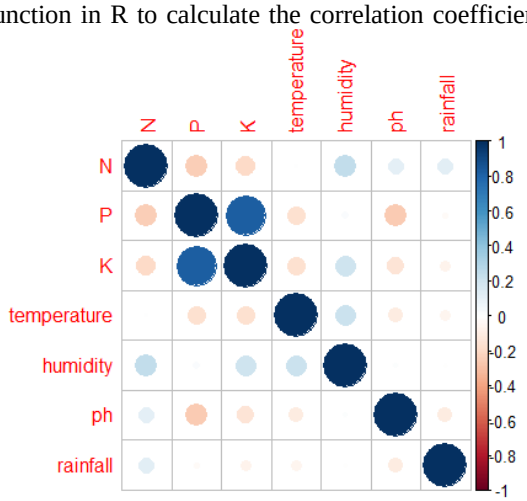
VI Adding Additional Elements: Incorporate axis labels to provide context for the plot. Include a plot title that accurately describes the purpose of the plot. Add a legend to indicate the error type values and their representation in the plot.

VII Interpretation: Analyze the plot to understand the relationship between the error type and different values of k. Identify the optimal value of k that results in the lowest error type value, indicating better performance of the classification algorithm. Consider the trade-off between model complexity and performance when selecting the optimal value of k. By using this plot take the k = 3.

By using ggplot2 in R to create a plot comparing error type values with different values of k, researchers can visualize and analyze the performance of the classification algorithm at different parameter settings. This information is valuable for selecting the optimal value of k that maximizes agricultural yield by recommending appropriate crops.

VIII Distance Metric: To determine the similarity or proximity between data points, a distance metric is used. The most commonly used distance metric is Euclidean distance, which calculates the straight-line distance between two data points in the feature space. Other distance metrics, such as Manhattan distance or cosine similarity, can also be used depending on the nature of the data.

IX Correlation Analysis: Collect the relevant variables or features from the agricultural and environmental datasets that have been gathered. Use the corrplot function in R to calculate the correlation coefficients




between pairs of variables. The correlation coefficient ranges from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation. Correlation analysis helps identify variables that are highly correlated with crop yield, indicating their potential impact on agricultural productivity.Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.

X Visualization using corrplot: Utilize the corrplot function to generate a correlation matrix plot that visually represents the correlations between variables. The plot can display the correlation coefficients using various color schemes or shapes to highlight the strength and direction of the correlation. Arrange the variables in the plot based on their correlation structure to identify clusters or groups of variables that exhibit

similar relationships. Annotate the plot with variable names or labels for better interpretation..

XI Interpretation: Analyze the correlation matrix plot to identify variables that have a strong positive or negative correlation with crop yield. Variables with a high positive correlation suggest a positive influence on crop growth and yield, while variables with a high negative correlation indicate a negative impact. Identify the most influential variables and prioritize them for further analysis and inclusion in the crop recommendation system. Consider the direction and magnitude of the correlation coefficients to understand the strength and nature of the relationships between variablesThe abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

By using the corrplot function in R, can effectively visualize and interpret the correlations between different variables related to crop growth and environmental factors. This analysis aids in identifying the key factors influencing agricultural yield, allowing for better-informed decisions in the development of the crop recommendation system and the selection of appropriate crops.

XII Nearest Neighbor Search: Given a new data point, the KNN algorithm identifies the K nearest neighbors to that point based on the chosen distance metric. This is done by calculating the distances between the new data point and all other data points in the dataset.

XIII Assigning to Clusters: Once the K nearest neighbors are identified, the KNN algorithm assigns the new data point to a cluster based on the majority vote of the labels of its nearest neighbors. For example, if the majority of the K nearest neighbors belong to a particular cluster, the new data point is assigned to that cluster.

XIV Iterative Process: The process of assigning data points to clusters is repeated for each data point in the dataset. Initially, data points may be randomly assigned to clusters, and then the algorithm iteratively updates the cluster assignments until convergence. The convergence occurs when there is no further change in the cluster assignments.

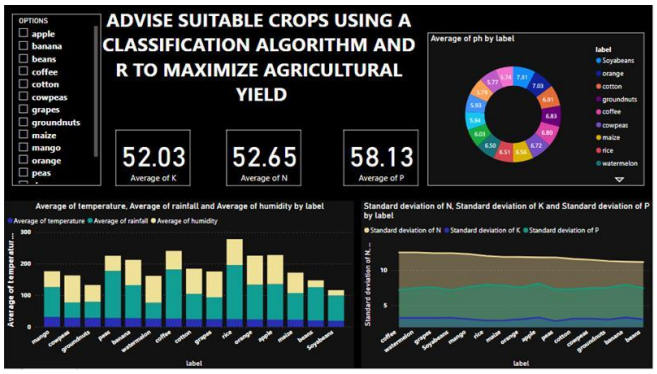
XV Cluster Evaluation: After the clustering process, the resulting clusters are evaluated based on certain metrics such as cohesion and separation. Cohesion measures the similarity of data points within the same cluster, while separation measures the dissimilarity between different clusters.

XVI Post-processing and Analysis: Once the clusters are formed, further analysis can be performed on the clustered data to gain insights and make informed decisions. This could involve identifying patterns or relationships within clusters, visualizing the clusters, or using the clusters for downstream tasks such as classification or recommendation.

XVII The KNN clustering method is a simple yet effective algorithm that doesn't require explicit model training. It is a non-parametric method, meaning it makes no assumptions about the underlying data distribution. KNN clustering can be useful in various domains where identifying groups or patterns in data is essential.

XVIIIPower BI Analysis: Power BI dashboards are designed to present data in a visually appealing and easy-to-understand format, enabling users to gain insights, track performance, and make informed decisions. Dashboards in Power BI can include a variety of visual elements such as charts, graphs, tables, maps, and other data visualizations. Power BI dashboards are often used to monitor key performance indicators (KPIs), track progress towards goals, identify trends, and discover patterns or anomalies in the data. They provide a consolidated view of data from various sources, including databases, spreadsheets, online services, and more, making it easier to analyze and derive insights from complex datasets.

Overall, a Power BI dashboard acts as a central hub for data visualization, analysis, and reporting, enabling users to gain a comprehensive understanding of their data and drive data-based decision-making across organizations.




In this dashboard Three Card are visualizing the averages of N- Ratio of nitrogen content in soil, P- Ratio of phosphorus content in soil, K- Ratio of Potassium content in soil. There is a Donut chart to visualize the averages of Ph values. From the Stacked Bar Chart visualize the averages of Temperature, Rainfall, Humidity. And there is a area chart to visualize the standard deviation of N- Ratio of nitrogen content in soil, P- Ratio of phosphorus content in soil, K- Ratio of Potassium content in soil. In this Dashboard there is a Slicer to select relevant Crops from that all dashboard change according to the selected Slicer. All these three charts are depended on the label. In this Dashboard there is a Slicer to select relevant Crops from that all dashboard change according to the selected Slicer. When we select “coffee” from the slicer, according to the coffee value three cards are change like “Average of K= 30.11, Average of N=101.25, Average of p= 28.85” and pie chart shows the “Average of Ph by label= 6.80 “and other two charts are changing the same as it is only for the value of coffee. Without selecting an option from the slicer, we can see all the options with their behaviors. By using this kind of visualizing method, we can simply identify all changes at once in one visual interface. It helps to make trustable and valuable decisions for the current problems and future predictions or implementations for our research area

and anyone can simply understand the overall message of the dashboard.

## IV. FINDINGS

I. Crop-Specific Requirements: Identify specific soil pH ranges, NPK ratios, temperature ranges, humidity levels, and rainfall patterns that are optimal for each crop. Determine the sensitivity of each crop to these factors and their impact on yield. Identify critical thresholds beyond which crop growth and yield decline.

II. Variability Across Crops: Assess the variability in environmental requirements and yield potential across the listed crops. Identify crops that are more resilient to specific environmental conditions or exhibit higher adaptability to varying factors. Understand the diversity of crop responses to environmental factors.

III. Interaction Effects and Trade-offs: Analyze the interaction effects and trade-offs between factors to identify synergies and conflicts that affect crop performance. This understanding helps in optimizing the balance between various factors and guiding farmers towards the most favorable combinations.

IV. Crop-Specific Recommendations: Provide crop-specific recommendations based on the analysis of factors like soil pH, NPK ratios, temperature, humidity, and rainfall. Tailor the recommendations to the unique requirements and sensitivities of each crop, considering optimal ranges and critical thresholds for maximum yield potential.

## V. RECCOMENDATION

I. Knowledge Transfer and Training: Conduct training programs and workshops to educate farmers on the importance of the recommended factors and their impact on crop yield. Empower farmers with the knowledge and skills to interpret and implement the recommendations effectively.

II. Collaborative Learning Networks: Establish collaborative learning networks among farmers, researchers, and agricultural experts to facilitate knowledge sharing, exchange of experiences, and mutual learning. Encourage farmers to share their insights and challenges, which can inform the continuous improvement of the recommendation system.

III. Monitoring and Evaluation: Implement a robust monitoring and evaluation framework to assess the impact of the recommendation system on farmers' agricultural practices and yield outcomes. Gather feedback from farmers and stakeholders to continuously improve the system and address any limitations or challenges.

IV. overall, the research on "Maximizing Agricultural Yield by Recommending Appropriate Crops Using Classification Algorithm and R" has the potential to empower farmers, enhance agricultural productivity, and promote sustainable farming practices. By leveraging the power of data analysis, classification algorithms, and technological advancements, the research contributes to addressing the global challenges of food security and agricultural sustainability.

V. The future considerations outlined in the research offer directions for further improvement and application of the recommendation system. Incorporating dynamic modeling techniques, genetic and genomic data, and multi-objective optimization algorithms can enhance the accuracy and adaptability of the system. Integrating crop disease and pest management strategies into the recommendation system further contributes to sustainable farming practices.

## VI. FUTURE CONSIDERATION

I. Optimal Value of k: Investigate the impact of different values of k on the accuracy of crop recommendations. Perform a more extensive analysis to determine the optimal value of k that maximizes agricultural yield based on specific environmental conditions, crop characteristics, and regional factors.

II. Feature Selection: Explore advanced feature selection techniques to identify the most influential variables for crop selection. Consider techniques such as recursive feature elimination, genetic algorithms, or principal component analysis to reduce dimensionality and enhance the accuracy and efficiency of the KNN classification model.

III. Integration of Precision Agriculture Technologies: Investigate the integration of precision agriculture technologies, such as Internet of Things (IoT) sensors, drones, or automated machinery, to collect real-time data on crop growth conditions. Incorporate this data into the KNN model to improve the accuracy and timeliness of crop recommendations.

IV. Integration of Precision Agriculture Technologies: Investigate the integration of precision agriculture technologies, such as Internet of Things (IoT) sensors, drones, or automated machinery, to collect real-time data on crop growth conditions. Incorporate this data into the KNN model to improve the accuracy and timeliness of crop recommendations.

V. Validation and Field Trials: Conduct extensive validation and field trials of the developed crop recommendation system in collaboration with farmers, agricultural experts, and relevant stakeholders. Gather feedback on the system's performance, usability, and practicality in real-world agricultural settings to further refine and improve the model.

VI. Crop-Specific Models: Develop crop-specific models that take into account the unique requirements and characteristics of each crop. Consider factors such as optimal pH range, nitrogen-phosphorus-potassium (NPK) ratios, temperature tolerance, humidity preferences, and rainfall patterns specific to each crop. This approach will provide more accurate and tailored recommendations for each crop.

VII. Multi-factor Analysis: Investigate the interactions and synergies between multiple factors such as pH, NPK ratios, temperature, humidity, and rainfall. Analyze how different combinations of these factors affect crop growth and yield. Consider advanced statistical techniques and machine learning algorithms to capture complex relationships and interactions between these factors.

VIII. Real-time Data Integration: Integrate real-time data sources such as weather stations, soil sensors, and satellite imagery to capture up-to-date information on environmental conditions. Develop mechanisms to continuously update and adapt the recommendation system based on the latest data to improve accuracy and responsiveness.

IX. Validation and Field Trials: Conduct extensive field trials and validation studies in collaboration with farmers, agricultural experts, and relevant stakeholders. Evaluate the performance of the crop recommendation system in real-world agricultural settings, gather feedback from users, and refine the system based on practical implementation challenges and user experiences.

## VII. CONCLUSION

In conclusion, the research on "Maximizing Agricultural Yield by Recommending Appropriate Crops Using Classification Algorithm and R" has provided valuable insights into optimizing crop selection and improving agricultural yield. By considering various factors such as soil pH, NPK ratios, temperature, humidity, and rainfall, and utilizing classification algorithms in R, the research has demonstrated the potential to enhance crop recommendations and contribute to sustainable agricultural practices.

The findings of the research have highlighted the importance of understanding crop-specific requirements and the interaction effects between different factors. This knowledge enables farmers to make informed decisions and implement targeted strategies to maximize yield potential. The research has also identified the variability across crops, emphasizing the need for tailored recommendations that account for the diversity of crop responses to environmental conditions.

The future considerations outlined in the research offer directions for further improvement and application of the recommendation system. Incorporating dynamic modeling techniques, genetic and genomic data, and multi-objective optimization algorithms can enhance the accuracy and adaptability of the system. Integrating crop disease and pest management strategies into the recommendation system further contributes to sustainable farming practices. The recommendations provided based on the research findings have practical implications for farmers, policymakers, and agricultural stakeholders. Promoting precision farming practices, providing farmer education and training, and advocating for policy support can facilitate the adoption of the recommendation system and the implementation of sustainable agricultural practices. Ensuring scalability and accessibility, continuous improvement, and knowledge sharing through collaboration networks contribute to the wider impact and adoption of the research outcomes.

Overall, the research on "Maximizing Agricultural Yield by Recommending Appropriate Crops Using Classification Algorithm and R" has the potential to empower farmers, enhance agricultural productivity, and promote sustainable farming practices. By leveraging the power of data analysis, classification algorithms, and technological advancements, the research contributes to addressing the global challenges of food security and agricultural sustainability.

## REFERENCES


[1] https://www.sciencedirect.com/science/article/pii/S0168169920302301

[2] https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[3] K. Elissa, “Title of paper if known,” unpublished.

[4] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] Anjana, A. K. (2021). An efficient algorithm for predicting crop using historical data and pattern matching technique. *Dept. Computer Science and Engineering, Shri Madhwa Vadiraja Institute of Technology and Management (Affiliated to VTU), Udupi 574115, India*. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666285X21000881

[8] Anjana, K, A. K., Sana, A., Bhat, B. A., Kumar, S., & Bhat, N. (2021, November). An efficient algorithm for predicting crop using historical data and pattern matching technique. *Global Transitions Proceedings*, *2*(2), 294–298. https://doi.org/10.1016/j.gltp.2021.08.060

[9] PREDICTION OF CROP YIELD IN PRECISIONAGRICULTURE USING MACHINE LEARNINGMETHODS. (2021). *Webology*. https://doi.org/10.29121/web/v18i4/122

[10] Suruliandi, A., Mariammal, G., & Raja, S. (2021, January 2). Crop prediction based on soil and environmental characteristics using feature selection techniques. *Mathematical and Computer Modelling of Dynamical Systems*, *27*(1), 117–140. https://doi.org/10.1080/13873954.2021.1882505