iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.51282

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Crop Prediction using PySpark

Dr.N.Usha Rani[1], Puvvadi Kavya[2], Shaik Afroze Sulthana[3], A.V Mahvith[4], Shaik Sohail[5], Ale Anil[6]
[1, 2, 3, 4, 5, 6]*Department of CSE, SVU College of Engineering*

Abstract: ***Agriculture is a crucial sector for many economies worldwide. The success of agriculture is largely dependent on the crops grown, which in turn rely on various factors such as soil properties. To aid farmers in making informed decisions about crop selection based on soil properties, crop prediction systems have been developed. This work presents a crop prediction system that utilizes soil properties and PySpark. The main aim of this work is to forecast the most suitable crop to be grown in a given area based on its soil properties. The system takes input data of different soil parameters, such as pH, Nitrogen, etc., and uses machine learning algorithm to create a model for crop prediction. PySpark is used to manage large-scale soil data. The work's results demonstrate the effectiveness of PySpark in handling and analyzing large-scale data for crop prediction. The accuracy and performance of the system can be further improved by incorporating additional features and refining the Random forest algorithm is used in the model.***

Keywords: Crop, soil parameters, PySpark, machine-learning algorithm, Random forest

I. INTRODUCTION

Agriculture is a vital sector of the Indian economy, employing more than half of the country's workforce and contributing significantly to its GDP. India is the world's second-largest producer of food after China and is one of the world's top producers of crops such as rice, wheat, sugarcane, cotton, and pulses.

The agriculture of India faces numerous challenges, including low productivity, lack of modern technology, inadequate irrigation facilities, and weather-related risks such as droughts and floods. Small-scale farmers also face challenges such as limited access to credit and market information, which can hinder their ability to improve their livelihoods.

The objective of the "Crop Prediction Using Soil Properties" work is to determine the most appropriate crop for a particular soil type by using PySpark, a fast and scalable data processing engine. Agriculture is a vital sector for many countries, and the success of crop growth depends on the quality of soil. The work involves collecting data on soil properties and crop types from various sources, which is then cleaned and preprocessed. The cleaned data is utilized to train a machine learning model using PySpark's MLlib library, which predicts the most suitable crop for a given set of soil properties. The work also includes the development of a user-friendly interface using Flask, a micro web framework in Python. The web application will be hosted on a cloud-based platform to ensure accessibility and scalability.

Several existing systems use different methods such as machine learning, statistical modeling, and remote sensing to forecast crop yields. Examples of popular systems include the Crop Yield Forecasting System (CYFS), MARS Crop Yield Forecasting System, Yield Prophet, and US Crop Explorer. These systems use various data such as remote sensing, weather, soil, and crop growth models to provide valuable insights into crop yield predictions, which can aid farmers, policymakers, and commodity traders in decision-making related to crop production and marketing. However, it is essential to note that crop yield predictions may not always be accurate and are affected by various factors such as weather, soil health, and management practices.

II. LITERATURE SURVEY

In paper [1] machine learning techniques for predicting crop yields. They highlighted the potential of these models to improve accuracy and identified areas for future research. The paper emphasized the importance of accurate crop yield prediction and explored various types of machine learning models used in this domain, including regression models, decision trees, support vector machines, artificial neural networks, and deep learning models. The authors provided an overview of each model's advantages and limitations and emphasized the importance of selecting appropriate features and pre-processing techniques. They also discussed the need for developing more interpretable models in agriculture. The authors identified challenges associated with crop yield prediction, such as data availability and quality, and suggested using remote sensing data, such as satellite imagery, to improve accuracy. The review concluded with suggestions for future research, such as exploring more advanced machine learning techniques and incorporating more diverse data sources like weather and socio-economic data.

The review provides valuable insights for researchers working in crop yield prediction and highlights the potential for further development and improvement of machine learning models in this field.

A system [2] extensively review the role of soil information in improving crop yield prediction accuracy. The authors emphasize the significance of incorporating soil properties into prediction models and discuss the advantages and limitations of various soil variables. They also address the challenges associated with soil data collection, quality, and spatial-temporal variability. Additionally, the paper explores the potential of using remote sensing data and machine learning techniques to further enhance crop yield prediction accuracy. Overall, the authors' insights provide valuable information on the current state of research on crop yield prediction models that are based on soil information and identify future research directions for developing more accurate and efficient prediction models. Such models can significantly contribute to improving agricultural production, food security, and sustainability efforts.

A systematic review to provide an all-inclusive summary of the current research state regarding the use of machine learning (ML) techniques for crop yield prediction [3]. The paper highlights the advantages and limitations of ML models and identifies the most widely used techniques, including regression-based models, decision trees, random forests, support vector machines, and neural networks. Additionally, the review discusses commonly used features for crop yield prediction, such as climate data, soil data, remote sensing data, and management practices. The authors emphasize the significance of feature selection and engineering to improve model accuracy. The review also identifies gaps in the literature and recommends future research directions, such as the inclusion of diverse data sources, enhancing data quality, and developing more interpretable models. The paper offers a valuable resource for researchers and practitioners interested in utilizing ML techniques to improve crop yield prediction accuracy, emphasizing the potential of these models to support decision-making and enhance crop management practices.

A model built, which focuses on crop yield prediction models that utilize soil properties [4]. The authors emphasize the importance of soil properties in predicting crop yield and discuss the various physical, chemical, and biological soil properties that have been used for this purpose. They also address the challenges associated with using soil properties, including data availability and quality, and suggest the potential of using remote sensing data and machine learning techniques to improve prediction accuracy. The paper further provides an overview of various machine learning techniques utilized for crop yield prediction, including linear regression, decision trees, random forests, and support vector machines. The authors emphasize the significance of selecting appropriate features, pre-processing techniques, and model selection for improving the accuracy of crop yield prediction models. The review concludes by identifying areas for future research, such as the development of more efficient methods for measuring soil properties and incorporating more diverse data sources. The paper serves as a valuable resource for researchers and practitioners in the field of crop yield prediction and identifies opportunities for further development and improvement of prediction models.

III. EXISTING SYSTEM

The conventional approach for crop prediction involves using soil samples and historical weather data, but it has limitations in terms of accuracy and scalability. Machine learning provides a more sophisticated approach to crop prediction, which can analyze data in real-time, providing more accurate predictions and quick responses to changes in soil conditions and weather patterns. One popular machine learning algorithm for crop yield prediction is the Random Forest algorithm, which creates an ensemble of decision trees to make a prediction about crop yield.

Several existing systems use different techniques such as machine learning, statistical modeling, and remote sensing for crop yield prediction. Some popular systems include the Crop Yield Forecasting System [5] (CYFS), MARS Crop Yield Forecasting System, Yield Prophet, and US Crop Explorer. These systems use a combination of remote sensing data, weather data, and crop growth models to forecast crop yields for various crops. The predictions provided by these systems can be useful for farmers, policymakers, and commodity traders to make informed decisions related to crop production and marketing. However, it is important to keep in mind that crop yield predictions are influenced by various factors and may not always be accurate.

IV. PREDICTION OF CROP USING PYSPARK

The proposed system indeed has the potential to be an efficient and user-friendly platform for farmers to predict crop yields based on soil data. The use of Hadoop and PySpark [6] can help in processing and analyzing large amounts of data in a scalable and distributed manner, while the website interface can make it easy for farmers to access and use the system.

Additionally, the incorporation of historical weather data and crop growth patterns into the analysis can provide valuable insights into crop yield predictions, which can help farmers optimize planting schedules and inform decisions about fertilizer application, leading to better crop production and reduced waste.

The system can also contribute to the development of precision agriculture, which aims to maximize the efficiency and productivity of crop production while minimizing environmental impact.

However, it is important to note that the accuracy of crop yield predictions is influenced by various factors, such as weather conditions, soil health, and management practices, and may not always be accurate. Therefore, farmers should use the predictions as a guide and not rely solely on them when making decisions related to crop production.

A. *System Design*

Algorithm written below illustrated the step-by-step approach for the crop prediction.

1) *Data collection:* Collect data on soil properties, such as pH, nitrogen content, moisture, texture, etc. Collect data on crop yield from past growing seasons or through ongoing monitoring.
2) *Data preprocessing:* Clean and preprocess the data by removing any missing or incorrect values and transforming it into a format that can be used by machine learning algorithms.
3) *Feature selection:* Identify the most important features that influence crop yield based on statistical analysis or domain knowledge.
4) *Model selection:* Choose a machine learning algorithm that is suitable for the problem at hand, such as regression, decision trees, or neural networks.
5) *Model training:* Use the preprocessed data to train the machine learning model. Split the data into training and testing sets and use techniques such as cross-validation to evaluate the performance of the model.
6) *Deployment:* Deploy the model in a web application, which takes input from users about the soil properties and predicts the crop yield output.

B. *Implementation Process*

Implementation of the work where the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, Investigation of the existing system and it's constraints on implementation, design of methods to achieve changeover and evaluation of changeover methods.

C. *Hadoop Framework*

Hadoop is an open-source software framework used for distributed storage and processing of large data sets across clusters of computers. Hadoop also supports various tools and technologies that enhance its functionality. For example, Apache Pig is a high-level scripting language that simplifies the processing of large data sets, while Apache Hive is a data warehousing tool that allows for querying and analysis of data stored in Hadoop.

Other technologies that can be used in conjunction with Hadoop include Apache Spark, which provides faster processing speeds than MapReduce, and Apache Storm, which is used for real-time data processing. Hadoop can also integrate with other data storage and processing technologies, such as Apache Cassandra and Apache Kafka.

Overall, Hadoop is a powerful tool for big data processing and storage, and its flexibility and scalability make it a popular choice for organizations dealing with large data sets.

D. *Flask Framework:*

Flask is a micro web framework written by using python language. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where a pre-existing third party library provides common functions. However, flask supports extensions that can add application features as if they were implemented in flask itself. Extensions exist for object relational mapper, form validation, uploading files handing various open authentication technologies.

E. *Random Forest model*

Random Forest is a machine learning algorithm that is implemented in PySpark, the Python API for Apache Spark. PySpark provides a distributed computing framework that enables processing of large datasets in parallel across a cluster of computers.

The Random Forest algorithm in PySpark works by creating an ensemble of decision trees, each of which is trained on a subset of the input data. Each decision tree makes a prediction about the output variable based on the input features, and the final prediction is made by combining the predictions of all the decision trees.

PySpark's implementation of Random Forest provides several tuning parameters that can be adjusted to optimize the algorithm's performance. These parameters include the number of decision trees to include in the ensemble, the maximum depth of the decision trees, and the number of input features to consider at each split. To use the Random Forest algorithm in PySpark, the input data must be in the form of a PySpark DataFrame, which is a distributed collection of data organized into named columns. The input DataFrame must have a column containing the output variable that is to be predicted, as well as columns containing the input features.

The PySpark Random Forest algorithm can be trained using the RandomForestRegressor or RandomForestClassifier classes, depending on whether the output variable is continuous or categorical, respectively. Once the model is trained, it can be used to make predictions on new input data using the transform method. Overall, PySpark's implementation of the Random Forest algorithm provides a powerful tool for predictive modelling on large datasets, with the ability to scale computations across a distributed cluster of computers.

V. RESULTS

Figure 1 shows that give soil parameters as input to the system to predict suitable crop

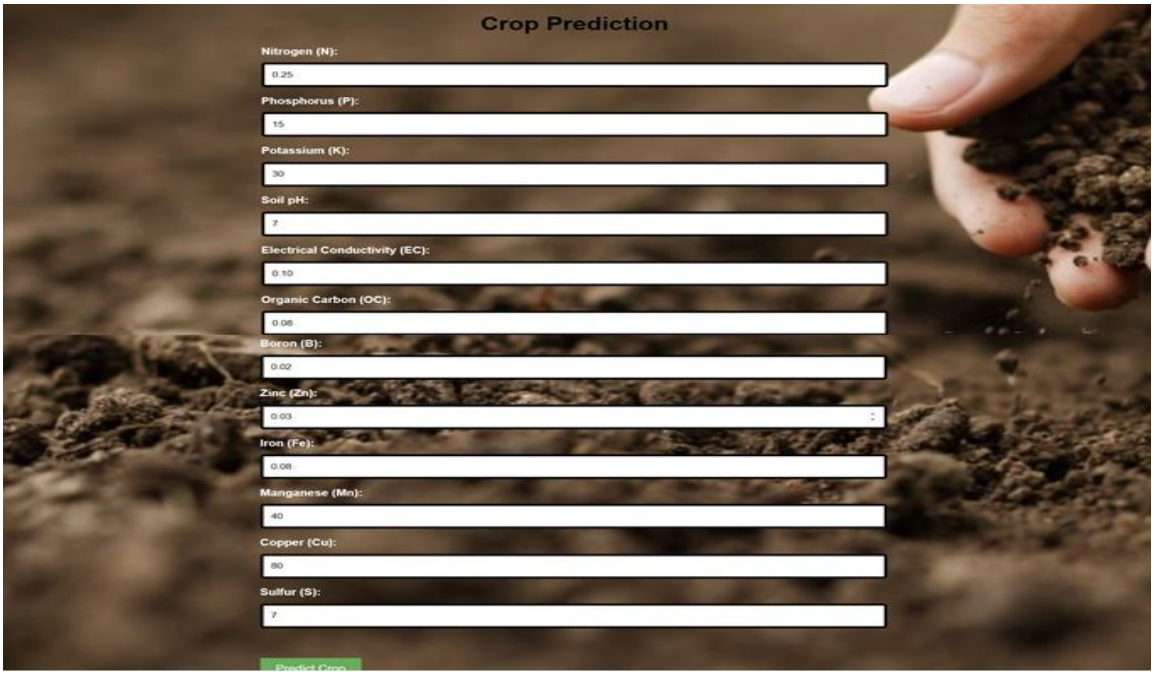


Figure 1: Soil input to predict Crop

Figure 2 shows that predict suitable crop based on given soil parameters to the system

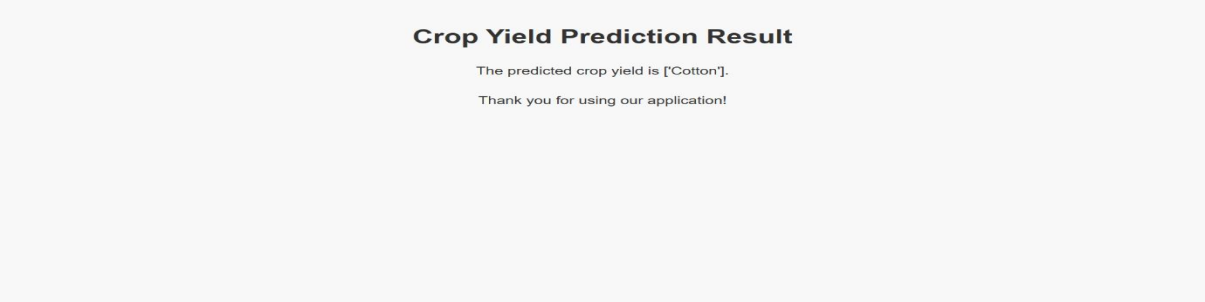


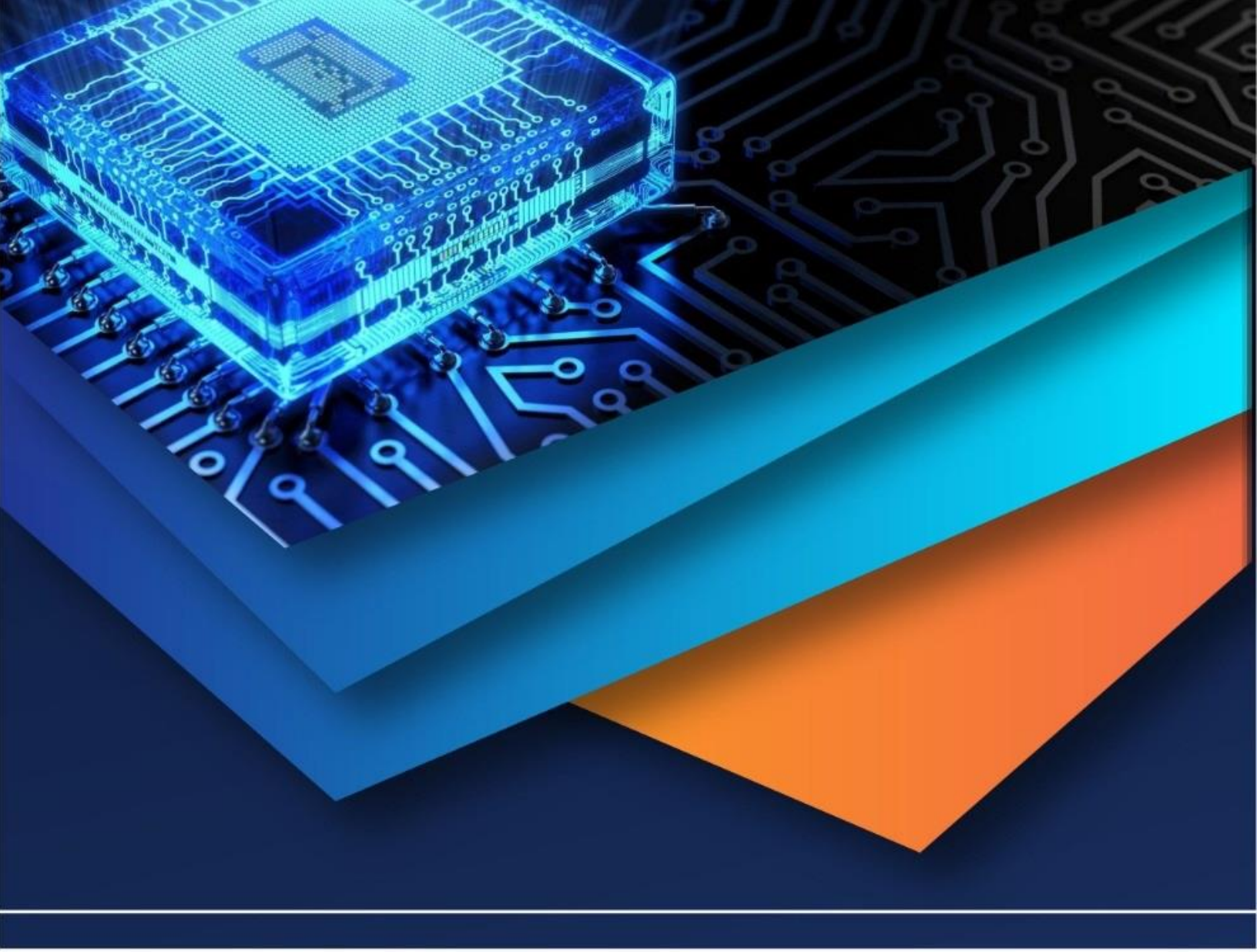
Figure 2: Predict suitable Crop based on parameters of soil

VI. CONCLUSION

This work aimed to develop a model using PySpark that could predict the type of crop based on soil properties. The process included data collection, preprocessing, feature extraction, model training, and evaluation. Soil properties such as pH, nitrogen, and phosphorus content were collected from multiple farms and preprocessed using PySpark. Feature extraction was performed to identify the most important soil properties that contribute to crop growth. The model was then trained using machine learning algorithms such as Random Forest, and evaluated using various performance metrics. The trained model was deployed to predict the crop type for future data based on soil properties. The results showed that the model successfully predicted the crop type based on soil properties. This has important implications for farmers and policymakers as it provides a tool to aid in decision-making and planning for the future.

REFERENCES

[1] "A review of crop yield prediction models based on machine learning techniques" by Wen et al. (2020)
[2] "A review of crop yield prediction models based on soil information" by Huang et al. (2019)
[3] "Crop yield prediction using machine learning: A systematic literature review" by Shaaban et al. (2021).
[4] "Crop yield prediction using soil properties: A review" by Nouri et al. (2020).
[5] M. van der Velde, L. Nisini, Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015, Agricultural Systems,Volume 168,2019,Pages 203-212,ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2018.06.009.
[6] https://sparkbyexamples.com/pyspark-tutorial/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY