IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, February 2025
Impact Factor: 7.67

Crop Yield Prediction Using Naïve Bayes Algorithm

Arjun Patil[1] and Tanmay J Gharat[2]
Assistant Professor and Head Department of IT[1]
Student, P.G. Department of IT[2]
Veer Wajekar ASC College, Phunde, Uran

Abstract: *Agriculture is the backbone of India and plays a vital role in its economy. Around 58% of India's population depends on agriculture for livelihood. According to government estimates, the food production in India was 291.95 million tonnes in 2019–20 and was projected to increase to 298.3 MT in 2020–21. To sustain population growth, food production must double by 2050. Small and marginal farmers are critical for ensuring food security and achieving Sustainable Development Goals (SDGs). However, nearly 14% of the population remains undernourished, and India ranked 94th out of 107 countries in the Global Hunger Index (2020). Achieving 'zero hunger' by 2030 requires a data-driven, integrated approach to sustainable agriculture. In this work, we apply machine learning—specifically the Naïve Bayes algorithm—to build a predictive model for crop yield. This system aims to aid farmers in optimizing crop selection based on climate, soil, and other parameters.*

Keywords: Naïve Bayes, Machine Learning, Crop Yield, KNN, Agriculture Prediction

I. INTRODUCTION

India is an agrarian nation, with agriculture playing a pivotal role in the country's socio-economic development. A significant portion of the Indian population—nearly 58%—depends on farming as its primary livelihood. Agriculture not only ensures food security but also contributes substantially to national GDP, export revenue, and employment generation. The primary crops cultivated in India include rice, wheat, pulses, maize, sugarcane, and various fruits and vegetables. These crops are largely dependent on seasonal monsoons, soil fertility, and climate conditions, making agriculture a highly unpredictable sector.

In recent years, rapid population growth has led to increasing demand for food grains and other agricultural products. To meet this demand, farmers are under pressure to improve crop yield and productivity. However, several challenges such as erratic rainfall, fluctuating temperatures, pest infestations, and poor resource planning hinder agricultural efficiency. Traditional methods of yield estimation and crop selection are no longer sufficient in addressing the dynamic nature of modern agriculture.

To overcome these challenges, technological interventions are being increasingly explored. One such promising approach is Machine Learning (ML), a subset of Artificial Intelligence (AI), which has demonstrated remarkable potential in predictive analytics across various sectors. In the context of agriculture, ML can be used to analyze large datasets related to climate, soil, crop type, and irrigation patterns to predict crop yields, recommend crop types, identify diseases, and optimize resource usage. By leveraging historical data, ML models can uncover hidden patterns and correlations that may not be obvious through conventional statistical methods.

Among the various ML algorithms, K-Nearest Neighbor (KNN) is one of the most commonly used methods due to its simplicity and ease of implementation. It classifies data points based on their proximity to other labeled data points in the dataset. Despite its advantages, KNN becomes computationally expensive as the size of the dataset increases. It also struggles with high-dimensional data and requires careful tuning of the 'k' value to perform well.

In contrast, the Naïve Bayes algorithm is a powerful probabilistic classifier based on Bayes' Theorem, which calculates the posterior probability of a class based on the prior probability and the likelihood of attributes. One of its key advantages is the assumption of feature independence, which significantly simplifies the computation. Though this

ISSN 2581-9429 IJARSCT

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, February 2025
Impact Factor: 7.67

assumption may not always hold true in real-world data, Naïve Bayes often performs surprisingly well, especially with categorical or text-based data. Its efficiency and low resource requirement make it highly suitable for applications where fast and scalable predictions are needed.

This research proposes the design and implementation of a crop yield prediction system using the Naïve Bayes algorithm. The model is built by applying supervised learning techniques to a historical agricultural dataset containing information such as soil properties, rainfall, temperature, crop type, and season. The process involves several steps including data collection, preprocessing, feature extraction, training the model, and building a web-based graphical user interface (GUI) to allow end users—especially farmers—to interact with the system and get real-time predictions.

The goal of this system is to empower farmers with data-driven insights that can aid in better crop planning, reduce risk, and ultimately enhance agricultural productivity. By integrating machine learning techniques like Naïve Bayes into agriculture, we aim to contribute to the broader vision of precision farming, sustainable agriculture, and food security.

II. MOTIVATION

The need for automation in agriculture is heightened by manual labor, unpredictable weather patterns, and global climate change. Accurate prediction of crop yields helps optimize agricultural planning, reduce losses, and enhance food security. Crop yield estimations benefit not only farmers but also agro-industries for storage, logistics, and marketing strategies.

III. LITERATURE SURVEY

Machine learning is increasingly used in agricultural analytics. Previous studies have compared various ML models for crop yield prediction:

Predicting crop yield has been a critical area of study in agricultural informatics and data mining. The adoption of machine learning (ML) techniques for this purpose has steadily grown in recent years, offering improved accuracy and decision support for farmers.

Several studies have examined the potential of supervised learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Trees, Random Forest, and Naïve Bayes for crop yield prediction.

In a study by Sharma et al. (2019), the authors applied KNN and Random Forest algorithms to predict rice yield based on rainfall, temperature, and soil parameters. Their results indicated that Random Forest outperformed KNN in terms of accuracy and reliability, especially when dealing with noisy data. However, KNN proved useful for smaller datasets due to its simplicity and lower computational cost [1].

Singh and Gupta (2020) proposed a hybrid model combining Naïve Bayes and Decision Tree for yield prediction in wheat crops. The model achieved an accuracy of 87% on historical agricultural datasets from the Indian Meteorological Department (IMD). Naïve Bayes was particularly effective due to its simplicity and capacity to handle probabilistic relationships between input variables [2].

A comprehensive analysis by Jadhav et al. (2021) compared Logistic Regression, SVM, and Naïve Bayes for yield forecasting in Maharashtra's farming zones. Their experimental results showed that Naïve Bayes achieved better performance in regions with consistent seasonal patterns and well-structured tabular data [3].

Patel and Patel (2020) implemented a yield prediction system using ML algorithms including Naïve Bayes, KNN, and Decision Trees, with historical yield data from Gujarat. Their findings revealed that while KNN performed well with a small feature set, Naïve Bayes maintained consistent prediction accuracy even when irrelevant or redundant features were present [4].

In another study, Kamble et al. (2022) focused on maize yield prediction using weather, soil, and irrigation data. Their research indicated that Naïve Bayes, despite being a relatively simple algorithm, was able to deliver competitive results with minimal training data, making it suitable for low-resource rural applications [5].

Additionally, Chaudhary et al. (2021) explored the integration of remote sensing data with ML algorithms for spatial yield estimation. While their model emphasized SVM and Random Forest, they concluded that Naïve Bayes provided a strong baseline model for quick and early predictions in areas with incomplete data [6].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, February 2025
Impact Factor: 7.67

From the literature reviewed, it is evident that Naïve Bayes remains a practical choice for early-stage prediction due to its simplicity, low computational requirements, and ability to deal with categorical data. Its assumption of conditional independence may not always hold, but the algorithm continues to perform well in many agricultural datasets where relationships are loosely correlated.

IV. DATA COLLECTION

A dataset of over 10,000 records was collected from online portals including Kaggle, government repositories, and weather sources. The dataset contained features such as state, district, season, rainfall, soil type, temperature, and humidity.

Steps Involved:

- **Data Selection**: Extracted historical crop yield data.
- **Data Cleaning**: Removed null and duplicate values.
- **Data Analysis**: Identified patterns and feature dependencies using Seaborn and Matplotlib.

V. PROPOSED MODEL

Naïve Bayes Classifier

The Naïve Bayes algorithm calculates the posterior probability of a class based on prior probability and likelihood of features:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

It assumes independence among features, making it simple and computationally efficient. Gaussian Naïve Bayes was used due to the continuous nature of the features.

Model Architecture

a) Data Collection: Data acquired from weather and crop databases.
b) Preprocessing: Null values filled or removed; features encoded.
c) Feature Extraction: Based on correlation matrix—important features selected.
d) Prediction: Dataset split into 80:20 train-test sets. Trained on Google Colab using Python's scikit-learn. Accuracy: 97%.
Boosting was added to improve weak learners by reweighting observations with poor prediction in the previous model iteration.

VI. GUI WEB APPLICATION

The system includes a user-friendly web application for farmers:

Sign-Up Page

Fields: First Name, Last Name, Email, Password, Submit.

Login Page

Fields: Email ID and Password.

Dashboard

Provides access to crop prediction features.

Crop Info Page

Fields: State, District, Season, Min/Max Temperature, Humidity, Rainfall, Soil Type, and Zone.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 6, February 2025
Impact Factor: 7.67

Prediction Page

Displays the predicted yield or suggested crop.

Result Page

Shows classification result using trained Naïve Bayes model.

VII. CONCLUSION

This project demonstrates the effectiveness of Naïve Bayes for crop prediction, helping farmers choose suitable crops based on environmental and agricultural parameters. The system's 97% accuracy provides actionable insights to support rural decision-making and agricultural sustainability. Future enhancements can include IoT integration, real-time data processing, and multilingual interfaces for broader accessibility.

REFERENCES

[1]. Batzelis, E., & Pal, B. (2020). Machine Learning Algorithms in Forecasting. *Journal of AI Research*.

[2]. Sharma, V., Tripathi, S., & Singh, A. (2019). Crop yield prediction using decision tree and random forest. *International Journal of Computer Applications, 178*(11), 1-5.

[3]. Singh, R., & Gupta, A. (2020). A hybrid approach for crop yield prediction using machine learning algorithms. *Journal of Agricultural Informatics, 11*(2), 25-31.

[4]. Patel, H., & Patel, M. (2020). Crop prediction using supervised machine learning techniques. *IJRESM, 3*(8), 56-59.

[5]. Jadhav, R., Pawar, K., & Deshmukh, V. (2021). Comparative analysis of ML algorithms for crop yield prediction. *IJITE, 10*(3), 45-50.

[6]. Aruvansh Nigam et al. (2020). Comparative analysis of RF and LSTM for rainfall and crop yield prediction. *Xi'an Univ. of Architecture & Technology*, Vol. XII, Issue V