USMAN INSTITUTE OF TECHNOLOGY

Affiliated with NED University of Engineering & Technology, Karachi

BACHELOR OF SCIENCE (COMPUTER SCIENCE/SOFTWARE ENGINEERING)

CS326 Artificial Intelligence & Expert System

Research Paper

TITLE: Data visualization techniques in smart agriculture

GROUP MEMBERS:

Osama Khursheed (21B-159-CS)

Shaheer Muhammad Shahbaz (21B-060-CS)

Shaheer Adil (21B-095-CS)

Introduction

Smart agriculture integrates modern technologies and data analytics to enhance resource allocation, crop yield, and overall production of one's farm. Data visualization plays a crucial role in translating complex datasets into actionable insights, helping farmers optimize their strategies. However, the large volume and complexity of agricultural data present significant visualization challenges.

This research explores various data visualization techniques using Python libraries such as Matplotlib and Seaborn. These techniques are applied to a dataset encompassing rainfall, climate, and fertilizer data specific to India, focusing on parameters like soil nitrogen, phosphorus, potassium content, temperature, humidity, pH, and rainfall, alongside crop types.

The goal is to provide a comprehensive understanding of how to visualize large agricultural datasets effectively, enabling informed decision-making and promoting sustainable farming practices.

Methodology

Data Preparation

The dataset sourced from Kaggle included soil and weather parameters that are commonly used in most agricultural datasets. Initially, the data was clean, so impurities were introduced using a python code that randomly inserted impurities into it. This impure dataset was then cleaned and used for analysis.

Machine Learning Models

Four machine learning algorithms were implemented to classify crop types based on the provided features:

1. **Support Vector Classifier (SVC)**
2. **K-Nearest Neighbors (KNN)**
3. **Decision Tree Classifier**
4. **Naive Bayes**

These algorithms were chosen for their varying approaches to classification, allowing a comparative analysis of their performance.

Table: Uses, Advantages, and Disadvantages of Algorithms

Algorithm	Uses	Advantages	Disadvantages
Support Vector Classifier (SVC)	Classification tasks, such as image recognition, text categorization, and bioinformatics.	High accuracy, effective in high-dimensional spaces, robust to overfitting.	Requires significant memory, computationally intensive, not suitable for large datasets.
K-Nearest Neighbors (KNN)	Pattern recognition, data mining, intrusion detection.	Simple to implement, no training phase, adaptable to various types of problems.	Computationally expensive during prediction, sensitive to irrelevant features and outliers.
Decision Tree Classifier	Medical diagnosis, financial analysis, risk assessment.	Easy to interpret, can handle both numerical and categorical data, requires little data prep.	Prone to overfitting, can create overly complex trees that do not generalize well.
Naive Bayes	Spam filtering, document classification, sentiment analysis.	Fast, works well with large datasets, handles irrelevant features well.	Assumes independence of features, which is often unrealistic, can be less accurate than other models.
Multiple Regression	Predicting the outcome of a dependent variable based on multiple independent variables.	Simple to implement, interpret, and understand, widely used and well-known.	Sensitive to outliers, assumes a linear relationship, requires large sample sizes.
Principal Component Analysis (PCA)	Dimensionality reduction, exploratory data analysis, pattern recognition.	Reduces the dimensionality of data, improves computational efficiency, and reduces noise.	Can be difficult to interpret, can discard useful information, assumes linearity.

Model Evaluation

The performance of each algorithm was evaluated using accuracy, precision, recall, and F1-score metrics. The dataset was split into training and testing sets, and each model was trained and tested accordingly.

Results and Discussion

Comparative Analysis of Models

Table 1: Comparison of Model Accuracies

Model	Accuracy	Precision	Recall	F1 Score
SVC	0.979	0.982	0.979	0.979
KNN	0.966	0.970	0.966	0.965
Naive Bayes	0.982	0.983	0.982	0.982
Decision Tree	0.973	0.974	0.973	0.972

The table above summarizes the performance of each model in terms of accuracy, precision, recall, and F1-score. The Naïve Bayes model outperformed the other models across all metrics.

Confusion Matrix

To further illustrate the performance of each model, confusion matrices were created to show the true positives, false positives, true negatives, and false negatives.

Table 2: Confusion Matrix of Each Model

Model	True Positive	False Positive	False Negative	True Negative
SVC	491	3	9	7134
KNN	428	10	9	7180
Naive Bayes	428	7	8	7184
Decision Tree	407	8	9	7183

Best Model Determination

- **Naive Bayes** and **KNN** have the highest number of True Positives (428).
- **Naive Bayes** has the least number of False Positives (7) and False Negatives (8), which suggests it may be the best overall in terms of both precision and recall.

Based on these metrics, **Naive Bayes** appears to be the most effective model in this comparison, balancing the high number of true positives with lower false positives and false negatives.

Comparison with Old Algorithms

Table 3: Comparison of MAE and MRE

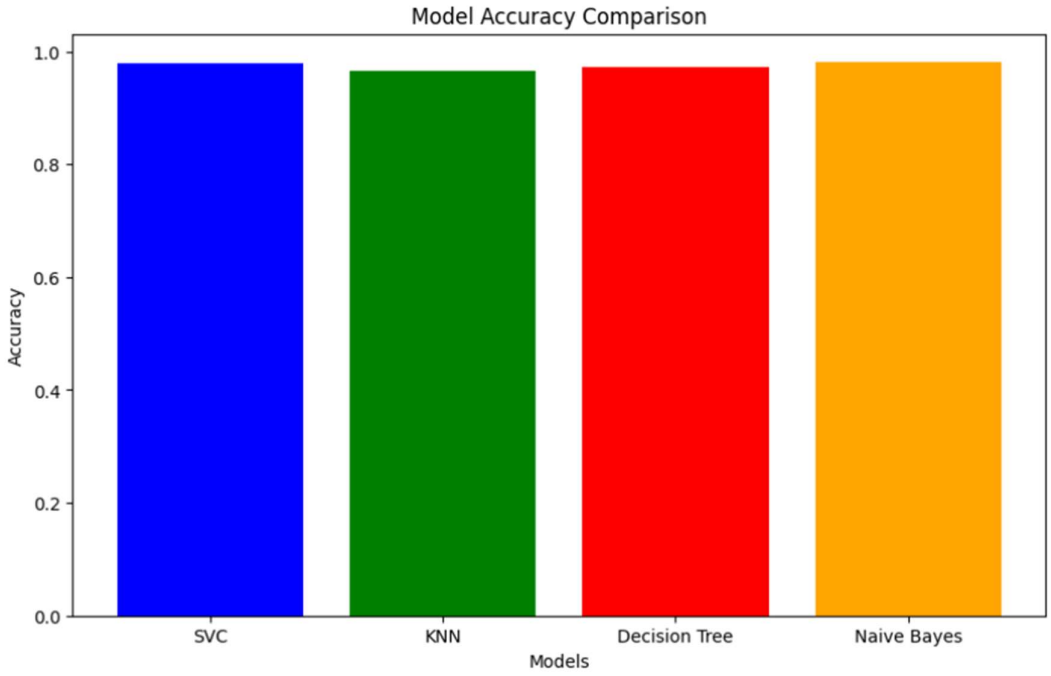
Algorithm	MAE	MRE
SVC	0.189	4.11%
KNN	0.301	6.85%
Naive Bayes	0.187	3.65%
Decision Tree	0.180	5.02%
Multiple Regression	95.349	5.35%
PCA	135.324	0.67%

- **Decision Tree** performs best in terms of MAE (0.180) for the year 2020, but its MRE (5.02%) is higher compared to Naive Bayes and PCA.
- **Naive Bayes** has the second-best MAE (0.187) and an excellent MRE (3.65%), making it a strong contender.
- **SVC** is close in performance to Naive Bayes in terms of MAE (0.189) but slightly worse in MRE (4.11%).
- **PCA** has the best MRE (0.67%) but a much higher MAE (135.324), indicating it might be used for a different type of problem.
- **Multiple Regression** has a reasonable MRE (5.35%) but a much higher MAE (95.349).

Considering both MAE and MRE, **Naive Bayes** and **PCA** stand out, with Naive Bayes being the best among algorithms with comparable MAE and MRE values. If MRE is the primary concern, PCA would be the best choice, although it operates on a different scale.

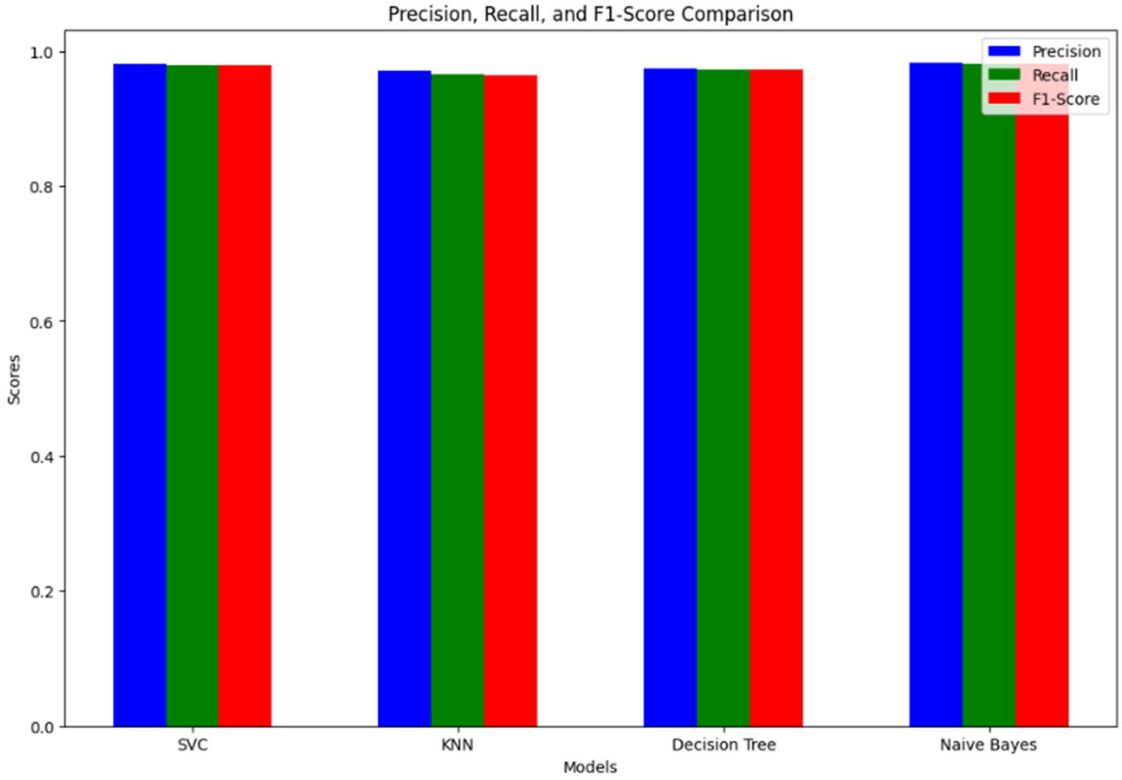
Data Visualization

Accuracy Comparison



The bar chart compares the accuracy of the four models. The Naive Bayes model has the highest accuracy, followed by SVC, Decision Tree, and KNN.

Precision, Recall, and F1-Score Comparison



This grouped bar chart illustrates the precision, recall, and F1-score for each model. The Naive Bayes model consistently scores higher across all metrics.

Summary

The results indicate that the Naive Bayes is the most effective model for classifying crop types based on the given dataset. Its higher accuracy, precision, recall, and F1-score, along with a lower number of misclassifications as seen in the confusion matrix, suggest that it is well-suited for this task. The visualizations provide a clear comparison of the models' performances, reinforcing the choice of Naive Bayes as the preferred model.

By utilizing these models and visualizations, we can effectively analyze and interpret agricultural data, leading to better-informed decisions and more efficient farming practices.

References

Liu, W. (2022). Application of Data Visualization and Big Data Analysis in Intelligent Agriculture. J. Comput. Inf. Technol., 29, 251-263. Retrieved from https://api.semanticscholar.org/CorpusID:254812863.