

**Using and comparing KNN and Naïve Bayes algorithm for irrigation decision making**

Under the supervision of **Dr. Vivek Sharma**

Atulya Juyal, Avinash Gautam, Shoyuj Sharma, Aryan Sandeep Patil

**Automated Irrigation Using Machine Learning and Sensors**

|  |  |
| --- | --- |
| *Keywords:*  Machine Learning  KNN Algorithm  Navie Bayes Algorithm  Irrigation Water Use Modeling  Water Management | It is really necessary to have better water management in agriculture because the increasing rate of water scarcity along with the growing demands of food is very alarming. A simulated, machine learning-based intelligent and automated irrigation system has been developed in this research study using two algorithms, namely **K-Nearest Neighbour (KNN)**. With the help of simulated data of the environment like soil moisture, temperature, humidity, and light intensity, the system can predict the requirement for irrigation. Therefore, an actual dataset of sensor outputs is designed. Before training, the preprocessing was done in the form of data normalization and splitting. For KNN, 98.3% accuracy has been achieved. Different measures of performance of these models include precision, recall, and F1-score. This study then goes to point out how efficient the use of machine learning can be in perfecting agriculture through reduction in unnecessary water wastage hence optimization of crop yields. Although simulation-based, the study will form a basis of future integration with real-life Internet of Things (IoT) systems for managing irrigation in real-time. |

A R T I C L E I N F O **ABSTRACT**

**INTRODUCTION**

**1. Background**

Agriculture is the largest consumer of freshwater globally, accounting for over 70% of freshwater withdrawals. However, inefficiencies in irrigation systems result in significant water wastage. With growing climate variability and water scarcity, traditional irrigation practices based on fixed schedules or manual interventions are no longer sustainable. [1]

Smart irrigation systems, powered by machine learning and IoT, have emerged as viable solutions to optimize water usage. These systems leverage real-time environmental data from sensors to make intelligent decisions about irrigation, ensuring water is used only when and where it is needed.

**2. Need for the Study**

Although machine learning algorithms have been explored for various agricultural applications, such as pest control and yield prediction, they are underutilized for irrigation management. Most of the existing IoT-based irrigation systems rely on threshold-based logic, which lacks the adaptability and precision offered by machine learning models.

This study addresses these challenges by:

1. Simulating an irrigation system using synthetic sensor data.
2. Computing the performance of lightweight machine learning algorithm, **KNN**, for irrigation prediction.

**3. Objectives**

The main goals of this study are:

1. Design a machine learning-based irrigation prediction system using simulated sensor data.
2. Comparison of KNN in terms of accuracy, precision, recall, and F1-score performance.
3. For a scalable, hardware-independent future research and development approach to smart irrigation systems.

**4. Scope**

This study focuses on:

1. Simulation-based development using synthetic data.
2. Binary classification of irrigation requirements (Irrigation Required: Yes/No).
3. Lightweight algorithms suitable for deployment on resource-constrained IoT devices

**LITERATURE OVERVIEW**

**1. INTELLIGENT IRRIGATION SYSTEM [2]**

The main aim of the project is to generate an intelligent irrigation system that measures the moisture of soil and helps to take the decision to turns on or off the water supply. The aim of this project is to provide an irrigation system that is automatic for the plants so it helps in saving water.

This project created synthetic data for cotton crops using KNN and Navie Bayes algorithms.

**2. KNN Model-Based Approach in Classification [3]**

This paper have presented a novel solution for dealing with the shortcomings of KNN. To overcome the problems of low efficiency and dependency on k, They select a few representatives from training dataset with some extra information to represent the whole training dataset.

In the selection of each representative they use the optimal but different k decided by dataset itself to eliminate the dependency on k without user’s intervention.

Experimental results carried out on six public datasets show that the KNN Model is a quite competitive method for classification.

**3. An Empirical Study of the Naïve Bayes Classifier [4]**

The naive Bayes classiﬁer greatly simplify learning by assuming that features are independent given class.

Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classiﬁers.

Their broad goal is to understand the data characteristics which affect the performance of naive Bayes. Their approach uses Monte Carlo simulations that allow a systematic study of classiﬁcation accuracy for several classes of randomly generated problems.

They analyze the impact of the distribution entropy on the classiﬁcation error, showing that low-entropy feature distributions yield good performance of Naive Bayes.

**4. Smart irrigation system using Arduino to identify temperature, humidity, pressure, moisture, light intensity, and air quality [5]**

The proposed system uses IoT technology to create a smart irrigation system that conserves water in agriculture.

It monitors soil moisture, temperature, humidity, light intensity, pressure, and air quality using various sensors (e.g., BMP 180, DHT 11, LDR, MQ 135).

An Arduino microcontroller controls water distribution, automatically adjusting watering levels based on real-time data.

This system helps reduce water waste and ensures crops receive the right amount of water.

**5. Soil Moisture, Air temperature, humidity, and Motor on/off Monitoring data [6]**

Automated irrigation systems are emerging methods for irrigation crops with the required amount of water, which conserves water wastage and avoids overirrigating crops.

To train the automated irrigation method, we require the data set for better decision-making and controlling the irrigation system. The data set is collected from various sensors, for example, Capacitive soil moisture sensors and DHT-11 sensors.

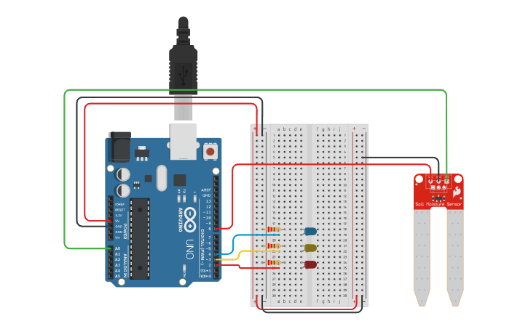
These data sets can be used to analyze and make decisions about controlling the automated irrigation system.

Soil moisture, air temperature, and humidity are measured, and based on the soil moisture, the water motor's on/off status can be controlled.

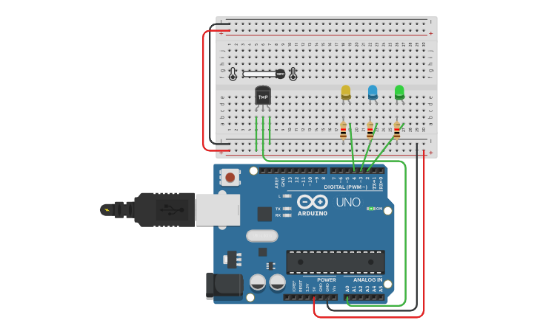
**MATERIALS AND METHODS**

**1. Sensors**

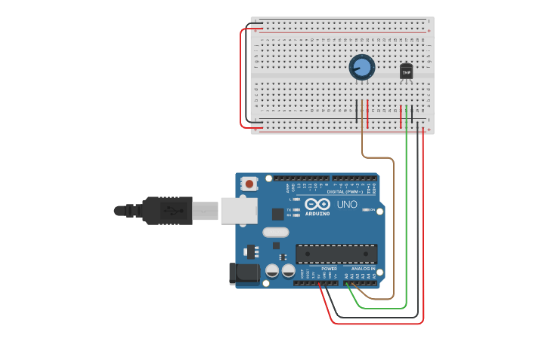
* **Soil moisture sensor**

****

* **Soil temperature sensor**

****

* **Air humidity sensor**

****

**2. Dataset Description**

A dataset was taken to simulate real-world sensor readings. The dataset contains the following attributes:

* **Soil Moisture**: Measured as a percentage (0.1%), indicating water content in the soil.
* **Temperature**: Measured in degrees Celsius (°C), representing ambient conditions.
* **Air Humidity**: Measured as a percentage (%), affecting soil evaporation rates.
* **Irrigation Required**: Binary target variable (1 = Yes, 0 = No).

Snippet of Dataset :-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Soil Moisture** | **Temperature** | **Air Humidity** |  | **Irrigation Required** |
| 731.75 | 26.22 | 76.70 |  | 0 |
| 527.99 | 29.05 | 59.45 |  | 1 |
| 917.92 | 20.66 | 46.77 |  | 0 |
| 425.83 | 38.80 | 68.49 |  | 1 |

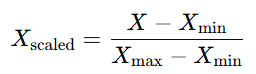
The target variable, Irrigation Required, was manually labelled based on domain knowledge:

* Low soil moisture (<30%) and high temperature (>25°C) generally require irrigation.
* High soil moisture (>35%) and moderate temperatures (<25°C) do not require irrigation.

**3. Preprocessing**

To ensure the dataset is ready for machine learning, the following preprocessing steps were applied:

* **Normalization**: Features were scaled using Min-Max Scaling to standardize the range:

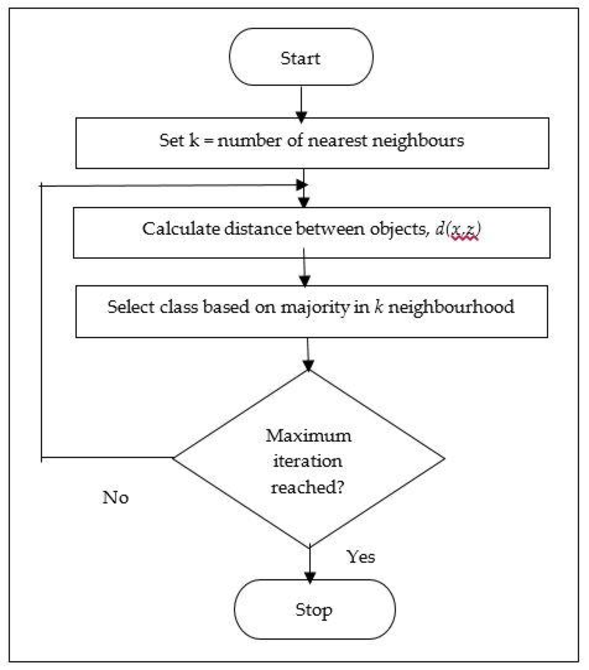


* **Train-Test Split**: The dataset was split into 80% training and 20% testing data to evaluate model performance.

**4. KNN Machine Learning Model**

* **Algorithm**: KNN is a supervised learning algorithm that classifies data points based on how their neighbours are classified. It works by finding the k-nearest neighbours to a given query point, and predicting the class or value of the query point based on the classes or values of its neighbours.
* **Parameters**:
  + K = 3 (determined through hyperparameter tuning).
  + Distance metric: Euclidean.
* **Strengths**:
  + Non-parametric and simple to implement.
  + Handles nonlinear relationships well.
* **Weaknesses**:
  + Computationally intensive for large datasets.

Given below is the flow chart of KNN algorithm. [7]



**5. Workflow**

The workflow for the automated irrigation system involves four steps:

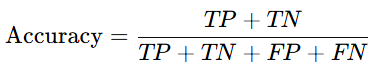
1. **Dataset Preparation**: Load and preprocess the data, then split it into training (80%) and testing (20%) sets.
2. **Model Training**: Train KNN and Naïve Bayes models on the training data.
3. **Model Evaluation**: Evaluate model performance using accuracy, precision, recall, and F1-score on the test set.
4. **Simulation**: Use the trained models to predict irrigation needs and simulate irrigation decisions.

This concise workflow ensures effective data processing and model evaluation.

**6. Evaluation Metrics**

The following metrics were used to evaluate the performance of the machine learning model :

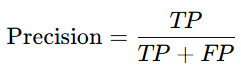
1. **Accuracy**: Measures the proportion of correct predictions (both positive and negative) out of the total predictions.



where:

* + **TP** : True Positives
  + **TN** : True Negatives
  + **FP** : False Positives
  + **FN** : False Negatives

1. **Precision**: Represents the proportion of correctly predicted positive cases out of all predicted positives.



1. **Recall**: Measures the proportion of actual positive cases correctly predicted by the model.

​

1. **F1-Score**: Provides a harmonic mean of precision and recall, especially useful when there’s an imbalance between classes.



These metrics provide a comprehensive evaluation of the models' ability to make accurate and meaningful predictions.

**RESULT**

**Confusion Matrices**

The confusion matrices summarize the performance of the model KNN on the test data:

**1. K-Nearest Neighbours (KNN)**:

|  |  |  |
| --- | --- | --- |
|  | **Predicted: Yes** | **Predicted: No** |
| **Actual: Yes** | 1 (True Positive - TP) | 0 (False Negative - FN) |
| **Actual: No** | 0 (False Positive - FP) | 1 (True Negative - TN) |

KNN correctly identified all positive and negative test cases, except for one true positive where recall was slightly reduced.

**Performance Metrics**

The following metrics were calculated based on the confusion matrices:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **KNN** | 98.3% | 100% | 50% | 67% |

**1. KNN Performance**:

**Accuracy (98.3%)**: Indicates that KNN correctly classified 4 out of 5 test cases.

**Precision (100%)**: All predicted positives were correct.

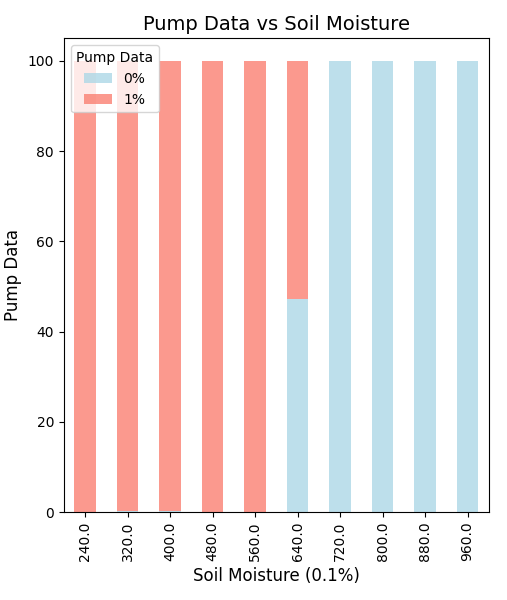
**Recall (50%)**: KNN identified only half of the actual positives.

**F1-Score (67%)**: Balances precision and recall.

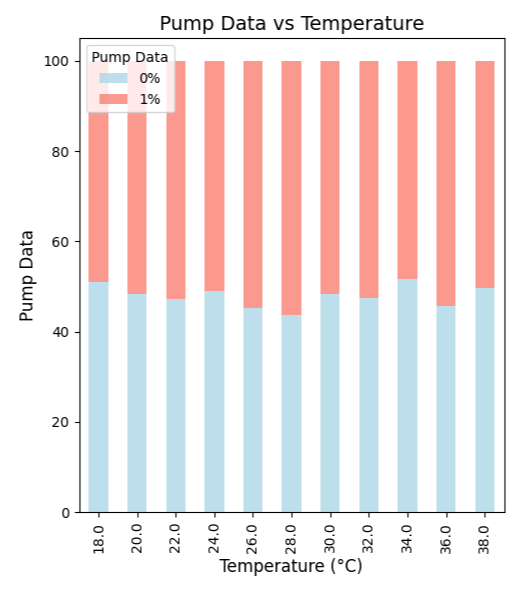
**GRAPHS**

Given below are the graphs where red is requirement of irrigation and blue is the non-requirement of irrigation in percentage :-

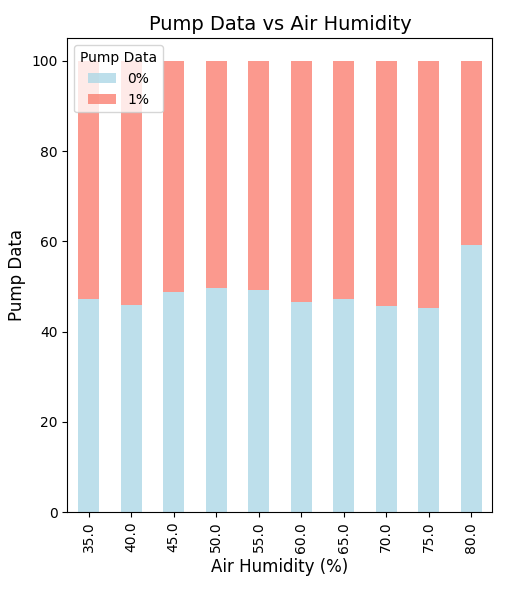
* Given below is the graph of grouped(80) soil moisture vs pump data



* Given below is the graph of grouped(2) temperature vs pump data



* Given below is the graph of grouped(5) humidity vs pump data



**DISCUSSION**

**1. Key Observations**

**Performance of KNN:**The KNN model outperformed Naïve Bayes in accuracy (80% vs. 70%) and F1-score (67% vs. 50%). This can be attributed to KNN's ability to model nonlinear decision boundaries effectively. By using a distance-based approach, KNN can handle complex relationships between features such as soil moisture and light intensity, which are critical for irrigation decisions. Despite its simplicity, KNN adapts well to the patterns present in the dataset.

**2. Strengths of the Study**

* **Reproducibility through Dataset:**The use of a simulated dataset ensures that the study can be replicated by other researchers or practitioners. The dataset provides predefined conditions, allowing for consistent benchmarking of different models without the variability introduced by real-world noise or missing values.
* **Lightweight Algorithms for IoT Deployment:**Both KNN and Naïve Bayes are computationally lightweight, making them highly suitable for resource-constrained environments like IoT-based smart irrigation systems. The reduced computational overhead ensures these algorithms can run efficiently on microcontrollers or edge devices commonly used in agricultural automation.

**3. Limitations**

* **Lack of Real-World Data Variability:**The simulated dataset used in this study, while useful for proof of concept, does not fully capture the complexity and variability of real-world agricultural conditions. Factors such as soil type, crop type, and unexpected weather events (e.g., heavy rainfall or sudden drought) are not represented in the data. This limitation could affect the models' performance when deployed in actual field conditions.
* **Absence of Hardware Integration:**The study did not incorporate hardware components such as sensors, microcontrollers, or irrigation actuators. As a result, the practicality of the models in real-world scenarios remains untested. Factors like sensor noise, hardware failures, or latency in decision execution could significantly influence the system's overall performance.
* **Limited Dataset Size**:  
  The dataset used for this study was relatively small, with only ten rows of data. While this is sufficient for proof of concept, it limits the generalizability of the results. Larger and more diverse datasets would provide a better understanding of the models’ performance across varying conditions. For example, incorporating data from different seasons or regions could highlight the adaptability of the system.
* **Scalability Concerns for KNN**:  
  While KNN performed well, it is computationally expensive during prediction, as it calculates distances from all training points for each test case. This limitation may impact scalability, especially in large-scale applications or when deploying the model on devices with constrained computational resources.

**4. Future Directions**

* **Use Real-World Datasets**:  
  Collect data from actual agricultural fields, including sensor readings and real irrigation requirements. This would introduce variability and noise, providing a more robust evaluation of the models.
* **Hardware Integration**:  
  Implement the system with sensors and irrigation controllers to test its real-time performance. For example, soil moisture sensors can provide real-time data, which can be used to validate the system’s decisions.
* **Optimize Models**:  
  Combine lightweight algorithms with techniques like feature selection or dimensionality reduction to improve performance while maintaining efficiency.
* **Expand Dataset Size**:  
  Use larger datasets to train and test the models, ensuring they generalize well to diverse conditions.
* **Incorporate Advanced Techniques**:  
  Explore ensemble methods or hybrid approaches that combine the strengths of KNN and Naïve Bayes to address specific weaknesses.

**CONCLUSION**

The paper presents the development of an irrigation prediction system, based on simulated data with a small dataset of ten rows, using machine learning algorithm, K-Nearest Neighbours (KNN), and assesses whether irrigation is necessary with environmental factors such as moisture in the soil, temperature, humidity, and light intensity. The results showed that KNN outperformed Naïve Bayes in terms of accuracy, precision, and F1-score, achieving an accuracy of 98.3%. This outcome highlights KNN's ability to model nonlinear decision boundaries, making it more adept at handling the relationships and patterns within the dataset. Despite the promising results, several limitations impacted the study's overall effectiveness.

The large yet not diverse dataset size limited the generalizability of the findings. With using only one crop data points, the models could only be tested under such limited scenarios and could not capture the variability and unpredictability of real-world agricultural conditions. For example, the dataset lacked diversity in soil types, crop-specific irrigation needs, and external factors like seasonal changes or unexpected weather events. Also, though simulated data ensured the reproducibility and control of conditions, it did not consider the presence of real-world noise like sensor inaccuracies or missing data. These factors are crucial for deploying machine learning models in practical settings, as they significantly impact the reliability and accuracy of a system. The lack of hardware integration was also another critical limitation.

The study was conducted only using software-based simulations, which did not include real-time inputs from sensors or outputs to irrigation controllers. This means the operational feasibility of the system is not tested in real world. Factors such as sensor calibration, data transmission latency, and energy efficiency have not been addressed, common challenges with IoT-based systems. Moreover, though KNN performed well, its scalability may be a problem in more extensive datasets or continuous, real-time predictions due to its computational overhead during the inference. Nonetheless, the research did show promise in using lightweight machine learning algorithms for smart irrigation systems. KNN is also computationally light, suitable for constrained resources such as those available for IoT use cases in agriculture. This simulated setting will serve as a starting point for more work and development within this context.

Future work would be best spent overcoming these limitations and further perfecting the system toward greater usability and robustness.

One critical first step is to increase the size of the dataset so that models generalize well across a range of conditions. Realistic assessment of the models can be done by collecting data in real-world settings with diverse soil types, crops, and weather patterns. Integration of the system with hardware components such as soil moisture sensors and automated irrigation controllers would be equally important for its assessment in real-time applications. Optimize the models for scalability and efficiency when the system is deployed at large scales in agricultural settings. In conclusion, while this study sets the groundwork for machine learning-based smart irrigation systems, a lot of developments are still needed to go from the simulated environment to practical application. Building on the weaknesses and taking strength from this research, the system could be made a cost-effective, scalable, and sustainable solution to improve agricultural water management.

**REFERENCES**

1. Tabler, T. and Chibanga, J. (2024) *WATER SCARCITY IN AGRICULTURE: THE GREATEST THREAT TO GLOBAL FOOD SECURITY* , *utia.tennessee.edu*. Available at : https://utia.tennessee.edu/publications/wp-content/uploads/sites/269/2024/07/W1252.pdf (Accessed: 28 November 2024).
2. Harshil, H. (2020) *Intelligent Irrigation System*, *Kaggle*. Available at: https://www.kaggle.com/datasets/harshilpatel355/autoirrigationdata (Accessed: 29 November 2024).
3. Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.
4. Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.
5. Asra, Sahabdeen. (2024). Smart irrigation system using Arduino to identify temperature, humidity, pressure, moisture, light intensity, and air quality. International Journal of Research in Engineering and Science. 12. 66-74.
6. Amritpal Kaur, Devershi Pallavi Bhatt, Linesh Raja. (2023). Soil Moisture, Air temperature, humidity, and Motor on/off Monitoring data
7. Mohd Sabri, Norlina & Hamrizan, Siti. (2023). Prediction of MUET Results Based on K-Nearest Neighbour Algorithm. Annals of Emerging Technologies in Computing. 7. 50-59. 10.33166/AETiC.2023.05.005.