WATER DEMAND PREDICTION IN IRRIGATION SYSTEM USING KNN ALGORITHM

WINT WAH LOON

M.C.Sc. **September 2022**

WATER DEMAND PREDICTION IN IRRIGATION SYSTEM USING KNN ALGORITHM

By

WINT WAH LOON

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

September 2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…..…… ……..……

Date Wint Wah Loon

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin**, Rector, University of Computer Studies, Yangon, for her kind permission to submit this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win and Dr. Thin Zar Thaw,** Professors, Faculty of Computer Science, University of Computer Studies, Yangon, for their superior suggestions, administrative supports and encouragement during my academic study.

My thanks and regards go to my supervisor, **Dr. Thin Lai Lai Thein,** Professor, Faculty of Information Science, University of Computer Studies, Yangon, for her support, guidance, supervision, patience and encouragement during the period of study towards completion of this thesis.

I also wish to express my deepest gratitude to **Daw Hnin Yee Aung**, Lecturer, Department of English, University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's degree course and my friends for their cooperation.

I especially thank to my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

ABSTRACT

Increasing food demand will challenge the agricultural sector globally over the next decades. A sustainable solution to this challenge is to increase crop yield without massive cropland area expansion. This can be achieved by identifying and adopting best management practices. The more detailed understanding of how crop yield is impacted by climate change and growing-season weather. Many factors influence irrigation water requirement in an agriculture field. Those factors are age of plant, humidity, temperature, soil moisture/soil water needed. Despite the multiple solution proposed, still the quantity of water over flood and underfloor in the agriculture felid. The artificial influence on irrigation requirement should be thought of an important impact factor, considering the requirement of water, the technology can help in preserving large quantity of water in agriculture felid. This system will predict the balance between water supply and demand requires efficient water supply system by using K-nearest neighbor (KNN). In this system, the C# programming language is implemented on Microsoft Visual Studio ID and Microsoft SQL Server is also used for Database Engine.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

LIST OF EQUATION

CHAPTER 1
INTRODUCTION

Weather forecasting and predicting the soil water that will happen in a space can be exceptionally dreary. It would include fastidious perception of the air conditions and cloud development alongside the formation of models to mimic barometrical circumstances and cumulus cloud connection which likewise prompts a serious level of intricacy [2]. How much water utilized for water system is assessed without thinking about the successful measure of precipitation that will be knowledgeable about a given region. At the point when unreasonable measure of water is utilized it might prompt over water system, Water logging and may likewise bring about saltiness in this way lessening crop yield. Anyway utilizing less water may likewise bring about under water system and diminish yield effectiveness.

1.1 Water Demand Prediction for Agriculture

Consequently, assessing the perfect proportion of water that will be provided for inundating the harvests is of central significance. The strategy includes dissecting different geological factors, for example, land geography, slant, channel elements, for example, soil surface, construction and profundity alongside meteorological boundaries, for example, Temperature, radiation, relative stickiness, wind speed to anticipate how much compelling precipitation that will be gotten over indicated geographic locales [1]. With this the harvest water not entirely set in stone consistently.

The objective of the strategy is to help ranchers in utilizing the water required or not really for water system and assist them with picking the right water system situation that ought to be carried out for the ideal development of yields. Information mining methods, for example, characterization are utilized to acquire information about how much water expected for water system which would go far in reinforcing the agrarian area. This additionally assists ranchers with knowing ahead of time how much water that ought to be put something aside for water system if there should arise an occurrence of disappointment of monsoon rain. This system will predict the water supply needed or not for specific crops by using K-nearest neighbor (KNN).

1.2 Objectives of the Thesis

The thesis aims at forecasting the demand of water needs to grow multiple crops. It also focuses to handle multiple indicators of water demand and analyses the effectiveness of KNN classifier in prediction water demand on irrigation. The other objectives of doing this thesis are to support crop yield without massive cropland area expansion, and agricultural production in a cost-effective way.

1.3 Motivation of the Thesis

When the climate varies, then automatically field parameters also suddenly changes. Whenever there is heavy rainfall or temperature varies this may become very hard to analyze the situation and it causes a major problem. Taking this as a problem into consideration, designing of water demand prediction on irrigation system is needed. Therefore, in order to obtain the most suitable representation one needs to select a time series model that respects these features.

This system used the k-nearest neighbor (KNN) approach to forecast the short-term water demand time series. The KNN approach is a pattern recognition algorithm where the forecasted values are directly determined by the most similar past observations.

1.4 Overview of the System

Many mathematical models have been proposed over the ages and many of them are still applied to find the effective water irrigation. Here the effective prediction is computed by performing a series of past climate and water requirement. Important factors such as land evaporation humidity, groundwater / soil moisture, and temperature have been considered to ascertain the effective water needed/or not prediction in a simpler way. This KNN technique can be used to predict based on the amount of effective training data and in turn can also be used to predict the crop / plants water needs for any particular area

1.5 Organization of the Thesis

The thesis is organized in five chapters. They are as follows:

In Chapter 1, introduction of the system, Motivation, objectives of the thesis, related works and thesis organization are described. **Chapter 2** presents the background theory of prediction and classification. **Chapter 3** discusses the methodology of the system. **Chapter 4** expresses the design and implementation of the system. Finally, **Chapter 5** presents the conclusions, benefits of the system, and further extensions of the system.

CHAPTER 2
BACKGROUND THEORY

Water demand expectation is urgent for the feasible administration of water appropriation frameworks. Likewise, a variable is considered by foundation leaders to guarantee powerful water use plans and timetables, particularly considering the continuous metropolitan extension, where customers are urged to decrease their energy and asset utilization. In the current water conveyance conditions, territorial water asset the board faces vulnerabilities and difficulties, for example, water deficiencies, generally development in water interest, occasional interest tops because of environmental change, provincial monetary contest, and general wellbeing prerequisites.

Rainfall and weather expectation include refined PC displaying and amusement for precise forecast [3]. Displaying of such non-direct frameworks have been achieved by utilizing fake brain networks which have been utilized in the framework proposed in papers [8] and [9]. An Counterfeit Brain Organization here is utilized to predict the way of behaving of such nonlinear frameworks. Demonstrating of such frameworks basically includes the utilization of delicate figuring

Delicate figuring has three fundamental parts, to be specific, Artificial Neural Network (ANN), Fluffy rationale and Hereditary Algorithm. [3] Delicate processing is a model which manages surmised models where an estimate reply or result is accomplished. Measurable signs picked are fit for separating the patterns, which can be viewed as components for making the models. Fake brain networks have been utilized by many individuals to display the cumulus cloud communication in different places, for example, Thailand [10] and in a lot more places to gauge the got precipitation. They have likewise been utilized in the approaches proposed in [8] and [9]. Their endeavors however are gathered in the making a proficient displaying framework to foresee got precipitation and relatively few propose ideas to utilize the precipitation got in a viable manner.

In this system[11] a framework to give data on precipitation qualities and its expectation from the verifiable informational collections, determination of harvests in light of the gauge on taluka premise has been depicted with a contextual investigation of Bijapur region of Karnataka. It gives a definite examination on the precipitation got

throughout recent years yet information mining methods have not been applied in an effective manner in order to mine sufficient information to take care of a portion of the issues looked by ranchers, for example, assessing how much water for water system. In Paper [4]; the assessment of powerful precipitation depends principally on 3 factors to be specific dampness, temperature and precipitation which is carried out with a neuro fluffy framework which comprises of two sections fluffy rationale and the brain organization. Different factors, for example, the land incline, soil surface, soil design and wind speed have not been thought of. This might permit little errors to crawl up while computing how much compelling precipitation. This may not be alluring if when the edge for blunder is little particularly in a nation like India where ideal yield is of fundamental significance.

2.1 Models for Predicting Water Demand

Choosing an appropriate water demand prediction model is a challenge because it involves many factors, such as technology, population, society, economy, climate, and public policy [3]. Predictive models can be broadly divided into two groups:

1) Statistical and
2) Machine learning models

Statistical models utilize likelihood hypothesis and numerical measurements to get the practical connection between various factors. Generally utilize measurable models for relapse incorporate straight relapse, edge relapse, and tether relapse. Interestingly, AI models do not need characterizing an unmistakable connection among reliant and informative factors. All things considered, they utilize calculations, for example, support vector machine, choice tree, and irregular timberland to gain designs from preparing information and use them to anticipate future results.

Measurable models are broadly utilized for water request expectation [5, 6]. The primary constraint of factual models is that they should have a foreordained design, making it hard to track down one numerical capability that would function admirably on various information. Besides, factual models frequently neglect to actually manage complex information connections; their forecast exactness likewise diminishes with an expansion in how much information. Different techniques ought to be utilized while managing large and complex information. For instance, Rozos et al.

[3] utilized an incorporated framework elements and cell automata model to foresee water interest under elective methodologies, including disseminated water foundation.

Machine learning models are turning out to be progressively famous and have exhibited high prescient execution in spaces, for example, metropolitan framework, credit risk, energy, biology, and water asset the executives [4].

AI models can be additionally partitioned into single indicators and troupe calculations, as per the quantity of utilized indicators. A solitary indicator contains just a single indicator (or calculation, for example, brain organization, support vector machine, or choice tree. Group calculations like irregular timberland, AdaBoost, stowing, and slope supporting tree total various indicators, all adding to the last forecast outcome.

Outfit learning is turning out to be progressively famous. It utilizes measurable inspecting standards to prepare different models. Each of these models is utilized to foresee another example independently. The worth of the last expectation result for the new example is chosen in view of the greater part casting a ballot system. All in all, group learning changes various theories given by single indicators into one speculation.

In the field of foreseeing water assets, [3] researched 12 measurable and AI models to anticipate everyday family water utilization because of private water interest. Applied irregular backwoods, counterfeit brain organization, and backing vector machine are used to savvy meter information to foresee the hourly water interest of 90 records. It helps vector relapse, counterfeit brain network with backpropagation, and outrageous learning machine to anticipate the month to month and day to day streams of four waterway bowls in the US. They utilized characterization and relapse trees and irregular backwoods to lay out a multivariate expectation model for water interest in Seville, Spain. It utilized help vector machine, counterfeit brain organization, and arbitrary backwoods to anticipate changes in stream channel morphology.

Water request forecast can help its chiefs to accomplish more proficient water asset distribution. Existing investigations have utilized one or a couple of models to foresee water interest. Conversely, this study gives, interestingly, a thorough similar examination of a few factual and AI models.

2.2 Common Techniques in Data Classification

In this section, a variety of methods that are normally utilized for information grouping will be mentioned. These techniques will likewise be related with the various sections in this book. It ought to be brought up that these strategies address the most well-known procedures utilized for information order, and it is challenging to examine every one of the techniques in a solitary book thoroughly. The most well-known strategies utilized in information arrangement are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks. Each of these methods will be discussed briefly in this chapter.

2.2.1 Feature Selection Methods

The first period of basically all grouping calculations is that of component choice. In most information mining situations, a wide assortment of elements are gathered by people who are in many cases not space specialists. Obviously, the unimportant highlights may frequently bring about unfortunate displaying, since they are not very much connected with the class mark. As a matter of fact, such highlights will normally deteriorate the characterization precision in light of overfitting, while the preparation informational collection is little and such elements are permitted to be a piece of the preparation model. For instance, consider a clinical model where the elements from the blood work of various patients are utilized to anticipate a specific infection. Obviously, a component, for example, the level of Cholesterol is prescient of coronary illness, while a feature1, for example, public service announcement level isn't prescient of coronary illness. Be that as it may, in the event that a little preparation informational index is utilized, the public service announcement level might have freak connections with coronary illness due to irregular varieties. While the effect of a solitary variable might be little, the total impact of numerous superfluous elements can be critical. This will bring about a preparation model that sums up ineffectively to concealed test examples. Consequently, it is basic to utilize the right highlights during the preparation process.

There are two broad kinds of feature selection methods:

1. Filter Models: In these cases, a fresh basis on a solitary component, or a subset of elements, is utilized to assess their reasonableness for characterization. This technique is autonomous of the particular calculation being utilized.

2. Wrapper Models: In these cases, the element choice cycle is implanted into a grouping calculation, to make the component determination process delicate to the order calculation. This approach perceives the way that various calculations might work better with various elements.

In request to perform highlight choice with channel models, various measures are utilized to evaluate the importance of a component to the order interaction. Normally, these actions figure the irregularity of the element values over various scopes of the property, which may either be discrete or mathematical. A few models are as follows:

• **Gini Index**: Let p1 ... pk be the fraction of classes that correspond to a particular value of the discrete attribute. Then, the gini-index of that value of the discrete attribute is given by:

$$G = 1 - \sum_{i=1}^{k} p_i^2 \qquad \textbf{2.1}$$

The value of G ranges somewhere is in the range of 0 and 1 − 1/k. More modest qualities are more demonstrative of class lopsidedness. This demonstrates that the component esteem is more discriminative for characterization. The in general gini-file for the quality can be estimated by weighted averaging over various upsides of the discrete trait, or by utilizing the greatest gini-list over any of the different discrete qualities. Various methodologies might be more alluring for various situations, however the weighted normal is all the more ordinarily utilized.

• **Entropy**: The entropy of a particular value of the discrete attribute is measured as follows:

$$E = -\sum_{i=1}^{k} p_i \cdot \log(p_i) \qquad \textbf{2.2}$$

The same notations used above are, as for the case of the gini-index. The value of the entropy lies between 0 and log(k), with smaller values being more indicative of class skew.

• **Fisher's Index**: The Fisher's index measures the proportion between the class disperse to the inside class dissipate. Hence, in the event that pj is the negligible portion of preparing models having a place with class j, μj is the mean of a specific

element for class j, μ is the worldwide mean for that component, and σj is the standard deviation of that element for class j, then the Fisher score F is registered as follows:

$$F = \frac{\sum_{j=1}^{k} p_j \cdot (\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \cdot \sigma_j^2} \qquad \textbf{2.3}$$

A wide assortment of different measures, for example, the χ2-measurement and common data are additionally accessible to evaluate the discriminative force of properties. A methodology known as the Fisher's discriminant [61] is likewise utilized to consolidate the various highlights into bearings in the information that are profoundly applicable to arrangement. Such techniques are obviously highlight change strategies, which are additionally firmly connected with include determination strategies, similarly as unaided dimensionality decrease techniques are connected with solo component choice strategies.

All the more for the most part, it ought to be brought up that many elements are frequently firmly related with each other, and the extra utility of a trait, when a specific arrangement of highlights have been chosen, is not quite the same as its independent utility. To resolve this issue, the Base Overt repetitiveness Most extreme Pertinence approach was proposed in [69], in which highlights are gradually chosen based on their steady addition on adding them to the list of capabilities. this technique is noted likewise a channel model, since the assessment is on a subset of highlights, and a fresh measure is utilized to assess the subset.

In covering models, the component choice stage is implanted into an iterative methodology with a characterization calculation. In every emphasis, the characterization calculation assesses a specific arrangement of highlights. This arrangement of highlights is then increased utilizing a specific (e.g., insatiable) methodology, and tried to see the nature of the characterization gets to the next level. Since the characterization calculation is utilized for assessment, this approach will largely make a list of capabilities, which is delicate to the order calculation. This approach has been viewed as helpful practically speaking, due to the wide variety of models on information grouping. For instance, a SVM will quite often lean towards highlights in which the two classes separate out utilizing a straight model, while a

closest neighbor classifier would favor highlights in which the various classes are grouped into round districts.

2.2.2 Probabilistic Methods

Probabilistic methods are the most key among all information order strategies. Calculations of probabilistic orders use measurable derivation to find the optimum class for a given model. Probabilistic characterization calculations will produce a related back likelihood of the test example being an individual from each of the probable classes, as well as effectively relegating the best class like other grouping calculations. The back likelihood is characterized as the likelihood subsequent to notice the particular attributes of the test occasion. Then again, the earlier likelihood is just the small portion of preparing records having a place with every specific class, without any information on the test occasion. In the wake of acquiring the back probabilities, these can be involved the choice hypothesis to decide class enrollment for each new example.

2.2.3 Time Series and Sequence Data Classification

Both of these information types are worldly information types in which the traits are of two kinds. The main sort is the logical quality (time), and the subsequent property, which relates to the time series esteem, is the social characteristic. The primary distinction between time series and grouping information is that time series information is nonstop, though succession information is discrete. In any case, this distinction is very critical, on the grounds that it changes the idea that normally involved models in two different situations.

Time series information is well known in numerous applications, for example, sensor organizations, and clinical informatics, in which involving enormous volumes of streaming time series information to play out the classification is alluring. Two sorts of characterization are conceivable with time-series information:

Classifying specific time-instants: These relate to explicit occasions that can be gathered at specific moments of the information stream. In these cases, the names are related with moments in time, and the way of behaving of one or additional time series are utilized to characterize these moments. For instance, the location of critical occasions progressively applications can be a significant application in this situation.

Grouping part or entire series: In these cases, the class names are related with segments or the series, and these are all utilized for characterization.

2.3 Variations on Data Classification

Many natural variations of the information arrangement issue compare to either little varieties of the standard characterization issue or are upgrades of grouping with the utilization of extra information. The critical varieties of the grouping issue are those of interesting class learning and distance capability learning. Metacalculations, additional information in tactics like move learning and co-preparing, active learning, and human mediation in visual learning are all used to improve the information grouping problem. In terms of information order, the topic of model evaluation is also important. This is due to the fact that the model evaluation issue is crucial for the strategy using compelling order meta-calculations.

2.3.1 Rare Class Learning

Rare class learning is a sizable subset of the characterization problem that is closely related to exception investigation. In actuality, it might be considered a regulated variation of the exception finding problem. In unusual class learning, the information is delivered in a way that is extremely unbalanced, making it typically more important to choose the correct positive class. Consider the scenario where categorizing patients into risky and routine categories is appealing. In these situations, the majority of patients may be considered the scenario where categorizing patients into risky and routine categories is appealing. In these situations, the majority of patients may be average, but misclassifying a seriously dangerous patient is typically much more expensive (misleading negative)average, but misclassifying a seriously dangerous patient is typically much more expensive (misleading negative). Accordingly, False negatives are more costly than false benefits. The issue is strongly linked to cost-delicate learning, because misclassification of multiple classes results in multiple classes.

The significant distinction with the standard arrangement issue is that the goal capability of the issue should be adjusted with costs. This gives a few roads that can be utilized to take care of this issue successfully:

- **Model Weighting**: Due to the cost of misclassification, the models are weighted in an unexpected manner in this circumstance. The majority of characterization calculations, which are typically simple to carry out, undergo minor modifications as a result.In a SVM classifier, for instance, the goal capability ought to be properly weighted with costs, whereas in a choice tree, the split rule measurement ought to weight the models with costs. When selecting the class with the greatest presence in a closest neighbor classifier, the k closest neighbors are appropriately weighted.
- **Model Re-examining**: In this case, the models are properly retested, with the goal of over-inspecting the interesting classes and under-inspecting the typical classes. The retested data are subjected to a standard classifier with little to no change. This method is compared to model weighting from a specialized perspective. However, from a computational perspective, such a method benefits from the fact that the recently retested data is significantly smaller.This is because the majority of the models in the data belong to the ordinary class, which has not been thoroughly tested, whereas the uncommon class is typically only slightly overexamined.

Many variations of the rare class detection problem are possible, in which either examples of a single class are available, or the normal class is contaminated with rare class examples.

2.3.2 Distance Function Learning

A key problem that is closely related to information characterization is distance function learning. In this situation, it is appealing to connect collections of informational occurrences over a distance by using either guided or unguided procedures. Consideration is given to the example of an image collection, where the similitude is described in accordance with the client-focused semantic rule. In this instance, using standard distance is effective since Euclidian measurements struggle to accurately reflect the semantic similarities between two images because they rely on human perception to attempt to change from one variety to another. Accordingly, expressing the fusion of human critique with learned experience is the most efficient way to address this problem.

This criticism is frequently consolidated by using either sets of images with clear distance values or evaluations of numerous images relative to a certain objective image. Such a methodology is used to prepare material that is used for the ultimate purpose of learning, and it may be applied to a wide range of information spaces.

2.3.3 Ensemble Learning for Data Classification

A meta-algorithm is a classification method that makes use of at least one previously performed arrangement calculation by using either different power models or combining the results of previous calculations with different pieces of data. The calculation's main goal is to combine the outcomes from several preparation models, either sequentially or independently, to provide more robust results. A grouping model's common mistake is focusing on predisposition and change, despite the information's inherent turbulence. The inclination of a classifier is based on how the constrained options of a certain model could not compare to the true choice limit. Of the classes, it is possible to see the Bayes likelihood of class I:

For example, even if the prepared data does not have a straight choice limit, an SVM classifier must have one. The arbitrary variants in the particular preparation information index determine the difference. Smaller preparatory information indexes will change more dramatically. Different group research endeavors are used to lessen and alter this tendency. An amazing conversation between inclination and change is expressed by the perceptual user. In order to obtain more precise results from various information mining issues, meta-calculations, such as grouping and exception examination, are frequently used.

Due to its novel assessment metrics and comparatively easy merging of the results of numerous calculations, the area of grouping is the most lavish one from the perspective of meta-calculations. The following are a few well-known meta-calculations:

- **Boosting**: Boosting is a common technique used in grouping. The idea is to focus on the development of challenging sections of the informative collection to create models that can more accurately sort the facts of interest in these bits and then use them for the outfit scores over each component. For each section of the informational index, the incorrectly grouped samples are selected using a hold-out strategy. The idea is to successively select stronger classifiers for

increasingly challenging informational components, combine the results, and then produce a meta-classifier that performs brilliantly on all components of the information.

- **Bagging**: Bagging is a process that combines the results from models built using numerous examples and works with arbitrary information tests. Each classifier's prepared models are selected by examining them using substitution. We refer to these as "bootstrap tests. This strategy has frequently been demonstrated to have favorable results in particular circumstances, although this is not typically the case. Due to the specific arbitrary components of the preparation knowledge, this strategy can minimize change but is unsuccessful at reducing predisposition.
- **Random Forests**: Using irregular subsets of the preparation information or sets of decision trees on either part with produced vectors, specifically, random forests, is a method that registers the score as a component of these varied sections. The irregular vectors are typically created based on an appropriate likelihood distribution. Therefore, either irregular split choice or irregular information determination can be used to create arbitrary timberlands. According to how the example is picked, irregular backwoods will be seen as a special example of irregular backwoods because irregular woods are closely associated with sacking (bootstrapping).
- **Model Averaging and Combination**: Perhaps the most well-known model used in troupe assessment is this one. In actuality, the unconventional wood method we looked at above is a remarkable illustration of this idea. There are many Bayesian strategies for the model mix process that can be used to address the grouping problem. The use of many models ensures that the error caused by a particular classifier's bias does not dominate the order results.
- **Stacking**: A second-level classifier is used to play out the mix in techniques like stacking that combine numerous models in different ways. A second-level classifier's element depiction is created using the output of different first-level classifiers. These first-level classifiers could be chosen in a variety of ways, such as by using various stored classifiers or various preparation models. The preparation data should be divided into two subsets for the first and second level classifiers in order to prevent overfitting.

The "wait" portion of the informative index is used in the "bucket of models" strategy to select the best model. The most accurate model is one in which the held-out informational index achieves the highest level of accuracy. Generally speaking, this strategy can be viewed as a competition or heat off challenge amongst the several models.

2.4 Bernoulli Multivariate Model

This group of algorithms views a record as a collection of recognizable words lacking any recurring information, where a component (term) may be present or absent. Permit us to assume that the terms' derived vocabulary is denoted by V = t1... tn. Let's assume that the class is drawn from "1... k" and that the "sack of words" (or text record) being referred to contains the terms Q = "ti1... tim". The system's procedures will then show the likelihood that the report, which is assumed to have been created from one class's term circulations, belongs to class I because it contains the terms Q = ti1... tim. The best way to understand the Bayes technique is to think of it as an examining/generative interaction from the fundamental combination model of classes. By looking at a number of terms T from the term distribution of the classes, it is possible to see the Bayes likelihood of class I:

What is the likelihood that I initially chose class I for checking if a term set T of any size was examined from the term conveyance of one of the randomly chosen classes, and the end result was the set Q? Class I's fragmentary presence in the assortment is equivalent to the deduced likelihood of choosing it.

The studied set T's class is denoted by CT, and the corresponding back likelihood is denoted by P(CT = i|T = Q). The fact that we are essentially selecting a subset of terms from V without any frequencies associated with the selected phrases must be kept in mind because there is no allowance for substitution. As a result, the set Q might not have copy components. This is essentially equivalent to either picking or not choosing each term with a likelihood that depends on the basic term circulation under the fallacious Bayes presumption of freedom between terms. It is also very important to remember that there is no restriction on the number of terms that can be chosen in this model. The key differences between these assumptions and the multinomial Bayes model. The Bayes approach requires us to record the following

two probabilities in order to group a given set Q in light of the likelihood that Q is an illustration of the information dispersion of class I, i.e., P (CT = i|T = Q).

How likely is it that a set T is an example of the class I term circulation earlier? This probability is denoted by the symbol P(CT = I). What is the likelihood that the set Q is the example when a set T of any size is studied from the term dispersion of class I? This probability is denoted by P(T = Q|CT = I).

A more numerical portrayal of Bayes demonstrating will presently be given. At the end of the day, what will be displayed is P(CT = i|Q is tested). This limited probability can be composed using the Bayes rule and evaluated more efficiently using the fundamental corpus. As a result, the streamlined will look like this:

$$P(C^T = i|T = Q) = \frac{P(C^T = i) \cdot P(T = Q|C^T = i)}{P(T = Q)} \qquad 2.4$$
$$= \frac{P(C^T = i) \cdot \prod_{t_j \in Q} P(t_j \in T|C^T = i) \cdot \prod_{t_j \notin Q} (1 - P(t_j \in T|C^T = i))}{P(T = Q)}$$

The final condition in the preceding succession makes use of the fallacious freedom presumption that the probability of events for the distinct terms is independent of one another. To transform the likelihood conditions into a structure that can be evaluated using the fundamental data, this is significantly needed.

The class assigned to Q has the most notable back probability in light of Q. It is clear that the denominator, the small possibility of recognizing Q, has no bearing on this decision. In other words, Q will receive the corresponding class.

$$\begin{aligned} \hat{i} &= \arg\max_i P(C^T = i|T = Q) \\ &= \arg\max_i P(C^T = i) \cdot \\ &\quad \prod_{t_j \in Q} P(t_j \in T|C^T = i) \cdot \prod_{t_j \notin Q} (1 - P(t_j \in T|C^T = i)) \end{aligned} \qquad 2.5$$

It is important to keep in mind that all phrases on the last condition's right handside can be evaluated using the prepared corpus. The value of P (t j T|CT = I) is the insignificant percentage of reports in the ith class that contain the term t j, whereas the value of P (CT = I) is the global portion of archives that belong to class I. The evaluations of the corresponding probabilities in the aforementioned notes are all at their most severe. In the end, Laplacian smoothing is eventually used, in which minor features are given to word frequencies to prevent the absence of terms that are merely

present. The normalizer P(T = Q) does not need to be calculated because the personality of the class with the highest likelihood esteem, as opposed to the genuine likelihood esteem associated with it, is taken into account in many applications of the Bayes classifier. By using the logarithm of the Bayes expression and removing terms that have no bearing on the requesting of class probabilities, significant gains are actually possible in registering these Bayes "likelihood" values due to parallel classes.

Despite the fact that the back probability P(CT = i|T = Q) must be calculated specifically for grouping P(T = Q), some applications do. For instance, the specified back likelihood esteem P(CT = i|T = Q) must logically examine the likelihood esteem over many test cases and score them for their irregular nature in order to be used for administered oddity finding (or uncommon class recognition). In these scenarios, we would need to calculate P(T = Q). Basically, adding up the grades from all the classes is one way to do this.

$$P(T = Q) = \sum_{i} P(T = Q|C^T = i)P(C^T = i) \qquad 2.6$$

This is dependent on how freely constrained each class's elements are. Since each class's boundary values are evaluated separately, the problem of an information-poor situation might be addressed. Making the assumption of (universal) freedom of terms and processing it as follows is an optional way to register it that might help with the information adequacy issue.

$$P(T = Q) = \prod_{j \in Q} P(t_j \in T) \cdot \prod_{t_j \notin Q} (1 - P(t_j \in T)) \qquad 2.7$$

where the term probabilities are based on global term transfers across all classes. A common question that arises is whether it is possible to design a Bayes classifier that models the conditions that exist between the words during the order interaction without using the guileless suspicion. Due to the greater processing costs and inability to accurately and robustly estimate the boundaries in the presence of limited information, methods that sum up the gullible Bayes classifier without using the freedom suspicion perform poorly. One limitation results in a Bayesian organization model that is computationally expensive due to a suspicion of total reliance.

Then again, it has been shown that permitting restricted degrees of reliance can give great tradeoffs among precision and computational expenses.

Even if the freedom assumption is just a reasonable guess, it has been demonstrated that the methodology has some fictitious support. Undoubtedly, many

exploratory experiments would have generally demonstrated the guileless classifier's excellent practical performance.

By creating new elements for each of these attributes, the Bayes methodology provides a characteristic mechanism for incorporating such additional data into the grouping system. The enlarged portrayal of order is then used in conjunction with the conventional Bayes approach. The Bayes approach has also been applied to the fusion of various types of spatial information, such as the integration of link data into the grouping system.

Progressive order and the Bayes approach both work well together, and the preparation data is arranged according to a scientific subject classification. For example, there are vast collections of reports that are arranged into different leveled groups on the Open Directory Project (ODP), Yahoo! Scientific classification, and many information websites. Since it has been demonstrated that using delicate component determination can produce more beneficial grouping results, it is possible to accomplish more successful arranging by taking advantage of the points' progressive design. For progressive grouping, a Bayes classifier is used at each hub, providing the next branch for ordering reasons. Two such solutions are put forward, in which the arrangement cycle makes use of hub explicit features. Since the elements chosen are relevant to that branch, it goes without saying that there will be many fewer highlights at a certain hub in the order.

2.5 Multinomial Distribution

This class of methods regards a record as a bunch of words with frequencies connected to each word. Subsequently, the arrangement of words is permitted to have copy components. As in the past case, the arrangement of words in archive is meant by Q, drawn from the jargon set V. The set Q contains the unmistakable terms {ti1 . . .tim} with related frequencies F = {Fi1 . . .Fim}. The terms and their frequencies indicated by [Q,F]. The complete number of terms in the archive (or record length) is signified by L = Σmj =1 F(i j). Then, it is likely to show the back likelihood that the record T has a place with class I, considering that it contains the terms in Q with the related frequencies F. The Bayes likelihood of class I can be demonstrated by utilizing the accompanying examining process:

If the sampled L terms sequentially from the term distribution are randomly chosen classes (allowing repetitions) to create the term set T, and the final outcome for sampled set T is the set Q with the corresponding frequencies F, then what is the posterior probability that we had originally picked class i for sampling? The a-priori probability of picking class i is equal to its fractional presence in the collection.

The previously mentioned likelihood is indicated by P(CT = i|T = [Q,F]). A supposition that is usually utilized in these models is that the length of the report is free of the class name. While it is effectively conceivable to sum up the strategy, with the goal that the report length is utilized as an earlier, freedom is normally expected for effortlessness. As in the past case, the two qualities to register the Bayes back will be appraised.

1. What is the prior probability that a set *T* is a sample from the term distribution of class *i*? This probability is denoted by $P(C^T = i)$.
2. If the sample *L* terms *from the term distribution of class i* (with repetitions), then what is the probability that the sampled set *T* is the set *Q* with associated frequencies *F*? This probability is denoted by $P(T = [Q,F]|C^T = i)$.

$$P(C^T = i|T = [Q,F]) = \frac{P(C^T = i) \cdot P(T = [Q,F]|C^T = i)}{P(T = [Q,F])} \propto P(C^T = i) \cdot P(T = [Q,F]|C^T = i). \quad 2.8$$

As in the past case, it isn't important to register the denominator, P(T = [Q,F]), to conclude the class name for Q. The worth of the likelihood P(CT = I) can be assessed as the negligible part of archives having a place with class I. The calculation of P([Q,F]|CT = I) is more confounded. At the point when the consecutive request of the L various examples is considered, the quantity of potential ways of testing the various terms in order to bring about the result [Q, F] is given by L! Πmi =1 Fi! . The likelihood of every one of these groupings is given by $\Pi t\ j \in QP(t\ j \in T)Fj$, by utilizing the gullible freedom supposition. In this way, there is the equation:

$$P(T = [Q,F]|C^T = i) = \frac{L!}{\prod_{i=1}^{m} F_i!} \cdot \prod_{t_j \in Q} P(t_j \in T|C^T = i)^{F_j}. \quad 2.9$$

Substitute Condition 3.20 in Equation 3.19 to get the class with the most noteworthy Bayes back likelihood, where the class priors are figured as in the past

case, and the probabilities P(t j ∈ T|CT = I) can likewise be effortlessly assessed as already with Laplacian smoothing. Note that to pick the class with the most elevated back likelihood; it does not actually need to register L! ΠmI =1 Fi!, as it is a consistent not relying upon the class name (i.e., the equivalent for every one of the classes). In addition, it could be noted that the probabilities of class nonattendance are absent in the above conditions in view of the manner by which the testing is performed.

Various varieties of the multinomial model have been proposed. In the work, it is demonstrated the way that a class order can be utilized to work on the gauge of multinomial boundaries in the gullible Bayes classifier to further develop grouping precision fundamentally. The key thought is to apply shrinkage methods to smooth the boundaries for information scanty kid classifications with their normal parent hubs. Subsequently, the preparation information of related classes are basically "shared" with one another in a weighted way, which works on the vigor and precision of boundary assessment when there are deficient preparation information for every individual youngster classification. The work has played out a broad examination between the

Bernoulli and the multinomial models on various corpora, and the accompanying ends were introduced:

The multi-variate Bernoulli model can in some cases perform better compared to the multinomial model at little jargon sizes.

The multinomial model outflanks the multi-variate Bernoulli model for huge jargon sizes, and quite often beats the multi-variate Bernoulli when jargon size is decided ideally for both. On the normal a 27% decrease in mistake.

The in advance of referenced results strongly imply that the two models might have various qualities, and may hence be valuable in various situations.

2.6 Water scarcity and climate change

Climate is expected to continue to change in the future, in spite of the fact that there are still many uncertainties, which will affect natural and human systems such as forestry, fisheries, water resources, human settlements, and human health (IPCC, 2001). (IPCC, 2001). Global surface temperature has risen by 0.74 oC in the past 100 years, with temperatures increasing more rapidly in the past 50 years. Heat and water

are closely linked, and in recent decades, warming trends have led to changes in the hydrologic cycle (Intergovernmental Panel on Climate Change (IPCC, 2007). Water scarcity can be brought on by disturbances in the hydrological cycle in a given area.

Water scarcity refers to the relative shortage of water in a water supply system that may lead to limits on consumption. The degree to which demand exceeds the amount of resources available is known as scarcity. It can be brought on by droughts as well as human activities like population growth, water waste, and unequal access to resources. Most of the Mediterranean countries are facing water scarcity. Whereas, drought is a recurrent feature of climate that is characterized by temporary water shortages relative to normal supply, over an extended period of time – a season, a year, or several years. The term is relative, since droughts differ in extent, duration, and intensity. According to World Resources Institute (WRI), the world's water systems face formidable threats. More than a billion people currently live in water-scarce regions, and as many as 3.5 billion could experience water scarcity by 2025. Increasing pollution degrades freshwater and coastal aquatic ecosystems. Economic expansion is probably causing an increase in global water use.

In most countries, except for a few industrialized nations, water use has increased over recent decades due to population and economic growth, changes in lifestyle, and expanded water supply systems, with irrigation water use being by far the most important cause. Irrigation accounts for about 70% of total water withdrawals worldwide and for more than 90% of consumptive water use (i.e., the water volume that is not available for reuse downstream) (Bates et al., 2008). (Bates et al., 2008). The water scarcity has even affected other temperate regions with generally copious resources, such as Europe and North America, where times of drought are growing more common and are lasting longer. The level of groundwater supplies has reached a critical point in several areas of Italy, France, Spain, and the United Kingdom as a result of repeated droughts over the past few decades. Some watercourses have dried up. Numerous forces, such as water pollution, water scarcity, and floods, have an impact on Europe's waters. Morphology and water flow are also impacted by significant changes to water bodies (EEA, 2012).

The Mediterranean region is undergoing rapid local and global social and environmental changes. All evidence point to an increase in environmental and water scarcity concerns with negative implications towards present and future sustainability. Pressures related to water scarcity do not apply equally to all sectors of water use in

the Mediterranean region. It is suggested that managing the risk of water scarcity through preparedness rather than a crisis approach and emphasizing local management at the basin scale (Iglesias et al., 2007).

Climate change raises water resource strains in various places of the world when runoff diminishes, particularly around the Mediterranean, in portions of Europe, central and southern America, and southern Africa. In other water-stressed parts of the world—particularly in southern and eastern Asia—climate change increases runoff, but this may not be very beneficial in practice because the increases tend to come during the wet season and the extra water may not be available during the dry season (Arnell, 2004). (Arnell, 2004). Future population growth will increase the pressure on available water resources in many countries as well as globally. The management of water resources by nations around the world is becoming more and more important.

Climate change will increase water temperature and the likelihood of flooding, droughts and water scarcity in the years to come. There are many indications that water bodies already under stress from pressures are highly susceptible to climate change impacts, and that climate change may hinder attempts to restore some water bodies to good status. Preparing for climate change is a big concern for water management in Europe. The Water Framework Directives (WFD) is the first piece of European environmental legislation that addresses hydro morphological pressures and impacts on water bodies. When hydro morphological pressures interfere with the ability to accomplish WFD goals by affecting the ecological status, it is necessary to take action. Execution of the WFD is to be performed through the River Basin Management Plan (RBMP) procedure, which mandates the preparation, implementation and assessment of an RBMP every six years. Water resource management needs an integrated aspect of the RBMP. In more arid river basins, such as in the Mediterranean, drought management plans are already partly integrated into RBM planning.

The major purpose of EU water policy is to ensure that throughout the EU, a sufficient quantity of good-quality water is accessible for people's needs and for the environment. The WFD, which came into force on 22 December 2000, establishes a new framework for the management, protection and improvement of the quantity and quality of water resources across the EU. EU Member States shall seek to achieve excellent status in all bodies of surface water and groundwater by 2015 unless there are grounds for derogation. Only in this case may achievement of good status be

extended to 2021 or by 2027 at the latest. Achieving good status involves meeting certain standards for the ecology, chemistry, morphology and quantity of waters. In general terms, 'good status' means that water shows only a slight change from what would normally be expected under undisturbed conditions. There is also a general 'no deterioration' clause to prevent deterioration in status (EEA 2012). (EEA 2012)

2.7 Water allocation

Water allocation describes a process whereby an available water resource is distributed to legitimate claimants and the resulting water rights are granted, transferred, reviewed, and adapted (Quesne et al., 2007). Allocation of water among competing uses (industry, agriculture, and municipal) is already under pressure due to demographic change, increased environmental awareness, and changing patterns of water demand. Some of the world's major rivers are now completely dry for stretches or periods of time. But the most effective means of allocating water will always be determined by local demographic, environmental, political, and social circumstances; there is no single approach that can simply be replicated globally. Meeting the new challenges on water resources management, implies the quantification of climate change impact on basin scale hydrology (Varies et al., 2004).

Hydrologic analysis of climate change scenarios indicate possible reductions in stream flows in some areas, increased flood frequencies in other areas, and changes in the seasonal pattern of flows, with reduced summer flows likely in many semi-arid and Mediterranean river basins. By the middle of the 21st century, annual average river runoff and water availability are projected to increase as a result of climate change at high latitudes and in some wet tropical areas, and decrease over some dry regions at mid-latitudes and in the dry tropics (Bates et al., 2008).

This phenomenon is making water allocation in summer more difficult and challenging. Initially, sufficient water is available to meet the needs of all water sectors within a catchment without jeopardising hydrological ecosystems. As a consequence, little management is required. An growth in agricultural and industrial activity coupled with population change contribute to growing water demands.

In Spain, there have been 9 River Basin Organizations (RBOs) since the 1920's for the development and allocation of water resources and the control of water use and pollution at basin level, with water user participation in governing bodies and advisory stakeholder participation at national and basin levels(GWP, 2002). (GWP,

2002). It is obvious that allocation processes usually arise from a familiar pattern in the development of water use. However, some augmentation of supply through engineering approaches is usually possible to meet increased demand, like alternated water resources, notably the construction of increased storage capacity and inter-basin transfer etc.

With the increased population growth rates, change in climate and improved life style, the competition over scarce water resources is increasing. The Mediterranean is expected to experience a rise in water scarcity, which will significantly increase the demand for effective water allocation systems. However in certain cases, a new and more sophisticated strategy to water management is necessary instead of engineering technique when water stress is achieved. These strategies aim to manage water resources in a multi-disciplinary and multi-stakeholder manner in order to efficiently distribute water and restore river flow.

2.8 Decision Support System

It is important to promote efficient use of water through better management of water resources, for social and economical sustainability in arid and semi-arid areas, under the conditions of severe water shortage. In a water-stressed area, however, the process of selecting the appropriate intervention for any circumstance involves multiple stages.In order to deal with complex issues pertaining to water resources, the disciplines, individuals, and institutions needed to do so must collaborate on the creation, development, and implementation of efficient decision support systems. The purpose of decision support systems, or DSS, is to provide decision makers with accurate and sufficient information.

Decision support systems are integral parts of Integrated Water Resource Management IWRM processes facilitating the use of science and technology advances in public policy. A key effort during the DSS development phase is to determine the necessary and necessary information that decision makers need to make good decisions. It is anticipated that this information set will vary depending on the type of decision (planning, management, or near real time), the management agency, and the stakeholder group. Important interaction should take place with the decision makers who are tasked with determining which information is most appropriate. These concerns are currently being addressed by multiple environmental evaluations and

research initiatives at all pertinent scales, frequently in conjunction with other disciplines. However, integrating this wealth of information across disciplines remains a considerable challenge (Millenium Ecosystem Assessment, 2003). (Millennium Ecosystem Assessment, 2003).

2.9 Climate Change Standards

Global warming and climate change have significantly increased the pressure on ecosystem services provided by the hydrological cycle. Over the past 20 years, a number of categories and criteria for climate change have been created to assess and quantify these problems. In 1988, the IPCC (Intergovernmental Panel on Climate Change) was established. It was established by the World Meteorological Organization (WMO) and the United Nations Environment Program (UNEP) to prepare assessments on all aspects of climate change and its effects based on available scientific information in order to formulate realistic response strategies. Today the IPCC's role is as defined in Principles Governing IPCC Work, "...to assess on a comprehensive, objective, open and transparent basis the scientific, technical and socio-economic information relevant to understanding the scientific basis of risk of human-induced climate change, its potential impacts and options for adaptation and mitigation".

IPCC reports should be neutral with respect to policy, although they may need to deal objectively with scientific, technical and socioeconomic factors relevant to the application of particular policies." Through the IPCC, climate experts from around the world synthesize the most recent climate science findings every five to seven years and present their report to the world's political leaders. The IPCC has issued comprehensive assessments in 1990, 1996, 2002 and most recently the Fourth Assessment Report (AR4) released in 2007 (IPCC, 2007).

The evaluation of climate change impacts on ecosystem services provision shows that water provisioning and erosion control are highly sensitive to climate change in the Llobregat river basin. A review study (Gosling, 2013) found a proportionally larger amount of evidence to suggest that ecosystem services are vulnerable to changes in the large-scale climate-earth system in the Mediterranean region. Services supply and delivery are likely to reduce by significant amounts, indicating that urgent measures must be taken to avoid future water stress in the basin.

The sub-watersheds from the Pyrenees region are responsible for most of the services provision, and are also the most impacted areas regarding climate change. Interventions to enhance the provision of regulating services should focus in certain areas where obtained benefits per surface area are estimated to be the highest. For the protection of these areas, interventions such as restoration and measures suitable for increasing or maintaining resilience in rivers are essential to assure future water use in the basin. The groundwater–surface water interplay and the temporal nature of water demands in the Mediterranean region also lend complexity to the system (Bangash et al., 2012). The aim of the study was the detection of change in trends over time and the quantification of ecosystem service provisioning under climate change impact. The results show clear trends over time, with decreases in water yield and the amounts of sediment retained being two orders of magnitude higher than that exported.

Climate change is the only variable and driving force considered in this study. Other important drivers of change, as land use and land cover, and the increase in demand by different sectors are considered constant for the whole catchment. However, it is obvious that the protection of water resources is not sufficient if the levels of consumption continue to increase in the future. Proactive management of basin should be implemented for adapting to climate change as mitigating measures taken in the present may avoid long-term future consequences.

CHAPTER 3
METHODOLOGY

The K-Nearest-Neighbors (KNN) is a non-parametric request computation, for instance, it makes no suppositions on the simple dataset. It is known for its ease and reasonability. It is a managed learning estimation. A noticeable planning dataset is given where the data centers are requested into various classes, so that class of the unlabeled data can be expected.

In Grouping, different properties conclude the class to which the unlabeled data has a spot. KNN is by and large used as a classifier. Gathering data considering closest or connecting getting ready models in a given locale is used. This procedure is used for its straightforwardness of execution and low computation time. For unending data, it uses the Euclidean distance to register its nearest neighbors. For another data the K nearest still up in the air and the larger part among the connecting data picks the game plan for the new data. Regardless of the way that this classifier is essential, the value of 'K' expects a huge part in requesting the unlabeled data. There are various approaches to pick the characteristics for 'K', but the programs can essentially run the classifier on different events with different characteristics to see which worth gives the best result. The computation cost is to some degree high since all of the assessments are made while the planning data is being organized, not when it is knowledgeable about the dataset.

It is a dormant learning estimation as not much is done when the dataset is being ready except for taking care of the planning data and holding the dataset in light of everything. It does not perform hypothesis on the arrangement dataset. So the entire focal dataset being arranged is required when in the testing stage. In backslide, KNN predicts steady characteristics. This value is the ordinary of the potential gains of its K - nearest neighbor.

3.1. Development of KNN

K-closest neighbor game plan was made to execute brand name assessment when clear parametric approximations of probability densities were dark or difficult to choose. In an unpublished US Flying corps School of Flight Medication report in

1951, Fix and Hodges introduced a non-parametric computation for configuration gathering that has since become realized the K-nearest neighbor rule.

3.2. Tasks of kNN

KNN is a classification algorithm. Mainly there are two steps in classification:

1. Learning Step: Using the training data a classifier is constructed.

2. Assessment of the classifier.

As shown the nearest neighbor technique, the new unlabeled data is organized by sorting out which classes its neighbors have a spot with. KNN estimation involves this thought in its calculation. In case of KNN estimation, a particular worth of K is fixed which helps us in requesting the dark tuple. When a new unlabeled tuple is knowledgeable about the dataset, KNN performs two undertakings:

- In the first place, it separates the K concentrates closest to the new information of interest, i.e., the K nearest neighbors.
- Second, using the neighbors' classes, KNN chooses concerning which class should the new data be organized into.

Right when a couple of new data is added, it describes the data likewise. It is more important in a dataset which is by and large parceled into gatherings and has a spot with a specific region of the data plot. Thusly this estimation gets more precision secluding the data inputs into different classes in an all the clearer way. KNN figures out the class having the best number of centers sharing insignificant division from the data guide that prerequisites toward be requested. In this way, the Euclidean distance ought not to set in stone between the test and the foreordained arrangement tests.

After K-Closest Neighbors have collected, we basically take a large portion of them to expect the class of the readiness model. The factors that impact the introduction of KNN are: the value of K, the Euclidean distance and the normalization of the limits. To understand the distinct working of the estimation, the means are according to the accompanying:

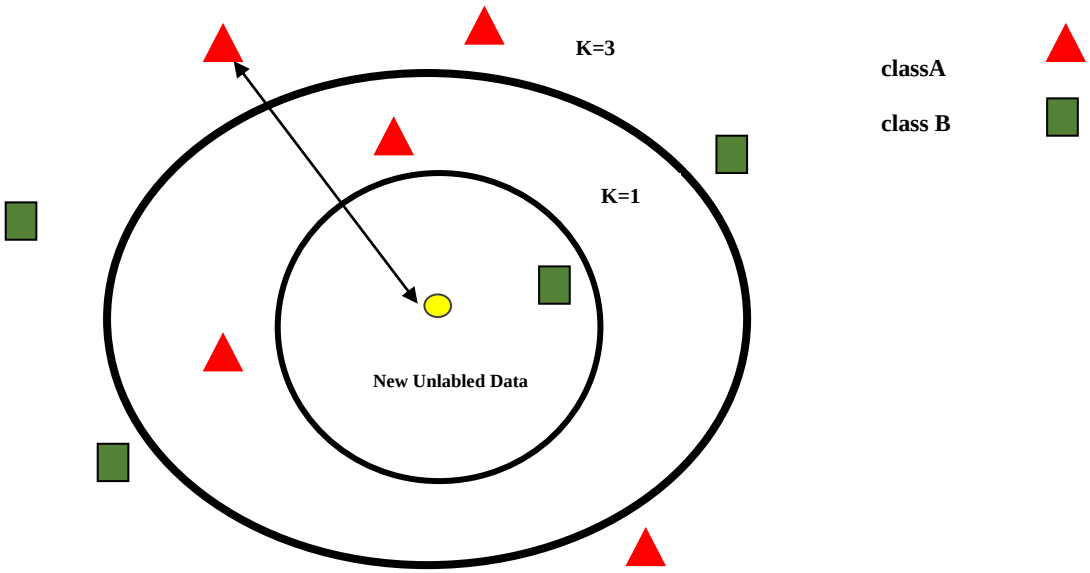


Figure 3.1 Sample Classification of KNN

Given the training dataset: { (x(1), y(1)) , (x(2), y(2)), ... , (x(m), y(m)) }

Step1: Store the training set

Step2: For each new unlabeled data,

A. Calculate Euclidean distance with all training data points using the formula

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad 3.1$$

B. Find the k- nearest neighbors

C. Assign class containing the maximum number of nearest neighbors.

In the wake of taking care of the readiness, put down all stopping points ought to be normalized, with the objective that the assessments become more direct. The outcome of the gathering is sensitive to the value of 'K'. The data variable 'K' finishes up the amount of neighbors that ought to be considered. The value of 'K' influences the estimation as using the 'K' regard that can create the constraints of each class.

TO DETERMINE K: The best worth of K is picked by first investigating the data Bigger potential gains of K are more definite as they diminish the net upheaval anyway this is not guaranteed. A nice worth of K can in like manner be settled using cross endorsement. If K=1, the data is fundamentally allocated to the class of its nearest neighbor. At K=1, the slip-up rate is dependably zero for the planning data. This occurs considering the way that the nearest feature any planning data point is itself. Thusly the best results are gained if the value of K=1. However, with K=1, the cutoff points are over fitted.

In case of small potential gains of 'k' the estimation is excessively fragile to try and consider noising. To get a decent worth of K, the readiness and endorsement set ought to be separated from the basic dataset. Expecting the two Closest neighbors (K=2) have a spot with two remarkable classes, the outcome is dark. Along these lines, we increase the amount of nearest neighbors to a greater worth (say 5-nearest neighbors). This will portray an earliest neighbor region and will give the clarity.

Greater potential gains of 'K' make as far as possible smoother, which presumably would not be appealing as then the signs of various classes could get associated with the area. While the readiness data centers are accessible in a scattered manner, the value of K is trying to choose.

3.3. Advantages

KNN is known for its ease, understandability and flexibility. It is easy to interpret. The assessment time is less. In like manner the judicious power is uncommonly high which makes it convincing and compelling. KNN is extraordinarily convincing for enormous planning sets. The means went on in the gathering done by this computation are modestly less puzzling than that followed by various estimations.

The mathematical computations are easy to understand and fathom. They do exclude calculations that give off an impression of being irksome. Basic thoughts like that of Euclidean distance assessment are used which update the ease of the estimation rather than choosing other composite techniques like that of consolidation or division. It is significant for non-direct data. KNN is convincing for portrayal as well as backslide.

3.4. Disadvantages

KNN can be luxurious in confirmation of K if the dataset is immense. It needs a more unmistakable storing than a strong classifier. In kNN the assumption stage is postponed for a greater dataset. Moreover, computation of careful distances expects a significant part in the confirmation of the estimation's accuracy. One of the critical stages in kNN is choosing the limit K. From time to time, it is not clear which sort of distance to use and what part will give the best result. The computation cost is exceptionally high as the distance of each getting ready model is not entirely set in stone. KNN is a lazy acquiring computation as it does not acquire from the planning data, it recommends holds it and a short time later uses that data to portray the new data.

3.5. kNN and its Variants

As examined before, the adequacy of the estimation can be improved by making changes in the factors that administer it. There are various varieties of KNN that have been concentrated before to make this computation more effective, some of them are:

(1) Locally Adaptive KNN:

Locally adaptable KNN estimations proposed by [1]. It picks the value of k that should be used to arrange a commitment by checking out at the delayed consequences of cross-endorsement estimations in the close by neighborhood of the unlabeled data.

(2) Weight Adjusted KNN:

The calculation by [2] suggests that the distances, on which the mission for the nearest neighbors is arranged in the underlying step, should be changed into practically identical measures, which can be used as burdens. The consigned loads finish up how much a quality effects the portrayal movement. This classifier is particularly significant for the circumstance where a dataset has numerous components, some of which can be considered to be un-fundamental, but it has high computational cost.

(3) Improved KNN for Text Categorization:

[3] proposes a refined KNN computation for text request, which fabricates the portrayal model by consolidating KNN text order and bound one pass gathering estimation. In case a consistent worth of K is used for all of the classes, the class with greater number of properties will partake in an advantage. In better KNN, a sensible number of nearest neighbors are used by the movement of data in planning set, to predict the class of an unlabeled data.

(4) Adaptive KNN:

KNN perceives the same number of nearest neighbors for each new data. Flexible KNN by [4] sorts out a fit worth of K for each test. Starting an ideal worth of K is found. Then, to expect the course of action of the unlabeled data, the value of K is set comparable to the ideal worth of K of it is nearest neighbor in the readiness dataset. The execution of the proposed estimation is then taken a stab at different datasets.

(5) KNN with Shared Nearest Neighbors

A better K-nearest neighbor computation is presented by [5] using split nearest neighbor equivalence which can figure resemblance between test tests with nearest neighbor tests. It uses Comparability judgment computation and works out the nearest neighbor likeness a motivator for each planning test. Then it learns the most outrageous between these characteristics.

(6) KNN with K-Means:

One all the additional method for managing the estimation is depicted by [6]. This estimation endeavors to detach a lot of concentrations into K sets or gatherings so the concentrations in each pack are close to each other. The focal points of these recently complete bunches are taken as the new arrangement tests. To predict the portrayal of an unlabeled data, its detachment from the as of late found planning not entirely set in stone, and the center which shares the base partition from the data is distributed to that class. Not at all like standard KNN, there is the data limit K is not passed. This records to be one of its benefits.

(7) SVM KNN

Support Vector Machine (SVM) is a request procedure that can be applied on straight as well as non-direct data. It is a composite variation of kNN mixed in with SVM for visual grouping affirmation, and is extended in [7]. In this estimation, the readiness is done with the help of K nearest neighbors to the relevant unidentified data. The K-nearest data are not set up permanently in the first place. After that, the pairwise distance that separates these K data centers is dealt with.. In this way we get a distance system from the decided distances. A Bit network is then arranged from the got distance system. This piece structure is dealt with as commitment to SVM classifier. The result gained is the class of the dark snippet of data. On the other hand, one could use SVMs yet time use is one of its disadvantages. Also, it incorporates assessment of pairwise distances.

(8) KNN with Mahalanobis Metric

The estimation distance is basic in gathering of one more information of interest. Mahalanobis is one more distance metric, approach of which is campaigned in [8]. The metric ensures that the K-nearest neighbors are contained in comparative class and the models having a spot with different classes are secluded by a colossal degree of difference.

(9) Generalized kNN

KNN can in like manner be used for consistent - regarded class credits. For this plan, the not entirely set in stone among neighbors is assigned to the class property of the unlabeled data. [9]Implement this computation to predict the steady - regarded class trademark.

(10) Informative kNN

Typically, the value of K relies upon the data, making it hard to pick the limit according to different applications. [10] introduced one more metric that activities the edifying ness of objects to be assembled. Instructiveness appraises the meaning of core interests. In this procedure, there are two data limits K and I. The larger part class of most instructive descending models will be the class of the new test.

(11) Bayesian KNN

The data values including the goal are made by a comparative probability movement, expanding outwards over the sensible number of neighbors. [11] recursively calculated the probability of the last change-point and moved towards the goal, and enrolled the back probability scattering over K.

\

CHAPTER 4
SYSTEM DESIGN AND IMPLEMENTATION

When the climate varies then automatically field parameters also suddenly changes Whenever there is heavy rainfall or temperature varies this may become very hard to analyze the situation and it causes a major problem. Taking this as a problem into consideration, designing of water demand prediction on irrigation system is needed. To get the best representation, it is necessary to choose a time series model that takes these factors into account. This system forecasts the short-term time series of water demand using the k-nearest neighbor (KNN) approach. The projected values of the KNN approach, a pattern recognition method, are directly decided by the prior observations with the highest degree of similarity.

4.1 Design of the System

The proposed water demand prediction on irrigation will be emphasized for two sample training datasets (Plants A dataset and Plant b dataset). Each attribute of the training dataset contents and testing dataset contents distance are calculated by Euclidean distance. Then, the resultant distance values are feed to KNN for prediction water demand by suitable selected K values. For the accuracy evaluation, the confusion matrix will be used. The calculation of confusion matrix is calculated based on true positive, false positive, true negative and false negative.

Data processing state in this system have four steps. Step (1) import the dataset. Step (2) verify the missing value. Dividing dataset into training and testing data in Step (3) and then using the Euclidean distance to predict the final result with selected on K-value in Step (4) .

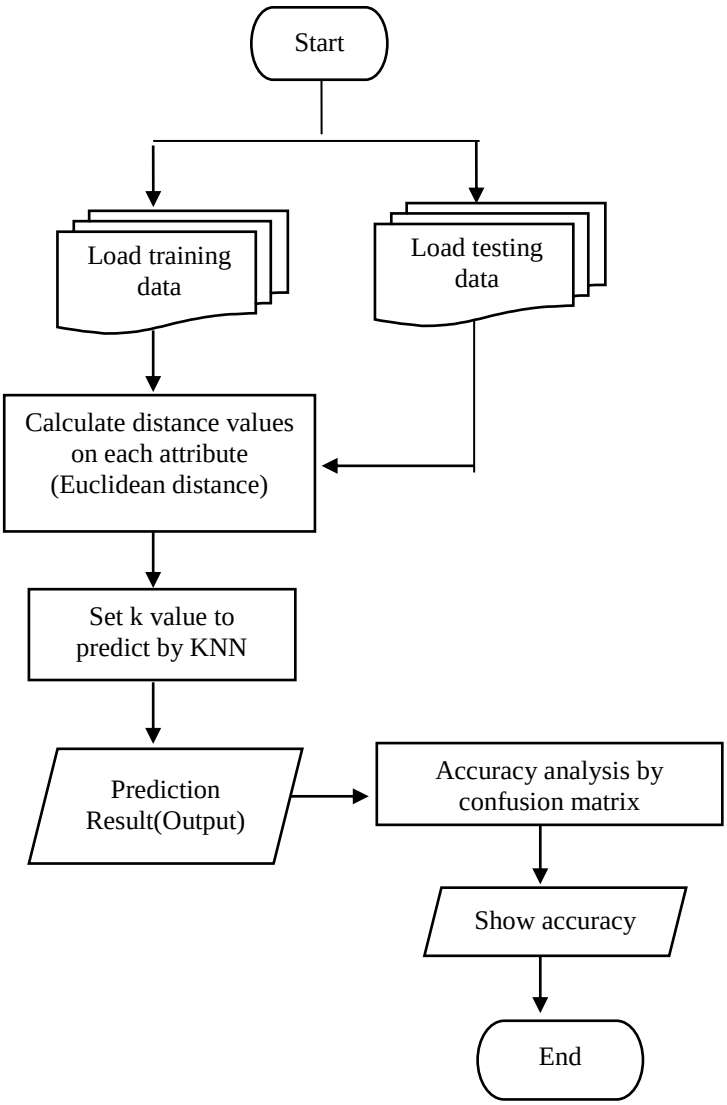


Figure 4.1 System Flow Diagram

4.2 Dataset

In the system two groups of datasets used. In group I, contains four attributes: Temperature, Humidity, Soil Moisture, Wind Speed and one class label: Water needed / not. In group II, contains three attributes: Temperature, Humidity, Soil Moisture, and one class label: Water needed / not. For the purpose of weather conditions: there are many types of attributes such as: Temperature, Humidity, Wind Speed, Press, Wind Direction, rain and so on. This system used only two groups of datasets are as follow:

Table (4.1) Some Data of Banana Dataset (Group I)

ID	Temp	Humidity	Soil Moisture	Wind speed	Plant Type	Remark
1	25	68	37	5	Banana	No Water Need
2	28	84	24.2	6	Banana	Water Need
3	27	94	37.5	5	Banana	No Water Need
4	28	94	21.5	7	Banana	Water Need
5	25	80	43.3	2	Banana	No Water Need
6	31	70	21.5	6	Banana	Water Need
7	29	84	42.5	3	Banana	No Water Need
8	31	66	21.4	5	Banana	Water Need
9	32	71	21.5	12	Banana	Water Need
10	33	59	21.3	2	Banana	Water Need

Table (4.2) Sample Dataset of Watermelon Dataset (Group II)

ID	Temp	Humidity	Soil Moisture	Wind speed	Plant Type	Remark
1	32	63	24.2	Watermelon	Water Need	1
2	31	66	21.2	Watermelon	Water Need	2
3	30	75	24.2	Watermelon	Water Need	3
4	25	70	52	Watermelon	No Water Need	4
5	28	74	24.2	Watermelon	Water Need	5
6	27	79	51.5	Watermelon	No Water Need	6
7	28	94	21.5	Watermelon	Water Need	7
8	25	80	52.3	Watermelon	No Water Need	8
9	30	70	21.5	Watermelon	Water Need	9
10	29	60	52.5	Watermelon	No Water Need	10

4.3 Water Demand Prediction Algorithm

It is a supervised machine learning algorithm which uses proximity to make classifications or predictions about the grouping of an individual data point. Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it The number of nearest neighbors to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'. Its aim is to locate all of the closest neighbors around a new unknown data point in order to figure out what class it belongs to. It is also called a lazy. learner. algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. If k=1, then the object is simply assigned to the class of that single nearest neighbor. The advantages of KNN are: Easy to implement, adapts easily ad few hyper-parameters.

KNN works as follows:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated **Euclidean distance**.

- The training tuples are described by n attributes.
- Each tuple represents a point in an n-dimensional space.
- When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple.
- "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say,
- $X_1 = (x_{11}, x_{12},..., x_{1n})$ and $X_2 = (x_{21}, x_{22},..., x_{2n})$, is

$$dist(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}. \qquad \textbf{4.1}$$

Where, x1, x2 = two points in Euclidean n-space

X_{1i}- X_{2i} = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Case Study

Data Source:

The NASA-USDA Enhanced SMAP Global soil moisture data provides soil moisture information across the globe at 10-km spatial resolution. This dataset includes: surface and subsurface soil moisture, soil moisture profile (%), surface and subsurface soil moisture anomalies (-).

The dataset is generated by integrating satellite-derived Soil Moisture Active Passive (SMAP) Level 3 soil moisture observations into the modified two-layer Palmer model using a 1-D Ensemble Kalman Filter (EnKF) data assimilation approach. Soil moisture anomalies were computed from the climatology of the day of interest. The climatology was estimated based on the full data record of the SMAP satellite observation and the 31-day-centered moving-window approach. The assimilation of the SMAP soil moisture observations help improve the model-based soil moisture predictions particularly over poorly instrumented areas of the world that lack good quality precipitation data.

This dataset was developed by the Hydrological Science Laboratory at NASA's Goddard Space Flight Center in cooperation with USDA Foreign Agricultural Services and USDA Hydrology and Remote Sensing Lab. The data for weather of Yangon region is collected from Myanmar Weather Forecasting, Yangon. Related weather conditions need for crops are referenced from Food and Agriculture Organization of the United Nations.

Table 4.3 Water Demand Prediction

Temp	Humidity	Soil Moisture	Euclidean Distance	Rank minimum distance
21	88	37	(21-19) ^2 +(88-74) ^2 +(37-27.6) ^2 = 4+196+88.36 = Sqrt (288.36) =16.98117	5
30	46	29	(30-19) ^2+(46-74) ^2+(29-27.6) ^2 =11+784+1.96 = Sqrt (736.96) =27.14701	8
20	88	25.6	(20-19) ^2+(88-74) ^2+(25.6-27.6) ^2 = 1+196+4.00 = Sqrt (201) =14.17745	4
31	43	32.5	(31-19) ^2+(43-74) ^2+(32.5-27.6) ^2 =144+961+24.01 = Sqrt (1129.01) =33.60074	9
19	94	35.6	(19-19) ^2+(94-74) ^2+(35.6-27.6) ^2 =0+400+64 = Sqrt (464) =21.54066	6

26	70	27.5	(26-19) ^2+(70-74) ^2+(27.5-27.6) ^2 =49+16+0.01 = Sqrt (65.01) =8.06288	2
25	69	28	(25-19) ^2+(69-74) ^2+(28-27.6) ^2 =36+25+0.16 = Sqrt (61.16) =7.82049	1
22	73	35.7	(22-19) ^2+(73-74) ^2+(35.7-27.6) ^2 =9+1+65.61 = Sqrt (75.61) =8.69540	3
33	41	26	(33-19) ^2+(41-74) ^2+(26-27.6) ^2 =196+1089+1.6 = Sqrt (1286.6) =35.86921	10
20	94	37.5	(20-19) ^2+(94-74) ^2+(37.5-27.6) ^2 =1+400+98.01 = Sqrt (499.01) =22.33853	7

4.4 Implementation of System

System Designs (Login Page)

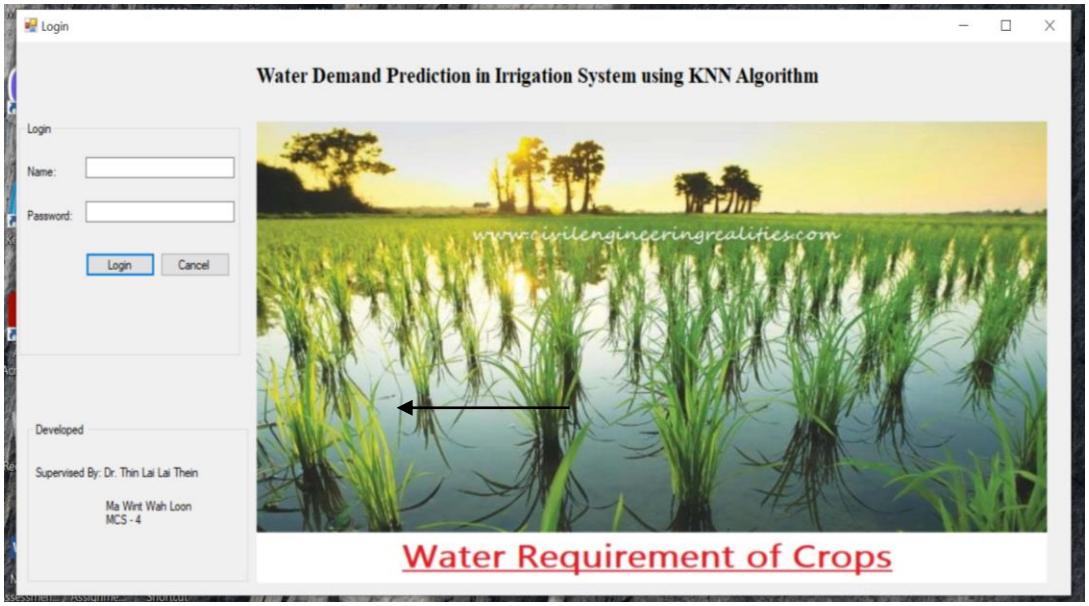


Figure 4.2 System Login Page

The proposed is implemented to predict the water demand in irrigation system using k nearest neighbor algorithm. The user must be login to use or to enter the system. Only the admin is allowed to add new user to the system. The system login page is shown in above figure 4.2. After the authentication process is success, the system main page will appear as shown in figure 4.3

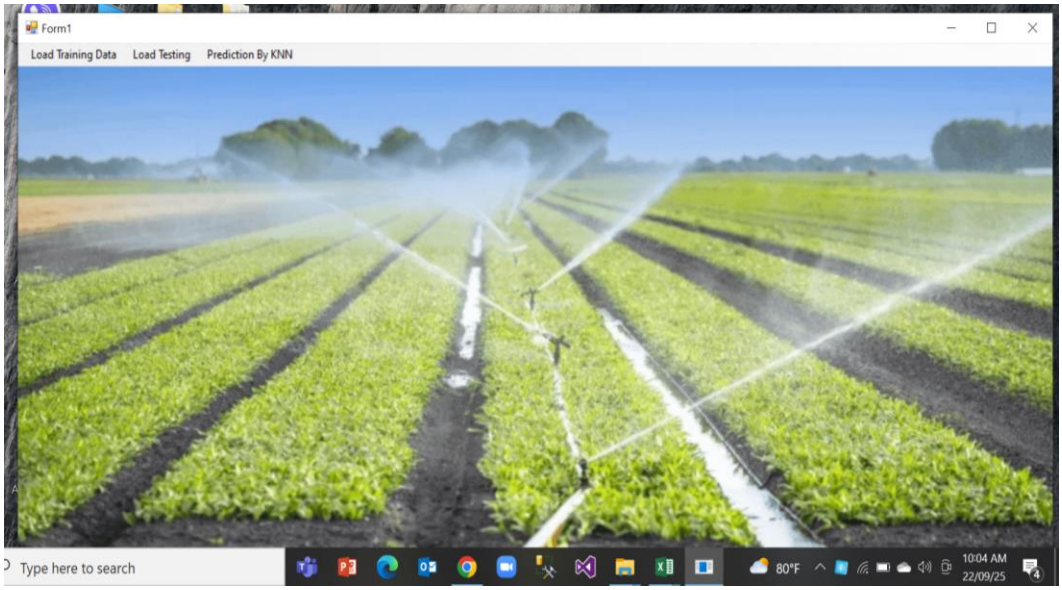
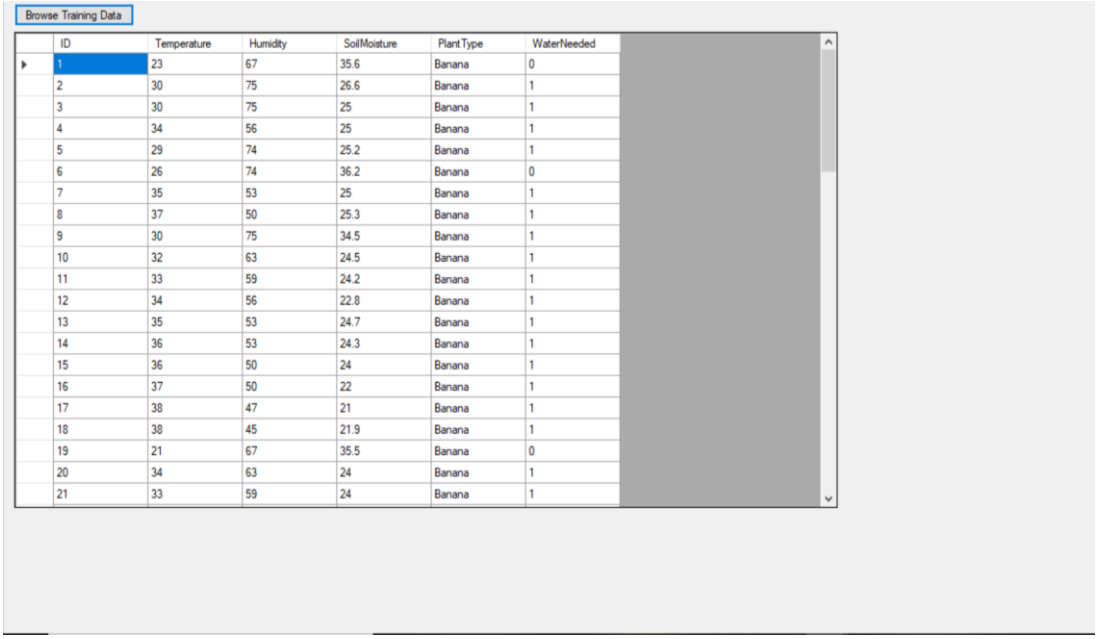


Figure 4.3 Main Page of System

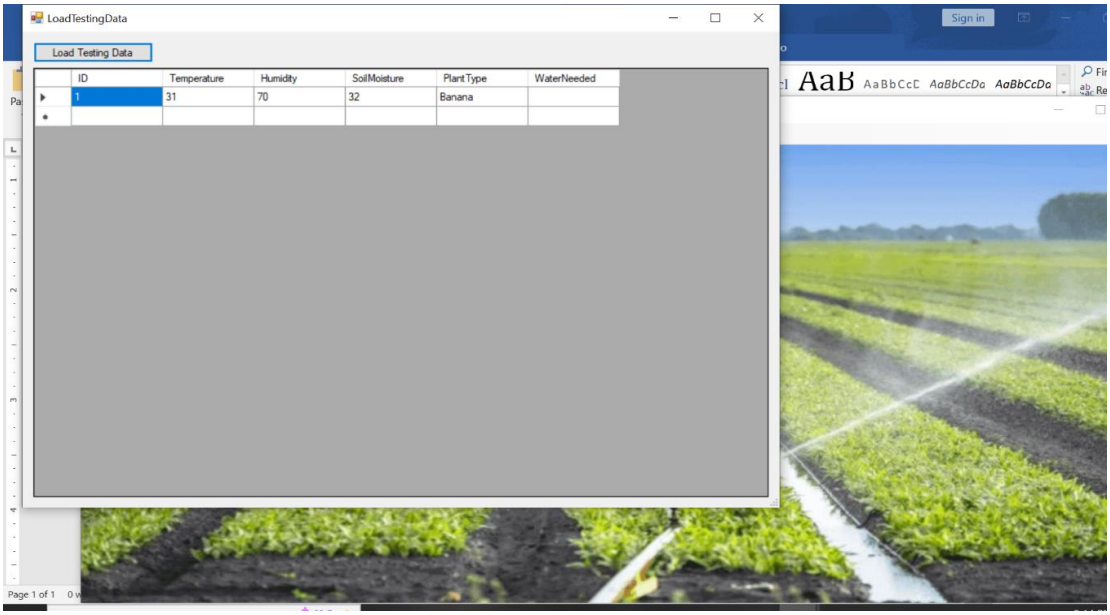
Training Data Loading Page



ID	Temperature	Humidity	SoilMoisture	PlantType	WaterNeeded
1	23	67	35.6	Banana	0
2	30	75	26.6	Banana	1
3	30	75	25	Banana	1
4	34	56	25	Banana	1
5	29	74	25.2	Banana	1
6	26	74	36.2	Banana	0
7	35	53	25	Banana	1
8	37	50	25.3	Banana	1
9	30	75	34.5	Banana	1
10	32	63	24.5	Banana	1
11	33	59	24.2	Banana	1
12	34	56	22.8	Banana	1
13	35	53	24.7	Banana	1
14	36	53	24.3	Banana	1
15	36	50	24	Banana	1
16	37	50	22	Banana	1
17	38	47	21	Banana	1
18	38	45	21.9	Banana	1
19	21	67	35.5	Banana	0
20	34	63	24	Banana	1
21	33	59	24	Banana	1

Figure 4.4 Load Training Data

In the main page, there three main menu: “Load Training Data” menu, “Load Testing Data” menu, and “Prediction By KNN” menu. Figure 4.4 shown the user interface design of the “Load Training Data” menu page and in this page contains a button “Browse Training Data” to load the user desire dataset for training phase. After the training data loading, “Load Testing Data” Page is used to load testing dataset. The testing dataset loading page is shown in figure 4.5.



ID	Temperature	Humidity	SoilMoisture	PlantType	WaterNeeded
1	31	70	32	Banana	

Figure 4.5 Load Testing Data

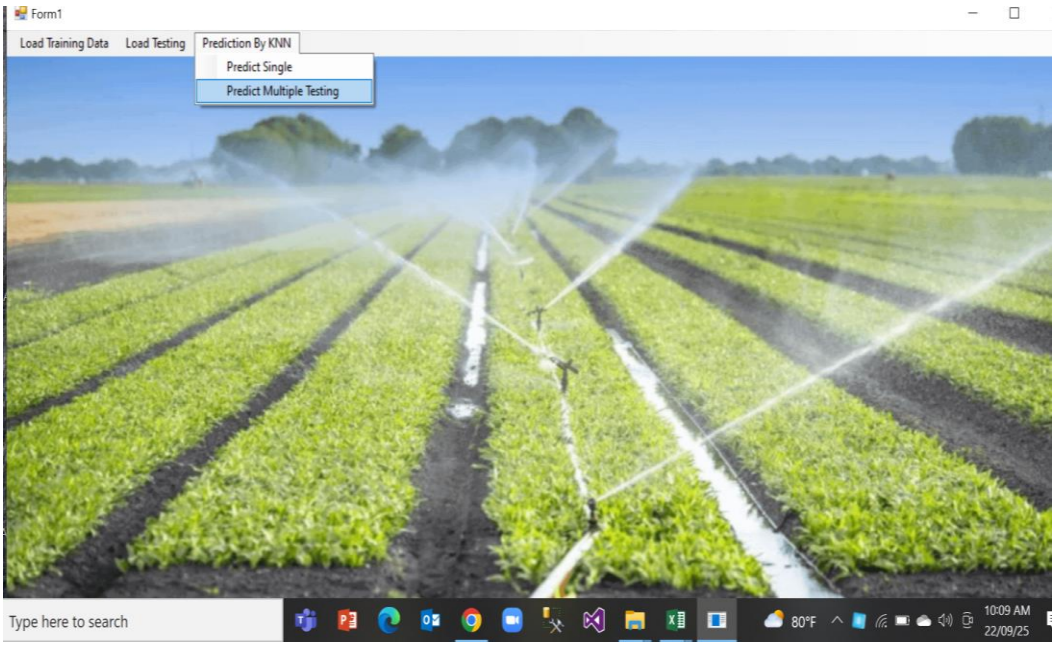


Figure 4.6 Prediction Page (single record testing)

In Fig 4.6 “Prediction By KNN” menu, there are two submenu: “Predict Single” menu which is used to predict the single record of data testing for prediction and “Predict Multiple Testing” submenu is to test batch data testing.

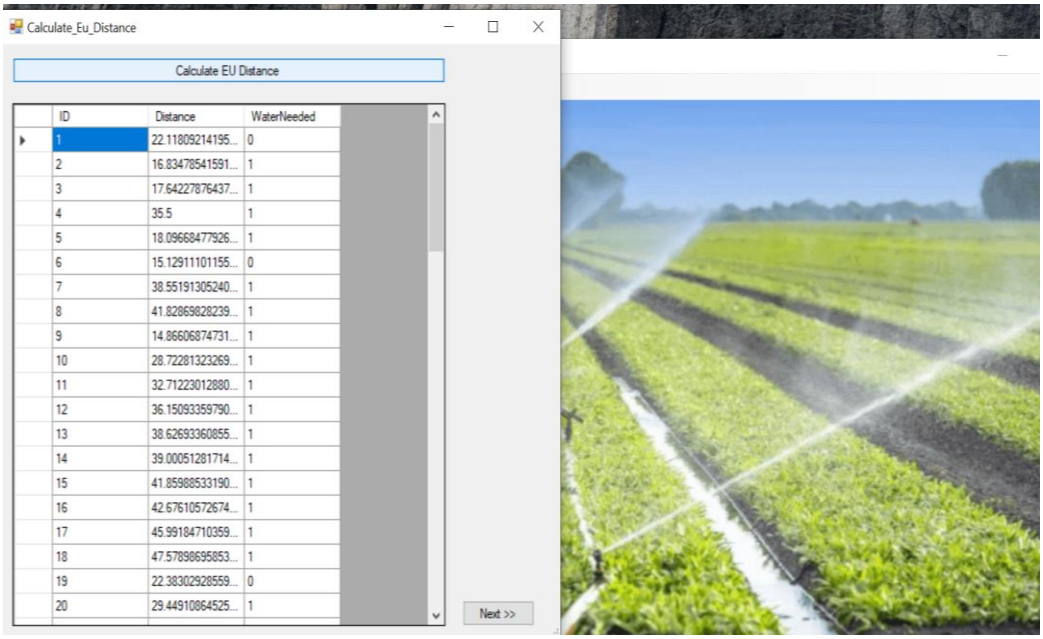


Figure 4.7 Calculation of Eu_Distance

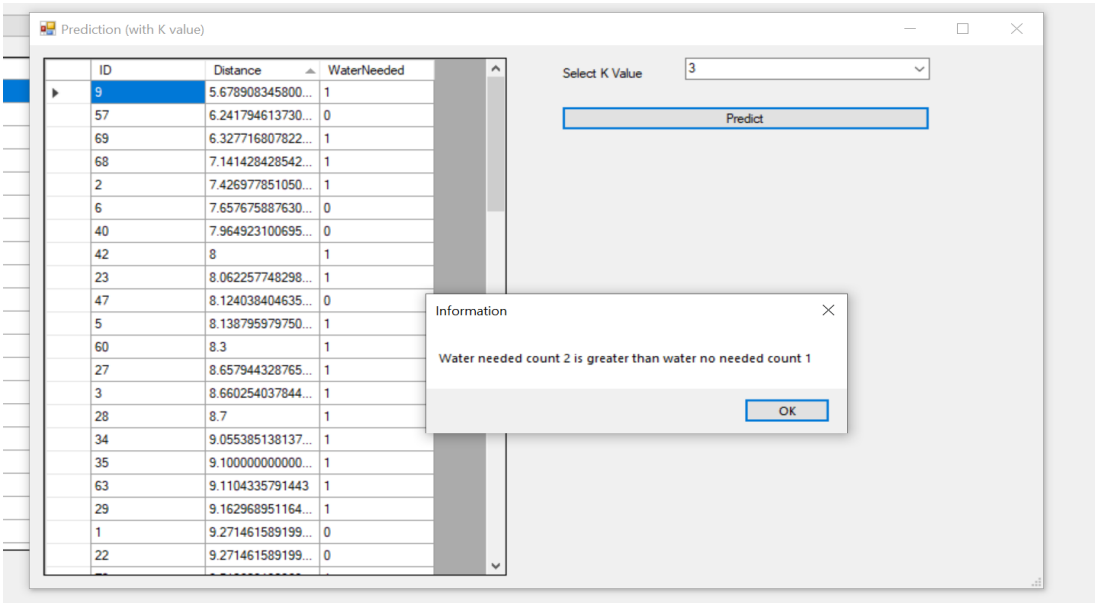


Figure 4.8 Prediction With selected K value for single record

In the prediction phase, the Euclidean Distance is calculated on each testing with respect to existing training dataset. The Euclidean Distance calculation is first phase of the prediction Figure 4.7 and then the system user must set the K value to predict the optimal result as shown in Figure 4.8.

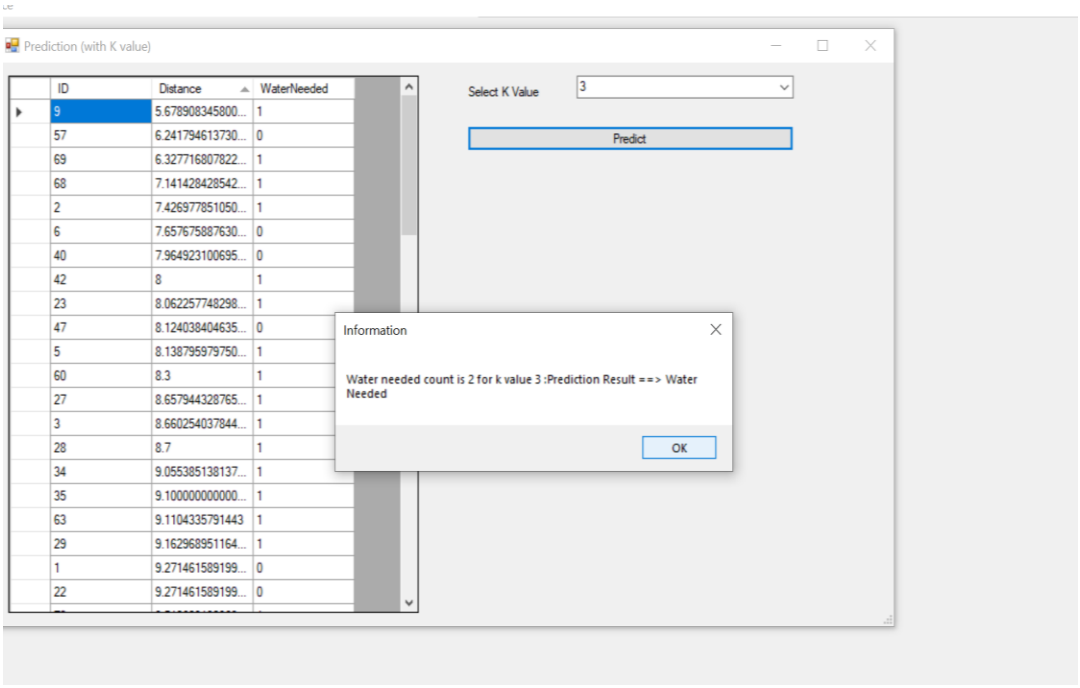


Figure 4.9 Prediction Result

After the single record testing is finished, the prediction result will appear as shown in Figure 4.9. The batch testing dataset prediction result will be shown as following Figure 4.10.

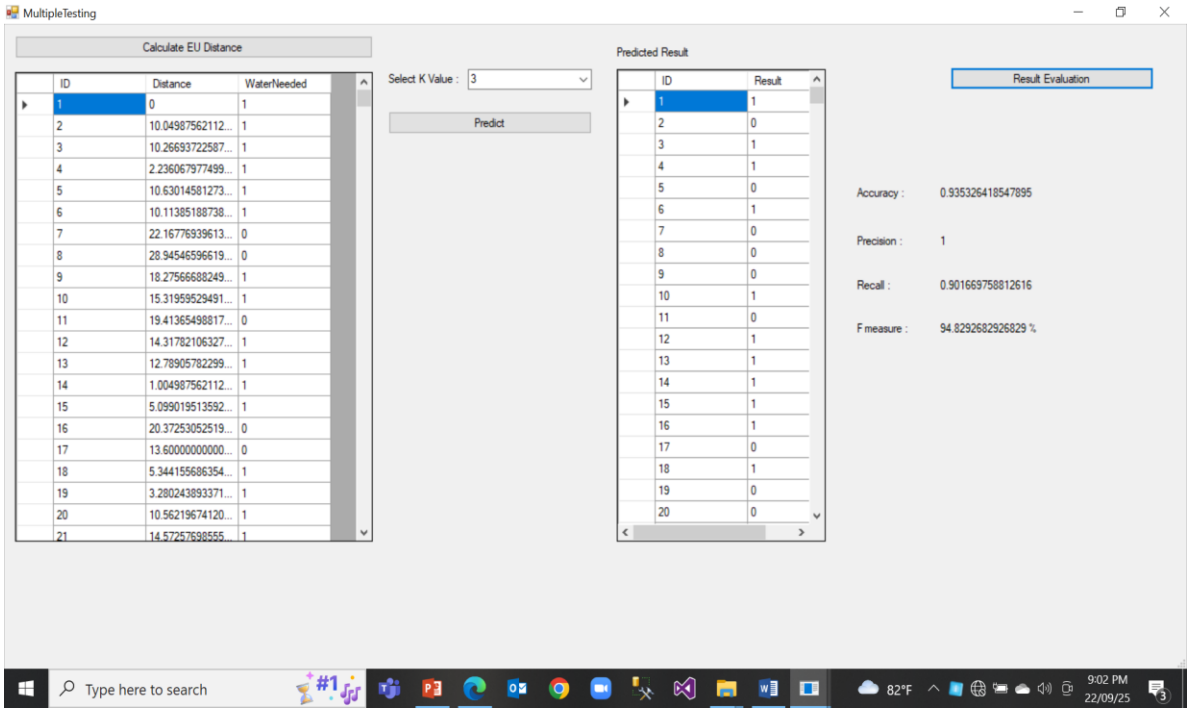


Figure 4.10 Accuracy on Multiple Testing Data Sample

4.5 Evaluation Metrics

Five evaluation metrics, which are precision, recall, F-measure, accuracy and failure-ratio, are used to evaluate the effectiveness of the system. These are calculated by using Eq. (5.1) - (5.4) respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad 5.1$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad 5.2$$

$$\text{F-measure} = \frac{\text{Precision*Recall}}{\text{Precision+ Recall}} \quad 5.3$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad 5.4$$

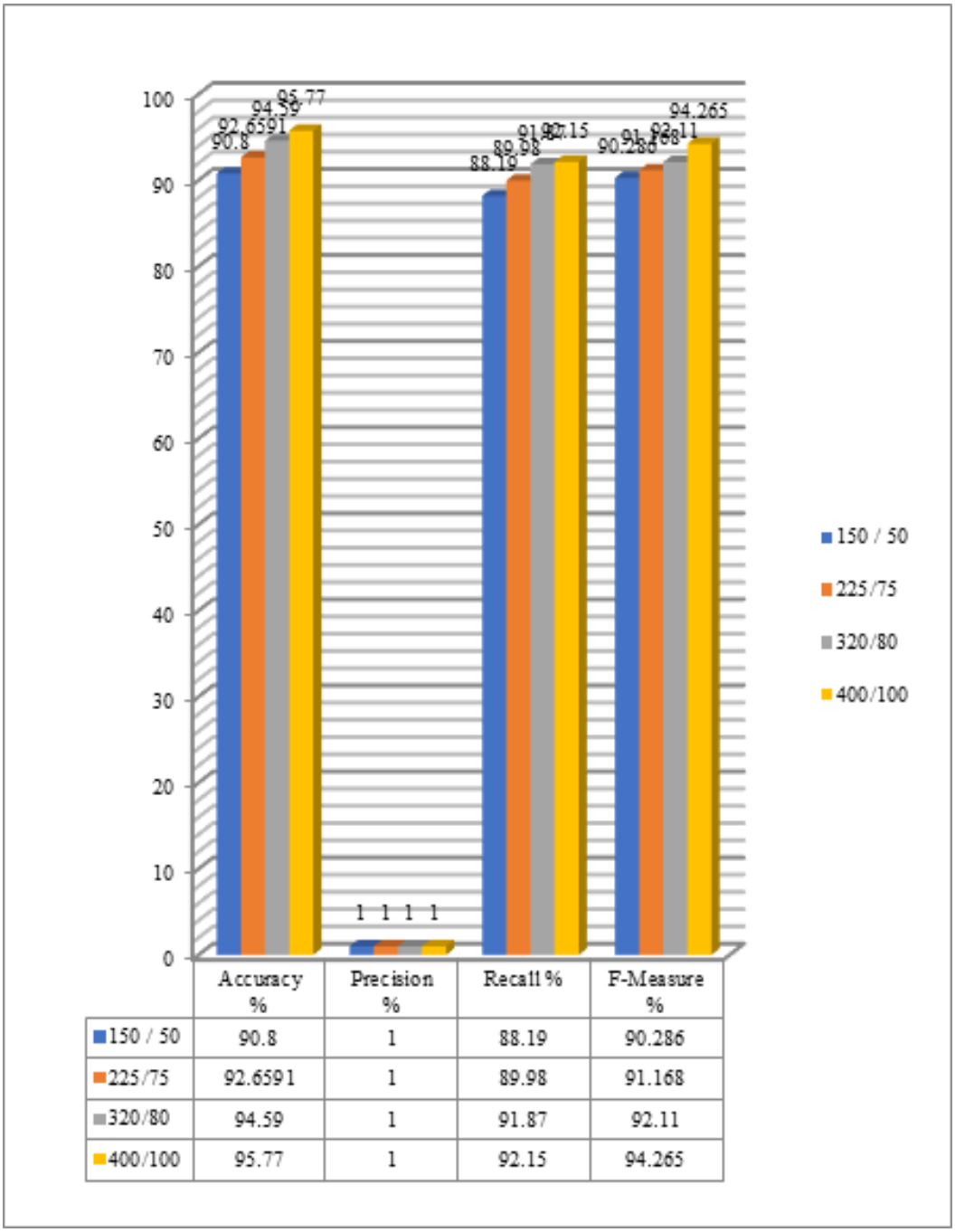
Where:

- TP refers to the number of true positive reviews.
- TN refers to the number of true negative reviews.
- FP refers to the number of false positive reviews.
- FN refers to the number of false negative reviews.
- Number of Misclassified Reviews refers to the reviews labelled to the class label which was not included in the actual class labels.
- Total Number of Reviews refers to the number of all reviews.

4.6 Experimental Result

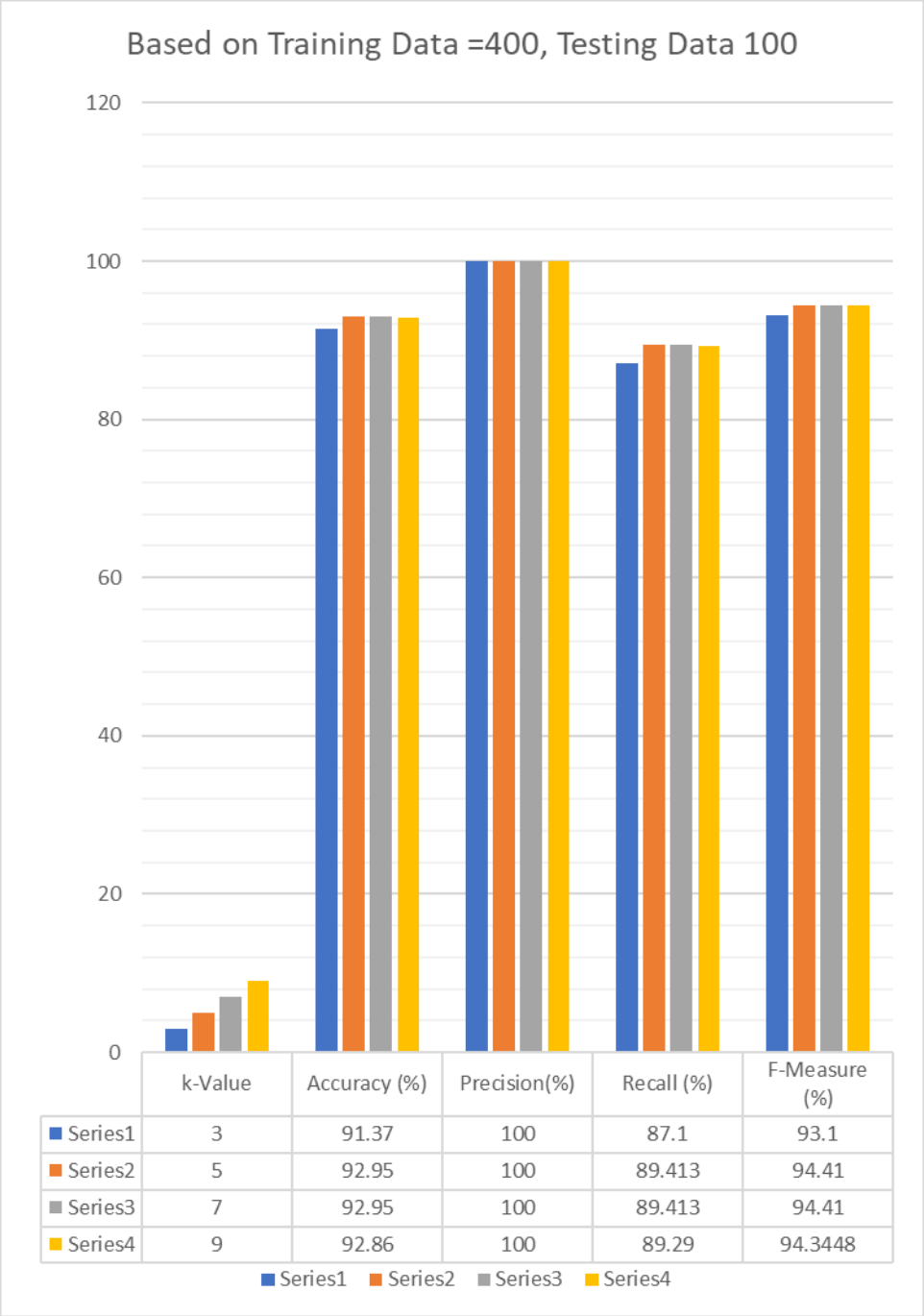
To make an effective analysis, 4 different training dataset and testing dataset pairs are used (Test1: Training Data 150 records and Testing Data 50 records; Test 2: Training Data 225 records and Testing Data 75 records; Test 3: Training Data 300 records and Testing Data 100 records; Test 4: Training Data 400 records and Testing Data 100 records. This system made the experiment results and performance evaluation based on Accuracy, Precision, Recall and F-measure of each analysis. The analysis results of 4 different dataset are shown in figure.

Based on the analysis, the system can give better detection result if the more trained data can feed to this system.



	Accuracy %	Precision %	Recall %	F-Measure %
150 / 50	90.8	1	88.19	90.286
225/75	92.6591	1	89.98	91.168
320/80	94.59	1	91.87	92.11
400/100	95.77	1	92.15	94.265

Figure 4.11 Experimental Results



	k-Value	Accuracy (%)	Precision(%)	Recall (%)	F-Measure (%)
Series1	3	91.37	100	87.1	93.1
Series2	5	92.95	100	89.413	94.41
Series3	7	92.95	100	89.413	94.41
Series4	9	92.86	100	89.29	94.3448

Figure 4.12 Test in Same Data Sample (400:100) with Different K Value

Figure 4.14 shows the analysis result of sample data size but different k values (k=3, 5, 7, 9) are used to predict water needed/not conditions of weather point of views and soil moisture. Based on above testing result, if the different k values are changed in testing, it will affect on four point of views: Accuracy, Precision, Recall and f-Measure.

CHAPTER 5
CONCLUSION AND FURTHER EXTENSION

The need for irrigation water is mostly influenced by human factor and climate change. The primary climatic factor is rainfall, and human factors such as irrigation area and technology level are also important. The impact of human factors is becoming increasingly noticeable with the advancement of water-saving irrigation equipment. The water demand forecasting model incorporates the idea of the water-saving improve coefficient based on the dual feature of “artificial-natural”. The suggested system has a better simulation effect than time series analysis and conventional regression, and it can more accurately depict the impact of water-saving technologies and adjusting planting structures on irrigation water.

By the use of this system, the result of sample data size is analyzed and different k values (k= 3, 5, 7, 9) are set in the equation to be able to predict the amount of water needed based on the conditions of weather point of views and soil moisture. According to the result of testing, if the different k values are changed in testing, it will have an effect on four points of views: Accuracy, Precision, Recall and f-Measure. After making an analysis on four different pairs of training and testing data, more precise result will be issued if the training data is well managed on the system.

5.1 Benefit of the System

A predictive model combining information about plant physiology, real-time soil conditions and weather forecasts can help make more informed decisions about when and how much to irrigate. This could save percentages of the water consumed by more traditional methods. The System supports the cultivators by increasing their awareness of the water demand and help to increase the crop yield.

5.2 Further Extension

This system is only emphasized on two groups of datasets: group 1 dataset contains three attributes and one class label such as temperature, humidity, soil moisture and class label. Group 2 dataset contains four attributes and one class label such as temperature, humidity, wind speed, soil moisture and class label. This system is only tested and analyzed for water melon and banana. This system can be extended the prediction based on various climates attributes with the respective crops and vegetables for each seasonal period of Myanmar.

AUTHOR'S PUBLUCATION

[1] Wint Wah Loon, Thin Lai Lai Thein, "Water Demand Prediction In Irrigation System Using KNN Algorithm", The National Journal of Parallel & Soft Computing (NJPSC 2022), 2022.

REFERENCES

[1] A.M. ASCE, “Water Demand Prediction Using Machine Learning”, https:// www.mifratech.com, 2022.

[2] A. pani and P. Mishra, “Hapa Irrigation for promoting sustainable agricultural intensification: experience from Bankura district of India, “geocoronal vol. 86, no 1, pp, 109-132, 2021.

[3] Bougadis, J.; Adamowski, K.; Diduch, R. Short-term municipal water demand forecasting. Hydrol. Process. 2005, 19, 137–148

[4] Bhuvana, Shashikala, “Water Demand Prediction using KNN Algorthim”, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology, 2022.

[5] C. Sutcliffe, J. Knox, and T. Hess, “Managing irrigation under pressure: how supply chain demands and environmental objectives drive imbalance in agricultural resilience to water shortages,” Agricultural Water Management, vol. 243, Article ID 106484, 2021

[6] Evaluation of Crop to Crop Water Demand Forecasting: Tomatoes and Bell Peppers Grown in a Commercial GreenhousDean C. J. Rice, Rupp Carriveau *, David S. -K. Ting and Mo’tamad H. BataTurbulence and Energy Laboratory, Ed Lumley Centre for Engineering Innovation, University of Windsor, Windsor

[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, “Machine learning for Prediction Outcomes: review, approaches and open research problems, ‖ Heliyon, vol. 5, no. 6, 2019, doi: 10.1016/ j.heliyon. 2019.e01802.

[8] Forecasting of irrigation water demand considering multiple factors Xuemei wang, Xiaohui Lei, Xuning guo, Jinjun you & Hao Wang

[9] Intelligent Control Of Agricultural Irrigation Through Water Demand Prediction Artificial Neural Network. Ouiyu Bo and Wuqun Cheng Institute of urbanrural construction, Agricultural University of Hebei. Baoding 071001, China.

[10] Orgaz, F.; Fernández, M.D.; Bonachela, S.; Gallardo, M.; Fereres, E. Evapotranspiration of horticultural crops in an unheated plastic greenhouse. Agric. Water Manag. 2005, 72, 81–96. [CrossRef]

[11] P. Cunningham and S. J. Delany, ―K - Nearest Neighbour Classifiers, ‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6...

[12] Paulo José A. Oliveira and Dominic L. Boccelli, “k-Nearest Neighbor for Short Term Water Demand Forecasting”, World Environmental and Water Resources Congress2017.