**Implementation of Advance Classifier With Machine Learning For Fruits Classification**

*Submitted in partial fulfillment of  
The requirements for the award of the degree of*

**Bachelor of Technology in Information Technology**

Submitted by

**Anuj Kumar (1873713013)**

**Dileep Kumar (1873713018)**

Under the Supervision of

**Dr. Ramesh Chand Pandey**

ASSISTANT PROFESSOR

Department of Information Technology

**RAJKIYA ENGINEERING COLLEGE, AMBEDKAR NAGAR**

Session 2021-2022

**Undertaking**

We declare that the work presented in this report titled **“Implementation of Advance Classifier with Machine Learning for Fruits Classification”**, submitted to the Information Technology Department, Rajkiya Engineering College, Ambedkar Nagar, for the award of the Bachelor of Technology degree in Information Technology, is our original work. We have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, we accept that our degree may be unconditionally withdrawn.

(Anuj Kumar, 1873713013)

(Dileep Kumar, 1873713018)

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**RAJKIYA ENGINEERING COLLEGE,**

**AMBEDKAR NAGAR-224122, UP, INDIA**

*AFFILIATED TO*

**DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY**

**LUCKNOW, UTTAR PRADESH, INDIA**

***CERTIFICATE***

This is to certify that this project report entitled **“Implementation of Advance Classifier with Machine Learning for Fruits Classification”,** submitted by Anuj Kumar(1873713013)and Dileep Kumar(1873713018), to the Rajkiya Engineering College, Ambedkar Nagar, towards the fulfillment of requirements for the award of the degree of Bachelor of Technology in Information Technology is a record of bonafide work carried out by them under my supervision and guidance during 2021-2022. The project, in our opinion, is worthy of consideration for the award of the degree of Bachelor of Technology in Information Technology with the rules and regulations of the Institute.

**Dr. Ramesh Chand Pandey Dr. Ashish Kumar Mishra**

***Assistant Professor Project Coordinator***

***Department of Information Technology Department of Information Technology (Supervisor)***

**External Examiner Dr. Sudhakar Tripathi**

***Associate Professor***

***HOD (IT)***

**Acknowledgment**

We wish to express our sincere gratitude to **Dr. Sudhakar Tripathi, HOD, IT Department of REC, Ambedkar Nagar**, for providing us all the facilities required for the completion of this research work project.

First, we would like to thank our guide, **Dr. Ramesh Chand Pandey**, for being a great mentor we could ever have. His advice, encouragement, and critics are a source of innovative ideas, inspiration, and causes behind the successful completion of this dissertation. The confidence shown on us by him was the biggest source of inspiration for us. His perpetual energy and enthusiasm in research had motivated us. Also, he was always accessible and willing to help his students with their work.

Finally, we are grateful to our parents, the rest of the family members, and all friends for their faith, help, and support. We couldn't complete this project work without their love, blessings, and encouragement.

(Anuj Kumar, 1873713013)

(Dileep Kumar, 1873713018)

**TABLE OF CONTENT**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Abstract** | | | | | **v** |
| **List of Figures** | | | | | ix-x |
| **List of Tables** | | | | | x |
| **1.** |  | **INTRODUCTION** | | | **11-16** |
|  |  | 1.1 |  | What is Classifier | 12 |
|  |  | 1.2 |  | How it works | 13 |
|  |  | 1.3 |  | Importance of Classifiers | 13 |
|  |  | 1.4 |  | How classification algorithms work | 13 |
|  |  | 1.5 |  | Type of Classification | 13-14 |
|  |  |  |  | 1.5.1 Binary Classifier | 13 |
|  |  |  |  | 1.5.2 Multi-Class Classifier | 13-14 |
|  |  | 1.6 |  | Name of Classification Algorithms | 14 |
|  |  | 1.7 |  | Related Research | 14-16 |
| **2.** |  | **METHODOLOGY** | | | **17-28** |
|  |  | 2.1 |  | Introduction | 17 |
|  |  | 2.2 |  | Machine Learning Application Models | 17-28 |
|  |  |  | 2.2.1 | Random Forest Classifier | 17-19 |
|  |  |  |  | 2.2.1.1 What is a random forest | 17-18 |
|  |  |  |  | 2.2.1.2 How it works | 18 |
|  |  |  |  | 2.2.1.3 Random Forest (Merits& Demerits) | 18-19 |
|  |  |  |  | 2.2.1.4 Application of Random Forest | 19 |
|  |  |  | 2.2.2 | SVM Classifier | 19-22 |
|  |  |  |  | 2.2.2.1 What is SVM | 19 |
|  |  |  |  | 2.2.2.2 Types of SVM | 20 |
|  |  |  |  | 2.2.2.3 How SVM Works | 20-21 |
|  |  |  |  | 2.2.2.4 SVM Classifier (Merits &Demerits) | 21-22 |
|  |  |  |  | 2.2.2.5 Application of SVM Classifier | 22 |
|  |  |  | 2.2.3 | K- Nearest Neighbour (K-NN) | 22-25 |
|  |  |  |  | 2.2.3.1 What is K-NN | 22-23 |
|  |  |  |  | 2.2.3.2 How K-NN works | 23-24 |
|  |  |  |  | 2.2.3.3 K-Nearest Neighbour (Merits& Demerits) | 24 |
|  |  |  |  | 2.2.3.4 Application of K-NN | 25 |
|  |  |  | 2.2.4 | Decision Tree (DT) | 25-28 |
|  |  |  |  | 2.2.4.1 What is Decision Tree | 25 |
|  |  |  |  | 2.2.4.2 Decision Tree Terminologies | 25-26 |
|  |  |  |  | 2.2.4.3 How Decision Tree Works | 26-27 |
|  |  |  |  | 2.2.4.4 Decision Tree (Merits & Demerits) | 27-28 |
|  |  |  |  | 2.2.4.5 Application of Decision Tree | 28 |
|  |  | 2.3 |  | Summary | 28 |
|  |  | **3.** |  | **IMPLEMENTATION** | **29-35** |
|  |  |  | 3.1 | Introduction | 29 |
|  |  |  | 3.2 | Dataset | 29-30 |
|  |  |  | 3.3 | Jupyter Notebook and Python | 31 |
|  |  |  | 3.4 | Principal Component Analysis (PCA) | 31 |
|  |  |  | 3.5 | Variance Explained Using PC | 32 |
|  |  |  | 3.6 | Proposed Methods | 33-35 |
|  |  |  |  | 3.6.1 Flow Chart of Implementation | 33 |
|  |  |  |  | 3.6.2 Flatten of Data | 34 |
|  |  |  |  | 3.6.3 Fruits Sample | 35 |
|  |  |  |  | 3.6.4 Visualization of Data | 35 |
|  |  |  |  |  |  |
| **4.** |  | **EXPERIMENTAL RESULTS** | | | **36-41** |
|  |  | 4.1 Normalized Confusion Matrix | | | 36-37 |
|  |  | 4.2 Performance Measure  4.2.1 Accuracy  4.2.2 Precision  4.2.3 Recall  4.2.4 F1 Score | | | 37-38  37  37  37  38 |
|  |  | 4.3 ROC Curve | | | 39-40 |
|  |  | 4.4 Performance with Respect to Runtime | | | 40 |
|  |  | 4.5 Result Summary | | | 41 |
| **5.** |  | **CONCLUSION AND FUTURE WORK** | | | **42** |
|  |  | 5.1 Conclusion  5.2 Future Work | | | 42  42 |
|  |  | **REFERENCES** | | | **43-45** |

**Abstract**

Machine learning algorithms can generally model complex object classes and accept various input predicting data class objects. In this project, Machine learning is a collection of various algorithms like supervised algorithms, unsupervised algorithms, and reinforcement learning. We used the supervised algorithms in this project. So that we can perform the certain operations on my dataset. This project totally based upon the implementation of advance classifier using machine learning. We are using my dataset contains 8 classes of different fruits and 934 total images. My idea is to perform different classification supervised machine learning algorithms, Such as Support Vector Machine (SVM), K-Nearest Neighbour(K-NN), Decision Tree, firstly for a binary classification task, then for a multi-class one. I will also apply Principal Component Analysis to reduce the dimensionality of the dataset, see the variance of each class and then I will try to apply classification algorithm having only two dimensions. At the end I will make a comparison between all methods to find which of them perform better on this dataset.

We have implemented this project using python programming language and their libraries such that Pandas, Numpy, OpenCV and Matplotliv. The maximum accuracy achieved for fruit detection is 98.49.00% (k-NN), 100% (SVM) and 91.57% (Decision Tree). SVM achieved 100% accuracy because of small dataset.

Keywords—dataset, classifier, algorithm, machine learning, analysis, KNN, SVM, python.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure**c**No.** | **Name** **of Figure** | **Page**c**No.** |
| Figure 2.1 | Block Diagram of Random Forest | 18 |
| Figure 2.2 | General Solution | 21 |
| Figure 2.3 | Optimal Solution | 21 |
| Figure 2.4 | Class Distribution of K-NN | 24 |
| Figure 2.5 | Block Diagram of Decision Tree | 27 |
| Figure 3.1 | Fruits Sample | 29-30 |
| Figure 3.2 | Diagram of Principal Component Analysis | 31 |
| Figure 3.3 | Representation of Variance | 32 |
| Figure 3.4 | Flow Char of Implementation | 33 |
| Figure 3.5 | Image unfolding and Preparation of Image Matrix | 34 |
| Figure 3.6 | Fruits Sample | 35 |
| Figure 3.7 | Data Representation in 2D | 35 |
| Figure 3.8 | Data Representation in 3D | 35 |
| Figure 4.1 | Normalized Confusion Matrix of Different Classifiers | 36-37 |
| Figure 4.2 | Plot Diagram of Precision, Recall, F-1 Factor and Accuracy | 38 |
| Figure 4.3 | ROC Curve of different Algorithms | 39-40 |
| Figure 4.4 | Runtime Vs Classifiers | 40 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Name** **of Table** | **Page**c**No.** |
| Table 3.1 | Number of Sample Data | 30 |

**CHAPTER 1**

**INTRODUCTION**

The method of data science projects, as well as several statistical and machine learning technique applications, were discussed. The paper discussed the use of numerous parameters in machine learning as well as the study's appropriate cases. The study mainly focuses on various algorithms like K nearest neighbors (KNN) [17], Support Vector Machine (SVM) [9][23], Decision Tree [20][27] and Random Forest [22][28]. These all algorithms used for the classification of the data. Which is helps to train a machine therefore machine can easily detect or classify the objects. The study primarily focused on the statistical and mathematical features of each method, as well as the algorithms' applicability for specific use cases and the primary shortcomings of the respective algorithms. We believe that our work provides a straightforward and effective narration of the described techniques and will assist prospective researchers as a starting point for machine learning and data science research in science areas. We are making this attempt for the sake of summarizing the various algorithms mentioned above, and many articles and research papers were there, but we believe that our work will provide a straightforward and effective narration of the described techniques and will assist prospective researchers as a starting point for machine learning and data science research in science areas. The outcome of the work is explanation of the algorithms in the extensive by covering the pros and cons of each aspect along with the suitable use cases. The results we'll present will be a comparison of the same dataset with all the techniques, with the best performer in that use case projected.

Several statistical and machine learning methodology applications, as well as the process of data science projects, were covered. The study's appropriate scenarios were explored in the paper, as well as the usage of a variety of factors in machine learning. The research primarily focuses on K closest neighbors (KNN) [17], Support Vector Machine (SVM) [9][23], Decision Tree [20][27], and Random Forest [22][28]. These are the algorithms that were utilized to classify the data. Which aids in the training of a machine so that it can readily recognize and classify objects. The study primarily focused on the statistical and mathematical features of each method, as well as the algorithms' applicability for specific use cases and the primary shortcomings of the respective algorithms. We are making this attempt for the sake of summarizing the various algorithms mentioned above, and many articles and research papers were there, but we believe that our work will definitely provide a straightforward and effective narration of the described techniques and will assist prospective researchers as a starting point for machine learning and data science research in science areas. The work's output is a detailed explanation of the algorithms, including the benefits and drawbacks of each feature as well as appropriate use cases. We'll show the results of a comparison of the same dataset using all of the aforementioned methodologies, with the best performer in that use case projected. The machine learning approach is a method of teaching machines how to handle data more effectively and accurately. In this scenario, machine learning techniques are used to classify the data [10]. There is a demand for machine learning because a large number of datasets are available from various sources. Machine learning is being used in a variety of industries, from medical to the military, to extract meaningful data from large databases. The basic purpose of machine learning is to learn from existing data. Many algorithms have been developed to enable machines to learn on their own [3][26]. To solve this problem, several mathematicians and programmers use a variety of ways. A handful of these are seen in Figure 1. All of the education that is supervised Machine learning classification approaches are discussed in Section 2. The binary classification is used in this categorization.

So that we can determine whether the image is there in the file. If the image leaves, set the output to 1 with respect to the image; otherwise, set it to 0. On the same dataset, we apply the identical operations to all the algorithms stated. Following that, we'll compare the performance and accuracy of the method to that of other machine learning techniques. Following that, we shall compare each other. As a result, we can determine which algorithm is superior with respect of performance and runtime. And he questioned why other algorithms couldn't outperform them. All these points are discussed in detail below, along with explanations.

* 1. **What is Classifiers**

Any algorithm that organizes data into labelled groups, or categories of information, is referred to as a classifier. Spam filters, for example, analyze incoming "raw" emails and identify them as "spam" or "not-spam" based on their content. In many types of machine learning, classifiers are a tangible implementation of pattern recognition.

**1.2 How it works**

The classifier algorithm is trained using the labeled data; In the image recognition example, for example, the classifier receives training data that labels the images. After sufficient training, the classifier can take unlabeled images as input and output the classification label for each image.

**1.3 Importance of Classifiers**

High-end machine theory meets practical application in classifiers. These algorithms are more than just a way to classify unlabeled data into discrete classifications. Classifiers include a unique collection of dynamic rules, including an interpretation mechanism for dealing with ambiguous or unknown values, all of which are suited to the type of inputs being analyzed. The majority of classifiers also use probability estimations, which allow users to adjust data classification using utility functions.

Classifiers are the backbone of cluster analysis in unsupervised learning, and they are how the system characterize and evaluates unlabeled input in supervised or semi-supervised learning.

**1.4 How classification Algorithms works**

The Classification algorithm is a Supervised Learning technique that determines the category of new observations using training data. The process of a software learning from a dataset or observations and then classifying new observations into one of several groups or groupings is known as classification. Yes or No, 0 or 1, Spam or Not Spam, cat or dog, and so on are all options. Classes can be described using phrases such as targets, labels, and categories.

**1.5 Type of Classification**

**1.5.1 Binary Classifier:** This sort of classifier is employed when a classification task only has two possible outcomes.

Some examples include YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, and so on.

**1.5.2 Multi-Class Classifier**: When there are more than two possible outcomes in a classification problem, a Multi-class Classifier is utilized.

For example, classifications of various types of crops or classifications of various types of music.

**1.6 Name of Classification Algorithms**

There are many algorithms in machine learning which are used for many purposes in classification. Which are listed below.

Common Types of Classification Algorithms:

1. Naive Bayes Classifier
2. Vector Support Machines (SVMs)
3. Decision Tree
4. Random Forest
5. Artificial Neural Networks
6. K-Nearest Neighbors

**1.7 Related Research**

Machine learning, particularly its area of Deep Learning, has made numerous incredible developments in recent years, and significant research publications may lead to technological breakthroughs [15]. that are utilized by billions of people. The research in this topic is progressing at a rapid pace, thus we've put together a list of the most essential recent scientific papers released since 2014. Which is relevant to the fields of my project. This section summarises the literature on paper categorization methods relevant to the study's research topic. This model based (KNN) approach differs from the DR and other Cimmerian closest neighbour algorithms. It builds a robust model by identifying a group of representatives with additional data from the training data based on similarity principle. Deep Learning. The produced representatives can be viewed as regions in the data space and classified further. In 5-fold cross-validation, the average classification accuracy of our proposed KNN model technique on six datasets is better than C5.0 and comparable to KNN [17]. However, by preserving only a few representatives for categorization, the KNN Model considerably enhances the efficiency of KNN. According to the results of the experiments, the average decrease rate is 90%. The following are some examples of document classification application areas:

To classify documents, a variety of classification algorithms have been utilized [11]. There are two types of automatic document classification: supervised and unsupervised [19,20,21]. Documents are classified using the supervised classification method. Many of the foundations of supervised learning approaches were found in machine learning, particularly its area of Deep Learning. These approaches evaluate training data (i.e., preset input–output pair data) and provide an inferred function that may be used to map other examples. Unsupervised categorization, on the other hand, groups documents based on their similarity rather than any specified criterion. There have been several types of automatic document categorization algorithms established, such as the Nave Bayes classifier, Support Vector Machine (SVM) [9][23], K-Nearest Neighbors (KNN), Decision Tree, and so on [20, 27].

[2] shows a way for creating synthetic images that look very much like empirical photos. This paper provides a method for creating large-scale semantic segmentation datasets on a computer. Automated per-pixel class and depth categorization at the plant-part level of realistic agriculture scenes. One of the purposes of a synthetic dataset like this would be too bootstrap or pre-train computer vision models, which would then be finetuned on a smaller empirical image dataset. Similarly, a network trained on synthetic images in paper [25] can count the quantity of apples in an image without really determining where they are in the image. Another paper [4] predicts the yield for the future season using two back propagation neural networks trained on photos of apple "Gala" variety trees. The total cross-sectional area of fruits, fruit number, total cross-sectional area of small fruits, and cross-sectional area of foliage were all collected from photos for this work. The paper [18] examines fruit detectability in relation to the camera angle at the time the image was shot. The fruit detectability was shown to be highest on front views with a zenith angle of 60 degrees upwards, according to this study. We can observe a technique to recognizing fruits based on color, shape, and texture in studies [4, 18, 19]. They demonstrate how difficult it is to correctly distinguish similar fruits from several species. They propose merging existing approaches for detecting regions of interest in 5 photos utilizing the texture, shape, and color of fruits. Similarly, to enhance the number of fruits, uses a method that combines the form, size, color, and texture of the fruits with a k nearest neighbour algorithm is employed to improve recognition accuracy. One of the most recent papers [18] describes an algorithm based on the modified Chaves level-set model, as well as the level-set notion and M-S mode. The proposed goal was to do green grape detecting at night. The picking point of the fruit stem was computed using the notion of the smallest circumscribed rectangle of fruit and the Hough straight-line detection method. After your paper is styled, save the equation as a graphic and insert it inside the text.

**CHAPTER 2**

**METHODOLOGY**

**2.1 Introduction**

We will use supervised machine learning methods in this part. We have both the input and output data in this form of machine learning method. The model (data) is trained by the algorithm to map the input to the output. For both classification and regression, supervised learning techniques are used. However, classification techniques are used for categorical data. The practice of breaking data sets into multiple categories or groups by assigning labels to them is known as classification. Different classification methods exist, including Decision Trees, Random Forest, K-NN, and SVM.

**2.2 Machine Learning Application Models**

Machine learning algorithms learn from historical data and use it to predict future results. Deep learning can be thought of as a subset of machine learning. In deep learning, the word ’deep’ means there are many layers used in neural networks.  
Machine and Deep learning have yielded results equivalent to and, in some cases surpassing, human expert performance in computer vision, recognition of speech, natural language processing (NLP)[15][32], automatic translation of machines, healthcare image analysis, bioinformatics, material inspection, and board game programs [26-27]. We used the four machine learning algorithms Random Forest, K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and Decision Tree Classifier. The explanations for these algorithms are listed below.

**2.2.1 Random Forest**

**2.2.1.1 What is Random Forest**

Random Forest is a well-known machine learning method that falls within the category of supervised learning algorithms. And random forest can be used for classification and regression. It is built on the resemble learning concept. This is a method of combining numerous classifiers to solve a difficult problem and increase the model's performance. Random Forest is a classifier that uses the average of many decision trees on distinct subsets of a dataset to increase the dataset's predictive accuracy. It takes each tree's prediction and predicts the final output based on the majority.

**Higher Accuracy ∝ Number of Trees**

**2.2.1.2 How it works**

The working procedure is outlined in the stages and picture below:

**Step 1:** At random, choose k data points from the training set.

**Step 2:** Make decision trees for the data points you've chosen.

**Step 3:** Determine how many decision trees you'd like to create.

**Step 4:** Go oversteps 1 and 2 again.

**Step 5:** Locate each decision tree's forecasts for new data points and assign the new data points to the most popular category.

Diagram

Description automatically generated

Figure 2.1 Block Diagram of Random Forest

**2.2.1.3 Random Forest (Merits & Demerits)**

The key advantage of the random forest technique is that it can handle large datasets with high dimensionality. It also increases the accuracy of the model and avoids the issue of overfitting. The most notable disadvantage, however, is that it is ineffective for regression tasks.

* **Merits**

1. Both classification and regression problems can be handled by Random Forest.
2. It is capable of handling large datasets with a high degree of dimensionality.
3. It enhances the accuracy of the model and eliminates the issue of overfitting.

* **Demerits**

Even though random forest may be used for both classification and regression problems, it is not better suited to regression.

**2.2.1.4 Application of Random Forest**

1. In the following four industries, random forest is most commonly used:
2. Banking: In the banking industry, this method is mostly used to identify loan risk.
3. This strategy can be used to detect disease trends as well as disease risks in medicine.
4. Land Usage: Using this method, we may locate areas with similar land use.
5. This algorithm can be used to identify marketing trends in the future.

**2.2.2. Support Vector Machine Classifier (SVM)**

**2.2.2.1 What is SVM**

Support Vector Machine (SVM) algorithm is an easy and better supervised machine learning algorithm. SVM algorithm can act well with both linearly dividable and non-linearly separable datasets. And support vector machine is a biased classifier normally defined by separating the hyperplane. In two-dimensional space, this hyperplane is a line separating a plane into two portions where in individually class lay on either side. It is effective in the higher dimension. And can be used for non-linear data sets with the help of kernel function. It works for both large and small datasets. But larger dataset requires a large amount of time to process.

**2.2.2.2 Types of SVM**

SVM classifier are divided into two categories:

1. **Linear SVM**

It is a classifier for linearly separable data, which means a dataset can be classified into two classes using a single straight line, and the classifier is called Linear SVM.

1. **Non-linear SVM**

Non-linear SVM is a classifier that is used for non-linearly separated data. This means that if a dataset cannot be classified using a straight line, it is non-linear data, and the classifier used is termed Non-linear SVM.

**2.2.2.3 How SVM Works**

Consider the following example to see how the SVM algorithm works. Assume there are two different features in the dataset: x1 (Circle) represented by Class A and x2 (Square) represented by Class B. For the pair of coordinates, we're looking for a classifier that can tell the difference between orange and blue (x1, x2). The SVM approach aids in the discovery of the best line or decision boundary, sometimes referred to as a hyperplane. The SVM method identifies the point at where the lines from both classes cross. The margin is the distance between the vectors and the hyperplane. The optimal hyperplane is the one with the largest margin. A basic straight line will work because it's a two-dimensional space. And the classified the data x1 and x2 in two classes represented in below diagram which have separated from the support vector machine.

Chart, scatter chart

Description automatically generated

Figure 2.2 General Solution

Chart, scatter chart

Description automatically generated

Figure 2.3 Optimization Solution

**2.2.2.4 SVM Classifier (Merits & Demerits)**

* **Merits**

1. When there is a clear margin of distinction between classes, SVM performs well.
2. In high-dimensional spaces, SVM is more effective.
3. When the number of dimensions is more than the number of samples, SVM is effective.
4. SVM uses a small amount of memory.

* **Demerits**

1. The SVM algorithm is ineffective for large data sets.
2. SVM does not perform well when the data set contains more noise, such as overlapping target classes.
3. If the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
4. Because the support vector classifier works by placing data points above and below the classifying hyperplane, there is no probabilistic reason for the classification.

**2.2.2.5 Application of SVM Classifier**

1. **Face detection** — SVM classifies sections of the picture as face or non-face and draws a square around the face.
2. **Text and hypertext categorization** — Both inductive and transudative models can use SVMs for text and hypertext classification. They classify articles into multiple groups using training data. It categorizes based on the generated score and compares it to a threshold number.
3. **Image classification** - The use of SVMs improves image classification search accuracy. It outperforms typical query-based searching techniques in terms of accuracy.
4. **Protein classification and cancer classification**- SVM is used to determine gene categorization, patient classification based on genes, and other medical difficulties.

**2.2.3 K- Nearest Neighbors**

**2.2.3.1 What is K-NN**

The k-nearest neighbour algorithm is a well-known supervised machine learning approach for pattern detection or categorization. Classifying objects based on the issue space's closest training examples in this scenario. KNN is a type of instance-based learning in which the function is only estimated locally, and all computation is deferred until classification. It is one of the most fundamental machine learning algorithms. A majority vote of an object's neighbour determines its classification, with the object being assigned to the most common class among its k closest neighbour (k is a positive integer). If k = 1, the object is simply assigned to the class of its closest neighbour.

**2.2.3.2 How K-Nearest Neighbors works**

The k-NN approach can be used to estimate continuous variables as well. An inverse distance weighted average of the k-nearest multivariate neighbour is one such implementation. And assume there are two different features in the dataset: x1 (orange color) represented by Class A and x2 (blue color) represented by Class B. For the pair of coordinates, we're looking for a classifier that can tell the difference between green and blue (x1, x2). following is how this algorithm works:

* 1. Calculate the Euclidean or Mahala Nobis distance between the target plot and the sampling plots.
  2. Place samples in order based on computed distances.
  3. Using the cross-validation technique, choose the k nearest neighbour that is heuristically ideal.
  4. Using the k-nearest multivariate neighbour, compute an inverse distance weighted average.

Chart

Description automatically generated

Figure 2.4 Class distribution of K-NN

**2.2.3.3 K-NN (Merits & Demerits)**

The K-Nearest Neighbors classification algorithms merits and demerits given to below.

* **Merits**

1. It is simple to put into action.
2. It can withstand noisy training data.
3. If the training data is large, it can be more effective.

* **Demerits**

1. The value of K must be calculated on a regular basis, which might be difficult at times.
2. The calculation cost is high since the distance between the data points for all the training samples must be determined.

**2.2.3.4 Application of K-NN**

The K-Nearest Neighbors algorithms used in many fields in real life. Which is listed below.

1. Text mining
2. Agriculture
3. Finance
4. Medical
5. Facial recognition
6. Recommendation systems (Amazon, Hulu, Netflix etc.)

**2.2.4 Decision Tree**

**2.2.4.1 What is Decision Tree**

Decision Tree is a well-known supervised machine learning technique. It can be used to tackle both classification and regression problems, however classification problems are the most popular. In this tree-structured classifier, internal nodes represent dataset properties, branches represent decision rules, and each leaf node delivers the conclusion. A Decision tree's two nodes are the Decision Node and the Leaf Node. Decision nodes are used to make any decision and have numerous branches, whereas Leaf nodes are the output of such decisions and do not have any more branches.

It's a graphical representation of all possible solutions to a problem or decision based on a set of criteria. To make a tree, the CART algorithm is employed.

**2.2.4.2 Decision Tree Terminologies**

* **Root Node:** The decision tree's root node is the starting point. It depicts the entire dataset before being divided into two or more homogeneous groups.
* **Leaf Node:** Leaf nodes are the tree's final output nodes, and they cannot be further separated.
* **Splitting:** The process of splitting the decision node/root node into sub-nodes based on the conditions stated is known as splitting.
* **Branch:** A tree that has been divided into subtrees or branches.
* **Pruning:** Pruning is the process of removing unwanted branches from a tree.
* **Parent/Child Node**: The parent node is the tree's root node, while the child nodes are the tree's remaining nodes.

**2.2.4.3 How it works**

The working of decision tree algorithm starts predicting class of mention dataset from the root node of the tree. And comparison starts the root value of attributes with record of the real dataset based upon the comparison. And it follows the branch and jumps to further node. This process continues until it is reaching the leaf node of tree. For the better understanding steps are defined below one by one:

**Step-1:** Start the tree from root node which contains the complete dataset.

**Step-2:** Get the best attributes in the dataset by the Attribute Selection Measure.

**Step-3:** Divide the root node into the subsets that contains possible values for the best attributes.

**Step-4:** Generate the decision tree node, which have the best attribute.

**Step-5:** Recursively build the new decision tree using the subsets of the dataset created in step-3. The process keeps going on till that where until you cannot further classify the nodes and called the final node as a leaf node.

Diagram

Description automatically generated

Figure 2.5 Block Diagram of Decision Tree

**2.2.4.4 Decision Tree (Merits & Demerits)**

The K-Nearest Neighbors classification algorithms merits and demerits given to below.

* **Merits**

1. It is simple to put into action.
2. It can withstand noisy training data.
3. If the training data is large, it can be more effective.

* **Demerits**

1. The value of K must be calculated on a regular basis, which might be difficult at times.
2. The calculation cost is high since the distance between the data points for all the training samples must be determined.

**2.2.4.5 Application of Decision Tree**

The K-Nearest Neighbors algorithms used in many fields in real life. Which is listed below.

1. Text mining
2. Agriculture
3. Finance
4. Medical
5. Facial recognition
6. Recommendation systems (Amazon, Hulu, Netflix etc.)

**2.3 Summary**

All the above mentioned algorithms are related to supervised machine learning algorithms. Which labels use to classify data and machine training. We have explained all the algorithms that we will use in the implementation part. Also discussed the functions, merits and demerits and applications. Machine learning is a very vast field. Which is very useful today. Most company and industry today use it to classify unstructured data for insights and fact or pattern detection for big data. Because nowadays a huge amount of unstructured data is being generated every second. That's why machine leaning play the important role to training the data and growth in automation and artificial intelligence. In the next section we will explain the implementation part.

**CHAPTER 3**

**IMPLEMENTATION**

**3.1. Introduction**

In this section we will talk about the implementation part. And which kind of data we are using and other factors which we need during the process. And how the principal component works and how we to calculate the variance and their works. And we are using the jupyter notebook and python programming language. And explained about it. And how to propose model works. All these factors have discussed here.

**3.2. Dataset**

We created our own dataset. which have a different variety of fruits images. The name of the dataset is dataset fruits. Which contains the 8 different types of fruits. And the total number of images in the dataset is 934. And perform the supervised machine learning algorithms on this dataset. The sample of dataset images is given below.

A close up of an orange

Description automatically generated with medium confidence A red apple with a stem

Description automatically generated

(a). Orange (b). Apple

A pineapple with a white background

Description automatically generated with medium confidenceA yellow banana with a white background

Description automatically generated with medium confidence

(c). Pineapple (d). Banana

A picture containing litchi, fruit

Description automatically generated A picture containing indoor, fruit

Description automatically generated

(e). Lychee (f). Mango

A picture containing kiwi

Description automatically generated A green apple with a stem

Description automatically generated with medium confidence

(g) Kiwi (h) Guava

Figure 3.1. Fruits Sample

The total number of images for each fruit listed in the table number 3.1. There are multiple varieties of fruits each of them being considered as a separate object.

|  |  |  |
| --- | --- | --- |
| **Variety** | **Number of training images** | **Number of test images** |
| Apple | 103 | 53 |
| Banana | 121 | 45 |
| Guava | 96 | 48 |
| Kiwi | 117 | 43 |
| Lychee | 94 | 40 |
| Mango | 109 | 32 |
| Orange | 121 | 35 |
| Pineapple | 173 | 36 |

Table No 3.1 Number of Dataset

**3.3. Jupyter And Python**

Jupyter Notebook is an open-source web application. Which is used for the run the code. And jupyter run the code statement one by one. And we also can share the documents which contains the live code and equations. And we used the python programming language for this project. And python is a common language which used for the machine leaning, deep learning and data analytics. It is most popular for the libraries. Because python have the large collection of libraries like pandas, OpenCV and Numpy etc.

**3.4. Principal Component Analysis (PCA)**

Principal Component Analysis is a technique of unsupervised machine learning algorithms. Which is reduce the order of dimensions of a dataset. The principal component analysis is most useful their where image is no class labels. It projects data again in lower dimensions. And at that time also minimize information loss. It is used before training the model. Principal Component Analysis increases the efficiency of algorithms and reduces the space required to store the data. And x1 and x2 is a data. which is helping to get covariance matrix.

Chart, pie chart

Description automatically generated

Figure 3.2 Diagram of PCA

**3.5. Variance Explained using PC**

The variance is the amount of information that may be described by a single major component (eigenvectors) The ratio of the eigenvalue of a single main component (eigenvector) to the total eigenvalues is used to compute explained variance.

Each primary component number's variance is accumulated in the cumulative explained variance. Each primary component's variance is described by the individual explained variance. And the variance explained in percentage with corresponding the principal component in below the figure 3.3.

Chart, bar chart, histogram

Description automatically generated

Figure 3.3 Representation of Variance

**3.6. Proposed Methods**

The processing of data have done by the above mention classification algorithms. Which is take the input of fruit dataset . And applied the supervised machine learning algorithms on the dataset. Which clearly show in the below flow chart diagram in figure 3.4. And after completion all the steps we calculated the results. And got the accuracy of all the algorithms respectively.

**3.6.1 Flow Chart of Implementation**

In this section to described that the how the dataset classify from the input to output. And represented the flow chart of all the algorithms and how the dataset is working and how to classification occurred**.**

**Timeline

Description automatically generated**

Figure 3.4 Flow Chart of Proposed Method

**3.6.2 Flatten Data**

In this part the input data convert into the one dimensional array. Because input data present in multi-dimensions. Then has been flatted in one single vector (Image Features Vector) and then scaled subtracting the mean of the dataset in order to perform classification algorithms. And the image matrix has the three basic or fundamental colors that is red, green, and blue. which is separated from the image matrix to the individual color. and further individual color matrix convert into the one-dimensional array. This is the working of the flatten of image.

**Chart, box and whisker chart

Description automatically generated**

Figure 3.5 Image unfolding and Preparation of Image Matrix

**3.6.3 Fruits Sample**

**A picture containing text, red

Description automatically generatedA picture containing diagram

Description automatically generated**

(a). Apple (b). Pineapple

Figure 3.6 Fruits Sample

**3.6.4 Visualization of Data**

**Chart, scatter chart

Description automatically generated Chart, radar chart, scatter chart

Description automatically generated**

Figure 3.7 Data Representation In 2D Figure 3.8 Data Representation In 3D

**CHAPTER 4**

**EXPERIMENTAL RESULTS**

**4.1. Normalized Confusion Matrix**

Confusion Matrix is a performance measurement of classification model. It is the table of 4 different combination of predicted label and true label. And all confusion matrix of SVM, K-NN, Random Forest and Decision Tree are represented below with corresponding name. And it is most useful for the measure the precision, recall and ROC Curve. And the normalized confusion matrix diagram of fruits data represented below.

Chart

Description automatically generated Chart

Description automatically generated

(a). K-NN (b) Random Forest

Chart

Description automatically generated Chart

Description automatically generated

(c) SVM (d) Decision Tree

Figure 4.1 Normalized Confusion Matrix of different Classifier

**4.2 Performance Measure**

**4.2.1 Accuracy**

Accuracy tells the how many times model predicted correctly out of all positive and negative classes in the given model.

**4.2.2. Precision**

Precision indicates the positive prediction of model. And it is calculated by the given formula.

**4.2.3 Recall**

Recall quantifies the number of positive class predictions made of all positive examples in the dataset. And calculated the given formula.

**4.2.4 F-1 Score**

F-Score useful to measure recall and precision at same time. And it is used the harmonic means at place of the arithmetic mean by sending the extreme values more.

Where-

* **TP:** True Positive
* **TN:** True Negative
* **FP:** False Positive
* **FN:**  False Negative

And have plotted the precision, recall, and F1-factor in below the diagram. The experimental results and their accuracy present in the form of percentage below the diagram 15.

Figure 4.2 Plot Diagram of Precision, Recall, F1-factor And Accuracy

**4.3. ROC Curve**

The Receiver Operating Characteristic (ROC) Curve is a graph that displays a classification model's performance across all categorization levels. The true positive rate and the false positive rate are plotted on the ROC curve. All SVM, K-NN, Decision Tree, and Random Forest ROC Curve algorithms are included here.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

1. (b)

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

(c) (d)

Chart, line chart

Description automatically generated

Figure 4.3 ROC Curve of different Algorithms

**4.4. Performance of Algorithms with Time**

This section discusses which classification methods performed the best in terms of runtime. The algorithms SVM, K-NN, Random Forest, and Decision Tree take 12.75, 13.44, 7.89, and 7.47 seconds, respectively. The graph between classifier and time has also been drawn. The Decision Tree approach took the least amount of time, whereas the K-NN algorithm took the most. And the classifier is defined on the x-axis. and time on the y-axis.

Figure 4.4 Runtime Vs Classifier

**4.5. Result Summary**

We have got all the results of the classification algorithm described above. and determined the accuracy for all KNN, SVM, Random Forest and Decision Tree algorithms consisting of different test images and number of different training images from the dataset. And the support vector machine given the better results. And the accuracy of SVM and Random Forest is 100% because of small dataset. And K-NN and Decision Tree accuracy is 98% and 90% respectively. And based upon the runtime Random Forest performed the better. Because it took the least time to execute. And K-NN took the more time.

**CHAPTER 5**

**CONCLUSION AND FUTURE WORK**

**5.1. Conclusion**

After implementation of all the algorithms we found that SVM algorithms performed the better, but it gave the 100% accuracy because of the small dataset. And the K-NN algorithm performed better than the decision tree. And Random Forest performed better that decision tree. Because K-NN given the more accuracy comparison to decision tree algorithms. The accuracy of K-NN and Decision Tree are 98.49% and 91.57% respectively.

**5.2 Future Work**

In this implementation you can improve the accuracy of this project. And can use the large dataset for the better results. Because we are taking the dataset which have created by own. That’s why SVM is giving 100% accuracy. So that in the future can work upon this problem. And you can improve the accuracy of other described algorithms.

**REFERENCE**

[1] Yann LeCun, Fu Jie Huang, and Leon Bottou, “Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, 2004.

[2] Sandhya N. dhage, Sandhya N. dhage “A review on Machine Learning Techniques” International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 4 Issue: 3

[3] M. Welling, “A First Encounter with Machine Learning”

[4] Cheng, H., Damerow, L., Sun, Y., and Blanke, M. Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. Journal of Imaging 3, 1 (2017). ⇒4

[5] Ciresan, D. C., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. CoRR abs/1202.2745 (2012). ⇒6

[6] Oltean, M., and Muresan, H. Fruits 360 dataset on github. [Online; accessed 27.10.2018]. ⇒1, 10, 28

[7] Oltean, M., and Muresan, H. Fruits 360 dataset on kaggle. [Online; accessed 27.10.2018]. ⇒1, 10

[8] Wikipedia. Deep learning article on Wikipedia. [Online; accessed 05.05.2018]. ⇒5

[9] Liu Y, Bi J W, Fan Z P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm [J]. Information Sciences, 2017, 394 (9): 38-52

[10] Singh J, Singh G, Singh R (2017) Optimization of sentiment analysis using machine learning classifiers. Hum-cent Comput Inf Sci 7:32

[11] Aggarwal CC, Zhai CX (2012) A survey of text classification algorithms. In: Mining text data, Springer, Berlin, pp 163–222

[12] Nanba H, Kando N, Okumura M (2011) Classification of research papers using citation links and citation types: towards automatic review article generation. Adv Classif Res Online 11(1):117–134

[13] R. Muralidharan and Dr.C. Chandrasekar, “Object Recognition using SVM-KNN based on Geometric Moment invariant” International Journal of Computer Trends and Technology, ISSN: 2231-2803 Page 215 July to Aug Issue 2011

[14] Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley & Sons; 2012 Nov 9.

[15] Mureşan H, Oltean M. Fruit recognition from images using deep learning. Acta Universitatis Sapientiae, Informatica. 2018 Aug 1;10(1):26-42.

[16] Kausar, A., Sharif, M., Park, J. and Shin, D.R., 2018, December. Pure-CNN: A Framework for Fruit Images Classification. In 2018 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 404-408). IEEE.

[17] Ninawe P, Pandey S. A completion on fruit recognition system using k-nearest neighbors’ algorithm. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2014 Jul;3(7):2352-6

[18] Priya MM, Sujatha N. Recongnition of fruits using multiple features with neural classifier. Asian Journal of Computer Science and Information Technology. 2017 Jul 3; 7:37-9.

[19] Haralick RM, Shanmugam K, Din stein IH. Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. 1973 Nov (6):610-21.

[20] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics. 1991 May;21(3):660-74.

[21] Dudani SA. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics. 1976 Apr (4):325-7.

[22] Liaw A, Wiener M. Classification and regression by random Forest. R news. 2002 Dec 3;2(3):18-22.

[23] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep 1;20(3):273-97.

[24] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2011, 5 (4): 1-18.

[25] Liu Y, Bi J W, Fan Z P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm [J]. Information Sciences, 2017, 394 (9): 38-52.

[26] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35 (1): 221.

[27] J. R. Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81–106, 1986.

[28] Random Forest - an overview | ScienceDirect Topics.

[29] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.

[30] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin, “A supervised machine learning algorithm for arrhythmia analysis,” in Computers in Cardiology 1997. IEEE, 1997, pp. 433–436.

[31] Rahnemoonfar, M., Sheppard, C.: Deep count: fruit counting based on deep simulated learning. Sensors 17(4), 905 (2017) CrossRefGoogle Scholar

[32] Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C.: Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of ‘Mango YOLO’. Precis. Agric. 20(6), 1107–1135 (2019) CrossRefGoogle Scholar

[33] Hossain, M.S., Al-Hammadi, M., Muhammad, G.: Automatic fruit classification using deep learning for industrial applications. IEEE Trans. Ind. Inf. 15(2), 1027–1034 (2018) CrossRefGoogle Scholar

[34] Hameed, K., Chai, D., Rassau, A.: A comprehensive review of fruit and vegetable classification techniques. Image Vis. Comput. 80, 24–44 (2018) CrossRefGoogle Scholar