Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Flower Classification using Supervised Learning

Asmita Shukla
Computer Science and Engineering
Babu Banarasi Das National Institute of Technology and Management Lucknow, India

Ankita Agarwal
Computer Science and Engineering
Babu Banarasi Das National Institute of Technology and Management Lucknow, India

Hemlata Pant
Computer Science and Engineering
Babu Banarasi Das National Institute of Technology and Management Lucknow, India

Priyanka Mishra
Computer Science and Engineering
Babu Banarasi Das National Institute of Technology and Management Lucknow, India

***Abstract*—Biodiversity of earth is very rich. About 360000 create a healthy biome within the environment of earth. Some of them are identical in physical appearance like shape, size and color. Hence it is difficult to recognize any species. Similarly Iris flower species has three subspecies Setosa, Versicolor and Virginica. We are using Iris dataset because it is frequently available. The dataset of Iris flower contains 3 classes of 50 instances each. With the help of Machine learning, Iris dataset identifies the sub classes of Iris flower. The paper focuses on how Machine Learning algorithms can automatically recognize the class of flower with the help of high degree of accuracy rather than approximately. There are three phases to implement this approach are segmentation, feature extraction and classification. Using Neural Network, Logistic Regression, Support Vector Machine and k-Nearest Neighbors**

Keywords—Iris dataset, k-nearest neighbors, Logistic Regression, Neural Network, Scikit-Learn, Support Vector Machine.

I. INTRODUCTION

Machine Learning is program that learns from past data set to perform better with experience. It is tools and technology that we can utilize to answer questions with our data. Machine Learning works on two values these are discrete and continuous. The use and applications of Machine Learning has wide area like Weather forecast, Spam detection, Biometric attendance, Computer vision, Pattern recognition, Sentiment Analysis, Detection of diseases in human body and many more. The learning methods of Machine Learning are of three types these are supervised, unsupervised and reinforcement learning. Supervised learning contains instances of a training data set which is composed of different input attributes and an expected output. Classification which is the sub part of supervised learning where the computer program learns from the input given to it and uses this learning to classify new observation. There are various types of classification techniques; these are Decision Trees, Bayes Classifier, Nearest Neighbor, Support Vector Machine, Neural Networks and many more. Some example of Classification tasks are Classifying the credit card transactions as legitimate or fraudulent, classifying secondary structures of protein as alpha-helix, beta-sheet or random coil and categorize the news stories as finance, weather, entertainment and sports.

Python is a programming language created by Guido van Rossum in 1989. Python is interpreted, object-oriented, dynamic data type of high-level programming language. The programming language style is simple, easy to implement and elegant in nature .Python language consists of powerful libraries. Moreover, Python can easily combine with other programming languages, such as C or C++ or Java.

Scikit-Learn use the sciPy library of python as a toolkit. Scikit learn was originally called as "Scikits.learn". It includes dataset loading, manipulation and pre-processing of pipelines and metrics. Scikit Learn has a huge collection of machine learning algorithms.

II. REVIEW CRITERIA

The Iris data set is present at University of California Irvine Machine Learning Repository. The data set was firstly acquainted by Edgar Anderson in 1935 but due to use of many classification methodologies, it was further generalized by Ronald Fisher in 1936 and also known as Fisher's Iris data set. Hence the characteristics of data are multivariate and based on real values.

[1] David W. Corne and Ziauddin Ursani proposed in their paper an evolutionary algorithm for nonlinear discriminant classifier, in which they mentioned that it was not appropriate for learning tasks with any individual single value. Hence they tested this method on two data sets, Iris Flower and Balance Scale, where decisions of class membership can only be affected collectively by individual lineaments of flower.

[2] Detlef Nauck and Rudolf Kruse have proposed a new approach in which they classify the data on the basis of fuzzy Neural Networks. They used backpropagation algorithm to define other class of fuzzy perceptron. They concluded that on increasing the number on hidden layer, increase the need of more training cycles and raises incorrect results. Hence the better result can be evaluated using 3 hidden layers also.

[3] To overcome the problem of data depth, long parameters, long training time and slow convergence of Neural Networks, two other algorithms Transfer Learning and Adam Deep Learning optimization algorithms were considered for flower recognition by Jing FENG, Zhiven WANG, Min ZHA and Xiliang CAO. Where, Transfer Learning was based on

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

features in isomorphic spaces. They concluded in their paper that if the pictures of flowers placed into model training in the form of batches, then it will meliorate the speed of updating the value of parameters and provide the best optimal result of parameter values.

[4] Rong-Guo Huang, Sang-Hyeon Jin, Jung-Hyun Kim, Kwang-Seog Hong focus on recognition of flower using Difference Image Entropy (DIE),which is based on feature extraction. According to their research, the experimental results give 95% of recognition rate as an average. The DIE based approach takes original image of flower as an input, and applies pre-processing and DIE computation to produce recognition result.

III. PROPOSED ALGORITHM

Segmentation is the process which is used to remove the inadmissible background and consider only the spotlight (foreground) object that is flower. The main objective is to simplify the representation of the flower and to provide something which is more significant and easier to analyze.

In **Feature Extraction** we extract characteristics or information from flower in the form of real values like float, integer or binary. The primary features to quantify the plants or flowers are color, shape, texture. We do not prefer only one feature vector because the sub species have many attributes which are common with each other and produce less effective result. Therefore we have to measure the image by merging different feature descriptors which identify the image more efficaciously. The first five Iris datasets are represented in the given table 1.

Table1. The five Iris datasets [5]

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

After extracting features and labels from Iris dataset, we need to train the system. With the help of scikit-learn we create machine models, which **classify** the Iris flower into their sub species. The following table2 represents the descriptive statistics of Iris dataset.

Table2. The description of Iris dataset

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Neural Network, Iris Species have less feature, therefore multilayer perceptron is used as the currently architecture of neural network to preclude overfitting. In multilayer perceptron model, there is one scaling layer, two perceptron layer and one probabilistic layer. Iris dataset has four attributes, hence input layer consists of four variables these are sepal_length, sepal_width, petal_length and petal_width. The below graphs represent the relationship between SepalLength vs. SepalWidth (Fig1), PetalLength vs. PetalWidth (Fig2), SepalLength vs. PetalLength (Fig3) and SepalWidth vs. PetalWidth (Fig4)

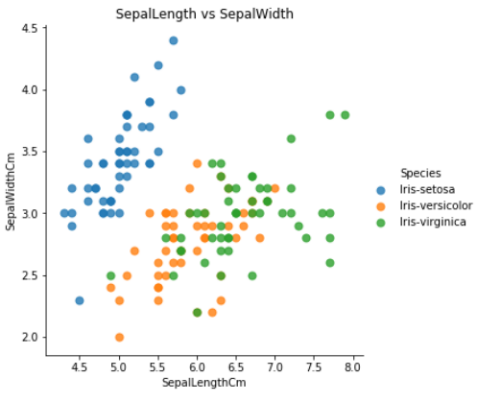


Fig1. Relationship graph of sepal length and width

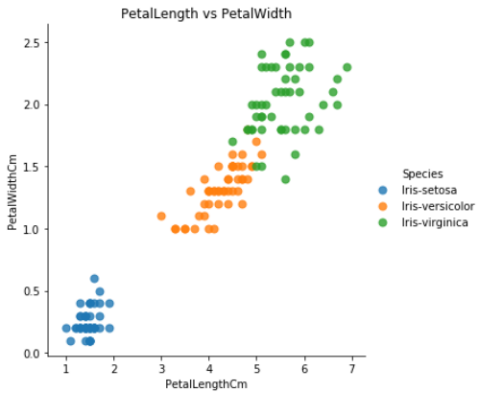


Fig2. Relationship graph of petal length and width

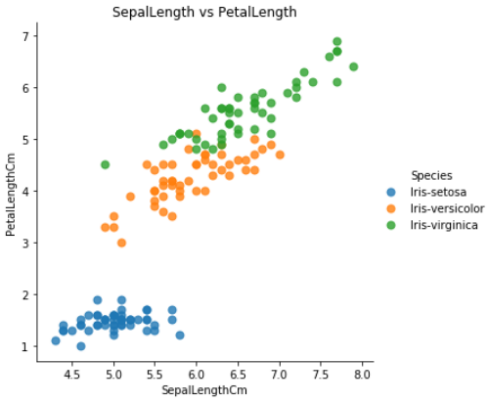


Fig3. Relationship graph of sepal and petal length

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

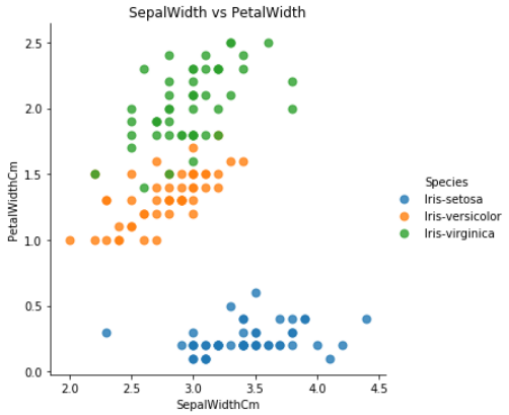


Fig4.Relationship graph of sepal and petal width

The scaling layer is used for normalizing the input values. We use mean and standard deviation scaling method to calculate normal distribution of sepal and petal, lengths and widths. The first perceptron layer consist 4 inputs and 3 neurons and the other perceptron layer consists 3 inputs and 3 neurons. Both two perceptron layer uses logistic activation function. The probabilistic layer permits the outputs to be represented as probabilities and normalizes the feature of each dataset to the range of 0-1. The neural network produces three outputs as Iris subspecies Setosa, Versicolor and Virginica.

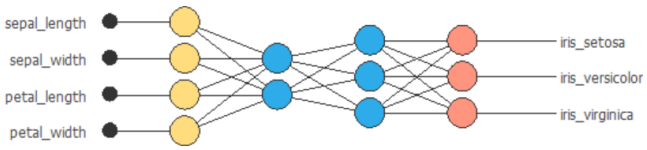


Fig5.Neural Network in the first perceptron layer [10]

Logistic Regression falls under the category of classification algorithms of machine learning. It provides a baseline for any binary classification problem and is also used for multinomial classification in which more than two classes are ascertained. Using link function, logistic model is transformed to predictor. Regularization [12] is used to identify error of overfitting and underfitting in the proposed model while training data.

Regularized Cost Function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

Regularized Gradient

$$\frac{\partial J(\theta)}{\partial \theta_j} = \left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j \quad \text{for } j \geq 1$$

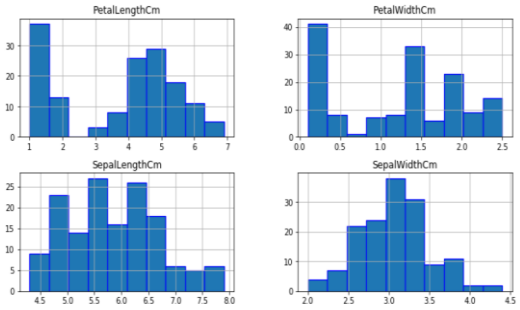


Fig6. The univariate plots are represented in the form of histograms

Support Vector Machine (SVM), In SVM dimensionality reduction techniques like Principal Component Analysis (PCA) and Scallers are used to classify dataset expeditiously. The first step towards implementation of SVM is data exploration. The initial configuration of hyper parameters like degree of polynomial or type of kernel are done by data exploration .Here we use two variables x and y, where x and y represent the features matrix and the target vector respectively. Dimensionality reduction is used to reduce the number of features in dataset which further reduces the computations. Iris dataset have four dimensions, with the help of dimensionality reduction it will be projected into a 3 dimensions space where the number of features is 3. We split the transformed data into two part, these are 80% of training set and 20% of test set.

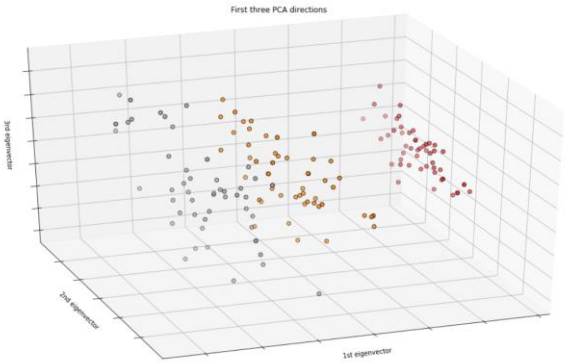


Fig7. The number of features in the new subspace is 3

k-Nearest Neighbors (KNN), KNN is used for both classification and regression. In KNN, we have labeled dataset which consists of training observations (X, Y) and would like to establish the relationship between X and Y

When KNN has unseen observation, then similarity is determine by distance metric between two data points. The distance can be measured by following methods [16]-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Where, n is number of dimensions, x is datapoint from dataset and y is new data point to be predicted.

KNN does not take string labels. Hence LabelEncoder is used to modify the string labels into numbers where Iris Setosa, Iris Versicolor and Iris Virginica are represented by 0, 1 and 2 respectively. Iris dataset are multivariate, therefore data visualization is done by several plotting methods like Parallel Coordinates, Andrew Curves, Pairplot, Boxplot. Implementing KNN with scikit-learn following three steps are performed: making decisions, evaluating predictions and using cross validation for parameter pruning.

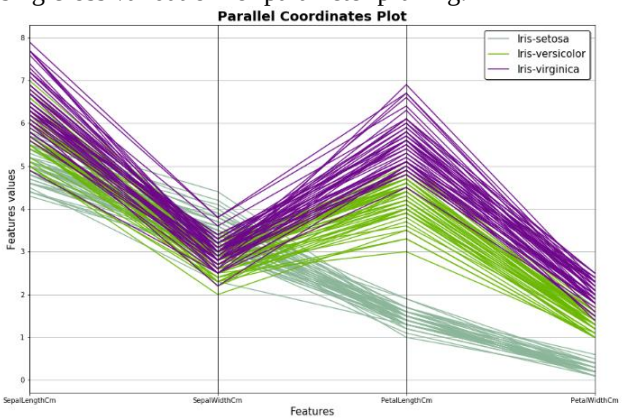


Fig8. For plotting multivariate data, Parallel Coordinates technique is used

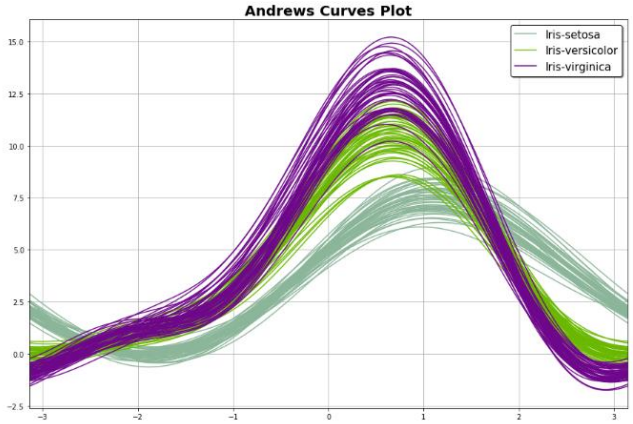


Fig 9.For visualizing the multivariate data Andrew Curves are used

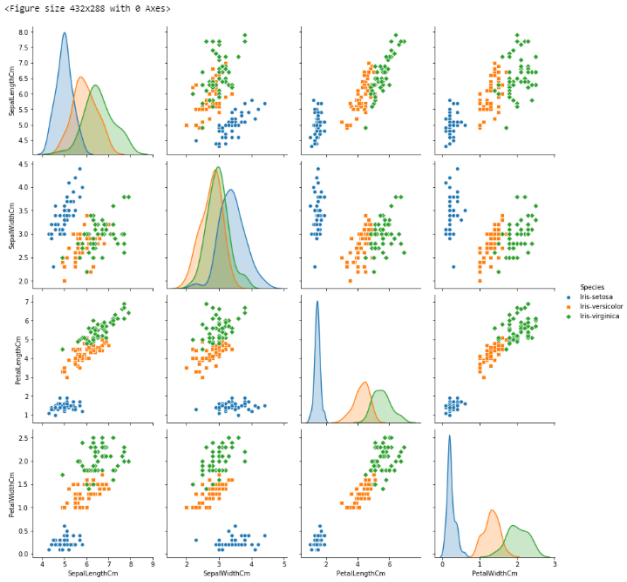


Fig 10.Pair Plot is used to visualize the distribution of the relationship between multiple variables separately

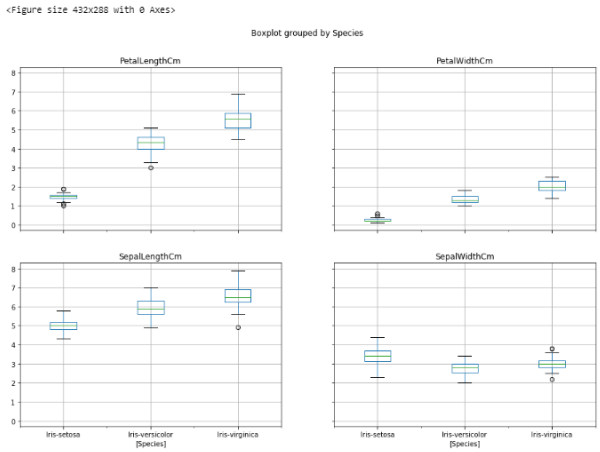


Fig11. The representation of data in from of Box Plots

IV. RESULT

1) An accuracy of 96.667% is achieved using Neural Network. We find that neural network learns from its existing feature and with the help of its weights and biases, it predicts the more accurate outcomes.

```
prediction=model.predict(X_test)
length=len(prediction)
y_label=np.argmax(y_test,axis=1)
predict_label=np.argmax(prediction,axis=1)

accuracy=np.sum(y_label==predict_label)/length * 100
print("Accuracy of the dataset",accuracy )
```

```
Accuracy of the dataset 96.66666666666667
```

2) The accuracy of logistic regression on calculating with scikit-learn is 96.6667%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

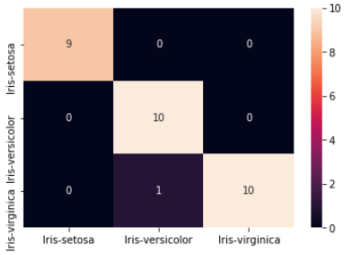
```
# Predictions
Prob = sigmoid(X_test.dot(all_theta.T)) # probability for each flower
pred = [Species[np.argmax(Prob[i, :])] for i in range(X_test.shape[0])]

print(" Test Accuracy ", accuracy_score(y_test, pred) * 100 , '%')
```

```
 Test Accuracy  96.66666666666667 %
```

```
# Confusion Matrix
cnfm = confusion_matrix(y_test, pred, labels = Species)

sb.heatmap(cnfm, annot = True, xticklabels = Species, yticklabels = Species);
```



Using sklearn model the accuracy is

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
print('Test Accuracy for Scikit-Learn model:', metrics.accuracy_score(y_test, y_pred)* 100,'%')
```

```
Test Accuracy for Scikit-Learn model: 96.66666666666667 %
```

3) The dataset consists of 4 dimensions, where PCA compress the data and provides the number of features in the new subspace. Using Linear SVM, the accuracy of training set is 0.97 and test set is 1.00 while using non-linear SVM, the accuracy of training set is 99.17% and test is 100. And on tuning the C parameter the best estimator accuracy on training set is 96 and on test set it is 100.

3.1 Using Linear SVC

```
Accuracy of linear SVC on training set: 0.97
Accuracy of linear SVC on test set: 1.00
```

3.2 Using Grid Search

```
The best parameters are {'C': 5.777360913698984} with a score of 96
Best estimator accuracy on test set 100.00
```

3.3 Using Non-Linear SVC

```
Accuracy of SVC on training set: 99.17
Accuracy of SVC on test set: 100.00
```

4) Using KNN classification, the accuracy of our model is evaluated to 96.67% and on finding the best k, the optimal numbers of neighbors is 9. (Fig12)

```
Accuracy of our model is equal 96.67 %.
```

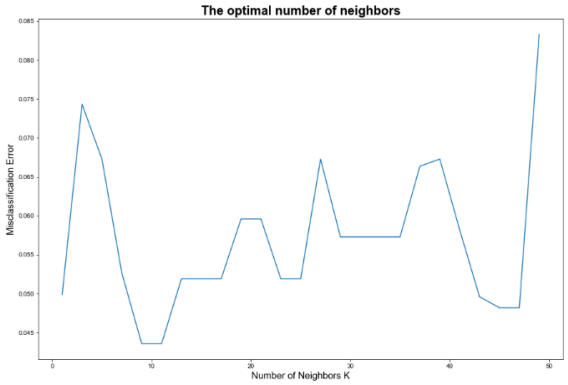


Fig12. The above diagram represents the optimal number of neighbours which is 9.

Table3. The accuracy of Classification model

	Algorithm used for Classification	Accuracy of Model
1.	Neural Network (NN)	96.6667%
2.	Logistic Regression (LR)	96.6667%
3.	Support Vector Machine (SVM)	98.00%
4.	k-nearest neighbors (k-NN)	96.67%

Table 4.The accuracy of model using SVM

	Support Vector Machine(SVM)	Accuracy of Training Set	Accuracy of Test Set
3.1	Linear SVC	97	100
3.2	On Tuning the C parameter	96	100
3.3	Non-Linear SVC	99.17	100

REFERENCES

1. Ziauddin Ursani and David W. Corne , “A Novel Nonlinear Discriminant Classifier Trained by an Evolutionary Algorithm” DOI: 10.1145/3195106.3195132
2. Detlef Nauck and Rudolf Kruse, “NEFCLASS-A Neuro-Fuzzzy approach for the classification of data” DOI: 10.1145/315891.316068
3. Jing FENG, Zhiwen WANG, Min ZHA and Xinliang CAO, “Flower Recognition Based on Transfer Learning and Adam Deep Learning Optimization Algorithm”. DOI: 10.1145/3366194.3366301
4. Roung– Guo Huang, Sang-Hyeon Jin, Jung –Hyun Kim and Kwang-Seck Hong, “Flower Image Recognition Using Difference Image Entropy”. DOI: 10.1145/1821748.1821868
5. UCI Machine Learning Repository- IRIS DATASET
6. Introduction of Machine Learning and scikit-learn, Available at- https://youtu.be/GwIo3gDZCVQ https://youtu.be/rvVkVsG49uU
7. Introduction to Python Programming Language, Available at https://www.ritchieng.com/machine-learning-iris-dataset/ https://www.geeksforgeeks.org/python-language-advantages-applications/

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 05, May-2020

https://medium.com/gft-engineering/start-to-learn-machine-learning-with-the-iris-flower-classification-challenge-
https://www.theseus.fi/bitstream/handle/10024/64785/yang_yu.pdf?sequence=1&isAllowed=y

8. Image Classification using Python and Scikit-learn https://gogul.dev/software/image-classification-python
9. Image Segmentation https://en.wikipedia.org/wiki/Image_segmentation
10. Classification of iris flowers from sepal and petal dimensions using Neural Designer https://www.neuraldesigner.com/learning/examples/iris-flowers-classification
11. Neural Network from Kaggle by Louisong available at https://www.kaggle.com/louisong97/neural-network-approach-to-iris-dataset
12. Logistic Regression from pluralsight available at https://www.pluralsight.com/guides/designing-a-machine-learning-model
13. Support Vector Machine from Kaggle by Moghazy available at https://www.kaggle.com/moghazy/classifying-the-iris-dataset-using-svms
14. k- Nearest Neighbor from Kaggle by skalskip available at https://www.kaggle.com/skalskip/iris-data-visualization-and-knn-classification
15. Exploratory Data Analysis of IRIS Data Set Using Python available at https://medium.com/@avulurivenkatasaireddy/exploratory-data-analysis-of-iris-data-set-using-python-823e54110d2d
16. Distance functions: https://www.saedsayad.com/k_nearest_neighbors.htm