International Journal of Scientific Research in Mathematical and Statistical Sciences **Research Paper**
Volume-9, Issue-6, pp.01-10, December (2022) **E-ISSN:** 2348-4519

Classification of Iris Flower Dataset using Different Algorithms

S.A. Mithy[1*], S. Hossain[2], S. Akter[3], U. Honey[4], S.B. Sogir[5]

[1]Center for Community Health and Research, Gonoshasthaya Samaj Vittik Medical College, Savar, Dhaka-1344, Bangladesh
[2]Center for Multidisciplinary Research, Gono Bishwabidyalay, Savar, Dhaka-1344, Bangladesh
[3]Dept. of Economics, Gono Bishwabidyalay. Savar, Dhaka-1344, Bangladesh
[4]Dept. of Computer Science and Engineering, Gono Bishwabidyalay. Savar, Dhaka-1344, Bangladesh
[5]Dept. of Statistics, Jahangirnagar University, Dhaka-1342, Bangladesh

[*]*Corresponding Author: sohel6944@gmail.com, Mob: +8801943798635*

Available online at: www.isroset.org | DOI: https://doi.org/10.26438/ijsrmss/v9i6.110

Received: 16/Oct/2022, Accepted: 19/Nov/2022, Online: 31/Dec/2022

Abstract— The Iris dataset is one of the most famous dataset containing data on four attributes named as Sepal.length, Sepal.width, Petal.length, Petal.width and three classes or subspecies named as Sentosa,Versicolor and Virginic each class has 50 samples. The measurement of four attributes in CM (centimeters). This data set was developed by Ronald Fisher in 1936. This is available on UCI data set. In this study we want to show that how to solve the classification problem using some algorithms like K-means clustering, Random Forest decision, SVM, Logistic Regression, KNN, K-medoids. In addition, we also worked on four features to a advanced feature. The scikit tool we use for implementation. In this study applies classification and regression algorithms on the iris dataset by discovering and analyzing the patterns.

Keywords— Iris dataset, Logistic Regression, k-nearest neighbors, Support Vector Machine, Random Forest, K-means clustering, K-medoids.

I. INTRODUCTION

Ronald Fisher world was born 17 February 1890 in England. He is one of the most famous biologists and statistician in the whole world. Who introduce Fisher's Iris data set / Iris flower data set with us by his paper named "The use of multiple measurements in taxonomic problems" [1]. The data was collected by the Edgar Anderson from the morphologic variation flowers of Iris different related species that's why sometimes it's called Anderson's Iris data set. Classification is the most regularly used tasks of decision making for human activity. A problem of classification occurs when an item needs to assigned into a class based on a number of observed assign related to that item.

Data mining is one of the most useful areas for researcher, statistician, computer and data scientists. Nowadays we have lot of data but there is no meaningful information. We need meaningful information from large number of data. For find out meaningful information we need some technique. One of the most well-known techniques is data mining. This technique is the process of takeout useful information from a collection or large number of data in order that, data could be grouped classified of prediction of future and past [2].

In last decade there are huge research work have been done by using several techniques of data mining in the part of marketing [3], business [4], stock market [5], agriculture [6], and pharmaceutical products [7]. Tools of data mining includes capable of mathematics, statistics and analytics. The purpose of this technique is identifying trends, relationship, patterns to keep up informed planning and making a decision.

In this work a new technique is presented for Iris flower identification. It works training phase and testing phase. Training phase dataset are loaded through MLC (Machine Learning Model) during training and assigned the labels. Also the model predictive, predicts to those species from Iris flower data set belongs to. Then Iris species get labeled. This paper we will focus on classification of iris flower species by using machine learning algorithms with scikit tools. For iris data set classify we should have to discover design by examining sepal and petal size of the iris flowers. Therefore, the prediction is done by examine the design to form iris flower class. In this research work we instruct the model of machine learning along with data. If any unknown data is finding then the model of predictive will predicts the species of flowers from what it learned through trained data. Our main task is identifying the species raised on iris flowers characteristics. At last, classification models and the dataset are discussed and analyzed.

II. RELATED WORK

Nowadays data mining is take part in significant role. Generally, data mining methods can be making out into

various categories such as prediction, regression, and classification and clustering. These various methods have been effectively used in various areas.

A model of the Iris species system is described from this work we can see that Patrick conducts the iris dataset to apply oneself to the statistical analysis. In his study he analyzes two different systems. The dataset is plotted for the shake of determine the different sample in classification. He was able take out statistical details by evaluating an application with java [15].

In the other study of Wong P. H. he used bunching calculations for iris flowers. Bunching is impartial by the theory of chart. The theory of chart is utilized. He connected various AI (artificial-intelligence) devices with K-NN. In his study he fined the result is 96% [16].

The model for the System of Iris Flowers is described as the proposed technique is carried out on Iris data and classifies the dataset among 4 classes. In this exposition, the network can select the well features and take out a small but suitable set of regulation for the classification job. For Class one data set we acquired 0 confuse on set of test and for any other set of data the outcome obtained are close to the results indicated in the literature [17].

A model of the Iris species system is described as implemented their technique can automatically identify the Irish flower class with three talk to segmentation, classification and extraction of feature. Using SVC, k-NN, Logistic Regression and Neural network [18].

III. METHODOLOGY

To implementation different types of machine learning methods for accept which method is given valid result for point out iris specie data set associate to. We use some algorithms like as K-medoids, K-means clustering, Random Forest decision, SVM (Support vector Machine), Logistic Regression, and KNN (k-Nearest Neighbors). We use all this algorithm in scikit tool kit depends on python.

Data Set:
In this paper we use a built-in data set from scikit project which is open source named as IRIS data set. In is It contain four (04) attributes named as Sepal-length, Sepal-width, Petal-length, Petal-width and three(03) classes or species named as Sentosa, Versicolor and Virginic each class has 50 samples. The total samples of iris data set are 150.

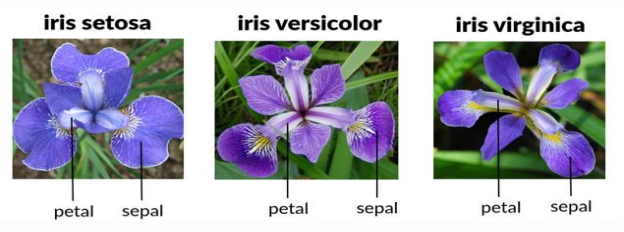


Figure 1: Three species of IRIS flower

Source of Data
As this dataset is built-in data set from scikit project which is open source.

Data Variables:
There are five variables in this data set:

1. Sepal length: (It is use as input data which is measured by centimeters)
2. Sepal width: (It is use as input data which is measured by centimeters)
3. Petal length: (It is use as input data which is measured by centimeters)
4. Petal width: (It is use as input data which is measured by centimeters)
5. Class: (It is use as target data which is measured by centimeters)
   The variable of class contains three (03) group or species named as iris Sentosa iris Versicolor and iris Virginic.

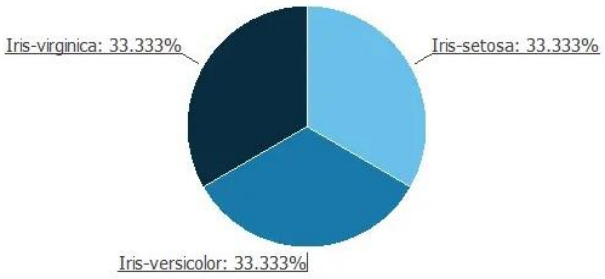


Figure 2: IRIS data set Phi-chart

Table 1: Characteristics the Data Set Feature with in Statistical information

Feature	Meaning	Range	Average
Sepal_ length	Lower length of leaf	4.3-7.9	5.843
Sepal_width	Lower width of leaf	2.0-4.40	3.057
Petal_ length	Upper length of leaf	1.0-6.9	3.758
Petal_width	Upper width of leaf	0.1-2.5	1.199

Source: Python coding

From table 1 we see that, the feature of Sepal length and width range is 4.3 to7.9 and 2.0 to 4.40, average range is 5.843 and 3.057, which is Lower length and width of leaf. The feature of Petal length and width range is 1.0 to6.9 and 0.1 to 2.25, average range is 3.758 and 1.199, which is Upper length and width of leaf.

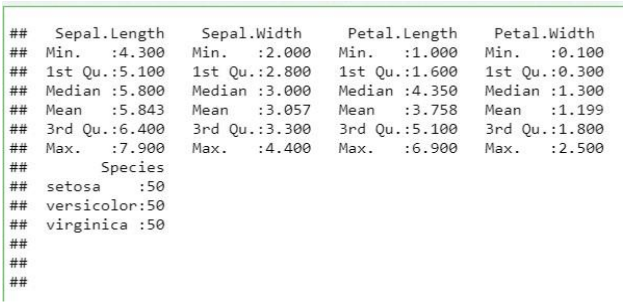
Table 2: IRIS data set samples

No.	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5	3.6	1.4	0.2	Setosa
…	…	…	…	…	…
146	6.7	3	5.2	2.3	Verginica

147	6.3	2.5	5	1.9	Verginica
148	6.5	3	5.2	2	Verginica
149	6.2	3.4	5.4	2.3	Verginica
150	5.9	3	5.1	1.8	Verginica

Source: Python coding

Table 1 represents that, 150 sample of Iris dataset. From 1st row we can see that, Sepal Length and width is 5.1 cm and 3.5 cm. Petal Length and Width is 1.4 cm and 0.2 cm.



```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Source: Python coding

Figure 3: Summary of IRIS data

Figure 3 shows that, summary of iris dataset. From figure we can see that, mean of Sepal Length and width is 5.843 and 3.053, mean of Petal Length and Width is 3.758 and 1.119. Median of Sepal Length and width is 5.80 and 3.00, Median of Petal Length and Width is 4.650 and 1.30.

Training and Testing Phase:

When we will recognize about the data set then we should have to use the data set to train our ideal for forecast output value exactly. At 1st we should have taken some sample for train our model. We divided our data set into two parts (training and testing). The rate of training and testing parts is 80:20. Where, training part use 80% data and tasting part use 20% data. Our focal point on the classify iris flower class with expulsion of data from iris dataset. RFC (Random Forest Classifier) method was used for classifying testing data and to get iris flower color codes. We find high level accuracy using algorithm of k-means method.

IV. ALGORITHMS USE

K-means clustering:

This is the easiest unsupervised algorithm for machine learning. For unlabeled or large data, we use k-means algorithm, where there is no prediction variable. This algorithm is faster than the other algorithms cluster, the aim of this algorithm to find out groups in the data. It allows us form clusters based on same feature. This algorithm is based on distance, where we evaluate the distances to allocate a point to a cluster. In K-Means, every cluster is associated with a distance. We can use many clusters if we want, depends on our objectives, although, the more clusters, the less likely we are to construct generalized assumptions over the data. To get accurate result we use Euclidean distance. We calculate distance among many points. To find out the minimum distance we compare all points with each other's. From the result, we predict the accurate result which stays 0 to 1. If the result is close to 1 then we say that this point is better accurate from others.

Mathematical equation of Euclidean distance

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}\;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where, p,q = two points, p_i,q_i = Euclidean vectors, n = n-space.

Suppose, we use 3 and 9 clusters for 3 species with the dimension of sepal and petal in iris data set where k=3 and 9. At 1st we input iris data set from sklearn, load k-means and libraries which we needed. We use some python programming code to get the figure 4.

Source: Python coding

Figure 4: K-Means pair-plot where K=3.

The Figure 4 shows that the species are fully separated into clusters build on their assessments. This is the occasion with all the combinations without when plotting sepal length and width. These are a bit difficult to different. When the species of setosa in this diagram is low clustered as it's another diagram, virginica and Versicolor are quite well compounded together here building it more complex to tell anyone group from another individually. This can be good because it will assume that it clustered the data similarly to the species labels but only by using their sepal and petal dimensions.

Table 3: K-Means clustered the samples against their respective species, where k=3

Col_0	0	1	2
Species			
Setosa	0	50	0
Versicolor	2	0	48
virginica	36	0	14

Source: Python coding

Table 3 shows that Setosa is completely grouped. Virginica and Versicolor twain have a number of irrational groupings. This is most possible to the sepal width of twain species existence relatively alike when compared to sepal length. Maybe if there are more clusters, they can help fall off the areas that is mixed among two species and their area.

Source: Python coding

Figure 5: K-Means pair-plot where K=9.

From figure 5 we can shows that, the pair-plot is slightly strong to compare in case of the actual colour and plot by species, although, we can see that there are several specific clusters that cluster at one place.

Table 4: K-Means clustered the samples against their respective species, where k=9

Cluster_2	0	1	2	3	4	5	6	7	8
Species									
Setosa	0	22	0	0	0	0	0	20	8
Versicolor	19	0	0	23	0	4	4	0	0
virginica	0	0	12	1	22	15	0	0	0

Source: Python coding

Table 4 shows that K-Means clustered where k=9 is better than K-Means clustered where k=3. The number of improperly clustered samples is much less than in the former cluster. By selecting the species by including the cluster while the determine species cluster we include all the clusters at once and see a great were not correctly clustered make use of the species as mentioned.

Table 5: K-Means clustered result

K=3 kMeans	89.333%
K=9 kMeans	96.666%

Source: Python coding

From table 5 we see that, the accuracy of every model is calculated. To do this we divided it by the all number of observations (i.e. Correct + incorrect predicted) and add all correct predicted observations. Also show that, the accuracy of K-means where k=3 and 9 is 89.333 % and 96.666 %. The result of k-means where k=9 is better than k=3. This algorithm discovers significant construction out of data and find out fundamental patterns. It reduces the all squared error.

Random Forest Classifier (RFC):

RFC is an algorithm of machine learning which was promoted by Leo Breiman [8]. This method may be used for classification (Classes), regression (mean prediction) and any other task, it often called random decision forests, which work by construction a large number of data. It gives us fast prediction and very good accuracy. It may be reduces repetition and removes duplicate values, it provides the outcome by taking on an mean of the decision trees. The general equation of RF model is

$$H(X,\theta_j)=\sum_{i}^{k} h_i(X,\theta_j) \text{ ---------- (2)}$$

Where, j= 1, 2, 3….. m

$H(X,\theta_j)$ = Meta determination tree classifier

X= input characteristic vector for training dataset

θ_j = self-made and systematically allocate for random vector that establish the growth operation of tree [9]. Classifies of RF work with k means together. At first we check the RF accuracy which depends on classification strength of individual trees. When the accuracy is nearly one then the program accept that and go to next step. Mutual OOB measurement depends on which is possible estimate only committed of error but also relevance of every attribute for classification's purpose.

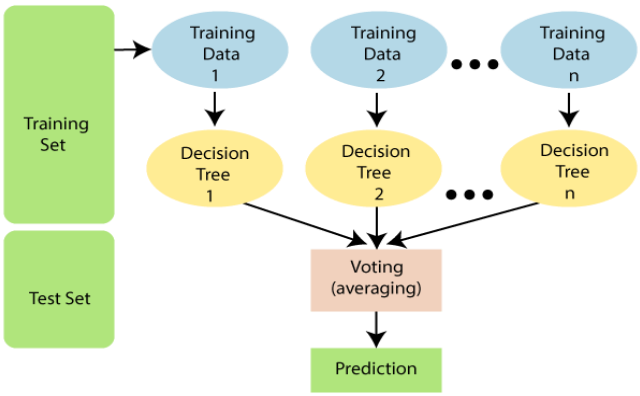


Figure 6: RF Algorithms Structure for training dataset

The important measurements of TRF made 3 attributes it's showed in figure 8. From the figure 8 the attribute "petal length" is most effective on train model which is high important parameter on the other hand the attribute "sepal width" is less high important parameter.

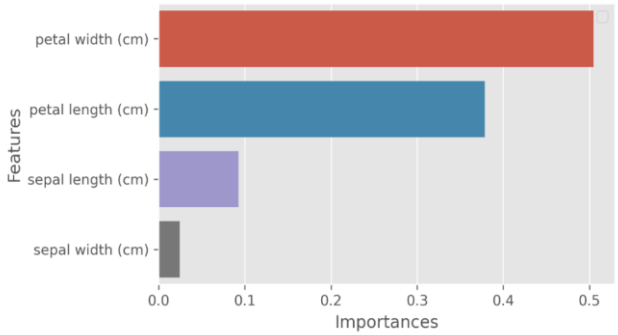


Figure 7: Attributes significant graph

The RF trained model tested for 30 data. The RF confusion matrix formed since estimation process and test in figure 8.

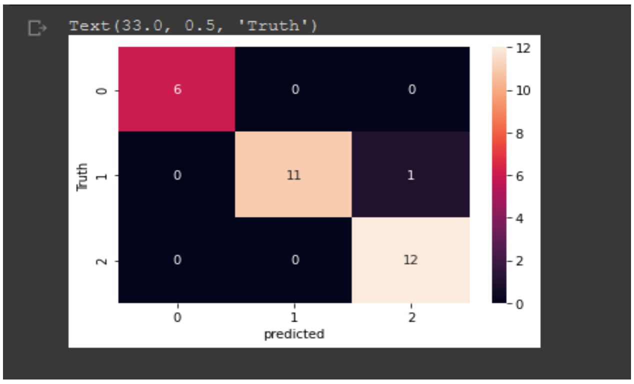


Figure 8: RF Confusion Matrix

From figure 8 we can see the correct and incorrect number which estimated by RF. The figure correctly predicted 29 data within 30 data. It is clear that this model get even so detects correctly approximately of accuracy is 97% with the classes. Also we used method of macro calculation. The methods of macro calculation result are shown in Table 5.

Table 6. RF Evaluation results

Criteria of Evaluation	Value
Accuracy	97% or 0.97
Sensitivity	96% or 0.96
Specificity	97% or 0.97
F1 score	97% or 0.97
AUC	97% or 0.97

Table 6 shows that the model get evaluation outcome are given in Table 2 When Table 2 is look into, it is notice that the model acquire is perfectly effective. Trained model accuracy was appointed as 97% or 0.97. The value of AUC is 97% or 0.97 which is so near to 1. We use curve of ROC graphically to check out model accuracy.

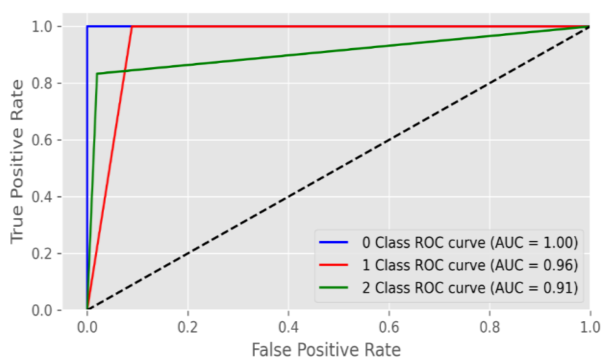


Figure 9: RF of AUC value and ROC curve

From the figure 9 we see that, the curve of ROC is given for every classes and values of AUC. The curve result is not far from to a perfect ROC curve to iris flower detection.

SVM (Support Vector Machine):

In modern times this algorithm (SVM) method are used extensively in different fields like as intrusion and face detection, image gene and email classification, Bioinformatics, Handwriting and text recognition, agriculture, engineering and so on [10]. Generally it use for solving classification and regression model and also modify nonlinear mapping. There are two main types of this method firstly SVC and secondly SVR which is used for learning regression and nonlinear regression [11].

$$\{f \mid f(x)\} = w^T x + b, w \in \Re^d, b \in \Re \text{...(3)}$$

Where, x = data set of training, w = weight vector and b = threshold value. We use equation (3) for find SVR highest exact hypothesis function set. The aim of SVM is divided the data into various classes by line, the best way is hyperplane. Which class point is close to hyperplane is familiar by SVM. There are much hyperplanes that will be individual classes, hence to choose exact hyperplane. The algorithm of SVM finds the closest point to hyperplane of classes and examines the distance among SV and hyper plane. Here the distance is carry out the margin. SVM select the hyper plane to gives the highest margin.

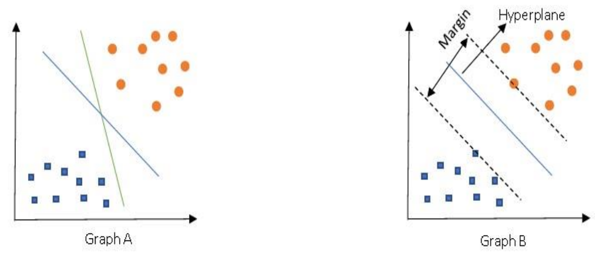


Figure 10: Hyper plane select

There are two graphs (Graph-A and Graph-B) in figure 10. Graph-A shows that, green line and blue line. We select blue line is hyperplane because it gives us highest margin which is support vectors and this line represents the space. We use Scallers and PCA (Principal Component Analysis) to quickly classify dataset for reduction dimensional method which reduce number of features and computations dataset. At first we implement data exploration for SVM. For data exploration we use class of kernel or hyper parameters configuration such as polynomial degree. We use variable "x" which deals with the matrix features and "y" which deals with target vector.

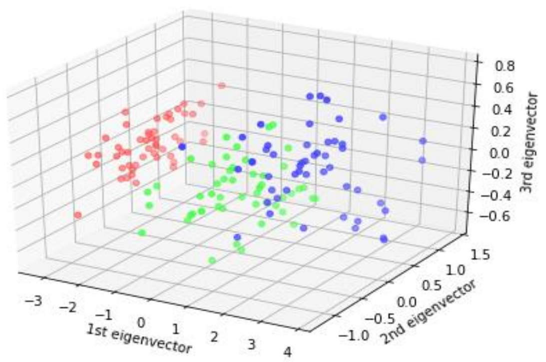


Figure 11: The three new subspace features of Irish dataset

Iris dataset is four dimensional using method of reduction dimension reduce Iris dataset in (03) three dimensions where the features number is 3. We divided our data set into two parts training and test parts 80% and 20% respectively.

Logistic Regression:

We use this algorithms for solving classification problems and to predict the probability either an instance going along with 1 class or other class. It apply numerical or categorical predictor variables. This regression is also known as Logit Regression or Model which is mathematical model use to estimate probability of event occurring given some foregoing data. Generally it's work with binary data 0 and 1. If the event can happens 1 or true otherwise event can't happen 0 or false. We use link function for this model transformed through predictor we also use Regularization [12] to point out over fitting error and under fitting the prefer model for training data.

Regularzed Cost

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\log(h_\theta(x^{(i)}))-(1-y^{(i)})\log(1-h_\theta(x^{(i)}))]+\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad \ldots\ldots 4)$$

Regularzed Gradient

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

for $j \geq 1$ ……(5)

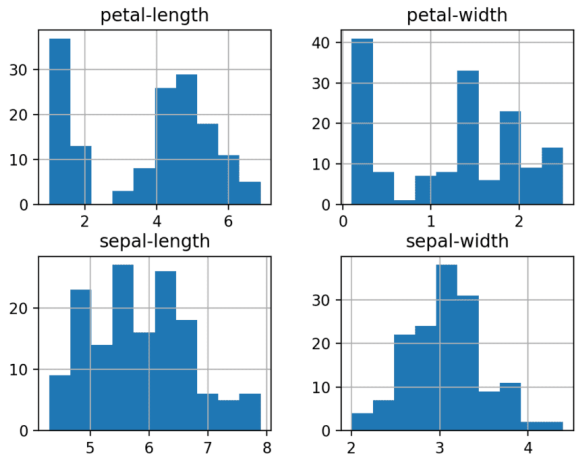


Figure 12: The histograms plots of univariate

KNN (k-Nearest Neighbors):

k-NN method is supervised classification or nonparametric algorithm . Naturally we used this method for prediction and classification. Also use for distribute a point depends on nearest neighbors observe by neighbors of Euclidian distance and most of vote as principle part for predication and classification. The basic idea of this method is to define data by computing the k-NN point of data. In another word, calculate a gap among the test data and its feedback to make the proper prediction. Therefore, the most common point of class is the assign to those k-NN [13]. Which is shown in the Figure.

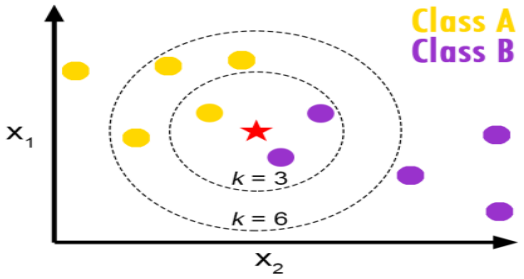


Figure 13: Classification of k-NN

There are a several different metrics to measuring the dissimilarity among the two variables. In this algorithm our dataset is labeled which involved observations of training (X, Y) for this we want to established a interrelationship among variable of X and Y. When KNN have invisible observation, then similitude are calculate through the distance metric among point of two data. We use some methods to calculate the distance. When k=1, the sample is just imposed to sample of their closest neighbor. All distance calculate only continuous variables. For categorical variables we use Hamming distance which conducted for the numerical variables standardization among 1 and 0 when the both categorical and numerical variables are mixed in dataset.

Euclidian distance

$$\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \quad \ldots\ldots (6)$$

Manhattan distance

$$\sum_{i=1}^{k}|x_i-y_i| \quad \ldots\ldots (7)$$

Minkowski distance

$$\left(\sum_{i=1}^{k}(|x_i-y_i|)^q\right)^{\frac{1}{q}} \quad \ldots\ldots (8)$$

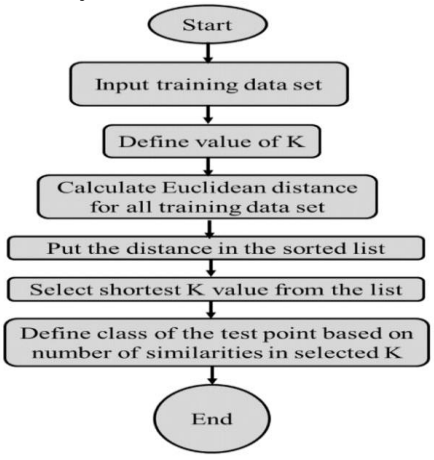
Hamming distance

$$D_H=\sum_{i=1}^{k}|x_i-y_i| \quad \ldots\ldots (9)$$

$$x=y \Rightarrow D=0$$

$$x\neq y \Rightarrow D=1$$

Where, x = data point and y= current point of data for forecast. KNN cannot take string tag, to change string tag into number we use LabelEncoder from sklearn. Where, Iris Setosa, Versicolor, Virginica which is represent 0,1 and 2 sequentially.



K-Nearest Neighbors Flow chart

We use some plotting system such as boxplot is used for data representation, we use Pair Plot for visualize the format of the interrelationship among many variables individually, Andrew Curves are used for visualizing the multivariate data, Parallel Coordinates are used for plotting multivariate data. Then there are three steps for implementation KNN for Iris dataset. The steps are evaluating predictions, making decisions and parameter pruning.

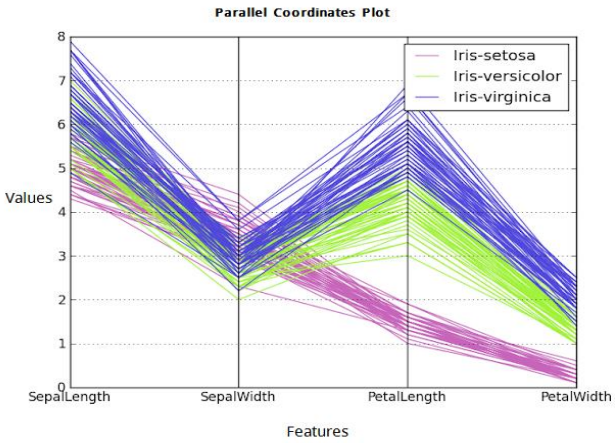


Figure 14: Parallel Coordinates plot

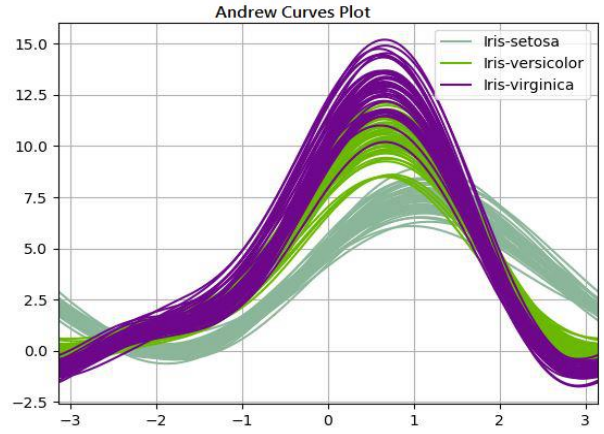


Figure 15: Andrew Curves Plot

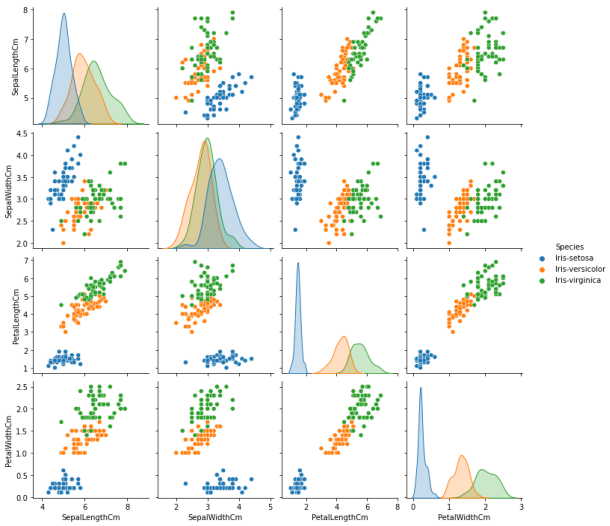


Figure 16: Pair Plot

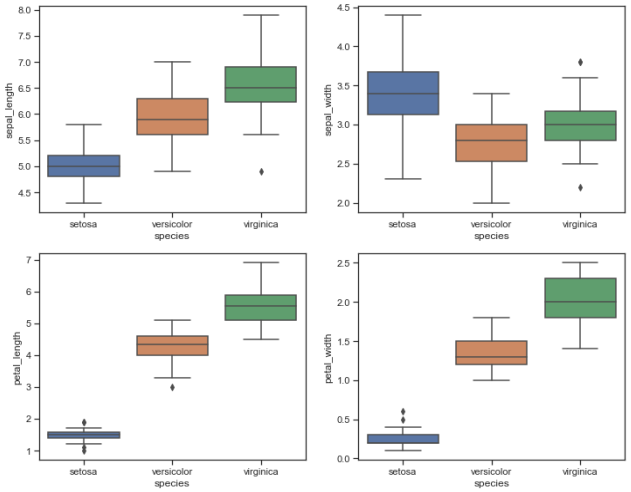


Figure 17: Pair Plot

K-medoids:

k-medoids is one of the most renowned clustering algorithm . The aim of this algorithm is find the grouping in data sets. It is very closely same to k means method but it has so little difference and also it is slightly different boost or optimization function from there on k-means. This method is better option to K-means method [14]. We use Manhattan distance equation for this algorithm. Manhattan distance

$$j = \sum_{i=1}^{k} \sum_{p \Omega c_j}^{n} \left\| P - O_j \right\| \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

The total variation among two points is minimized. A point describes in a cluster and others point selects as the midpoint of the cluster are minimized. Using mesh function we made a plot of mesh which provides us a 3D surface. This method provides a well prediction and minimize over fitting of this models.

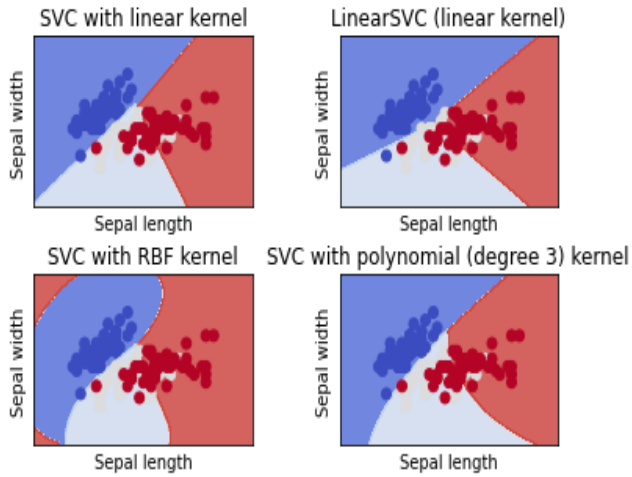


Figure 18: Graph of K-medoids

V. RESULTS AND DISCUSSION

Logistic regression algorithm:

Calculate the logistic regression with scikit we get the result of accuracy is 96.6667%. Which is showing the bellow.


```
from sklearn.linear_model import LogisticRegression
from sklearn.linear_model 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: 96.66666667
```

SVM (Support Vector Machine):

The Irish dataset involved 4 dimensions where Principle components analysis compress the data which provides features numbers for new subspace. We get accuracy 1.00 for test set and 0.97 for training set while we use Linear SVM. While we use non-linear SVM get accuracy 100 for test set and 99.17% for training set.

SVM with Linear

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

SVM with Non Linear

```
Accuracy of Non-linear SVC on training set: 99.17
Accuracy of Non-linear SVC on test set: 100.00
```

KNN classification:

Calculate for the KNN classification we get the result of accuracy is 96.67% and find k which is best optimal number for neighbors is showing in figure.

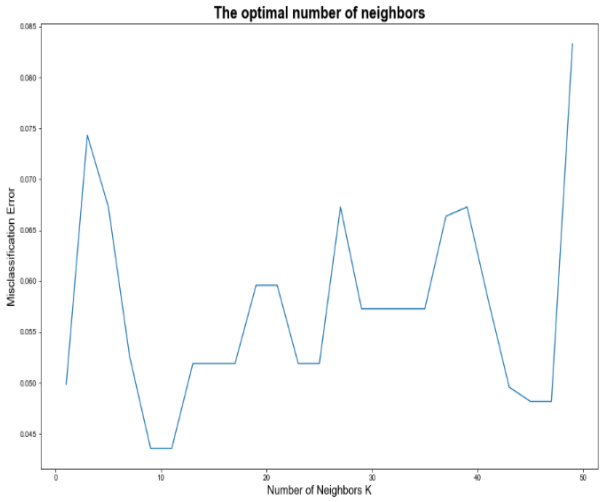


Figure 19: Neighbors optimal number is 9.

Table 07: Classification model accuracy

Algorithms	Accuracy
KNN	96.67%
Linear SVM	97%
Non Linear SVM	99.17%
Logistic regression	96.6667%.
Random Forest	97%
k-Means (K=3)	89.333%
k-Means (K=9)	96.666%

VI. CODE

KNN Algorithm

```
from sklearn.neighbors import KNeighborsClassifier
model_1=KNeKNeighborsClassifier(n_neighbors=5)
model_1.fit(x_train,y_train)
```

Random Forest

```
from sklearn.ensemble import RandomForestsClassifier
model_2=RandomForestsClassifier(n_estimators=5)
model_2.fit(x_train,y_train)
```

Support Vector Machine

```
from sklearn.svm import SVC
model_3= SVC(kernel='linear')
model_3.fit(x_train,y_train)
```

To find the model accuracy of predictions for test data and y test.

```
print("Accuracy of logistic regression algorithom:", accurecy_score(y_test, prediction))
print("Accuracy of KNN algorithom:", accurecy_score(y_test, KNN_prediction))
print("Accuracy of Random Forest algorithom:", accurecy_score(y_test, RF_prediction))
print("Accuracy of SVM algorithom:", accurecy_score(y_test, SVM_prediction))
```

```
Accuracy of logistic regression algorithm: 0.966666666667
Accuracy of KNN algorithm: 0.966666666667
Accuracy of Random Forest algorithm: 1.0
Accuracy of SVM algorithm: 1.0
```

Logistic regression

```
model=LogisticRegression()
model.fit(x_train,y_train)
```

VII. CONCLUSION AND FUTURE SCOPE

At the present time, classification is one of the most often uses in machine learning problems with a number of applications like as flower classification, biometric identification ,clustering, identity verification and others. In order to construct a model, the classification algorithms make a relation between the input and output characteristics and attempts to predict the target population with the more accuracy. In this study we want to make forecast on invisible data which are not accustomed to train of the model, therefore the model of machine learning construct need to accurately predict the Irish flowers classification of future on behalf of the accurately of forecasting the label of previously trained data. The principle objective of this study was to come to a consensus on how well K-means clustering, Random Forest decision, SVM, Logistic Regression, KNN, K-medoids algorithms performed in IRIS flower classification. In the analysis we found that when the percentage of training data improves, so does the degree of

precision. In comparison to random forest, which achieved 97% accuracy, Logistic regression algorithm decision, which achieved 96.6667% accuracy SVM with Linear, which achieved 97% accuracy, SVM with Non Linear, which achieved 99.17% accuracy, KNN, which achieved 96.67% accuracy. In the future, analyses on separate data sets will be generated, and different methods will be utilized and mixed to produce improved distinction results and also expand the accuracy by using dissimilar models.

REFERENCES

[1] problems". Annals of Eugenics. Vol.**7**, Issue.**2**, pp.**179–188**,**1936**.

[2] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[3] Elsalamony, Hany A. "Bank direct marketing analysis of data mining techniques." International Journal of Computer Applications Vol.**85**, Issue.**7**, pp.**12-22**, **2014.**

[4] Jeyapriya, A., and CS KanimozhiSelvi. "Extracting aspects and mining opinions in product reviews using a supervised learning algorithm." 2015 2nd International Conference on Electronics and Communication Systems (ICECS). **IEEE, 2015**.

[5] Sharma, Manik, Samriti Sharma, and Gurvinder Singh. "Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining" Vol.**3**, Issue.**4**, pp.**54**, **2018**.

[6] Shakoor, MdTahmid, et al. "Agricultural production output prediction using supervised machine learning techniques." 2017 1st International Conference on Next Generation Computing Applications (NextComp). IEEE, **2017**.

[7] Sadarina, P., M. Kothari, and J. Gondaliya. "Implementing data mining techniques for marketing of pharmaceutical products." International Journal of Computer Applications & Information Technology, 2.1 (2013). Vol.**3**, Issue.**4**, pp.**54**, **2018**.

[8] Breiman, L. Rastgele Ormanlar. Machine Learning 45, **5-32, 2001.**

[9] Özkan, İ. N. İ. K., & Ülker, E. (2017). Derin Öğrenme ve Görüntü Analizinde Kullanılan Derin Öğrenme Modelleri. Gaziosmanpaşa Bilimsel Araştırma Dergisi, Vol.**6**, Issue.**3**, pp.**85-104, 2017.**

[10] Mohammadi, K., Shamshirband, S., Anisi, M. H., Alam, K. A., & Petković, D. (2015). Support vector regression based prediction of global solar radiation on a horizontal surface. Energy Conversion and Management, 91, **433-441, 2015.**

[11] Zhang, L., Zhou, W. D., Chang, P. C., Yang, J. W., & Li, F. Z. (2013). Iterated time series prediction with multiple support vector regression models. Neurocomputing, 99, **411-422, 2013.**

[12] J. Gou, T. Xiong, and Y. Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule," *J. Comput.*, Vol.**6**, no.**5**, pp.**833–840, May 2011.** doi: 10.4304/jcp.6.5.833-840.

[13] Kaufman, L. and Rousseeuw, P.J.(1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the –Norm and Related Methods, edited by Y.Dodge, North-Holland, **405-416, 1987.**

[14] K R Rathy, Arya Vaishali, "Classification of Dataset using Efficient Neural Fuzzy Approach", Vol. **099**, August **2019**

[15] Wong P. H. S, Theis N. T, "A new beginning for information technology by Moore's law," Comput. Sci. Eng., Vol.**19**, no.2, pp.**41–50, 2016.**

[16] Vaishali Arya, R K Rathy, "An Efficient Neura-Fuzzy Approach For Classification of Dataset", International Conference on Reliability, Optimization and Information Technology, Feb **2014**.

[17] Asmita Shukla, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra, "Flower Classification using Supervised Learning," *Int. J. Eng. Res.*, vol. Vol.**9**, no. 05, pp.**757–762, 2020.**

[18] Hossain S, Aktar S, and Mithy SA. Solution of large-scale linear programming problem by using computer technique, *Int. J. Mat. Math. Sci*., Vol.**4**, Issue.1, pp.**15-34, 2021.**

AUTHORS PROFILE

Samme Akter Mithy was born in Patuakhali District, Galachipa Upazila, Bangladesh. She received her B.Sc. degree in Applied Mathematics from Gono Bishwabidyalay (University) and Masters in Applied Statistics and Data Science from Jahangirnagar University, Dhaka-1342, Bangladesh. She is currently working at Center for Community Health and Research, Gonoshasthaya Samaj Vittik Medical College, Dhaka 1344, Bangladesh since 2016. She has published three research papers in reputed international journals; her research interests include Linear Programming Problem, Numerical analysis Public Health, Social Sciences and Social economics. She has one year of teaching experience and five years of research experience.

Shohal Hossain was born at Ramganj Upazila, lakshmipur district, Bangladesh. He received his B.Sc. degree in Applied Mathematics from Gono Bishwabidyalay (University), and Masters in Applied Mathematics from Islamic University, Kushtia , Bangladesh. He is currently working as Research Assistant at Center for Multidisciplinary Research, Gono Bishwabidyalay (University), Dhaka-1344, Bangladesh since 2021. He has published five research papers in reputed international journals, His research interests include Linear Programming Problem, Numerical analysis, Infectious Disease, Regression analysis, Time Series Analysis, Statistical Modelling, Public Health, Social Sciences, and Social economics. He has three years of teaching experience and two years of research experience.

Shamima Aktar was born at Jamalpur District, Melandaha Upazila, Bangladesh. She received her B.S.S. and M.S.S degree in Economics from Shahjalal University of Science and Technology, Bangladesh. She is currently working as Sr. Lecturer of Economics at Gono Bishwabidyalay (University), Dhaka 1344, Bangladesh since 2011. She has published three research papers in reputed international journals, His research interests include Health Economics, Micro-Economics, Health Economics and Social economics. He has near about 12 years of teaching experience and three years of research experience.

Umme Honey was born at Jhalokathi District, Bangladesh. She received her B.Sc. and M.Sc. degree in Statistics from Jahangirnagar University, Dhaka 1342, Bangladesh. She is currently working as Asst. Lecturer of Statistics, Gono Bishwabidyalay (University), Dhaka 1344, Bangladesh since 2022. She has published two papers in reputed international journals, her research interests include Bio- statistics, Applied Statistics, Bayesian Inference, Infectious Disease, Regression analysis, Time Series Analysis, Mental Health, Clustered and Count Data Analysis, Public Health, Mixed Models, and Statistical Modelling. She has one years of teaching experience and research experience.

Sajjad Bin Sogir was born in Feni District, Sonagazi Upazila, Bangladesh. He received his B.Sc. Degree in Statistics and doing master's on Statistics from Jahangirnagar University, Dhaka-1342, Bangladesh. He has summited one research paper in reputed international journal and working on three research papers. His research interests include Machine learning, Bio-stat, Numerical analysis, Multivariate analysis, Design of experiment, Epidemiology and survival analysis, Time series, Stochastic process.

