Prediction Model based on Iris Dataset Via Some Machine Learning Algorithms

Chya Fatah Aziz
Food Science and Quality Control
Technical College of Applied Science
Sulaimani Polytechnic University
Sulaimani, Iraq
chia.aziz@spu.edu.iq
Orcid.org/0009-0005-4231-3587

Banan Jamil Awrahman
Information Technology
Halabja Technical Institute
Sulaimani Polytechnic University
Sulaimani, Iraq
banan.awrahman@spu.edu.iq

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100210

Received Mar. 20, 2023. Accepted for publication Jun. 11, 2023

Abstract— *Supervised Machine Learning algorithm has an important approach to Classification. We are predicting the deal type of the Iris plant using various algorithms of machine learning. Iris plants are determined by numerous factors such as the size of the length and width of the property. A horticultural skill announces that some of the plants are different in some physical appearances like size, shape, and color. Hence it is difficult to recognize any species. Versicolor, Setosa, and Virginica have three identical subspecies of The Iris flower species. This paper uses machine learning algorithms to recognize all classes of the flower with an accuracy degree of %100 for KNN, %95 for RF, %97 for DT, and %98 for LR. The Iris dataset is frequently available, and it is implemented using Scikit tools. and build the prediction model for Plants. Here, algorithms of machine learning such as Logistic Regression (LR), Decision Tree (DT), K Nearest Neighbor (KNN), and Random Forest (RF) are employed to construct a predictive model.*

Keywords: ***Machine Learning, Iris Flower, Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest***

I. INTRODUCTION

A most fascinating research topic is Machine Learning today [1], today many people work in the machine learning field, and some publishers or researchers discovering or updating new algorithms and methods for Machine Learning, Basically, machine learning is the process of causing the machine to make the same decisions as the human brain. machine learning which is the main part of artificial intelligence has two main categories[2]supervised learning and unsupervised learning. Hence, the learning phase is classified as Supervised learning, unsupervised learning, and Reinforcement Learning [[3], [4]]. As part of the supervised learning process, an output target is presented, which assists or makes the system in learning also it contains instances of training data that consist of different input attributes and an output. A subpart of supervised learning is Classification, where the program

learns from the input data given to it and uses this process to classify new observations. Classification techniques have various types like Decision Trees, Neural Networks, Bayes Classifiers, Support Vector Machine, K-Nearest neighbors, and many more. Here are some examples of Machine learning classification tasks with both discrete and continuous data: Classifying credit card transactions, Detection of diseases in the human body classifying protein as alpha-helix as secondary structures of beta-sheet or random coil, weather forecasts, and categorizing news stories as finance, sports, and entertainment[5]

Unsupervised learning, on the other hand, does not have a target output, but the system must learn without any guidance. The last category of learning is how external agents or individuals are involved to learn from mistakes and take action.

Python is a powerful programming language tool created by Guido van Rossum in 1989. Python is interpreted, as the object-oriented, dynamic data type of high-level programming language. Python stands out among other programming languages due to it is uncomplicated and user-friendly writing process, and it is elegant facilitates straightforward implementation while still maintaining a refined structure. Python language consists of powerful libraries like Numpy, Seaborn, and Matplotlib.pylot, Panda, and many more. Moreover, It is combined with C or C++ or Java or other programming languages easily. The sciPy library of python

utilized by Scikit-Learn as a toolkit, and it has a big collection of algorithms in machine learning. In Scikit learn library there are several steps that include loding datasets, manipulation them, and creating pre-processing of pipelines and evaluating metrics.

Two researchers are proposed an algorithm for a nonlinear classifier, in which any individual single value was not appropriate for learning tasks. they tested this method on two different data sets, Balance Scale and Iris Flower, where the decisions of class members can only be affected collectively by individual lineaments of the flower [6].

Testing the accuracies of Machine learning algorithms on the Iris dataset while varying the test data and train data split between 75-25, 70-30, and 80-20. In all three training and testing cases, SVM performed the best and KNN performed the second best and Decision Tree has the worst accuracy and worst performance [7].

Wine quality dataset has higher variability amongst the values of feature parameters we use to predict its quality which has a huge effect on the performance used algorithms. in this study, some models are used but the final SVM Classification method performed the best for this data set after greatly removing certain parameters and reducing the variability of values from consideration [8].

Researchers proposed a paper to identify and organize objects in the dataset. They use the Machine Learning algorithms such as decision trees (j48), K-nearest neighbors, and random forests, and then compare their accuracies or performances using the IRIS dataset. As a result of the comparison, analysis gained that the K-nearest neighbors have the best performance than the other algorithm classifiers which have 100% performance and no error rate. Also, the second algorithm is a random forest classifier and the last classifier is a decision tree (j48) [9].

Some researchers proposed that if you use a complex model on a simple dataset or vice versa, you might not get the best results, and this can cost you time and money. In order to demonstrate this, we use the IRIS dataset and Wine quality dataset in order to identify which type of Algorithm gives the highest accuracy for each type of dataset. Based on this research, our results show that algorithms like KNN are the most effective (95.5% accuracy) for evenly distributed and simple datasets such as the Iris dataset. For the Wine Quality dataset, algorithms like Decision Tree give a high accuracy of 100% and K-nearest neighbors give a minimum accuracy which was 82.29% [10].

Flower classes were recognized automatically based on three approaches: segmentation, feature extraction, and classification Using Logistic Regression, Neural network, K-Nearest Neighbors and Support Vector Machine [11]. Different tools and libraries were used, such as Scikit and Pandas, Numpy, etc. All these tools were used to test the dataset of iris flowers, using all these tools and a variety of methods were discussed, [12].

A Decision Tree (DT) consists of a tree-based strategy in which every path between the root and leaf nodes represents a data separation series before getting a Boolean result[13], [14]. An information relationship is depicted hierarchically by nodes and links. A node represents a use, whereas a tie is used to distinguish them [15]. Classification and regression can both be performed using DT, which is a form of ML algorithm. In a binary tree, nodes make decisions by comparing functions to thresholds and dividing the decision route according to the result. A leaf node can contain an actual value, a choice, or a class name, depending on whether the task is a grouping or regression [16], [17]. By using Random Forest (RF), trees are created at random using input vectors to estimate output vectors, the equivalent of producing a random range of weights that stays unchanged from earlier weight sequences [18].

An expert botanist's identification of some iris species [19] belonging to The Iris database is the most commonly used database for machine-learning algorithms and it has subclasses of the species Setosa, Versicolor, or Virginica used in this paper to develop a machine-learning model from measurements of some of these iris species. Based on these measurements, we can predict which species the iris belongs to. Under the supervised learning classification, a three-class classification problem comes with each type of species. Classification provides a list of possible class types, and a single iris is called a label. Various tools, packages, and libraries are used to predict iris species type. Jupiter Notebook editor with Scikit, Numpy, Seaborn, pandas, matplotlib, and Python programming tools.

II. RELATED WORK

The Machine Learning Repository at UCI contains the iris dataset. Besides Edgar Anderson's 1935 description, Ronald Fisher's 1936 generalization of the data set called Fisher's Iris was based on a variety of classification methodologies. In this way, real values and multivariate data are characteristic of data.

Using different supervised machine learning techniques proposed in their paper, by using the %80 of training data and %20 of testing data, as an average the experimental results gave an accuracy rate for the training set with 96.66 for Neural network and Logistic Regression, % 98 for Support Vector Machine, and % 96.67 for k-Nearest Neighbors of recognition rate with the accuracy average of a testing set of %100[20].

The authors proposed an evolutionary algorithm for nonlinear discriminant classifiers but noted that it wasn't suitable for learning tasks with individual values. Hence they used two datasets for testing the method, Balance Scale and Iris Flower, An individual lineament of a flower cannot affect decisions of class membership [21].

The fundamental component of Artificial Intelligence exactly Machine Learning is the prediction process. The data mining prediction process must be performed by machines using algorithms. Their study used the k-NN algorithm to estimate the classification of the Iris data set by using the Orange tool. Changing the optimum k value in k-NN method

and using the correct attribute are two key successes of the k-NN algorithm, they conclude that the k-NN method provides the classification prediction on the Iris dataset as %98.67 if you use the k value as 15 [22].

Some ML techniques was used like Adaptive Boosting (AdaBoost), Support Vector Machines (SVM), Random Forest and K-Nearest Neighbors (KNN) machine learning algorithms for predicting the Coruh river streamflow monthly[23].

Convolutional Neural Network (CNN) was used to extract and classify features of the iris dataset patterns. the experiments were made on MATLAB and a Graphics Processing Unit (GPU). The training sets accuracy was determined, with the %95.33 with a time of 17.59 minutes. also get %100 in just 12 seconds. There has been a success with the application of the iris recognition system [24].

III. METHODS AND MATERIALS

Machine learning will be used to recognize the species of iris. In general, we can use four types of algorithms in machine learning: Logistic Regression (LR), Random Forest (RF) k-Nearest neighbors (K-NN), and Decision Tree (DT) classifier.

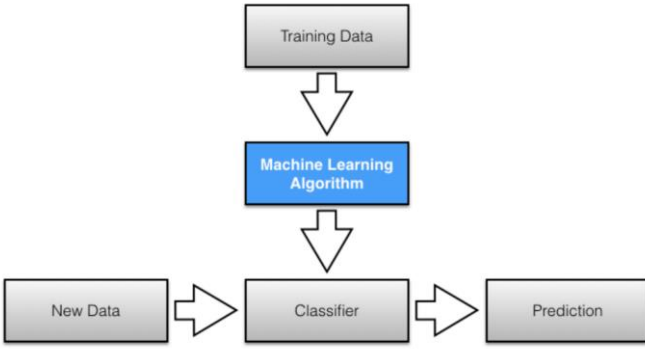


Fig. 1. The general pattern classification model is illustrated in this diagram in a simplified form [25].

A. Dataset Visualization:

The dataset for this paper comes from the scikit open-source project, name the Iris Flower data set, which is already built into it.

Figure 2 illustrated that this dataset comprises of a total of 150 samples with 50 samples of each of virginica, Setosa, and versicolor.

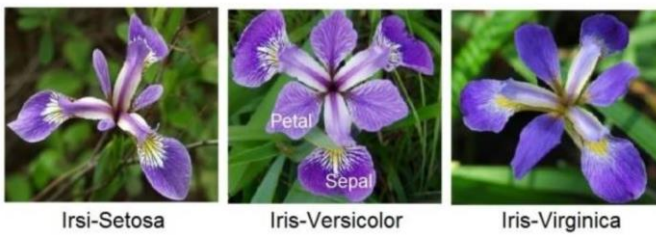


Fig. 2. Iris Flower Types

Each of the three samples has four features. also, these are the properties of each sample's sepal width, sepal length, petal width, and petal length.

Figure 3 shows the measurements for each property with non-missing data in each type of Iris flower. A flower measurement is represented by each row. With this model, we will be able to predict a new species of iris based on these measurements. As a supervised learning problem, this appears to be a classification problem.

```
1 Iris_df.head(7)
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4
6	4.6	3.4	1.4	0.3

Fig. 3. seven samples of the Iris_df dataset

Figure 3 shows that the petal width for the first five flowers is 0.2 cm, and the longest sepal for the first flower is 5.1 cm. We now place each flower in the target array according to its iris species. There is a one-dimensional array in the target array, which is a NumPy array type. A species is encoded as a number between 0 and 2. The encoded results mean that the number 0 is setosa, number 1 is versicolor, and number 2 means virginica.

In general, this paper focused on random forests, decision trees, logistic regression, and k-Nearest Neighbor algorithms.

B. Data Processing

Our model can be developed to predict the iris flower species based on new measurements. But before applying this model to new measurements, let's see if it works. The available data can only predict the correct target based on existing measurements. In other words, it doesn't work well with new data. Split the data into two parts for building the model and estimating performance. One is the training set or training data and the other is the test set or test data. It's a good idea to visually inspect the data before building the model. The training sets actual characteristics are shown in Figure 4 through a pair plot, data points are color-coded according to the species to which the iris belongs.

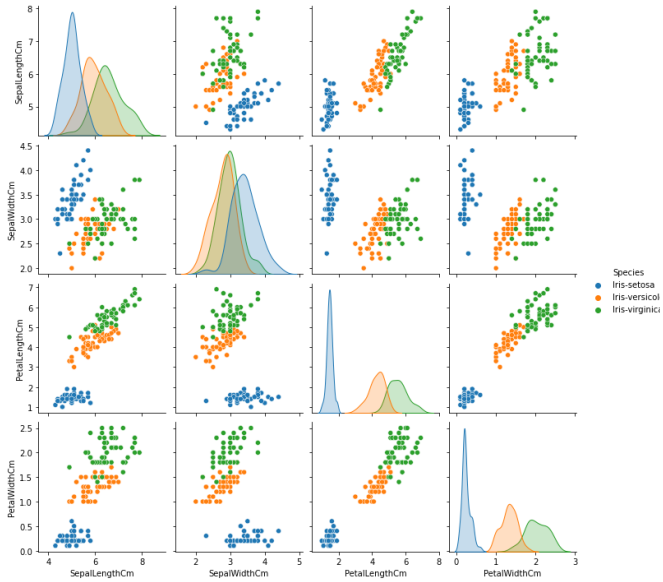


Fig. 4. Pair Plot of Iris data set real features

A. *Decision Tree Classifier*

One widely used technique in data mining is the system of creating classifiers. DT is a previously used text and data mining classification algorithm. Decision tree classifiers (DTCs) have proven effective in a variety of classification applications [26]. train_test_split() function is in scikit-learn . This function extracts %27 as his test data and %73 of the data as training data. Figure 5 below shows the division of train and test data.

```
X = iris.drop(['species'],axis=1)
y = iris['species']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.27, random_state=25)
```

Fig. 5. train test function for DT algorithm.

After extracting the train-test-split function and fitting the method we prove that the accuracy of the Decision Tree (DT) algorithm gets %98 as shown in figure 6 with the confusion matrix array.

```
#score
print(DT.score(X_test,y_test))
```

```
0.975609756097561
```

```
#confusion_matrix
print(confusion_matrix(y_test,predictions))
```

```
[[14  0  0]
 [ 0 15  1]
 [ 0  0 11]]
```

Fig. 6. score and confusion matrix function for DT algorithm

C. *Random Forest Classifier*

After training the Random Forest (RF) model we are doing the predictions and evaluation for test data finally we prove that by making (600) trees randomly we obtain the %95 accuracy of the DT classification method which you can see in figure 7.

```
#score
print(rfc.score(X_test,y_test))
```

```
0.9512195121951219
```

```
#confusion_matrix
print(confusion_matrix(y_test,predictions))
```

```
[[14  0  0]
 [ 0 15  1]
 [ 0  1 10]]
```

Fig. 7. score and confusion matrix function for RF algorithm

And also, you can see the heatmap of the confusion matrix method for Random Forest and Decision Tree algorithms of machine learnings that they have the same result in this case of classification problem, that is shown in figure 8.

```
sns.heatmap(confusion_matrix(y_test, predictions), annot=True, cmap='Reds')
```

```
<AxesSubplot:>
```

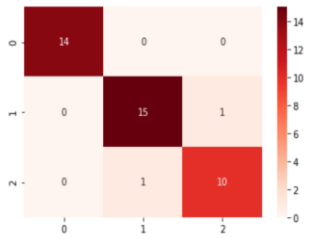


Fig. 8. Heatmap of confusion matrix function for RF algorithm

D. *Logistic Regression*

A common method in data mining for constructing a classifier through a classification algorithm is Logistic Regression (LR) algorithm. Scikit-learn includes the functions of train_test_split. As a part of the test data 27% is extracted by this function and 73% of the data as training data. figure 9 below shows the division of train and test data.

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.27, random_state=25)
```

```
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
```

Fig. 9. train test function for Logistic Regression algorithm

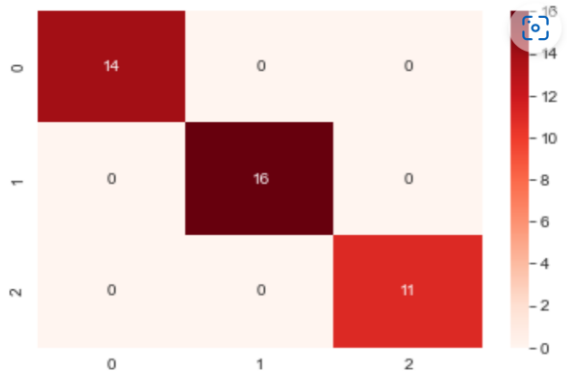
Here we are putting different percentages between the train data and test data according to the DT and RF method just because we calculate and obtained %97 Accuracy and classified the iris flower types.

E. *K-Nearest Neighbors Classifier*

The K-NN approach is non-parametric and can be used both for regression and classification. Guided learning uses K-NN methods as one of its methods. A data point is identified by computing its k nearest neighbors based on the fundamental idea behind this approach. The gap between the test data and the feedback must be calculated and then

correctly predicted. the point is the most common class assigned to those neighbors of k. Different metrics can be used to measure the disparity between two samples. In equation 1, we can see how to calculate Euclidean distance vectors.

Xi shows the i'th sample value, which is entered from the outside and whose class will be determined. In a data set, each specific class represents by Yi. In this case, n represents the number of attributes. Euclidean distance(X, Y), gives distance values between Yi and Xi [27], and It is an extremely common metric. In the training set, each unknown instance is classified according to the K Nearest Neighbor rule. The heatmap of the K-NN confusion matrix is shown in below Figure 10.



No.	Dataset	Machine Learning Algorithms	Accuracy Predictions	Test Data and Training data Average
1	Iris Flower Dataset	Logistic Regression (LR)	%98	%27 Test data %73 Training data
2		K Nearest Neighbor (KNN)	%100	
3		Decision Tree (DT)	%97	
4		Random Forest (RF)	%95	%27 Test data %73 Training data And Creating 600 Tree randomly

Fig. 10. Heatmap of confusion matrix K-NN

work, calculating the distance between the training set and the test set. After that, list all the determined distances in ascending order, and select an appropriate distance for our work. Finally, the most unique class is the instance class of choosing the first five teaching instances in a list ordered

$$\text{Euclidean distance}(X,Y) = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2} \quad (1)$$

(k=15), and the weighted average accuracy is 100%.

IV. RESULT

The dataset is obtained from the scikit open-source project, this data set contains three classes of Virginica, Setosa, and Versicolor participating in the same five features:
-Sepal Length and Width.
-Pital Length and Width
-Species
The experiments are created on Windows 11 platform using the Python Jupiter Notebook tool.

TABLE.1. DESCRIPTION OF INPUT PARAMETERS

Dataset Name	No. of Instances	No. of columns	Missing values
Iris-dataset	150	5	None

This data set as shown in table (1) has no missing value and distinguishes the dependent and independent dataset features then the mentioned machine learning methods have been used and the result are obtained and given the performance as shown in below table:

TABLE .2. EXPLAINING MACHINE LEARNING METHODS WITH THEIR PERFORMANCES WITH THE IRIS DATASET

Figure 11 shows the graph of the distribution of accuracy predictions percentage for the dataset features based on the given training set and testing set.

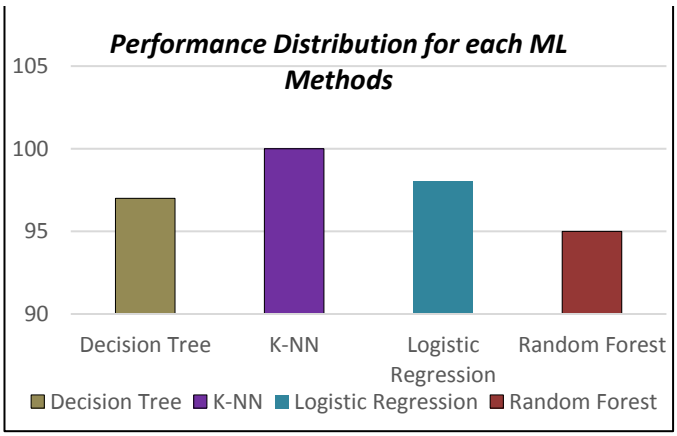


Fig. 11. Performance distribution for each ML method

V. CONCLUSION

Today, the most widely used method in machine learning algorithms and data mining is classification, and it has more applications and uses such as face recognition,

flower classification, clustering, and so on. We built a model using some different algorithms of Machine Learning techniques to predict Iris flower types. Our accuracy was %100 for KNN when the chosen neighbors of k value were equal to 15, %95 for RF, %97 for DT, and %98 for LR, which means that we successfully prepared and managed the sizes of the test set and training set for the dataset to make a good prediction for all three Iris flower types.

REFERENCE

[1] J. Cutler and M. Dickenson, *Computational Frameworks for Political and Social Research with Python*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-36826-5.

[2] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *Applied Sciences (Switzerland)*, vol. 9, no. 20. MDPI AG, Oct. 01, 2019. doi: 10.3390/app9204396.

[3] A. ÇELİK, "Improving Iris Dataset Classification Prediction Achievement By Using Optimum k Value of kNN Algorithm," *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, Mar. 2022, doi: 10.53608/estudambilisim.1071335.

[4] "Machine Learning Classifiers Based Classification For IRIS Recognition", doi: 10.48161/Issn.2709-8206.

[5] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Syst Appl*, vol. 42, no. 5, pp. 2670–2679, Apr. 2015, doi: 10.1016/j.eswa.2014.11.009.

[6] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory," Oct. 2014, [Online]. Available: http://arxiv.org/abs/1410.5329

[7] B. K. O. C. Alwawi and A. F. Y. Althabhawee, "Towards more accurate and efficient human iris recognition model using deep learning technology," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 4, pp. 817–824, Aug. 2022, doi: 10.12928/TELKOMNIKA.v20i4.23759.

[8] A. H. Reddy Kohir, A. Shukla, A. Agarwal, M. Lucknow, I. H. Pant, and P. Mishra, "IJERT-Flower Classification using Supervised Learning Flower Classification using Supervised Learning." [Online]. Available: www.ijert.org

[9] A. ÇELİK, "Improving Iris Dataset Classification Prediction Achievement By Using Optimum k Value of kNN Algorithm," *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, Mar. 2022, doi: 10.53608/estudambilisim.1071335.

[10] F. TOSUNOĞLU, S. HANAY, E. ÇİNTAŞ, and B. ÖZYER, "Makine Öğrenimi Kullanarak Aylık Akarsu Akışı Tahmini," *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 13, no. 3, Dec. 2020, doi: 10.18185/erzifbed.780477.

[11] M. Czajkowski and M. Kretowski, "Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach," *Expert Syst Appl*, vol. 137, pp. 392–404, Dec. 2019, doi: 10.1016/j.eswa.2019.07.019.

[12] Z. Ursani and D. W. Corne, "A novel nonlinear discriminant classifier trained by an evolutionary algorithm," in *ACM International Conference Proceeding Series*, Feb. 2018, pp. 336–340. doi: 10.1145/3195106.3195132.

[13] A. Mohsin Abdulazeez, D. Zeebaree, D. M. Abdulqader, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review Deep Learning View project Gait recognition with wavelet transform View project Machine Learning Supervised Algorithms of Gene Selection: A Review," 2020. [Online]. Available: https://www.researchgate.net/publication/341119469

[14] G. L. Bajaj, M. Syamala Devi, A. Guleria, K. Rai, and M. Syamala Devi Professor, "Decision Tree Based Algorithm for Intrusion Detection Scheduling in VANETs View project Multiagent Integrated scheme for Intrusion Detection View project Kajal Rai Decision Tree Based Algorithm for Intrusion Detection", [Online]. Available: https://www.researchgate.net/publication/298175900

[15] A. S. Eesa, A. M. Abdulazeez, and Z. Orman, "A DIDS Based on The Combination of Cuttlefish Algorithm and Decision Tree," *Science Journal of University of Zakho*, vol. 5, no. 4, p. 313, Dec. 2017, doi: 10.25271/2017.5.4.382.

[16] N. Mahdi Abdulkareem and A. Mohsin Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," 2021, doi: 10.5281/zenodo.4471118.

[17] P. Grewal, P. Sharma, A. Rathee, and S. Gupta, "EPRA International Journal of Research and Development (IJRD) COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS", doi: 10.36713/epra2016.

[18] P. Prathima and R. Kumar, "COMPARISON ON IRIS DATASET USING CLASSIFICATION TECHNIQUES," JETIR, 2021. [Online]. Available: www.jetir.org

[19] K. Thirunavukkarasu, A. S. Singh, P. Rai, and S. Gupta, "Classification of IRIS dataset using classification based KNN Algorithm in supervised learning," in *2018 4th International Conference on Computing Communication and Automation, ICCCA 2018*, Dec. 2018. doi: 10.1109/CCAA.2018.8777643.

[20] A. H. Reddy Kohir, A. Shukla, A. Agarwal, M. Lucknow, I. H. Pant, and P. Mishra, "IJERT-Flower Classification using Supervised Learning Flower Classification using Supervised Learning." [Online]. Available: www.ijert.org

[21] "Machine Learning Classifiers Based Classification For IRIS Recognition", doi: 10.48161/Issn.2709-8206.

[22] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," in *Procedia Computer Science*, 2018, vol. 125, pp. 305–312. doi: 10.1016/j.procs.2017.12.041.

[23] A. Shukla, A. Agarwal, M. Lucknow, I. H. Pant, and P. Mishra, "Flower Classification using Supervised Learning." [Online]. Available: www.ijert.org

[24] IEEE Computational Intelligence Society. and Institute of Electrical and Electronics Engineers, *IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI2017) : November 23-24, 2017, Republic of Mauritius.*

[25] M. A. Turk and A. P. Pentland, "Face Recognition Using Eigenfaces."

[26] S. Halakatti, "Traffic Sign Symbol Recognition Using Single Dimension PCA." [Online]. Available: www.computerscijournal.org

[27] E. Alpaydin, *Introduction to machine learning*.