iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** I **Month of publication:** January 2022

DOI: https://doi.org/10.22214/ijraset.2022.40097

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

IRIS Species Predictor

D. Rajkumar[1], Dr. J. Sreerambabu[2], S. Kalidasan[3]
[1]Assistant Professor, [2]Head of the Department, [3]Assistant Professor, Master of Computer Applications Department, Thanthai Periyar Government. Institute of Technology, Vellore-2

Abstract: In Machine Learning, we are using semi-automated extraction of knowledge of data for identifying IRIS flower species. Classification is a supervised learning in which the response is categorical that is its values are in finite unordered set. To simply the problem of classification, scikit learn tools have been used. This paper focuses on IRIS flower classification using Machine Learning with scikit tools. Here the problem concerns the identification of IRIS flower species on the basis of flowers attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to from the class of IRIS flower. In this paper we train the machine learning model with data and when unseen data is discovered the predictive model predicts the species using what it has been learnt from the trained data.
Keywords: MATLAB, Machine learning, Neural Network.

I. INTRODUCTION

The Machine Learning is the subfield of computer science, according to Arthur Samuel in 1959 told “computers are having the ability to learn without being explicitly programmed”. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicitly algorithms with good performance is difficult or unfeasible; example applications include email filtering, detection of network intruders, learning to rank and computer vision.

Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. It is a research field at the intersection of statistics, artificial intelligence and computer science and is also known as predictive analytics or statistical learning. There are two main categories of Machine learning. They are Supervised and Unsupervised learning and here in this, the paper focuses on supervised learning. Supervised learning is a task of inferring a function from labeld training data. The training data consists of set of training examples. In supervised learning, each example is a pair of an input object and desired output value. A supervised learning algorithm analyze the training data and produces an inferred function. Supervised learning problems can be further grouped into regression and classification problems. Classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

Regression problem is when the output variable is a real value, such as “dollars” or “weight”. In this paper a novel method for Identification of Iris flower species is presented. It works in two phases, namely training and testing. During training the training dataset are loaded into Machine Learning Model and Labels are assigned. Further the predictive model, predicts to which species the Iris flower belongs to. Hence, the expected Iris species is labeled. This project focuses on IRIS flower classification using Machine Learning with sci-kit tools. The problem statement concerns the identification of IRIS flower species on the basic of flower attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS flower and how the prediction

II. EXISTING SYSTEM

Many methods have been presented for Identification of Iris Flower Species. Every method employs different strategy. Review of some prominent solutions is presented. The methodology for Iris Flower Species System is described . In this work, IRIS flower classification using Neural Network. The problem concerns the identification of IRIS flower species on the basis of flower attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS flower and how the prediction was made from analyzing the pattern to form the class of IRIS flower. By using this pattern and classification, in future upcoming years the unknown data can be predicted more precisely. Artificial neural network shave been successfully applied to problems in pattern classification, function approximations, optimization, and associative memories. In this work, Multilayer feed-forward networks are trained using back propagation learning algorithm .

The model for Iris Flower Species System is described . Existing iris flower dataset is preloaded in MATLAB and is used for clustering into three different species. The dataset is clustered using the k-means algorithm and neural network clustering tool in MATLAB. Neural network clustering tool is mainly used for clustering large data set without any supervision. It is also used for pattern recognition, feature extraction, vector quantization, image segmentation, function approximation, and data mining. Results/Findings: The results include the clustered iris dataset into three species without any supervision. The model for Iris Flower Species System is described . The proposed method is applied on Iris data sets and classifies the dataset into four classes. In this case, the network could select the good features and extract a small but adequate set of rules for the classification task. For Class one data set we obtained zero misclassification on test sets and for all other data sets the results obtained are comparable to the results reported in the literature

A. Drawbacks of the Existing System

1) Existing system face several difficulties like computational power is increase when run deep learning.
2) It requires large amount of data.
3) In existing system, IRIS flower classification done by neural network.
4) Neural network contain many clusters of data so it accuracy is less.
5) The dataset is clustered using the K-means algorithm.
6) And the IRIS flower dataset preloaded in MATLAB.
7) It is very expensive.

B. Proposed System

It is observed from the literature survey that the existing algorithms face several difficulties like the computational power is increases when run Deep Learning on latest computation, requires a large amount of data, is extremely computationally expensive to train, they do not have explanatory power that is they may extract the best signals to accurately classify and cluster data, but cannot get how they reached a certain conclusion. Neural Networks cannot be retrained that is it is impossible to add data later.

To address these problems the current work is taken up to develop a new technique for Identification of Iris Flower Species using Machine Learning. The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson’s Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris Flower of three related species.

Two of the three species were collected in Gaspe Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with same apparatus. The data set consists of 50 samples from each of three species of Iris that is 1) Iris Setosa 2) Iris Virginica 3) Iris Versicolor. Four features were measured from each sample. They are 1) Sepal Length 2) Sepal Width 3) Petal Length 4) Petal Width. All these four parameters are measured in Centimeters. Based on the combination of these four features, the species among three can be predicted.

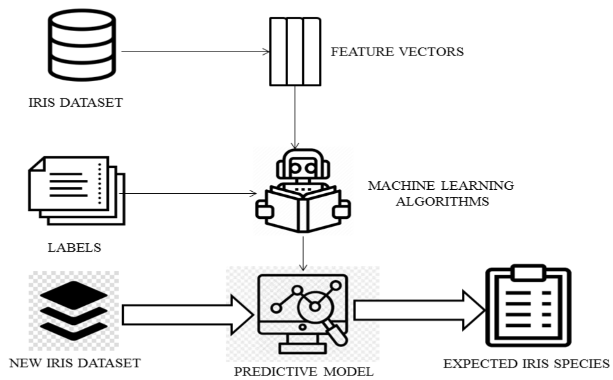
Various datasets of Iris Flower are collected. There are totally 150 datasets belonging to three different species of Iris Flower that is Setosa, Versicolor and Virginca. The collected Iris Datasets are loaded into the Machine Learning Model. Scikit-learn comes with a few standard datasets, for instance the Iris Dataset for Classification. The load_iris function is imported from Scikit-learn. The load_iris function is run and save the return value in an object called “iris”. The iris object is of type “sklearn.datasets.base.bunch”, here bunch is a special object in scikit-learn to store datasets and attributes. The few attributes of iris object are data, feature names, target, target names etc.

C. Advantages of Proposed System

1) In proposed system various datasets of IRIS flower are collected
2) There are totally 150 datasets belonging to three different species of IRIS flower that is Setosa, Versicolor, Virginca.
3) Four features were measured from each sample(sepal length, sepal width, petal length, petal width).
4) Using scikit learn we can store the attributes.

D. ML Model Training

In this module, prepared IRIS flower data are trained by particular machine learning algorithms (i.e)Logistic regression, K-Nearest Neighbour algorithm. The collected IRIS flower data are modelled and trained by using these algorithms. In this module we are already inserted the 150 IRIS flower datasets for model training.



E. Image Processing

One of themost successfulapplicationsof imageanalysisandunderstanding,face recognition hasrecentlyreceived a significantattention, especially during thepast few years. In addition to this, theproblem of machinerecognitionof human facescontinues to attract researchersfrom disciplines such as imageprocessing, patternrecognition, neural networks, computervision, computergraphics and psychology. The strongneed for user-friendlysystems that can secureour assets and protect our privacy without losing our identityin a sea of numbers isobvious. We as humansuse faces to recognizeand identify our friends andfamily. Computerscan now also identify peopleautomatically using stored information such asfigure, iris or faceto identify a particularperson. Earlier manyface recognitionalgorithms were used to achievefully automated faceidentification process. The first face recognitionsystem was createdin ther1960s.

It was not fullyautomated and it required animal inputs of the locationof theeyes, ears, noseand mouth on the images then itcalculates a distance to some common point thenit compares it to the stored data. Thestill image problemhas severalinherent advantages anddisadvantages. For applicationssuch as driver’s license, due to the controlled nature of the imageacquisition process, the segmentationproblem is rather easy. However, if only astatic picture of an airportscene is available, automatic location and segmentation of a face could poseserious challenges to any segmentationalgorithm.

On the other hand, if a video sequenceis available, segmentation of a movingperson can be more easilyaccomplished using motionas a cue. But the small sizeand low image qualityof faces captured from video cansignificantly increasethe difficulty in recognition. Face recognitionand sometime is called faceidentifying is simply putting a label to known faces just like human asmentioned above, we learn the facesof our family andcelebrities just by lookingat their faces. Since the 1970s there was manytechniques andalgorithms were developed for amachine to learn torecognizeknown faces. Most of the recent techniquesinvolve at least threesteps:

1) Human detection
2) Human Image reprocessing
3) Human recognition

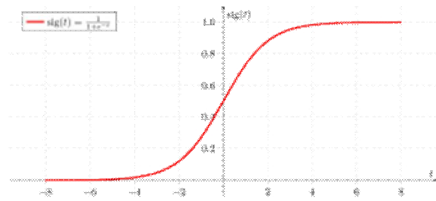
III. LOGISTIC REGRESSION ALGORITHM

Logistic Regression is a type of regression that predicts the probability of occurrence of an event by fitting data to a logit function (logistic function). Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. For instance, the probability that a person has a heart attack within a specified time period might be predicted from knowledge of the person's age, sex and body mass index. This regression is quite used in several scenarios such as prediction of customer's propensity to purchase a product or cease a subscription in marketing applications and many others.

Logistic Regression uses the logistic function to find a model that fits with the data points. The function gives a 'S' shaped curve to model the data. The curve is restricted between 0 and 1, so it is easy to apply when y is binary. Logistic Regression can then model events better than linear regression, as it shows the probability for y being 1 for a given x value.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

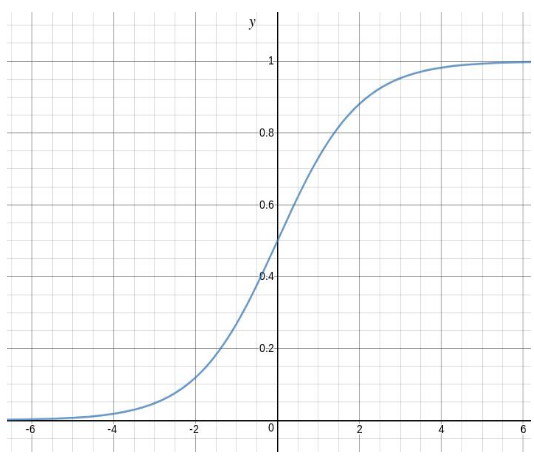
Logistic Regression is used in statistics and machine learning to predict values of an input from previous test data. A mesh when drawn over the plot shows the three classes of the logistic regression. Supervised learning consists in learning the link between two datasets: the observed data X and an external variable y that we are trying to predict, usually called “target” or “labels”. Most often, y is a 1D array of length n_samples. All supervised estimators in scikitlearn implement a fit(X, y) method to fit the model and a predict(X) method that, given unlabeled observations X, returns the predicted labels y.
Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.
Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.



The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case. In case of a Precision-Recall trade-off we use the following arguments to decide upon the threshold:-

A. Low Precision/High Recall

In applications where we want to reduce the number of false negatives without necessarily reducing the number false positives, we choose a decision value which has a low value of Precision or high value of Recall. For example, in a cancer diagnosis application, we do not want any affected patient to be classified as not affected without giving much heed to if the patient is being wrong.



We can infer from above graph that:

1) g(z) tends towards 1 as
2) g(z) tends towards 0 as
3) g(z) is always bounded between 0 and 1

IV. KNN ALGORITHM

K-Nearest Neighbors Algorithm The k-Nearest Neighbors algorithm (or kNN for short) is an easy algorithm to understand and to implement, and a powerful tool to have at your disposal. The implementation will be specific for classification problems and will be demonstrated using the Iris flowers classification problem. The model for kNN is the entire training dataset.
When a prediction is required for a unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances.

The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other types of data such as categorical or binary data, Hamming distance can be used. In the case of regression problems, the average of the predicted attribute may be returned. In the case of classification, the most prevalent class may be returned.

The kNN algorithm is belongs to the family of instance-based, competitive learning and lazy learning algorithms. Instance-based algorithms are those algorithms that model the problem using data instances (or rows) in order to make predictive decisions. The kNN algorithm is an extreme form of instance-based methods because all training observations are retained as part of the model. It is a competitive learning algorithm, because it internally uses competition between model elements (data instances) in order to make a predictive decision. The objective similarity measure between data instances causes each data instance to compete to "win" or be most similar to a given unseen data instance and contribute to a prediction. Lazy learning refers to the fact that the algorithm does not build a model until the time that a prediction is required. It is lazy because it only does work at the last second. This has the benefit of only including data relevant to the unseen data, called a localized model. A disadvantage is that it can be computationally expensive to repeat the same or similar searches over larger training datasets. Finally, kNN is powerful because it does not assume anything about the data, other than a distance measure can be calculated consistently between any two instances. As such, it is called non-parametric or non-linear as it does not assume a functional form.

V. CLASSIFICATION

In this module trained IRIS flower data are classified.In this project IRIS flower species is predicted by image preprocessing method. Inserted IRIS flower datasets are modelled and trained by algorithm. And we finally predict the IRIS flower species by these datasets. The predict method is used on the KNeighbours Classifier Class object and Logistic Regression Class object and pass the features of Unknown iris as a Python list. Actually, expects numpy array but it still works with list since numpy automatically converts it to an array of appropriate shape. The predict method returns a object of type numpy array with predicted response value. The model can predict the species for multiple observations at once.

VI. SUPERVISED LEARNING

It is the learning where the value or result that we want to predict is within the training data (labeled data) and the value which is in data that we want to study is all the other columns in the dataset are known as the Feature or Predictor Variable or Independent Variable.

A. Supervised Learning is Classified into Two Categories

1) *Clarification:* Here our target variable consists of the categories.
2) *Regression:* Here our target variable is continuous and we usually try to find out the line of the curve.

As we have understood that to carry out supervised learning we need labeled data. How we can get labeled data? There are various ways to get labeled data:

a) Historical labeled Data
b) Experiment to get data: We can perform experiments to generate labeled data like A/B Testing.
c) Crowd-sourcing

Now it's time to understand algorithms that can be used to solve supervised machine learning problem. In this post, we will be using popular **scikit-learn** package.

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images

VII.CONCLUSION

The primary goal of supervised learning is to build a model that "generalizes" .Here in this project we make predictions on unseen data which is the data not used to train the model hence the machine learning model built should accurately predicts the species of future flowers rather than accurately predicting the label of already trained data.

REFERENCES

[1] https://www.python.org/
[2] http://deeplearning.net/software/theano/
[3] http://scikit-learn.org/stable/
[4] https://en.wikipedia.org/machinelearningwiki/
[5] DiptamDutta, Argha Roy, KaustavChoudhury, "Training Aritificial Neural Network Using Particle Swarm Optimization Algorithm", International Journal on Computer Science And Engineering(IJCSE), Volume 3, Issue 3, March 2013.
[6] Poojitha V, Shilpi Jain, "A Collecation of IRIS Flower Using Neural Network CLusterimg tool in MATLAB", International Journal on Computer Science And Engineering(IJCSE).
[7] VaishaliArya, R K Rathy, "An Efficient Neura-Fuzzy Approach For Classification of Dataset", International Conference on Reliability, Optimization and Information Technology, Feb 2014.