Iris Flower Classification
ML PROJECT 2020
RISHABH, SITARAM, KANISHK
2018255, 2018197, 2018044

Anonymous CVPR submission

Paper ID ****

1. KEYWORDS

IRIS dataset, KNN,SVM,Decision Tree,Logistic Regression

Abstract

The classification is an ML technique used to predict group membership for data instances. Simplify the problem of classification, and a neural network introduces. This paper focuses on IRIS plant classification using Neural Network. The issue concerns the identification of IRIS plant species based on plant attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS plant and how the prediction made from analyzing the way to form the class of IRIS plant

1.1. Motivation

Can we actually classify iris flower with machine learning? Since now a days there are huge data sets and since machine learning is very popular in classify or predicting things we should now apply many machine learning techniques and algorithm on iris flowers to classify each of them using their sepal petals. As we all know from the nature, most of creatures have the ability to recognize the objects in order to identify food or danger. Human beings can also recognize the types and application of objects. An interesting phenomenon could be that machines could recognize objects just like us someday in the future. This project mainly focuses on machine learning in pattern recognition applications.

1.2. Related work

LITERATURE REVIEW- Paper 1

Objectives: The main objective of this paper is supervised learning. 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.

Model Selection: This paper has used two models.

K-NN: This paper has used 2 different values of N(N=1 and 5). Accuracy for K-NN without train-test split method are 96.67 per and 100 per respectively. Accuracy for K-NN with train-test split method are 95 per and 96.67 per respectively.

Logistic Regression: Accuracy without train-test split method is 96 per and with train-test split method is 95 per.

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.

LITERATURE REVIEW-Paper 2

Objective: The objective of their methodology is to choose the best classification model which performs well on iris flower species identification.

Model Selection: In this step we have imported the model, based on which species prediction should be done.

Support Vector Machine(SVM): SVC() function of sklearn is used to import the model. Predictions are made on the input test set. After comparing the actual response values with predicted response values, we got the SVM model accuracy as 96 per.

Logistic Regression:

LogisticRegression() function from sklearn.linear model is used here to import the model. After examining the Iris dataset on the basis of logistic regression we got the model accuracy as 91 per.

K-NN Classifier: They have used KneighborsClassifier(n neighbors=3) function from sklearn.neighbors package. Based on the predictions, we got the model accuracy as 93 per.

Conclusion: With the help of three methods, we have found that SVM classification method is more effective than the other methods. This paper also mentions the important functions of scikit- learn software tool that are applied to learn machine learning.

1.3. DATASET DESCRIPTION

DATASET DESCRIPTION		
S . NO	Attribute	Description
1	Sepal Length	length of sepal in cm
2	Sepal Width	width of sepal in cm
3	Petal Length	length of petal in cm
4	Petal Width	width of petal in cm
5	Species	species which the flower belongs

Visualization

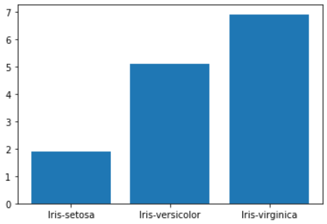


Figure 1. Species vs petal length

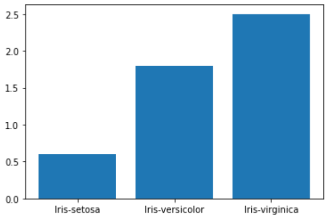


Figure 2. Species vs petal width

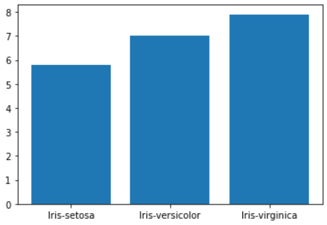


Figure 3. Species vs sepal length

Correlation matrix

After seeing the correlation matrix, We find that petal features are highly correlated.

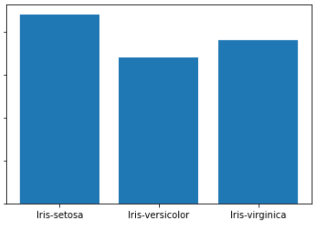


Figure 4. Species vs sepal width

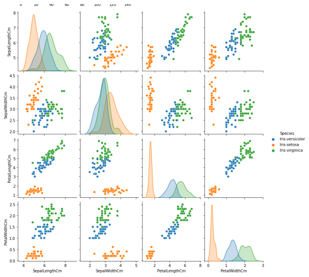


Figure 5. Visualization

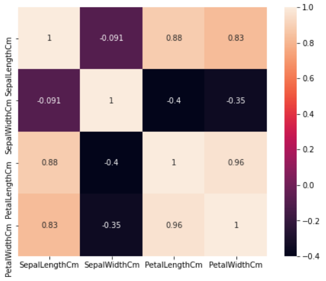


Figure 6. Visualization

Preprocessing:

1. We have removed id attribute

2. Relabeled the output as (Setosa='0', Versicolor='1', Virginica='2')

3. After checking the correlation matrix, and seeing the above visualisation we find that petal features are highly correlated.

4. We use standard scaling and stratified sampling in logistic regression to avoid outliers and overfitting. Used k-cross validation in logistic model.

5. We use only petal features and all features as our data to see that if there is difference in accuracy.

There will be no extra page charges for CVPR 2020.

1.4. METHODOLOGY

Logistic Regression:

Train:Test=75:25

We used K-cross validation with k=5 to find the average accuracy of this model.

Decision Tree:

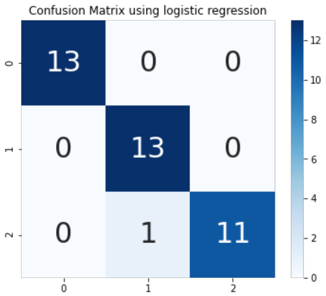
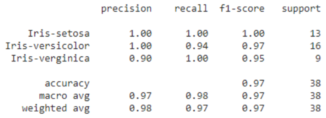


Figure 7. confusion matrix using log. regr.

Train:Test=75:25 **features importance:** [0. 0. 0.42232109 0.57767891]

Classification report



```
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00        13
Iris-versicolor       1.00      0.94      0.97        16
 Iris-verginica       0.90      1.00      0.95         9

       accuracy                           0.97        38
      macro avg       0.97      0.98      0.97        38
   weighted avg       0.98      0.97      0.97        38
```

Figure 8. Classification report

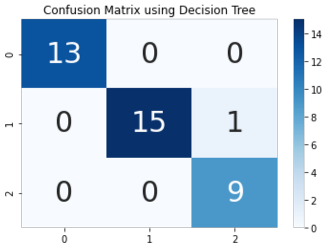


Figure 9. confusion matrix using DT

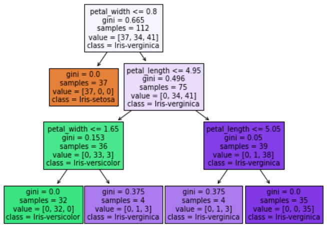


Figure 10. DT

Naive Bayes:

Train:Test=60:40

We have applied Naive bayes although we know Naive bayes assumes independent features but in our case from the results, petals are highly correlated. We want to know how robust is naive bayes assumption is?

Classification report

Random Forest:

Train:Test=60:40

features importance:

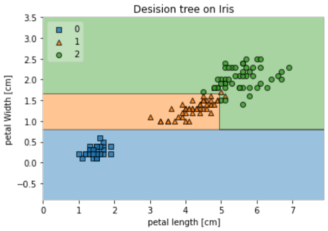


Figure 11. DT

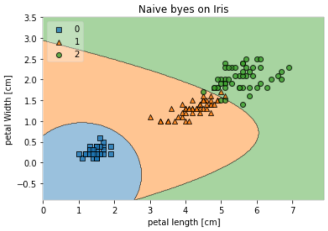


Figure 12. Naive Bayes

[0.09509096 0.01489166 0.42963573 0.46038165]

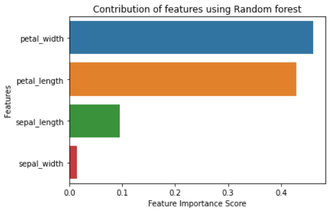


Figure 13. RF

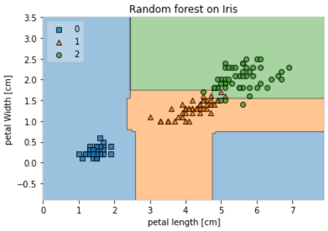


Figure 14. RF

KNN:

Train:Test=60:40

We have performed 10-fold cross validation. The optimal number of neighbors are: 1

There is no difference in accuracy if we take only petals and all features.

SVM:

Train:Test=60:40

We have implemented SVC using initial regularization parameter C = 10 and randomstate = 0.

There is no difference in accuracy if we take only petals and all features.

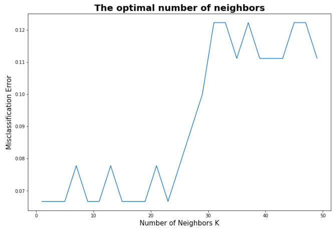


Figure 15. RF

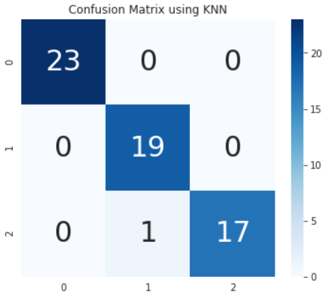


Figure 16. RF

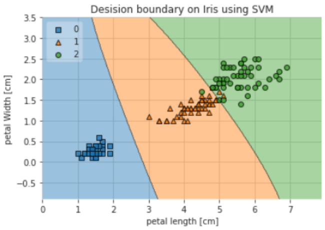


Figure 17. RF

1.5. Result

Accuracy obtained in Logistic regression is 97.34

Accuracy obtained in Decision Tree is 97.37

Accuracy obtained in Naive Bayes for all features is 96.67

Accuracy obtained in Naive Bayes for only petal features is 98.33

Accuracy obtained in Random Forest is 98.33

Accuracy obtained in KNN is 98.33

Accuracy obtained in SVM is 100

1.6. CONTRIBUTION

Sitaram: Project Report, Logistic Regression,Naive bayes, Random forest, Data visualization, SVM Literature review

Rishabh: Project Report, Literature review,Data preprocessing, Naive bayes, Decision Tree ,KNN,Balanced Random Forest,SVM

Kanishk: Project Report, Random forest, Data preprocessing, Literature review, Logistic Regression,KNN,Decision Tree

learning from the project

Rishabh:- From this project, I learned how to handle single and multi-class and handle the redundant data in a dataset. Since our project is on classification, I knew all the classification techniques and how to implement it. All the preprocessing work and everything I learned very quickly.

Sitaram: -How to apply different ML models that we learned from our lecture to a dataset and see their pros and cons on that particular dataset. If we know the Models' behavior applied to a given dataset, we can improve the accuracy by using the appropriate model on the given problem.

Kanishk:- Learnt how to preprocess on a large dataset and different types of preprocessing techniques. Also learnt some new techniques in random forest. Results showed that random forest can give good results without scaling.

References

https://towardsdatascience.com/exploring-classifiers-with-python-scikit-learn-iri https://ipasj.org/IIJCS/Volume5Issue8/IIJCS-2017-08-18-18.pdf https://www.ijream.org/papers/SSJ2019017.pdf http://airccse.org/journal/ijsc/papers/2112ijsc07.pdf