# Classification of IRIS Dataset using Weka

Kalpana Sharma, SD College, Rajhasthan

**Abstract:**

IRIS is an open access flower based dataset and is normally available on UCI dataset. The major objective of this research work is to examine the IRIS data using data mining techniques available supported in WEKA. In this work, four different classifier viz. Bayes Network Classifier, J48, Random Forest and OneR has been succefully used to classify the IRIS dataset. The dataset consist of five different attributes viz. sepallength, sepalwidth, petallength, petalwidth and class. The number of instaces in 150. It has been observed that the rate of correctly classified instances using J48 is better than bayes network, random forest and oneR classifier. The use of J48 assist us in getting 96% of accuracy. Whereas, the mininmum rate of classification achieved is with bayes network classifier.

**Keywords**: Data Mining, Classfiers, IRIS data set and Kappa statistics.

## 1. Introduction

Data mining is an important area for computer sciencists and researchers. Nowadays, there is no problem of data. however, the main problem lies in extracting meaningful information from the large volume of data. data mining techniques assists in mining large volume information and converting data into meaningful information so that the data can be classified, grouped or past and future prediction can be made[1][2]. In last few years, lot of research work has been done using differet data mining techniques in the area of agriculture[3][4][5], business & marketing [6][7], medical science [8-15], stock market[16][17] and pharmaceutical products [18][19]. The root of data minig techniques lie in three different subjects viz. Statistics, Artificial Intelligence and Machine Learning. Several heuristics have been projected to perk up the competence of the data mining process. As stated earier, clustering, association mining and prediction are four major tasks of data mining technique.

## 2. Literature Review

Data mining is playing significant role in the current days. In general, data mining techniques can be described into different categories known as classification, clustering, association, regression and prediction. These different techniques have been successfully used in different area viz. agriculture, health science, business, fincance, engineering, weather forecasting etc. It has been found that different researchers have used different classification clustering, association and predictive techniques for mining their massive data of different domain. In agriculture, people work on finding the relationship between spray and food/vegetables, Prediction of problematic wine fermentation, plants disease diagnosis, optimizing pesticides etc [18][19][20]. In medical science, different classification and clustering techniques have been used to diagnose different human diseases like diabetes, cardio, stroke, stress, cancer etc. [21][22][23]. Moreover, some of the people have also find the association between medicine and health of the person. In-text processing, opinion mining, web mining and sentiment analysis are on the top list[24][25].

## 3. Methods and Results

**IRIS is flower based multivariate dataset.** This is perhaps the best known database to be found in the pattern recognition literature. It has 150 instances and 4 attributes. In this dataset, there are three different classes of 50 instances each, where each class refers to a type of iris plant.

Table 1: Characterstics of Dataset

| Attribute | Value |
|---|---|
| Data Set Characteristics | **multivariate** |
| Attribute Characteristics: | **Real** |
| Number of instances | **150** |
| Number of attributes | **4** |
| Missing value | **No** |
| Domain | **Life science** |

The visualization of all five different attributes viz. sepallength, sepalwidth, petallength, petalwidth and class are shown in Figure 1.

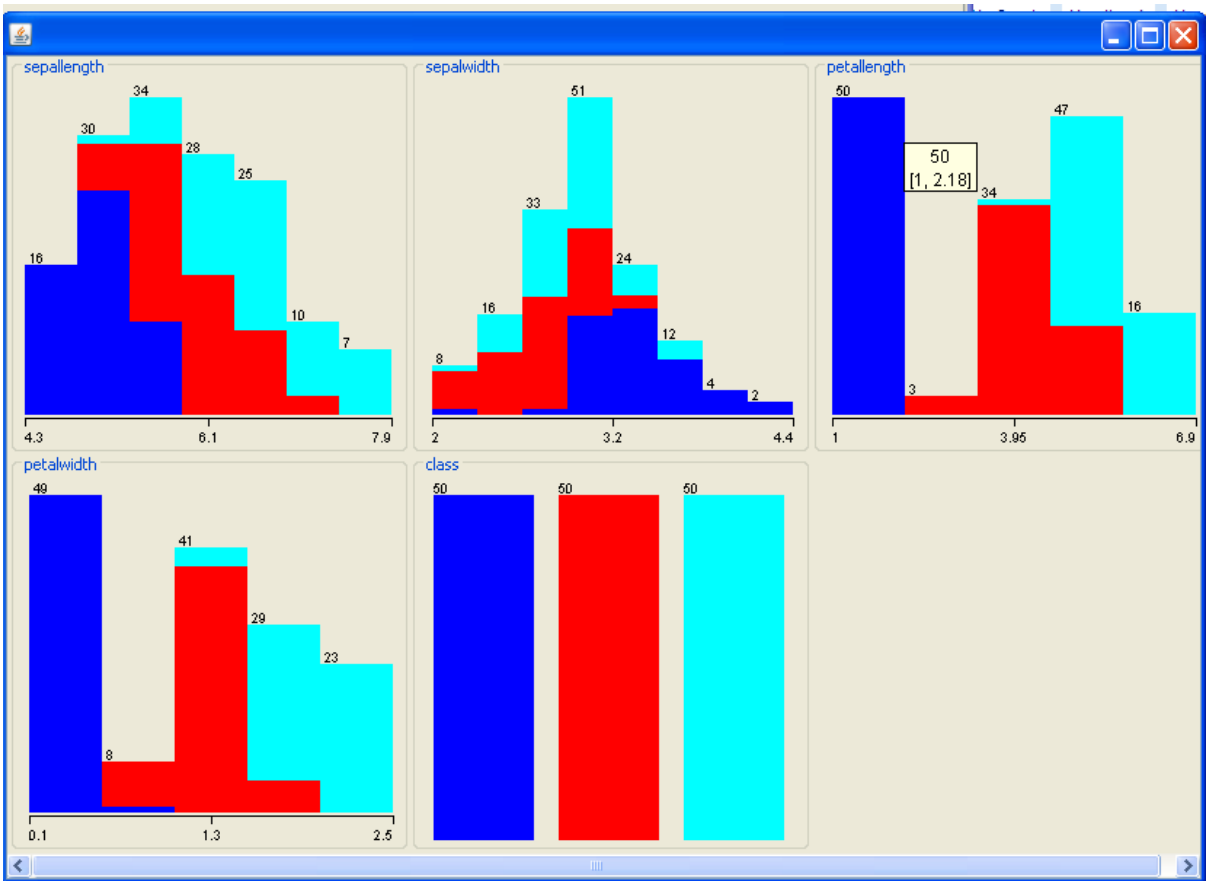


**Figure 1: Visualization of all IRIS attributes**

The four different classifier viz. Bayes Network Classifier, J48, Random Forest and OneR have been succesufully employed using WEKA for IRIS dataset. Table 2, 3, 4 and 5 presents the performance of these classifier in classifying the IRIS dataset.

**Table 2: Performance metric of Bayes Network Classifier**

| **Attribute** | Value |
|---|---|
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 139 (92.6667 %) |
| **Incorrectly Classified Instances** | 11 (7.3333 %) |
| **Kappa statistic** | 0.89 |
| **Mean absolute error** | 0.0454 |
| **Root mean squared error** | 0.1828% |
| **Root relative squared error** | 38.7793 % |
| **Time Taken** | 0.02 seconds |

**Table 3: Performance metric of J48**

| **Attribute** | Value |
|---|---|
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 144 (96 %) |
| **Incorrectly Classified Instances** | 06 (4%) |
| **Kappa statistic** | 0.94 |
| **Mean absolute error** | 0.035 |
| **Root mean squared error** | 0.1586 |
| **Root relative squared error** | 33.6353 % |
| **Time Taken** | 0.02 seconds |

**Table 4: Performance metric of Random Forest**

| **Attribute** | Value |
|---|---|
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 143 (95.3333 %) |
| **Incorrectly Classified Instances** | 07 (4.66%) |
| **Kappa statistic** | 0.93 |
| **Mean absolute error** | 0.04 |
| **Root mean squared error** | 0.1655 |
| **Root relative squared error** | 35.0999 % |
| **Time Taken** | 0.02 seconds |

**Table 5: Performance metric of OneR**

| **Attribute** | Value |
|---|---|
| **Total Number of Instances** | 150 |
| **Correctly Classified Instances** | 141 (94 %) |
| **Incorrectly Classified Instances** | 09 (6%) |
| **Kappa statistic** | 0.91 |
| **Mean absolute error** | 0.04 |
| **Root mean squared error** | 0.2 |
| **Root relative squared error** | 42.4264 % |
| **Time Taken** | 0.02 seconds |

## 4. Conclusion

The objective of this research work is to present the use of WEKA classifiers in categorizing the IRIS dataset. In this work, four different classifier viz. Bayes Network Classifier, J48,

Random Forest and OneR has been succefully used to classify the IRIS dataset. different performance metric such as correctly classified instances, incorrectly classified instances, kappa statistics, mean absolute error, root mean squared error, Root relative squared error along with execution time has been computed and examined. The rate of correctly classified instances using J48 is better than bayes network, random forest and oneR classifier. The use of J48 assist us in getting 96% of accuracy. Whereas, the mininmum rate of classification achieved is with bayes network classifier.

## 5. References

1. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
2. Sharma, M., G. Singh, and R. Singh. "Stark assessment of lifestyle based human disorders using data mining based learning techniques." IRBM 38.6 (2017): 305-324.
3. Corrales, David Camilo, Juan Carlos Corrales, and Apolinar Figueroa-Casas. "Towards detecting crop diseases and pest by supervised learning." Ingeniería y Universidad 19.1 (2015): 207-228.
4. 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.
5. Liakos, Konstantinos G., et al. "Machine learning in agriculture: A review." Sensors 18.8 (2018): 2674.
6. 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.
7. Elsalamony, Hany A. "Bank direct marketing analysis of data mining techniques." International Journal of Computer Applications 85.7 (2014): 12-22.
8. Vijiyarani, S., and S. Sudha. "Disease prediction in data mining technique–a survey." International Journal of Computer Applications & Information Technology 2.1 (2013): 17-21.
9. Kaur, Prableen, and Manik Sharma. "A survey on using nature inspired computing for fatal disease diagnosis." *International Journal of Information System Modeling and Design (IJISMD)*8.2 (2017): 70-91.
10. Meng, Gilliar, and HebaSaddeh. "Performance Analysis of Different Classifier for Diabetes Diagnosis." *International Journal of Computer Applications & Information Technology*11.2 (2019): 265-270.
11. Sharma, Manik, Gurvinder Singh, and Rajinder Singh. "An Advanced Conceptual Diagnostic Healthcare Framework for Diabetes and Cardiovascular Disorders." *arXiv preprint arXiv:1901.10530* (2019).
12. Gautam, Ritu, Prableen Kaur, and Manik Sharma. "A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings." *Progress in Artificial Intelligence* (2019): 1-24.
13. Kaur, Prableen, and Manik Sharma. "Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review." *Int. J. Pharm. Sci. Res* 9 (2018): 2700-2719.

14. Fatima, Meherwar, and Maruf Pasha. "Survey of machine learning algorithms for disease diagnostic." *Journal of Intelligent Learning Systems and Applications* 9.01 (2017): 1.
15. Diwani, Salim Amour, and Anael Sam. "Diabetes Forecasting Using Supervised Learning Techniques." *Advances in Computer Science: an International Journal* 3.5 (2014): 10-18.
16. Sharma, Manik, Samriti Sharma, and Gurvinder Singh. "Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining." *Data* 3.4 (2018): 54.
17. Kannan, K. Senthamarai, et al. "Financial stock market forecast using data mining techniques." *Proceedings of the International Multiconference of Engineers and computer scientists*. Vol. 1. 2010.
18. 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).
19. Ranjan, Jayanthi. "Data mining in pharma sector: benefits."*International journal of health care quality assurance* 22.1 (2009): 82-92.