INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Iris Identification with Precision: Machine Learning Classifiers-Based Classification

G. Bhargavi [1], S. Sailaja [2]
[1] Associate professor, Associate Professor [2]
Department of Computer Science Engineering
RISE Krishna Sai Prakasam Group of Institutions

Abstract—Classification is the most widely applied machine learning problem today, with implementations in face recognition, flower classification, clustering, and other fields. The goal of this paper is to organize and identify a set of data objects. The study employs K-nearest neighbors, decision tree (j48), and random forest algorithms, and then compares their performance using the IRIS dataset. The results of the comparison analysis showed that the K-nearest neighbors outperformed the other classifiers. Also, the random forest classifier worked better than the decision tree (j48). Finally, the best result obtained by this study is 100% and there is no error rate for the classifier that was obtained.

Keywords—Data Mining, Classification, Decision Tree, Random Forest, K-nearest neighbors

1. INTRODUCTION

Nowadays, the data online is massive, and it is growing on a daily basis. It is essential to handle such vast amounts of data and to view the most relevant queries on the user's computer. Since manually analyzing and retrieving relevant data from vast databases is impossible, automatic extraction tools are needed, which enable user-queried data to be retrieved from billions of sites on the internet and relevant knowledge to be discovered. Search engines such as Yahoo, Bing, MSN, and Google are commonly used by users to obtain data from the World Wide Web [1], [2]. Data mining is also used to explore and derive information from data warehouses. Data mining is a method of processing user data and extracting data from vast data warehouses that employ a variety of trends, intelligent processes, algorithms, and software. This approach will assist companies in evaluating results, forecasting potential patterns, and predicting user behavior. For relevant data extraction, data mining involves four techniques phases [3], [4], [5]. A data base is a set of information from different sources, a vast database that can include issue definitions. Data discovery is the

ISSN: 2057-5688

method of collecting valuable knowledge from vast volumes of unfamiliar data [6]. The third stage is modeling, which entails creating and evaluating various templates. Finally, in the final phase of data mining techniques, validated models are implemented [7]. Data mining methods may be used by businesses to turn raw data into useful facts. By understanding all about consumer actions, it will also assist companies in enhancing their communication campaigns and growing revenues [8], [9]. Moreover, this data should be properly classified to benefit from its great use. Classification tries to predict the target category with the highest accuracy. The classification algorithm establishes a connection through the input and output attributes in order to build a model [10], [11]. The volume of data collected in data mining environments is massive. Using the decision tree method is optimal if the data set is properly classified and contains the fewest number of nodes [12], [13]. A Decision Tree (DT) is a tree-based strategy in which every direction between the root and the leaf node is represented by a data separating series before a Boolean outcome is obtained [14], [15]. It is a hierarchical exemplification of nodes and links in information relationships. Nodes reflect uses as ties are used to distinguish [16]. DT is a form of ML algorithm that is applicable to both classification and regression. It typically makes use of the shape of a binary tree, with each node making a decision by comparing a function to a threshold and dividing the decision route there. Depending on whether the task is grouping or regression, leaf nodes include choices, actual values, or class names [17], [18]. Random Forest (RF) utilizes an ensemble of trees to create trees at random using the training input vector to estimate the output vector, equivalent to producing a random range of weights that is unchanged by previous weight sequences [19]. The best tree is then voted in, and the procedure is replicated a certain number of times, with the best tree being chosen as the corresponding classifier [20]. The K-Nearest Neighbors (KNNs) classifier, also known as the Nearest Neighbor Classifier, is a kind of supervised ML method that is used to classify or predict data. K-NN is incredibly easy to set up and use, yet it excels at specific grouping tasks like economic forecasting. Since it is a non-parametric approach, it does not have a particular training phase. Instead of classifying a question data point, it observes all of the data. K-NN can no longer make any assumptions regarding the underlying results. This property corresponds to the underlying trend in the vast majority of real-world datasets . The aim of this study is to evaluate the efficiency of the used methods that are based on classification. Besides, the researchers have

ISSN: 2057-5688

highlighted the most widely employed techniques as well as the strategies with the best precision.

2. RELATED WORK

The term data mining is a process of assigning individual objects in a database to one or set of categories or groups. In the phase of classification, the aim is to correctly classify the target class for each instance. This portion offers a survey of the most current and useful approaches to classification in different fields of ML that have been established by researchers in the last two years. Also, it only focuses on decision trees, random forests, and k-Nearest Neighbors as classifiers.

Lakhdoura and Elayachi compared the performance of two classifiers methods: J48 (c4.5) and RF on the IRIS features, and the test was executed by the WEKA 3.9. Therefore, the IRIS plant dataset, one of the most common databases for classification issues, is gained from the ML library at the University of California, Irvine (UCI). In addition, the investigators compared the results of both classifiers on various efficacy assessment measures. The findings showed that the J48 classifier outperforms the Random Forest (RF) classifier for IRIS variety prediction using various metrics such as classification precision, mean absolute error, and time to construct the technique. The J48 classifier has an accuracy of 95.83%, while the Random Forest has an accuracy of 95.55%. Mijwil and Abttan proposed using a C4.5 decision tree to reduce the effects of overfitting. The datasets used were IRIS, Car Assessment, Bottle, and WINE, both of which may be included in the UCI ML library. The trouble with this classifier is that it has so many nodes and divisions, which contributes to overfitting. This overfitting has the potential to sabotage the classification mechanism.

The experimental findings showed that the genetic algorithm was efficient in pruning the impact of overfitting on the four datasets and maximizing the trust Confidence Factor (CF) of the C4.5 decision tree, with an accuracy of about 92%. Rana et al. performed the comparison between Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with Linear Regression (LR) and Random Forest (RF) for IRIS Flower Classification. The suggested distinction between the results of both machine learning and the dimensionality reduction processes. According to the findings, both approaches provide decent classification performance, though the accuracy varies depending on the number of

ISSN: 2057-5688

principal components chosen. LDA, on the other hand, outperforms PCA for a defined collection of principal components. The analysis also showed that when the percentage of training data improves, so does the degree of precision.

LR and RF were used to classify the data. Both the RF and PCA approaches behave similarly to PCA and LDA. In comparison to the 86% of results provided by PCA, the LDA performs much better, providing 100% accuracy. Gong et al. [28] presented a new evidential clustering algorithm centered on the discovery of "cumulative belief peaks" and the application of the irrefutable K-NN principle. This method's basic assumption is that a cluster center in its neighborhood has the greatest accumulated probability of becoming a cluster center, and that its neighborhood is relatively big. Iris, Pima, Seeds, Waveform, WDBC, Wine, and Pen-based datasets were all included in the analysis. In the context of belief functions, a new notion of accumulated belief is proposed to quantify such cumulative probability. The scale of the comparatively wide neighborhood is calculated by optimizing an objective function. The cluster centers are then immediately detected as the objects with the highest collective confidence among their own neighborhood of this magnitude.

Finally, a creedal partition is formed using the evidential K-NN base and the constant cluster core. Experiment results show that when working with datasets with a limited number of data items and measurements, in a reasonable amount of time, the suggested evidence gathering method will easily classify cluster nodes and reveal data structure in the form of doctrinal sections. When using seeds as a dataset, the best accuracy is obtained, which is 95.26%. Shukla et al. focused around how machine learning algorithms can automatically identify the flower class with a high degree of precision rather than roughly. They used the IRIS dataset, and it is divided into three groups, each with 50 instances. The Iris dataset utilizes deep learning to classify the subclasses of Iris flower. Segmentation, function extraction, and classification are the three steps of this method's implementation. To identify the flower class, Neural Networks (NN), Logistic Regression (LR), Support Vector Machine (SVM), K-NN are utilized. The results showed that the accuracy achieved by each algorithm was as follows: Both NN, LR, and K-NN have an equal precision of 96.67% while a SVM has higher precision than all of which is 98%.

ISSN: 2057-5688

Ogundokun et al. investigated the diagnosis of longsightedness employing three techniques, namely NN, DT, and Back Propagation, resulting in the creation of an Expert System. Furthermore, the information area was extracted from detailed discussions with specialists in the area of eye examination (ophthalmologists) as well as various studies of the literature. The specialist framework was built from the ground up using the C# programming language and MySQL as the database. The NN was trained using back propagation and DT algorithms. According to the signs of the patient, a DT was used to identify and categorize the illness using an information extraction rule. The designed system's outcome demonstrated how the illness was detected in order to eliminate the neural network's impenetrability. Also, they showed that the hybridization of the three algorithms made the system model accurate and efficient, and eventually, the strategy was validated after implementation.

Sarpatwar et al. offered an end-to-end method to support privacy-enhanced decision tree classification using an open-source Homomorphic Encryption Library (HELib). They demonstrated the classification use case for decision trees with the iris dataset (150 samples, 4 functions, and 3 classes). The comparator and other associated processing in the first stage enable the function values to be within a certain range. Use a number of options to create a decision node, in addition to the ignorant accounts and the argmax feature in g Fully Homomorphic Encryption (FHE). The findings revealed that a highly stable and trustworthy decision tree service can be implemented, and the achieved precision was 98%, meaning that the private solution suited the non-private variant nearly exactly.

3. DATASET

In this article, three data mining algorithms on classifications are applied to the IRIS dataset from the UCI ML library. There are five characteristics in the data collection, each of which corresponds to a different iris flower species. Class (Species), Petal Length, Petal Width, Sepal Width, and Sepal Length are the characteristics [35]. There were 50 samples of each genus, totaling 150 examples. For the four non-species defining characteristics, this data form is broken down numerically in (cm) volume. Furthermore, it offers a clear and easy-to-manage presentation [36]. Data mining and deep learning have been extensively applied to clustering for several years for the iris dataset. It was postulated by the British statistician and evolutionary biologist Ronald Fisher in his publication, "On the Analysis of Covariance of

ISSN: 2057-5688

Taxonomic Studies," in which he argued that multiple measure testing ought to be preferred to one over one measure for character classification.

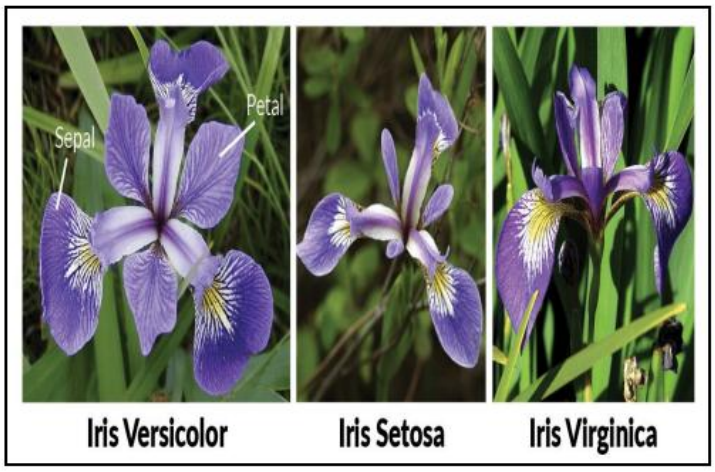


Fig. 1: IRIS flower types

TABLE 1: SAMPLE INSTANCES FROM IRIS DATASET

	Sepal length	Sepal width	petal length	petal width	species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

4. METHODOLOGY

Classification is a data mining strategy for categorizing data instances into one of a few classes. Machine learning classification algorithms are made up of many algorithms that have been designed to outperform one another. They all use statistical methods such as decision trees, linear programming, support machine vectors, and neural networks, among others. To make a guess, these methods examine the available data in a variety of ways

ISSN: 2057-5688

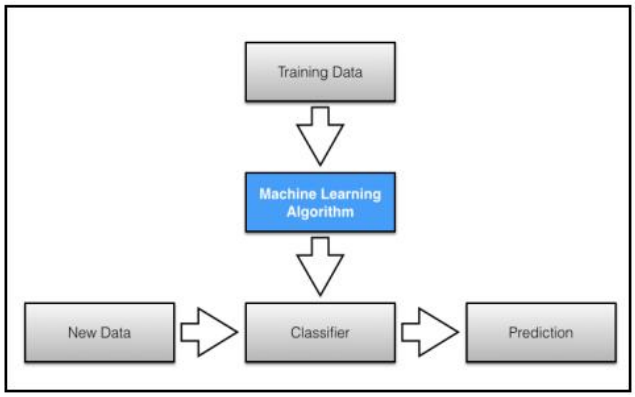
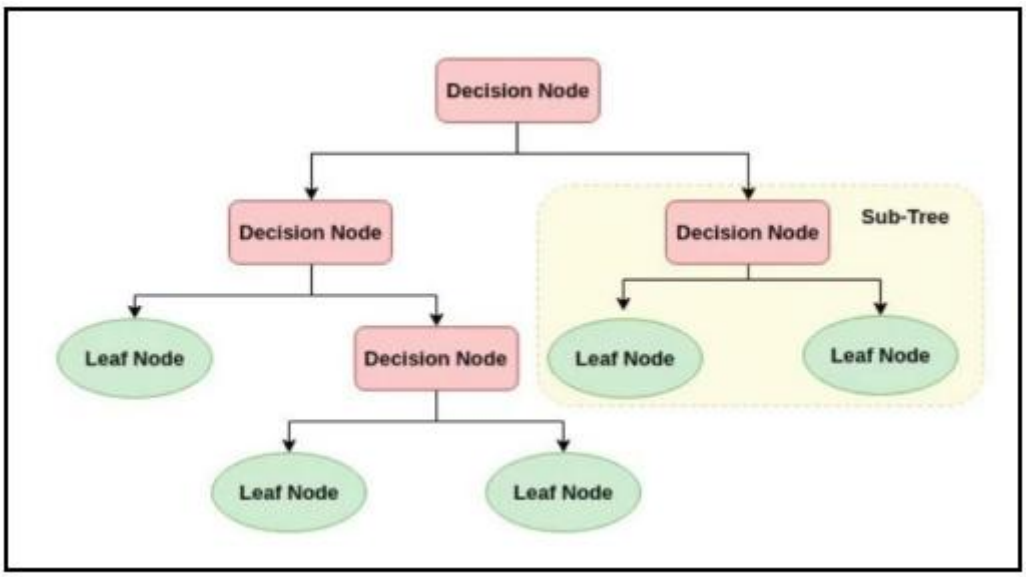


Fig. 2: Simplified diagram of the procedures for building the general pattern classification model

This work focused on decision tree, random forest, and k-Nearest neighbourhoods algorithms in general, and they are implemented by the data mining tool known as Weka. “Fig. 2” depicts a simplistic illustration of the procedures for building the general pattern classification technique.

Decision Tree Classifier: One of the techniques widely used in data mining is the systems that create classifiers [41]. DT is a text and data mining classification algorithm that was used previously. Decision Tree classifiers (DTCs) have been shown to be effective in a variety of classification applications. A hierarchical decomposition of the data space is the framework of this methodology. D. Morgan first suggested, and J.R. Quinlan established the DT as a classification task. The basic concept is to create a tree with classified data points dependent on attributes, but the main problem of a DT is deciding which attributes or features should be at the parent level and which should be at the child level. De Mántaras suggested statistical modeling for feature selection in trees as a solution to this issue.

ISSN: 2057-5688



5. CONCLUSION

Nowadays, classification is the most often utilized in machine learning problems with a number of applications such as face recognition, flower classification, clustering, and so on. In order to construct a model, the classification algorithm creates a connection between the input and output characteristics and attempts to predict the target population with the greatest accuracy. The main objective of this study was to come to a consensus on how well K-nearest neighbors, decision tree (j48), and random forest algorithms performed in IRIS flower classification. According to the findings, both approaches yield strong classification outcomes, and the precision is calculated by the number of principal components used. The analysis also found that when the percentage of training data improves, so does the degree of precision. In comparison to random forest, which achieved 99.33% accuracy, and decision tree (j48), which achieved 98% accuracy, the experimental findings revealed that K-nearest neighbors performed significantly better, achieving 100% 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.

REFERENCES

[1] M. J. H. Mughal, “Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 6, 2018, doi: 10.14569/IJACSA.2018.090630.

ISSN: 2057-5688

[2] D. Q. Zeebaree, A. M. Abdulazeez, O. M. S. Hassan, D. A. Zebari, and J. N. Saeed, Hiding Image by Using Contourlet Transform. press, 2020.

[3] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," J. Appl. Sci. Technol. Trends, vol. 1, no. 2, pp. 56–70, 2020.

[4] M. A. Sulaiman, "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications," J. Soft Comput. Data Min., vol. 1, no. 2, pp. 11–25, 2020.

[5] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and region growing for breast cancer segmentation," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 88–93.

[6] S. H. Haji and A. M. Abdulazeez, "COMPARISON OF OPTIMIZATION TECHNIQUES BASED ON GRADIENT DESCENT ALGORITHM: A REVIEW," PalArchs J. Archaeol. Egypt Egyptol., vol. 18, no. 4, Art. no. 4, Feb. 2021.

[7] I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, 2021.

[8] P. Galdi and R. Tagliaferri, "Data mining: accuracy and error measures for classification and prediction," Encycl. Bioinforma. Comput. Biol., pp. 431–6, 2018.

[9] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," J. Appl. Sci. Technol. Trends, vol. 1, no. 4, pp. 140–147, 2020.

[10] G. Gupta, "A self explanatory review of decision tree classifiers," in International conference on recent advances and innovations in engineering (ICRAIE2014), 2014, pp. 1–7.

[11] N. S. Ahmed and M. H. Sadiq, "Clarify of the random forest algorithm in an educational field," in 2018 international conference on advanced science and engineering (ICOASE), 2018, pp. 179–184.

[12] T. Bahzad and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20– 28, 2021.

ISSN: 2057-5688

[13] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in 2018 International Conference on Advanced Science and Engineering (ICOASE), 2018, pp. 145–150.

[14] N. M. Abdulkareem and A. M. Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," Int. J. Sci. Bus., vol. 5, no. 2, pp. 128–142, 2021.

[15] 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, 2015.

[16] A. S. Eesa, A. M. Abdulazeez, and Z. Orman, "A DIDS Based on The Combination of Cuttlefish Algorithm and Decision Tree," Sci. J. Univ. Zakho, vol. 5, no. 4, pp. 313–318, 2017. [17] K. Rai, M. S. Devi, and A. Guleria, "Decision tree based algorithm for intrusion detection," Int. J. Adv. Netw. Appl., vol. 7, no. 4, p. 2828, 2016.

[18] 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, 2019.

[19] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," Mach. Learn., vol. 62, no. 03, 2020. [20] S. Dahiya, R. Tyagi, and N. Gaba, "Comparison of ML classifiers for Image Data," EasyChair, 2020.