IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Case Study On Mushroom Classification Using Data Mining Techniques: An Experimental Approach

Dr P Devaraju
Assistant Professor,
Department of Computer Science &Technology,
Sri Krishnadevaraya University,
Anantapuramu

ABSTRACT

Mushrooms are diverse fungi that can be either edible or poisonous, posing a significant health risk if misidentified. This study explores the application of data mining techniques to classify mushrooms based on their characteristics, such as cap shape, color, odor, and more, to determine their edibility. The dataset used for this research includes 8124 mushroom instances, each described by 23 features, with 4208 labeled as edible and 3916 as poisonous. Two popular data mining algorithms, Artificial Neural Networks (ANN) and Logistic Regression, were employed to build predictive models. These models were evaluated using classification performance metrics including accuracy, precision, and recall. The results indicate that the ANN outperforms Logistic Regression, achieving an accuracy of 98.62%, demonstrating its effectiveness in distinguishing between edible and poisonous mushrooms. This paper highlights the potential of data mining techniques for improving classification accuracy and ensuring safe mushroom consumption.

1. Introduction

The objective of data mining is to extract meaningful patterns and knowledge from large datasets, with the goal of categorizing data into different classes and making predictions for future instances. Unlike traditional machine learning, where a single model is used, ensemble learning techniques generate multiple models. These models work in concert to make predictions by combining their outputs, typically using methods like averaging or voting. Ensemble learning is widely applicable across various types of models and learning tasks, making it a powerful approach for improving prediction accuracy and generalization [1][2]. A key insight in supervised learning is that ensembles can often outperform individual models by capturing more complex patterns and reducing the risk of overfitting [3][4]. In developing a data mining model, one crucial step is to determine the type of base model and learning technique to use. The effectiveness of an ensemble largely depends on how well the individual base models complement each other.

Data mining plays a vital role in transforming large datasets into valuable information by identifying hidden patterns. It is especially relevant in fields like healthcare, fraud detection, banking, and marketing, where large amounts of data need to be analyzed to extract actionable insights [6]. In the context of classification, data mining algorithms are used to group data into predefined categories based on their attributes. These techniques allow for the creation of predictive models that can be applied to unseen data for future classification tasks [7]. Classification is one of the primary data mining tasks, and it involves the process of learning a model from training data and using that model to predict the class labels of new data.

2. Classification

Classification is the process of identifying which category or class an instance belongs to, based on a set of input features. It is typically a two-step process: first, a classifier is trained on a labeled dataset, where each instance is associated with a known class label. In the second step, the trained model is used to classify unseen data by predicting the class labels of instances whose class membership is unknown. The model's performance is evaluated by comparing its predictions against the true labels of the test data. If the classifier performs at an acceptable level of accuracy, it can be used to classify future instances [4][5].

Classification can be seen as a form of predictive modeling, where the goal is to develop a model that maps input data to specific class labels. The quality of a classifier is often measured by accuracy, precision, recall, and F1 score, which help evaluate how well the model is generalizing to unseen data. In data mining, classification techniques are widely used for various tasks, including medical diagnosis, fraud detection, and customer behavior analysis [6][8].

3. Methodology

This section outlines the key data mining techniques used in the research, including Logistic Regression and Artificial Neural Networks (ANN), which were selected for mushroom classification tasks.

3.1 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the human brain, designed to recognize patterns by simulating the way neurons interact. In ANNs, data is passed through layers of interconnected nodes (neurons) that process and transmit information, typically represented as binary values (0s and 1s). Each neuron is assigned a weight that reflects its importance in making predictions, and the network learns by adjusting these weights to minimize errors during training [4][5]. ANNs are composed of an input layer, one or more hidden layers, and an output layer. The input layer receives the data, the hidden layers perform computations, and the output layer provides the final predictions. Activation functions play a crucial role in determining whether a neuron should be activated, helping to model non-linear relationships in the data.

3.1.1 Multi-layer Perceptron (MLP)

The Multi-layer Perceptron (MLP) is a type of ANN with one or more hidden layers between the input and output layers. It is a powerful model for supervised learning tasks and is used in a variety of applications, including classification. The MLP learns to map input features to the appropriate class labels by adjusting weights between layers through backpropagation, which is an optimization technique that minimizes prediction error. MLPs are particularly effective for complex, non-linear classification tasks and have been applied in fields ranging from image recognition to medical diagnosis [6][7][9].

3.2 Logistic Regression

Logistic Regression is a statistical method used to model the relationship between one or more independent variables and a binary dependent variable. It is often used when the outcome variable is categorical and can take one of two possible values (e.g., 0 or 1). The logistic regression model estimates the probability of an event occurring by applying the logistic function, which outputs a value between 0 and 1. In the context of mushroom classification, logistic regression is used to predict whether a mushroom is edible or poisonous based on its features, such as cap shape, color, and odor [10]. This method is particularly suitable for classification tasks where the dependent variable is binary.

4. Experimental Results

The experiments in this study were conducted using Python and the Scikit-learn library, which provides a wide range of data mining algorithms. Two supervised learning algorithms—Artificial Neural Networks (ANN) and Logistic Regression—were applied to classify mushrooms based on their characteristics, such as cap shape, cap color, and odor. The dataset used for training and testing contains 8124 instances of mushrooms, each described by 23 features. The mushrooms are classified into two categories: edible (4208 instances) and poisonous (3916 instances) [10].

The experimental results for the selected algorithms are summarized in Figure 1:

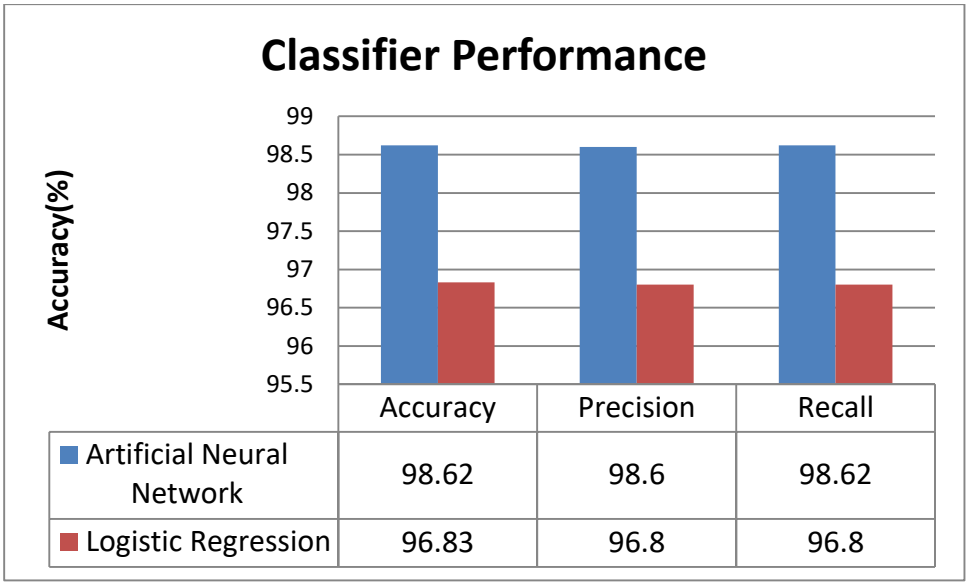

Figure-1: Classifier Performance

As shown in Figure 1, the Artificial Neural Network outperforms Logistic Regression in terms of accuracy, precision, and recall. The ANN achieved an impressive accuracy of 98.62%, while Logistic Regression achieved 96.83%. This indicates that the ANN model is more effective in distinguishing between edible and poisonous mushrooms based on the provided features.

4.1 Discussion

The results highlight the effectiveness of machine learning techniques, particularly ANNs, in accurately classifying mushrooms. Key points from the discussion include:

- **ANN Superiority:** The Artificial Neural Network demonstrated superior performance over Logistic Regression, achieving higher accuracy, precision, and recall. This suggests that neural networks can capture more complex relationships between features, resulting in better classification results.
- **Practical Implications:** The high accuracy of the ANN model (98.62%) makes it highly valuable for practical applications, such as helping foragers, mycologists, and consumers make informed decisions about mushroom edibility.

5. Conclusion

This study demonstrated the potential of machine learning, specifically Artificial Neural Networks and Logistic Regression, in classifying mushrooms as either edible or poisonous. The ANN model, with an accuracy of 98.62%, outperformed Logistic Regression and showed promise as a reliable tool for mushroom classification. These models can play a crucial role in ensuring safe mushroom consumption, which is vital for foragers and consumers. Further research could explore the incorporation of additional features or more advanced data mining techniques to improve classification performance even further.

This work emphasizes the power of data mining techniques in addressing real-world classification challenges and their potential to enhance safety and decision-making in various domains.

References

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G Ravi Kumar, K Tirupathaiah and B Krishna Reddy, "Client Churn prediction of banking and fund industry utilizing machine learning techniques",IJCSE,Volume-7,Issue- 6,PP:842-846, 2019

[3] G. Ravi Kumar, K. Venkata Sheshanna, S. Rahamat Basha, and P. Kiran Kumar Redd, "An Improved Decision Tree Classification Approach for Expectation of Cardiotocogram", Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing, Lecture Notes on Data Engineering and Communications Technologies 62, https://doi.org/10.1007/978-981-33-4968-1_26

[4] Dr. G. Thippanna1, Dr. D. William Albert, E. Ramachandra "A REVIEW ON BIG DATA INTEGRATION's DIFFICULTIES WITH AI", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 14 No. 4 July-Aug 2023.

[5] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[6] J.Han and M.Kamber, Data Mining concepts and Techniques, the Morgan Kaufmann series in Data Management Systems, 2nded.San Mateo, CA; Morgan Kaufmann, 2006.

[7] N. Michael, "Artificial Intelligence - A Guide to Intelligent Systems",2nd edition, Addison Wesley, 2005.

[8] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[9] M. V. Lakshmaiah, G. Ravi Kumar and G. Pakardin, "Frame work for Finding Association Rules in Bid Data by using Hadoop Map/Reduce Tool", International Journal of Advance and Innovative Research, Volume 2, Issue1(1), PP:6-9,2015, ISSN: 2394-7780

[10] UCI machine learning repository. http://archive.ics.uci. edu/ml/