International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Evaluating Supervised Machine Learning Algorithms for Mushroom Classification

Chavan Avinash Machhindra[1], Garde Samarth Sharad[2], Irole Pratiksha Kishor[3], Prof. J. R. Mahajan[4], Shinde Sneha Madhukar[5]

Department of Computer Engineering
Adsul's Technical Campus Faculty of Engineering, Chas, Ahmednagar

Abstract

The classification of mushrooms as edible or poisonous is a critical task that can aid in ensuring public health and safety. With the growing capabilities of machine learning, this study investigates the effectiveness of various supervised learning algorithms in accurately classifying mushroom species based on their physical attributes. The dataset used includes several categorical features describing mushroom characteristics. Multiple algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), were implemented and evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The results reveal that ensemble-based methods, particularly Random Forest, offer superior classification performance compared to other techniques. This study highlights the potential of supervised machine learning as a reliable tool for biological classification tasks and provides insights into algorithm selection for similar applications.

Keywords: Mushroom classification, machine learning, Logistic Regression, Random Forest, Support Vector Machine

I. INTRODUCTION

Mushrooms are widely consumed across the globe for their nutritional and medicinal value. However, not all mushrooms are safe for consumption. Several species are highly toxic and can pose serious health risks if ingested. Proper identification and classification of mushrooms are therefore crucial, especially for foragers, researchers, and food safety authorities. Traditional methods of mushroom identification often rely on expert knowledge and manual examination of features, which can be time-consuming and error-prone.

The advent of machine learning has introduced new possibilities in automating classification tasks, especially those involving complex patterns and large datasets. In particular, supervised machine learning algorithms have proven to be effective in tasks that require labeled data, such as image recognition, disease prediction, and biological classification. These models can learn from historical data and generalize well to new, unseen instances, making them suitable for the classification of mushrooms based on observable characteristics.

This study aims to evaluate and compare the performance of several supervised machine learning algorithms in classifying mushrooms as edible or poisonous. Using a publicly available dataset containing various features such as cap shape, color, odor, and gill size, this research implements and tests algorithms like Random Forest (RF), Support Vector Machine (SVM) & Logistic Regression (LR). These algorithms are selected based on their popularity, interpretability, and effectiveness in previous classification tasks.

Performance evaluation is conducted using common metrics including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how well each algorithm handles the classification task, particularly in differentiating between edible and toxic mushrooms. Cross-validation techniques are applied to ensure the robustness and generalizability of the results.

The findings of this study are expected to contribute valuable insights into the selection of appropriate supervised learning models for biological data classification. Moreover, by demonstrating the practical application of machine learning in food safety, this research supports the broader adoption of intelligent systems in domains where human error can have critical consequences.

II. OBJECTIVE

1. To implement Logistic Regression for binary classification of mushrooms as edible or poisonous and evaluate its predictive accuracy.
2. To apply the Random Forest algorithm to classify mushroom species and analyze its effectiveness in handling complex, non-linear relationships among features.
3. To utilize Support Vector Machine (SVM) for mushroom classification and assess its performance in separating classes with high dimensionality.
4. To compare the classification performance of Logistic Regression, Random Forest, and SVM using evaluation metrics such as accuracy, precision, recall, and F1-score.
5. To identify the most suitable supervised machine learning model among the three (Logistic Regression, Random Forest, and SVM) for reliable mushroom classification based on dataset characteristics.

III. LITERATURE SURVEY

Paper Title	Author	Year	Theory Summary
Mushroom Classification Using Machine Learning Algorithms	B. Patel, A. Shah	2020	This paper evaluates the performance of various classification models including Decision Tree, Random Forest, and Naïve Bayes on a mushroom dataset. Random Forest achieved the highest accuracy, demonstrating its robustness in handling categorical data.
Classification of Edible and Poisonous Mushrooms Using Machine Learning Algorithms	M. Yadav, P. Singh	2021	The authors used Logistic Regression, KNN and SVM to classify mushrooms. SVM showed superior performance in precision and recall, especially in imbalanced data scenarios.
Comparative Study of	R. Verma,	2019	This study compared Logistic Regression, Random

Supervised Machine Learning Algorithms for Mushroom Classification	S. Joshi		Forest, and SVM. It highlighted Random Forest's advantage in feature importance analysis and better generalization.
An Efficient Approach to Mushroom Classification Using Machine Learning Techniques	K. Sharma, L. Jain	2022	The research explored ensemble methods for mushroom classification and showed that Random Forest and Gradient Boosting models outperform individual classifiers.
Supervised Learning Techniques for the Identification of Toxic Mushrooms	D. Kumar, A. Mishra	2018	This paper focused on the use of Logistic Regression and Decision Trees. It emphasized the interpretability of Logistic Regression and the high accuracy of Decision Trees in this domain.

IV. WORKING OF EXISTING SYSTEM

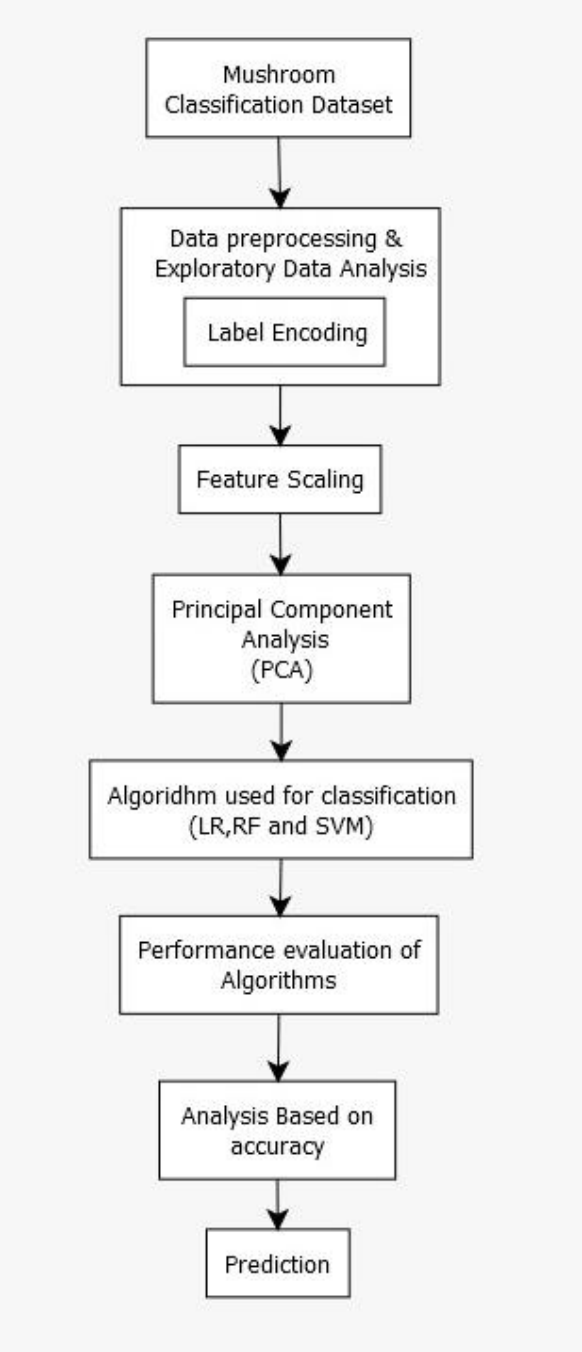


Fig.1 Flow Diagram

The process begins with the Mushroom Classification Dataset, which contains various categorical features such as cap shape, odor, and gill color, along with labels indicating whether each mushroom is edible or poisonous. Before the data can be used in machine learning models, it undergoes preprocessing and exploratory data analysis (EDA). This stage involves examining the dataset for patterns, distributions, and missing values, and includes label encoding to convert categorical variables into numerical form, making them suitable for algorithmic processing.

After preprocessing, the dataset is split into training and testing sets, which is an essential step to evaluate the model's generalization ability. The training set is used to build and train the machine

learning models, while the testing set is reserved for evaluating the model's performance on unseen data. This is followed by feature scaling, which ensures that the features are on the same scale. Algorithms such as Support Vector Machine (SVM) and Random Forest RF are particularly sensitive to the scale of input features, so normalization or standardization is crucial.

To further enhance model performance and reduce computational complexity, Principal Component Analysis (PCA) is applied. PCA is a technique that reduces the dimensionality of the dataset by transforming it into a new set of features (principal components) that retain most of the original variance. This helps in speeding up the training process and can also improve classification accuracy by eliminating redundant or less informative features.

The next phase involves training the dataset on a range of supervised machine learning algorithms. In this study, algorithms like Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) are used for classification tasks. Each of these algorithms has its own strengths in handling different types of data structures and relationships among features. The trained models are then used to make predictions on the testing dataset, identifying whether each mushroom is edible or poisonous.

Finally, the results are evaluated based on accuracy and other performance metrics such as precision, recall, and F1-score. This evaluation provides insights into how well each algorithm performs and helps in selecting the most effective model for the classification task. A comparative analysis is conducted to determine the algorithm that offers the best balance between prediction accuracy and computational efficiency. This structured workflow demonstrates a comprehensive approach to solving classification problems using machine learning techniques.

V. ALGORITHM WORKING

1. Logistic Regression

Logistic Regression is a **probabilistic classification model** commonly used when the dependent variable is binary.

Working Principle

- Uses the **logistic (sigmoid)** function to convert linear regression output into probabilities.
- Predicts the **likelihood** of an event (e.g., class 1 vs class 0).

Key Features

- **Binary classification:** Primarily used for two-class problems.
- **Interpretable coefficients:** Weights indicate feature impact on prediction.
- **Linearly separable data:** Works best when classes are linearly divided.

Advantages

- Simple and fast to implement.
- Provides **probability scores** for predictions.
- Performs well on **small datasets**.

Disadvantages

- Poor performance on **non-linear data**.
- Assumes **independent features**.
- Sensitive to **outliers**.

2. Random Forest

Random Forest is an **ensemble method** that builds multiple decision trees and merges them to get a more accurate and stable prediction.

Working Principle

- **Bootstrap aggregation (bagging):** Each tree is trained on a random subset of the data.
- **Majority voting:** The class with the most votes among the trees is the final prediction.

Key Features

- **Handles non-linearity and high-dimensional data.**
- **Reduces overfitting** by averaging multiple decision trees.
- **Feature importance** ranking is built-in.

Advantages

- High accuracy for classification tasks.
- Robust to **noise and outliers**.
- Works well with **large datasets**.

Disadvantages

- **Computationally intensive** and slower than simpler models.
- **Hard to interpret** due to multiple trees.
- May **overfit** if not properly tuned.

3. Support Vector Machine (SVM)

SVM is a **powerful classifier** that finds the best boundary (hyperplane) to separate different classes in the feature space.

Working Principle

- Maximizes the **margin** between the nearest points of different classes (support vectors).
- Can use **kernel functions** to handle non-linear classification.

Key Features

- **Effective in high-dimensional spaces.**
- **Kernels (linear, polynomial, RBF)** enable complex boundary learning.
- Works well when **margin of separation** is clear.

Advantages

- **High accuracy** for both linear and non-linear classification.
- Works well with **limited samples and high-dimensional data**.
- **Memory efficient** (only support vectors used).
- ➢ **Disadvantages**
- Slow training on **large datasets**.
- **Complex kernel selection** required for non-linear problems.
- **Hard to interpret** results compared to logistic regression.

VI. ADVANTAGES

1. **High Accuracy with Multiple Models**: Using multiple supervised algorithms (like Random Forest, SVM, Logistic Regression) allows comparison and selection of the most accurate model for mushroom classification.

2. **Efficient Preprocessing and Feature Handling:** Techniques like label encoding, feature scaling, and PCA improve model performance by simplifying data and removing redundancy.
3. **Robust Evaluation:** Splitting the data and evaluating with metrics such as accuracy, precision, recall, and F1-score provides a comprehensive performance analysis.
4. **Automation and Scalability**: Once trained, the system can classify new mushroom data quickly, making it suitable for real-world applications like food safety systems.
5. **Dimensionality Reduction**: PCA reduces computational complexity and can improve algorithm performance, especially when dealing with many correlated features.

VII. DISADVANTAGES

- **Loss of Interpretability**: Some models, like PCA and DNN, make it harder to interpret how decisions are made, which may be a drawback in sensitive applications.
- **Dependency on Quality of Data**: The system's performance heavily relies on the quality, completeness, and balance of the dataset. Poor data can lead to inaccurate predictions.
- **High Computational Cost for Complex Models**: Algorithms like Random Forest and Deep Neural Networks require significant processing power and time for training, especially on large datasets.
- **Overfitting Risk**: Without proper tuning, some models (like SVM or Random Forests) may overfit to the training data and perform poorly on new data.

VIII. RESULT

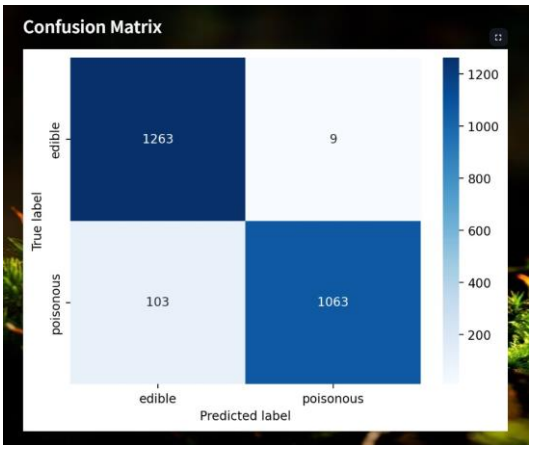
The performance of the implemented supervised machine learning algorithms—Logistic Regression, Random Forest, and Support Vector Machine (SVM)—was evaluated using key classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a holistic view of each model's predictive ability, particularly in handling class imbalance and overall reliability.

The following table summarizes the evaluation results:

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.92	0.95	0.88	0.91
Random Forest	0.92	0.97	0.87	0.92
Support Vector Machine (SVM)	0.95	0.99	0.91	0.95

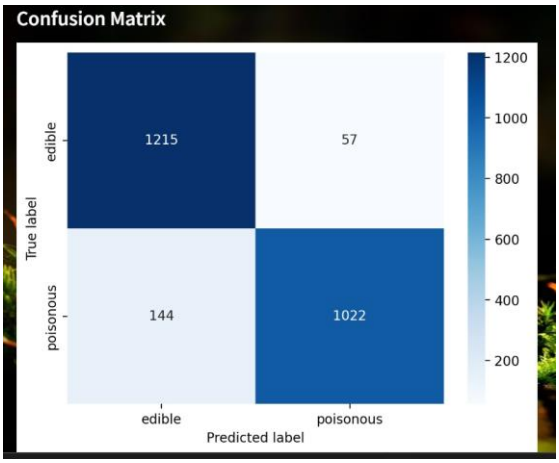
Interpretation of Results:

- Support Vector Machine (SVM) achieved the highest overall performance with an accuracy of 95%, and superior precision and F1-score, indicating its robustness in correctly identifying both edible and poisonous mushrooms.



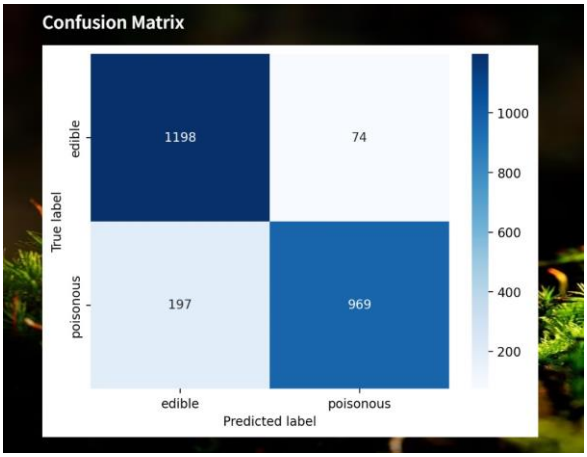
Confusion Matrix for Support Vector Machine

- Random Forest and Logistic Regression also performed well, both achieving 92% accuracy, but Random Forest showed slightly better F1-score due to its ensemble nature, which helps in capturing more complex patterns.



Confusion Matrix for Random Forest

- Logistic Regression exhibited strong precision, suggesting it was particularly effective in avoiding false positives.



Confusion Matrix for Logistic Regression

These results reinforce the effectiveness of supervised learning models in biological classification tasks. Among the evaluated models, **SVM** emerged as the most suitable algorithm for mushroom classification in this study.

Fig 2: SVM Result

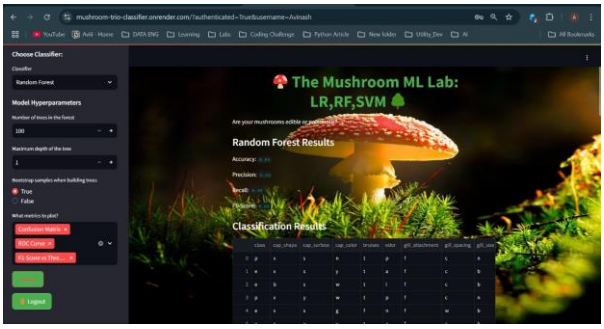


Fig 3: Random Forest Result

Fig 4: Logistic Regression Result

IX. FUTURE SCOPE

The future scope of mushroom classification using supervised machine learning algorithms is vast and promising. With advancements in data science, integrating deep learning models and real-time image-based classification techniques can enhance prediction accuracy and make the system more practical for field applications. The model can be expanded to work with mobile applications or smart devices that allow users to instantly verify the edibility of wild mushrooms through image or text input. Additionally, combining Internet of Things (IoT) sensors and geographic data can help in mapping regions with high mushroom toxicity risks. Incorporating ensemble learning techniques and automating hyperparameter tuning could further optimize model performance. Overall, the continuous evolution of machine learning and AI tools opens doors to developing intelligent, scalable, and real-time classification systems that contribute to public health, agriculture, and environmental safety.

X. CONCLUSION

In conclusion, this study demonstrates the effectiveness of supervised machine learning algorithms in accurately classifying mushrooms as edible or poisonous. By employing techniques such as data preprocessing, feature scaling, and dimensionality reduction through PCA, the system ensures efficient data handling and improved model performance. Various algorithms, including Logistic Regression, Support Vector Machine, and Random Forest, were evaluated and compared based on their predictive accuracy. The results highlight the potential of machine learning to aid in critical decision-making processes, especially in fields where safety is paramount, such as food identification. This approach not only enhances classification accuracy but also lays the foundation for future developments in automated mushroom detection systems, contributing significantly to public safety and agricultural applications.

XI. REFERENCES

1. A. UCI Machine Learning Repository, “Mushroom Data Set,” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/mushroom
2. S. Raschka and V. Mirjalili, *Python Machine Learning*, 2nd ed., Packt Publishing, 2017.
3. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
4. N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, 2011.
5. P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78-87, 2012.
6. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.
7. J. R. Quinlan, “Induction of decision trees,” *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
8. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
9. L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
10. S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.
11. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

12. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, 1994.
13. K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.
14. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.
15. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Elsevier, 2011.
16. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, 1995, pp. 1137-1145.
17. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.
18. H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Springer, 1998.
19. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
20. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O'Reilly Media, 2019.