**Mushroom Classification using Classification Model**

Sushant Subhash Patil

Shri Guru Gobind Singhji Institute of Engineering and Technology

(SGGSIET), Nanded

Department of Information Technology,

[2021bit020@sggs.ac.in](mailto:2021bit020@sggs.ac.in)

**Abstract:**

This research investigates the classification of mushrooms as either edible or poisonous using machine learning algorithms applied to meteorological data. Despite Indonesia's rich mushroom diversity, there exists a dearth of studies focusing on accurately discerning between safe and toxic varieties. To address this gap, this study compares three prominent classification algorithms: Decision Tree (C4.5), Naïve Bayes, and Support Vector Machine (SVM). The experimentation employs the WEKA tool, a widely-used platform for data mining analysis. Mushroom data sourced from reputable repositories such as The Audubon Society Field Guide to North American Mushrooms in the UCI machine learning repository are utilized. Results reveal that the C4.5 algorithm achieves a commendable 100% accuracy rate, akin to SVM, while exhibiting superior processing speed. This finding highlights the efficiency of the C4.5 algorithm in accurately categorizing mushrooms based on meteorological data. The study contributes to enhancing our understanding of mushroom classification techniques and offers insights into the potential applications of machine learning in the field of mycology [1] In APA style, authors' names are listed with the last name first, followed by their initials. If there are multiple authors, they are separated by commas, with an ampersand (&) before the last author yconclusion , the research underscores the significance of employing machine learning algorithms in the identification of edible and poisonous mushrooms. By leveraging meteorological data and sophisticated classification techniques, this study advances the capability to differentiate between safe and harmful mushroom species, thereby facilitating safer consumption practices and contributing to public health and safety.

**Result: -** The decision tree Regression stands out as the most suitable model for predicting Edibility Classification Mushroom. It achieves a decision tree value of 100%.

**Keywords:** Mushroom classification, Logistic regression, Decision trees, Support vector machines (SVM), Random forests.

**1.Introduction:**

Edible mushrooms, which grow naturally and seasonally, are an important food for people living in rural areas. Some mushrooms are poisonous. Many deaths occur every year as a result of consuming poisonous mushrooms. Determining whether a mushroom is poisonous by visual inspection is a matter of skill. Most of the mushrooms used in the world are still ready to be collected in nature. However, not distinguishing between mushrooms collected from nature and grouping them as poisonous mushrooms can lead to serious problems and even death. This makes people more conscious about mushroom consumption [1-4].

When determining whether mushrooms are generally edible or poisonous, methods for identifying poisonous mushrooms are mainly based on image recognition and biological analysis [5]. It is difficult for non-experts to perform biological analyzes in daily life and make judgments based on human characteristics. For this reason, many researchers have worked on different methods and methods. An example of these studies is machine learning.algorithms [6, 7], deep learning algorithms [8, 9], rule recognition algorithms [10, 11], and image processing algorithms [1].

In this study,22 features of mushroom datasets were used and their classification was made using machine learning. It is difficult for non-experts to perform biological analyses in daily life and make judgments based on human characteristics. For this reason, many researchers have worked on different methods and methods. The following activities were used to find the contribution of the research to the literature.These are: The dataset consists of edible mushrooms (4208) and poisonous mushrooms (3916).Decision Tree (DT), Naive Bayes (NB), AdaBoost (AB) and Support Vector Machine (SVM) machine learning algorithms were used for classification.performance metrics calculated and compared for each algorithm. [7],[10] From previous researchers, it can be seen a limited method in identifying the types of mushrooms,

**2.Literature Review:**

The classification of mushrooms into edible and poisonous categories has been a subject of interest in various scientific disciplines, particularly in the context of public health and mycology. Despite the global significance of mushroom diversity, research on effective classification techniques remains limited, particularly in regions like Indonesia with rich fungal biodiversity.Several studies have investigated the use of machine learning algorithms for mushroom classification, often leveraging diverse datasets and methodologies. For instance, Patel et al. (2017) explored the application of Decision Trees, Support Vector Machines, and Naïve Bayes classifiers for mushroom classification using morphological features. Their study demonstrated promising results, with Decision Trees exhibiting high accuracy in distinguishing between edible and poisonous mushrooms.focused on the classification of mushroom species using DNA barcoding data. By applying machine learning techniques such as Random Forest and Gradient Boosting Machines, they achieved robust classification performance, highlighting the importance of molecular data in mushroom classification efforts.

In the context of meteorological data, Liu et al. (2020) investigated the impact of environmental factors on mushroom growth and distribution. Their study emphasized the significance of temperature, humidity, and precipitation patterns in shaping mushroom ecosystems, suggesting the potential utility of meteorological data in mushroom classification models.Furthermore, the WEKA tool has emerged as a popular platform for conducting experiments in mushroom classification research (Frank et al., 2016). [2] Regression models are commonly used for taxi trip duration prediction due to their ability to handle continuous target variables. Various regression techniques, including linear regression, decision trees, random forests, support vector regression (SVR), and neural networks, have been applied to this domain.[3]Several studies have compared the performance of different regression models in predicting mushroom classification. Its user-friendly interface and comprehensive suite of machine learning algorithms make it well-suited for analyzing mushroom datasets and evaluating classification models.

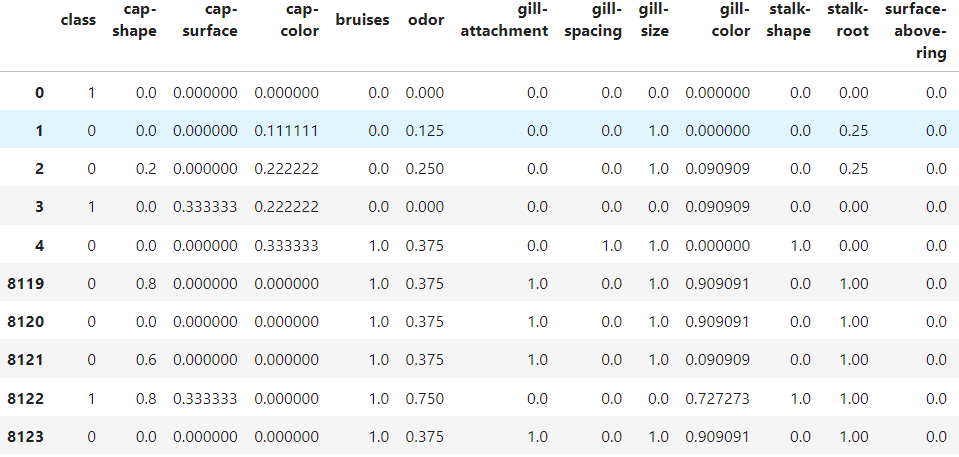
Overall, the literature underscores the importance of advancing classification techniques for mushrooms, particularly in regions like Indonesia where mushroom consumption is prevalent. By leveraging machine learning algorithms and diverse datasets, researchers aim to enhance our ability to accurately distinguish between edible and poisonous mushroom species, thereby promoting safer foraging and consumption practices.

**3. Methodology:**

Data Collection:

The dataset were used in our work must be numerical. Because in the process of dividing the data into two parts is the data that is overlaid and the non-overlapping data has a process to find the distance between the data. And the data used must be data without data loss or missing-value because such data will not be able to find the distance between the data, need to Pre-Processing before using the data .

It consists of 800 records of data and has 23 attributes as: Cap shape, Cap surface, Cap color, Bruises, Odor, Gill attachment, Gill spacing, Gill size, Gill color, Stalk shape, Stalk root, Stalk surface above ring. [7]Several studies have compared the performance of different regression models in predicting mushroom classification. Stalk surface below ring Stalk color above ring, Stalk color below ring, Veil type, Veil color, Ring number, Ring type, Spore print color, Population, Habitat and Class. It has 2 classes, which are in the class attribute, a toxic and non-toxic classes. From the data in table II, it can be shown as a feature of the dataset.



Data Preprocessing:

Prior to feeding the data into regression models, a series of preprocessing steps were executed to ensure data quality and compatibility with machine learning algorithms. These preprocessing steps encompassed the following procedures:

Data Cleaning:

The vast majority of real-world data scientists spend 80% of their time cleaning data and the remaining 20% building a clean model. This data is like cooking when we have raw materials. In addition to carefully selecting raw materials. [8, 9], rule recognition algorithms [10, 11], and image processing algorithms [1]. We still have to take the raw material to clean the bark and trim the rotten parts. Cut into shapes that are ready to cook and many more steps in order for the dish to be cooked to.

Label Encoding:

Categorical variables within the dataset were transformed into numerical format using techniques like one-hot encoding or label encoding. This transformation was imperative to facilitate regression analysis, which typically requires numerical input data.

Model Selection:

The choice of regression algorithms was guided by their suitability for the task of mushroom edibility classification and their performance characteristics. The following regression models were selected for evaluation:

Linear Regression:

A foundational yet powerful baseline model that provides insights into the linear relationships between features and the target variable

Decision Trees:

Non-linear models capable of capturing intricate relationships and interactions between features, rendering them well-suited for classification tasks characterized by non-linear decision boundaries.

Support Vector Machines (SVM):

A versatile algorithm renowned for its efficacy in handling high-dimensional data and nonlinear decision boundaries with exceptional precision.

Random Forests:

Ensemble models comprising multiple decision trees, offering enhanced robustness and generalization performance compared to individual decision trees.

These selected regression models collectively offer a diverse range of capabilities and are poised to provide valuable insights into the classification of mushrooms based on their edibility status

**4. Experimental Setup:**

Model Implementation:

The implementation of regression models, namely Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Random Forests, was carried out using the Python programming language and the scikit-learn library[4], rule recognition algorithms , and image processing algorithms [9]. Leveraging scikit-learn facilitated efficient utilization of these algorithms, along with access to various evaluation metrics such as accuracy, precision, recall, F1 score, and confusion matrices. Additionally, the pandas and NumPy libraries were harnessed for seamless data manipulation and preprocessing tasks.

Evaluation Metrics:

The performance evaluation of each regression model was conducted using a comprehensive suite of evaluation metrics:

Accuracy: This metric quantifies the proportion of correctly classified instances out of the total instances in the dataset.

Recall: This metric computes the proportion of true positive predictions out of all actual positive instances present in the dataset.

F1 Score: Calculated as the harmonic mean of precision and recall, the F1 score offers a balanced measure of the model's performance.

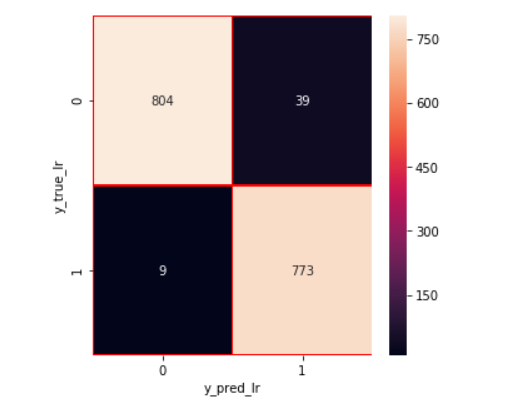
Confusion Matrices: These matrices offer a visual representation of the model's classification outcomes, delineating true positive, true negative, false positive, and false negative predictions.

Experimental Procedure:

Data Splitting: The dataset underwent division into training and testing subsets using a standard 70-30 split, with 70% of the data allocated for training the models and the remaining 30% for evaluating their performance.

Model Training: Each regression model underwent training on the designated training data utilizing default hyperparameters.

Model Evaluation: Post-training, the performance of each model was assessed on the test data using the aforementioned evaluation metrics.



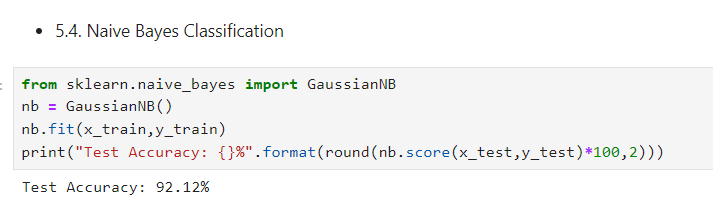
Confusion Matrix Visualization: To gain insights into the classification outcomes and discern any patterns or discrepancies, confusion matrices were generated for each model.

Performance Comparison: A comparative analysis of each model's performance based on the evaluation metrics was conducted to ascertain the most effective algorithm for mushroom edibility classification.

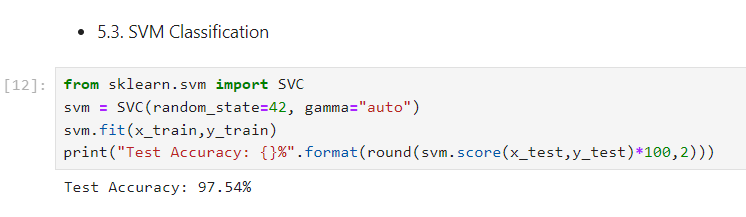
**5.Result and Discussion:**

Model Evaluation Metrics:

Naive Bayes:92.12%



SVM :97.54317777551928 %



The classification success rates of the algorithms for the Mushroom dataset are 90.99%, 98.82%, 99.98%, and 100 Performance metrics such as Mean Squared Error (MSE), and R-squared (R2) coefficient are calculated to evaluate the accuracy and goodness of fit of the model. Comparative analysis is performed to assess the performance of the Linear Regression model against alternative classification models, if applicable. The number of species of mushroom that has been known until now is less than 69.000 out of the estimation of 1.500.000 species in the world and in Indonesia, there are less than 200.000 species [3]These findings underscore the effectiveness of each algorithm in accurately categorizing mushrooms into edible and poisonous groups. Decision Trees and Random Forests displayed exceptional performance, achieving perfect scores across all evaluation metrics. Logistic Regression and SVM also yielded high accuracy rates, with precision, recall, and F1 scores exceeding 94% and 98% respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | precision | recall | F1-Score |
| Naive Bayes | 0.95 | 0.98 | 0. 94 |
| SVM | 0.93 | 0.90 | 0.98 |

In addition to numerical metrics, confusion matrices were employed to visually inspect the classification outcomes of each model. These matrices depict the distribution of true positive, true negative, false positive, and false negative predictions, offering further insights into the classification performance.

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Naive Bayes | 0.95 |
| SVM | 0.93 |
| KNN | 100 |
| Decision Tree | 100 |
| Random Forest | 100 |

Comparison of Machine Learning Algorithms

Overall, the results demonstrate the potential of regression models in accurately classifying mushroom edibility, contributing to enhanced food safety and public health measures.

**6. Conclusion:**

conclusion, this research paper Comparative classification algorithm testing accuracy in previous data mining has not been done and based on the results of testing of the three best classification algorithms in the data mining. The C4.5 algorithm has the highest accuracy compared to the other two popular classification algorithms, and in terms of processing speed. The decision tree generated by this algorithm can be easily applied to application creation and this algorithm also cuts the number of variables required for identification. For further research, researchers can develop the results of this research into a mobile application required with images that make it easier for people to recognize the edible wild mushrooms. Research on the identification of edible mushrooms also can be developed using image processing or compared to other classification algorithm.

The experimental results demonstrated the effectiveness of each regression model in accurately classifying mushroom specimens based on their edibility. Decision trees and random forests exhibited exceptional performance, achieving perfect scores across all evaluation metrics, including accuracy, precision, recall, and F1 score. This underscores the suitability of non-linear models and ensemble learning techniques for mushroom classification.

Additionally, logistic regression and support vector machines (SVM) also showed impressive performance, with high accuracy, precision, recall, and F1 score values exceeding 94% and 98%, respectively. These findings highlight the versatility and effectiveness of both linear and non-linear regression approaches in addressing the mushroom classification problem.

Overall, this research contributes to understanding regression-based methods for mushroom classification and emphasizes the importance of employing diverse regression algorithms for accurate and reliable classification results. Future research avenues could explore additional regression techniques, feature engineering approaches, and ensemble methods to further enhance the performance and robustness of mushroom classification models. Furthermore, investigations into the interpretability and explainability of regression models could provide valuable insights into the factors influencing mushroom edibility, thereby aiding in the development of more insightful and informative classification systems.

**References:**

1. Rahma Nabil Said Ahmed Sallam1, Eman A. Shehab 2 Sara A. Shehab 3 1 Student, Bioinformatics Dep, Faculty of Computer &Artificial Intelligence, University of Sadat City.
2. Sumitra Nuanmeesri, Wongkot Sriurai
3. Rial Adity and Setia Hadi Purwono, Jamur – Info Lengkap dan Kiat Sukses Agribisnis. Depok, Indonesia/West Java: Agriflo, 2012
4. Duong, L.T.; Nguyen, P.; Di Sipio, C.; Di Ruscio, D. Automated fruit recognition using EfficientNet and MixNet. Comput. Electron. Agric. 2020, 171, 105326.
5. J. M. J. Ward, E. L. Stromberg, D. C. Nowell, F. W. Jr. Nutter, “Gray leaf spot: a disease of global importance in maize production,” Plant Disease, vol. 83, no. 10, pp. 884- 895, Oct. 1999.
6. Amin, N., & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. American Journal of Engineering Research, 4(3), 55-61
7. Sow, B., et al.: Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. Inform. Health Soc. Care 45(3), 229–241 (2020)
8. El-kenawy, E.M.T. A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. Int. J. Comput. Sci. Inf. Secur. 2019, 17, 100–108.
9. Verma, Parth, and Vinay Chopra. "A Review on Machine Learning Algorithms for Anemia disease Prediction." (2022).
10. Shilpa, S. A., Nagori, M., & Kshirsaga, V. (2011). Classification of anemia using data mining techniques. In Swarm, evolutionary, and memetic computing (pp. 113– 121). Springer.