

From Spore to Score: Performance Analysis of Naive Bayes and K-Nearest Neighbors on the UCI Mushroom Classification Dataset

Melca M. Abogado¹, Gardenia B. Concillo¹, Hazel D. Bantugan¹,
Jose C. Agoylo Jr.¹, Jimson A. Olaybar², Joedee Mark Rodriguez²

¹ BSIT Department, Southern Leyte State University – Tomas Oppus Campus, Southern Leyte, Philippines

² FCIS, Southern Leyte State University – Main Campus, Sogod, Southern Leyte, Philippines

Abstract - This study used two machine learning algorithms (Naive – Bayes & K-Nearest Neighbors (KNN)) to predict whether a mushroom is edible or not, based on its morphological features. The study attempts to use a public mushroom dataset containing 8,124 instances and 22 categorical attributes to determine the model that can show more accuracy and reliability in the classification of mushrooms as edible or poisonous. Pre-processed data was used, and the results showed that the KNN classifies with an accuracy of 99.63% whereas the Naive Bayes classifies with an accuracy of 92.18%, which is still a good result, although lower than that of KNN. These results indicate that KNN performs effectively on the selected benchmark dataset and may serve as a useful baseline approach for future mushroom classification studies.

Keywords -mushroom classification, Naive Bayes, K-Nearest Neighbors (KNN), predictive modeling, edibility prediction, Machine learning.

DOI: 10.18421/SAR92-07

<https://doi.org/10.18421/SAR92-07>

Corresponding author: Jose C. Agoylo Jr.,
BSIT Department, Southern Leyte State University – Tomas Oppus Campus, Southern Leyte, Philippines

Email: jagoylo@southernleytestateu.edu.ph

Received: 05 May 2026

Revised: 17 June 2026.

Accepted: 22 June 2026.

Published: 27 June 2026.

 ©2026 Melca M. Abogado et al.; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License.

The article is published with Open Access at <https://www.sarjournal.com/>

1. Introduction

It is important to determine the edibility of a mushroom, as consuming poisonous mushrooms can cause many health issues and may even be fatal. Mushroom identification is traditionally done by experts or by relying on physical features. This may be arduous for the common people, however, and at times result in mis-classification. This makes machine learning an attractive tool to enhance mushroom classification accuracy and efficiency. Machine learning algorithms have been used in various applications, such as agriculture and environmental studies, and they have demonstrated their effectiveness in prediction and classification tasks in these areas [5], [8].

The reviewed studies collectively demonstrate the growing application of artificial intelligence, machine learning, and computer vision in classification, identification, and quality assessment tasks across various domains. [1] highlighted how computer vision and machine learning technologies have significantly improved mushroom cultivation, species recognition, disease detection, and quality evaluation through automated image analysis. Similarly, [4] showed that deep learning models can effectively extract complex leaf features and achieve high accuracy in plant classification, reducing dependence on manual feature engineering. In the agricultural sector, [6] discussed the use of optical sensing and nondestructive inspection technologies for automated fruit and vegetable grading, emphasizing their role in improving efficiency and product quality. Beyond agriculture, [7] demonstrated the effectiveness of the Random Forest algorithm in accurately classifying neuroimaging data for Alzheimer's disease diagnosis, highlighting its robustness in handling complex datasets.

Furthermore, [11] introduced the Positive and Unlabeled (PU) learning approach, which enables effective classification even when only limited labeled data are available, making it valuable for real-world applications where complete datasets are difficult to obtain.

The mushroom dataset used in this study was downloaded from Kaggle, which was originally from the UCI Machine Learning Repository. It has 22 mushroom properties (cap shape, cap colour, gill size, smell, etc.) which were used to identify edible and poisonous mushrooms [12]. It is a widely-used dataset in the field of machine learning, where it is used as a benchmark for evaluating and comparing classification algorithms [9], [10].

In this study the two widely used machine learning algorithms, Naive Bayes and K-Nearest Neighbours (KNN) are compared. Naive Bayes is dubbed as the simple probabilistic approach, and KNN is also well known for its similarity among instances classification. Such algorithms have been used to tackle different predictive and classification problems successfully, such as in [5] and [8]. Estimation of tree volume using forest health indicators was also used [15] and applied similar machine learning methods.

Although the UCI Mushroom dataset has been extensively used in machine learning research, comparative evaluations of simple baseline classifiers remain valuable for educational purposes and for establishing benchmark performance levels. This study focuses on examining the behavior of probabilistic and instance-based learning approaches under identical experimental conditions and discusses their suitability and limitations for mushroom edibility classification.

2. Methodology

The process of comparing the performance of the Naive Bayes and the K-Nearest Neighbours (KNN) algorithm for the classification of edibility of mushrooms (edible vs. poisonous) is explained in this section. The methodology consists of data collection, data preprocessing, model training and model evaluation, as shown in Figure 1.

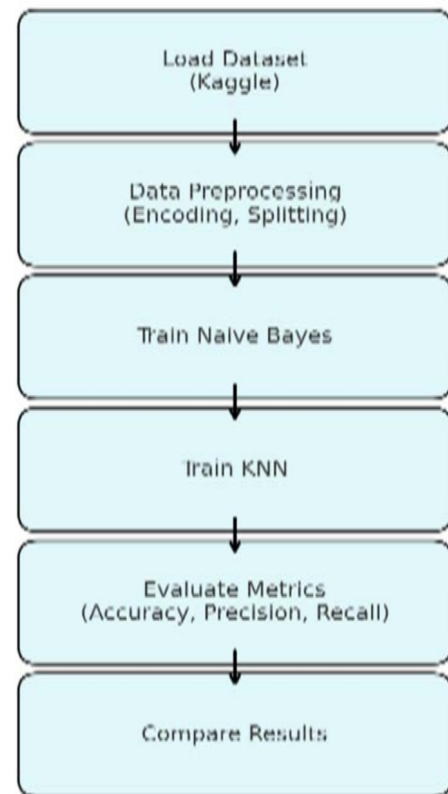


Figure 1. Workflow of Mushroom Edibility Classification Using Machine Learning

2.1. Data Collection

In this research, the dataset is downloaded from Kaggle and obtained from UCI Machine Learning Repository [12]. Its 8,124 mushroom samples are categorized by 22 features, such as cap shape, cap color, odor, gills characteristics, and spore print color. Every record is marked e (for edible) or p (for poisonous). Due to its completeness and reliability, it is widely used in past mushroom classification research [9] and [10]. Figure 2 shows the distribution of mushrooms by the classes is relatively even, with some being edible and others poisonous.

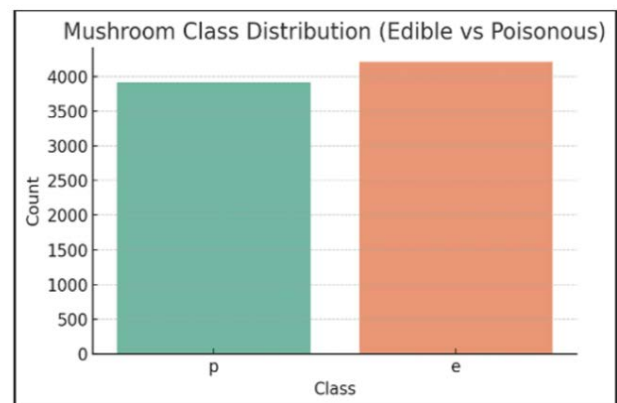


Figure 2. Class Distribution of Mushrooms

2.2. Data Preprocessing

Table 1 shows a representation of the data set. All the categorical attributes were encoded into numbers prior to model training. This step enabled the machine learning algorithms to efficiently process the data.

There were no missing data in the dataset, so no further data cleaning procedures were required. The encoded data was split into a training set (70%) and a testing set (30%). There were a total of 5,687 records for training and 2437 records for testing and evaluation.

Table 1. Sample Entries from the Mushroom Dataset

class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ground
p	x	s	n	t	p	f	c	n	k	e	e	s
e	x	s	y	t	a	f	c	b	k	e	c	s
e	b	s	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	n	e	e	s
e	x	s	g	f	n	f	w	b	k	t	e	s
e	x	y	y	t	a	f	c	b	n	e	c	s
e	b	s	w	t	a	f	c	b	g	e	c	s
e	b	y	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	p	e	e	s
e	b	s	y	t	a	f	c	b	g	e	c	s
e	x	y	y	t	l	f	c	b	g	e	c	s
e	x	y	y	t	a	f	c	b	n	e	c	s
e	b	s	y	t	a	f	c	b	w	e	c	s
p	x	y	w	t	p	f	c	n	k	e	e	s
e	x	f	n	f	n	f	w	b	n	t	e	s
e	s	f	g	f	n	f	c	n	k	e	e	s
e	f	f	w	f	n	f	w	b	k	t	e	s
p	x	s	n	t	p	f	c	n	n	e	e	s
p	x	y	w	t	p	f	c	n	n	e	e	s
p	x	s	n	t	p	f	c	n	k	e	e	s
e	b	s	y	t	a	f	c	b	k	e	c	s

2.3. Algorithms Used

• Naive Bayes

Categorical Naive Bayes (CategoricalNB) was employed because the mushroom dataset consists entirely of categorical attributes. Unlike Gaussian Naive Bayes, which assumes continuous normally distributed features, CategoricalNB estimates class probabilities directly from the frequency of categorical feature values. The classifier applies Bayes' theorem under the assumption of conditional independence among features and predicts the class with the highest posterior probability. Laplace smoothing was used to prevent zero-probability estimates for unseen feature-category combinations. The probabilistic approach to Naive Bayes is based on the principles of statistical communication and information theory developed [14].

Categorical Naive Bayes Formula: Categorical Naive Bayes is designed for discrete categorical features and is based on Bayes' Theorem:

$$P(C_k | \mathbf{x}) = \frac{P(C_k)P(\mathbf{X} \setminus C_k)}{P(\mathbf{X})} \quad (1)$$

• K-Nearest Neighbors (KNN)

KNN is an instance-based learning algorithm that classifies the data based on the majority class of its KNN. [13] first introduced the method and it is still widely used in pattern recognition and classification problems.

In this study, the value of $k = 5$, and Euclidean distance was used:

$$d(\mathbf{x}', \mathbf{x}_i) = \sqrt{\sum_{j=1}^n (x'_j - x_{ij})^2} \quad (2)$$

The predicted class $\hat{y}$ is the most frequent among the k nearest neighbors:

$$\hat{y} = \text{mode}([y_1, y_2, \dots, y_k]) \quad (3)$$

Unlike Naive Bayes, KNN does not assume any underlying distribution and adapts well to complex boundaries in the dataset.

2.4. Model Evaluation

Several evaluation metrics were used to assess the models performance. To measure the overall correctness of the prediction, accuracy was used, and to measure the quality of the positive classifications, precision and recall were used. To get a balance of precision and recall, the F1-score was added. Further, confusion matrices were created to analyse the classification errors, and to detect the correct and incorrect prediction of the instances. The experiments were performed with Python and Scikit-learn library.

2.5. Ethical Consideration

This research only used data that was opened to the public on Kaggle, which came from the UCI Machine Learning Repository. The data is anonymous and synthetic and does not include any personal details or sensitive characteristics. Hence, no human subjects were used in this study and there was no need for ethical clearance.

Furthermore, all machine learning models were created and tested for academic and research work only. Implementation is based on ethical principles in handling data, transparency and reproducibility, and the responsible use of publicly available data. An analysis is carried out without any manipulation of data or amplification of bias, with the aim of helping to design safe and intelligent systems for food classification.

3. Results

The classification accuracy of Naive Bayes and K-Nearest Neighbours (KNN) algorithms on the mushroom dataset are discussed in this section. Both models were trained using 70% of the data and tested using 30% of the data. They were evaluated using the metrics of accuracy, precision, recall, F1-score and ROC AUC.

- **Naive Bayes Results:** The Naive Bayes Classifier gave an accuracy of 92.18%. The mushroom classification is summarized by its performance in edible and poisonous mushrooms in Table 2. It showed good precision and recall, but feature independence resulted in some misclassifications.
- **K-Nearest Neighbors (KNN) Results:** The classification accuracy for KNN was 99.63% which was a lot more accurate than Naive Bayes. Precision and recall were very high, showing its effectiveness in dealing with this kind of categorical data. The model was able to accurately classify nearly all instances.

Table 2. Performance Comparison of Naive Bayes and K-Nearest Neighbors

Algorithm	Accuracy	Precision (Class 0)	Recall (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Weighted)
Naive Bayes	92.18%	0.93	0.91	0.91	0.93	0.92
K-Nearest Neighbors	99.63%	1.00	0.99	0.99	1.00	1.00

Class 0 = Poisonous, Class 1 = Edible

Table 2 shows the comparative performance metrics for both algorithms used in this study using Class 0 as poisonous and Class 1 as edible.

Figure 3 shows the understanding of the models' classification capabilities, the confusion matrix for Naive Bayes and K-Nearest Neighbors (KNN).

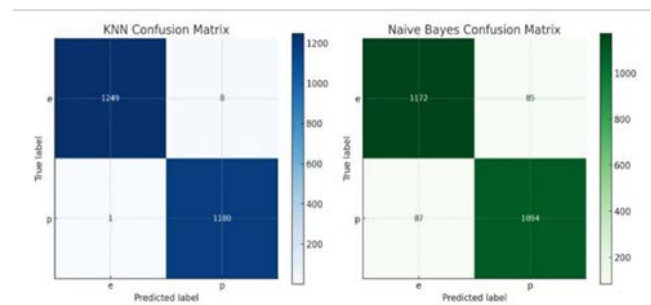


Figure 3. Confusion matrices for Naive Bayes and KNN classifiers

For the Naive Bayes model, there were 1172 poisonous mushrooms correctly classified out of 2401 test instances which were classified as edible, and 1094 edible mushrooms correctly classified with 87 misclassified as poisonous.

In contrast, the KNN model correctly classified 1249 poisonous mushroom and 1180 edible mushroom with only 9 incorrect classifications.

The confusion matrix analysis indicates that KNN has much lesser false positives and false negatives as compared to Naive Bayes. This supports the increased reliability of KNN in applications such as mushroom edibility determination, where the accuracy is crucial.

Figure 4 shows the effectiveness of each model to differentiate between classes. The ROC (Receiver Operating Characteristic) curves in KNN also obtained the AUC value of 1.00 (perfect separation) and Naive Bayes obtained the AUC value of 0.95. Note that KNN showed perfect classification (AUC = 1.00), and Naive Bayes was also good (AUC = 0.95).

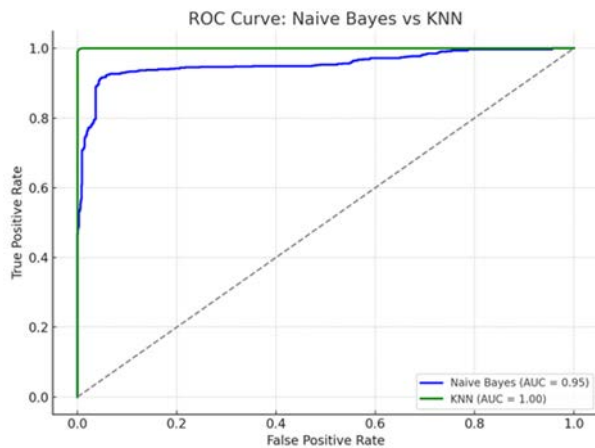


Figure 4. ROC Curve Comparison between Naive Bayes and K-Nearest Neighbors Comparison Result

Table 3. Comparison results between studies

Study	Dataset	Algorithm	Accuracy
Arslan et al. (2024)	UCI Mushroom	RF	100%
Ortega et al. (2020)	UCI Mushroom	KNN	99.5%
Proposed Study	UCI Mushroom	KNN	99.63%

Table 3 shows that several studies have evaluated machine learning algorithms for mushroom edibility classification using the widely adopted UCI Mushroom dataset.

4. Discussion

The findings of this study showed that there was a significant difference between the ability of Naive Bayes and K-Nearest Neighbors (KNN) in classing the edibility of mushrooms. However, even though Naive Bayes had a decent accuracy score of 92.18%, KNN had attained a higher accuracy of 99.63% and the values of precision, recall and F1 score were also better.

The lower performance of Naive Bayes may be due to the fact that it assumes that all features are independent. This is not necessarily true for the mushroom data set; several attributes of mushrooms may be correlated with each other. This will consequently make it difficult for the algorithm to grasp the relationships between attributes. [2] also reported similar findings that Naive Bayes may not be as effective in the presence of feature dependencies.

Unlike KNN, there is no assumption about the distribution of data in KNN. Rather, it labels samples based on their resemblance to other samples. This flexibility enables KNN to fit the data set.

This study's results are in accordance with the previous studies which showed that KNN is a good classification method in classification problems of categorical and high dimensional data [3], [9] and [10].

The ROC-AUC (receiver operating characteristic-area under the curve) analysis also showed that KNN was effective. The algorithm did well with an AUC of 1.00, and Naive Bayes came in at an AUC of 0.95. From this finding, it can be seen that KNN can classify poisonous mushrooms and edible mushrooms with higher accuracy, particularly in the application where the classification error has serious consequences.

For instance, [9] reported that ensemble-based methods, particularly Random Forest (RF), achieved perfect classification performance with an accuracy of 100%, demonstrating the effectiveness of tree-based learning techniques in handling categorical mushroom attributes. Similarly, [10] conducted a comparative analysis of classification algorithms and found that the K-Nearest Neighbors (KNN) algorithm achieved an accuracy of 99.5%, highlighting its capability to classify mushroom samples based on feature similarity. Consistent with these findings, the present study obtained an accuracy of 99.63% using KNN, slightly outperforming the result reported [10]. Although the performance of KNN remains highly competitive, the similarity of results across studies suggests that the UCI Mushroom dataset has become a mature benchmark problem where several machine learning algorithms can achieve near-perfect classification accuracy. Therefore, future research should focus on more challenging datasets, real-world mushroom samples, image-based classification, and explainable artificial intelligence techniques to enhance the practical applicability of mushroom edibility prediction systems [9], [10]. The results indicate the effectiveness and high level of accuracy of the mushroom classification using the KNN technique. The results demonstrate the effectiveness of KNN within the context of the UCI Mushroom dataset. However, further validation using real-world mushroom samples and image-based datasets is required before practical deployment can be considered.

Limitation of the Study: This study has several limitations. First, the experiments were conducted using a single publicly available benchmark dataset that may not fully represent the variability of mushrooms encountered in natural environments. Second, all predictor variables were structured morphological attributes rather than images or field observations. Third, only two machine learning algorithms were evaluated.

Finally, the use of encoded categorical variables with Gaussian Naive Bayes introduces methodological limitations, as the algorithm assumes continuous feature distributions. Therefore, the findings should be interpreted as a comparative benchmark study rather than evidence of practical mushroom identification capability.

5. Conclusion and Recommendation

It was found from the results that KNN performed better than Naive Bayes on all the important evaluation parameters. Naive Bayes obtained an accuracy of 92.18%, and KNN a 99.63% as well as a perfect ROC-AUC score. The results have shown that KNN can better deal with complex relationships between mushroom attributes and more reliable classification results.

KNN demonstrated superior performance on the benchmark dataset evaluated in this study. Nevertheless, the findings should not be interpreted as evidence of reliable mushroom identification in natural environments, since the experiments were conducted exclusively on a structured public dataset. Real-world deployment and optimization of these models in mobile and edge computing scenarios could also be investigated. For additional improvement of the same study, it is recommended to increase its scientific value by adding multiple algorithms for comparison, such as decision tree, random forest, SVM and XGBoost.

References:

- [1]. H. Yin, W. Yi, and D. Hu, "Computer vision and machine learning applied in the mushroom industry: A critical review," *Computers and Electronics in Agriculture*, vol. 198, Art. no. 107015, 2022. doi: 10.1016/j.compag.2022.107015.
- [2]. Y. Dou and W. Meng, "Comparative analysis of weka-based classification algorithms on medical diagnosis datasets," *Technology and Health Care*, vol. 31, no. 1_suppl, pp. 397–408, 2023. doi: 10.3233/THC-236034.
- [3]. Y. J. Sheela and S. H. Krishnaveni, "A comparative analysis of various classification trees," in *Proc. 2017 Int. Conf. Circuit, Power and Computing Technologies (ICCPCT)*, 2017, pp. 1–8. doi: 10.1109/ICCPCT.2017.8074403.
- [4]. S. H. Lee, C. S. Chan, S. J. Mayo, and P. Remagnino, "How deep learning extracts and learns leaf features for plant classification," *Pattern Recognition*, vol. 71, pp. 1–13, 2017. doi: 10.1016/j.patcog.2017.05.015.
- [5]. X. Chen and B. Jiang, "Analysis of the application of machine learning algorithms in data mining," in *J. Phys.: Conf. Ser.*, vol. 1550, no. 3, Art. no. 032039, 2020. doi: 10.1088/1742-6596/1550/3/032039.
- [6]. B. Zhang, B. Gu, G. Tian, J. Zhou, J. Huang, and Y. Xiong, "Challenges and solutions of optical-based nondestructive quality inspection for robotic fruit and vegetable grading systems: A technical review," *Trends in Food Science & Technology*, vol. 81, pp. 213–231, 2018. doi: 10.1016/j.tifs.2018.09.018.
- [7]. A. Sarica, A. Cerasa, and A. Quattrone, "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review," *Frontiers in Aging Neuroscience*, vol. 9, Art. no. 329, 2017. doi: 10.3389/fnagi.2017.00329.
- [8]. A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021. doi: 10.1109/ACCESS.2020.3048415.
- [9]. M. Arslan, M. Azam, M. Ali, M. U. Hashmi, A. Kousar, and Z. Zafar, "A comparative study of machine learning methods for optimizing mushroom classification," *Journal of Computing & Biomedical Informatics*, vol. 8, no. 1, 2024.
- [10]. J. H. J. C. Ortega, A. C. Lagman, L. R. Q. Natividad, E. T. Bantug, M. R. Resurreccion, and L. O. Manalo, "Analysis of performance of classification algorithms in mushroom poisonous detection using confusion matrix analysis," *International Journal*, vol. 9, no. 1.3, 2020. doi: 10.30534/ijatcse/2020/7191.32020.
- [11]. B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proc. 3rd IEEE Int. Conf. Data Mining (ICDM)*, 2003, pp. 179–186. doi: 10.1109/ICDM.2003.1250918.
- [12]. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017.
- [13]. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. doi: 10.1109/TIT.1967.1053964.
- [14]. C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.
- [15]. M. Abogado, R. Acaso, J. Olaybar, J. Tagud, and A. Bacalla, "Gradient-Boosted Machine Learning Models for Tree Volume Estimation Using Forest Health Indicators," *Insights in Computer Science*, vol. 1, pp. 26–33, 2025.