DOI: http://dx.doi.org/10.26483/ijarcs.v13i3.6830

ISSN No. 0976-5697

Volume 13, No. 3, May-June 2022

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Edibility detection of mushroom using Logistic Regression and PCA

Sai Charan Gangu
Dept of Computer Science Engineering
ANITS
Visakhapatnam, India

Madhu Nitesh Bandi
Dept of Computer Science Engineering
ANITS
Visakhapatnam, India

Dr Sangeetha Viswanadham
Professor, Dept of Computer Science Engineering
ANITS
Visakhapatnam, India

Chintala Chandrasekhar Sivaji
Dept of Computer Science Engineering
ANITS
Visakhapatnam, India

Toyaka Sai Kiran
Dept of Computer Science Engineering
ANITS
Visakhapatnam, India

Abstract: Mushroom is found to be one of the best nutritional foods with high proteins, vitamins and minerals. Only some of the mushroom varieties were found to be edible. Some of them are dangerous to consume. To distinguish between the edible and poisonous mushrooms, we use machine learning algorithms to classify them. Classification is performed using various machine learning classifiers and Logistic regression showed better results compared to other algorithms. A survey of various algorithms resulted in KNN giving an accuracy of 100% at k=1 using 800 samples. A change k value is leading to a decrease in accuracy. By using hybrid algorithms (i.e., using two or more algorithms) which includes a combination of dimensionality reduction techniques such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) along with existing classifiers better performance is achieved. Logistic Regression along with Principal Component Analysis is used to increase the accuracy. The results are shown in form of bar plots.

Keywords: Mushrooms, Poisonous, Classification, Logistic Regression, Linear Discriminant Analysis, Principal Component Analysis, Bar plot.

I. INTRODUCTION

[1],[2],[3] Mushroom is well-liked food rich in vitamins and minerals, normally good for many reasons and also helps in killing cancer cells. It contains antioxidants that prevent people from heart disease. Around 45000 species of mushrooms are found to be existing worldwide. And some of the mushrooms have poisonous properties which can result in death. To classify mushrooms into poisonous or non-poisonous, the existing system has used machine learning algorithms like Naive Bayes, SVM, KNN and other existing classifiers but the accuracies are observed to be less. Therefore, it is proposed to use Dimensionality reduction techniques to extract prominent features from a multi-dimensional dataset and then applying Machine Learning classifiers which produces better results and also improve the accuracy and efficiency of existing classification models.

[4],[5],[6],[7]In the existing system, the accuracies are below 90% for most of the classification algorithms only few algorithms namely K-Nearest Neighbours (100%), Naïve Bayes (91.5%) and Logistic Regression (94.91%) resulted in accuracy more than 90%. The KNN algorithm is 100% accurate at k=1, but having k value as 1 lead to overfitting of the system and a change in k value is leading to a decrease in accuracy. On applying Naïve Bayes, Logistic regression with LDA, only a slight increase in accuracy is observed. By applying Logistic regression along with PCA better accuracy is recorded compared to the existing models.

II. METHODOLOGY

[8],[9],[10] We used methodologies like Linear Discriminant analysis (LDA), Principal Component analysis (PCA) for feature extraction (i.e., to extract prominent features from large datasets) and Logistic Regression for classification. The dataset consists of 8124 instances with 23 attributes each which are divides into 2 classes (edible-e, poisonous-p). Before Classification, all the data in the dataset is of object type but machine learning algorithms accept numerical data. In order to convert object data in numerical data Label Encoder is used.

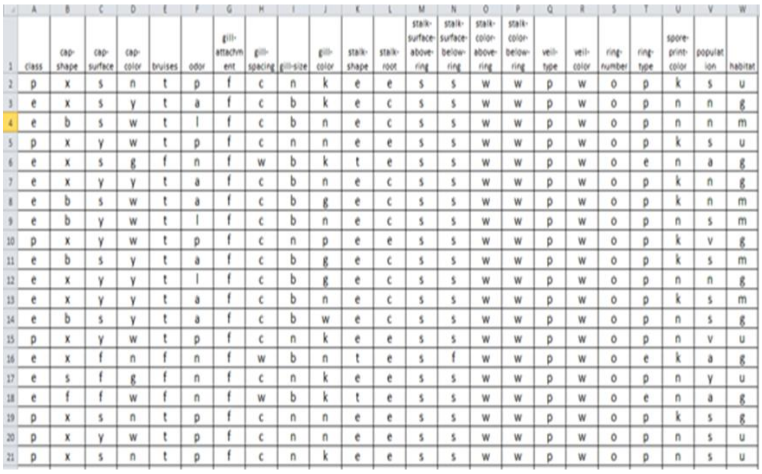
A. *Encoding the data into numerical values*

1. Input Data: The dataset contains 23 features for each mushroom type and their variants. The features differ for different types of mushrooms. The dataset contains data in the short hand notations and the shortcut notations are mentioned in table 1.

Table 1: Dataset description

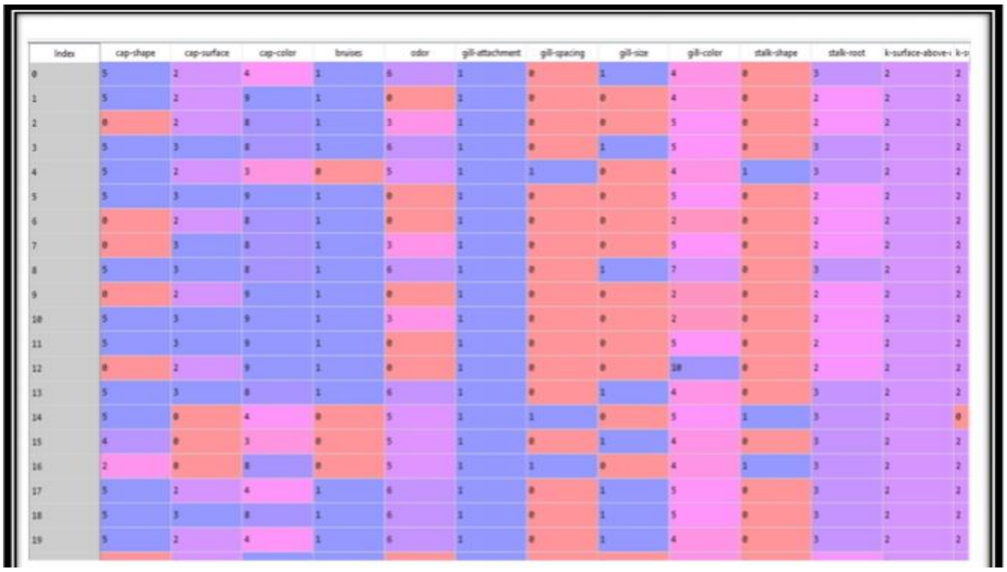
S.No	***Input Variable***	***Domain***
1	cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2	cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
3	cap-color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4	Bruises	bruises=t, no=f
5	Odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6	gill-attachment	attached=a, descending=d, free=f, notched=n
7	gill-spacing	close=c, crowded=w, distant=d
8	gill-size	broad=b, narrow=n
9	gill-color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
10	stalk-shape	enlarging=e, tapering=t
11	stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12	stalk-surface-above-ring	fibrous=f, scaly=y, silky=k, smooth=s
13	stalk-surface-below-ring:	fibrous=f, scaly=y, silky=k, smooth=s
14	stalk-color-above-ring:	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15	stalk-surface-below-ring:	fibrous=f, scaly=y, silky=k, smooth=s
16	veil-type	partial=p, universal=u
17	veil-color	brown=n, orange=o, white=w, yellow=y
18	ring-number	none=n, one=o, two=t
19	ring-type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20	spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21	Population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22	Habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

2. Dataset Acquisition and cleaning: Dataset is the collection of data related information. The mushroom dataset is obtained from Kaggle and the data contains 8124 instances with 23 attributes. All of the data is used to show the performance of the suggested model.
   The obtained dataset is shown in Fig1. The attribute name veil-type has no effect in classification so the attribute is removed to increase the efficiency of the model.
3. Label encoding: As the machine learning models can only be trained and tested using numerical data, the object type data is converted into numerical data using Label encoder.
4. Label encoder accepts categorical data, hence the object type data is converted into categorical data and later converted into numerical data. The final data after encoding is shown in Fig 2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring	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
2	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
3	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
4	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
5	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
6	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
7	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	n	g
8	e	b	s	w	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	n	m
9	e	b	y	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	s	m
10	p	x	y	w	t	p	f	c	n	p	e	e	s	s	w	w	p	w	o	p	k	v	g
11	e	b	s	y	t	a	f	c	b	g	e	c	s	s	w	w	p	w	o	p	k	s	m
12	e	x	y	y	t	l	f	c	b	g	e	c	s	s	w	w	p	w	o	p	n	n	g
13	e	x	y	y	t	a	f	c	b	n	e	c	s	s	w	w	p	w	o	p	k	s	m
14	e	b	s	y	t	a	f	c	b	w	e	c	s	s	w	w	p	w	o	p	n	s	g
15	p	x	y	w	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	v	u
16	e	x	f	n	f	n	f	w	b	n	t	e	s	f	w	w	p	w	o	e	k	a	g
17	e	s	f	g	f	n	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	y	u
18	e	f	f	w	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
19	p	x	s	n	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	g
20	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	n	s	u
21	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	n	s	u

Fig 1: Dataset



Index	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	k-surface-above-	k-s
0	5	2	4	1	6	1	0	1	4	0	3	2	2
1	5	2	9	1	0	1	0	0	4	0	2	2	2
2	0	2	8	1	3	1	0	0	5	0	2	2	2
3	5	3	8	1	6	1	0	1	5	0	3	2	2
4	5	2	3	0	5	1	1	0	4	1	3	2	2
5	5	3	9	1	0	1	0	0	5	0	2	2	2
6	0	2	8	1	0	1	0	0	2	0	2	2	2
7	0	3	8	1	3	1	0	0	5	0	2	2	2
8	5	3	8	1	6	1	0	1	7	0	3	2	2
9	0	2	9	1	0	1	0	0	2	0	2	2	2
10	5	3	9	1	3	1	0	0	2	0	2	2	2
11	5	3	9	1	0	1	0	0	5	0	2	2	2
12	0	2	9	1	0	1	0	0	10	0	2	2	2
13	5	3	8	1	6	1	0	1	4	0	3	2	2
14	5	0	4	0	5	1	1	0	5	1	3	2	0
15	4	0	3	0	5	1	0	1	4	0	3	2	2
16	2	0	8	0	5	1	1	0	4	1	3	2	2
17	5	2	4	1	6	1	0	1	5	0	3	2	2
18	5	3	8	1	6	1	0	1	5	0	3	2	2
19	5	2	4	1	6	1	0	1	4	0	3	2	2

Fig 2: Dataset after Label encoding

B. *Dimensionality Reduction*

After encoding the data into numerical values using Label Encoder, Dimensionality reduction is a technique used to identify the prominent features by projecting higher dimensional data into a lower dimensional space. This process removes unnecessary features that do not impact the classification thereby decreasing the complexity and increasing the efficiency of the model which in turn increases the accuracy.

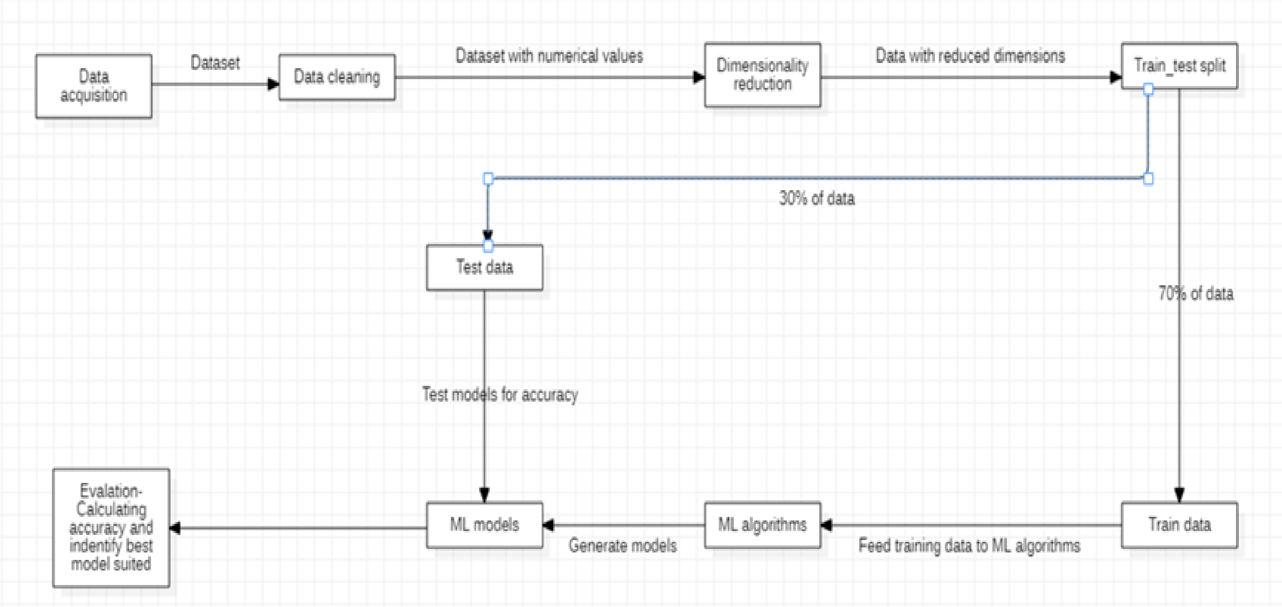
The dimensionality reduction techniques that are used in this system are:

1. Linear Discriminant analysis (LDA)
2. Principal Component analysis (PCA)

C. *Logistic Regression*

After completion of Dimensionality reduction, the Logistic Regression algorithm takes reduced data. It then performs the classification and classifies the data into two classes 0 -edible, 1 – poisonous.

III. SYSTEM ARCHITECTURE



IV. EXPERIMENTAL RESULTS &DISCUSSION

After performing logistic regression, LDA and PCA, the following are theexperiment values.

Table 2: Result of Logistic regression without dimensionality reduction techniques.

S.NO	***Training Size (%)***	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1_Score (%)***
01	60	94.35	95.34	95.03	95.18
02	70	94.91	95.20	94.24	94.72
03	80	94.76	94.39	94.75	94.57

Table-2 shows experimental results of logistic regression without linear discriminant analysis and principal component analysis. At 60% of training dataset and 40% of testing datathe accuracy is 94.35%. At 70% of training dataset and 30% of testing data, the accuracy is 94.91%. At 80% of training dataset and 20% of testing data, the accuracy is 94.76%. The highest accuracy usinglogistic regression classifier without dimensionalityreduction techniques is 94.76%.

Table 3: Result of Logistic regression with Linear Discriminant Analysis.

S.NO	***Training Size (%)***	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1_Score (%)***
01	60	95.13	95.43	94.46	94.94
02	70	94.95	95.52	93.98	94.75
03	80	94.58	94.71	93.98	94.35

Table-3 shows experimental results oflogistic regressionwith linear discriminant analysis. At 60% of training dataset and 40% of testing data, the accuracy is 95.13%. At 70% of training dataset and 30% of testing data, the accuracy is 94.95%. At 80% of training dataset and 20% of testing data, the accuracy is 94.58%. The highest accuracy in logistic regression classifier with linear discriminant analysis is 95.13%.

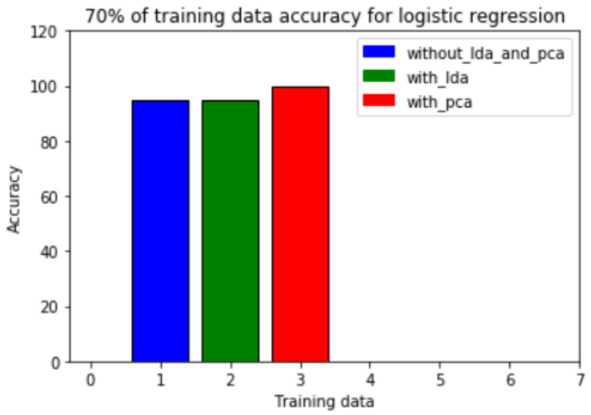
Table 4. Result of Logistic regression with Principal Component Analysis.

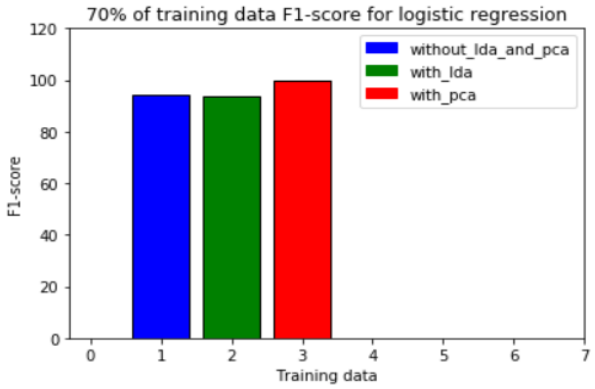
S.NO	***Training Size (%)***	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1_Score (%)***
01	60	99.47	99.24	99.68	99.46
02	70	99.67	99.66	99.66	99.66
03	80	99.56	99.36	99.74	99.55

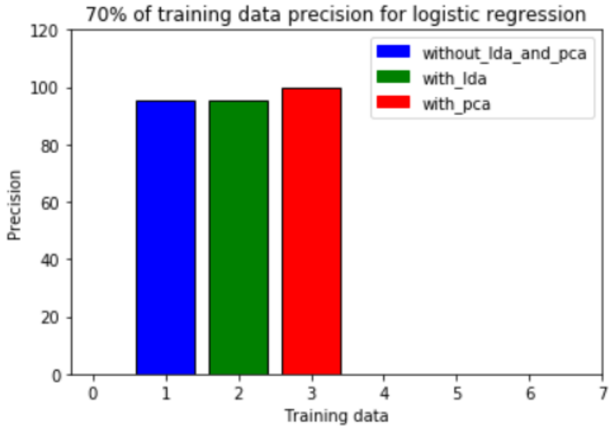
Table 4 shows experimental results of logisticregression with principal component analysis. At 60% of training dataset and 40% of testing data, the accuracy is 99.47%. At 70% of training dataset and 30% of testing data, the accuracy is 99.67%. At 80% of training dataset and 20% of testing data, the accuracy is 99.56%. The highest accuracy in Logistic regressionclassifierwith principalcomponent analysis is 99.56%.

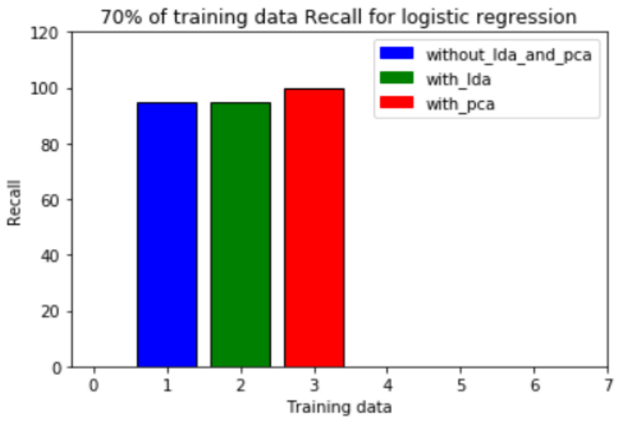
On comparing table 1, table 2 and table 3, it is clearly visible that, the accuracy of the Logistic Regression is increased in the presence of Dimensionality reduction techniques. At the 70% of training dataset, the accuracy of the Logistic Regression increased compared to other splitting ratios.

V. COMPARISON GRAPHS









The above graphs give the average of 70% training data and 30% testingdata.By comparing logistic regression with dimensionality reduction techniques, the average of logistic regression with PCA is greater than the average of Logistic regression without dimensionality reduction and with LDA.F1_Scorewithout dimensionality reduction is 94.72 and with LDA is 94.75 and with PCA is 99.66.Recall withoutdimensionality reduction is 94.24 and in with LDA is 93.98 and with PCA is 99.66. Precision without dimensionality reduction is 95.20 and with LDA is 95.52 and with PCA is 99.66. Accuracy without dimensionality reduction is 94.91 and with LDA is 94.95 and with PCA is 99.67.

VI. CONCLUSION

Classification using Logistic Regression along with Dimensionality reduction techniques like LDA(Linear Discriminant analysis) and PCA(Principal Component analysis) gave us an increase in accuracy. The accuracies recorded when applying Logistic Regression without dimensionality reduction, Logistic Regression with Linear Discriminant Analysis and Logistic Regression with Principal component analysis is 94.91%, 94.95% and 100% respectively. The accuracy of the Logistic regression is increased in the presence of LDA, but the difference in accuracy is almost negligible. When using Logistic Regression with PCA the model is 99.66% accurate with 8 principal components.

VII. REFERENCES

[1] Jong, S. & Birmingham, J. Medicinal benefts of the mushroom ganoderma. In Advances in Applied Microbiology, vol. 37, 101–134 (Elsevier, 1992).

[2] Lincof, G. Field Guide to North American Mushrooms. National Audubon Society (Alfred A. Knopf, 1997).

[3] Miles, P. G. & Chang, S.-T. Mushroom Biology: Concise Basics and Current Developments (World Scientifc, 1997).

[4]Ottom, Mohammad Ashraf. (2019). Classification of Mushroom Fungi Using Machine Learning Techniques. International Journal of Advanced Trends in Computer Science and Engineering,8(5),2378-2385. 10.30534/ijatcse/2019/78852019.

[5] Ottom, M. A., Alawad, N. A. & Nahar, K. M. Classifcation of mushroom fungi using machine learning techniques. Int. J. Adv. Trends Comput. Sci. Eng. 8, 2378–2385 (2019).

[6] Chelliah, B. J., Kalaiarasi, S., Anand, A., Janakiram, G., Rathi, B., & Warrier, N. K. (2018). Classification of Mushrooms using Supervised Learning Models. International Journal of Emerging Technologies in Engineering Research(IJETER),6(4).

[7] Verma, S. K., & Dutta, M. (2018). Mushroom classification using ANN and ANFIS algorithm. IOSR Journal of Engineering (IOSRJEN), 8(01), 94-100.

[8] Maurya, P. & Singh, N. P. Mushroom classifcation using feature-based machine learning approach. In Proceedings of 3rd International Conference on Computer Vision and Image Processing, 197–206 (Springer, 2020).

[9] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18, 1-8.

[10] Tony Cai, T., & Zhang, L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(4), 675-705.