JISA (Jurnal Informatika dan Sains)
Vol. 05, No. 02, December 2022

e-ISSN: 2614-8404
p-ISSN: 2776-3234

The Classification of Mushroom Types Using Naive Bayes and Principal Component Analysis

Deby Rianasari [1], Meina Noor Triana[2], Milla Rosiana Dewi[3], Yulia Astutik[4], Rio Wirawan5*
[1,2,3,4]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
[5]InformationSystem, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
Email : [1]debyr@upnvj.ac.id, [2]meinant@upnvj.ac.id, [3]millard@upnvj.ac.id, [4] yuliaa@upnvj.ac.id,
*Corresponding Author:
rio.wirawan@upnvj.ac.id

Abstract. Indonesia is one of tropical countries with high humidity and makes it possible for various plants and microorganisms to grow properly. One of the microorganisms that can grow well in Indonesia is mushroom. They have several types including poisonous and edible mushrooms that can be consumed by human beings.There was a previous research on a classification of mushroom species using the Naive Bayes method with the title "Implementation of the Naïve Bayes Classifier Method for Identification of Mushroom Types" and obtained pretty good accuracy results. This research conducted a classification using Naive Bayes which was improved by applying PCA as a dimension reduction technique to see the accuracy results obtained from improving this method. The dataset used is the Mushroom dataset from the official website of the UCI Machine Learning Repository. The Mushrooms dataset consists of 22 features and 1 class. After classification using Naive Bayes with Principal Component Analyst and evaluation using the 10-Fold Cross Validation technique, the results obtained are pc = 10 with an accuracy of 84%.

Keywords: *Dimension Reduction*, *Mushrooms Dataset, Principal Component Analyst*

I. INTRODUCTION

Indonesia is a tropical country which has high humidity and allows various plants and microorganisms to grow properly. One of the microorganisms that can grow well in Indonesia is mushroom [1].

Natural wealth and abundant biodiversity in a country is something to be grateful for. One of these natural wealth that we know is mushrooms. In subtropical areas with cold temperatures to tropical areas with warm temperatures there have been more than thousands of mushrooms with various types. Several types of mushrooms scattered in the world considerably can cause disease in humans and plants, which even some of them are contain toxins [2].

In agriculture, the applicable case of classification is determining whether the species of oyster mushroom (gilled mushroom) from the Agaricus and Lepiota families are either classified as poisonous or safe for consumption [3]. The importance of classifying a mushroom to be a topic in this research is to determine whether a mushroom is either classified as poisonous class or safe for consumption based on the physical characteristics of the features themselves. This identification is very important because the process of retrieving a plant species in data storage is considered difficult and requires quite a long time to go [4].

There have been many researches on the classification of mushrooms using computer equipment and data mining methods. Researchthat has been done previously was classification using Naïve Bayes algorithm for the Mushroom dataset from the UCI Machine Learning Repository with the title "Implementation of the Naïve Bayes Classifier Method for Identifying Mushroom Types" by Septian Ari Prayoga, Ismasari Nawangsih, and Tri Ngudi Wiyatno. From the results of the classification process, it can be concluded that the accuracy using Naïve Bayes method is 86.06%, and classified into good classification. Shows that the naive Bayes classifier method is well applied in the identification of agaricus and lepiota mushrooms in the edible or poisonous category [5].

Based on the results of previous research, researchers wanted to conduct research on the classification of mushroom species using Naive Bayes and improved by applying dimension reduction techniques using Principal Component Analysts. The purpose of improving this method is to see the results of the accuracy obtained from the application of the dimensional reduction technique used.

II. RESEARCH METHOD

In this research, there are several systematic stages are implied, as shown in the picture below.

2.1 Data Preparation

The dataset used is a mushroom dataset of 8124 records with large dimensions of 22 attributes and 1 class.

Table 1. Mushroom Data Variables

Variables	Attributes Information
cap-shape	*bell=b,conical=c,convex =x,flat=f, knobbed=k,sunken=s*
cap-surface	*fibrous=f,grooves=g,scal y=y,smooth=s*
cap-color	*brown=n,buff=b,cinnam on=c,gray=g,green=r, pink=p,purple=u,red=e, white=w,yellow=y*
bruises	*bruises=t,no=f*
odor	*almond=a,anise=l,creos ote=c,fishy=y,foul=f, musty=m,none=n,punge nt=p,spicy=s*
gill-attachment	*attached=a,descending= d,free=f,notched=n*
gill-spacing	*close=c,crowded=w,dist ant=d*
gill-size	*broad=b ,narrow=n*
gill-color	*black=k,brown=n,buff=b*

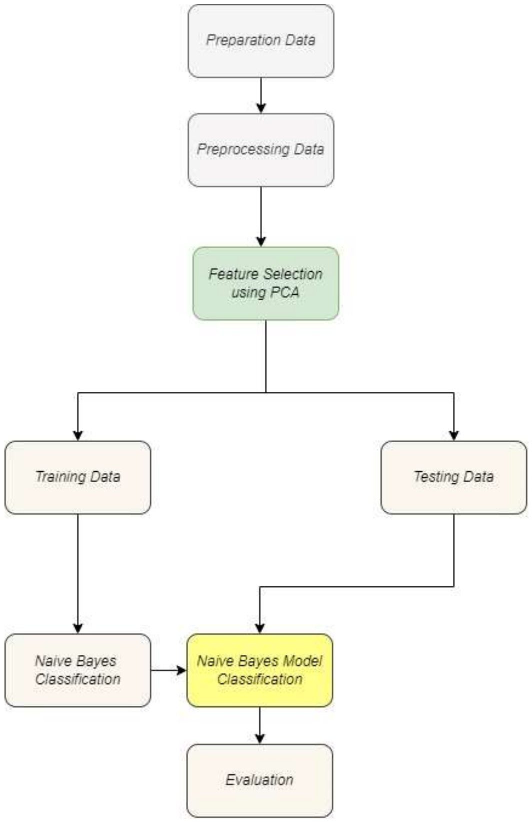

Preparation Data
Preprocessing Data
Feature Selection using PCA
Training Data
Testing Data
Naive Bayes Classification
Naive Bayes Model Classification
Evaluation

Image 1. Research methodology scheme.

The dataset was obtained from the official UCI Machine Learning Repository website. The data used is called Mushroom Data Set. The purpose of this data collection is to distinguish between edible and poisonous mushrooms using the Naïve Bayes classification technique and the Principal Component Analyst using mushroom data, and to determine the level of accuracy.

	,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
stalk-shape	*enlarging=e,tapering=t*
stalk-root	*bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?*
stalk-surface-above-ring	*fibrous=f,scaly=y,silky=k,smooth=s*
stalk-surface-below-ring	*fibrous=f,scaly=y,silky=k,smooth=s*
stalk-color-above-ring	*brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y*
stalk-color-below-ring	*brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y*
veil-type	*partial=p,universal=u*
veil-color	*brown=n,orange=o,white=w,yellow=y*
ring-number	*none=n,one=o,two=t*
ring-type	*cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z*
spore-print-color	*black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y*
population	*abundant=a,clustered=c, numerous=n, scattered=s,several=v,solitary=y*
habitat	*grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d*

2.2 Preprocessing Data

Pre-processing is a stage that is carried out on a set of data that will be used as training data before the mining process, both for classification and clustering [6]. The preprocessing stages carried out in this research used the process of cleansing and transforming data [7].

2.3 K-fold Cross Validation

The technique of cross validation is k-fold cross validation, which breaks the data into 'k' parts of the data set with the same size. The use of k-fold cross validation is to eliminate bias in the data. Training data and data testing are carried out as many as the specified number of k [8].

In k-fold cross validation, the original data is partitioned into subsets. This model then build by using data from subnet K-1 (2,3,4,5, etc.), and any other part of the subset used for the test set. The subset (dataset) itself should be more than just a test set, carried out iteratively to have a different model. The results of each K model (accuracy) are then combined by using an average to obtain accuracy results from the entire data. The benefit of using k-fold cross-validation is that each record can be recorded as once of a time, therefore, the disadvantage is that the validation task needed tends to be more difficult in progres [9]. Whereas the steps are as follow :

- Dataset being used is the mushroom around of 8124 records are already processed at the preprocessing, so the data shall be ready to process further.
- The distribution of data training and data testing or usually be called as Split, which is using the 10-fold cross validation method, the dataset will be divided into 10 parts from 8124 on category edible and poisonous mushrooms.

2.4 Dimension Reduction

At this stage, feature reduction is performed on the mushroom dataset using the Principal Component Analyst. Dimensional reduction is the process of reducing the number of features (dimensions) without eliminating important information from the data. Usually dimension reduction is applied as a preprocessing method to create a better classification model [10]. Principal Component Analysis (PCA) is a technique that used for the pre-processing process to perform feature scaling and data reduction [11]. The Principal Component Analysis (PCA) method is very useful for data that has a large number of features and has a correlation (interconnected) between the variables. The calculation of the principal component analysis (PCA) is based on the calculation of the eigenvalues and eigenvectors which express the distribution of data from a dataset [12].

The reduction function is used to reduce the number of variables (which were initially very large) to be smaller so as to facilitate analysis at a later stage. Meanwhile, the transformation function is used to change variables that are initially correlated to be uncorrelated. The following are the steps needed to perform Principal Component Analysis, as such:

1. Preparing data (data standardization),
2. Calculate the covariance matrix or correlation matrix,
3. Calculating the eigenvalues and eigenvectors from the correlation matrix,
4. Choose the principal component,
5. Output visualization,
6. Calculates a new score.

2.5 Naive Bayes Classification

At this stage, the classification model is made using Naive Bayes algorithm with the following stages [5].

1. Training is the determination of data that will be used as input for testing using the Naive Bayes method.
2. Classification is carried out for data testing and data training.
3. Accuracy. At this stage, accuracy calculations will be carried out using rapid miner tools on each test based on the confusion matrix of each test.

One of the probabilistic classification algorithms, the naive Bayes algorithm is a simple probabilistic classification based on the application of the Bayes theorem with strong assumption of independence, in other words the Naive Bayes algorithm assumes that the existence of a certain value of an attribute is not related to the presence of other attribute values. Naive Bayes classification is very suitable when the input dimensions are high, and has comparable performance with several other classification methods such as decision trees and neural network classifiers. Here are the equation of the theory of Naive Bayes:

$$p(D) = \frac{p(D)\,p(H)}{p(D)} \tag{1}$$

Where :

P : Probability
D : Data with an unknown class.
H : The data hypothesis is class specific.
p(□|□) : Hypothesis H probability based on condition D (posterior probability).
p(□) : Hypothesis H probability (prior probability).
p(□|□) : Probability D is based on the conditions in hypothesis H.
p(□) : D probability.

2.6 Confusion Matrix

Evaluation to measure the performance of the model is using confusion matrix. Where the confusion matrix is obtained from the validation process. Evaluating the performance of classification algorithms generally uses the overall results on the test dataset. A matrix of predictions that will be compared with the original class of input or in other words contains information on the actual and predicted values of the classification [13] . Confusion matrix can help show the details of classifier performance by providing information on the number of features of a class that are correctly and incorrectly classified[14]. Confusion matrix is a table that represents the performance of an algorithm or model specifically as shown in Table X[15].

4. True Positive = the number of actual data with positive class and prediction of the positive model.
5. True Negative = the number of actual data with negative class and the prediction of the negative model.
6. False Positive = the number of actual data with a negative class and the predictions of the positive model.
7. False Negative = the number of actual data with a positive class and the predictions of the negative model.

Based on the data above there are other data that can be used to measure the model performance, namely :

1. Accuracy = the total of all data that correctly classifies.
   (TP + TN) / Total
2. Precision = The total of all correct data with positive predictions.
   TP / (FP + TP)

3. Recall = The total of how often the model produces positive predictions with the actual positive class.
   TP / (FN + TP)
4. F1-Score = average harmonics of Precision and Recall
   2 x ((precision x recall) / (precision + recall))

2.7 Evaluation

At this stage, validate the model obtained by using k-Fold Cross Validation. Cross-validation (CV) is a statistical method that can be used to evaluate the performance of a model or algorithm where the data is separated into two subsets, namely learning process data and validation / evaluation data. The model or algorithm is trained by the learning subset and validated by the validation subset. Furthermore, the selection of the type of CV can be based on the size of the dataset. Usually CV K-fold is used because it can reduce computation time while maintaining the accuracy of the estimate.

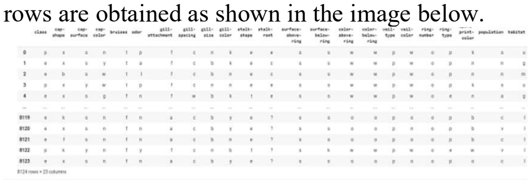
III. RESULTS AND DISCUSSION

3.1 Pre-Processing

At this preprocessing stage, the missing value is replaced in the stalk-root variable with the mean value and the data is normalized.

a. Replacing the missing value in the stalk-root variable with the mean value

In the mushroom dataset there are 2840 missing value on the stalk-root variable and 5644 valid data. the value in the missing value “?” need to replace the value of "?" becomes "nan" and then deletes every row that has the value "nan". This process is to get the mean value. determining the value of mean value, the first thing to do is replace the value of "?" with nan then delete the row contain the value nan. After replacing the value on the missing value, a new value is obtained as shown in the following image. 5644 rows are obtained as shown in the image below.



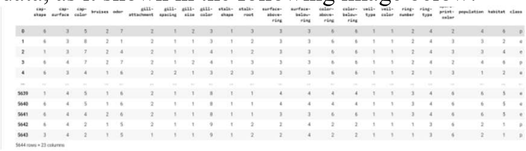
	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	surface-above-ring	surface-below-ring	color-above-ring	color-below-ring	veil-type	veil-color	ring-number	ring-type	print-color	population	habitat
0	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
1	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
2	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
3	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
4	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8119	e	k	s	n	f	n	a	c	b	y	e	?	s	s	o	o	p	o	o	p	b	c	l
8120	e	x	s	n	f	n	a	c	b	y	e	?	s	s	o	o	p	n	o	p	b	v	l
8121	e	f	s	n	f	n	a	c	b	n	e	?	s	s	o	o	p	o	o	p	b	c	l
8122	p	k	y	n	f	y	f	c	n	b	t	?	s	k	w	w	p	w	o	e	w	v	l
8123	e	x	s	n	f	n	a	c	b	y	e	?	s	s	o	o	p	o	o	p	o	c	l

8124 rows × 23 columns

Table 2. removing missing values

After obtaining a new value, the categorical data is converted into numerical data as shown in the image below.

Then change the categorical data to be numerical data, as it shown in the following image below.

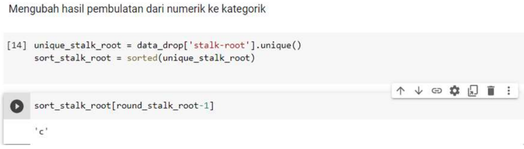


	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	surface-above-ring	surface-below-ring	color-above-ring	color-below-ring	veil-type	veil-color	ring-number	ring-type	print-color	population	habitat	class
0	6	3	5	2	7	2	1	2	3	1	3	3	3	6	6	1	1	2	4	2	4	6	p
1	6	3	8	2	1	2	1	1	3	1	2	3	3	6	6	1	1	2	4	3	3	2	e
2	1	3	7	2	4	2	1	1	4	1	2	3	3	6	6	1	1	2	4	3	3	4	e
3	6	4	7	2	7	2	1	2	4	1	3	3	3	6	6	1	1	2	4	2	4	6	p
4	6	3	4	1	6	2	2	1	3	2	3	3	3	6	6	1	1	2	1	3	1	2	e
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5639	1	4	5	1	6	2	1	1	8	1	1	4	4	4	4	1	1	3	4	6	6	5	e
5640	6	4	5	1	6	2	1	1	8	1	1	4	4	4	4	1	1	3	4	6	6	5	e
5641	6	4	4	2	6	2	1	1	8	1	1	3	3	6	6	1	1	3	4	6	6	5	e
5642	6	4	2	1	5	2	1	1	9	1	2	2	4	2	2	1	1	1	3	6	2	1	p
5643	3	4	2	1	5	1	1	1	9	1	2	2	4	2	2	1	1	1	3	6	2	1	p

5644 rows × 23 columns

Table 3. Numerical data

Then for the second stage, which is to calculating the mean on the stalk-root variable. The mean value obtained is 1.59 which is then rounded in integer form to 2. In the stalk-root variable, the value of 2 is the value for c or it can be called CLUB.



Mengubah hasil pembulatan dari numerik ke kategorik

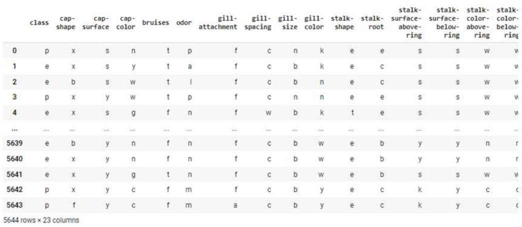
```
[14] unique_stalk_root = data_drop['stalk-root'].unique()
     sort_stalk_root = sorted(unique_stalk_root)
```

```
sort_stalk_root[round_stalk_root-1]
```

```
'c'
```

Image 2. Rounding results from numeric to categorical.

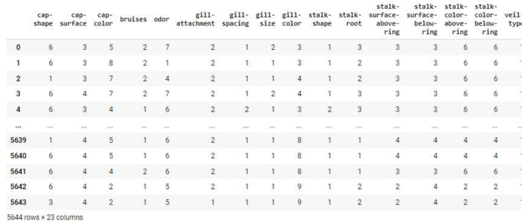
After the c or CLUB value is obtained, then the missing value will be filled in by using the categorical attribute value. Then after that, it is being converted from the categorical data into numerical data.



	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
0	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w
1	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w
2	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w
3	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w
4	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5639	e	b	y	n	f	n	f	c	b	w	e	b	y	y	n	r
5640	e	x	y	n	f	n	f	c	b	w	e	b	y	y	n	r
5641	e	x	y	g	t	n	f	c	b	w	e	b	s	s	w	w
5642	p	x	y	c	f	m	f	c	b	y	e	c	k	y	c	c
5643	p	f	y	c	f	m	a	c	b	y	e	c	k	y	c	c

5644 rows × 23 columns

Tabel. 4. Data Categorical



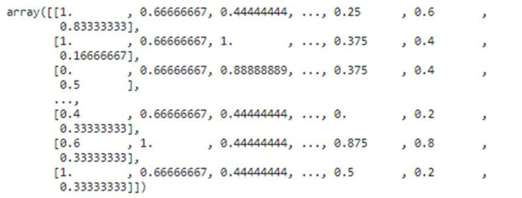
	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-surface-above-ring	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring
0	6	3	5	2	7	2	1	2	3	1	3	3	3	6	6
1	6	3	8	2	1	2	1	1	3	1	2	3	3	6	6
2	1	3	7	2	4	2	1	1	4	1	2	3	3	6	6
3	6	4	7	2	7	2	1	2	4	1	3	3	3	6	6
4	6	3	4	1	6	2	2	1	3	2	3	3	3	6	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5639	1	4	5	1	6	2	1	1	8	1	1	4	4	4	4
5640	6	4	5	1	6	2	1	1	8	1	1	4	4	4	4
5641	6	4	4	2	6	2	1	1	8	1	1	3	3	6	6
5642	6	4	2	1	5	2	1	1	9	1	2	2	4	2	2
5643	3	4	2	1	5	1	1	1	9	1	2	2	4	2	2

5644 rows × 23 columns

Table 5. Numerical Data

b. Data normalization

The numerical data from the previous process is then being normalized by using min-max normalization method.



```
array([[1.        , 0.66666667, 0.44444444, ..., 0.25      , 0.6       ,
         0.83333333],
       [1.        , 0.66666667, 1.        , ..., 0.375     , 0.4       ,
         0.16666667],
       [0.        , 0.66666667, 0.88888889, ..., 0.375     , 0.4       ,
         0.5       ],
       ...,
       [0.4       , 0.66666667, 0.44444444, ..., 0.        , 0.2       ,
         0.33333333],
       [0.6       , 1.        , 0.44444444, ..., 0.875     , 0.8       ,
         0.33333333],
       [1.        , 0.66666667, 0.44444444, ..., 0.5       , 0.2       ,
         0.33333333]])
```

Image 3. Data Normalization

3.2 Principal Component Analyst

After the pre-processing stage, a classification process will be carried out using Naive Bayes. before starting the classification process, the mushroom dataset is divided into training data and testing data with 90% for training data and 10% for testing data. In the classification process this test uses 10-fold cross validation and the number of pc = 10. The following is a visualization of the number of principal components.

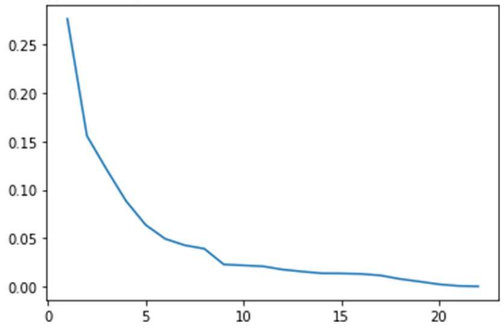
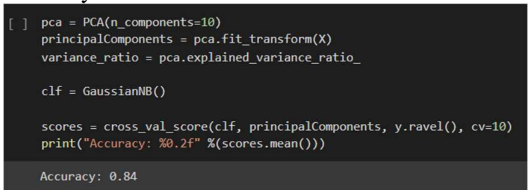


Image 4. Schematic of research methodology.

The results of this research were obtained from the calculation of the correlation matrix and feature selection using PCA with 10 components, the accuracy results is 84%



```
[ ]  pca = PCA(n_components=10)
     principalComponents = pca.fit_transform(X)
     variance_ratio = pca.explained_variance_ratio_

     clf = GaussianNB()

     scores = cross_val_score(clf, principalComponents, y.ravel(), cv=10)
     print("Accuracy: %0.2f" %(scores.mean()))

     Accuracy: 0.84
```

Image 5. Program calculation results.

3.3 Confusion Matrix

In this research, after classifying the mushroom data based on data sharing using K-fold Cross Validation method, 90% of the training data and 10% of the testing data of the entire data were random and balanced. In order to obtain 8124 data as test data which is evaluated by looking at the information contained regarding the classification which is correctly predicted by a classification system in the confusion matrix

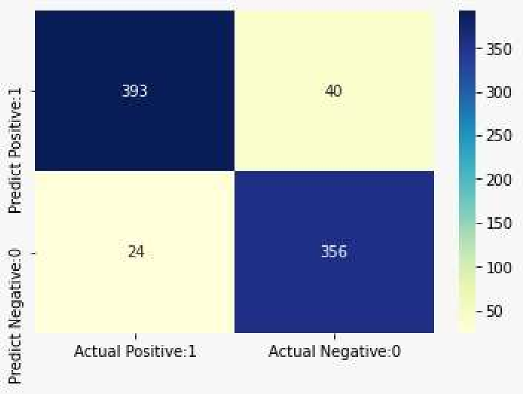


Image 6. Confusion Matrix

Based on the results of the confusion matrix, precision and recall values can be calculated manually as follows :

Precision value
Precision= TN/(TP+FP)=356/(393+40)=0.8221

Recall value
Recall= TP/(TP+FN)=393/(393+24)=0.9424

IV. CONCLUSION

Based on the results of this research, to determine the types of edible mushrooms and poisonous mushrooms using Naive Bayes classification method and applying the Principal Component Analyst with the number of pc = 10 and applying the evaluation technique using 10-Fold Cross Validation the accuracy value is 84%. so that it can be concluded that the Naive Bayes algorithm method using principal component analyst is good enough to be applied in classifying the types of poisonous and edible mushrooms.

REFERENCES

[1] Fitriani, L., Krisnawati, Y., Anorda, M. O. R., & Lanjarini, K. (2018). JENIS-JENIS DAN POTENSI JAMUR MAKROSKOPIS YANG TERDAPAT DI PT PERKEBUNAN HASIL MUSI LESTARI DAN PT DJUANDA SAWIT KABUPATEN MUSI RAWAS. *Jurnal Biosilampari: Jurnal Biologi*. https://doi.org/10.31540/biosilampari.v1i1.49

[2] Parjimo, H., & Andoko, A. (2007). *Budi Daya Jamur (Jamur Kuping, Jamur Tiram, Jamur Merang)*.

[3] Prihatini, R. (n.d.). Penerapan Data Mining

sebagai Evaluasi Ketepatan Akurasi terhadap Klasifikasi Mushroom Data Set. *Academia.Edu*, 11.

[4] Aruan, T. (2017). *IDENTIFIKASI JENIS TANAMAN JAMUR BERACUN MENGGUNAKAN PENDEKATAN K-NEAREST NEIGHBOR*.

[5] Prayoga, Septian Arie, I. N. and T. N. W. (2019). Implementasi Metode Naive Bayes Classifier Untuk Identifikasi Jenis Jamur. *Ilmiah Informatika, Arsitektur Dan Lingkungan*, *14*(2), 134–144. https://jurnal.pelitabangsa.ac.id/index.php/pelitatekno/article/view/239/191

[6] Misra, P., & Yadav, A. S. (2019). Impact of Preprocessing Methods on Healthcare Predictions. *SSRN Electronic Journal*, *Ml*. https://doi.org/10.2139/ssrn.3349586

[7] Muttaqin, F. A., & Bachtiar, A. M. (n.d.). *IMPLEMENTASI TEKS MINING PADA APLIKASI PENGAWASAN PENGGUNAAN INTERNET ANAK " DODO KIDS BROWSER " Jurnal Ilmiah Komputer dan Informatika (KOMPUTA) Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*.

[8] Tuntun, R. (2022). *Analisis Perbandingan Kinerja Algoritma Klasifikasi dengan Menggunakan Metode K-Fold Cross Validation*. *6*, 2111–2119. https://doi.org/10.30865/mib.v6i4.4681

[9] Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*. https://doi.org/10.1002/9781118874059

[10] Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *Journal of Information Engineering and Educational Technology*. https://doi.org/10.26740/jieet.v5n2.p49-54

[11] Sartika, D., & Saluza, I. (2022). Penerapan Metode Principal Component Analysis (PCA) Pada Klasifikasi Status Kredit Nasabah Bank Sumsel Babel Cabang KM 12 Palembang Menggunakan Metode Decision Tree. *Generic*, *14*(2), 45–49.

[12] Nasution, M. Z., Nababan, A. A., & Syaliman, K. U. (2019). *PENERAPAN PRINCIPAL COMPONENT ANALYSIS (PCA) DALAM PENENTUAN FAKTOR DOMINAN YANG MEMPENGARUHI PENGIDAP KANKER SERVIKS (Studi Kasus : Cervical Cancer Dataset)*. *3*(1), 204–210.

[13] Telaumbanua, K., Sudarto, S., Butar-Butar, F., & Bilqis, P. S. (2021). Identifikasi Sampah Berdasarkan Tekstur Dengan Metode GLCM dan GLRLM Menggunakan Improved KNN. *Explorer*. https://doi.org/10.47065/explorer.v1i2.94

[14] Bramer, M. (2016). *Introduction to Data Mining*. https://doi.org/10.1007/978-1-4471-7307-6_1

[15] Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*. https://doi.org/10.24076/citec.2019v6i1.178