Mushroom Classification Using Random Forest Classifiers with Python

Pankaj Kumar Singh
PG Scholar
Alliance College of Engineering and Design
Alliance University
Bangalore, India

Hitesh Kumar
PG Scholar
Alliance College of Engineering and Design
Alliance University
Bangalore, India

Dr. Senbagavalli M
Associate Professor/IT
Alliance College of Engineering and Design
Alliance University
Bangalore, India

Abstract- *Random forest classifier which is mostly used in the era of machine learning with classifiers, which has more accuracy while transmission of the line image processing of the model. The main aim of this paper is to classify mushroom which is safe to eat and poisonous. The process of classifying the edible or poisonous mushrooms is by using the specification or the features and color of the mushroom cap. We have used the Random Forest classifier algorithm with the help confusion matrix we are able training the dataset to identify the required features with the accuracy of 100%.*

Keywords- Random Forest, Integrated Classifier, Confusion Matrix.

I. Introduction

Classification is a task that requires the different use of machine learning algorithms that learn how to assign a class label to examples from the specific problem domain. For example, classifying emails as "spam" or "not spam." There are many different types of classification tasks that you may meet in machine learning and dedicated approaches to modeling that may be used for each. In this research, we have built different machine learning models that will identify if the mushroom is safe to eat or poisonous by tuning its specifications like its shape, color, gill color, etc. using random forest classifiers. Random forest classifier which is mostly used in the era of machine learning with classifiers, which has more accuracy in transmission line image processing of the model. Random Forest is purely Supervised Learning Algorithm and the supervised classification steps are included here:

Step1: user collects trainig data

Step2: user specifies training sites to be used for classification process

Step3: Assign pixels to closest class based on training data

Step4: Evalting Results

Most of the effort is done prior to the actual classification process. Once the classification is run in supervised classification the output is a thematic image with classes that are labeled and correspond to

information classes or land cover types. When compared to unsupervised classification, the Supervised classification can be much more accurate, but depends heavily on the training sites, specific skill of the individual image processing and spectral distinctness of classes.

II. Literature Survey

[2] Wei Zhengtao important applications of classifier integration. But in the improved the resampling method by adding constraints to the process of classification, when the classification results of each sampling results, and strengthened the classification ability of base classifier have similar error distribution, the final the Algorithm. [8]Hinton, G.E., Krizhevsky, A., Srivastava, N., Sutskever, I., & Salakhutdinov proposed the key idea is to randomly drop units from the neural network during training. This prevents units from co-adapting too much. This significantly reduces overfitting and gives major improvements over other regularization methods. [9]Senbagavalli, M & TholkappiaArasu,G proposed Decision Tree based Feature Selection algorithm for Opinion Mining. [11] Senbagavalli, M & Tholkappia Arasu,G proposed a Competent approach for extracting and visualizing web opinions using Clustering. [10] This survey is featured by in-depth analysis and discussion in various aspects, many of which, to the best of our knowledge, are the first time in this field. Above all, it is our intention to provide an overview how different deep learning methods are used rather than a full summary of all related papers.

III. Proposed method.

We propose and study the approach to the formation of the recommendations for determining the search ranges of the parameters values of the RF-classifier using the results of experimental studies to develop the appropriate models of this classifier based on various datasets from machine learning data repositories. Moreover, when forming the recommendations, the specifics of the decision trees development are considered. The choice of the RF algorithm from the total number of the advanced machine learning algorithms is due, inter alia, to the clarity of the presentation of the decision rules, as well as the ease of implementation and interpretation of the results. [6] We are using specifications like shape, color, gill color, etc. of the mushrooms cap to classify the mushrooms that is suitable or safe to eat and poisonous.
The python libraries and packages which we are considering in this project are below:

NumPy
Pandas
Seaborn
Matplotlib
Graphviz
Scikit-learn

The dataset used in these project contains 8124 instances of mushrooms with 23 features like cap-shape, cap-surface, cap-color, bruises, odor, etc.

We are importing the python packages and libraries to train the data set and examine to learn and identify the data frame by printing the information in the form or columns and rows.

1. We are using **count**, which shows the number of responses.
2. We also consider, **Unique** which shows the number of unique categorical values.
3. **Top** shows that the highest-occurring value.
4. And the **freq** shows the frequency of the highest categorical value.

Random Forrest Algorithm

RF algorithm actively applied for the RF classifier development in many scientific spheres for solving different classification problems. [7]

Random Forest is also called as Random Decision Forest (RFA) which is used for Classification, Regression and other tasks that are performed by constructing multiple decision trees. This Random Forest Algorithm is based on supervised learning and the major advantage of this algorithm is that it can be used for both Classification and Regression. Random Forest Algorithm gives you better accuracy when compared with all other existing systems and this is most used algorithm. [5]

Steps involved in Random Forest Classification Algorithm:

1. The first step is to Start with the selection of random samples from a given dataset information.
2. Next, construct a decision tree for every sample by using this algorithm. Then it will provide the prediction result from each decision tree.
3. In this third Step, Voting will be performed for every predicted result.
4. At last, select and fix the most voted prediction result as the final prediction result.

Following diagram illustrate its working:

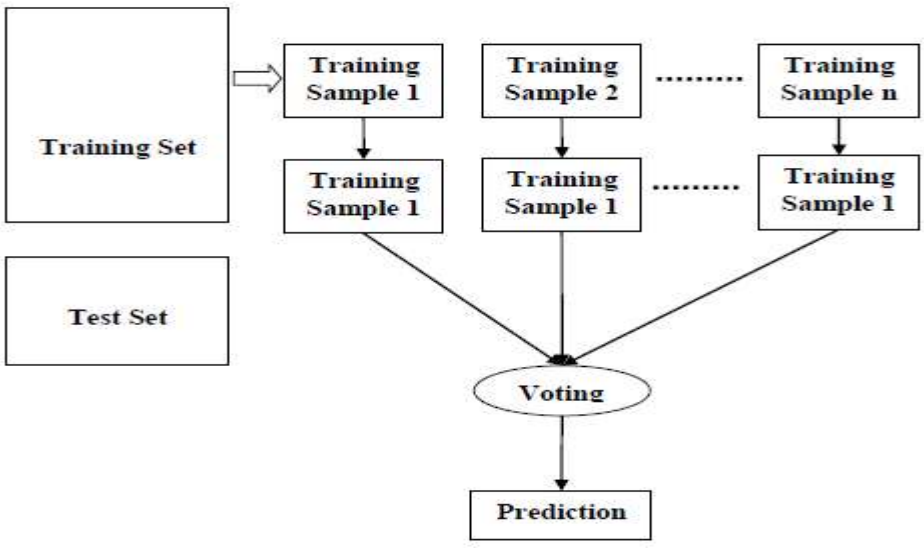

Fig. 1 working Model on Random Process Classifiers

IV. Experimental data and results.

100% using Random Forest Algorithm.

Here is the sample dataset that we are referring for our research, which is openly made available by Kanchi Tank[4]. With the referred database and analysis, the result that will bring the final conclusion by using of Random forest algorithm.

	A	B	C	D	E	F	G	H	I
1	class	cap-shape	cap-surfac	cap-color	bruises	odor	gill-attach	gill-spacin	gill-s
2	p	x	s	n	t	p	f	c	n
3	e	x	s	y	t	a	f	c	b
4	e	b	s	w	t	l	f	c	b
5	p	x	y	w	t	p	f	c	n
6	e	x	s	g	f	n	f	w	b
7	e	x	y	y	t	a	f	c	b
8	e	b	s	w	t	a	f	c	b
9	e	b	y	w	t	l	f	c	b
10	p	x	y	w	t	p	f	c	n
11	e	b	s	y	t	a	f	c	b
12	e	x	y	y	t	l	f	c	b
13	e	x	y	y	t	a	f	c	b
14	e	b	s	y	t	a	f	c	b
15	p	x	y	w	t	p	f	c	n
16	e	x	f	n	f	n	f	w	b
17	e	s	f	g	f	n	f	c	n
18	e	f	f	w	f	n	f	w	b
19	p	x	s	n	t	p	f	c	n

Fig. 2 Sample data set

The shape of the dataset

So our dataset contains 8124 rows which is instances of mushrooms and 23 columns indicates the specifications of mushroom cap which are shape, surface, color, bruises, odor, gill-size, etc.

```
print("Dataset shape:", df.shape)
```

Dataset shape (8124, 23)

The shape of the dataset

Unique occurrences of 'class' column

```
df['class'].unique()
```

The .unique() method will deliver the unique occurrences in the 'class' column of the dataset.

Here is the output:

```
array(['p', 'e'], dtype=object)
```

Unique values

'e' -> edible and 'p' -> poisonous

The count of unique occurrences of 'class' column

```
df['class'].value_counts()
```

The .value_counts() method will deliver the count of the unique occurrences.

Here is the output:

```
e    4208
p    3916
Name: class, dtype: int64
```

Value counts

There are 4208 occurrences of eatable or safe to eat mushrooms and 3916 occurrences of poisonous mushrooms in the dataset.

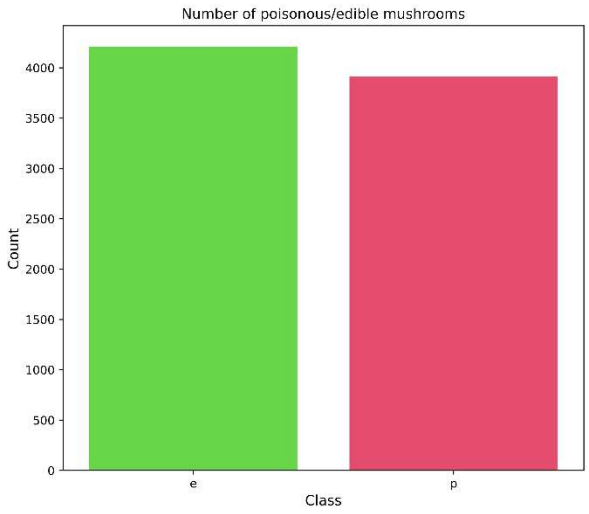

Fig. 3 Bar plot to visualize the count of edible and poisonous mushrooms

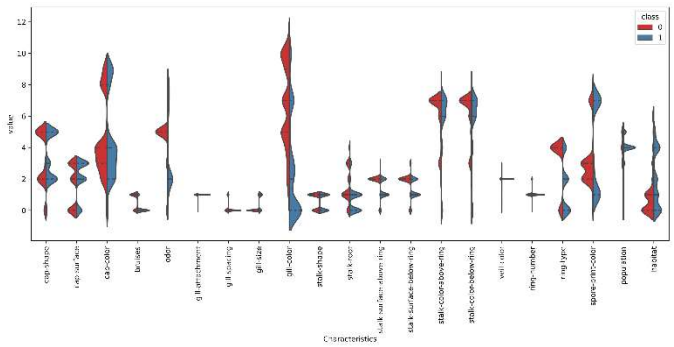

Fig. 4 Violin plot representing the distribution of the classification characteristics

Confusion Matrix for Random Forest Classifier

A confusion matrix is the best technique which is used in the research for summarizing the performance of a classification algorithm in Random Forest.

```
cm = confusion_matrix(y_test, y_pred_rf)x_axis_labels = ["Edible", "Poisonous"]
y_axis_labels = ["Edible", "Poisonous"]f, ax = plt.subplots(figsize =(7,7))
sns.heatmap(cm, annot = True, linewidths=0.2, linecolor="black", fmt = ".0f", ax=ax,
cmap="Purples", xticklabels=x_axis_labels, yticklabels=y_axis_labels)
plt.xlabel("PREDICTED LABEL")
plt.ylabel("TRUE LABEL")
plt.title('Confusion Matrix for Random Forest Classifier');
#plt.savefig("rfcm.png", format='png', dpi=500, bbox_inches='tight')
plt.show()
```

Here is the **output**:

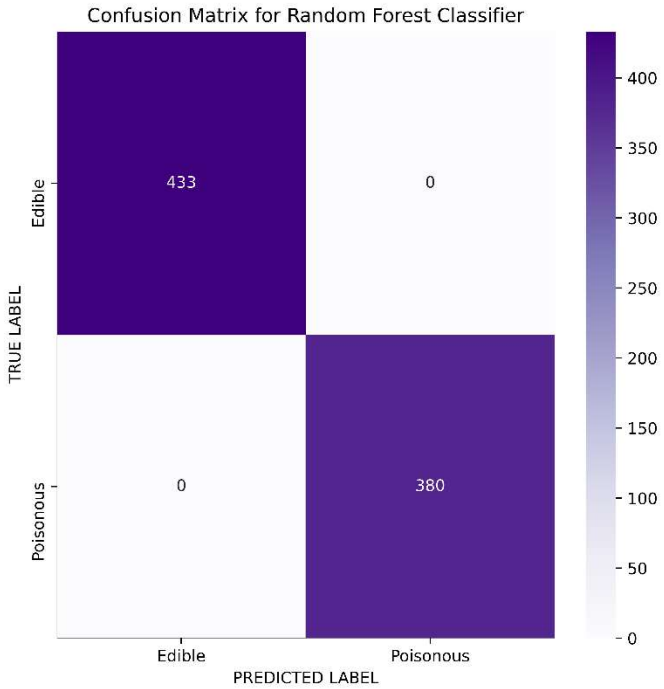

Fig. 5 Output of the confusion Matrix for Random Forest classifier.

Classification accuracy alone can be misleading if you have an unequal number of observations in each class that's where the confusion matrix comes in the picture.

Predictions

Predicting some of the X_test results and matching it with true i.e. y_test values using Decision Tree Classifier.

```
preds = dt.predict(X_test)print(preds[:36])
print(y_test[:36].values)# 0 - Edible
# 1 - Poisonous
```

```
[0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0 1]
[0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0 1]
```

Predictions

As we can observe that the predicted value are true and it has 100% match.

V. Conclusions.

In this paper, we have used the methodology to classify the eatable and poisonous mushrooms using its features. As the classification is difficult for human when the quantity is high, so the proposed system will identify the bunch of mushrooms based on the specification to get the edible or poisonous. As we have considered using Random Forest classifier to test the data to

Acknowledgments.

This research paper is executed with the help of Random Forest algorithm and other past works. The datasets are used for training and educational purposes to learn the concept and bring the best in it.

References:

[1] Zhang Bingzhen; Qiao Xiaoming; Yang Hemeng; Zhou Zhubo : "*A Random Forest Classification Model for Transmission Line Image Processing"*, 2020.

[2] Zhengtao Wei. "*Stochastic Forest Algorithm Research based on Non-equilibrium data [D]". Xidian University*,2017.

[3] Xu Zhang; Ding Han; Fengshan Bai; Ziyin Ma; *IEEE conference paper on Flower Recognition Based on Convolutional Neural Network*, 2019.

[4] Kanchi Tank, *Introduction to classification using Decision Tree, Logistic Regression, KNN, SVM, Naive Bayes, Random Forest Classifiers with Python, Blogger post*, 2020

[5] Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S., & Aswini, E. (2019). Credit Card Fraud Detection Using Random Forest Algorithm. 2019 3rd International Conference on Computing and Communications Technologies (ICCCT). doi:10.1109/iccct2.2019.8824930.

[6] Liliya Demidova, Mariya Ivkina. "*Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier*", 2019.

[7] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, Horst Bischof*, "On-line Random Forests*", 2009.

[8] Hinton, G.E., Krizhevsky, A., Srivastava, N., Sutskever, I., & Salakhutdinov,"Dropout: a simple way to prevent neural networks from overfitting", in Journal of Machine Learning Research, 15, 1929-1958,2014

[9] Senbagavalli, M &TholkappiaArasu,G , 'Opinion Mining for Cardiovascular Disease using Decision Tree based Feature Selection' , Asian Journal of Research in Social Sciences and Humanities –Vol. 6, No. 8, August 2016, pp. 891-897, ISSN 2249-7315.

opinions using Clustering', in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, ISSN: 2277 128X, 2014.

[11] Sergey Ioffe Christian Szegedy,'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'.

[12] D. A. M´ely, J. Kim, M. McGill, Y. Guo, and T. Serre. A systematic comparison between visual cues for boundary detection. Vision research, 120:93–107, 2016.