**MUSHROOM CLASSIFICATION PROJECT**

* **ABSTRACT:**

Mushroom is one of the fungi types’ foods that has the most potent nutrients on the plant. Mushrooms have major medical advantages such as killing cancer cells. This study aims to find the most appropriate technique for mushroom classification, and mushroom will be classified into two categories, poisonous and nonpoisonous. The proposed approach will implement a different techniques and algorithms like neural network (NN), Support Vector Machines (SVM), Decision Tree, and k Nearest Neighbors (KNN), on dataset of mushroom images, where the dataset contains images with background and without background.



* **INTRODUCTION:**

This study aims to find new approach working to classify the mushrooms images based on different features using the different techniques of Machine Learning (ML). The purpose of classification process is to predict categorical class labels or the target value, for example, feed-forward Artificial Neural Network (ANN), and the purpose of classifier is to map data to predefined classes or groups.

In the proposed approach, we used the training dataset that contain the mushroom images to classify it into poisonous and nonpoisonous. Where our approach aims to classifies and predict for the class (groups) of mushrooms when submit the features of the mushrooms to different techniques of machine learning.

A mushroom is one of the fungi types’ food that has the most potent nutrients on the plant. Mushrooms have major advantages such as kill cancer cells, viruses and enhancing

the human immune system. Currently, the mushroom refers to the process that performed by robot in food industry. This technique used to limit the features such as color. Recently, mushroom system used specific characteristics that improve the selection process of mushrooms. Such system depends on analyzing and investigating the features in order to get better classification based on the well-known features.

* **ALGORITHM AND TECHNIQUES:**

In this study, we will use different machine learning algorithms and techniques for mushroom classification, some of them are listed below:

• Neural Network (NN): is a distributed matrix structure, it used in different applications, such as classifying data and patterns, , predicting new cases or examples, and in pattern recognition applications. NN simulates human biological cells and human capability of thinking and learning.

• Decision Tree: is one of the most popular classification techniques in machine learning, where it used in decision support system. Decision aims to classify objects (instances) to find a track from the great parent node.

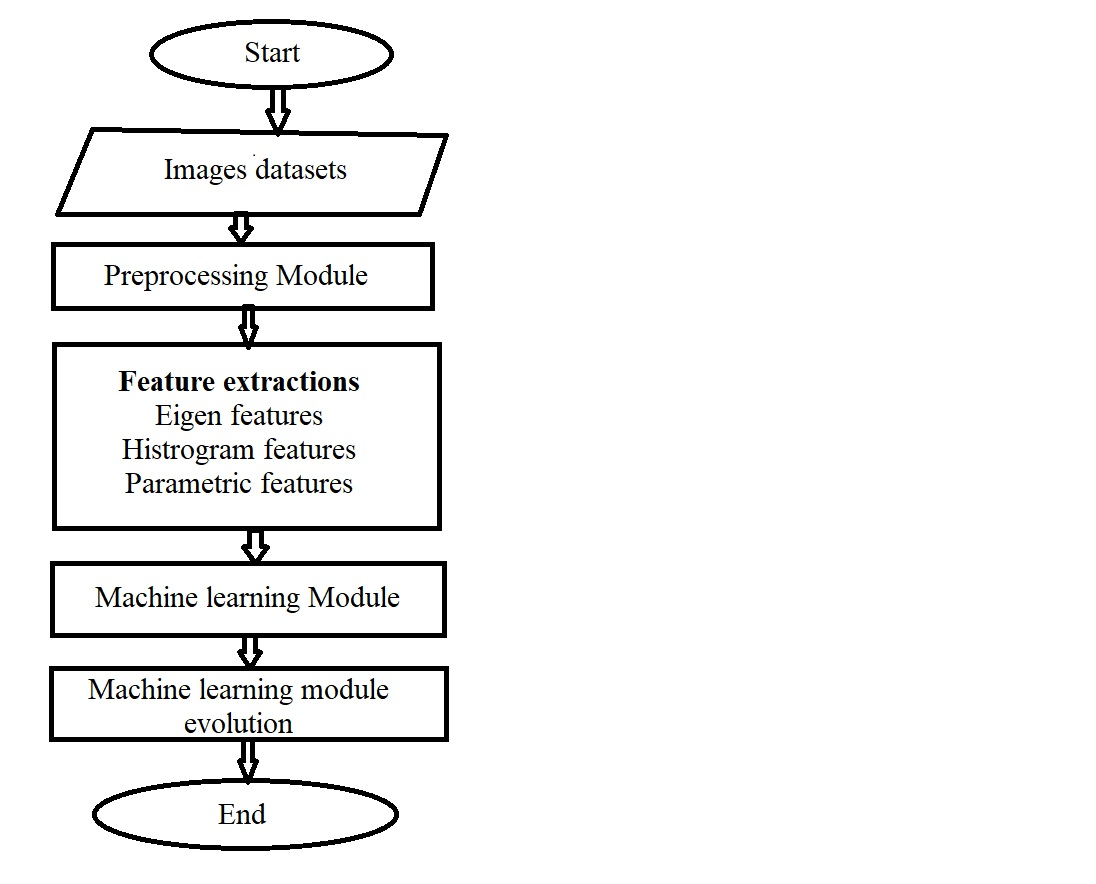
• kNN: is an algorithm and classified under machine learning. The kNN featured the low number of training parameters, where the computational complexity is not high, and the performance is satisfactory.

* **METHODOLOGY:**

The aim of this study is to identify Mushroom images and classify it into two categories (poisonous and nonpoisonous) using machine learning techniques.

* **Research Phases**

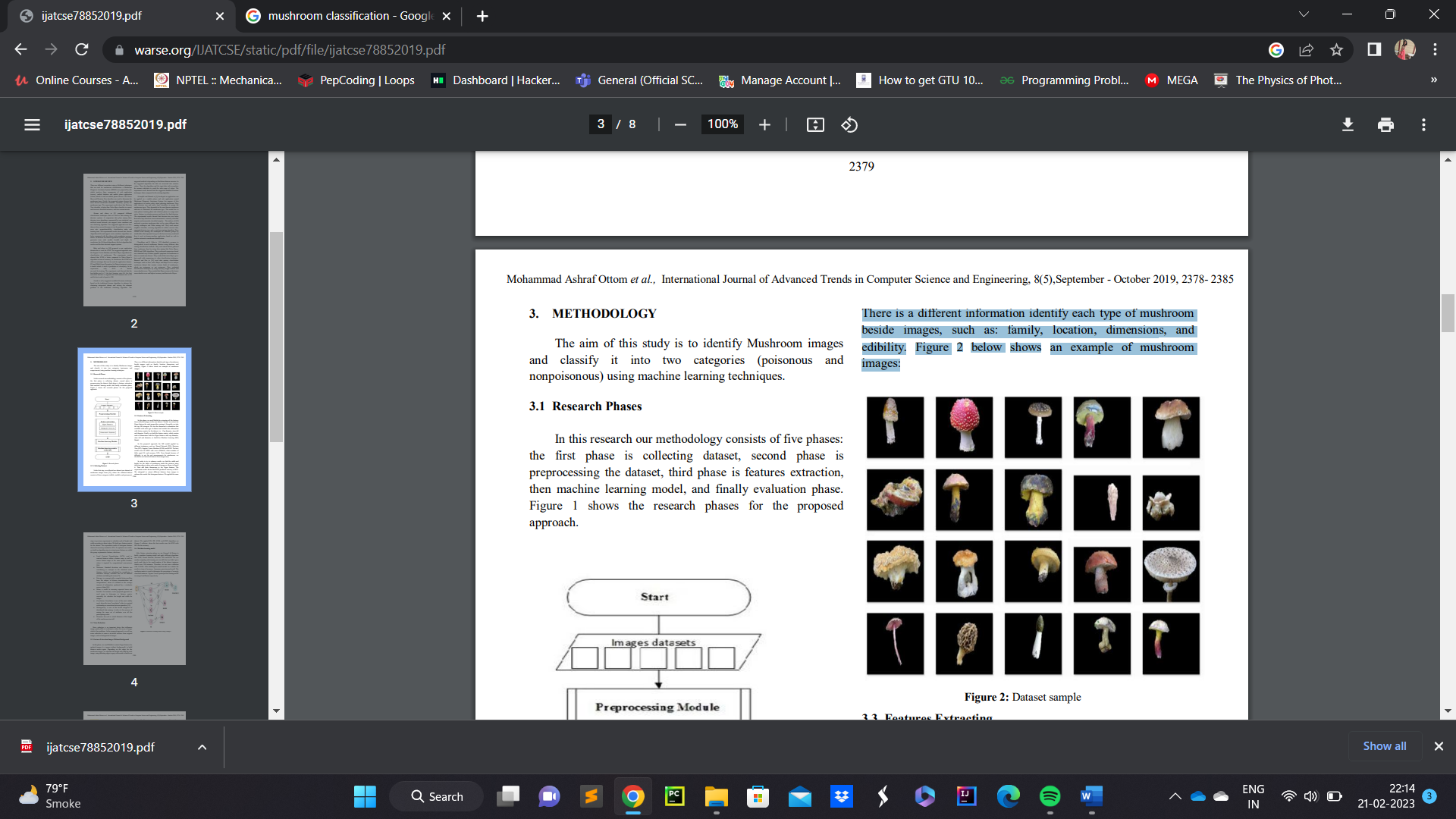
In this research our methodology consists of five phases: the first phase is collecting dataset, second phase is preprocessing the dataset, third phase is features extraction, then machine learning model, and finally evaluation phase. Figure shows the research phases for the proposed approach.



* **Collecting Dataset**

In the first step, we collected our dataset (raw dataset) of mushroom images from data, where the collected dataset consists of three categories (edible, inedible, and poisonous). There is a different information identify each type of mushroom beside images, such as: family, location, dimensions, and edibility. Figure

below shows an example of mushroom images:



* **Features Extracting**

In this phase, we used Matlab for extracting all the features from collected images in the raw dataset. Firstly, we extract the Eigen features for each image after resizing it. Secondly, we take the top 100 strongest. We use the dimension’s information that available with each type in dataset and include this information with feature matrix for the dataset, i.e. Cap diameter, stem tall and diameter. Finally, we build the feature matrix, which contain each of dimensions with the Eigen features with cap diameter, stem tall and diameter, to build the Machine Learning (ML) Model.

In the proposed approach, the ML model applied by different techniques, such as: Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), and KNN. In order to try to enhance results, we find the width and height for the shape of mushroom inside the pictures.

To extract more features we called this group as parametric features, which are:

• **Local Contrast Normalization (LCN):** used to contrast features within a feature map, as well as across feature maps at the same spatial location, where it inspired by computational neuroscience.

• **Skewness, Standard deviation, and Kurtosis:** are considering as concepts in the statistical meta features, which are calculated by considering a statistical concept, calculate this for all numeric attributes and taking the mean.

• **Entropy:** is a concept with a complex history and has been the subject of diverse reconstructions and interpretations, where it’s defined as the average amount of information produced by a stochastic source of data.

• **Mean:** is useful in assessing expected losses and benefits. For instance, in the proposed approach we used mean to determine in features matrix, especially for calculate the height and width for images.

• **Correlation:** Correlation is one of the most widely used, where the term "correlation" refers to a mutual relationship or association between quantities.

• **Homogeneity:** is one of the broad categories of distributed data mining, it refers to the process of mining the same set of attributes over all the participating nodes.

• **Diameter**: the real or virtual diameter of the length of the mushroom stem tall.

* **Noise Reduction**

Noise reduction is an important factor that influences image quality, its working to reduce the errors of image which it has problems. In the proposed approach, we will use noise reduction to remove un-useful sections from original images, such as background of images.

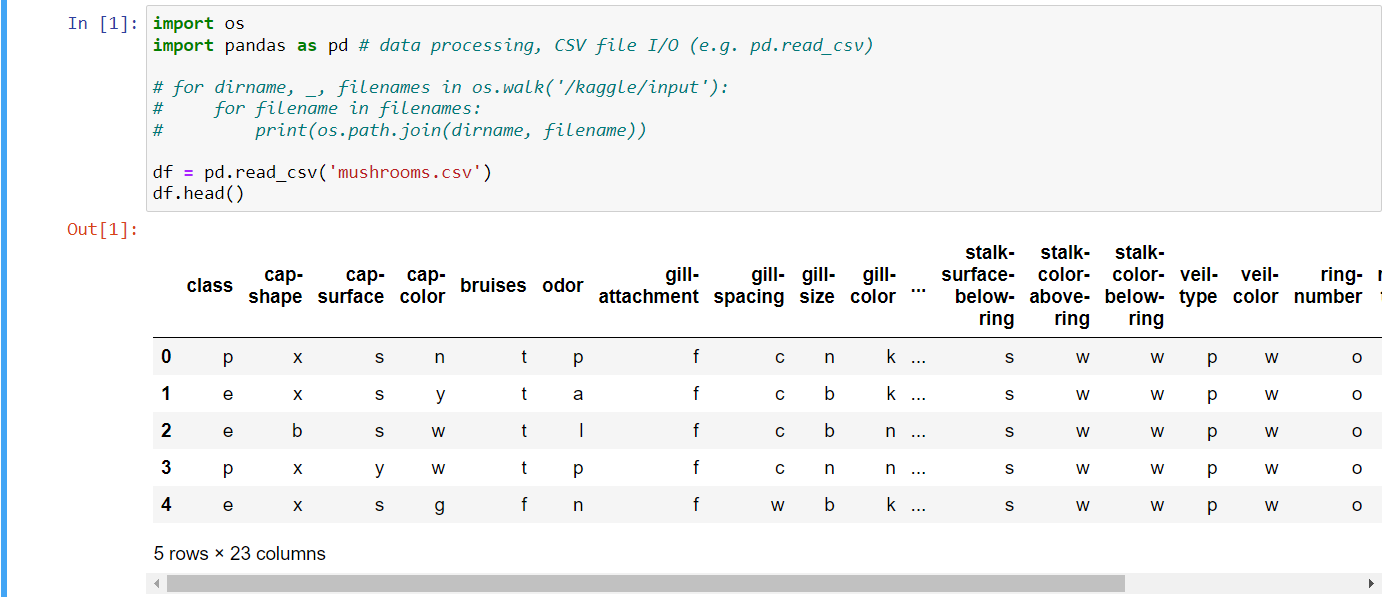
* **Features Extraction Images Without Background**

To extract Eigen features for updated images (i.e. images without background), to build features matrix again. Depending on the edges for the mushroom images, we calculate the height and width for each image, using detecting edges in gray scale mode to build new dataset. We applied NN, DT, SVM, and KNN algorithms.

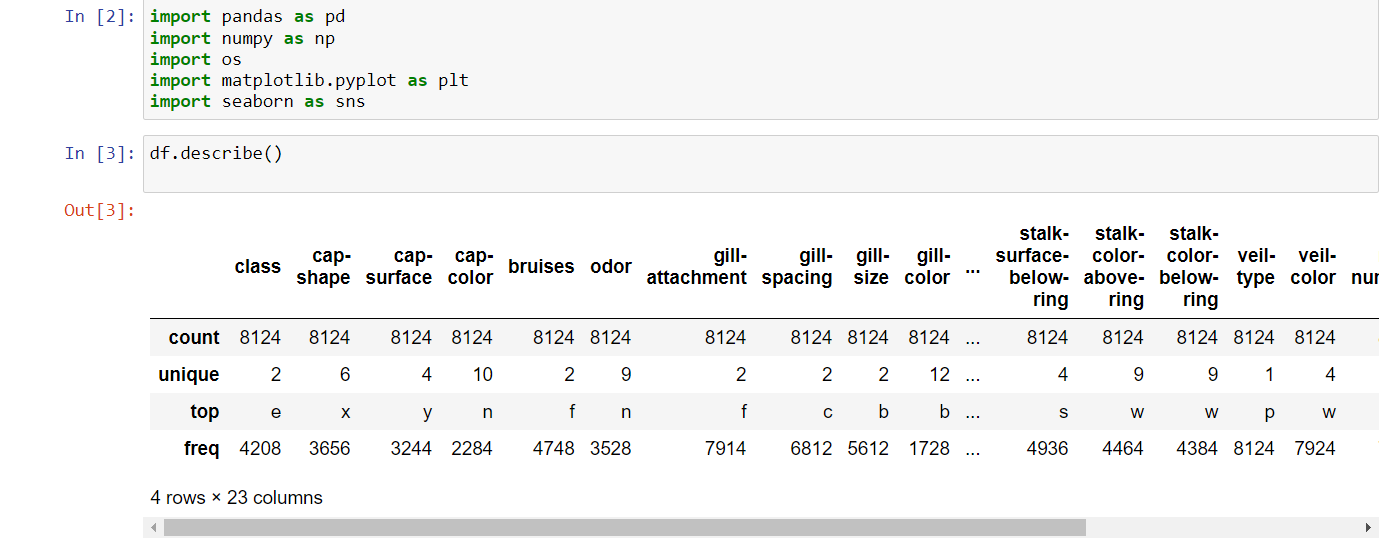
**Equipment/Instruments:** Jupyter Notebook

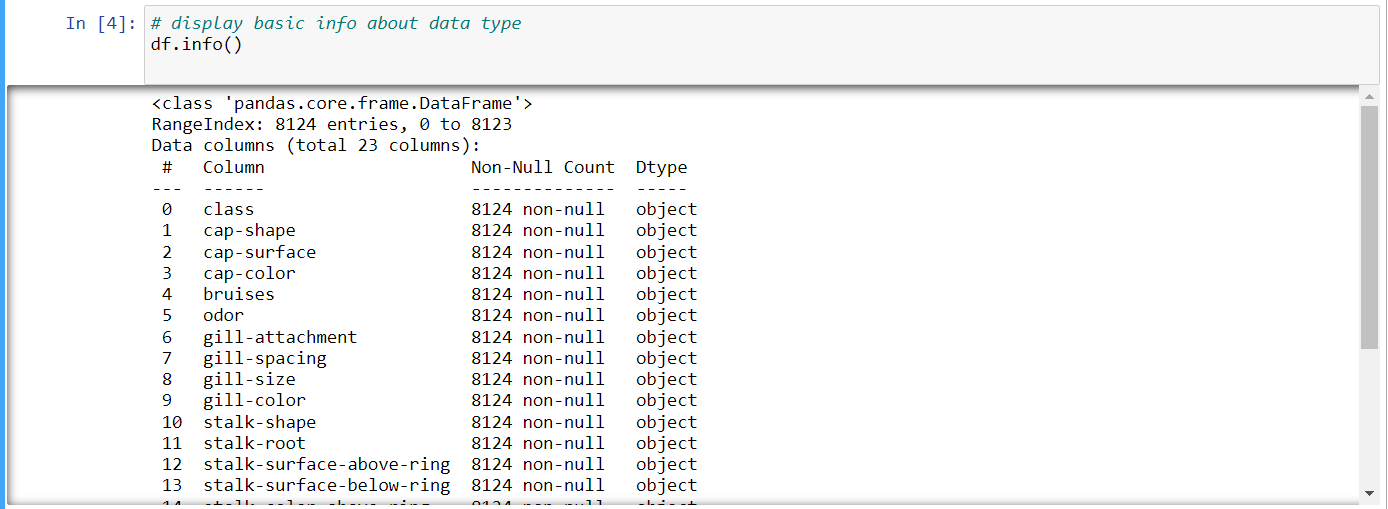
**Program:**

1. Loading the dataset

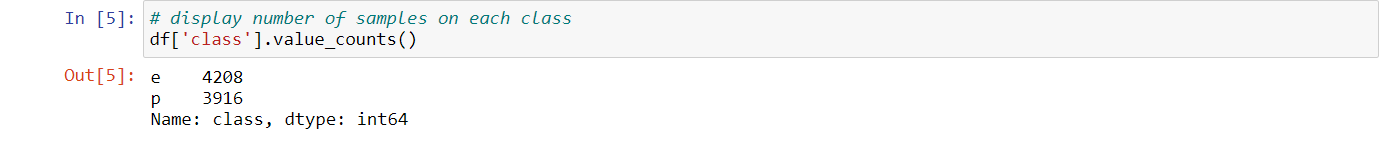


1. Import modules and describe the dataset
2. Display basic information

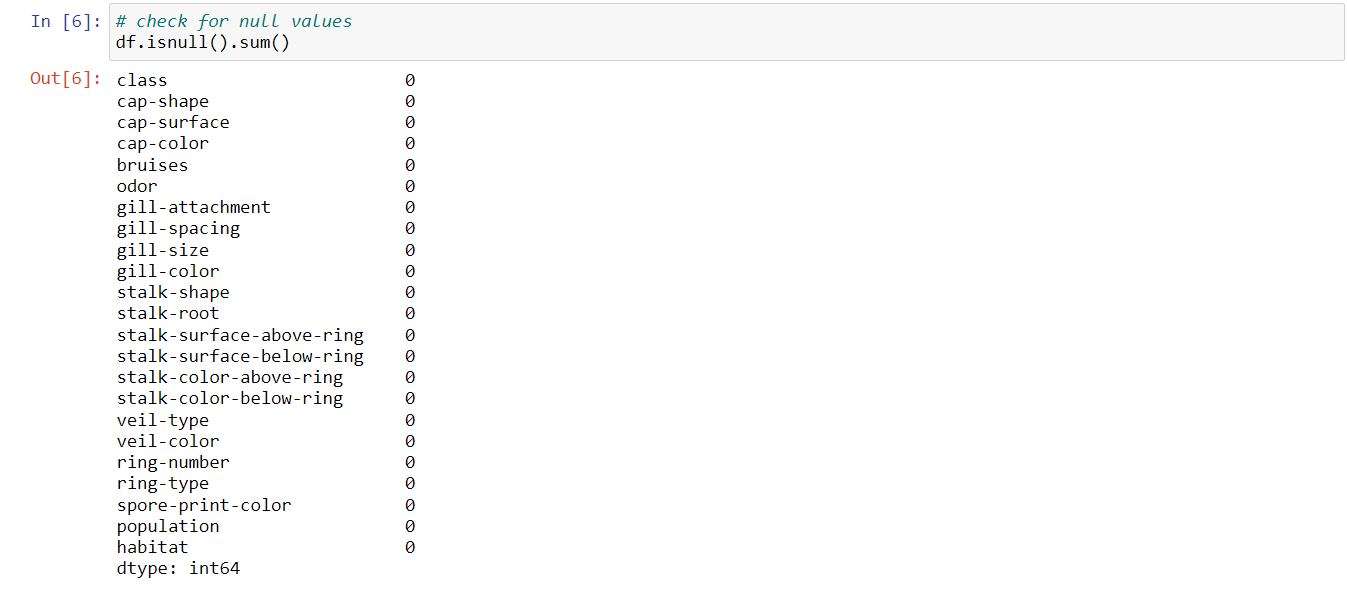


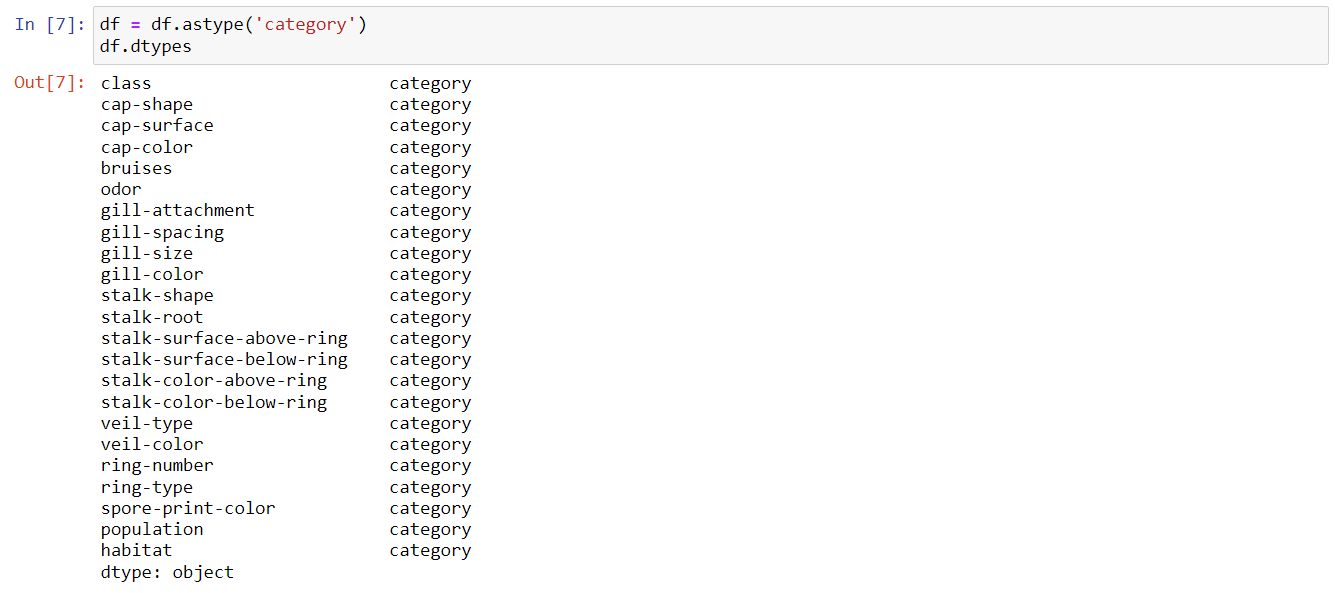


1. Display number of samples

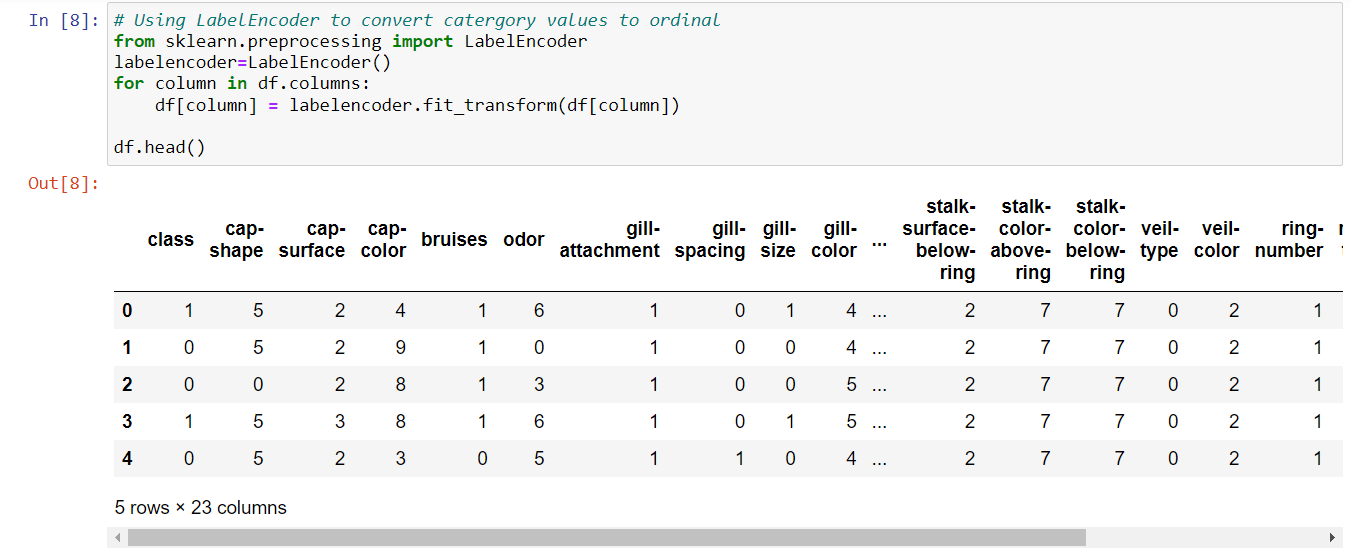


1. Check for null values





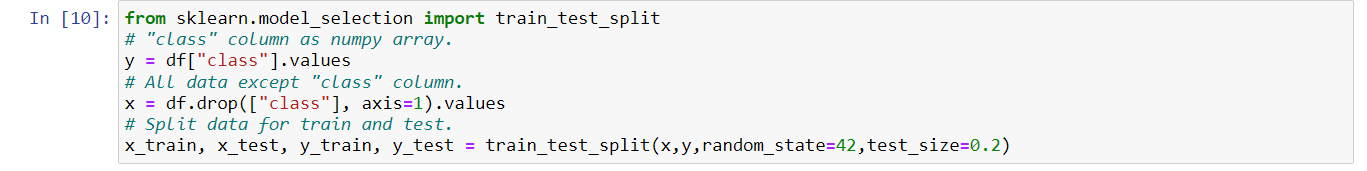
1. Now that we have converted the columns to be of category type, we can use LabelEncoder to make the columns into machine understandable format.



1. From the above figure we can see that veil-type has only one unique value and hence won't contribute anything to the data. So we can safely remove it.

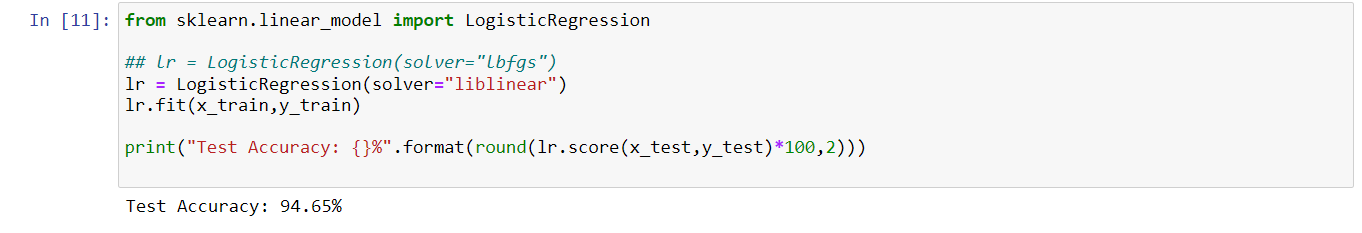


1. We can make use of scikit-learn's train\_test\_split method for creating the training and testing data.



**Classification Methods:**

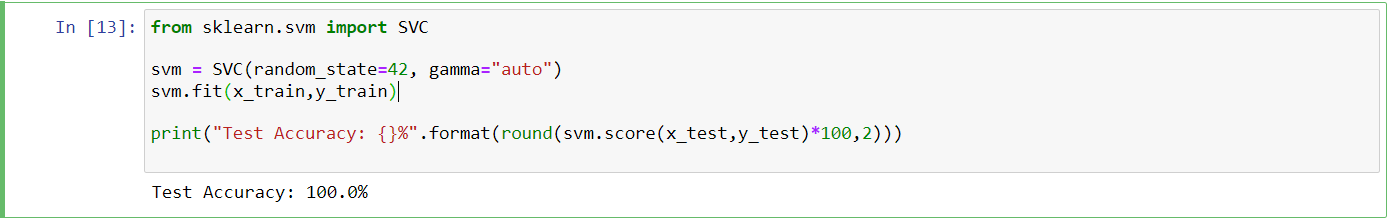
Logistic regression



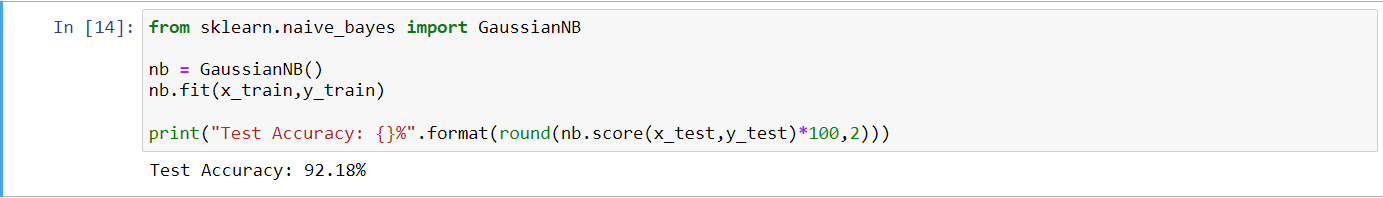
K-Neighbours classification



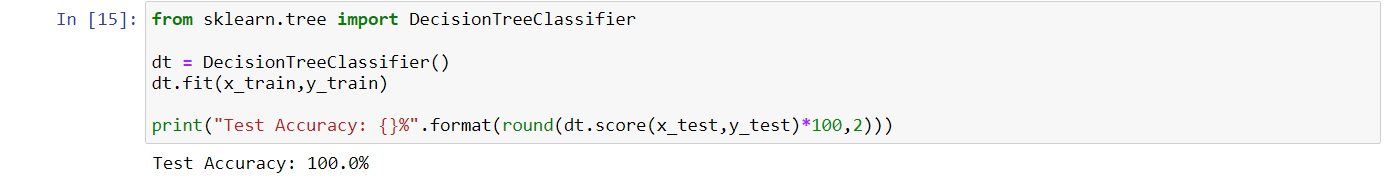
SVM classification



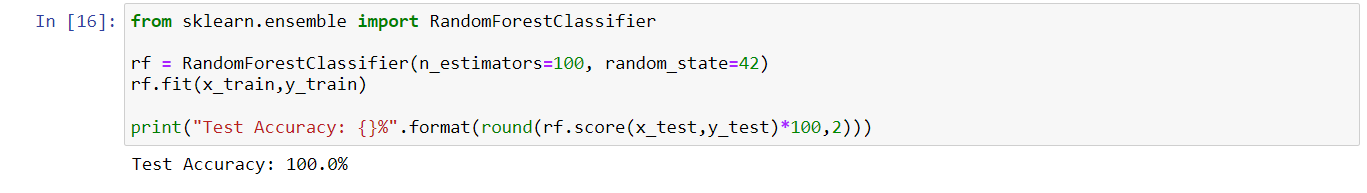
Naïve Bayes classification



Decision tree classification



Random forest classification

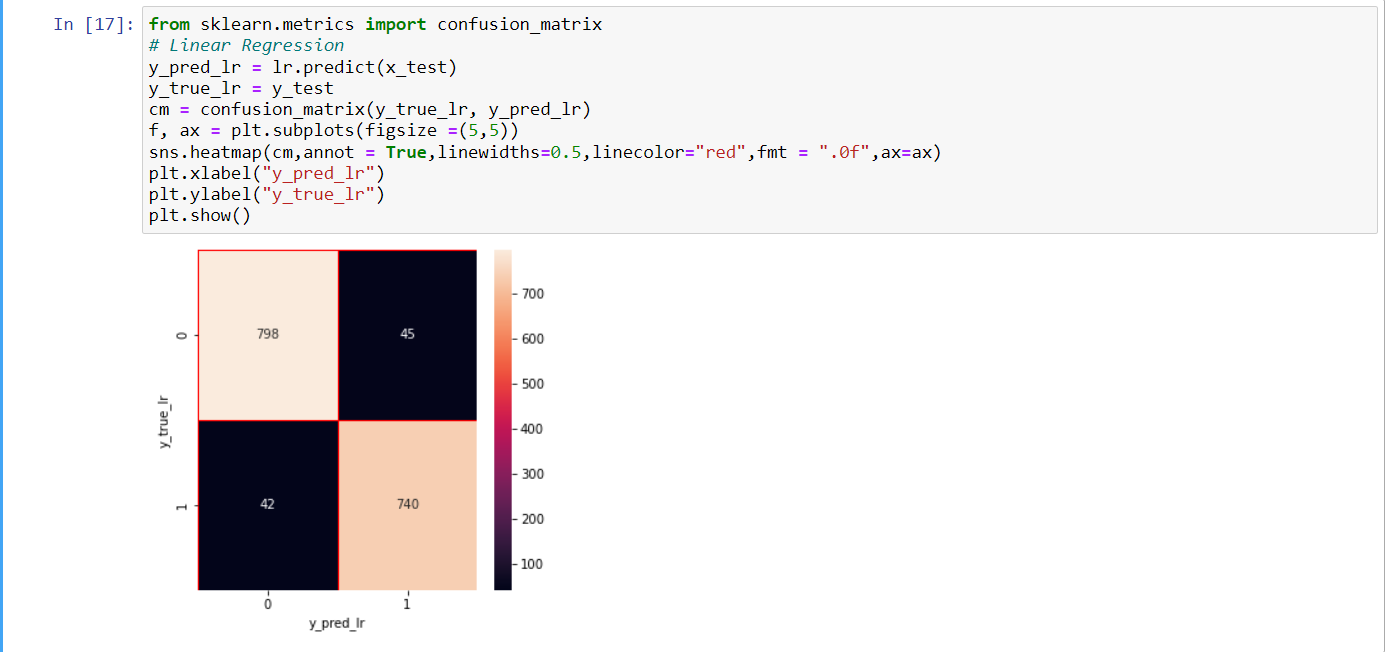


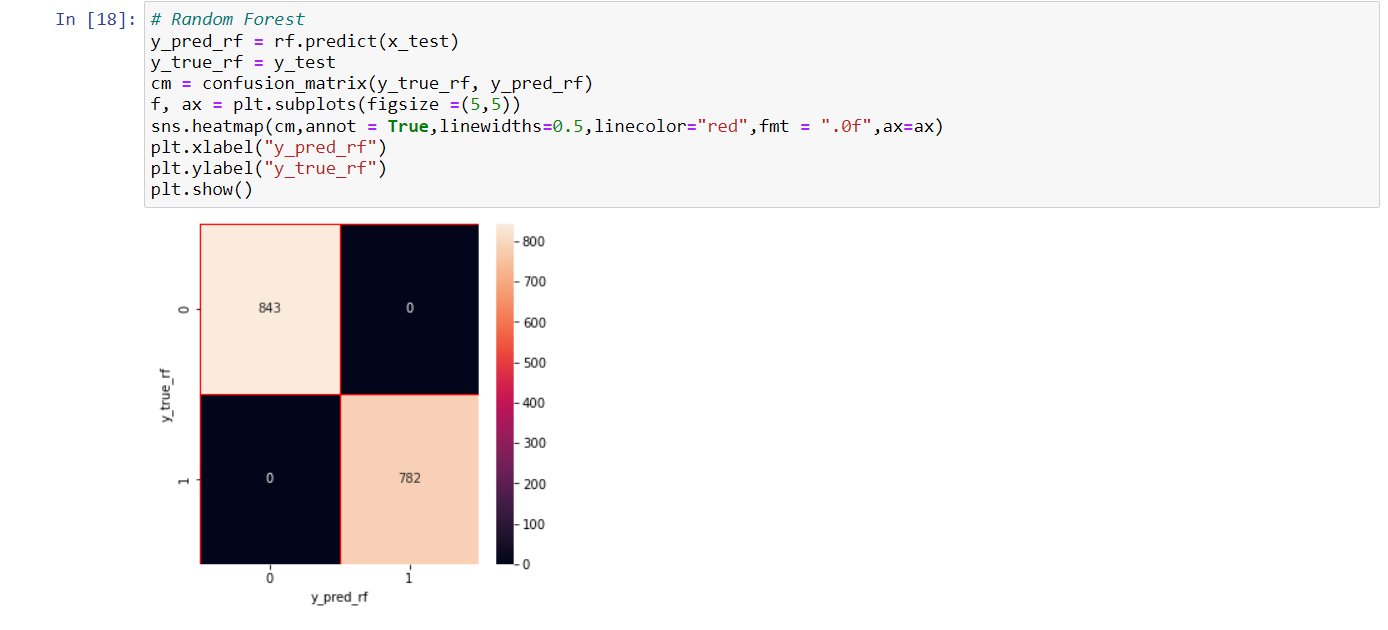
**Checking Classification Results with Confusion Matrix**

Check the results with confusion matrix on Logistic Regression and KNN Classification.

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

Logistic Regression's accuracy was 94.65% and KNN's was 100%.





**Observations:**

Logistic Regression's accuracy was 94.65% and KNN's was 100%.

**Conclusion:**

From the confusion matrix, we saw that our train and test data is balanced.

Most of classfication methods hit 100% accuracy with this dataset.

* **REFRENCES:**
* <https://medium.com/analytics-vidhya/mushroom-classification-using-different-classifiers-aa338c1cd0ff>
* <https://www.kaggle.com/datasets/uciml/mushroom-classification>
* <https://github.com/kanchitank/Mushroom-Classification>
* <https://projectsbasedlearning.com/apache-spark-machine-learning/machine-learning-project-on-mushroom-classification-whether-its-edible-or-poisonous-part-1/>