MUSHROOM DATASET- A CLASSIFICATION MODEL BUILD USING MACHINE LEARNING

ABSTRACT

Mushrooms have been consumed since earliest history. Ancient Greeks believed that mushrooms provide strength because of its high nutritional values. Mushroom has been considered as a ingredient of gourmet cuisine across the globe because of its unique flavour. Most of the cultivated mushrooms are of Agaricus family. Based on the Agaricus family of mushrooms, we will be developing one such model using machine learning techniques where we will be predicting the class of mushroom i.e they are edible or poisonous to eat based on its various features like cap shape, odour, gill spacing, stalk surface, rings, veil and many more. Through various analysis and plotting techniques we have created graphical representations of various features. We have proposed an efficient Decision Tree Classification (DTC) Model with almost 99% accuracy to predict the class of mushrooms. The other metrices like Precision, F1 score, Recall, ROC_AUC score are also evaluated.

Source

Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knop.

Donor

Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

DATASET

Raw data is something which has been collected from source but in its very initial phase. It has not been processed or cleansed to provide any useful information. The dataset that we are using is called as a mushroom dataset is available on UCI machine learning repository.

The dataset has total of 8125 rows and 23 columns/features. The dataset is in a raw form i.e we will be performing EDA (Exploratory Data Analysis) i.e some pre-processing and data cleaning techniques to covert the data into useful information for our machine learning model. We can observe that the all the features are of categorical type which makes it a classification problem.

In this Dataset, we will be predicting the class of mushroom i.e whether the mushroom is edible to eat or poisonous to eat based on the various features provided. We have to develop one such machine learning model which will predict the class of mushroom efficiently.

Details of various features:

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-colour**: brown=n, buff=b, cinnamon=c, grey=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
4. **bruises**: bruises=t, no=f
5. **odour**: 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-colour**: black=k, brown=n, buff=b, chocolate=h, grey=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-colour-above-ring**: brown=n, buff=b, cinnamon=c, grey=g, orange=o, pink=p, red=e, white=w, yellow=y
15. **stalk-colour-below-ring**: brown=n, buff=b, cinnamon=c, grey=g, orange=o, pink=p, red=e, white=w, yellow=y
16. **veil-type**: partial=p, universal=u
17. **veil-colour**: 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-colour**: 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

Details of label:

23. **class**: poisonous=p, edible=e.

Observations:

- Dataset has multivariate characteristics.
- Attributes are of categorial type.
- We have some missing values in the dataset.
- As the attributes are of categorial type we need to convert them to numeric type.
- As we have only two classes i.e edible and poisonous which makes it binary classification.
- We will also check the value counts of each class if there is major difference b/w two classes we will try to balance the set using sampling techniques.

DATA ANALYSIS

In this Phase we have created some graphical representations of various features on the basis of which we will come to know what kind of effect each feature has on predicting the class of mushroom.

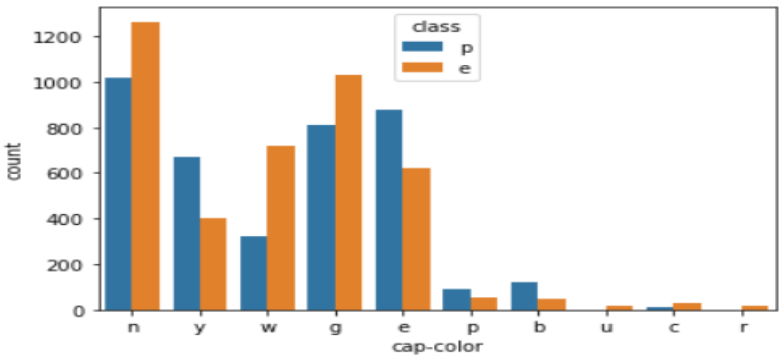


Figure 1: says that if the colour of the cap is White then it is highly edible to eat.

If the colour of the cap is yellow or red then the mushrooms are highly poisonous to eat.

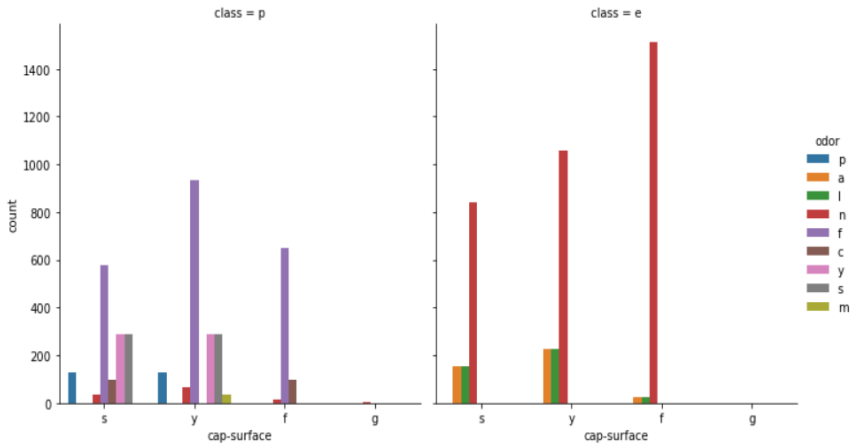


Figure 2: If the cap surface is fibrous and the odor is none then mushroom is highly edible to eat.

If the cap surface is scaly and the odor of mushroom is foul then mushroom is highly edible to eat.

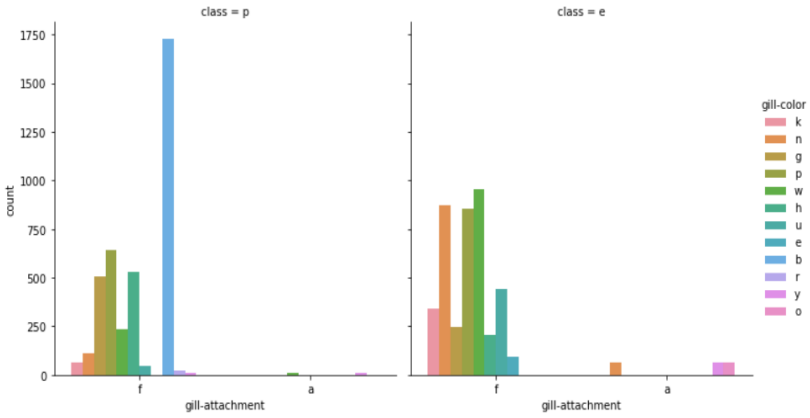
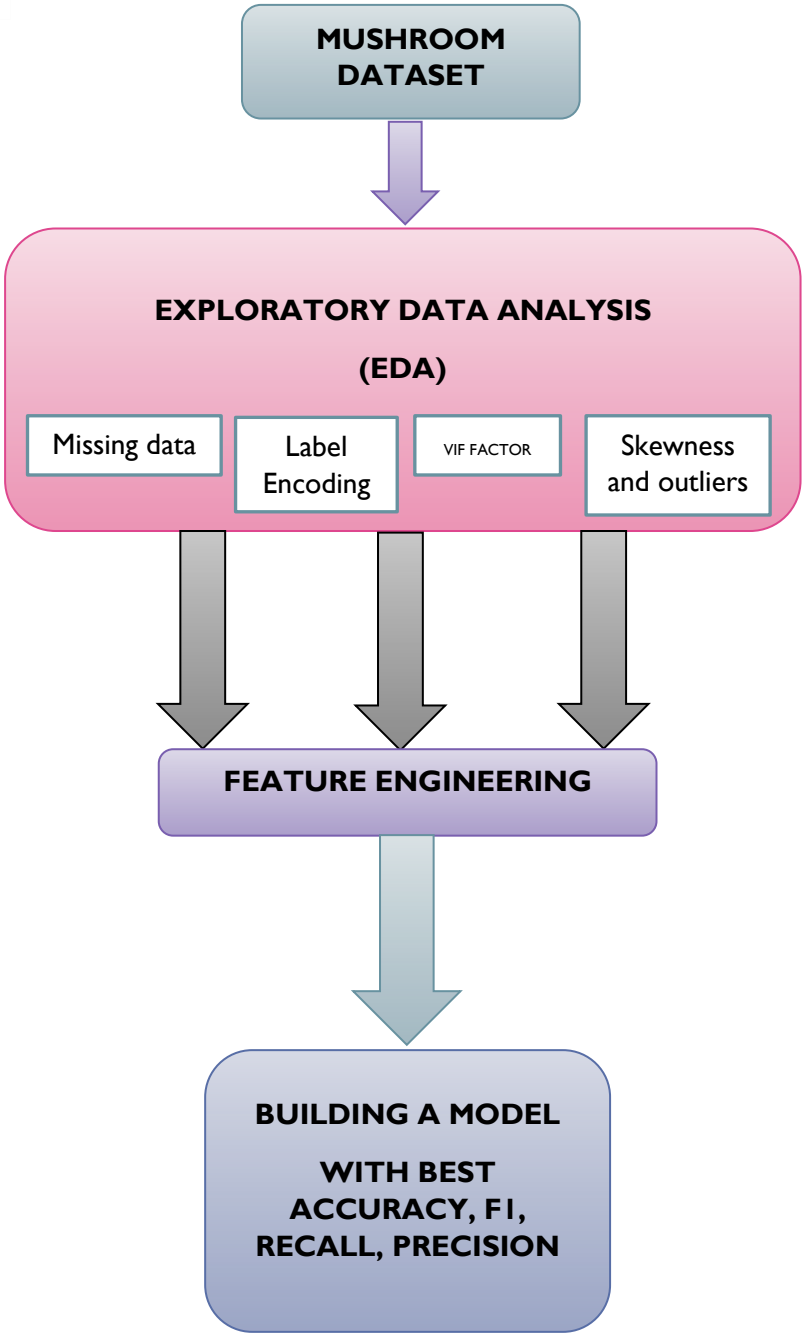


Figure 3: If the gill attachment is free and the gill color is buff then mushroom is poisonous to eat.

In the similar manner we have created some more pictorial representations while building model so as analyze the various features and label relationships.

Methodology for Mushroom Classification:-



Pre-Processing Pipeline

Preprocessing is one of the important steps in building a model. In this phase we usually deal with missing values if any or if there are any unrealistic values. In case of any irrelevant value we will remove that data. We will try to improve the quality of data in this phase so that we can develop a model with high accuracy score. For our dataset, we will first of all fill the missing values, then we have most of the data in categorial form we will covert the data in numerical form so that we are able to fed the data into classification algorithms. We can also check its there exists any multicollinearities through VIF FACTOR calculation.

EDA Concluding Remarks

- ✓ Replaced the missing data with the mode value of feature as it's a categorial data.
- ✓ Performed Label Encoding as all the features were in categorical format.
- ✓ The skewness and the outliers were removed using various techniques like log and sqrt transformation and zscore respectively.
- ✓ VIF factor was calculated to check the multicollinearities.
- ✓ The value count of label class was calculated to check if the dataset is balanced. As the dataset was balanced no sampling techniques were followed.
- ✓ As the values of the dataset were scaled so no scaling was performed.

Classification Experiments

We have performed various classification algorithms like Logistic Algorithm, Decision Tree Algorithm, Multinomial Naive Bayes, Support Vector Mechanism, Random Tree Classifier to check the accuracy. The main algorithm which has predicted the best results is the Decision Tree Algorithm(DTC). DT has a tree like structure where we start from a root of a tree. It has root node, decision node and a terminal node. Root node splits into decision node which further splits into terminal nodes or leaf. For the mushroom dataset, the Decision Tree Classifier is predicting the best scores.

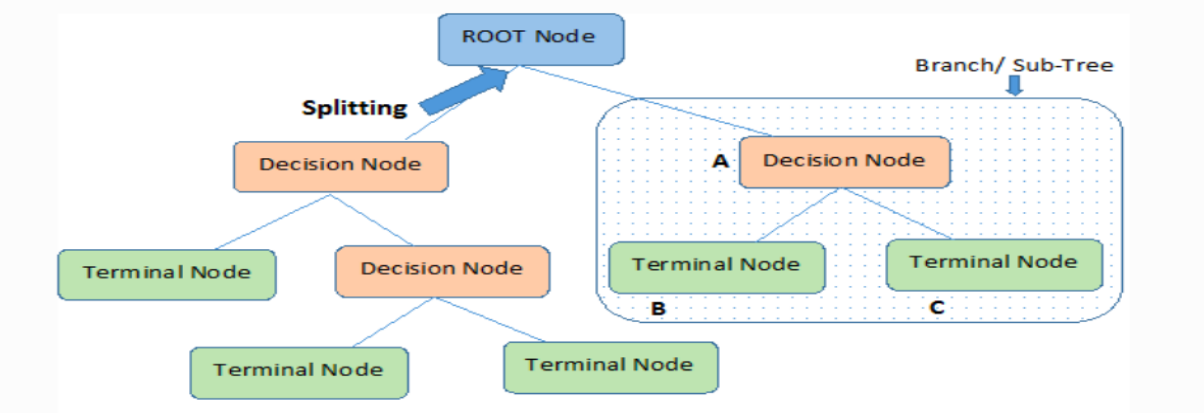


Figure 4: Decision Tree Algorithm

Evaluation Matrix

Evaluation matrix plays a very important role in evaluating the performance of any classification model. The metrices that are evaluated here are the accuracy score, ROC_AUC score, classification matrix which includes F1, recall, Precision and the confusion matrix. Time taken to test the model on dataset plays a very crucial role.

Here, for Decision Tree Classifier the accuracy score we are getting is 100% and the classification matrix i.e precision, F1, recall also scores to 100% and the confusion matrix clearly shows that there is no error in predicting the class of mushroom i.e it is predicting classes of mushrooms accurately.

```
Accuracy Score   1.0
Confusion Matrix:
 [[1234    0]
 [   0 1204]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1234
           1       1.00      1.00      1.00      1204

    accuracy                           1.00      2438
   macro avg       1.00      1.00      1.00      2438
weighted avg       1.00      1.00      1.00      2438
```

Figure 5: Results

As the classes of mushroom we will be predicting is either edible(e) or poisonous(p) which makes it a binary classification we have here plotted the ROC_AUC curve and ROC_AUC score. Higher the ROC_AUC score better the performance of model is. In this case AUC_ROC score is 1 which depicts that it model is performing accurately.

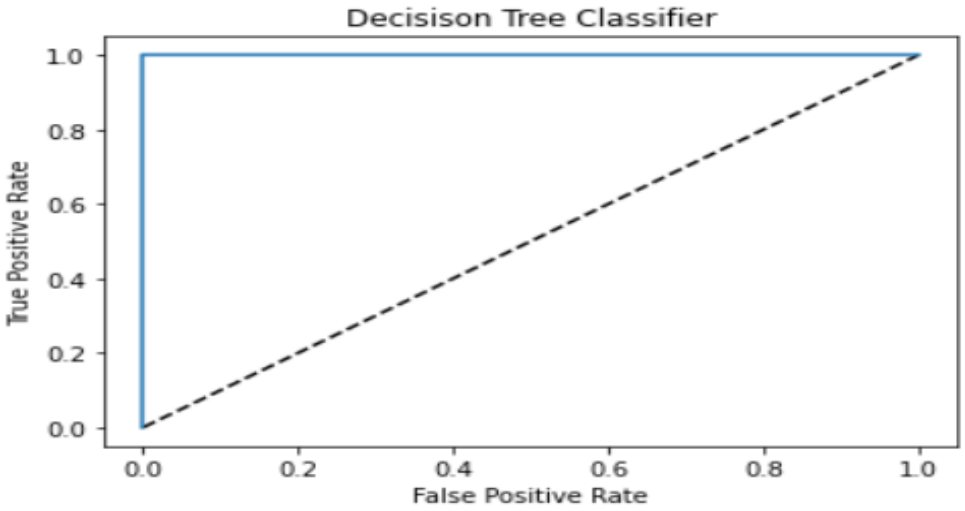


Figure 6: ROC_AUC CURVE

Conclusion

The above article states the methodology and the classification algorithms used for predicting the class of mushroom efficiently based on the various categorical features provided. We have analyzed the data through various graphical representations and cleaned the data before passing it to a classification model. We have developed 4-5 classification model and the best results were predicted by the Decision Tree Algorithm(DTC) with 100% accuracy and the confusion matrix shows that there is no error in predicting the class of mushroom.