Mushroom Edibility

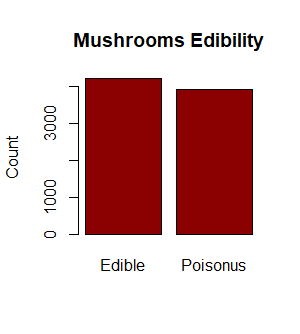
# Abstract:

We are doing the classification task using supervised learning model called Naive Bayes. The task is to classify the mushroom based on the features like cap shape, gill size and order of the mushroom plant for edibility purposes. We first go through the data discovery phase in which we find the types of our data. Next, we go to data preparation phase in which we discovered about the missing values, then we go to model planning phase in which we selected important features to be used in model training. In model building phase we applied the Naïve Bayes algorithm and got the 94% accuracy.

# Data Discovery Phase:

Our goal is to find the edibility of mushroom plant based on the given features. We hope to find the good features that fairly classify our data, with high accuracy. We also want to avoid the misclassification of the poisonous mushroom plant as edible plant.

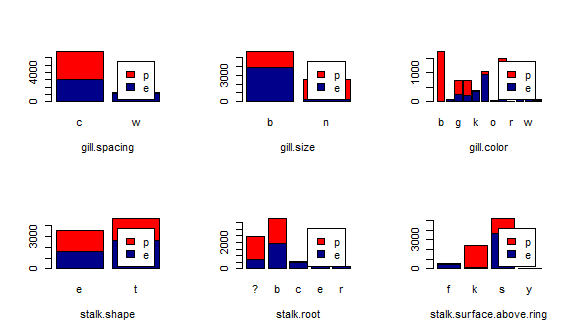
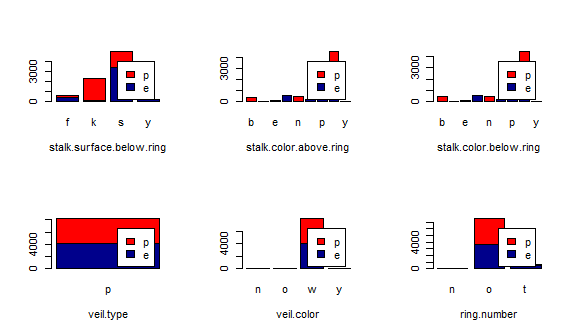
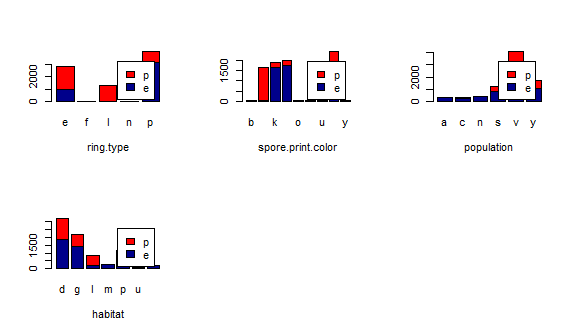
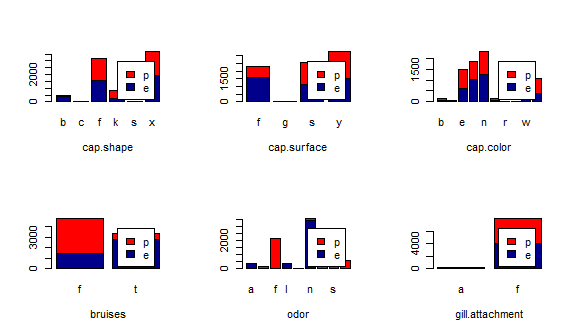
In this phase we tried to find inner insight of our data. First of all, we find the types of our features we are going to use in our model. We find out that all of our features are of categorical value, which is good because we do not have to deal with outliers in our data. The data is clean and have distinct vale of every feature.

As we know that we our target variable in class feature which tells us about the edibility of the mushroom plant. We plotted this feature to check, if we have the equal number of data example in our data. By checking this we can void the biasness of our model towards one class.

Figure

We can see in Figure 1, we have relatively same ratio of both edible and poisonous class in our data. In our cases we our model should have more poisonous example than edible.

As we know that we have categorical features, we can find some good feature that fairly divide in into edible and poisonous class. For this purpose, we plotted every feature with respect to our target variable to find good and bad features.

In figure 1, we can see that the order feature can be good featrue bacause there is less impurity in this feature. Each class of poisonus or edibile is can be classified based on oder. For example the foul smelling mushroom can be classified as poisonus because no mushroom that is eidbile have a foul smell.(In figure 1, f stand for foul smelling plant.)

Figure

Figure

Figure

Figure

In figure 3, the veil. Type feature will be useless because it only has one value which is not useful when classifying our data.

# Data Preparation Phase:

We have discovered our data and now in this phase we will prepare this data for our model training. We check the empty values for every column. There are no empty values in any column, so we do not have to deal with the null values.

We have made sure that we have supported type of our features in our data (for model building).

We have a perfectly type of data which do not need any type of data cleaning. There is no class type that is outside of data description, so every feature category has a meaning.

We only have a categorical data, so we do not have need to apply normalization in our data. This method is useful when we have numerical data.

# Model Planning Phase:

In this phase we have tired to find the usefulness of features using some syntactical method. We use the Information Gain to find the impurity in our dataset. This method is used in the Decision Tree for selection of the pure node, by apply this method we were able to sort the pure and impure feature. After that we deleted those features which have the lowest Information Gain for increasing our model accuracy and deceasing training load.

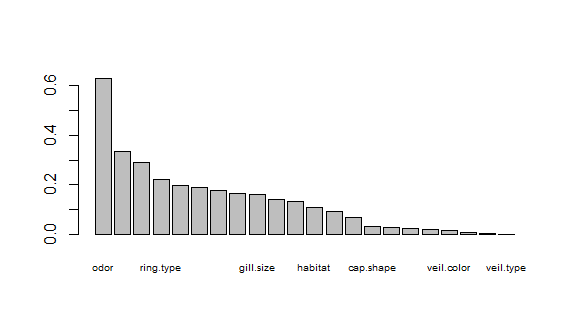
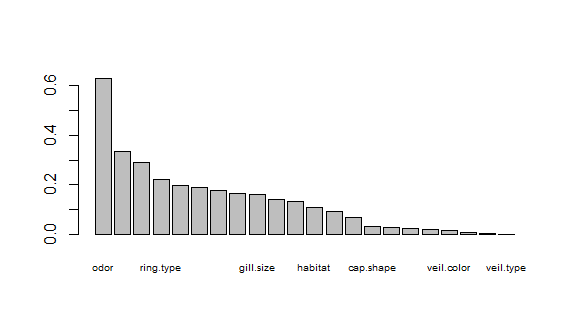


Figure Information Gain Chart

In the above figure we can see the features with the highest Information Gain and lowest Information Gain. By using this visualization, we selected the number of features to keep with high Information Gain value.

You can see in the figure 7, there are 14 features remain in our dataset out of 22 features. In this case our model will perform better.

After this we selected the Naïve Bayes algorithm to train our model.

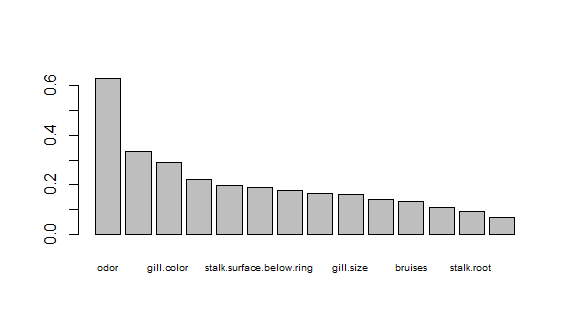


Figure Information Gain Chart (Remaining Features)

# Model Building Phase:

As we know that our problem is about two class classification and the Naïve Bayes classifier works best for these types of problem. Moreover, Naïve Bayes classifier require us to input the categorical features our data is already categorical (so do not need to convert continuous features to categorical). This algorithm is faster than other classification algorithm such as Logistic Regression, which require to compute the equation in the higher dimensions. The proclitic type is easy to compute and relatively faster than other algorithms.

# Project Results:

We splatted our data into 80% training set and 20% testing set, and trained our model on the training data. We have about 6500 training examples, which are enough for the good model to train. We trained our model and tested our model accuracy which comes about 94%. Which means about model is able to correctly classify most of the testing example in our testing data.

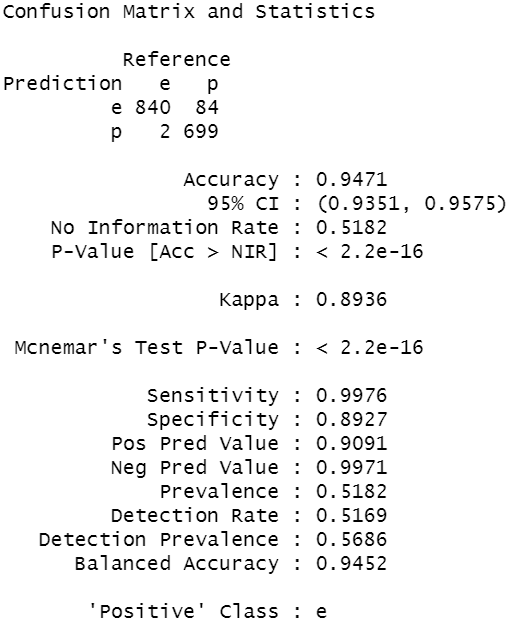
In figure 8, we can see our result of false positive which is high than false negative (which is not good in our case).

Figure Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | True Class | |
|  |  | Edible | Poisonous |
| Predicted Class | Edible | 840  (TP) | 84  (FP) |
| Poisonous | 2  (FN) | 699  (TN) |

Figure Confusion Matrix

Most of the testing example were correctly classified and some got misclassified be based on biasness in the training data.

We faced the problem regarding finding the accuracy of our model and building the confusion matrix for our model.

# Conclusion:

We can interpret our result by looking at confusion matrix of our model, which can tell us a lot about model reliability. If the same predicted and true class have larger number which means our model is predicting most example true. If the predicted and true class are not same and have large value it means our model is not doing well. In our case we have to correctly classify the poisonous mushroom because they can be lethal if we eat them. In our model are able to misclassify 84 testing examples as edible as oppose to misclassifying only 2 examples of edible class as poisonous which is not good. To make the model more reliable we have to introduce more poisonous training example in our data. For our problem we can use Decision Tree classifier as well, which can we a good alternative of Naïve Bayes classifier.