Renee Joyal
CSCI 347 - Final Project
05/05/2020

Exploration of Classification Algorithms

Data Description

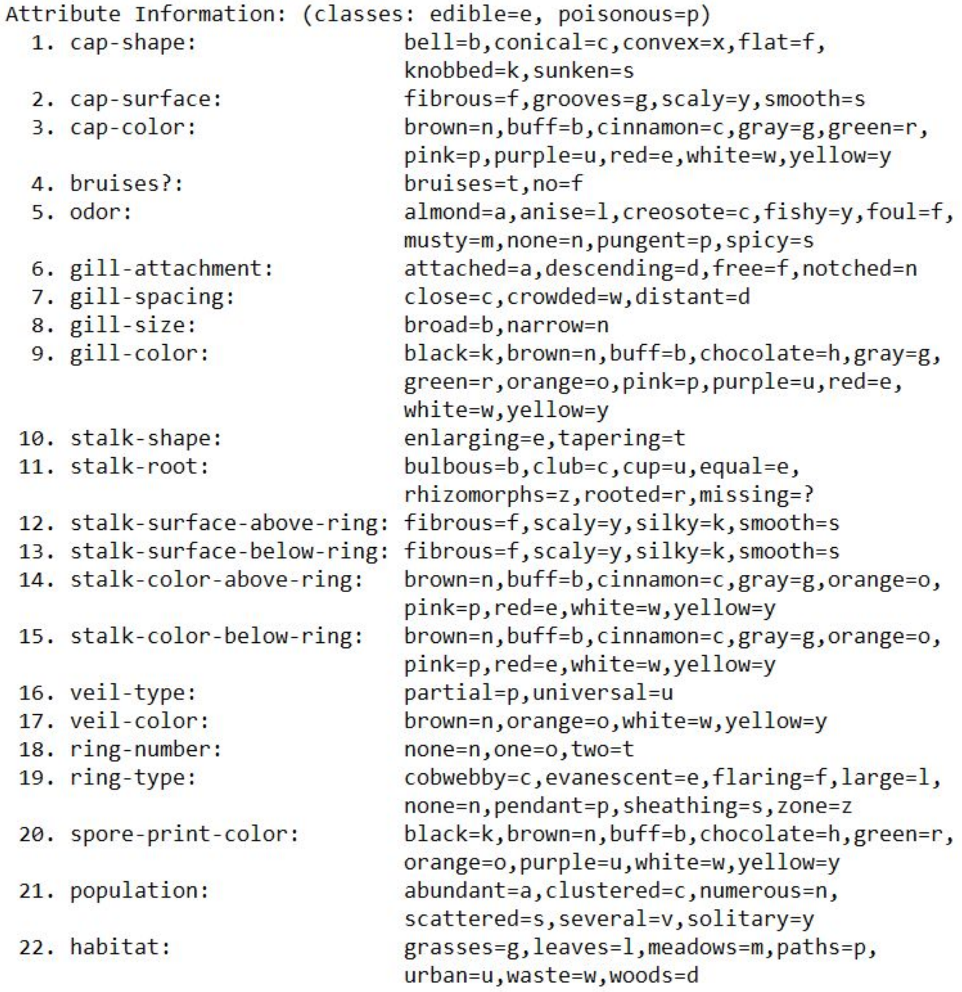
The problem that I will explore in this project is that of automizing the classification of mushrooms as poisonous or edible based on their observed physical attributes, habitat, and other various characteristics. I will be using the UCI Machine Learning Repository's "Mushroom Data Set" for this project. The data set is comprised of 8,124 instances, each with 22 categorical attributes. One of these categorical attributes is the classifier of poisonous or edible. There are 2,480 missing attribute values all located in attribute 11 (stalk-root).

Pre-Processing Techniques

To pre-process the data, I applied label encoding. I chose to use label encoding over one-hot encoding because each attribute has a number of different categories associated with them (see Figure A). There are many attributes such as 'cap-color' or 'gill-color' that have a myriad of possible colors. One-hot encoding would add over a hundred columns to the data. Thus, I determined that label encoding was the best way to pre-process this data.

The data includes the class of each mushroom as a categorical attribute. As a part of pre-processing the data, I extracted this class column into a separate pandas DataFrame to be used in the classification algorithms.

The data has 2,480 missing values located in attribute 11. Since all the missing values were located in a concentrated column, I decided to drop attribute 11 from the data to eliminate this problem.



```
Attribute Information: (classes: edible=e, poisonous=p)
  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-color:                brown=n,buff=b,cinnamon=c,gray=g,green=r,
                               pink=p,purple=u,red=e,white=w,yellow=y
  4. bruises?:                 bruises=t,no=f
  5. odor:                     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-color:               black=k,brown=n,buff=b,chocolate=h,gray=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-color-above-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,
                               pink=p,red=e,white=w,yellow=y
 15. stalk-color-below-ring:   brown=n,buff=b,cinnamon=c,gray=g,orange=o,
                               pink=p,red=e,white=w,yellow=y
 16. veil-type:                partial=p,universal=u
 17. veil-color:               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-color:        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
```

Figure A: Attribute Information

Data Mining Techniques

I will further explore classification algorithms with this project by using the Naïve Bayes and Support Vector Machine (SVM) algorithms to classify this data. In particular, I will be using Categorical Naive Bayes and SVC with a RBF kernel.

The argument for using Categorical Naive Bayes over other types is rather simple because this data is categorical rather than numerical. For the SVM, I had to decide between using SVC or LinearSVC. I decided to use SVC since it offers more flexibility for non-linear support vectors. Additionally, SVC has different kernel options (see Figure B). I went with the default RBF kernel as it appears to be a fairly dynamic kernel type.

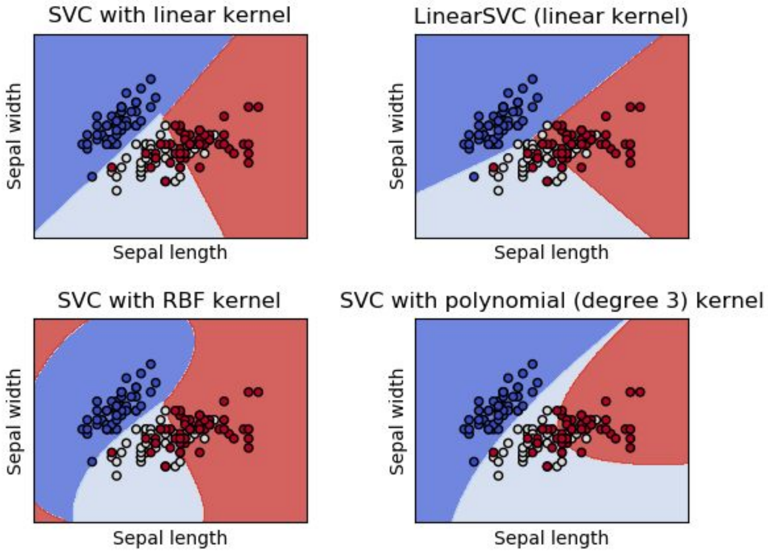


Figure B: Visualization of SVC kernel types
Citation: https://scikit-learn.org/stable/auto_examples/svm/plot_iris_svc.html

Analysis

Through my analysis, I discovered that the SVM algorithm performed slightly better than Naive Bayes. The accuracy of Naive Bayes on this data set is 95.2% while the accuracy of SVM is 98.8%. This means that SVM correctly predicted the classes of about 99% of the test data set. I am pleasantly surprised by how accurate both algorithms were. I expected at least one algorithm to be more than 90% accurate, but both were over 95% accurate.

These results are reflected in the confusion matrix of each algorithm's predictions (see Figures C and D). Naive Bayes predicted 4 edible mushrooms were poisonous and 74 poisonous mushrooms were edible. SVM predicted only 1 edible mushroom was poisonous and 19 poisonous mushrooms were edible.

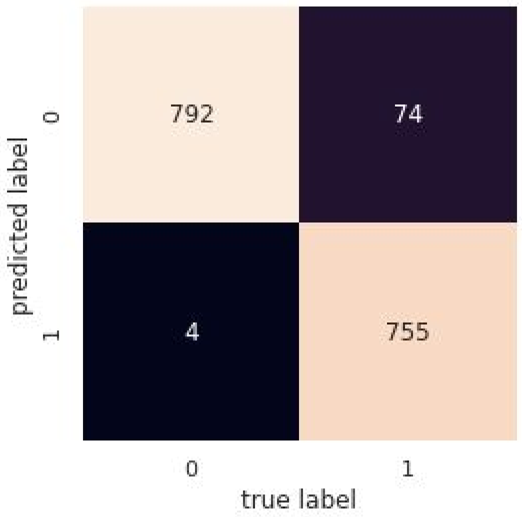


Figure C: Confusion Matrix of Naive Bayes (where ‘edible’ = 0 and ‘poisonous’ = 1)

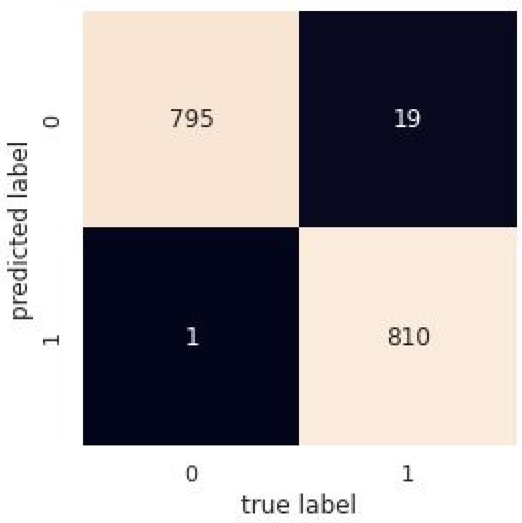


Figure D: Confusion Matrix of SVM (where ‘edible’ = 0 and ‘poisonous’ = 1)

Conclusion

Based on the analysis of the results of the Naive Bayes and SVM classification algorithms, SVM is the slightly more accurate method. It is certainly more desirable for as few poisonous mushrooms to be misclassified as possible, so SVM seems to be the better classification algorithm for this particular data set. Based on the results of this research, new mushrooms could potentially be tested using SVM and classified as edible or poisonous correctly. While not 100% accurate, the automated nature of the algorithm could speed up the classification of mushrooms. An advantage of using SVM to classify mushrooms is that it can easily use the data from thousands of other mushrooms to determine the edibility of a new sample.

The difference between the two classification algorithms is that Naive Bayes is based on probability while SVM is based on support vectors. Naive Bayes treats features as independent of each other. That might not be the most useful way of analysing this data set. There might be common sets of features that edible or poisonous mushrooms possess. SVM views features as less independent and bases its support vectors on the collection of data points. Things like common features among edible or poisonous mushrooms would be taken into account and improve the accuracy of SVM over Naive Bayes.

Further Work

There are several things that could be changed or improved about this project for further research. There could be additional research on comparing the types of SVM, for example. For this project, I removed the entire attribute that had missing values, but another project could experiment with different missing data handling techniques. For this project, I used the default parameters on the sklearn Naive Bayes and SVC modules. However, if I had more time, I would have liked to tweak the parameters to further improve the performance of the algorithms.

Acknowledgements

Schlimmer, J.S. (1987). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Mushroom]. Irvine, CA: University of California, School of Information and Computer Science.