Machine Learning Project Report
Course No: CSE-476

Problem Name: Mushroom Classification

Submitted By-

M. Tahmid Hossain
Reg. No: 2012331017

Md. Moshiur Rahman
Reg. No: 2012331049

Content

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom.

Inspiration

What types of machine learning models perform best on this dataset?

Acknowledgements

This dataset was originally contributed to the UCI Machine Learning repository.
Time period: Donated to UCI ML 27 April 1987.

About Data Set:

Total Data: 8125
Total Attributes: 22
Attribute Information: (classes: edible=e, poisonous=p)
cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y
bruises: bruises=t,no=f
odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s
gill-attachment: attached=a,descending=d,free=f,notched=n
gill-spacing: close=c,crowded=w,distant=d
gill-size: broad=b,narrow=n
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
stalk-shape: enlarging=e,tapering=t
stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?
stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
stalk-color-above-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
stalk-color-below-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y
veil-type: partial=p,universal=u
veil-color: brown=n,orange=o,white=w,yellow=y
ring-number: none=n,one=o,two=t
ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z
spore-print-color:
black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y

population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y
habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

Our Work Procedure:

If we just read and give the data to the classifiers and run our experiments then the result comes is not satisfactory. So we shuffled the data just after reading it from the file. Then we divided the data according to its class properly. Then we ran our test on different classifiers and the results are mentioned below.

Classifier Details and Result:

Naive Byes:

Classifier Details:

GaussianNB(priors=None)

Result:

Naive Byes accuracy score = 89.2036124795%

Naive Byes precision score = 89.402455468%

Naive Byes recall score = 89.3483883509%

Naive Byes f1 score = 89.2029556426%

Support Vector Machine:

Classifier Details:

SVC(C=25, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape=None, degree=3, gamma='auto', kernel='linear', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)

Validation Result:

SVM accuracy score = 97.5349219392%

SVM precision score = 97.5919650192%

SVM recall score = 97.5811637433%

SVM f1 score = 97.5349069598%

Testing Result:

SVM accuracy score = 96.3054187192%

SVM precision score = 96.3826366559%

SVM recall score = 96.4898595944%

SVM f1 score = 96.3045194628%

Decision Tree:

Classifier Details:

DecisionTreeClassifier(class_weight='balanced', criterion='gini', max_depth=20, max_features='log2', max_leaf_nodes=None, min_impurity_split=1e-07, min_samples_leaf=1, min_samples_split=5, min_weight_fraction_leaf=0.0, presort=False, random_state=251254, splitter='best')

Result:

Decision tree accuracy score = 100.0%

Decision tree precision score = 100.0%

Decision tree recall score = 100.0%

Decision tree f1 score = 100.0%

Neural Network:

Classifier Details:

MLPClassifier(activation='logistic', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=False, epsilon=1e-08, hidden_layer_sizes=(20, 20, 20), learning_rate='constant', learning_rate_init=0.001, max_iter=200, momentum=0.9, nesterovs_momentum=True, power_t=0.5, random_state=False, shuffle=True, solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False, warm_start=False)

Validation Result:

Neural network accuracy score = 100.0%

Neural network precision score = 100.0%

Neural network recall score = 100.0%

Neural network f1 score = 100.0%

Testing Result:

Neural network accuracy score = 100.0%

Neural network precision score = 100.0%

Neural network recall score = 100.0%

Neural network f1 score = 100.0%

Voting System:

We developed a voting system with five different types of neural network classifiers, naive byes classifier, svm classifier. Details of these classifiers are mentioned below:

Neural Network:

clf1 = (activation='logistic', alpha=.0001, hidden_layer_sizes=(20, 20, 20,), random_state=False)

clf2 = (activation='relu', alpha=.0001, hidden_layer_sizes=(20, 20, 20,), random_state=False)

clf3 = (activation='logistic', alpha=.0001, hidden_layer_sizes=(15, 15, 15,), random_state=False)

clf4 = (activation='logistic', alpha=.0001, hidden_layer_sizes=(100, 100,), random_state=False)

clf5 = (activation='logistic', alpha=.0001, hidden_layer_sizes=(50, 50,50), random_state=False)

SVM:

clf6 = (kernel="linear", C=25)

Naïve Byes:

clf7 = GaussianNB

Validation Result:

Voting system accuracy score = 100.0%

Voting system precision score = 100.0%

Voting system recall score = 100.0%

Voting system f1 score = 100.0%

Testing Result:

Voting system accuracy score = 100.0%
Voting system precision score = 100.0%
Voting system recall score = 100.0%
Voting system f1 score = 100.0%