Report Title
CSE 454 Data Mining

- Subtitle -
Mushroom Classification

Project Report

Student Name/Number
Muhammed Alperen Karaçete
171044052

Gebze Technical University
Computer Engineering

Contents

Chapter 1

1-)Project Topic

1.0.1 Problem Definition and Details

The problem is determining which mushroom is given edible which is poisonous. The data objects are the mushrooms, the attributes are the features of the mushrooms and the class attribute is the poisonous. There are 23 species of mushrooms.

For this project I found a dataset created for a research. The dataset is a dataset taken from Irvine University archive. The dataset format is like below. Each row represents a mushroom. The last attribute determines the class and rest of the attributes are the features of the mushroom.

Attributes are: cap-shape,cap-surface,cap-color,bruises,odor,gill-attachment,gill-spacing,gill-size,gill-color,stalk-shape,stalk-root,stalk-surface-above-ring,stalk-surface-below-ring,stalk-color-above-ring,stalk-color-below-ring,veil-type,veil-color,ring-number,ring-type,spore-print-color,population,habitat,poisonous

In Poisonous attribute, there are 2 classes. Poisonous (p) and Edible (e).

1.0.2 How To Solve Problem

The dataset has give very good results on Decision Tree,Random Forest algorithm.But I wanted to less good results for trying increase of accuracy rate. I choose logistic regression classifier as my method of solution. The reason why I chose logistic regression classifier method because dataset is already properly labeled, and it does not give as good results as like decision tree classifier.And because my my poisonous attribute has two value this means thath is a binary classification problem and logistic regression classifier is a suitable choice for a binary classification problem.

1.0.3 Problem Solution

In my project I used various methods to improve the initial results. The initial training set results are from using logistic regression classifier directly on the given dataset. After that I applied different methods for increasing accuracy and after every method I calculate and compare accuracies to each other.It is expected to get better results after each method.

In the end it is expected to get a better accuracy compared to the initial accuracy.The accuracies might slightly differ after each different training because of the random oversampling and undersampling methods.

Chapter 2

2-) Problem Solution

2.0.1 Data Preprocessing

While examining values in columns, I saw that there is null values in stalk-root column. There are total 2408 instances with missing values. My database is consist of categorical data so I used mode (most frequent value). For "stalk-root",
bulbous is represented as b,club is represented as c,cup is represented as u,equal is represented as e, rhizomorphs is represented as z,rooted is represented as r,missing values are is represented as ?
b = 3776 rows
c = 556 rows
e = 1120 rows
r = 192 rows
So I replaced ? value with b.

My dataset is consist of Categorical variables (e.g., mushroom cap color).So I converted them to the numerical values with Label Encoder.

When examining datas for each attribute I see that all values in veil-type is same so I deleted that column because It did not effect anything.

Finally I have control for class balance of poisonous attribute. I found that they are very close to each other.
edible: 4208 (51.8%)
poisonous: 3916 (48.2%)
total: 8124 instances
Figure is in below.

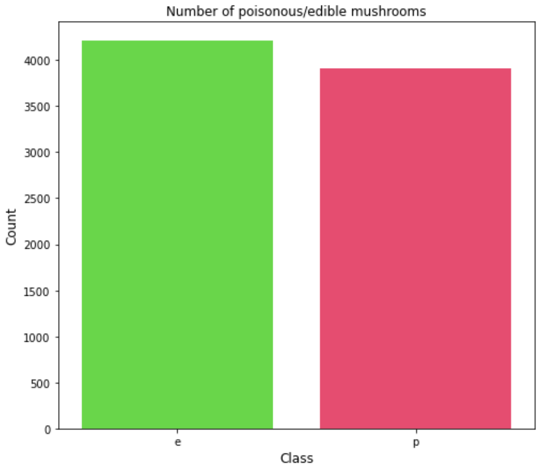


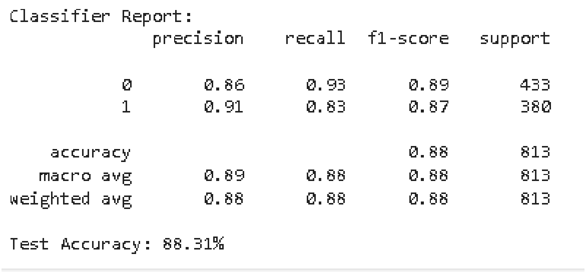
Figure 2.1: Classes

2.1 Logistic Regression Classifier

I implemented the Logistic Regression Classifier myself. It is implemented in my notebook. This algorithm is used for binary classification problems, where the target variable has two classes (0 or 1). The logistic regression model learns to map input features to probabilities and makes predictions based on a threshold (commonly 0.5). If the predicted probability is greater than the threshold, the instance is classified as class 1; otherwise, it is classified as class 0.

Firstly, I created the initialization function. Then I created the sigmoid function, which maps any real-valued number to a value between 0 and 1.

Figure 2.2: Classifier Accuracy



```
Classifier Report:
              precision    recall  f1-score   support

           0       0.86      0.93      0.89       433
           1       0.91      0.83      0.87       380

    accuracy                           0.88       813
   macro avg       0.89      0.88      0.88       813
weighted avg       0.88      0.88      0.88       813

Test Accuracy: 88.31%
```

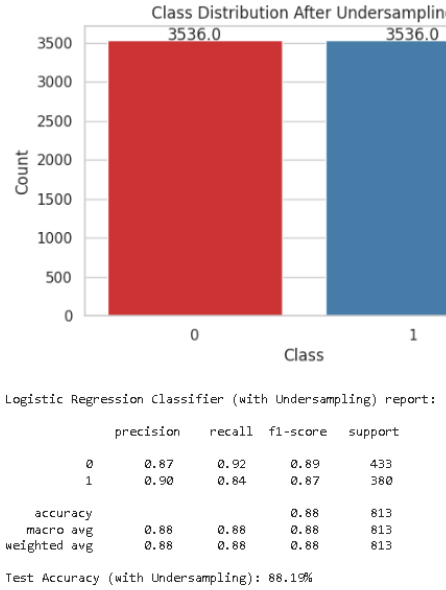
Chapter 3

Enchantment for Results

3.0.1 Undersampling:

Randomly remove samples from the majority class, with or without replacement. ** Under Sampling is decreased accuracy by %0.12

Figure 3.1: Undersampling Accuracy



```
Logistic Regression Classifier (with Undersampling) report:

              precision    recall  f1-score   support

           0       0.87      0.92      0.89       433
           1       0.90      0.84      0.87       380

    accuracy                           0.88       813
   macro avg       0.88      0.88      0.88       813
weighted avg       0.88      0.88      0.88       813

Test Accuracy (with Undersampling): 88.19%
```

3.0.2 Oversampling:

Random Oversampling involves supplementing the training data with multiple copies of some of the minority classes. ** Over Sampling is increased accuracy by %3.20

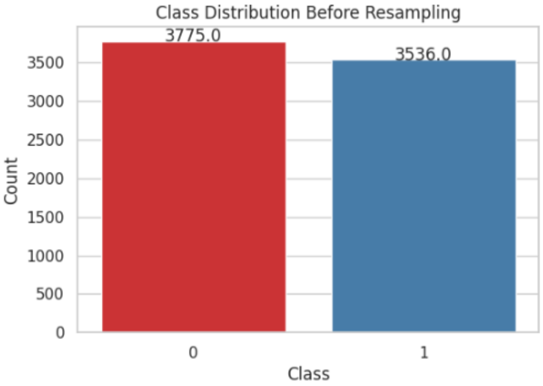
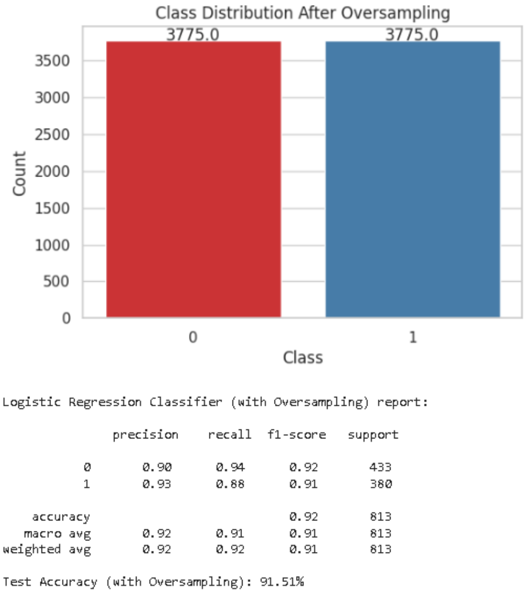


Figure 3.2: Datas Before Oversampling

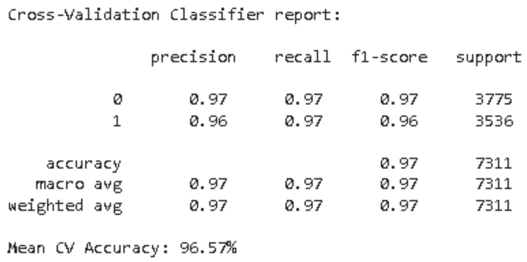


```
Logistic Regression Classifier (with Oversampling) report:

              precision    recall  f1-score   support

           0       0.90      0.94      0.92       433
           1       0.93      0.88      0.91       380

    accuracy                           0.92       813
   macro avg       0.92      0.91      0.91       813
weighted avg       0.92      0.92      0.91       813

Test Accuracy (with Oversampling): 91.51%
```

Figure 3.3: Oversampling Accuracy

3.0.3 Crossvalidation:

To have even better results both for training and testing I decided to create my model with cross validation instead. I created 5 different test sets and 5 different corresponding train sets. After training a model with each of them, combined their results and used the combined results as a new model. Thanks to this, biases are reduced. Cross validation is increased accuracy by %8.37



```
Cross-Validation Classifier report:

              precision    recall  f1-score   support

           0       0.97      0.97      0.97      3775
           1       0.96      0.97      0.96      3536

    accuracy                           0.97      7311
   macro avg       0.97      0.97      0.97      7311
weighted avg       0.97      0.97      0.97      7311

Mean CV Accuracy: 96.57%
```

Figure 3.4: CrossValidation Accuracy

3.0.4 Conclusion:

My class was nearly balanced so oversampling and undersampling did not increased or decreased accuracy so much. Cross validation with folds=5 increased model accuracy so much. So I use Cross Validation for my final analysis.

Chapter 4

Resources

- Logistic Regression on ScienceDirect: ScienceDirect Link
- Mushroom Dataset: Mushroom Dataset Link
- Data Mining Techniques Evaluation Metrics: Data Mining Techniques Evaluation Metrics Link