**Secondary Mushroom Data Set**

**Introduction:**

The Secondary Mushroom Data Set is a collection of information about various mushroom species, including their physical characteristics, habitat, and toxicity. It was originally created by Jeff Schlimmer in 1987 to test machine learning algorithms for classification tasks.

The dataset contains 61609 instances of mushrooms, each described by 21 attributes such as cap shape, cap color, gill size, and stem shape. The target variable is whether the mushroom is edible or poisonous, with 2 possible values: "e" for edible and "p" for poisonous.

The dataset has been widely used in machine learning research as a benchmark for classification algorithms. Its high dimensionality, nonlinearity, and class imbalance make it a challenging problem for machine learning algorithms to solve.

**Research problem:**

The research problem of the Secondary Mushroom Data Set is to develop accurate classification models that can distinguish between edible and poisonous mushrooms based on their physical characteristics. The high dimensionality and nonlinearity of the dataset make it a challenging problem for machine learning algorithms to solve.

The dataset poses several research questions, such as:

1. What are the most important physical features of mushrooms that can be used to predict their edibility?
2. Which classification model is most effective for accurately predicting the edibility of mushrooms?
3. How does the accuracy of the classification model vary with different feature selection techniques?

**Objective of this project:**

The aim of the Secondary Mushroom Data Set is to use the physical attributes of mushrooms to classify them as either edible or poisonous. The dataset comprises 21 variables and includes 61,069 hypothetical mushrooms, with each cap based on one of 173 species (averaging 353 mushrooms per species). The response variable, or target variable, is a binary class indicating whether the mushroom is edible or poisonous (coded as "e" or "p"), and there are 20 predictor variables included in the dataset.

**Project Dataset:**

In this project, we will use the **Secondary mushroom data.** [**http://archive.ics.uci.edu/ml/datasets/Secondary+Mushroom+Dataset**](http://archive.ics.uci.edu/ml/datasets/Secondary+Mushroom+Dataset)**.**

Dataset of simulated mushrooms for binary classification into edible and poisonous.

The dataset contains a binary class variable that indicates whether a mushroom is edible ("e") or poisonous ("p"). The poisonous class also includes mushrooms with unknown edibility. There are 20 additional variables, which include:

1. Cap diameter (measured in centimeters)
2. Cap shape: bell ("b"), conical ("c"), convex ("x"), flat ("f"), sunken ("s"), spherical ("p"), or other ("o")
3. Cap surface: fibrous ("i"), grooved ("g"), scaly ("y"), smooth ("s"), shiny ("h"), leathery ("l"), silky ("k"), sticky ("t"), wrinkled ("w"), or fleshy ("e")
4. Cap color: brown ("n"), buff ("b"), gray ("g"), green ("r"), pink ("p"), purple ("u"), red ("e"), white ("w"), yellow ("y"), blue ("l"), orange ("o"), or black ("k")
5. Whether the mushroom bruises or bleeds: yes ("t") or no ("f")
6. Gill attachment: adnate ("a"), adnexed ("x"), decurrent ("d"), free ("e"), sinuate ("s"), pores ("p"), none ("f"), or unknown ("?")
7. Gill spacing: close ("c"), distant ("d"), none ("f")
8. Gill color: same as cap color plus none ("f")
9. Stem height (measured in centimeters)
10. Stem width (measured in millimeters)
11. Stem root: bulbous ("b"), swollen ("s"), club ("c"), cup ("u"), equal ("e"), rhizomorphs ("z"), rooted ("r")
12. Stem surface: same as cap surface plus none ("f")
13. Stem color: same as cap color plus none ("f")
14. Veil type: partial ("p") or universal ("u")
15. Veil color: same as cap color plus none ("f")
16. Whether the mushroom has a ring: yes ("t") or no ("f")
17. Ring type: cobwebby ("c"), evanescent ("e"), flaring ("r"), grooved ("g"), large ("l"), pendant ("p"), sheathing ("s"), zone ("z"), scaly ("y"), movable ("m"), none ("f"), or unknown ("?")
18. Spore print color: brown ("n"), buff ("b"), gray ("g"), green ("r"), pink ("p"), purple ("u"), red ("e"), white ("w"), yellow ("y"), blue ("l"), orange ("o"), or black ("k")
19. Habitat: grasses ("g"), leaves ("l"), meadows ("m"), paths ("p"), heaths ("h"), urban ("u"), waste ("w"), or woods ("d")
20. Season: spring ("s"), summer ("u"), autumn ("a"), or winter ("w")

**Challenges:**

1. To handle the class variable, dummy variables were created using the as.factor() function.
2. Due to the large size of the dataset, consisting of 61069 observations, it is challenging to build a regression model using all 20 predictor variables.
3. The selection of predictor variables is another challenge that needs to be addressed.
4. Identifying the optimal model is also a challenging task.

**Models Developed:**

|  |  |  |  |
| --- | --- | --- | --- |
| model | AIC | BIC | Accuracy |
| cap.color | 80155.00 | 80263.24 | 0.5805 |
| cap.diameter | 81704.07 | 81722.11 | 0.5836 |
| cap.shape | 81551.35 | 81614.49 | 0.5805 |
| cap.surface | 80402.01 | 80510.24 | 0.5836 |
| does.bruise.or.bleed | 83901.41 | 83919.45 | 0.4450 |
| gill.attachment | 80822.46 | 80894.61 | 0.6184 |
| gill.spacing | 83133.72 | 83169.80 | 0.5780 |
| Gill.color | 81666.87 | 81775.10 | 0.5940 |
| Stem.height | 81666.87 | 81775.10 | 0.5940 |
| Stem.width | 81446.70 | 81464.74 | 0.5982 |
| Stem.root | 79519.65 | 79573.77 | 0.5722 |
| Stem.surface | 78622.89 | 78704.07 | 0.5724 |
| Stem.color | 78821.98 | 78939.23 | 0.6219 |
| Veil.type | 83751.86 | 83769.90 | 0.4450 |
| Veil.color | 81384.42 | 81447.56 | 0.5635 |
| Has.ring | 83721.91 | 83739.95 | 0.4450 |
| Ring.type | 80456.88 | 80538.06 | 0.5746 |
| Spore.print.color | 80640.87 | 80713.03 | 0.5639 |
| habitat | 81702.84 | 81775.00 | 0.5741 |
| season | 83096.91 | 83132.99 | 0.5795 |

1. Since the problem at hand is a classification problem, logistic regression is deemed the most suitable model for the mushroom dataset. Initially, logistic regression was performed separately for all 20 variables, such as cap color, to examine the single variable's impact on the model's outcome. The outcomes of the 20 models are available.

The accuracy of the single variable models ranges from around 44% to 62%. Additionally, the AIC and BIC values for these models are around 80000.

2.A model was created based on similar variables, such as cap color and diameter combined, to observe their combined impact on the model's outcome. The outcomes of these models are available.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | AIC | BIC | Accuracy |
| CAP1 | 79697.49 | 79769.65 | 0.5956213 |
| CAP2 | 76405.14 | 76612.59 | 0.6234096 |
| GILL | 76396.27 | 76576.67 | 0.6592972 |
| STEM1 | 75737.77 | 75827.97 | 0.6340697 |
| STEM2 | 78553.32 | 78679.59 | 0.6301724 |
| STEM3 | 79519.65 | 79573.77 | 0.5722543 |
| VEIL | 80401.53 | 80473.69 | 0.5808839 |
| RING | 79369.24 | 79459.44 | 0.5746451 |
| SPORE+DOES | 80642.63 | 80723.80 | 0.5639686 |
| HABITAT+SEASON | 81010.89 | 81110.11 | 0.588269 |

In comparison to the single variable models, the AIC and BIC values have been improved to around 75000, and the accuracy has increased to 65%.

3. By considering p-values less than 0.05, a reduced model consisting of only seven variables was created, which includes cap diameter, cap shape, cap color, stem width, stem height, veil type, and has ring.

The result of the model is.

AIC: 74731

BIC:74938.31

Reduced model Logistic Accuracy=0.66922

The AIC and BIC values for the reduced model are 74731 and 74938.31, respectively. The accuracy of the reduced model is 0.66922, which is relatively high compared to the accuracy of the single variable models and grouped models. Therefore, we can conclude that the reduced model is an improvement over the other models.

4. QDA accuracy=0.65156

5. LDA accuracy=0.6658

6. Naïve bayes accuracy=0.6011069

7.ROC curves for Naïve bayes, LDA, QDA and logistic are

Graphical user interface, chart

Description automatically generated

logistic regression is represented by RED color curve.

NB is represented by Black color curve.

LDA is represented by Blue color curve

QDA is represented by Green color curve.

8. AUC scores for the reduced models of logistic , NB,LDA and QDA are

Chart, line chart

Description automatically generated

**Result:** From the confusion matrix of the reduced model, we can observe that the model has correctly predicted 15660 instances of edible mushrooms and 25209 instances of poisonous mushrooms, while it has misclassified 11521 instances of edible mushrooms as poisonous and 8679 instances of poisonous mushrooms as edible. The AUC scores of logistic regression, LDA, QDA, and NB are 0.7196, 0.7196, 0.7147, and 0.681, respectively. Based on these results, we can conclude that the reduced model is performing better than other models, and it has an acceptable accuracy of around 67%. However, there is still room for improvement in the accuracy of the model, which can be achieved by further optimizing the feature selection and model parameters.

**Conclusion:** Based on the analysis of the mushroom dataset, we have found that logistic regression and LDA models have the highest AUC scores of 0.7196, followed closely by the QDA model with an AUC score of 0.7147. The reduced model, which includes only seven predictor variables, has improved AIC and BIC values of 74731 and 74938.31, respectively, compared to the other models. Moreover, the reduced model has a relatively high accuracy of 0.66922, which is higher than the accuracy of single variable and grouped models. However, the accuracy of the Naïve Bayes model is relatively low at 0.6011069. Therefore, we can conclude that the logistic regression, LDA, and QDA models are suitable for predicting the edibility of mushrooms, while the reduced model with seven predictor variables is an improvement over other models in terms of accuracy and AIC/BIC values.

**Future applications:**

1. There are several potential future applications and uses of the secondary mushroom data set, including:
2. Developing more accurate predictive models: Future researchers can use this dataset to develop more accurate predictive models for determining the edibility of mushrooms based on their physical characteristics. These models could be useful for mushroom hunters and consumers who want to avoid poisonous mushrooms.
3. Enhancing food safety: The data set could be used to develop tools for identifying potentially harmful mushrooms in the food supply chain. This could help enhance food safety by preventing the consumption of poisonous mushrooms.
4. Analyzing trends in mushroom growth: The data set could be used to study trends in the growth and distribution of different mushroom species over time. This could be useful for understanding the ecology of mushrooms and their role in different ecosystems.
5. Advancing machine learning techniques: The data set could be used to develop and test new machine learning algorithms and techniques for classification and regression analysis.
6. Educating the public: The data set could be used to create educational materials about mushrooms and their characteristics, such as field guides for mushroom hunters or resources for schools teaching biology or ecology.

**Refrence :** Dennis Wagner, Dr. G. Hattab, 'Mushroom data creation, curation, and simulation to support classification tasks' in Scientific Reports on 14.04.2021