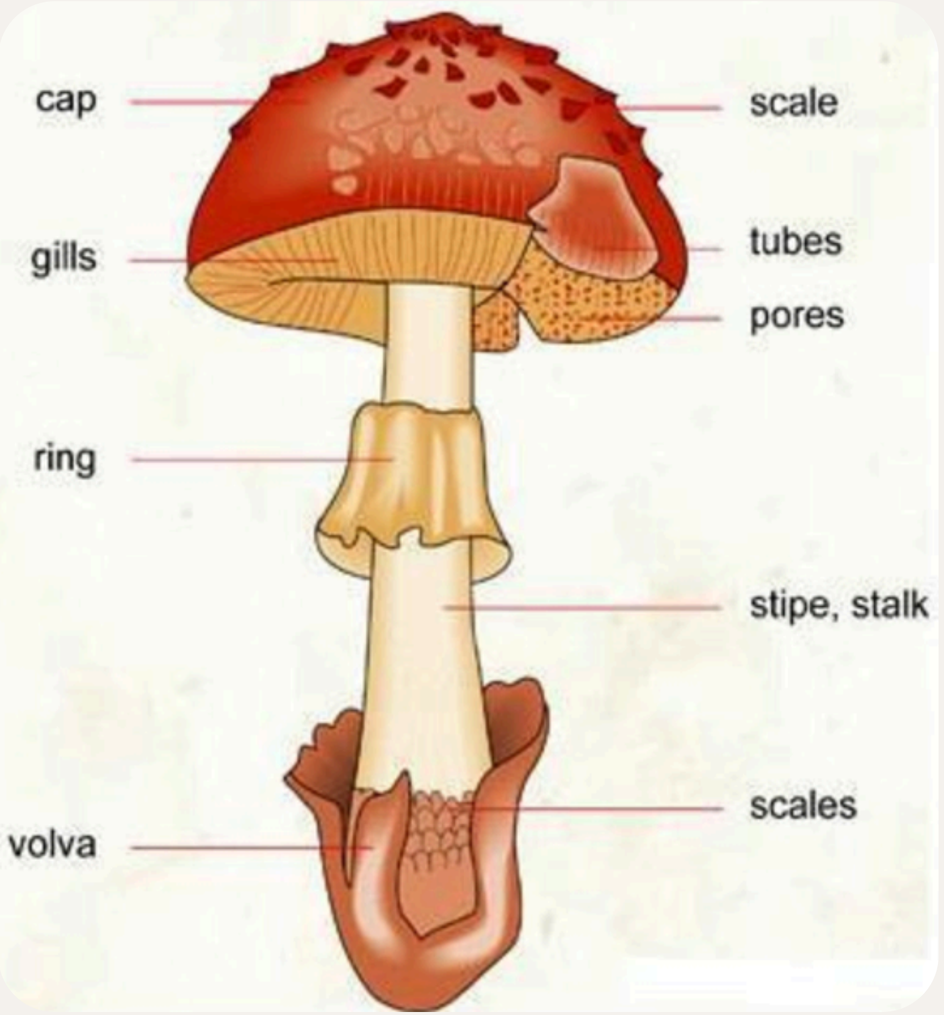
Classification of Poisonous Mushrooms

BESTE ÜNAL
ZEYNEP AFRA SEZER

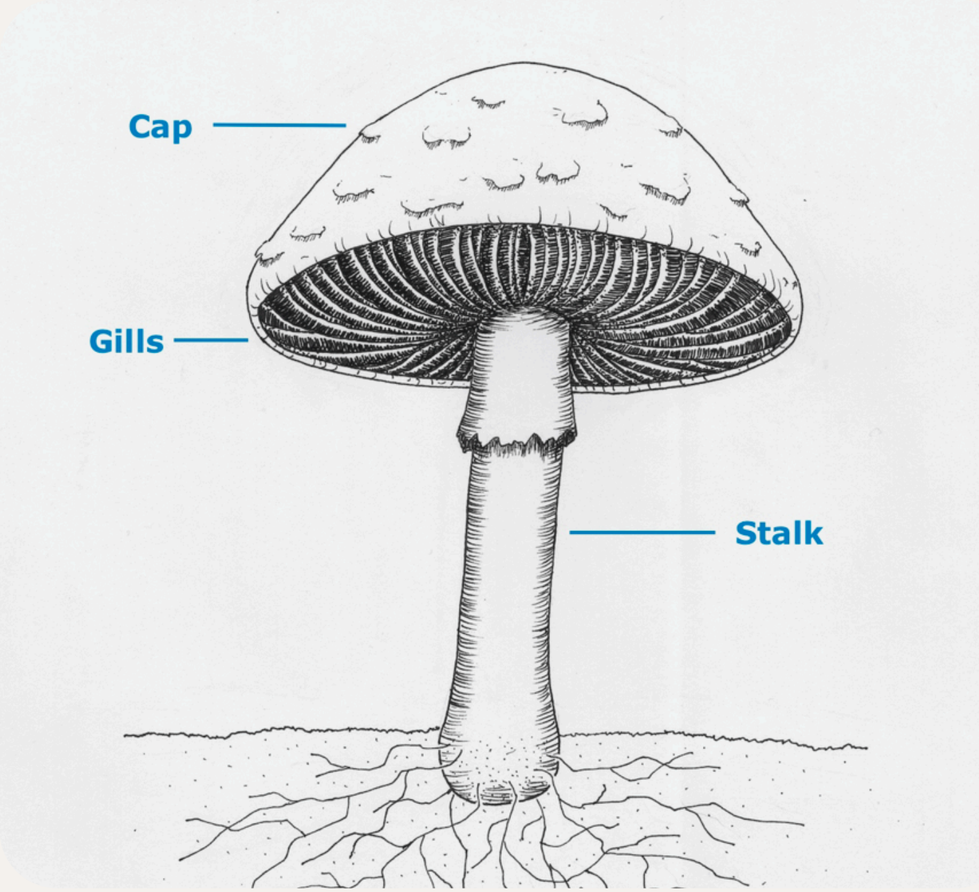
MUSHROOM DATASET

- class,
- hat shape,
- hat surface,
- what color,
- bruises,
- odor,
- gill attachment,
- gill spacing,
- gill size,
- gill color,
- stem shape,
- stem root,
- above-ring stem surface,
- below-ring stem surface,
- above-ring stem color,
- below-ring stem color,
- cover type,
- cover color,
- number of rings,
- ring type,
- spore print color,
- population,
- habitat.

The Mushroom Dataset includes 8124 mushrooms of various species, both edible and poisonous.



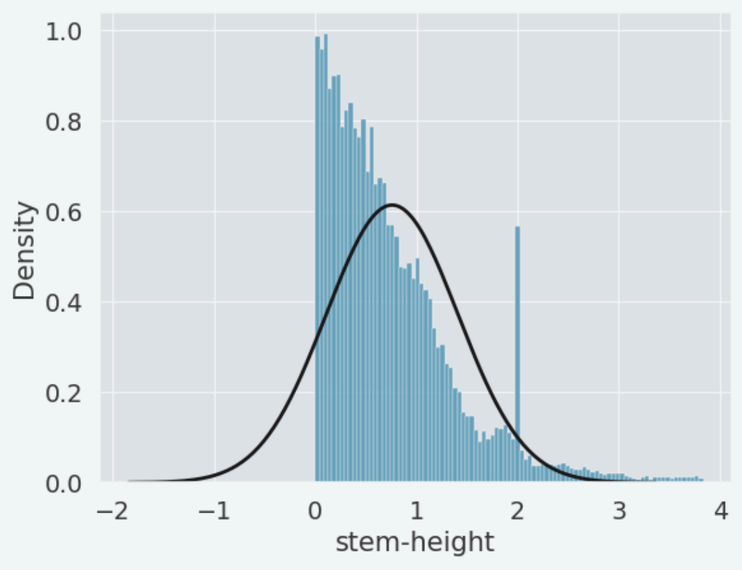
MUSHROOM_CLANED DATASET

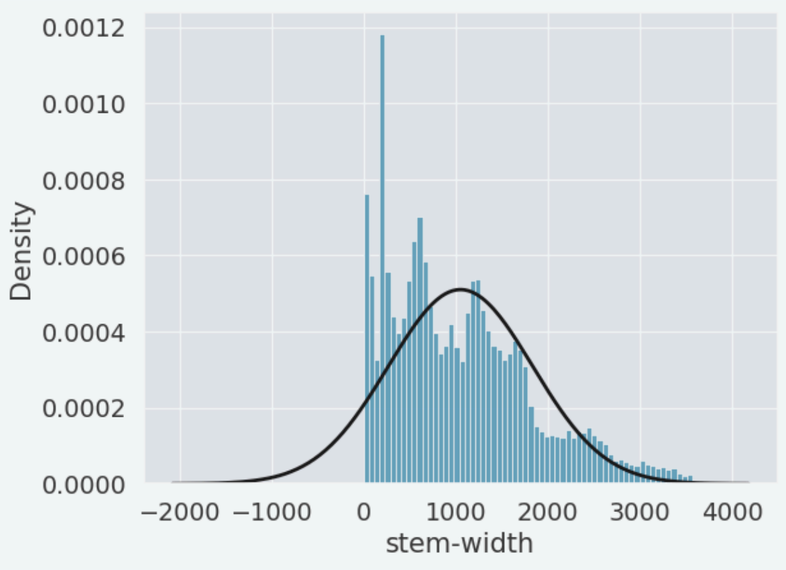


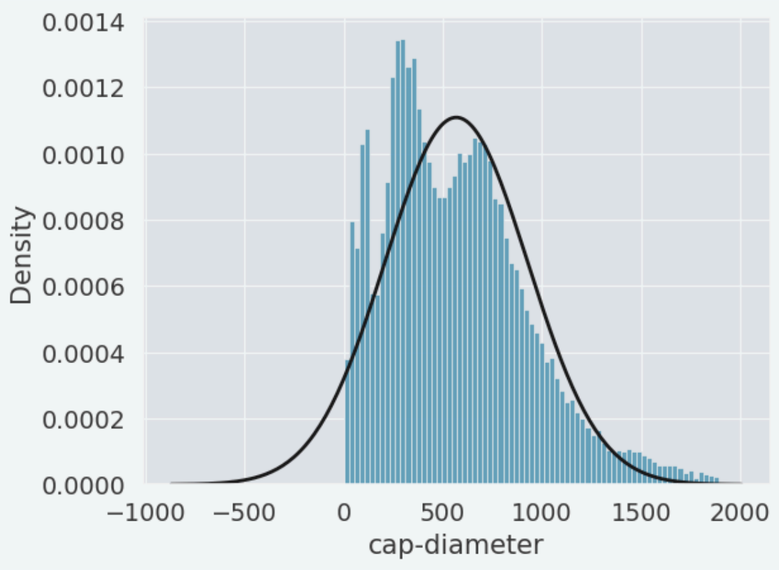
This dataset is a cleaned up version of the original Mushroom Dataset to identify which features are most indicative of a poisonous mushroom.

·cap-diameter: Diameter of the cork cap
·cap-shape: Mushroom cap shape
·gill-attachment: Attachment of the lamellae (thin leaves of the cork) to the cork
·gill-color: Color of the lamellae
·stem-height: Height of the mushroom stem
·stem-width: Width of the mushroom stem
·stem-color: Color of the mushroom stem
·season: Season.
·class: Mushroom class The Target Class contains two values- 0 or 1- where 0 refers to edible and 1 refers to poisonous.

MUSHROOM DATASET

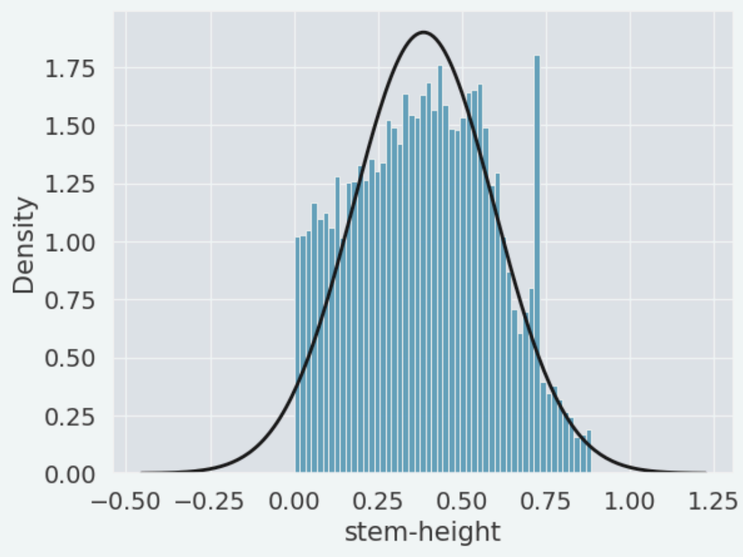


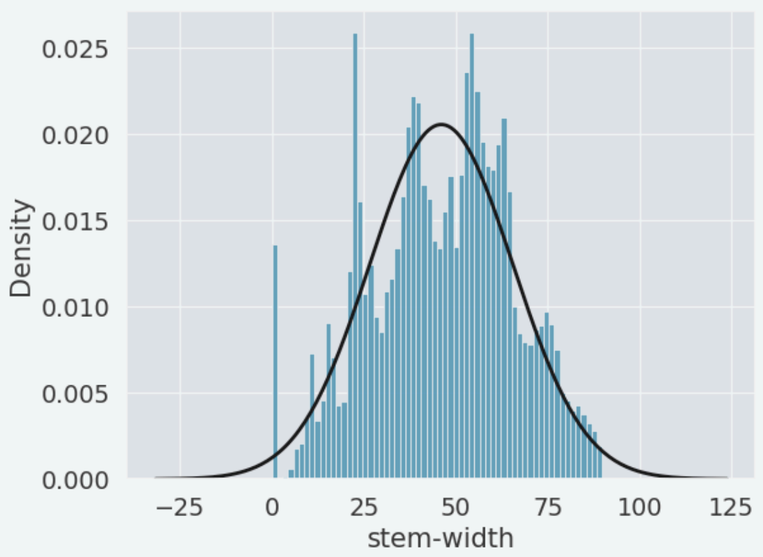


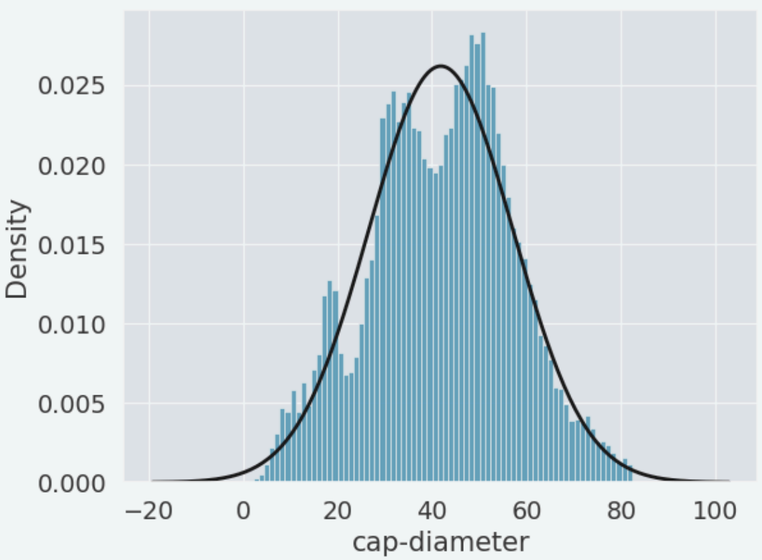


In the Mushroom Dataset, Modal imputation was used to fill in the missing data and one-time encoding to represent the categorical data in binary format. The data was then cleaned up using various techniques, such as z-score normalization to scale the mean to 0 and the standard deviation to 1, and feature selection to improve the model's performance and make it run faster.

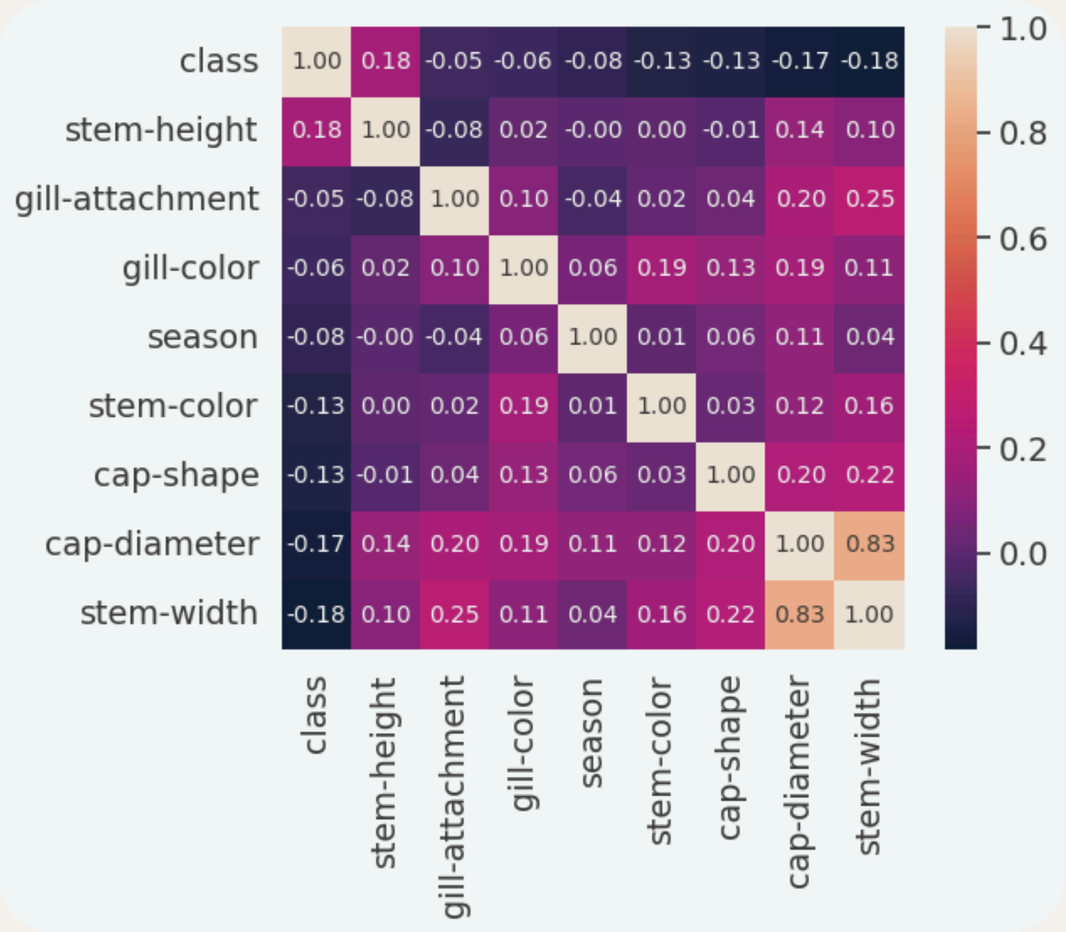
MUSHROOM_CLANED DATASET







THE CORRELATION BETWEEN THE VARIABLES

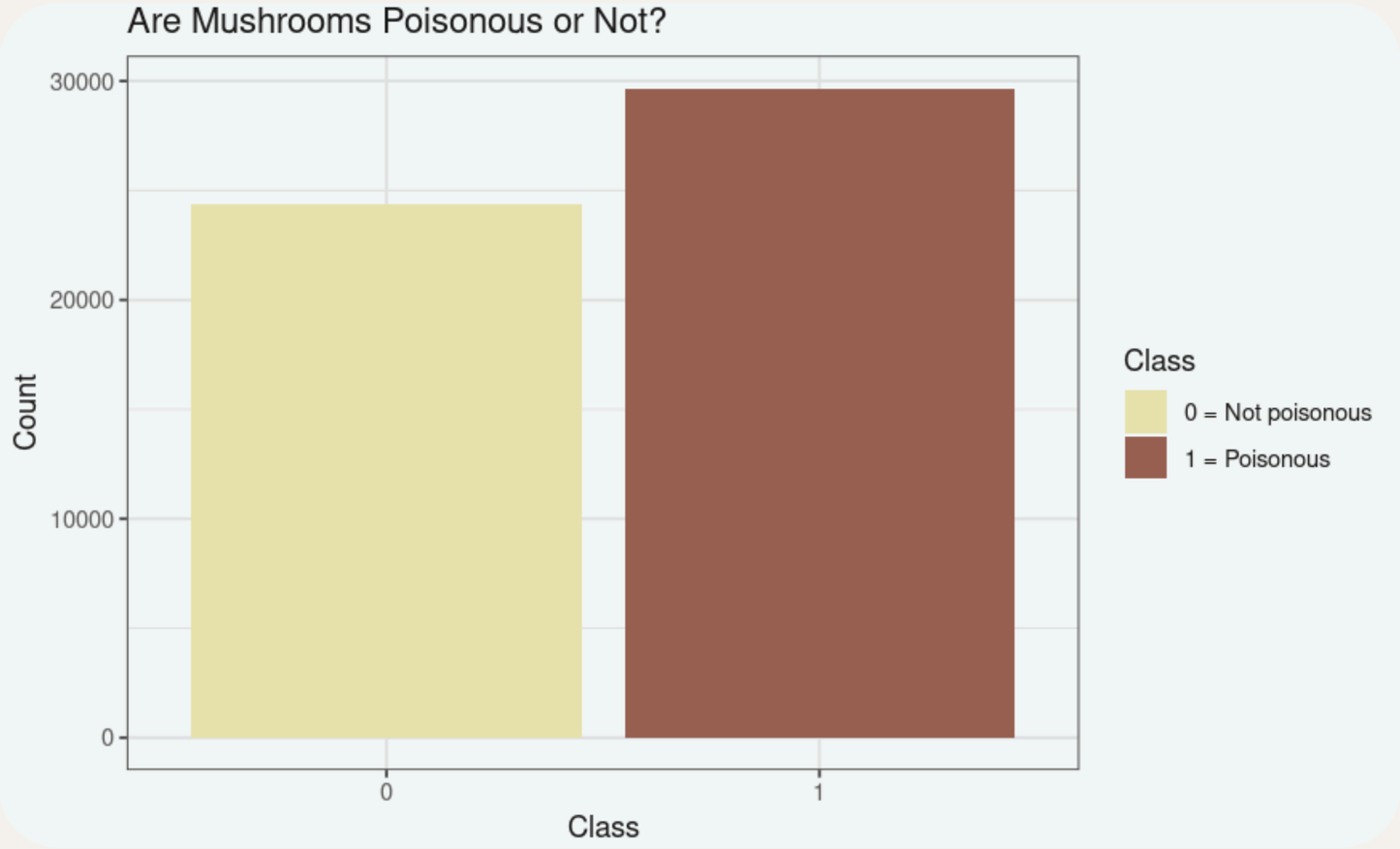


There is a high correlation (0.83) between the cap-diameter and the stem-width. In this case, it is likely that the wider the body, the wider the hood.

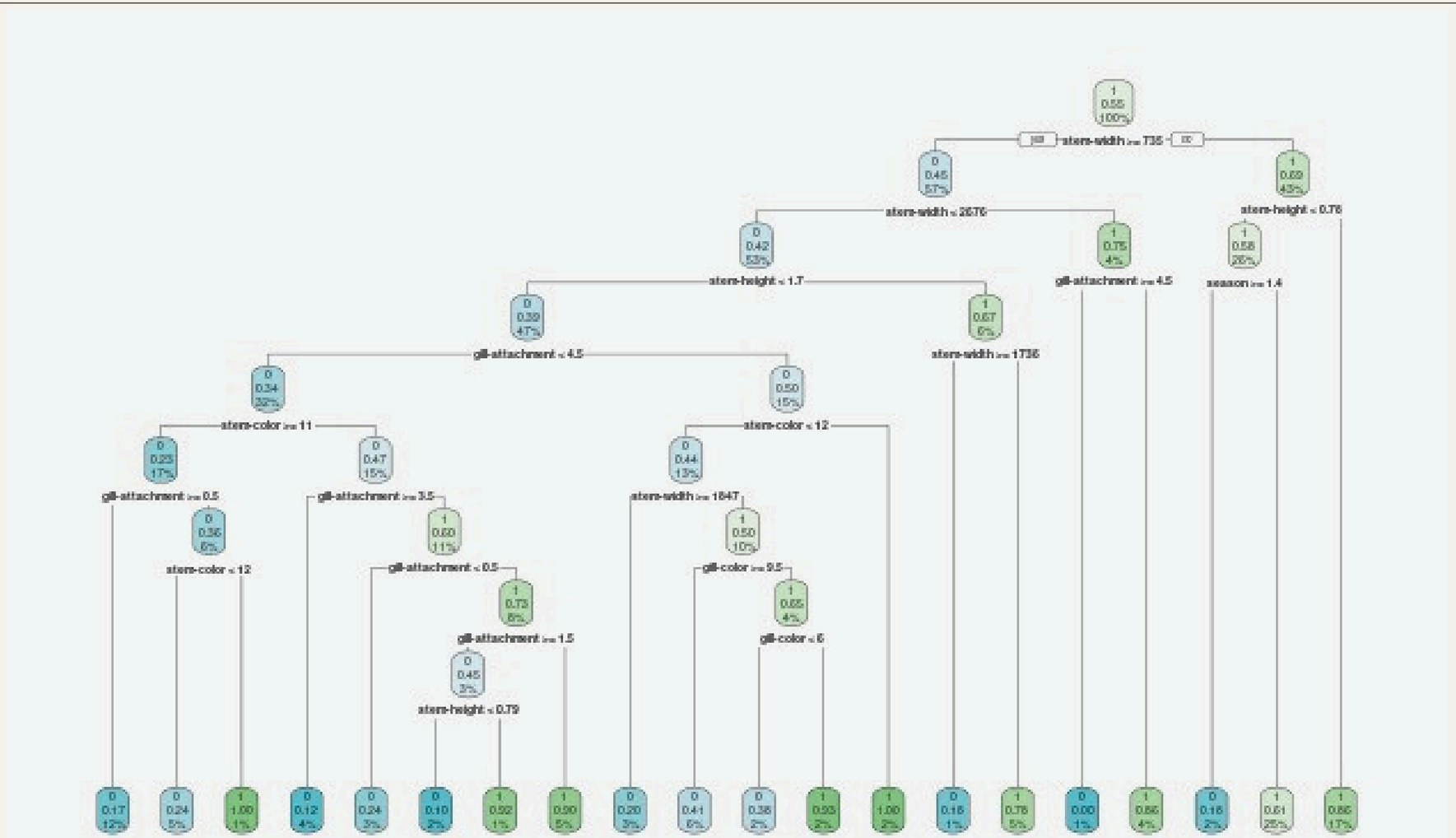
The variables with the lowest correlation (-0.18) are calss and stem-width variables.

POISONOUS OR NOT ?

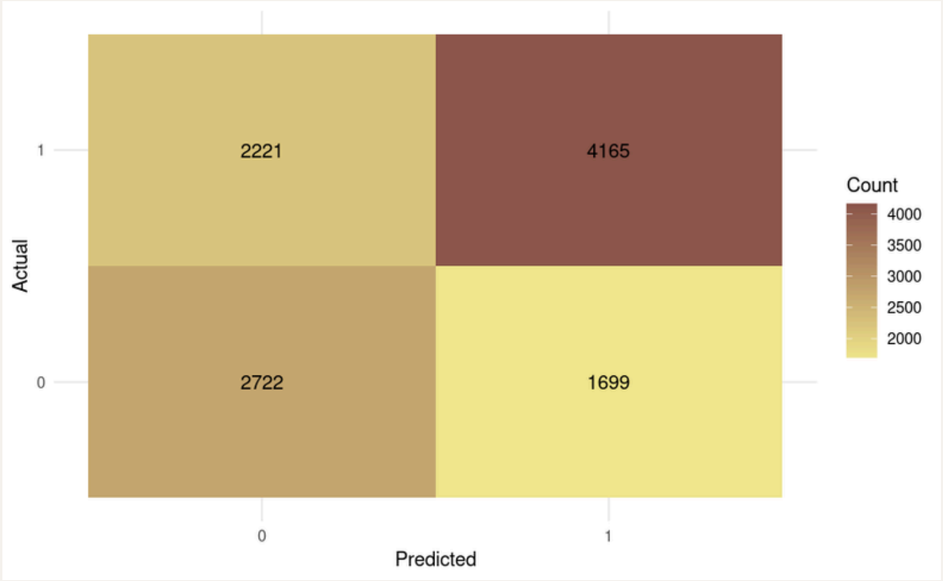
Out of 54,035 observations, there are 24,360 observations with a value of 0 and 29,675 observations with a value of 1.



DECISION TREE VISUALIZATION

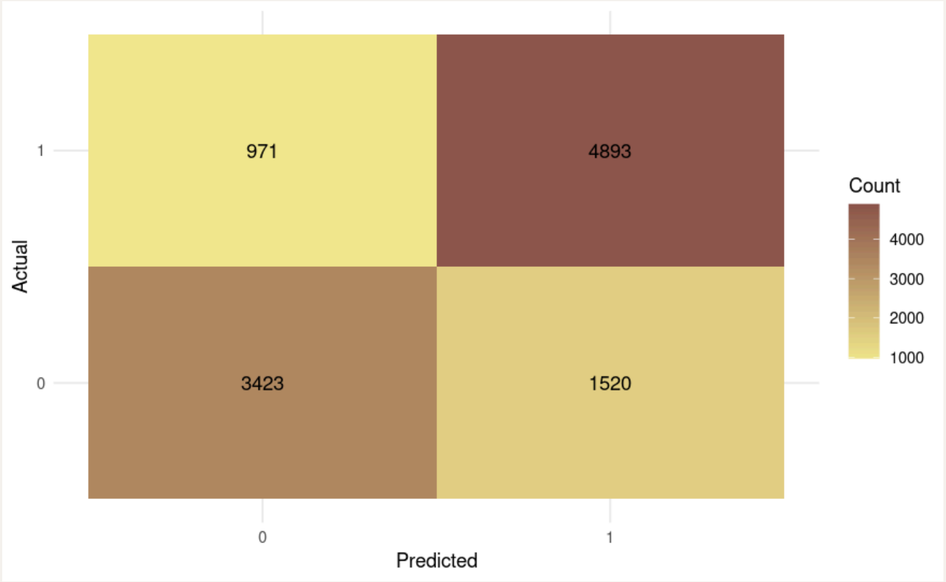


LOGISTIC REGRESSION



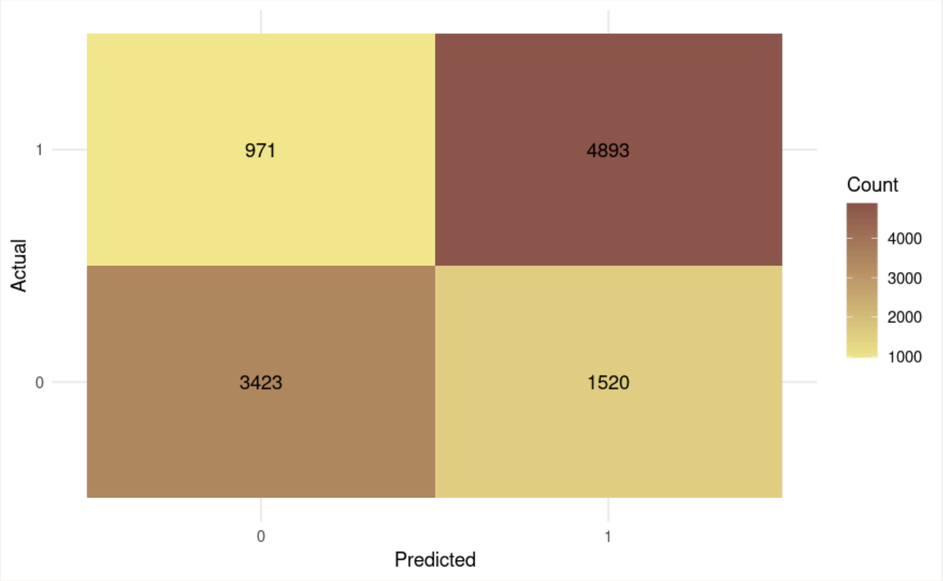
The overall accuracy of the model is 63.73%, meaning that it correctly classified 63.73% of the instances. The model correctly identifies 65.22% of the actual malignant mushrooms.

DECISION TREE

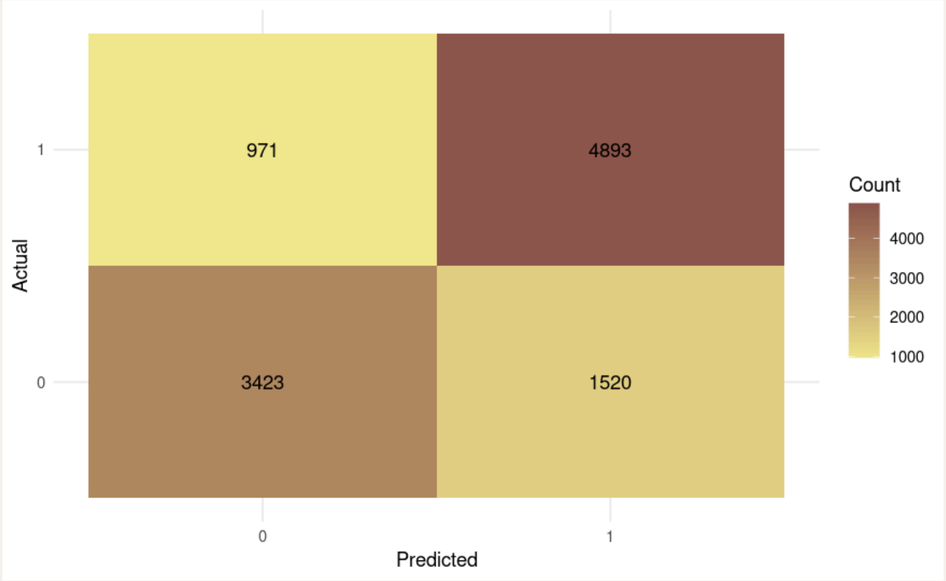


The accuracy of this model is about 77%. Sensitivity; the model correctly predicts the positive class (1: poisonous) about 83% of the time, indicating that the model is quite good at detecting the positive class.

BAGGING TREE



RANDOM FOREST



RANDOM FOREST

cap.diameter	1797.7995
cap.shape	1572.6344
gill.attachment	3405.4751
gill.color	2873.6631
stem.height	2841.8048
stem.width	4946.8626
stem.color	3154.8573
season	665.2981

According to the importance scores showing the contribution of each variable to the classification performance of the model, it is emphasized that some variables such as stem width, gill attachment and stem color are more effective in predicting whether the fungus is harmful or harmless. On the other hand, the influence of other variables such as season is less important.

Accuracy	0.7695
Precision	0.8344
Recall	0.7629
F1 Score	0.7971

The accuracy of this model is about 77%. For this value we can say that the model predicts mostly correctly. Sensitivity; the model correctly predicts the positive class (1: poisonous) about 83% of the time, indicating that the model is quite good at detecting the positive class.

Balanced Accuracy: 0.7634