Will Chan

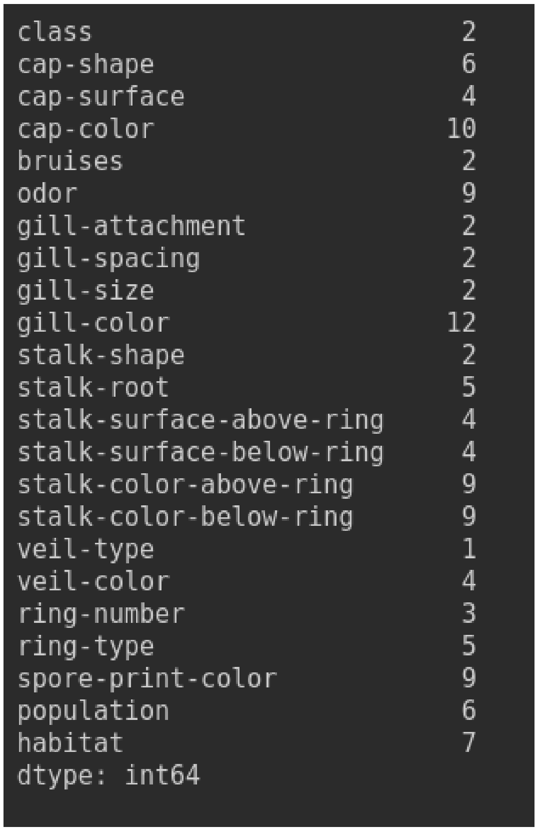
Final Report
Mushroom Edibility Classification

Context: Mushrooms are described as the fleshy, spore-bearing fruiting body of a fungus that is often produced above ground, soil, or on its own food source. There are over 14,000 species discovered and identification of mushrooms are commonly determined via spore print identification and other notable features (bruising, odor, taste, shades of color, habitat, tasting etc..); whereas modern identification is stemming towards microscopic examination. Also along with the value of mushrooms in nutrition, seasoning, and medicinal-properties, mushroom picking for consumption has become increasingly popular especially in the activity of mushroom hunting. Due to lack of microscopic tooling and need for quick assessment in this hobby, mushroom identification via mushroom features are important in determining edibility of mushrooms.

Objective: Usage of mushroom features to determine edibility of mushrooms.

Data Collection: UCI Machine Learning Repository. The source of this dataset was derived from the 1981 Reference book, *National Audubon Society Field Guide to Mushrooms*

Data Wrangling: The dataset originally consisted of 8124 observations and 23 features.
Below Figure displays a table showing the features in the dataset

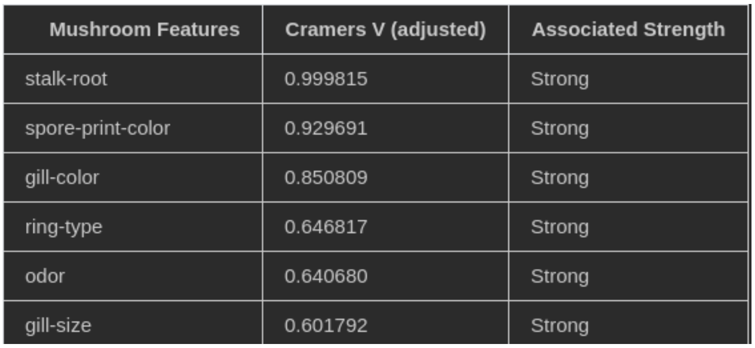


```
class                        2
cap-shape                    6
cap-surface                  4
cap-color                   10
bruises                      2
odor                         9
gill-attachment              2
gill-spacing                 2
gill-size                    2
gill-color                  12
stalk-shape                  2
stalk-root                   5
stalk-surface-above-ring     4
stalk-surface-below-ring     4
stalk-color-above-ring       9
stalk-color-below-ring       9
veil-type                    1
veil-color                   4
ring-number                  3
ring-type                    5
spore-print-color            9
population                   6
habitat                      7
dtype: int64
```

Low variety columns were dropped; veil-type was dropped due to only having 1 unique value; this resulted in having 22 features to work with

In observing missing values of the dataset a search function for missing values returned 0; however, from further observation the ‘Stalk Root’ feature of the dataset had the value ‘?’ used to signify missing values. From this column, 2480 values were found missing, which comprised approximately 30% of data which appeared too significant to drop; thus imputation was needed

However, in order to perform imputation, the likelihood of being able to impute the data needs to be assessed. This would include determining the type of missingness that this dataset is facing. The data was found to be MNCAR via the metric Adjusted Cramer’s V.



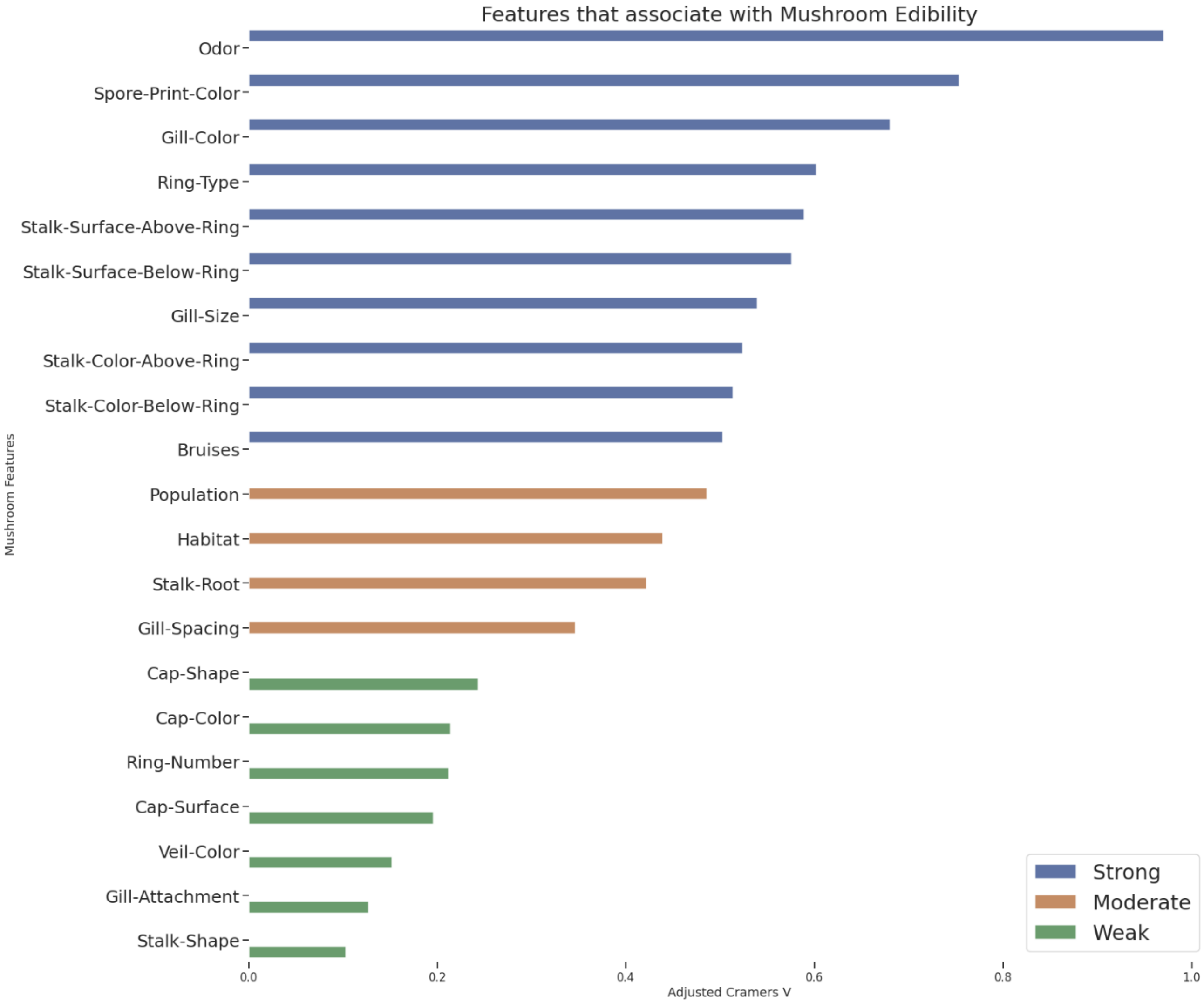
Mushroom Features	Cramers V (adjusted)	Associated Strength
stalk-root	0.999815	Strong
spore-print-color	0.929691	Strong
gill-color	0.850809	Strong
ring-type	0.646817	Strong
odor	0.640680	Strong
gill-size	0.601792	Strong

The above table shows Mushroom Features that have strong association strength with Adjusted Cramer’s V with the feature having missing values, Stalk Root. The Cramer’s V metric shows whether nominal variables are related to each other. Thus the table shows that there are features that are dependent and can be used in imputing the missing values of the Stalk Root column.

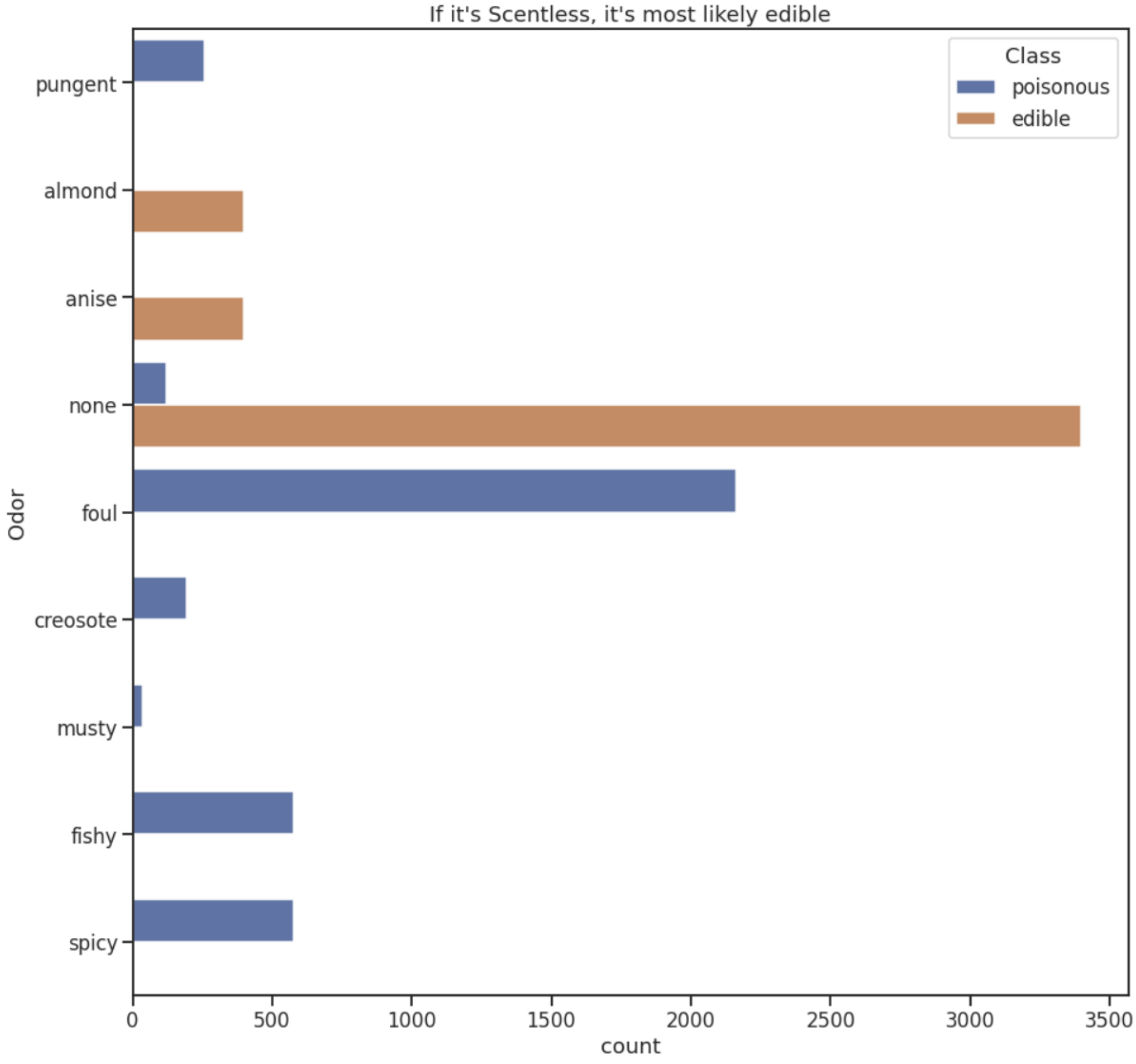
In imputation, the 7 strongly associated features were used (this omitted usage of Mushroom Edibility the target feature for the objective of this report). The dataset was first separated into known and unknown data (datasets not missing observation and datasets missing observations) . The non-missing data was preprocessed and split into train and test sets. A model was generated from the non-missing data; and this model was used to predict the missing values in the missing dataset. This imputation filled in the missing observations for ‘Stalk Root’ feature.

For data types of the features, no changes were necessary as they were appropriate for classification of the objective. Only upper-casing and renaming of the observation data were utilized for exploratory data convenience.

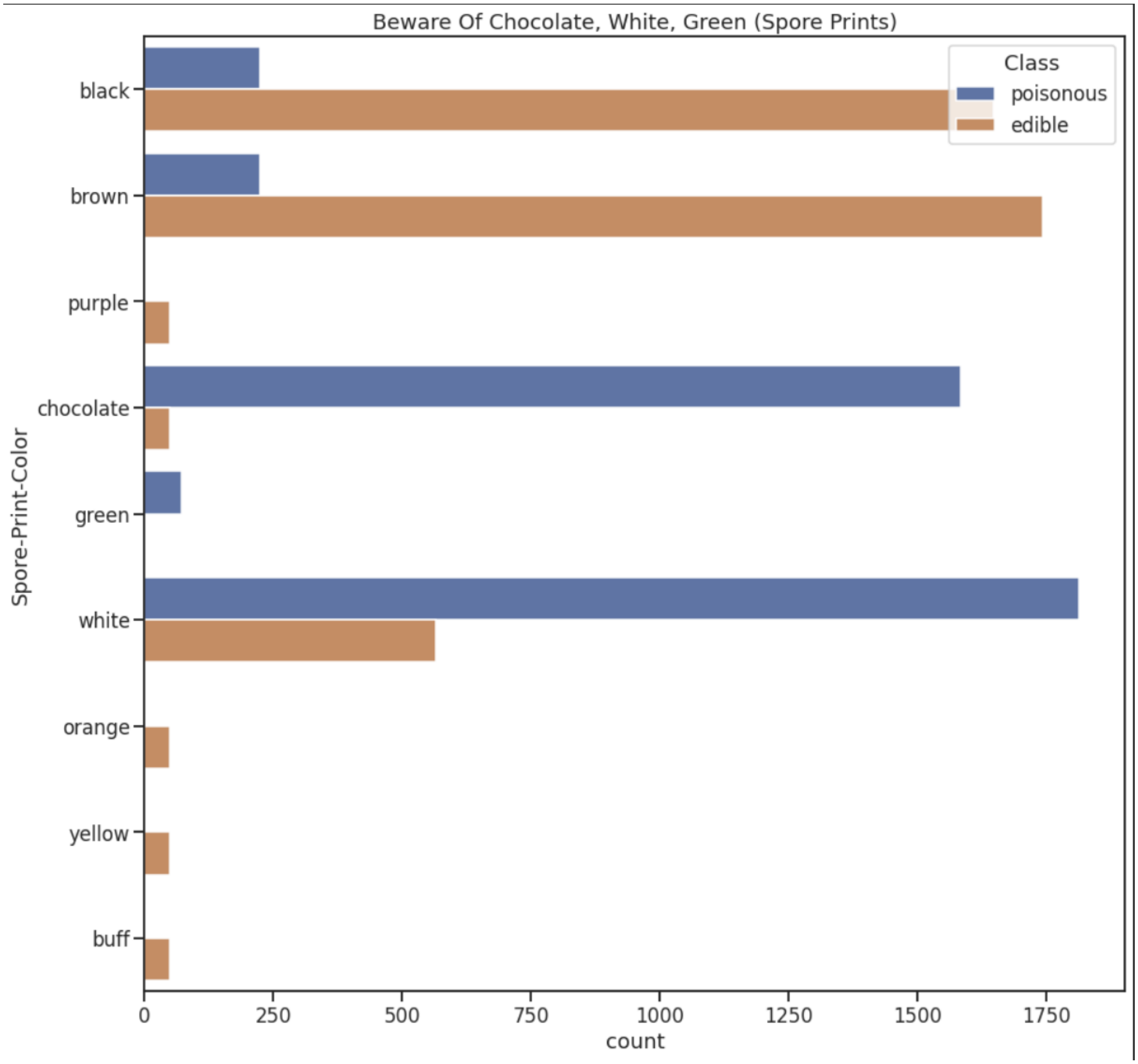
Exploratory Data Analysis: In exploring data, Adjusted Cramer’s V was utilized to determine features most associated with Mushroom edibility.



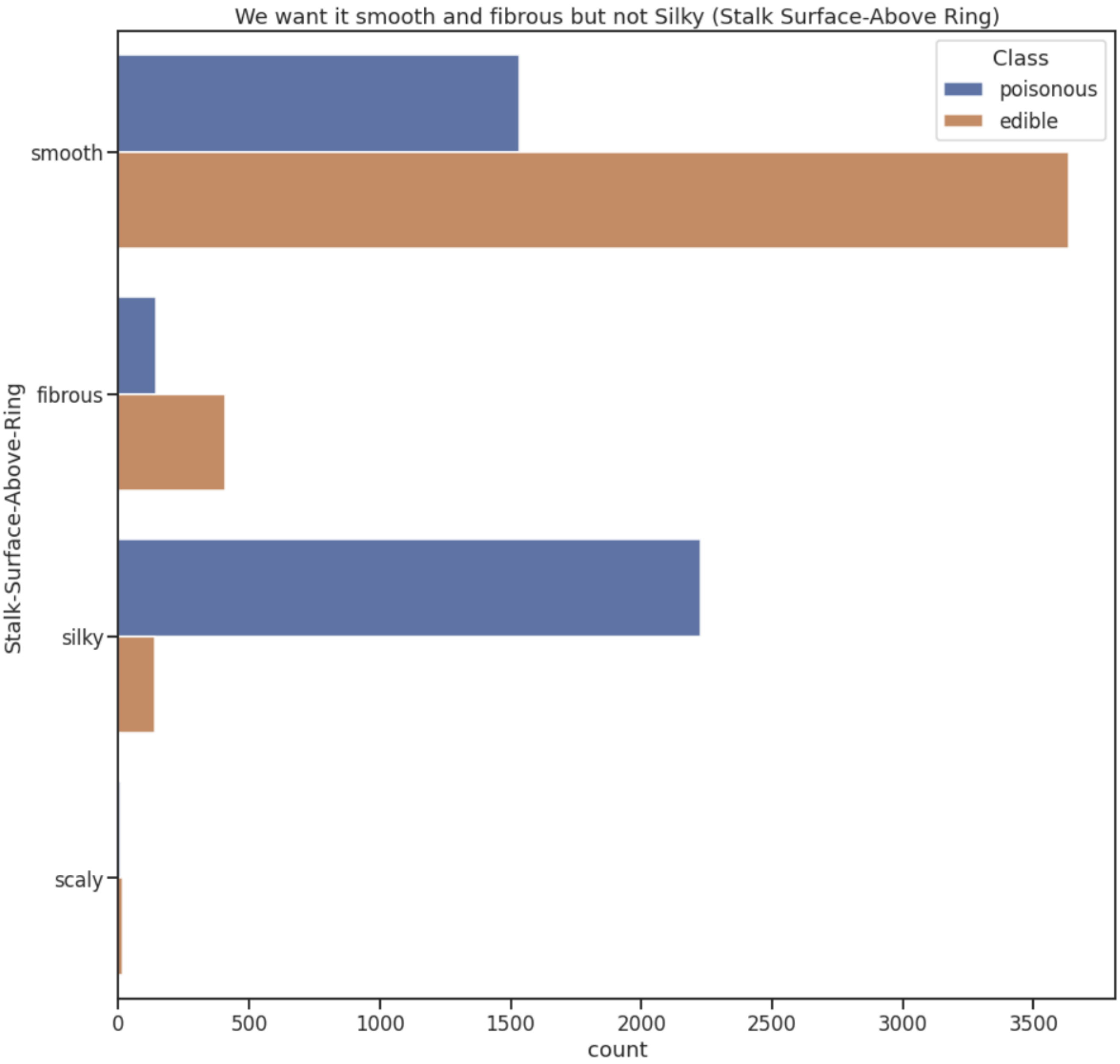
The above bar graph shows the strongly associated features with Mushroom edibility as well as moderate and weakly associated features with edibility.



The above count plot depicts of edibility of mushrooms among the varying odors from the dataset. The data above shows that scentless mushrooms and almond and anise mushrooms are least associated with poisonous mushrooms.



The above count plot depicts of edibility of mushrooms among the varying spore print colors from the dataset. It appears to show that chocolate, white, and green spore prints most often associated with poisonous mushrooms; whereas black and brown appear to be most associated with edible mushrooms. It also shows that there are not many mushrooms with the spore print colors purple, orange, yellow and buff.



The above count plot depicts of edibility of mushrooms among the varying Stalk Surface-Above Ring from the dataset. It shows that smooth and fibrous stalk surfaces above ring are most associated with edibility whereas silky textures are associated with poisonous.

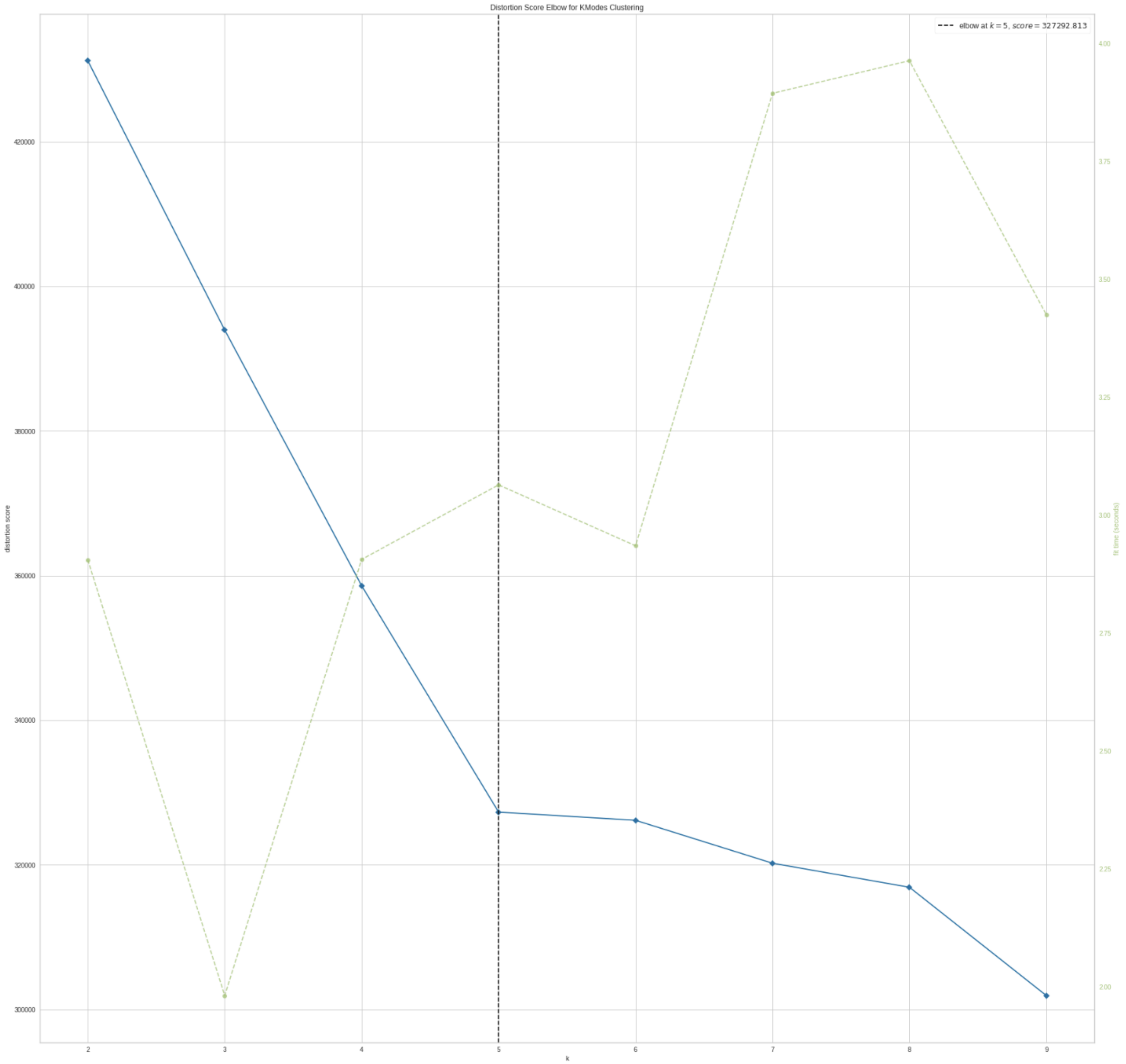
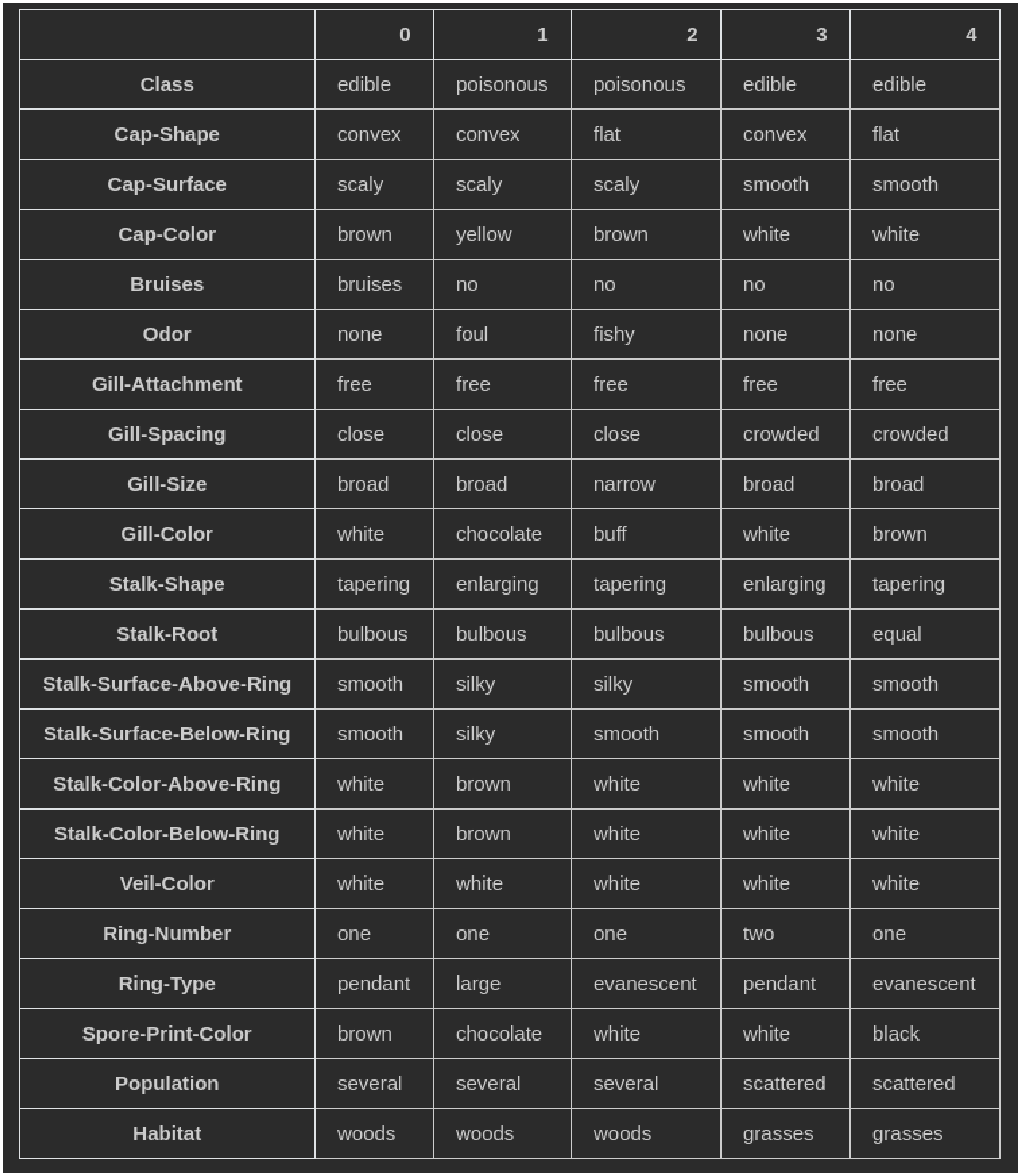


Figure depicts the Distortion Score Elbow for KModes Clustering



	0	1	2	3	4
Class	edible	poisonous	poisonous	edible	edible
Cap-Shape	convex	convex	flat	convex	flat
Cap-Surface	scaly	scaly	scaly	smooth	smooth
Cap-Color	brown	yellow	brown	white	white
Bruises	bruises	no	no	no	no
Odor	none	foul	fishy	none	none
Gill-Attachment	free	free	free	free	free
Gill-Spacing	close	close	close	crowded	crowded
Gill-Size	broad	broad	narrow	broad	broad
Gill-Color	white	chocolate	buff	white	brown
Stalk-Shape	tapering	enlarging	tapering	enlarging	tapering
Stalk-Root	bulbous	bulbous	bulbous	bulbous	equal
Stalk-Surface-Above-Ring	smooth	silky	silky	smooth	smooth
Stalk-Surface-Below-Ring	smooth	silky	smooth	smooth	smooth
Stalk-Color-Above-Ring	white	brown	white	white	white
Stalk-Color-Below-Ring	white	brown	white	white	white
Veil-Color	white	white	white	white	white
Ring-Number	one	one	one	two	one
Ring-Type	pendant	large	evanescent	pendant	evanescent
Spore-Print-Color	brown	chocolate	white	white	black
Population	several	several	several	scattered	scattered
Habitat	woods	woods	woods	grasses	grasses

The above table depicts the results of K-Modes Clustering of the imputed mushroom dataset. Because the categories are not ordinal of nature, a mean average was not used. Rather, the mode statistic was utilized for the average, due to the frequency of features being more interpretable for results.

From the Elbow method, it was determined that 5 clusters should be utilized.

And from the 5 clusters, some prominent features noted are:

- Cluster 0: edible, no odor, white gill-color, and stalk-surface-above-ring is smooth
- Cluster 1: poisonous, foul odor, gray gill-color, and stalk-surface-above-ring is silky
- Cluster 2: poisonous, spicy odor, buff gill color, and stalk-surface-above-ring is silky
- Cluster 3: Edible, no odor, white gill color, and stalk-surface-above-ring is smooth

- Cluster 4: Edible, no odor, brown gill color, and stalk-surface-above-ring is smooth

This pattern in analysis leads to belief that edible mushrooms will often have no smell, and have a stalk-surface-above-ring that is smooth. Other notable patterns from looking at the clusters are that mushrooms in wood habitats have several population, whereas grass habitat mushrooms have scattered populations, and crowded gill-spacing rather than close.

Note: Not all exploratory analysis results are discussed in this report, more analysis can be assessed from the EDA notebooks of the project directory.

Pre-Processing: For preprocessing, the dataset was split into 75-25 train-test split; resulting in 6093 training observations and 2031 testing observations.
The categorical data was also encoded via Ordinal Encoding. Although the categories are not of ordinal nature, ordinal encoding was utilized due to anticipation that Tree classifiers were to be utilized; and tree classifiers have been shown to perform well even when no ordering relationship is present. In addition, Ordinal Encoding was prove to save more memory than using One-Hot Encoding.

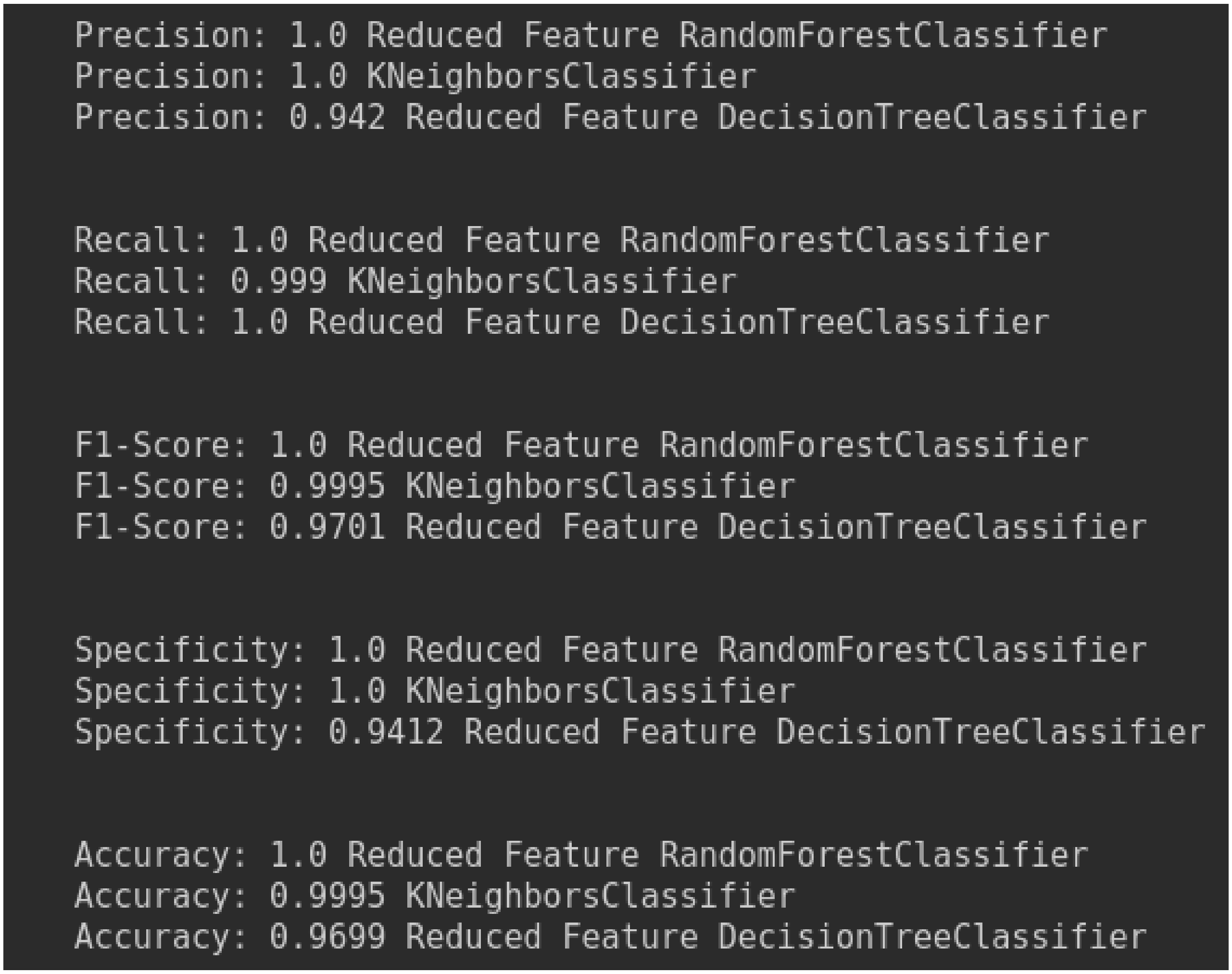
Note: For order of operations, Ordinal Encoder was trained on training set and this encoder was used in encoding both train and test set.

Modeling & Evaluations: Three models were employed in classifying edibility of mushrooms from its outside features and five-fold cross validation was used in model evaluation.

The three models employed were Random Forest Classifier, K-Nearest-Neighbor Classifier, and Decision Tree Classifier.

In resulting models, the Random Forest Classifier utilized 10 features, Decision Tree Classifier utilized 3 Features, whereas KNN utilized 21 Features,

The results of the 3 classifiers are displayed in the table below.



```
Precision: 1.0 Reduced Feature RandomForestClassifier
Precision: 1.0 KNeighborsClassifier
Precision: 0.942 Reduced Feature DecisionTreeClassifier


Recall: 1.0 Reduced Feature RandomForestClassifier
Recall: 0.999 KNeighborsClassifier
Recall: 1.0 Reduced Feature DecisionTreeClassifier


F1-Score: 1.0 Reduced Feature RandomForestClassifier
F1-Score: 0.9995 KNeighborsClassifier
F1-Score: 0.9701 Reduced Feature DecisionTreeClassifier


Specificity: 1.0 Reduced Feature RandomForestClassifier
Specificity: 1.0 KNeighborsClassifier
Specificity: 0.9412 Reduced Feature DecisionTreeClassifier


Accuracy: 1.0 Reduced Feature RandomForestClassifier
Accuracy: 0.9995 KNeighborsClassifier
Accuracy: 0.9699 Reduced Feature DecisionTreeClassifier
```

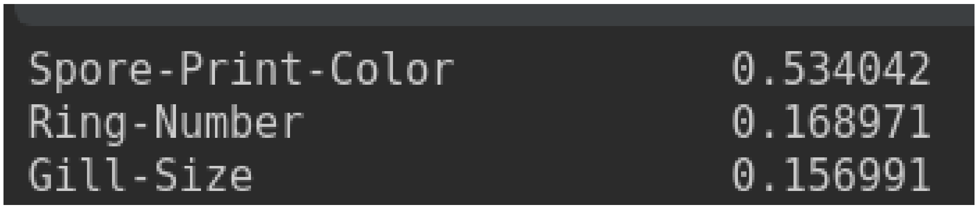
Precision scores refers how likely we are to predict an edible mushroom given that we said the mushroom was edible; this is because precision catches false positives, so looks at how accurate we can catch mushrooms classified as edible when it was noted as edible.

Recall score refers to how many true edible mushrooms were caught from the pool of total mushrooms that were actually edible From our situation, we want to reduce a type I error (false positives) and thus want a higher precision. So Random Forest, K-Nearest-Neighbor, and Decision Tree classifier had good precision of 1, which is optimal in oursituation.

We look more at F1 score than accuracy even though the two metrics are identical in our scenario, we look at F1 because it is a better metric when the false negative and false positive are more crucial than the true positive and true negatives, as we do not want to detect a false positive of saying a mushroom is edible when it is not. The three models mentioned earlier does well in this category.

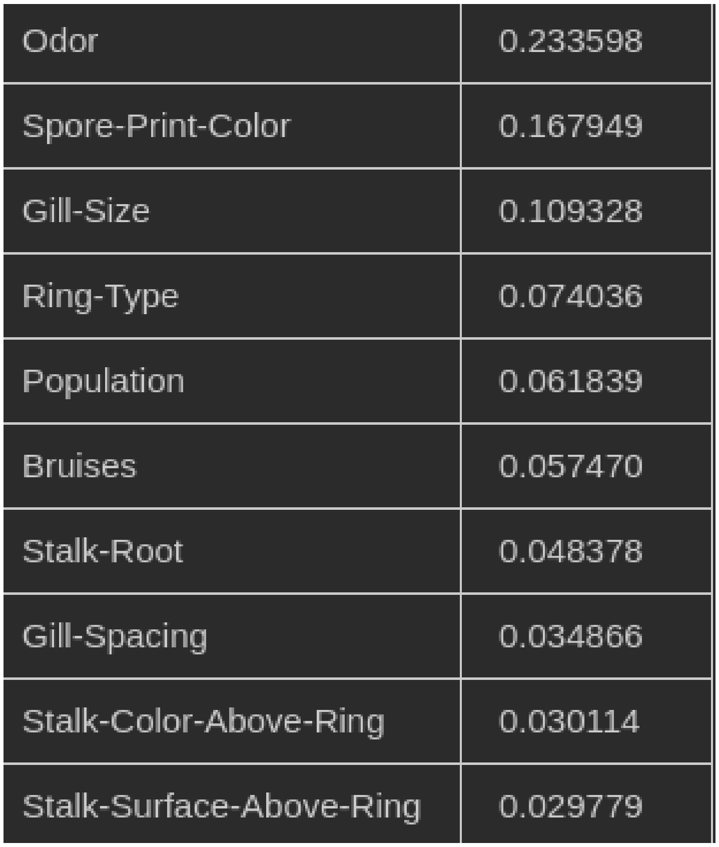
Specificity refers to how many mushrooms were referred to as poisonous were actually poisonous over the total number of actual poisonous mushrooms, and while this is an important metric, precision takes priority over this metric and we shall remain on deciding either choosing Decision Tree, Random Forest, or K-Nearest-Neighbor as our model of choice.

Note: ROC curves does not seem necessary to view as precision and recall are both 1, which signify a perfect AUC of 1.



```
Spore-Print-Color          0.534042
Ring-Number                0.168971
Gill-Size                  0.156991
```

The table above depicts the feature importance scores for Decision Tree Classifier. The features selection of using 3 features were determined from feature importance scores, which demonstrated that approximately 86% of the feature importance can be drawn from the 3 features listed above.



Odor	0.233598
Spore-Print-Color	0.167949
Gill-Size	0.109328
Ring-Type	0.074036
Population	0.061839
Bruises	0.057470
Stalk-Root	0.048378
Gill-Spacing	0.034866
Stalk-Color-Above-Ring	0.030114
Stalk-Surface-Above-Ring	0.029779

The table above depicts the feature importance scores for Random Forest Classifier. The features selection of using 10 features were determined from feature importance scores, which demonstrated that approximately 85% of the feature importance can be drawn from the 3 features listed above.

Note: The K-Nearest Neighbor Classifier does not use feature importance and had used all 21 features in classifying the edibility of mushrooms.

Overall, the Random Forest Classifier was selected as the optimal model for classifying edibility due to having best metrics out of the other models.

Limitations & Future Improvements & Potential Insights:

Challenges were that data was all nominal and thus presented challenges in determining different schemes of statistic analysis such as using Cramer’s V and K-Modes for analysis over common numerical statistical analysis techniques.

Future improvements would be to scrutinize or further assess the accuracy the mushroom reference book where the dataset was sourced from, which appears to be outdated (1981). Also producing own mushroom dataset with more applicable/descriptive features such as more frequent odors and habitats would provide more insight. From looking at reviews of the book, it also shows that some of the data is not as accurate and thus a more up-to date dataset would be more useful to study and grab information from.

Potential Insight that can be drawn mushroom classification would be useful for novice mushroom hunters who are hunting for edible mushrooms and lack microscopic equipment and knowledge. For future prospects, species classification via microscopic observation of spores would be highly valuable.