**Introduction**

**an introduction/overview/executive summary section that describes the dataset and summarizes the goal of the project and key steps that were performed**

In this report, our goal is to predict the edibility (class: edible / poisonous) of mushroom basing on attribution information. Data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). The reason of selecting this dataset is that this problem is related to classification which is a large part of application in data science. And, it is also a complement to project – MovieLens that we can cover each part of what we have learnt from the course.

The mushroom dataset has already been well formatted from the source already. Process of data cleaning is only removing 2 attributes prior to splitting the data to training set and test set. 10 algorithms are applied and an ensemble model combining the prior 10 different algorithms to see if it can provide improvement to our predictions.

1. glm
2. lda
3. Naïve Bayes
4. svmLinear
5. classification
6. knn
7. gamLoess
8. multinom
9. rf
10. adaboost
11. ensemble

[Talk about the results here]

**Method**

**a methods/analysis section that explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained, and your modeling approach**

**1. Data Cleaning**

Mushroom data set contains 23 columns of 1 class and 22 attributes related to cap, bruises, odor, gill, stalk, veil, ring, spore color, population and habitat of 8,124 observations. According to description from the source, there is data missing in the attribute of stalk root. The missing data point is marked “?” from the source already. Veil type is reported 1 level only. As a result, both stalk root and veil type are removed before we start data exploration & modeling.

[show str of mushroom]

[show summary of mushroom]

**2. Data Exploration**

We will use Data Explorer package in the process of data exploration. By using this package, it provides a standardized method to get the insights from the dataset.

From below 2 charts, we can see that the mushroom data is discrete with no data missing.

[2 charts]

From below 2 charts, we can see how many mushrooms belong to each category in each attribute and among those how many are edible and how many are poisonous. In the attribute of class which we are going to predict, we can say this feature is roughly equal distributed. However, observations are clustered in one gill\_attachment, gill\_spacing, veil\_color, ring\_number

[2 charts]

**3. Modeling Approach**

**Result**

**a results section that presents the modeling results and discusses the model performance**

**Conclusion**

**a conclusion section that gives a brief summary of the report, its limitations and future work**