PROJECT PLAN

1. Students' name and Purdue e-mail.

a. Nikita Rajaneesh — nrajanee@purdue.edu

b. Swaraj Bhaduri — sbhadur@purdue.edu

c. Utkarsh Jain — jain192@purdue.edu

d. Ishaan Saxena — isaxena@purdue.edu

2. Definition of the problem, possibly relevant to your interests.

Analysis of Mushroom Species Data by classification into groups of poisonous and edible based on features such as caps, odor, stalks, etc. and comparing different classification models and algorithms to determine which is best suited to this problem.

3. Description of the dataset (or datasets) to be used.

Size of dataset (before encoding): $(n = 8124, d = 22)$

Size of dataset (after encoding): $(n = 8124, d = 107)$

Classes: edible=e, poisonous=p (y-values)

Attribute Information and Encoding:

1. Nominal Categorical Variables:
These variables will be encoded as binary one-hot features. As a result, each feature in this category would be replace by the k features in the encoded dataset if the feature has k possible values. These featues include:

   i. **cap-shape**: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
   ii. **cap-surface**: fibrous=f, grooves=g, scaly=y, smooth=s
   iii. **cap-color**: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
   iv. **bruises**: bruises=t, no=f
   v. **odor**: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
   vi. **gill-attachment**: attached=a, descending=d, free=f, notched=n
   vii. **gill-color**: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
   viii. **stalk-root**: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
   ix. **stalk-surface-above-ring**: fibrous=f, scaly=y, silky=k, smooth=s
   x. **stalk-surface-below-ring**: fibrous=f, scaly=y, silky=k, smooth=s
   xi. **stalk-color-above-ring**: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
   xii. **stalk-color-below-ring**: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
   xiii. **veil-type**: partial=p, universal=u
   xiv. **veil-color**: brown=n, orange=o, white=w, yellow=y
   xv. **ring-type**: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
   xvi. **spore-print-color**: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
   xvii. **habitat**: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

2. Ordinal Categorical Variables:
   These variables will be encoded in place by encoding labels, as the data here has ordinal meaning to it. These variables include:

   i. **gill-spacing**: close=c→0, crowded=w→1, distant=d→2
   ii. **gill-size**: broad=b→0, narrow=n→1
   iii. **stalk-shape**: enlarging=e→0, tapering=t→1
   iv. **ring-number**: none=n→0, one=o→1, two=t→2
   v. **population**: abundant=a→0, clustered=c→1, numerous=n→2, scattered=s→3, several=v→4, solitary=y→5

4. URL where the above dataset(s) is(are) available.

https://www.kaggle.com/uciml/mushroom-classification

5. Which machine learning algorithm(s) is(are) going to be used?

We are looking to compare the performance of different algorithms, including Support Vector Machines, Adaboost, and Logistic Regression using different metrics. We will also use feature selection to ensure additional complexity control and gain interpretability.

6. Cross-validation technique (e.g., training/validation/testing, k-fold cross-validation, bootstrapping)

k-Fold Cross Validation would be used with about 5 to 10 folds.

7. Which hyperparameter(s) is(are) going to be tuned.

Hyperparameters of the chosen algorithm would be tuned after it is determined. For instance, if a Support Vector Machine is chosen, the value of the penalty parameter C for the error terms (slack variables) will be adjusted, or the regularization penalty for Logistic Regression.

8. Description of the experimental results, e.g., plots of number of samples versus accuracy (you can use different subsets of the same dataset), regularization parameter versus accuracy, ROC curves, plots of different datasets, etc.

Graphs for Hyperparameter tuning would include hyperparameter value vs accuracy, plots of descion boundary and margins obtained by different hyperparameter values.

We will add plots for correlations between features, and correlations between features and class labels, etc. for feature selection.

To compare models (and different kernels), we will use the sensitivity vs. specificity curves, area under ROC curve, number of folds used vs. training accuracy for each model, number of features (size of feature subset) vs accuracy.

9. Which programming language are you going to use?

Python-2.7