Classification of Two Genera of Mushrooms from the Agaricaceae Family

Aaron A. Gauthier and Pedro Uria Rodriguez
Machine Learning I
GWU

December 4, 2018

Table of Contents

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom
- Very difficult and time consuming for non-experts

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom
- Very difficult and time consuming for non-experts

Thus we want to...

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom
- Very difficult and time consuming for non-experts

Thus we want to...

- Classify mushrooms using Machine Learning

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom
- Very difficult and time consuming for non-experts

Thus we want to...

- Classify mushrooms using Machine Learning
- Identify top relevant features so that

Mushroom Classification

- Guides state there is no simple rule for determining the edibility of a mushroom
- Very difficult and time consuming for non-experts

Thus we want to...

- Classify mushrooms using Machine Learning
- Identify top relevant features so that
- Novices can quickly identify a poisonous mushroom, stop wasting time and move onto another potentially edible mushroom

Motivation

- People in the USA do not know the difference between an edible mushroom and a poisonous mushroom

Motivation

- People in the USA do not know the difference between an edible mushroom and a poisonous mushroom
- Educate people on the top significant features that determine whether a mushroom is edible or not

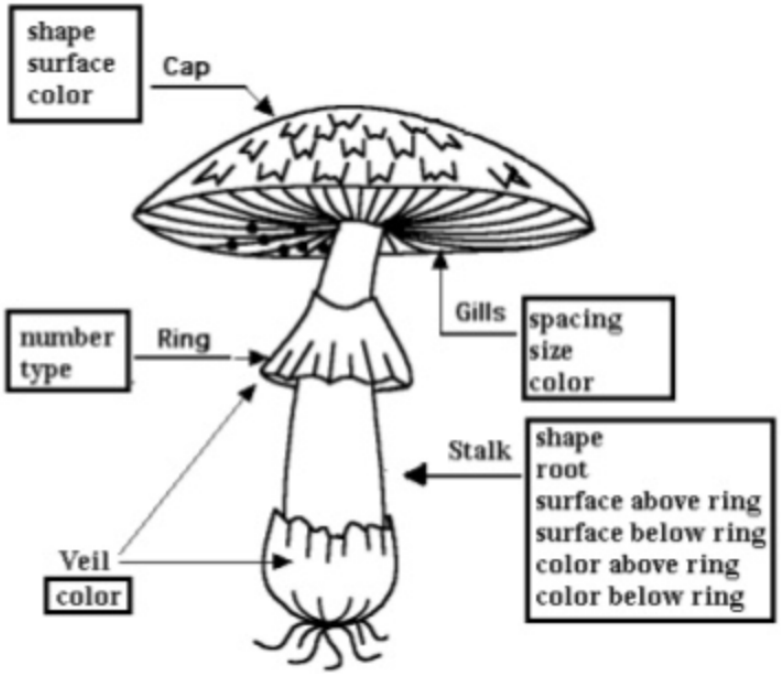
Motivation

- People in the USA do not know the difference between an edible mushroom and a poisonous mushroom
- Educate people on the top significant features that determine whether a mushroom is edible or not
- We hope this will help avoid tragedy, especially with children who are curious

Machine Learning Models

- Logistic Regression
- Decision Tree
- Random Forest
- K-Nearest Neighbors
- Gaussian Naive Bayes
- Support Vector Machine
- K-Means Clustering

Mushroom Features



Go to the jupyter notebook for an explanation of each of them.

Easy VS Hard Features

Easy:

- Cap Shape
- Cap Color
- Bruises
- Gill Color
- Stalk Shape
- Stalk Root
- Stalk Color
- Veil Color
- Ring Number
- Ring Type
- Spore Print Color
- Population
- Habitat

Hard:

- Cap Surface
- Gill attachement
- Stalk Surface
- Odor
- Gill spacement

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows $\Rightarrow$ Only Dummy Features (after one-hot-encoding) $\Rightarrow$ no Standardization required

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows $\Rightarrow$ Only Dummy Features (after one-hot-encoding) $\Rightarrow$ no Standardization required
- Target: Edible (e) or Poisonous (p) $\Rightarrow$ Label Encode into "0"s and "1"s

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows $\Rightarrow$ Only Dummy Features (after one-hot-encoding) $\Rightarrow$ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) $\Rightarrow$ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values $\rightarrow$ Various approaches:

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows ⇒ Only Dummy Features (after one-hot-encoding) ⇒ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) ⇒ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values → Various approaches:
  1. Use the dataset without the column

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows ⇒ Only Dummy Features (after one-hot-encoding) ⇒ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) ⇒ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values → Various approaches:
  1. Use the dataset without the column
  2. Use the raw dataset with "?" → `np.NaN`

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows $\Rightarrow$ Only Dummy Features (after one-hot-encoding) $\Rightarrow$ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) $\Rightarrow$ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values $\rightarrow$ Various approaches:
  1. Use the dataset without the column
  2. Use the raw dataset with "?" $\rightarrow$ `np.NaN`
  3. Use the raw dataset (`stalk-root_?` dummy feature)

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows ⇒ Only Dummy Features (after one-hot-encoding) ⇒ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) ⇒ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values → Various approaches:
  1. Use the dataset without the column
  2. Use the raw dataset with "?" → `np.NaN`
  3. Use the raw dataset (`stalk-root_?` dummy feature)
  4. Dropping rows with "?" (losing 30 % of our data)

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows ⇒ Only Dummy Features (after one-hot-encoding) ⇒ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) ⇒ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values → Various approaches:
  1. Use the dataset without the column
  2. Use the raw dataset with "?" → `np.NaN`
  3. Use the raw dataset (`stalk-root_?` dummy feature)
  4. Dropping rows with "?" (losing 30 % of our data)
  5. Mode imputation

Exploratory Data Analysis and Preprocessing

- 23 Categorical Features and 8123 rows ⇒ Only Dummy Features (after one-hot-encoding) ⇒ no Standardization required
- Target: Edible (`e`) or Poisonous (`p`) ⇒ Label Encode into "0"s and "1"s
- Feature `stalk-root` has 2480 missing values → Various approaches:
  1. Use the dataset without the column
  2. Use the raw dataset with "?" → `np.NaN`
  3. Use the raw dataset (`stalk-root_?` dummy feature)
  4. Dropping rows with "?" (losing 30 % of our data)
  5. Mode imputation
- 70-30 Train-Test Split

Table of Contents

Models Performance

- Perfect cross-validated scores after hyperparameter tuning for Logistic Regression, Decision Tree and Random Forest for every treatment of the missing values

Models Performance

- Perfect cross-validated scores after hyperparameter tuning for Logistic Regression, Decision Tree and Random Forest for every treatment of the missing values
- Thus found that `stalk-root` was not needed for perfect predictions

Models Performance

- Perfect cross-validated scores after hyperparameter tuning for Logistic Regression, Decision Tree and Random Forest for every treatment of the missing values
- Thus found that `stalk-root` was not needed for perfect predictions
- Also perfect scores for KNN and SVM without using `stalk-root`

Models Performance

- Perfect cross-validated scores after hyperparameter tuning for Logistic Regression, Decision Tree and Random Forest for every treatment of the missing values
- Thus found that `stalk-root` was not needed for perfect predictions
- Also perfect scores for KNN and SVM without using `stalk-root`
- Naive Bayes gave a 0.999 accuracy

Models Performance

- Perfect cross-validated scores after hyperparameter tuning for Logistic Regression, Decision Tree and Random Forest for every treatment of the missing values
- Thus found that `stalk-root` was not needed for perfect predictions
- Also perfect scores for KNN and SVM without using `stalk-root`
- Naive Bayes gave a 0.999 accuracy
- "Supervised" Clustering gave 0.904 accuracy

Features Importances

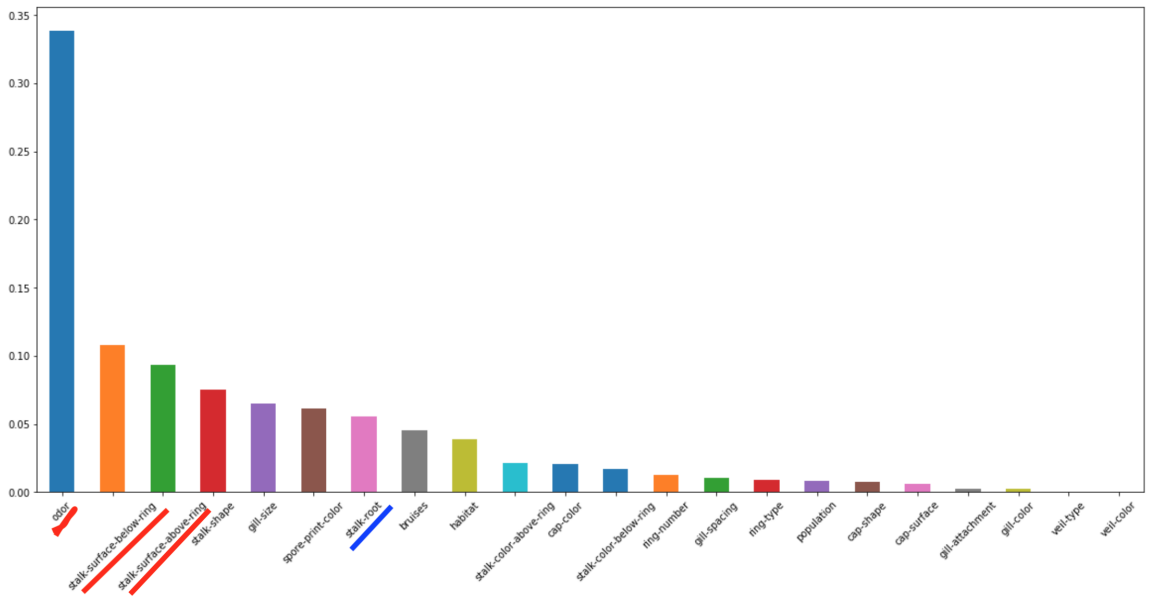
- Logistic Regression ranked the features by importance through Recursive Feature Elimination

Features Importances

- Logistic Regression ranked the features by importance through Recursive Feature Elimination
- Random Forest provided numerical feature importances by averaging the gini impurity at each node across all trees

Features Importances

- Logistic Regression ranked the features by importance through Recursive Feature Elimination
- Random Forest provided numerical feature importances by averaging the gini impurity at each node across all trees
- Similar results (Random Forest shown below)



Models Performance

- Without `stalk-root`, Logistic Regression tuned to give only 1 false positive (poisonous mushroom classified as edible) out of 1175 poisonous mushrooms, while Decision Tree tuned to give 8

Models Performance

- Without `stalk-root`, Logistic Regression tuned to give only 1 false positive (poisonous mushroom classified as edible) out of 1175 poisonous mushrooms, while Decision Tree tuned to give 8
- The rest of the missing values approaches gave perfect scores for Logistic Regression and Decision Tree

Models Performance

- Without `stalk-root`, Logistic Regression tuned to give only 1 false positive (poisonous mushroom classified as edible) out of 1175 poisonous mushrooms, while Decision Tree tuned to give 8
- The rest of the missing values approaches gave perfect scores for Logistic Regression and Decision Tree
- Decided to drop the missing values to train the rest: Random Forest, KNN and SVM gave perfect scores

Models Performance

- Without `stalk-root`, Logistic Regression tuned to give only 1 false positive (poisonous mushroom classified as edible) out of 1175 poisonous mushrooms, while Decision Tree tuned to give 8
- The rest of the missing values approaches gave perfect scores for Logistic Regression and Decision Tree
- Decided to drop the missing values to train the rest: Random Forest, KNN and SVM gave perfect scores
- Naive Bayes classified 16.6% of poisonous mushrooms as edible

Models Performance

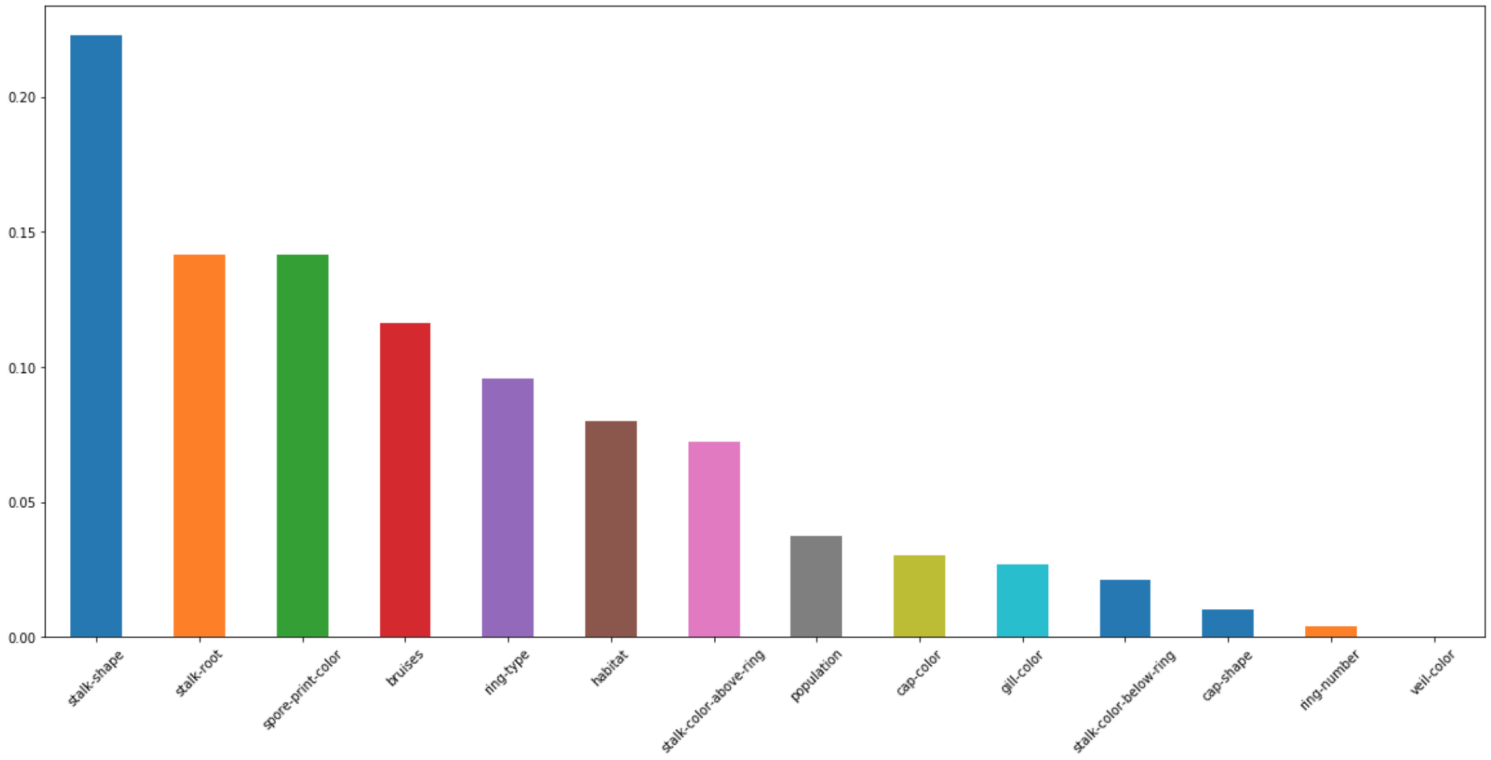
- Without `stalk-root`, Logistic Regression tuned to give only 1 false positive (poisonous mushroom classified as edible) out of 1175 poisonous mushrooms, while Decision Tree tuned to give 8
- The rest of the missing values approaches gave perfect scores for Logistic Regression and Decision Tree
- Decided to drop the missing values to train the rest: Random Forest, KNN and SVM gave perfect scores
- Naive Bayes classified 16.6% of poisonous mushrooms as edible
- Clustering accuracy decreased to 0.855

Features Importances

- Previous features move up in the ranks

Features Importances

- Previous features move up in the ranks
- `stalk-root` is the top 2



Decision Tree

- Offers simple rules to follow

Decision Tree

- Offers simple rules to follow

  Meaning that...

Decision Tree

- Offers simple rules to follow

  Meaning that...

- There is no need to bring your phone with you... a piece of paper with the tree will do

Decision Tree

- Offers simple rules to follow

  Meaning that...

- There is no need to bring your phone with you... a piece of paper with the tree will do
- With `stalk-root`, simplest tree: 17 nodes

Decision Tree

- Offers simple rules to follow

  Meaning that...

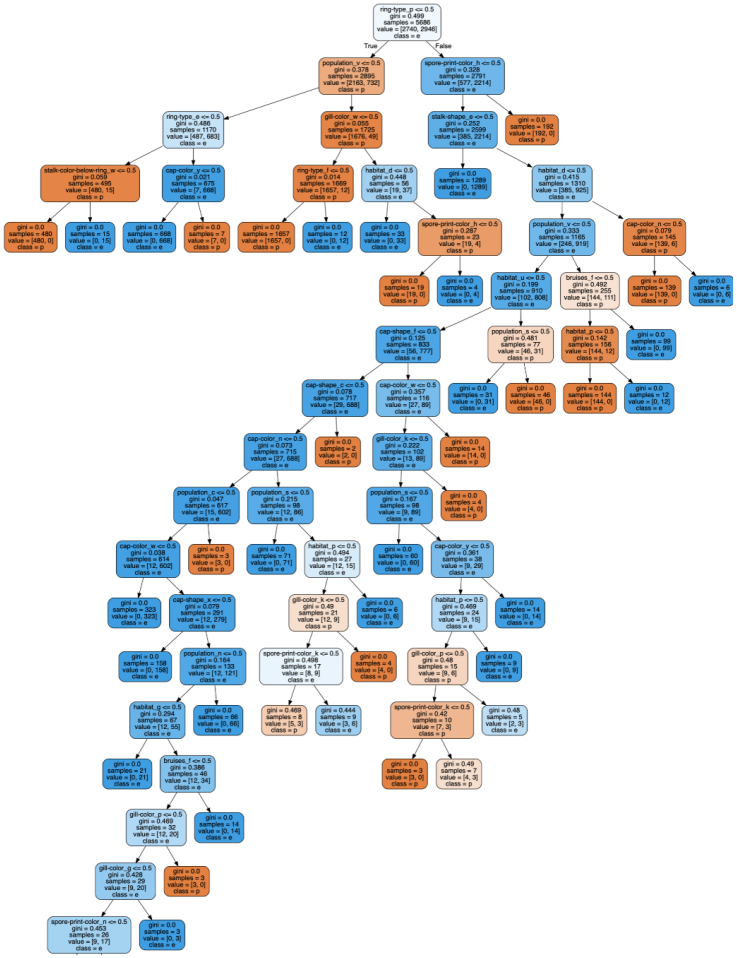
- There is no need to bring your phone with you... a piece of paper with the tree will do
- With `stalk-root`, simplest tree: 17 nodes
- Without `stalk-root` $\longrightarrow$

Decision Tree

- Offers simple rules to follow

  Meaning that...

- There is no need to bring your phone with you... a piece of paper with the tree will do
- With `stalk-root`, simplest tree: 17 nodes
- Without `stalk-root` $\longrightarrow$



Some Features to Look at First

- Computed % of poisonous mushrooms for top *easy-to-tell* features

Some Features to Look at First

- Computed % of poisonous mushrooms for top *easy-to-tell* features

  Some features that most likely make a mushroom poisonous:

Some Features to Look at First

- Computed % of poisonous mushrooms for top *easy-to-tell* features

  Some features that most likely make a mushroom poisonous:

- `spore-print-color` = chocolate, white or green
- Mushroom does not bruise
- Mushroom with large or no rings
- Mushrooms living on leaves, cities or paths
- `stalk-color-above-ring` = brown, buff, cinnamon or yellow

Table of Contents

Advice/Warning and Legal Disclaimer

- Although most models gave perfect cross-validated scores...

Advice/Warning and Legal Disclaimer

- Although most models gave perfect cross-validated scores...
- These are only accurate for the two genera (*Agaricus* and *Lepiota*) the data came from
- The data we used could also be compromised

Advice/Warning and Legal Disclaimer

- Although most models gave perfect cross-validated scores...
- These are only accurate for the two genera (*Agaricus* and *Lepiota*) the data came from
- The data we used could also be compromised

Thus the authors...

Advice/Warning and Legal Disclaimer

- Although most models gave perfect cross-validated scores...
- These are only accurate for the two genera (*Agaricus* and *Lepiota*) the data came from
- The data we used could also be compromised

Thus the authors...

- Take no responsibility on the usage of these models by other people
- Always advise to consult with an expert before ingesting wild mushrooms

Features Importances

- Identified most relevant *easy-to-tell* features

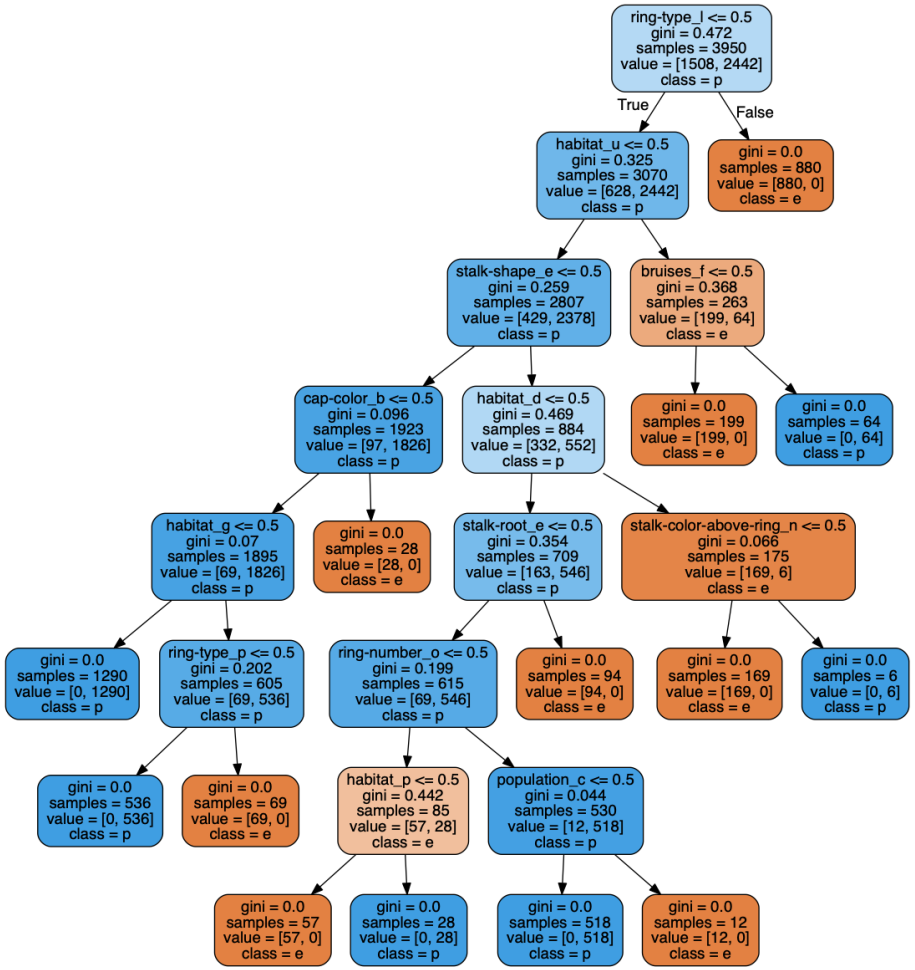
Features Importances

- Identified most relevant *easy-to-tell* features
- Found some interesting features for mushroom hunters to look at first...
- So that they can quickly identify a likely poisonous mushroom...
- Thus saving them time to focus on other potential edible mushrooms

Features Importances

- Identified most relevant *easy-to-tell* features
- Found some interesting features for mushroom hunters to look at first...
- So that they can quickly identify a likely poisonous mushroom...
- Thus saving them time to focus on other potential edible mushrooms
- Gave guidelines as to how to find more of these

Simplest Easiest Decision Tree



Happy Mushroom Hunting!

References

- Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt, Birmingham, 2017
- Machine Learning I's notes, slides, exercises and homework
- Wikipedia's entries for the various classifiers
- `https://arxiv.org/pdf/1410.5329v3.pdf`
- A bunch of other articles and coding questions, all referenced in the *jupyter notebook*

Github Repository: `https://github.com/QuirkyDataScientist1978/GWU-Machine-Learning-1-Fall-2018-Mushroom-Classification-Project`