DATA 240 Data Mining and Analytics

“Mushroom Classification”

Group: 4

Greeshma Venkatesh

Neelima Jagtap

Rishikumar Ravichandran

Rushikesh Jagtap

Sougandhika Ayathamaraju

Introduction

- Mushrooms contain protein, vitamins, minerals, and antioxidants.
- Mushrooms are formed from macrofungi that can be both hypogeous and epigeous.
- There are currently an estimated 1.5 million species of mushrooms in the world.
- Mushrooms are used as medicine in Eastern medicine for centuries as antioxidant, antidepressants, antidiabetic, etc

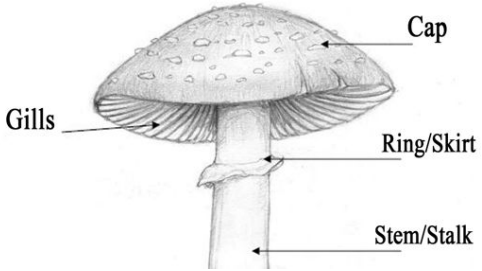
Motivation

- Mushrooms have more medicinal values that helps in avoiding chronic heart disease and Cancer.
- Not all mushrooms are edible. Some may causes allergic reaction to death.
- It is recorded that poisonous mushrooms are the reason for 9 out of 10 deaths caused by fungi each year.
- Biologists have classified the list of poisonous mushrooms based on their morphology and the scientific family.
- With the advancement in the field of data science it is easy to classify them into poisonous and non-poisonous.
- That paved the way for us to classify mushrooms into poisonous and non-poisonous based on the behavioral features.

Data Summary

1. The data is collected from the UCI machine learning repository.
2. In this dataset, different species of mushrooms are taken in count from the Agaricus and Lepiota family.
3. Total of 22 attributes like cap-shape, cap-color, gill-size, gill-color, stalk-root, ring-type, odor of the mushroom and 8124 instances.
4. Class Distribution after pre-processing:

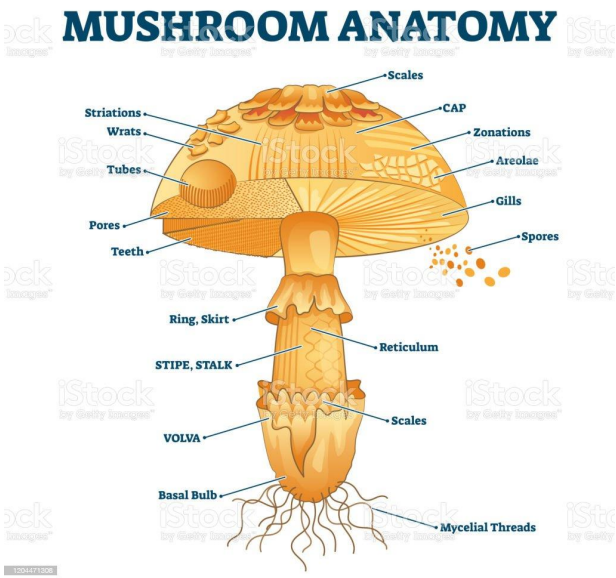
   -- edible: 4208 (51.8%) -- poisonous: 3916 (48.2%)



cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat
x	s	n	t	p	f	c	n	k	...	s	w	w	p	w	o	p	k	s	u
x	s	y	t	a	f	c	b	k	...	s	w	w	p	w	o	p	n	n	g
b	s	w	t	l	f	c	b	n	...	s	w	w	p	w	o	p	n	n	m
x	y	w	t	p	f	c	n	n	...	s	w	w	p	w	o	p	k	s	u
x	s	g	f	n	f	w	b	k	...	s	w	w	p	w	o	e	n	a	g

Understanding Mushroom Features

1. Odor
2. Stalk-surface- (above, below) - ring
3. Gill_color
4. Gill_size
5. Spore_print_color - Powder from gills
6. Ring_type - A skirt type ring tissue around the stem (large, pendant)
7. Bruises - Looks like someone pressed the mushroom. Bleeds blue or black.



Feature Selection

Why Feature Selection

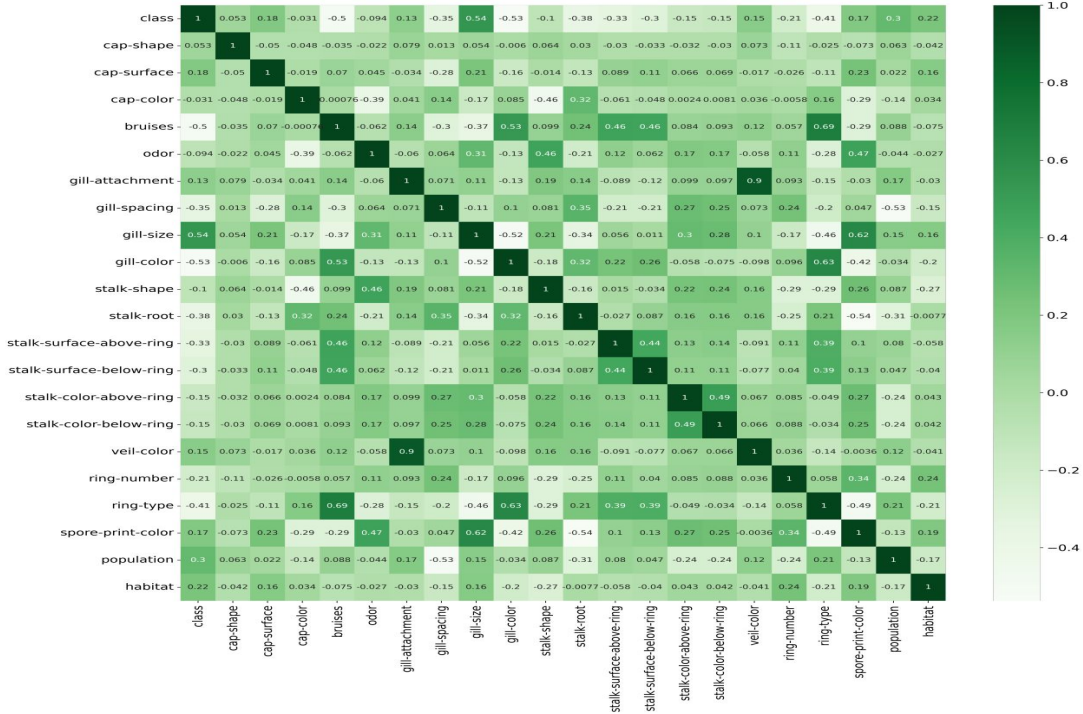
- Speed up the operation
- Overfitting Risk reduction
- Improved the model performance
- Improved Data Visualization

Methods Implemented for Feature Selection

- Correlation Matrix
- Feature importance
- Univariate feature selection Importance value

Feature Selection

feature	importance	std
odor_n	0.116294	0.059909
gill-size_b	0.070898	0.059909
odor_f	0.070394	0.059909
gill-size_n	0.055678	0.059909
stalk-surface-above-ring_k	0.047470	0.059909
gill-color_b	0.046713	0.059909
spore-print-color_h	0.045219	0.059909
ring-type_p	0.030550	0.059909
stalk-surface-below-ring_k	0.030436	0.059909
ring-type_l	0.029066	0.059909



Importance of selected features from Domain Knowledge

1. According to the literature review feature importance for mushroom classification V. Vanitha(etal) mentioned that higher score indicates that the particular features have high impact on model predictions.The main purpose of this paper is to feature selection is to eliminate the features that have the least impact on mushroom categorization for the better model performance. (https://bbrc.in/wp-content/uploads/2021/01/Galley-Proof-009.pdf)

2. According Nuanmeesri, Sriurai they mentioned feature selection criteria for the mushroom classification. In this paper they implemented the chi-square method and information gain for feature selection. They got the highest accuracy for the random forest method by using fewer features. (https://www.ijeat.org/wp-content/uploads/papers/v9i2/B4115129219.pdf)

Results with different metrics

Accuracy Comparison Table

Accuracy	Random Forest	K-Nearest Neighbor	Naive Bayes	Decision Tree
Without Feature Selection	0.9581	1.0000	0.9643	1.0000
With Feature Selection	0.9735	0.9649	0.9298	0.9766

F-1 Score Comparison Table

F-1 Score	Random Forest	K-Nearest Neighbor	Naive Bayes	Decision Tree
Without Feature Selection	0.9548	1.0000	0.9644	1.0000
With Feature Selection	0.9722	0.9617	0.9213	0.9755

Results with different metrics

Training Time Comparison Table

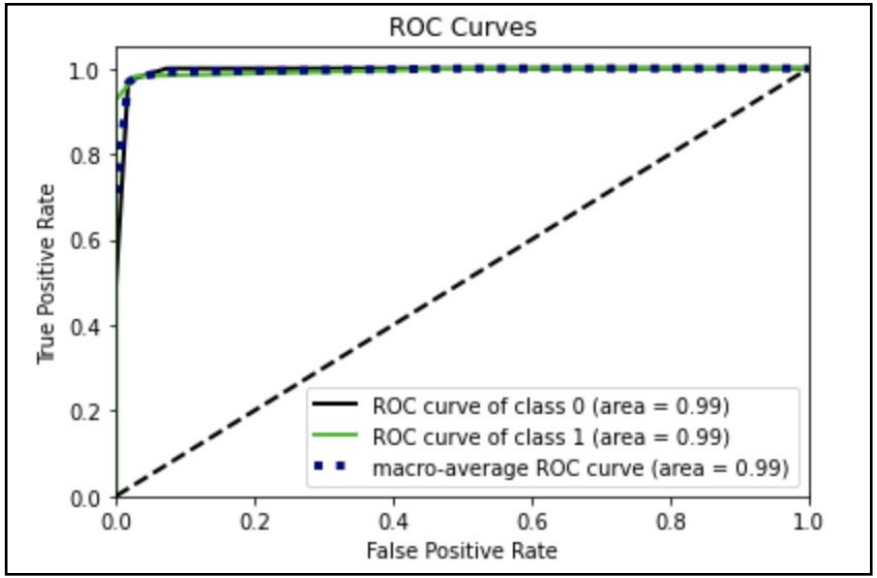
Training Time	Random Forest	K-Nearest Neighbor	Naive Bayes	Decision Tree
Without Feature Selection	2 min 14 sec	1 min 88 sec	1 min 87 sec	2 min 10 sec
With Feature Selection	1 min 54 sec	1 min 95 sec	1 min 98 sec	1 min 53 sec

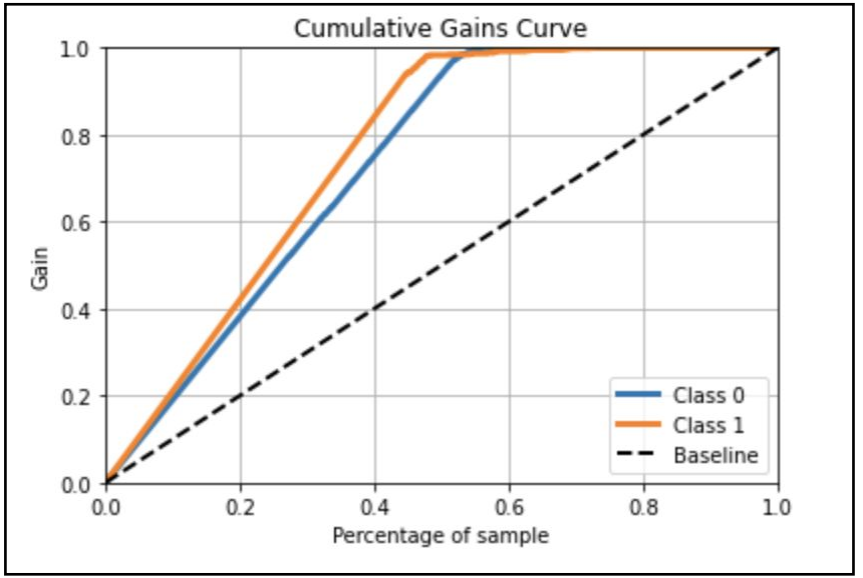
Confusion Matrix for Decision Tree Classifier

	Predict Positive	Predict Negative
Actual Positive	828	24
Actual Negative	14	759

Results with different metrics

AUC-RoC Curve and Cumulative Gain curve for Decision Tree Classifier





Why is a certain method better than other methods?

Decision Tree

- Decision Tree is better when the dataset have a “Feature” that is really important to take a decision.
- Easy interpretability
- Adapts quickly to the dataset

Why Decision Tree is Best Model for Our Dataset

1) Accuracy - 97.66%
2) F1 score - 97.55
3) Training Time - Lowest
4) False positive Rate (Incorrect predictions of poisonous mushrooms as edible) - 0.018 (FP/TN+FP)

Extracted knowledge from Data

- Generally, Green colour indicates poisonous, if the gill colour or spore print colour is green the mushrooms are not edible.
- we found that the mushrooms with orange caps (label ‘n’) are mostly poisonous and the ones with red caps (label ‘r’) are less poisonous.
- Mostly foul smelling mushrooms are poisonous.

Thanks So Mush

References

[1] https://archive.ics.uci.edu/ml/datasets/Mushroom

[2] https://www.webmd.com/diet/health-benefits-mushrooms

[3] https://en.wikipedia.org/wiki/Edible_mushroom

[4] https://ieeexplore.ieee.org/abstract/document/8405508

[5] https://www.betterhealth.vic.gov.au/health/healthyliving/fungi-poisoning