Learnbay Project Presentation

Edi-Fungi:

Safeguard foragers by distinguishing poisonous from edible mushroom

PRESENTED BY:

Theveetha T R

Agenda

Introduction

- Mushrooms : Diverse group of fungi used in cuisines.
- Challenge we pose : Difficulty in distinguishing between edible and poisonous mushrooms.
- Need : This classification or identification plays a crucial role in ensuring food safety and preventing the consumption of toxic or poisonous varieties

Objectives

Fundamental Goal

Create a model to classify mushrooms (edible or harmful) based on their attributes using Machine Learning

Purpose

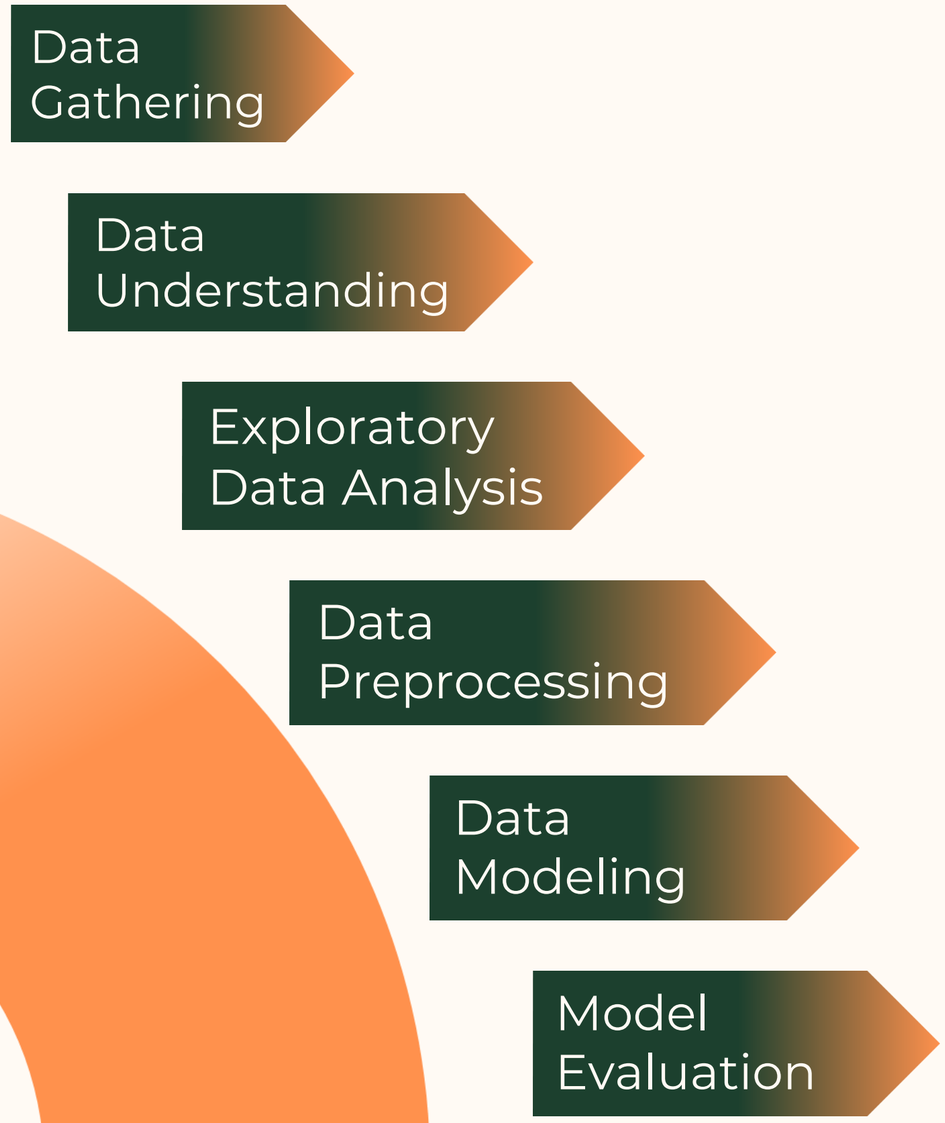
To prevent accidental intake of poisonous mushrooms and promoting safe intake of edible ones

Usage

Helps mushroom foragers to make an informed decision during mushroom foraging, ultimately ensuring safety and well being of consumers

EDIBLE
OR POISONOUS

Analysis Method



Given that the mushroom in the dataset is either classified as edible or poisonous this is a **classification problem**

Data Gathering

- Source of data: Dataset was provided as part of the project work
- The dataset provided consists of comprehensive information on labelled mushroom samples, including information on their edibility and relevant characteristics.
- The dataset covers a diverse range of mushroom species, including both edible and non edible varieties.

Data Description

Variable Name	Variable Information	Description	Additional Information
class	Binary	Mushroom Classification	poisonous = p, edible = e
cap-diameter	Metrical	Cap Diameter	float number in cm
cap-shape	Nominal	Cap Shape	bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
cap-surface	Nominal	Cap Surface	fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e
cap-color	Nominal	Cap Color	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
does-bruise-or-bleed	Nominal	Bruise or Bleed (Yes/No)	bruises-or-bleeding=t, no=f
gill-attachment	Nominal	Gill Attachment	adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, none=f, unknown=?
gill-spacing	Nominal	Gill Spacing	close=c, distant=d, none=f
gill-color	Nominal	Gill Color	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k, none=f
stem-height	Metrical	Stem Height	float number in cm
stem-width	Metrical	Stem Width	float number in mm
stem-root	Nominal	Stem Root	bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r
stem-surface	Nominal	Stem Surface	fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e, none=f
stem-color	Nominal	Stem Color	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k, none=f
veil-type	Nominal	Veil Type	partial=p, universal=u
veil-color	Nominal	Veil Color	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k, none=f
has-ring	Nominal	Has Ring (Yes/No)	ring=t, none=f
ring-type	Nominal	Ring Type	cobwebby=c, evanescent=e, flaring=r, grooved=g, large=l, pendant=p, sheathing=s, zone=z, scaly=y, movable=m, unknown=?
spore-print-color	Nominal	Spore Print Color	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k
habitat	Nominal	Habitat	grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season	Nominal	Season	spring=s, summer=u, autumn=a, winter=w

No of Rows: 61,609
No of Columns: 21

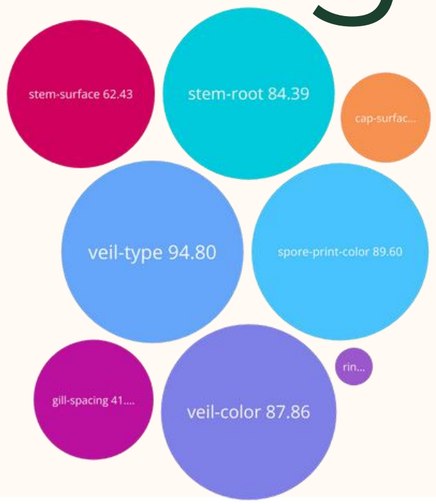
Categorical Columns: 18
Numeric Columns: 3

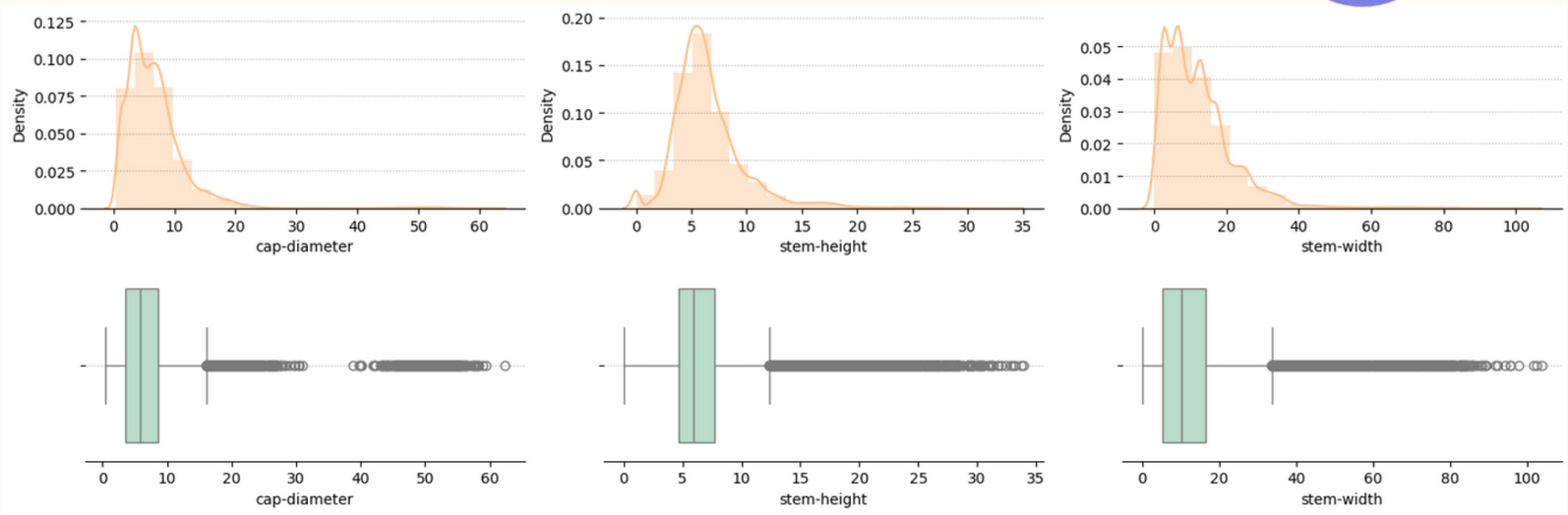
Data Understanding

Variables with missing values : 9

Variables having >60% of missing values : 5

No of duplicate records : 146





All the numerical variables are highly skewed - presence of outliers detected

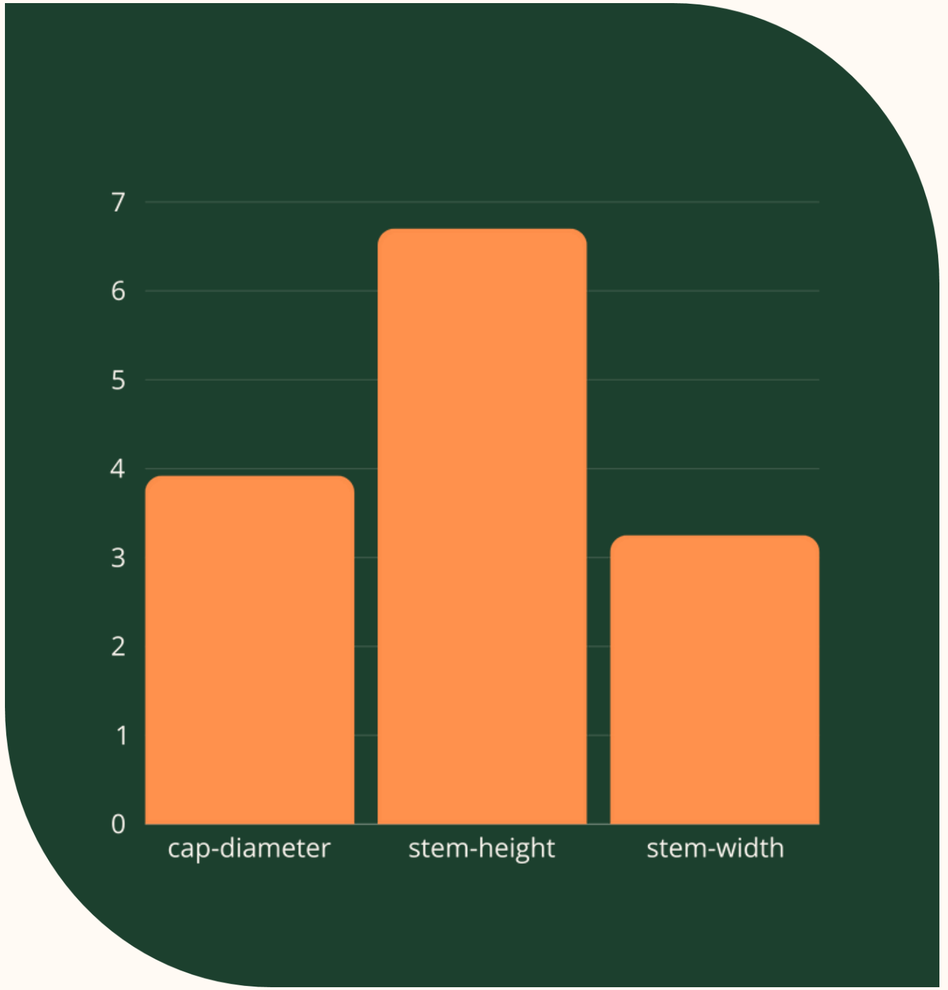
Exploratory Data Analysis

- Clean and Transform Data
- Univariate Analysis
- Bi-variate Analysis

Clean and Transform Data

- Handling missing Data:
  - > 40% - delete the variable
  - <= 40% - replace with mode
- Outliers percentage < 7% – no significant impact on analysis or business decision making.
- Treat outliers via Capping

Post Clean and Transform Data
No of Rows: 60,923
No of Columns: 15



Univariate Analysis

Single category accounts for > 50% of the data, indicating a significant class imbalance within these features.

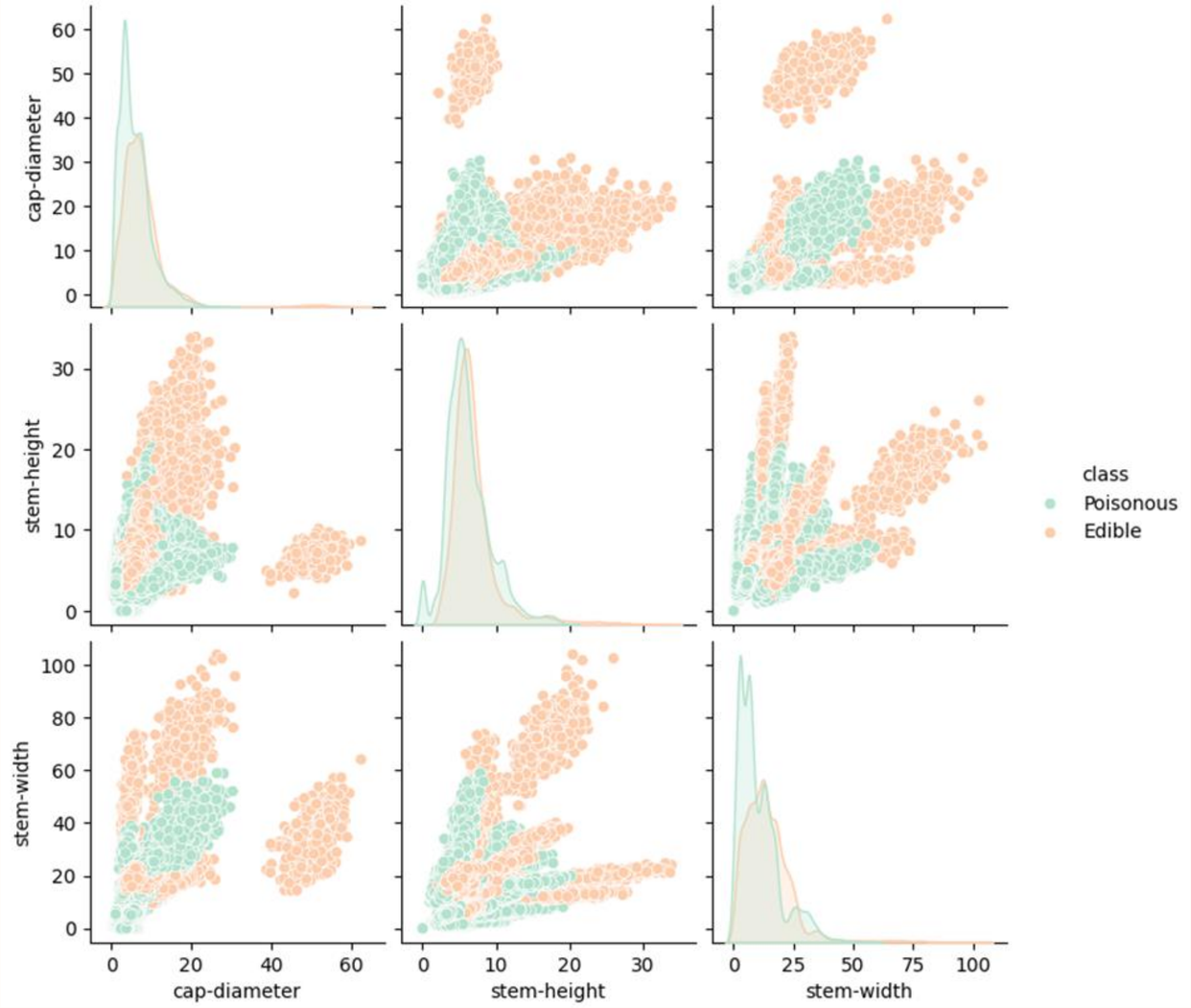
- ring-type - none (f) - 83.2
- bruise-or-bleed - false (f) - 82.62
- has-ring - false (f) - 75.1
- habitat - woods (d) - 72.33

Variable Name	Most common category	Percentage
ring-type	f - none	83.2
does-bruise-or-bleed	f - false	82.62
has-ring	f - false	75.1
habitat	d - woods	72.33
class	p - poisonous	55.38
season	a - autumn	49.47
cap-shape	x - convex	44.21
cap-color	n - brown	39.73
stem-color	w - white	37.63
gill-attachment	a - adnate	37.02
cap-surface	t - sticky	36.61
gill-color	w - white	30.35

Bi-variate Analysis

Numeric variables

- Numeric variables strong correlation between them
- Large mushrooms - cap-diameter, stem-height and width - good indicator for edible mushroom
- Medium-sized mushrooms - equal distribution of edibles and poisoners
- Minimum stem height, width, and cap diameter - poisonous category



Bi-variate Analysis

Categorical variables

Habitat and Season

- Poisonous:
  - Without rings/ zoned rings
  - Woods and Paths
  - Autumn season
- Edible:
  - Movable ring type
  - 100% - Urban and Waste habitat
  - Spring and Winter

Cap specifications

- Poisonous:
  - Bell shaped cap
  - Silky or Fibrous cap
  - Green cap colored
- Edible:
  - Spherical shaped cap
  - Shiny, scaly, or smooth cap
  - Brown, gray, and buff

Other Specifications

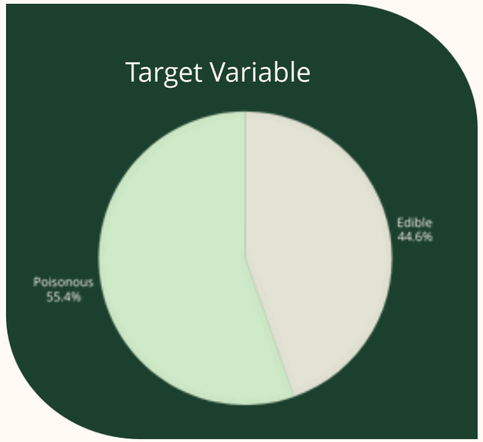
- Poisonous:
  - Adnate gill attachment
  - Yellow and Brown (gill/ stem color)
- Edible:
  - Free or Pores gill attachment
  - White, Buff and Grey (gill/ stem color)

Data Pre-processing

Imbalance Data

Balanced Distribution – no need of SMOTE

Edible: 44.62%
Poisonous: 55.4%



One hot Encoding

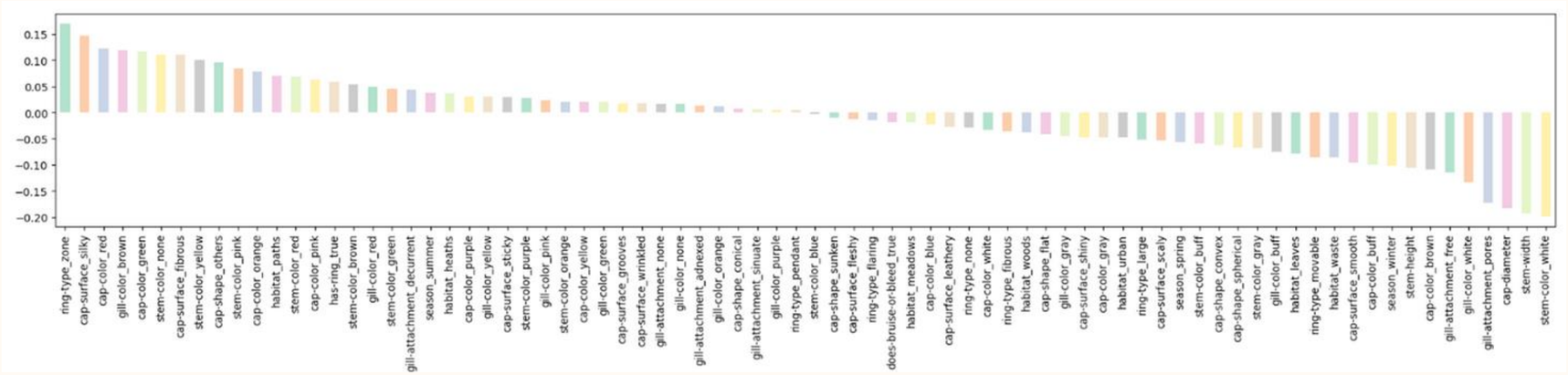
Convert categorical variable to numerical variable via one hot encoding

Problem statement - increased column count (78)

Feature Selection

Threshold-based feature selection – evaluate correlation matrix and eliminate less significant variables

Purpose - enhance model efficiency and performance.
No of rows: 60,923; No of columns: 38



Data Modelling

- Train-Test split : 80%-20%
- The chosen algorithms:
  - Decision Tree
  - Random Forest
  - K Nearest Neighbour
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machine
  - XGBoost

- Comparison of predicted metrices for chosen algorithms

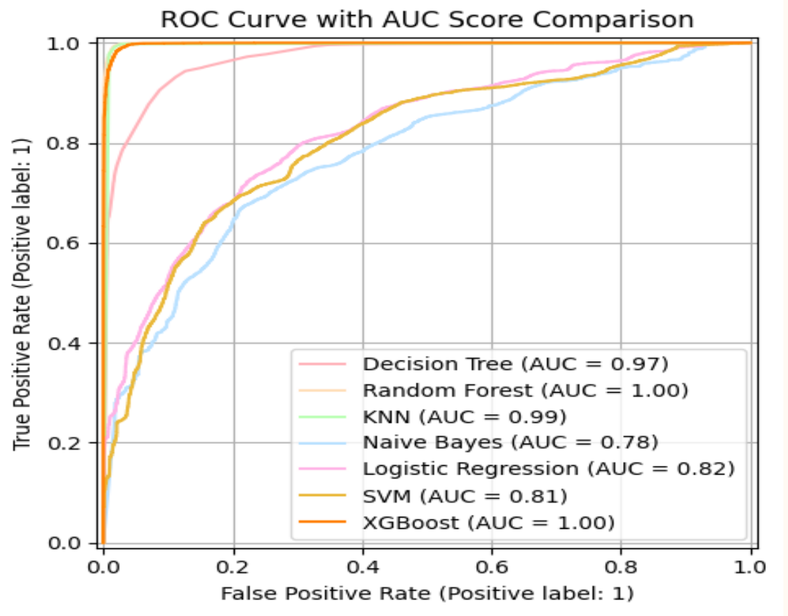
method	accuracy	precision	recall	f1	roc_auc
Decision Tree	91.24	91.41	93.09	92.24	90.98
Random Forest	98.84	98.6	99.34	98.97	98.78
K-Nearest Neighbour	98.76	98.64	99.15	98.9	98.71
Naive Bayes	67.78	82.77	53.59	65.06	69.7
Logistic Regression	74.72	77.91	76.54	77.22	74.48
Support Vector Machine	72.54	77.25	72.2	74.64	72.59
XG Boost	98.17	97.87	98.89	98.37	98.07

Model Evaluation

- Primary evaluation metric - **Precision**

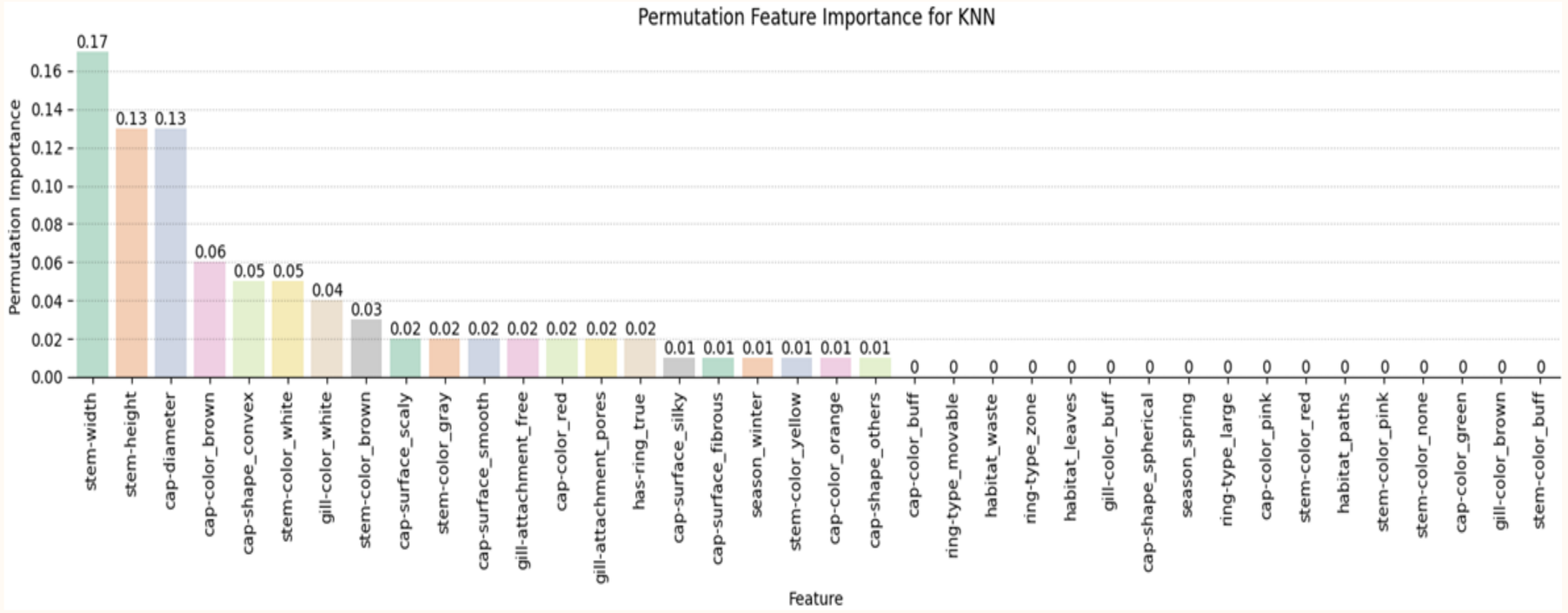
- Reason - we can prioritize minimizing the risk of misclassifying poisonous mushrooms (crucial safety concern).
  Precision = True Positive / (True Positive + False Positive)

- Random Forest – Eliminated
  - Overfitted as ROC curve with AUC Score = 1 though it is best fit in Precision metric



Conclusion

- Inference based on primary metric precision and AUC score - K Nearest Neighbour (KNN) algorithm achieved highest precision score.
- **KNN is the most effective model** for this task, offering the best balance of accuracy and safety in classifying mushrooms.
- Features contributing for KNN model classification
  - **Stem-width**
  - **Stem-height**
  - **Cap-diameter**



Thank You