Swinburne University of Technology Hawthorn Campus
Dept. of Computer Science and Software Engineering

COS10022 Introduction to Data Science

Assignment 2 - *Semester 1, 2021*

Assessment Title: Prediction Model Creation and Evaluation

Assessment Weighting: 30%

Due Date: Saturday, 8th May 2021 at 11:59 pm (AEDT)

Assessable Item:

- A written report of **no more than** 10-page long with the Assignment Cover Sheet. The plagiarism check must be **lower than 20%** in the main content part. Otherwise, it is not considered a qualified submission and the submission will not be marked.

Purpose of Assignment

This assignment aims at evaluating students' achievement of the following unit learning outcomes:

1. **Appreciate (and explain) the key concepts, techniques, and tools for handling the data and create prediction models.**
2. **Experiencing the model selection and implementation for carrying out a data science project.**

This is an **individual** assignment. You need to start with the data cleaning process, make decisions on which model to use, implement the model, and get the outcome for analysis.

Refer to the Unit Outline for the late submission penalty and group work policy.

"Creating a Prediction Model based on the Training Data"

Key Lessons:

In many practical situations, data scientists are required to build a forecasting/prediction model based on the existing data. The prediction result helps the business to be prepared ahead for the foreseeable scenarios. In this assignment, the goal is to identify the type of the collected mushroom to prevent accidentally poisonous cases.

Introduction

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

Inspiration

- What types of machine learning models perform best on this dataset?
- Which features are most indicative of a poisonous mushroom?

Acknowledgements

This dataset was originally donated to the UCI Machine Learning repository. You can learn more about past research using the data here.

Assignment Goal

The goal is to build a predictive model to predict whether the input mushroom is eatable. The predictive model will help the users to prevent accidentally be poisoned caused by the lack of knowledge in identifying the mushroom type.

Assignment Task

Your task is to select and create a predictive model for identifying whether a collected mushroom is eatable. The length of the report should not be more than 10 pages, including the title page and the reference page (single line space; 11pt font; Arial). Table of Content is not required.

There are 100 marks in this assignment. Your report must address the following tasks.

---

1. The files "mushrooms.csv" contains 23 attributes. The first attribute identifies whether the mushroom is eatable. The remaining attributes are describing the characteristics of a mushroom. Identify missing values (report the number of missing values and the attributes) and come up with strategies for applying treatments. **(9 marks)** Identify the attribute type from the given dataset and point it out in your report. If the model you chose in the next step requires a particular data type, convert the attributes to fit the requirements and explain how you did it. If there is no need to convert your attributes, explain the reason as well. **(6 marks)** You need to decide what attributes you are going to use with reasons. **(6 marks)** After selecting attributes, shuffle the data **(3 marks)** and then partition the dataset into three equal size pairs of training

set and test set with the training set size being 90% of the raw data. (Each pair would contain 1/3 of the records.) Explain how you do it in the report. **(6 marks)** The concept of splitting the dataset is depicted below. Report the number of records you have in each set.

The Whole Shuffled Dataset					
Training 1	Test 1	Training 2	Test 2	Training 3	Test 3

[30 marks]

2. Create a predictive model and train it with the training data and test the accuracy with the corresponding test set. Repeat the same process for three pairs and record the results. For every repeat (using different training sets), your model should be trained from scratch. Create the **confusion matrix** and find the **accuracy**, **true positive rate, false positive rate, false negative rate, and precision** for every repeat. The descriptions above are all for the outcome from the test sets.
   [50 marks]

3. Use any data visualisation method to present your findings and compare the outcomes obtained from each test result.
   [10 marks]

4. Present all your answers in the form of a high-quality written report.
   [10 marks]

The raw data in Assignment 2 is from Kaggle website, which is available at: https://www.kaggle.com/uciml/mushroom-classification. There are already abundant works dedicated to studying the problem of predicting the outputs. Similar works can be found online. You are encouraged to explore some of the existing literature and, where applicable, adapt their ideas but not copying their results into your work. When you do so, please include all the necessary **in-text citations** and the **end-of-report reference list**.

The Harvard Referencing format must be used when citing and referencing external information resources: http://www.swinburne.edu.au/library/referencing/harvard-style-guide/

Submission Requirement

To fulfil the requirement of this assignment, the following item must be submitted in **pdf** format and named as: ***COS10022_Assignment2_YourStudentID***. The submission should include a single file:

- A written report of **not more than** 10-page long with the Assignment Cover Sheet (digitally signed).

Failure to adhere to the submission requirements will immediately result in "N" grade for this assignment.

Data Dictionary

Attribute	Information
class	edible=e, poisonous=p
cap-shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
cap-surface	fibrous=f, grooves=g, scaly=y, smooth=s
cap-color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
bruises	bruises=t, no=f
odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
gill-attachment	attached=a, descending=d, free=f, notched=n
gill-spacing	close=c, crowded=w, distant=d
gill-size	broad=b, narrow=n
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
stalk-shape	enlarging=e, tapering=t
stalk-root	bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
stalk-surface-above-ring	fibrous=f, scaly=y, silky=k, smooth=s
stalk-surface-below-ring	fibrous=f, scaly=y, silky=k, smooth=s
stalk-color-above-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
stalk-color-below-ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
veil-type	partial=p, universal=u
veil-color	brown=n, orange=o, white=w, yellow=y
ring-number	none=n, one=o, two=t
ring-type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

Rubric: Subtask 1

Marks	Data Preparation (30 marks)
9.0	Missing values are all identified and corresponding treatments are made with reasons.
6.0	The attribute types are identified correctly. The process of how to convert the attributes is explained if any. Otherwise, the reason of not converting the attributes should be given.
6.0	Reveal the attributes selected to be used in your model with sufficient reasons. If all attributes are used, the reason should still be revealed.
3.0	Correctly shuffle the records and partition the dataset into correct pairs.
6.0	The procedures and methods used to shuffle and partition the data are properly explained.

Rubric: Subtask 2

Marks	Predictive Model (50 marks)
10.0	The selection of the predictive model is reasonable.
10.0	The description of the training/testing procedures are all correct. All accuracies from different test set are all revealed.
10.0	The confusion matrix is revealed.
5.0	The true positive rates are revealed.
5.0	The false positive rates are revealed.
5.0	The false negative rates are revealed.
5.0	The precision for all repeats is revealed.

Rubric: Subtask 3

Marks	Visualisation (10 marks)
7.1 - 10.0	The outcome is presented with splendour and informative visualisation. The outcomes from different test results are easy to be compared in the visualisation.
5.1 – 7.0	The outcome is presented with quality visualisation and contains sufficient information. The outcomes from different test results are somehow can be compared in the visualisation.
0.1 – 5.0	The outcome is presented with general and limited visualisation with basic information. The comparison between different test results is not straight forward to be observed in the visualisation.
0.0	There is no visualisation for presenting the findings and comparison.

Rubric: Subtask 4

Marks	Report Writing (10 marks)
5.1 - 10.0	The report looks professional and the content is meaningful, informative, and neat.
0.1 – 5.0	The report is not easy to be read and understand or some key elements are missing.
0.0	The submitted report is incomplete, with a high plagiarism rate, or not readable.