Logo

Description automatically generated

University of Science

Faculty of Information Technology

. . . . . . .o0o. . . . . . .

University of Science

Faculty of Information Technology

. . . . . . .o0o. . . . . . .

University of Science

Faculty of Information Technology

. . . . . . .o0o. . . . . . .

University of Science

Faculty of Information Technology

. . . . . . .o0o. . . . . . .

# Vietnam National University - Ho Chi Minh City

# Vietnam National University - Ho Chi Minh City

# Vietnam National University - Ho Chi Minh City

# Vietnam National University - Ho Chi Minh City

# LAB REPORT 02:

# DECISION TREE WITH SCIKIT-LEARN

Course: Artificial Intelligence

**Theory teacher: Nguyen Ngoc Thao**

**Instructors practice: Nguyen Thanh An**

**Ho Thi Thanh Tuyen**

**Le Ngoc Thanh**

**Student: Hoang Huu Minh An ID: 20127102**

**Class: 20CLC01**

Contents:

[I. Information: 2](#_Toc111382526)

[II. Introduction: 2](#_Toc111382527)

[III. Assignment requirements: 2](#_Toc111382528)

[**a)** **Preparing the data sets:** 2](#_Toc111382529)

[**b)** **Building the decision tree classifiers:** 2](#_Toc111382530)

[**c)** **Evaluating the decision tree classifiers:** 5](#_Toc111382531)

[**d)** **The depth and accuracy of a decision tree:** 10](#_Toc111382532)

[IV. List of References: 14](#_Toc111382533)

1. **Information:**

This Lab02: Decision tree presented by Hoang Huu Minh An-20127102.

1. **Introduction:**

About the Mushroom data set:

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. Each species is identified as either edible or poisonous. There are 8124 samples, each of which is characterized by 22 attributes and a target attribute.

* The target attribute is located at the first column
* The original data comes from the UCI Machine learning repository. You may refer to this page for more information

1. **Assignment requirements:**
2. **Preparing the data sets:**

* Read data from file ‘mushrooms.csv’ using read\_csv . function.
* Count the number of output ouputs
* Visualize columns of data
* Representing character data into corresponding representation data
* From the table above it can be seen that the column "veil-type" is 0 and not contributing to the data so I remove it.
* Shuffle data using the function shuffle
* Split the given data into the first column is output, the remaining 21 columns are data to train
* Split the data into train test sets using the split\_split\_test function with the x, y parameters respectively the data that I have split before, test\_size is the percentage of data that I want to split .Text

  Description automatically generated with low confidence

1. **Building the decision tree classifiers:**

* 40/60:

Timeline

Description automatically generated

* 60/40

Timeline

Description automatically generated

* 80/20:

Timeline

Description automatically generated with medium confidence

* 90/10:

Timeline

Description automatically generated

1. **Evaluating the decision tree classifiers:**

Classification tree predict a qualative response, in contrast to regression tree that predict quantitive response. Classification tree predicts the observation that belong to the most common occurring class from the training data for each region. In order to grow the classification tree the algorithm uses recursive binary splitting.

* 40/60:

Chart, treemap chart

Description automatically generated

Calendar

Description automatically generated with low confidence

* 60/40:

Chart, treemap chart

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

* 80/20:

Chart

Description automatically generated

Graphical user interface

Description automatically generated

* 90/10:

Chart

Description automatically generated

A picture containing text

Description automatically generated

* They are some of the mostly used measures in evaluating how good your system works.
* **Precision:**It is calculated with respect to the predicted values. For class-A, out of total predictions how many were really belong to class-A in actual dataset, is defined as the precision. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] column.
* **Recall:** It is calculated with respect to the actual values in dataset. For class-A, out of total entries in dataset, how many were actually classified in class-A by the ML model, is defined as the recall. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] row.
* **F1-score:** It is the harmonic mean of precision and recall.
* **Support:**It is the total entries of each class in the actual dataset. It is simply the sum of rows for every class-i

1. **The depth and accuracy of a decision tree:**

* **Max\_depth: 2**

**Diagram

Description automatically generated**

* **Max\_depth: 3**

Diagram

Description automatically generated

* **Max\_depth: 4**

Diagram

Description automatically generated

* **Max\_depth: 5**

A picture containing text, sign

Description automatically generated

* **Max\_depth: 6**

A picture containing timeline

Description automatically generated

* **Max\_depth: 7**

Timeline

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Max\_deapth | None | 2 | 3 | 4 | 5 | 6 | 7 |
| Accurary | 1.0 | 0.9046 | 0.9606 | 0.9655 | 0.9823 | 0.9969 | 1 |

* Length max tell me how deep can the tree be so when there is one node when there is one decision that's depth ones so when you make the decision or when there's only one node then that's a depth of one when you have a decision node there and then you split. Sometimes decision trees can grow to be very long so if you wanted something these things are not really things that you can.

1. **List of References:**

[1] Scikit-learn decision trees: <https://scikit-learn.org/stable/modules/tree.html>

[2] Analysis and classification of Mushrooms: https://www.kaggle.com/haimfeld87/analysis-and-classification-of-mushrooms

Thank you for waching !!