CIDM 6355 Data Mining Methods HW1

(60 points in total; Due 11:59 PM Central Time, February 12, 2024)

**Requirements:** Follow the instruction, take the required screenshots with date and time (see the examples in RapidMiner Lab instruction), and answer all the questions. Sharing your queries, screenshots, or answers with other students is considered as cheating, which will be reported to the university authority. A screenshot without showing reliable date and time will receive a penalty of 50% of points. If identical screenshots are found from two or more students, such a misconduct will be reported to the university authority. Please type your name as below to indicate that you understand and comply with all the requirements in this homework.

Name: Mehnaz Afrose

Part 1: Answer all the questions in Week 4 RapidMiner Lab (Step 1.3, 1.8.1-1.8.5, and 2.2.3) and an additional question via HW1-Part 1 Submission (30 points) on WTClass. You have two attempts and the higher one will be counted into your grade. Please DO NOT include them here; otherwise, they won’t be graded here.

Part 2 Lab Screenshots and Deliverables (30 points)

Take the required screenshots with date and time and answer all the questions. Windows and MacBook show the date and time differently, so your screenshot is acceptable as long as it displays the date and date, no matter how. MacBook displays the date and time on the top right corner. If you do not know how to take a screenshot, please check this website <https://www.take-a-screenshot.org/> for more instructions. If you do not know how to show the date and time on your MAC Book, Google your question or try [this site](https://osxdaily.com/2014/06/23/show-date-menu-bar-mac-os-x/). Sharing your queries, screenshots, or answers with other students is considered as cheating, which will be reported to the university authority.

1) Screenshots in RapidMiner Lab (6 points)

* Screenshot 1: A screenshot of the decision tree graph with date and time at Step 1.8 (3 points)

A computer screen shot of a computer screen

Description automatically generated

* Screenshot 2: A screenshot of prediction results for the 19 observations with date and time in Step 2.2 (3 points)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

2) Deliverables in R Lab (24 points)

Please first indicate which method you use, Library rpart or party? [If you miss this question, a penalty of 5 points will be applied.

I used the first method: **Library rpart.**

* Deliverable R1: take a screenshot of your decision tree model with date and time (4 points).

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Deliverable R2: take a screenshot of your decision tree graph with date and time and briefly describe it. Your description must include the root node, split nodes, and leaf nodes. (7 points: 4 points for your screenshot and 3 points for your description).

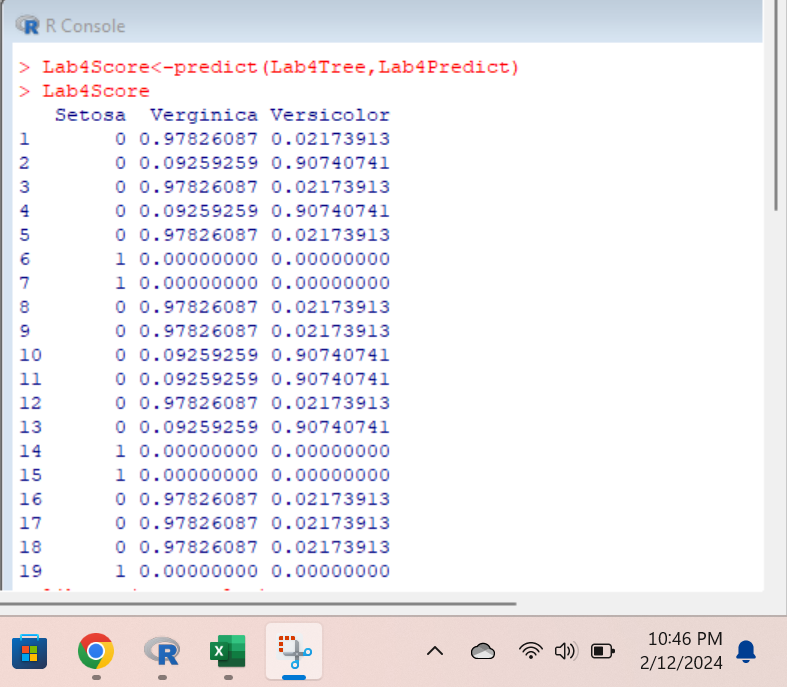
**Description:** The decision tree model is generated using one category: petal width. The model consists of one split node and three leaf nodes along with one root node. The root node is governed by the condition "Petal\_width < 0.8," while the split node is contingent upon the condition "Petal\_width >= 1.8." According to the graph, flowers with a petal width less than 0.8 can be identified as Setosa. Conversely, flowers with a petal width greater than or equal to 1.8 can be further divided into two categories: Virginica and Versicolor.

For Setosa, it shows 100% pure. On the other hand, Verginica is 98% and Versicolor is 91% pure.

A screenshot of a computer

Description automatically generated

* Deliverable R3: after you apply the decision tree model to your prediction dataset, take a screenshot of the prediction result with date and time and briefly describe how the result help you determine the predicted class of each case. (7 points: 4 points for your screenshot and 3 points for your description).



**How the result help you determine the predicted class:** The prediction results comprise 19 outputs categorizing flowers into three classes: Setosa, Virginica, and Versicolor. Upon examination, the initial observation is labeled as Virginica; however, this classification is not pure, as the confidence level for Virginica is 0.978. In contrast, result number 6 unambiguously identifies the category as Setosa, demonstrating purity with a confidence score of 1. The third category, represented by result observation number 13, is labeled as Versicolor, yet it is not entirely pure, with a confidence score of 0.907. Analyzing the total of 19 outputs, 5 are classified as Setosa, 5 as Versicolor, and 9 as Virginica. This breakdown aids in determining the predicted class within the decision tree model results.

* Deliverable R4: Compare the decision tree models generated in RapidMiner and R, and then point out at least two differences you observe. (6 points: 3 points for each qualified difference). When discussing each difference, please include both R and RM. For example, R does ….., but RM does not …… .

**Answer:** RapidMiner offers a user-friendly, visual interface for decision tree modeling, while R provides more flexibility and control through coding, making it suitable for users with programming skills. However, they may not provide the same results with the same dataset. In this instance, the models diverged, RapidMiner classified 6 flowers as Setosa and 4 as Versicolor, while R indicated 5 Setosa and 5 Versicolor.

**Difference 1:** RapidMiner produced 4 split nods and 6 leaf nodes along with one root nod. On the other hand, R only shows one split nod and three leaf nodes along with the root nod.

**Difference 2:** In the RapidMiner, it classified based on two categories: petal width and petal length. R classified based on one category: petal width.