CIDM 6355 Data Mining Methods

(100 points in total; Due 11:59 PM Central Time, July 19, 2025)

**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: Bolaji Muyiwa

Part 1: Answer all the questions in Week 4 RapidMiner Lab (Step 1.3, 1.8.1-1.8.5, and 2.2.3 – 2.2.4) and an additional question via HW1-Part 1 Submission (40 points). 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 (60 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 (10 points)

* Screenshot 1: A screenshot of the decision tree graph with date and time at Step 1.8 (5 points)

A screenshot of a computer

AI-generated content may be incorrect.

* Screenshot 2: A screenshot of prediction results for the 19 observations with date and time in Step 2.2 (5 points)

A screenshot of a computer

AI-generated content may be incorrect.

2) Deliverables in R Lab (50 points)

* **Deliverable R1**: take a screenshot of your decision tree model with date and time (5 points).

A screenshot of a computer

AI-generated content may be incorrect.

* **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. (10 points: 5 points for your screenshot and 5 points for your description).

A screenshot of a computer

AI-generated content may be incorrect.

For this decision tree graph, the flowers are classified into three species which are Setosa, Verginica and Versicolor

**Root Node**: The root node here starts with the Petal\_width < 0.8 which is Setosa

**Spilt Node**: The spilt node for this graph consists of Verginica or Versicolor with the Petal\_width >=1.8

**Leaf Node**: For this graph, the leaf node consists of Setosa with Petal\_width < 0.8, Verginica with Petal\_width >=1.8 and Versicolor with Petal\_width <1.8

* **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. (10 points: 5 points for your screenshot and 5 points for your description).

A screenshot of a computer

AI-generated content may be incorrect.

The Prediction result helps determine the confidence level for each of the flower Setosa, Verginica and Versicolor and more importantly help determine the **Species name** using the Confidence level to determine individual Species name.

Referencing the screenshot above, the first Species name based on the Confidence score is Verginica with 0.97826087, Follow by Versicolor with the Confidence level of 0.90740741 etc.

* **Deliverable R4**: take a screenshot of your decision tree model with date and time. Try to use the resources provided to understand its output (5 points).

A screenshot of a computer

AI-generated content may be incorrect.

* **Deliverable R5**: take a screenshot of your decision tree graph with date and time (5 points).

A screenshot of a computer screen

AI-generated content may be incorrect.

* **Deliverable R6**: after you apply the decision tree model to your prediction dataset, and take a screenshot of the prediction result with date and time (5 points).

A screenshot of a computer

AI-generated content may be incorrect.

* **Deliverable R7:** Choose one of the two decision tree models generated in R and compare it with the decision tree model generated in RapidMiner. Identify and discuss at least three differences between the two models. When discussing each difference, please include both R and RM. For example, "R does …, but RM does not …." (10 points).

|  |  |
| --- | --- |
| Rapid Miner | R |
| For the decision tree mode, RapidMiner does not show the variable importance in the result and output, however prediction shows dependance on Petal\_width and Petal length | For R model, it does list and shows the variable importance such as Petal\_width being 34% and Petal length being 31% followed by the Sepal\_length and Sepal\_width with 21% and 14% respectively |
| RapidMiner predictions do not have clear details on splits | R provide extensive view of the tree structure and splits, referencing the Decision Tree model, The tree splits first on Petal\_width < 0.8 to separate Setosa, and further splits on Petal\_width < 1.75 to set apart Versicolor and Virginica. |
| RM shows the confidence scores for each individual prediction and mainly focuses only on the final prediction and confidence scores. | R instead provide probabilities for each class at the nodes. |