CIDM 6355 Data Mining Methods HW1

(60 points in total; Due 11:59 PM Central Time, September 17, 2023)

**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: Jordan Unfred

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 screenshot of a computer

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

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.

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

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).

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).

A screenshot of a computer

Description automatically generated

To determine the predicted class of each case based on this data, you would typically choose the class with the highest predicted probability. In other words:

* For each row, compare the predicted probabilities in the columns for Setosa, Virginica, and Versicolor.
* Select the class with the highest probability as the predicted class for that case.

For example, if a row has the highest probability in the Setosa column, then the predicted class for that case would be "Setosa." Similarly, if another row has the highest probability in the Virginica column, then the predicted class for that case would be "Virginica."

The data helps determine the predicted class of each case by providing the predicted probabilities for each class, and the class with the highest probability is chosen as the prediction for that case.

* 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 …… .

***Data Presentation***

**R Data:** The R data is presented in a structured, tabular format. It resembles a comprehensive report, featuring various sections and detailed statistics. It meticulously illustrates the construction of the decision tree, highlighting elements like complexity parameters, variable importance, node numbers, class counts, and probabilities. This comprehensive approach provides an in-depth view of the decision tree model's structure and performance.

**RapidMiner Data:** In contrast, the RapidMiner data employs a more straightforward, textual representation. It simplifies the complex decision tree into a step-by-step narrative, making it more accessible to read. However, this simplicity comes at the cost of omitting many of the numerical details present in the R data.

***Data Format:***

**R Data:** The R data is meticulously organized, with clear headers and a logical flow. It embodies a rich dataset designed explicitly for decision tree analysis, ensuring all pertinent information is readily accessible.

**RapidMiner Data:** The RapidMiner data takes on a more concise, textual format. It adopts a hierarchical structure, like a written story, detailing the conditions and outcomes at different branches of the decision tree. While this format simplifies understanding it lacks the extensive statistical data provided by R.