CSIT-558: Data Mining

Assignment 4: Descriptive Data Mining

In this assignment, students will implement concepts in the realm of descriptive data mining on data sets using either association rules or sequence mining or cluster analysis.

1. Students should work with data of their choice, e.g. related to their research topics, MS project or simply their own areas of interest.
2. Work individually or with teams of two to three people. If you work in a team, you still need to submit the project report individually, and the individual report needs to be different.
3. Collect suitable data of your choice from any online repository, e.g. UCI, Kaggle etc. This forms the given dataset for the assignment.
4. Use the python-based popular industry data mining software packages and development/analysis tools - **Panda, Scikit-Learn, Seaborn,** etc. Study the related documentation from the software user guide.
5. On the given dataset, execute any predictive data mining technique such as decision trees, Bayesian classifiers, neural networks, k-nearest neighbors, ensemble learning, linear regression etc.
6. You can implement multiple techniques, one followed by another on the same dataset if you prefer to use this for your analysis.
7. If needed, you should convert the format of your dataset as needed by the respective technique(s). If you have already performed conversion in the assignment on data preprocessing, you can use the converted data here.

8. If you would like to use the results of your descriptive data mining techniques and conduct further analysis with predictive data mining, that is fine as well.
9. You should aim to achieve robustness and generalization in the mining, e.g. by altering seeds in the algorithm.
10. You must modify the concerned parameters, e.g. learning rate and error threshold for neural networks, the value of k for k-nearest neighbors etc. to get good results. Execute at least 3 different combinations of parameters and present the experimental results accordingly for at least one technique.
11. Observe the experimental results and draw useful conclusions from the data.
12. Based on the hypothesis obtained by the learning in the predictive data mining technique(s), write a simple program that uses the learned hypothesis to classify new data, and thus serves as a prototype mini classification tool.
13. If you have used multiple techniques, you can select the one that gives greatest accuracy or the one that is most suitable to your data and domain etc.
14. The program should communicate with the user to give outputs based on new unseen data. For example, consider that the learned hypothesis is based on a dataset that uses weather data to predict if it is okay to play tennis. This can include various attributes such as “temperature”, “chance-of-rain” and so forth to estimate the target “playing tennis” as being “yes”, “no” or “maybe”. It can be used to manually derive rules such as “if temperature = medium and chance-of-rain = low then play-tennis = yes” which can be coded into the program. Thereafter, when a user inputs new values for parameters such as “temperature”, the program can use these rules to output the classification

target and thereby suggest to the user “Yes, you can surely play tennis today” or “No, you should not play tennis today” or “Maybe, it seems okay to play tennis today”. This example can be modified based on your dataset and domain.

15. Show at least 3 different runs of such inputs and outputs based on user interaction. The final goal is to have simple communication with the user based on the learning done via predictive data mining for classifying the target attribute. Hence, please work accordingly based on your respective technique(s) and application. You can tune this as per the needs of your MS project, research topics or any other area of interest.
16. Please make slides based on your work in this assignment and upload them on Canvas.
17. There should be minimum 10, maximum 20 slides. ‘
18. The slides should include:
    - The dataset with a short explanation about the goals of your classification
    - Mention of the predictive data mining technique(s) with justification
    - Execution of the experiments with this technique, using at least 3 distinct combinations of suitable parameters
    - Relevant code snapshots of the program (in any language) for the prototype mini classification tool implemented using the learned hypothesis
    - Demo of the tool with suitable inputs and outputs, showing at least 3 runs of user interaction
    - Any references used including data sources, software tools, text books etc.

- Mention of the descriptive data mining technique and algorithm with justification (e.g. sequence mining because you are working with DNA data etc.)
- Execution of the experiments with this algorithm, using at least 3 distinct combinations of parameters (show inputs & outputs)
- Conclusions drawn from all the observations with the descriptive data mining technique
- Any references used including data sources, software tools, text books etc.

19. Please turn in the slides on the due date. Slides should be uploaded on Canvas individually by each student, even though you are working in groups.The slides should be different even if you work in a team.