CMP 466 – Machine Learning and Data Mining

Project

Project report and presentation

You need to submit the following by the deadline:

A. Report:

Your report should have the following sections:

1- **Abstract:** A summary of your project

2- **Introduction:** Introduce the problem and its importance, articulate a precise research question, summarize your novel contributions, and provide a brief roadmap of the report's sections.

3- **Literature Review.** Recent research papers related to your project should be summarized (2 – 3 papers by each team member). Papers using the same datasets are preferred. Other papers using similar datasets/ideas can be included. When summarizing any paper, try to focus on the purpose, methods used, dataset (public or private, its name), the performance metrics used and their values, and the overall findings of the work. Provide a critical review of each paper and an overall conclusion of the shortcomings of the previous work that helped you identify your contribution.

4- **Methods and Datasets:**

4.1 Dataset: Description of the dataset to be used in your project: source, type, size (number of samples), dimensionality (number of features), class labels, the number of samples in each class, descriptive summary of the features, etc.

4.2 Data Preprocessing Methods: Theoretical description of the methods used such as missing value handling, encoding, scaling, standardization, feature selection, etc.

4.3 Machine Learning Methods: Theoretical description of the methods used. You need to cover at least the following methods: KNN, logistic regression, decision trees, SVM (using linear or non-linear kernels).

4.4 Evaluation metrics: Theoretical description of the methods used: accuracy, F1-score, precision, and recall per class, AUC, ROC curve, confusion matrix, etc.

5- **Results:**

5.1 Data Preprocessing: Report the data preprocessing results

5.2 Machine Learning Results: Report the results of applying at least the following classifiers introduced in Section 4.3 to your dataset:

a. KNN
b. Logistic regression
c. SVM (linear or non-linear kernels)
d. Decision trees and ensemble methods

- For each of the classifiers, report the evaluation metrics as mentioned in Section 4.4. Choose the best models by optimizing the hyper parameters. Include plots and/or tables showing the models performance for different hyper parameter values and choose the optimal hyper parameters' values that make your models well-fitted (not overfitted or underfitted).

- For each of the classifiers, report the results with/without feature selection and dimensionality reduction.

6- **Discussion:** Discussion of the results and comparison with related works mentioned in the literature review.

7- **Conclusion:** Main conclusions of this work.

B. Contribution of each member

The contribution of each team member must be stated clearly in a separate file and/or in the report (by having a different font color for the parts done by different members).

C. Presentation

Prepare a presentation summarizing the work.

Submission: Please submit your assignment to iLearn by the deadline. Late submissions will be penalized according to the syllabus late policy. No submissions will be accepted after four days of the deadline.