

CSCS 460 – Machine Learning

Assignment 1 (Task 3)

Deadline: Thursday 9th March 2023 11:59 PM

Submission Guidelines:

- Submit all three tasks together as a single zip file on Moodle. All three tasks should have separate folders. Submit actual code files and not the snapshots.
- Only one member of the group has to submit on Moodle.
- Mention Roll numbers in the zip file name e.g., 23123123_2351551.zip
- A viva will be conducted during office hours after the submission.

Task 1 Submission Checklist

- Submit your labeled dataset in excel file.
- Do NOT submit your photos. Just labels along with images paths and three kinds of labels (i.e., you vs not you, expression, age) would suffice.
- Do NOT submit images numpy arrays either.
- Submit code to read, resize, and reshape images into vectors.

Task 2 Submission Checklist

- Submit code to read, clean, preprocess and vectorize the text.
- The code should print top 10 most important and bottom 10 (least important) feature names (i.e., words).
- Fix random seed of mutual_info_classif module to reproduce results on each run.
- Include the preprocessed data that is used to calculate feature importance.

Task 3

- 1) From task 1, you should have your images dataset with binary labels (you vs not you).
- 2) From task 2, you should have two version of the dataset
 - a) Train and test split without feature selection
 - b) Train and test with feature selection i.e., select k or p% most important features and remove useless features. (Select number of features as per your liking)
- 3) For all three datasets variations above, do:
 - a) Train decision tree with entropy as criterion
 - b) Output the accuracy of all three decision trees on respective test splits
 - c) Output Precision, Recall, and F1 score of all three decision trees on respective test splits

Task 3 Submission Checklist

- Submit code to train, test, and evaluate decision trees.
- Code should print the performance metrics mentioned above (accuracy, precision, recall, f1 score).

Some Helpful Libraries and Modules

- **Sklearn:** `SelectKBest`, `SelectPercentile`,
`mutual_info_classif`, `classification_report`, `accuracy_score`,
`cohen_kappa_score`, `confusion_matrix`, `f1_score`,
`precision_score`, `recall_score`
- **Numpy:** `flatten`, `reshape`
- **Cv2:** `resize`, `imread`
- **Pillow (PIL)**