| | |
|---|---|
| The BRITISH UNIVERSITY IN EGYPT<br>Faculty of Informatics and Computer Science | **22CSAI03H**<br>**Assignment 2**<br>**2022-2023** |
| Module Title **Machine Learning** | |
| Module Leader **Associate Pro. Nahla Barakat** | Semester<br>**One** |
| Assessment Weight<br>**25% of the total course mark** | Due Date<br>**09/12/2022** |

**Instructions to students:**

1. This is a group assignment; each group consists of 3-4 students.
2. Submission: The submission is via the e-learning system only

3. Assessment: Assessment will be based on the code submitted, the report, in addition to scheduled discussion with team members if needed.

4. Feedback: Feedback for each team will be given through discussions.

5. Along with the submitted assignment, you need to submit: a fully completed and signed *Coursework submission form* *and* a *Statement of Academic Honesty Form*. You can only submit your own work. Any student suspected of plagiarism will be subject to the procedures set out in the academic university regulations.

**Objectives:**

This assignment objective is to demonstrate the knowledge and skills required to build an end-to-end machine learning project; that helps solving or improving solutions for real life problem domain, and report the obtained results. The scope of this assignment is the, supervised, unsupervised and ensemble machine learning algorithms.

**Assignment resources**

A repository of different data sets from different domains will be provided, where you can choose your project data set(s) from. ***You can also find your own project data sets, however, you need to get the approval of the data set from one of the teaching team***.

***Python programming language should be used in all your implementations***.

***For teams of three students, select three data sets. For teams of four students, select four data sets***. One of your data sets should be imbalanced data set.

## Assignment Tasks:

---

1-Describe your data set, and explain why you think it is interesting: (Features' types, percentage of missing values, outliers; if any, dimensionality, target class(es) **[4 Marks]**

2- Select and utilize a clustering algorithm that you can use for data preparation. **[4 Marks]**

3- Use promising algorithms to build individual and ensemble models; (train 3 or 4 ***individual*** and ***ensemble models*** of those individual; from different categories, and compare results of individual classifiers and ***corresponding ensemble*** (e.g., ML methods to be used SVMs, Logistics Regression, Random Forest, etc.), at least 4 ensembles needed for a team of 4. **[15 Marks]**

4- Comment on its bias / variance outcome of your models. **[15 Marks]**

5- For the imbalanced data sets, and try 3 different methods to handle data imbalance and compare their results, using confusion matrix and ROC curves for cost-sensitive classification. **[12 Marks]**

6- Analyze and comment on the obtained results. Is Accuracy a reliable performance measure in all cases? Elaborate on your answer **[20 Marks]**

7- Document and report the tasks 6 as appropriate (1500-2500) words. **[20 Marks]**

*There will be a presentation for each team* **[10 Marks]**

**[Total Mark 100]**

**Data Sets:**

1. https://www.openml.org/search?type=data&sort=runs&status=active&id=1504
2. https://www.openml.org/search?type=data&sort=runs&status=active&id=31
3. https://www.openml.org/search?type=data&sort=runs&status=active&id=3
4. https://www.openml.org/search?type=data&sort=runs&status=active&id=1494
5. https://www.openml.org/search?type=data&sort=runs&status=active&id=1510
6. https://www.openml.org/search?type=data&sort=runs&status=active&id=1487
7. https://www.openml.org/search?type=data&sort=runs&status=active&id=1479
8. https://www.openml.org/search?type=data&sort=runs&status=active&id=1063
9. https://www.openml.org/search?type=data&sort=runs&status=active&id=1471
10. https://www.openml.org/search?type=data&sort=runs&status=active&id=1467
11. https://www.openml.org/search?type=data&sort=runs&status=active&id=44
12. https://www.openml.org/search?type=data&sort=runs&status=active&id=1067
13. https://www.openml.org/search?type=data&sort=runs&status=active&id=1461
14. https://www.openml.org/search?type=data&sort=runs&status=active&id=1220
15. https://www.openml.org/search?type=data&sort=runs&status=active&id=4534
16. https://www.openml.org/search?type=data&sort=runs&status=active&id=16
17. https://www.openml.org/search?type=data&sort=runs&status=active&id=32
18. https://www.openml.org/search?type=data&sort=runs&status=active&id=458
19. https://www.openml.org/search?type=data&sort=runs&status=active&id=188
20. https://www.openml.org/search?type=data&sort=runs&status=active&id=1497
21. https://www.openml.org/search?type=data&sort=runs&status=active&id=1466
22. https://www.openml.org/search?type=data&status=active&id=5