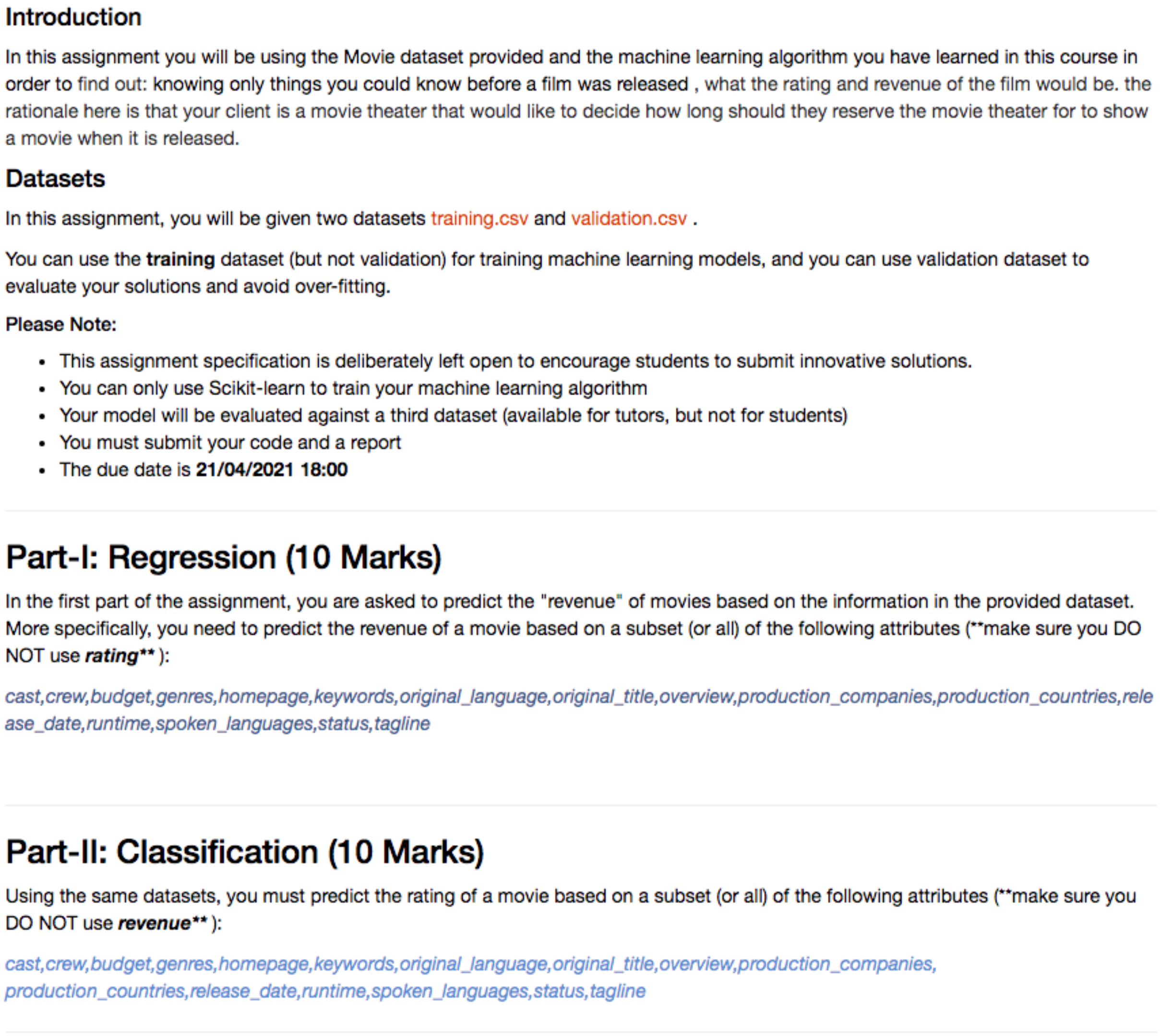
INSTRUCTIONS



Introduction

In this assignment you will be using the Movie dataset provided and the machine learning algorithm you have learned in this course in order to find out: knowing only things you could know before a film was released , what the rating and revenue of the film would be. the rationale here is that your client is a movie theater that would like to decide how long should they reserve the movie theater for to show a movie when it is released.

Datasets

In this assignment, you will be given two datasets training.csv and validation.csv .

You can use the **training** dataset (but not validation) for training machine learning models, and you can use validation dataset to evaluate your solutions and avoid over-fitting.

Please Note:

- This assignment specification is deliberately left open to encourage students to submit innovative solutions.
- You can only use Scikit-learn to train your machine learning algorithm
- Your model will be evaluated against a third dataset (available for tutors, but not for students)
- You must submit your code and a report
- The due date is **21/04/2021 18:00**

Part-I: Regression (10 Marks)

In the first part of the assignment, you are asked to predict the "revenue" of movies based on the information in the provided dataset. More specifically, you need to predict the revenue of a movie based on a subset (or all) of the following attributes (**make sure you DO NOT use ***rating****):

cast,crew,budget,genres,homepage,keywords,original_language,original_title,overview,production_companies,production_countries,release_date,runtime,spoken_languages,status,tagline

Part-II: Classification (10 Marks)

Using the same datasets, you must predict the rating of a movie based on a subset (or all) of the following attributes (**make sure you DO NOT use ***revenue****):

cast,crew,budget,genres,homepage,keywords,original_language,original_title,overview,production_companies, production_countries,release_date,runtime,spoken_languages,status,tagline

Submission

You must submit two files:

- A python script z{id}.py
- A report named z{id}.pdf

Python Script and Expected Output files

Your code must be executed in CSE machines using the following command with three arguments:

```
$ python3 z{id}.py path1 path2
```

- **path1** : indicates the path for the dataset which should be used for training the model (e.g., ~/training.csv)
- **path2** : indicates the path for the dataset which should be used for reporting the performance of the trained model (e.g., ~/validation.csv); we may use different datasets for evaluation

For example, the following command will train your models for the first part of the assignment and use the validation dataset to report the performance:

```
$ python3 YOUR_ZID.py training.csv validation.csv
```

Your program should create 4 files on the same directory as the script:

- z{id}.PART1.summary.csv
- z{id}.PART1.output.csv
- z{id}.PART2.summary.csv
- z{id}.PART2.output.csv

For the first part of the assignment:

" z{id}.PART1.summary.csv " contains the evaluation metrics (MSE, correlation) for the model trained in the first part of the assignment. Use the given validation dataset to compute the metrics. The file should be formatted exactly as follow:

```
zid,MSE,correlation
YOUR_ZID,6.13,0.73
```

- **MSE** : the mean_squared_error in the regression problem
- **correlation** : The **Pearson correlation coefficient** in the regression problem (a floating number between -1 and 1)

" z{id}.PART1.output.csv " stores the predicted revenues for all of the movies in the evaluation dataset (not the training dataset), and the file should be formatted exactly as:

```
movie_id,predicted_revenue
1,7655555
2,75875765
...
```

For the second part of the assignment:

" z{id}.PART2.summary.csv " contains the evaluation metrics (average_precision, average_recall, accuracy - the unweighted mean) for the model trained in the second part of the assignment. Use the given validation dataset to compute the metrics. The file should be formatted exactly as:

```
zid,average_precision,average_recall,accuracy
YOUR_ZID,0.69.71,0.89
```

- **average_precision** : the average precision for all classes in the classification problem (a number between 0 and 1)
- **average_recall** : the average recall for all classes in the classification problem (a number between 0 and 1)

" z{id}.PART2.output.csv " stores the predicted ratings for all of the movies in the evaluation dataset (not the training dataset) and it should be formatted exactly as follow:

```
movie_id,predicted_rating
1,1
2,4
...
```