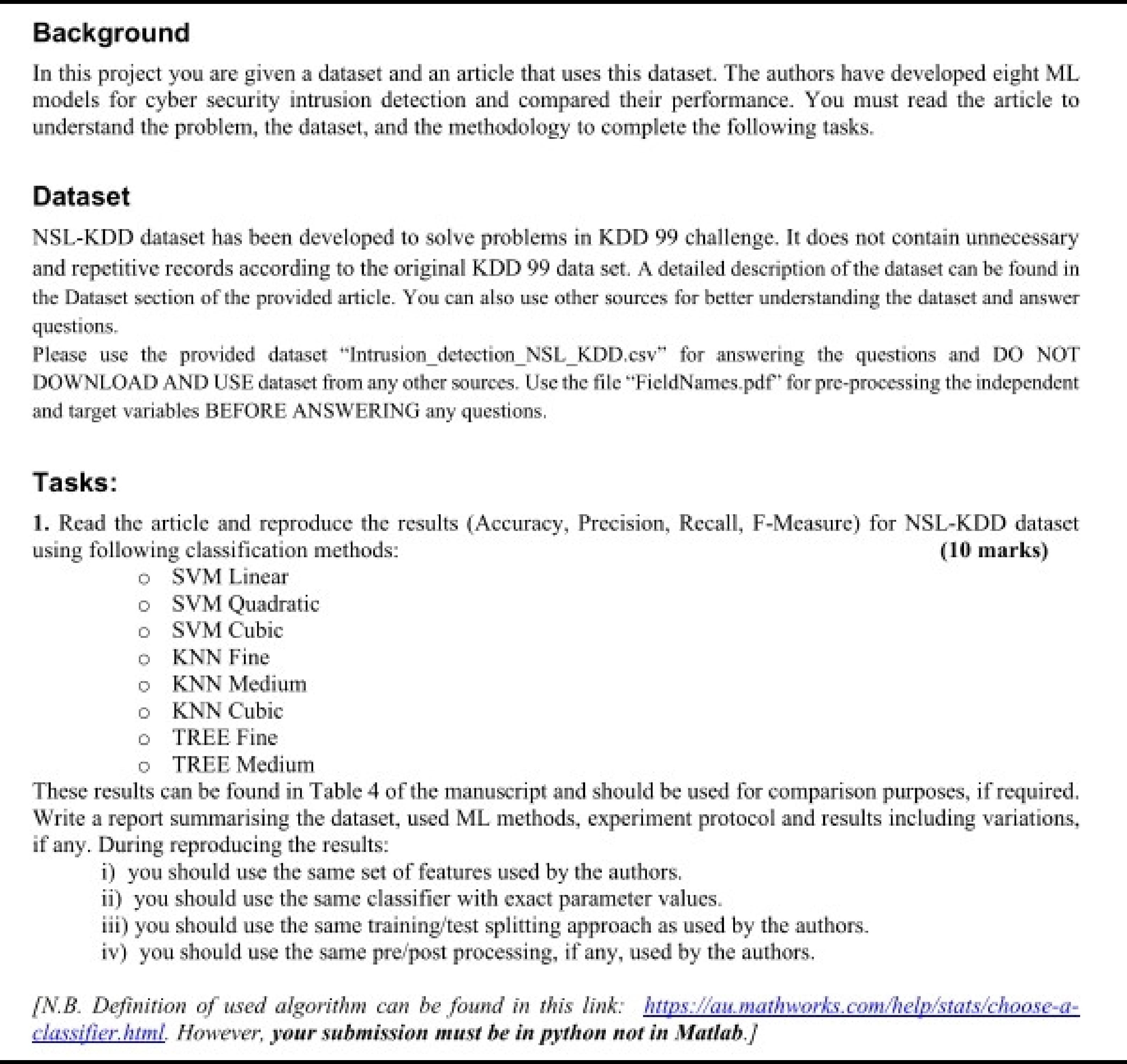

Background

In this project you are given a dataset and an article that uses this dataset. The authors have developed eight ML models for cyber security intrusion detection and compared their performance. You must read the article to understand the problem, the dataset, and the methodology to complete the following tasks.

Dataset

NSL-KDD dataset has been developed to solve problems in KDD 99 challenge. It does not contain unnecessary and repetitive records according to the original KDD 99 data set. A detailed description of the dataset can be found in the Dataset section of the provided article. You can also use other sources for better understanding the dataset and answer questions.

Please use the provided dataset "Intrusion_detection_NSL_KDD.csv" for answering the questions and DO NOT DOWNLOAD AND USE dataset from any other sources. Use the file "FieldNames.pdf" for pre-processing the independent and target variables BEFORE ANSWERING any questions.

Tasks:

1. Read the article and reproduce the results (Accuracy, Precision, Recall, F-Measure) for NSL-KDD dataset using following classification methods: **(10 marks)**

- SVM Linear
- SVM Quadratic
- SVM Cubic
- KNN Fine
- KNN Medium
- KNN Cubic
- TREE Fine
- TREE Medium

These results can be found in Table 4 of the manuscript and should be used for comparison purposes, if required. Write a report summarising the dataset, used ML methods, experiment protocol and results including variations, if any. During reproducing the results:

i) you should use the same set of features used by the authors.
ii) you should use the same classifier with exact parameter values.
iii) you should use the same training/test splitting approach as used by the authors.
iv) you should use the same pre/post processing, if any, used by the authors.

*[N.B. Definition of used algorithm can be found in this link: https://au.mathworks.com/help/stats/choose-a-classifier.html. However, **your submission must be in python not in Matlab**.]*

(iii) Similarly, variation in results due to randomness of data splitting will also be considered during evaluation based on your explanation.
(iii) Obtained marks will be proportional to the number of ML methods that you will report in your submission with correctly reproduced results.
(iv) Make sure your submitted Python code segment generates the reported results, otherwise you will receive zero marks for this task.

Marking criteria:

- *i) Unsatisfactory (x<4): tried to implement the methods but unable to follow the approach presented in the article. Variation of marks in this group will depend on the quality of report.*
- *ii) Fair (4<=x<5): appropriately implemented 50% of the methods presented in the article. Variation of marks in this group will depend on the quality of report.*
- *iii) Good (5<=x<7): appropriately implemented 70% of the methods presented in the article. Variation of marks in this group will depend on the quality of report.*
- *iv) Excellent(x>=7): appropriately implemented >=90% of the methods presented in the article. Variation of marks in this group will depend on the quality of report.*

2. Design and develop your own ML solution for this problem. The proposed solution should be different from all approaches mentioned in the provided article. This does not mean that you must have to choose a new ML algorithm. You can develop a novel solution by changing the feature selection approach or parameter optimisations process of used ML methods or using different ML methods or different combinations of them. This means, the proposed system should be substantially different from the methods presented in the article but not limited to only change of ML methods. Compare the result with reported methods in the article. Write a technical report summarising your solution design and outcomes. The report should include: **(20 marks)**

i) Motivation behind the proposed solution.
ii) How the proposed solution is different from existing ones.
iii) Detail description of the model including all parameters so that any reader can implement your model.
iv) Description of experimental protocol.
v) Evaluation metrics.
vi) Present results using tables and graphs.
vii) Compare and discuss results with respect to existing literatures.
viii) Appropriate references (IEEE numbered).

N.B. This is a HD (High Distinction) level question. Those students who target HD grade should answer this question (including answering all the above questions). For others, this question is an option. This question aims to demonstrate your expertise in the subject area and the ability to do your own research in the related area.

Marking criteria:

Quality of solution	***Quality of report***	***Overall score***
Unsatisfactory	Unsatisfactory	Unsatisfactory; Score<5
Unsatisfactory	Fair	Unsatisfactory; Score<7
Unsatisfactory	Good	Unsatisfactory; Score<10
Fair	Unsatisfactory	Unsatisfactory; Score<10
Fair	Fair	Fair; Score<12
Fair	Good	Fair; Score<14
Good	Unsatisfactory	Fair; Score <14