Term Project Template

Group#
Names:
IDs

General guidelines

- This slide will provide you general guidelines and structure of your presentations for the term project (please remove slides including instruction and information):
  - The answers of the term project will be presented in a presentation format (not in a report), which will be provided in this template.
  - DO NOT copy and paste any materials.
  - While 85 marks is dedicated to presentation style and results, remaining 15 marks is dedicated to actual presentation and the answer of the question.
  - Presentation time will be 8 minutes and Q&A session will be 2 minutes. You are supposed to limit your time usage according to the allowed time period. Otherwise marks reduction will be applied.
  - DO NOT include your code in this presentation. If you need to explain any piece of your code, then you shall use a flowchart instead.
  - You are allowed to use different color for the presentation's template, however, the contrast of the colors should be considered.
  - **All figures, tables and results should have titles, explanations, axis names and legends. Otherwise, marks reduction will be applied.**
- **If you wish, you can find your own dataset with minimum 10 features. But this dataset must be research oriented not well known any dataset. Please send me an email to explain your problem and dataset before working on your problem because you need my approval to start working on it.**
-
- **Dataset collected 6 days long. To use the dataset given below, the first 5 days data should be used as training set and remaining one day data should be used as test data. After this selection, feature "day" should be eliminated from the both training and test set.**

Dataset

Mobile CrowdSensing Dataset is used in this project. You can find more information in the link below:

http://nextconlab.academy/MCSData/MCS-FakeTaskDetection.html

Dataset is generated by CrowdSenSim simulation tool. Dataset contains legitimate tasks and fake tasks. The task attributed are as follows: {'ID', 'latitude', 'longitude', 'day', 'hour', 'minute', 'duration', 'remaining time', 'battery requirement %', 'Coverage', 'legitimacy', 'GridNumber', 'OnpeakHour'}. Location of tasks are specified by 'latitude' and 'longitude' together. Furthermore, 'day', 'hour' and 'minute' describe the task publish time. 'Duration' denotes task active duration in terms of minutes. 'Remaining time' denotes the residual time of a sensing task till its completion. 'Battery requirement' is percentage of battery required to complete a task. 'Coverage' denotes task sensing distance. 'Legitimacy' describes whether a task is illegitimate one or legitimate one. This feature is used only in training of the machine learning models as the MCS platform is unaware of task legitimacy when a task is submitted. 'GridNumber' is obtained by splitting sensing city map to small grids with numbers beginning at 1. 'OnpeakHour' is a binary flag to indicate if task start time occurs during 7am to 11am. We define 7am to 11am as the peak hour and other hours are non-peak for the sake of simplicity in simulations. Based on the configuration of task generation in Table 1, the dataset is created including total 14,484 tasks, with 12,587 legitimate tasks and 1,897 fake tasks, respectively.

Features	Fake Tasks	Legitimate Tasks
Day	Uniformaly distributedly in [1, 6]	Uniformaly distributedly in [1, 6]
Hour	80%: 7am to 11am; 20%: 12pm to 5 pm	8%: 0am to 5am; 92%: 6pm to 23pm
Duration (min)	70% in {40, 50, 60}; 30% in {10, 20, 30}	Uniformly distributed over {10, 20, 30, 40, 50, 60}
Battery usage	80% in {7%-10%}; 20% in {1%-6%}	Uniformly distributed in {1%-10%}
Recruitment Radius	Uniformly distribute in 30m to 100m	Uniformly distribute in 30m to 100m
Movement Radius	[10m, 80m]	[10m, 80m]
Number of Tasks	1,897	12,578

Slide 1: Problem's overview (5 marks)

Provide a conceptual figure to explain the problem in hand

- The figure should show an end-to-end dataflow, and provide insights on the problem
- You can write a few sentences to further explain the problem, if needed
- Copying and pasting the figure from online resources is not allowed. You are supposed to add more intelligence onto internet based materials.

Slide 2: Dataset’s overview (EDA) (5 marks)

- Please introduce your dataset giving some numerical information about it. Input-output relationship should be given through machine learning model (rectangular box). Features and outputs should be seen on the figure.

Slide 3: General Flowchart to summarize all process (10 marks)

- Provide an end-to-end flowchart, where you show every step in the process of the project's implementation.
  - The flowchart should be clear and the font's size is visible
  - You can use color code for different part of your methodology.
  - Flowchart consists of process of each part of your solution.

Slide 4: Visualize the training and test set to understand problem nature (5 marks)

- TSNE plot should be given for training and test set.

Q1) Obtain a baseline performance (15 marks)

Apply all ML methods below on the provided dataset to obtain baseline performance.

Plot confusion matrix and calculate the accuracy for each methods, and plot them in a bar-chart as baseline.

- ❏ KNN
- ❏ LogisticRegression
- ❏ SVM
- ❏ DecisionTreeClassifier
- ❏ Naive Bayes Classifier

The best baseline performance will be used as the first baseline result for remaining analysis.

Q2) First Improvement strategy : Comparing dimensionality reduction to feature selection (20 marks)

Q2.1) Dimensionality reduction

Use two dimensionality techniques to find the best dimension (the number of features) to increase the performance comparing with baseline performance seen in the previous slide. (8 marks)

- Use PCA analysis
- Use Autoencoder
- Compare two methods with baseline performance on the same figure (x-axis indicates the number of feature and y axis indicates accuracy. Baseline result should be given constant dotted line because it is not changing with the number of features.)

Q2.2) Feature selection (8 marks)

- Select 2 types of feature selection techniques amongs wrapper, filter and embedded types methods.
- Compare the 2 feature selection accuracy performances with the best performance in Q2.1

Update your dataset, **to be used in the next steps**, based on the technique that provides you with highest test accuracy (either dimensionality reduction or feature selection), **and provide the confusion matrix** and tsne plot of the highest test accuracy.
(4 marks)

Q3) Adding more machine learning model (10 marks)

- Random Forest
- Ensemble techniques
- Any other techniques
  Compare the performance of new techniques with the first improvement through confusion matrix. If the new results are better than the first improvement, the new results will be assumed as the second improvement. Otherwise, the first improvement should be kept for the remaining analysis.

Q4) Applying parameter fine tuning to get better performance from the previous best performance. (10 marks)

You are supposed to try this strategy onto two parameters. You should try at least 3 points such as low than default value and high than default value. The previous best accuracy will be constant while x axis indicates parameters values and y axis indicates the accuracy. If you get maximum performance from the parameter tuning better than the second improvement, this new result will be the third improvement. Plot each parameter performance separately.

Q5) Writing the conclusion (5 marks)

You are supposed to summarize all results as conclusion section. Considering all results from your analysis, please list your concluded comments here.