# 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 90 marks is dedicated to presentation style and results, remaining 10 marks is dedicated to actual presentation. 10 marks will be given for not only high quality visualized results but also well prepared and organized slides.
  - Presentation time will be 10 minutes. You need to record your presentation and submit together with your presentation and codes. 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.

# Dataset

## ➢Forest Cover Type Prediction Dataset:

Reference link: https://www.kaggle.com/competitions/forest-cover-type-prediction/data

The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest cover type. The seven types are:

1 - Spruce/Fir
2 - Lodgepole Pine
3 - Ponderosa Pine
4 - Cottonwood/Willow
5 - Aspen
6 - Douglas-fir
7 – Krummholz

Important: Use our provided dataset instead of downloading from the reference link.

# Dataset

- Features:
  - **Elevation** : Elevation in meters
  - **Aspect** : Aspect in degrees azimuth
  - **Slope** : Slope in degrees
  - **Horizontal_Distance_To_Hydrology** : Horz Dist to nearest surface water features
  - **Vertical_Distance_To_Hydrology** : Vert Dist to nearest surface water features
  - **Horizontal_Distance_To_Roadways** : Horz Dist to nearest roadway
  - **Hillshade_9am** (0 to 255 index) : Hillshade index at 9am, summer solstice
  - **Hillshade_Noon** (0 to 255 index) : Hillshade index at noon, summer solstice
  - **Hillshade_3pm** (0 to 255 index) : Hillshade index at 3pm, summer solstice
  - **Horizontal_Distance_To_Fire_Points** : Horz Dist to nearest wildfire ignition points
  - **Wilderness_Area** (4 binary columns, 0 = absence or 1 = presence) : Wilderness area designation
  - **Soil_Type** (40 binary columns, 0 = absence or 1 = presence) : Soil Type designation
  - **Cover_Type** (7 types, integers 1 to 7) : Forest Cover Type designation

# 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.
  - Each improvement strategy should be emhasized in the flowchart.

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

- ❏ TSNE plot should be given for training and test set to understand complexity of the problem.

# 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. In the next phase. the best 2 ML models should be used as performance calculations.

# Q2) First Improvement strategy : Comparing dimensionality reduction to feature selection (10 marks)

**Feature selection (10 marks)**

- Apply feature selection methods.
- One filter based approach and one wrapper based approch.
- For wrapper based approach, you need to provide training set, test set, number of features and ML model for training and test performances. You need to select **2 best performer ML models from Q1 stage** for wrapper and filter based feature selection methods.
- Performance comparisons will be the number of features vs the accuracy and also you need to add baseline performance of ML model you obtained by Q1 stage as constant dotted line with different color.
- You need to provide 4 figures and 2 for each feature selection method. 2 figures for each method are required because you need to show ML performance is better than baseline performance of the ML models.
- **After determining the best performance in this stage that will be first improvement, you need to use the best feature subset and the best ML model for remaining part of the project.**

# Q3) Adding more machine learning model (10 marks)

- Apply Random Forest and 2 ensembles techniques to get better performance.
- 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) Supervised & Unsupervised Combination via PKI (15 marks)

- Using the knowledge in the appendix, please find the best number of cluster to improve the supervised model. You use previous best supervised performance in stage Q3 and you need to show if SOFM provides improvement through PKI Strategy. You need to use 6*6, 7*7, 8*8, 9*9, 10*10, 11*11 and 12*12 and PKI model as given below. Please show the best SOFM structure using number of neurons vs accuracy figure with given the second improvement as red dotten line. If you get better result than the second one please use this accuracy as **the third improvement** in the next phase.

- PKI model will be realized via Deep Neural Network (DNN) which has 4 hidden layers having 30 neurons of each. Use Adam optimizer, set learning rate as 0.001 and choose tanh as the activation function. **You don’t have to use this DNN structure. You can find your best PKI model to provide better result than the stage Q3.**

## Q5) Applying parameter fine tuning to get better performance from the previous best performance. (10 marks)

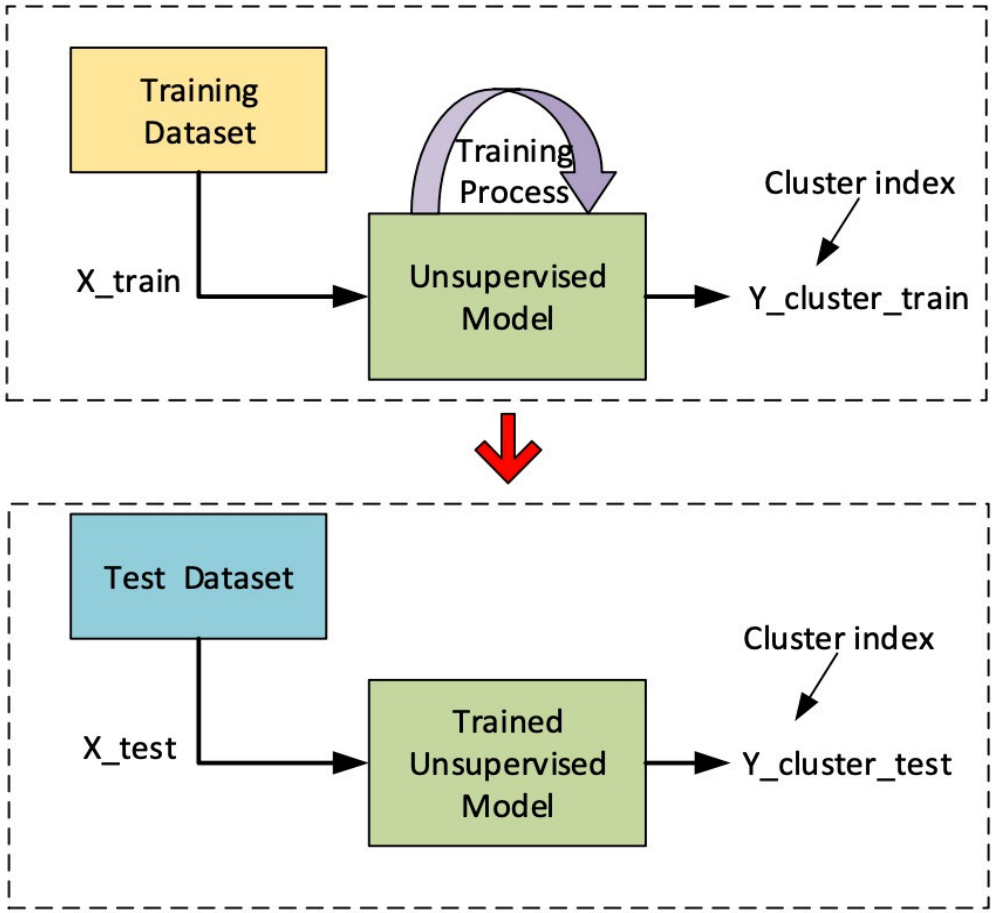
- This is final stage for obtaining the improvement. You need to try different structures for Deep Neural Network for PKI structure. Please try **more or less** hidden layers and neurons of each hidden layers which you find in the stage Q4 to show the best performance and the improvement. Please plot each tunning performance of hidden layers and neurons of each hidden layers with given the best accuracy so far as red dotted line.

## 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. We are expecting minimum one sentence for each improvement strategy. Please share your experience and knowledge about that.

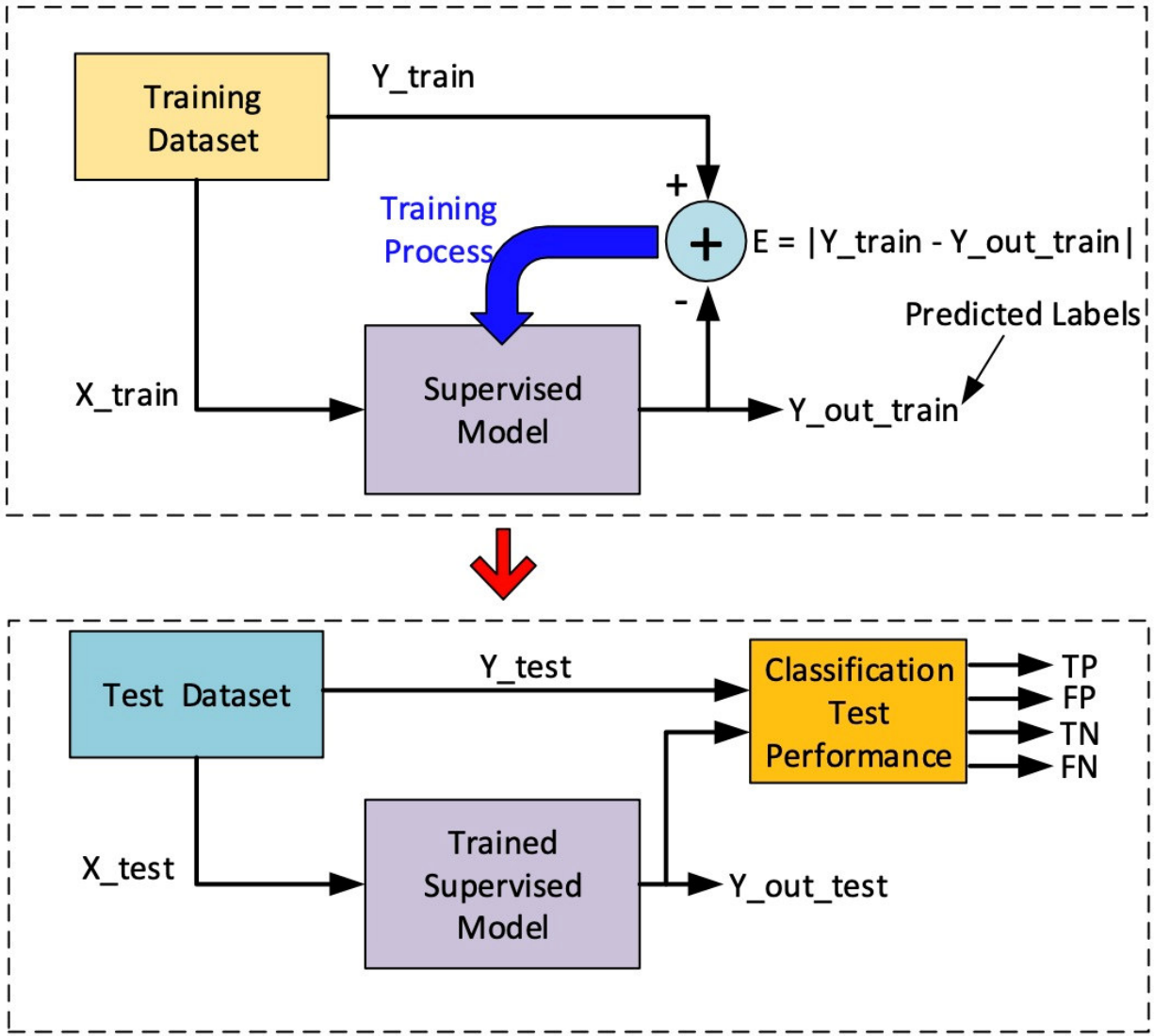
# Appendix:

- Unsupervised Learning



# Appendix:

- Supervised Learning



# Appendix:

- PKI Learning