

CSE422 Lab Project Report Template

- ★ Cover page
- ★ Table of contents
- ★ Page no.

1. Introduction

A small introduction on what the project aims to do, what problem it's aiming to solve, the motivation behind the project.

2. Dataset description

- Dataset Description
 - How many features?
 - Classification or regression problem? Why do you think so?
 - How many data points?
 - What kind of features are in your dataset? (Quantitative / Categorical)
 - Do you need to encode the categorical variables, why or why not?
 - Correlation of all the features (input and output features) (apply heatmap using the seaborn library)
 - What do you understand after the correlation test?
- Imbalanced Dataset
 - For the output feature, do all unique classes have an equal number of instances or not?
 - Represent using a bar chart of N classes (N=number of classes you have in your dataset).
- Perform exploratory data analysis to extract some important relationships from your data. [Reference: ■ EDA Lab CSE422]

3. Dataset pre-processing

- Faults
 - Null / Missing values
 - Categorical values
 - Feature Scaling
- Solutions
 - Delete rows/columns, Impute values [show cause]
 - Encoding(as required) [show cause]
 - Scaling as per requirement

Note: Firstly, discuss one problem, and then write about the solutions or pre-processing techniques you have applied to solve that problem. Afterward, proceed to the next problem.

4. Dataset splitting

- Random/Stratified (as required)
- Train set (80% / 70%) (Use Validation Set as required)
- Test set (20% / 30%)

5. Model training & testing (Supervised)

- [KNN](#) (for classification problem)
- [Decision Tree](#) (for classification/regression problem)
- [Logistic Regression](#) (for classification problem)
- [Linear Regression](#) (for regression problem)
- [Naive Bayes](#) (for classification problem)
- [Neural Network](#) (for classification/regression problem) [This can be applied using any library you feel comfortable with - sklearn, tensorflow, pytorch etc]

****** Treat the problem as an unsupervised learning problem, apply Kmeans and showcase the clusters******

Remember you have to apply a Neural Network and at least 2 other models

6. Model selection/Comparison analysis

- Bar chart showcasing prediction accuracy of all models (for classification)
- Precision, recall comparison of each model. (for classification)
- Confusion Matrix (for classification)
- AUC score, ROC curve for each model (for classification)
- R^2 score and Loss (for regression)

Compare the results of all models based on all of the above described metrics

7. Conclusion

- What do you understand from the results
- Make useful comments regarding the performance of your model
- Why do you think you are getting such results
- What are some of the challenges that you have faced