**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](https://drive.google.com/drive/folders/14Kb3UsdEsw2IYvjYq8FVkm7i_r0QDXUk?usp=sharing)]

1. **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.

1. **Dataset splitting**

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

1. **Model training & testing (Supervised)**

* [KNN](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html) (for classification problem)
* [Decision Tree](https://scikit-learn.org/stable/modules/tree.html) (for classification/regression problem)
* [Logistic Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) (for classification problem)
* [Linear Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) (for regression problem)
* [Naive Bayes](https://scikit-learn.org/stable/modules/naive_bayes.html) (for classification problem)
* [Neural Network](https://colab.research.google.com/drive/1lNnalGV-4_o6TNKwTTE5Yet9IO6SkH9T?authuser=1#scrollTo=GWqNBiMN94Gk) (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**

1. **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)
* R2 score and Loss (for regression)

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

1. **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