Project – CSN344, Machine Learning, ODD Semester, III Year

Student Name: **ID:** **Major:**

Instructions

- This project is given to exercise your brains w.r.t a ML-based data pipeline from starting to end; students are urged to choose a relevant problem statement and supportive data set.
- A complete ML pipeline supported by data centric exercises will be appreciated.
- Optional: I am also accepting any hackathon/challenge-based research problem statement, condition applied you are able to defend the project and it is not associated with any other firm/professor and is independent in nature
- References and citations should be added and any AI-based dependency/tools/GPTs should be declared (use Zotero/Mendeley); with minimum possible plagiarism and AI similarity index
- Edit the entire project in this document and add supportive documents (images/videos/files/links etc.) and save the updated word file; try to make a supportive 5-slide ppt and upload a PDF **version** of this document on or before November 30, 2025

Due Date: **November 30, 2025**

Submitting this Project: It is expected to submit (upload) this project on the LMS. Rename the final PDF document as SAPID_FULLNAME.pdf

Project Title:

End-to-End Machine Learning Pipeline Development (**feel free to re-name your project**)

Project Overview:

This project aims to help students design, implement, and document a complete machine learning pipeline starting from dataset selection to model evaluation and reporting. Students are expected to choose a relevant, real-world problem statement and an appropriate dataset that enables them to demonstrate all core ML components such as data ingestion, preprocessing, exploratory data analysis (EDA), model development, performance evaluation, and result interpretation. A thoroughly documented data-centric ML workflow supported by code, figures, and insights will be valued.

Task Requirements:

1. Dataset selection & handling:

A. Select any dataset containing at least 200 samples and 5 features.

Project – CSN344, Machine Learning, ODD Semester, III Year

Student Name: **ID:** **Major:**

B. Clearly describe the problem statement and justify the choice of dataset.
C. Import the dataset and summarize it with number of rows and columns, data types, basic structure and shape etc.

2. **Data pre-processing**:

A. Perform all preprocessing steps required for your ML pipeline, including handling missing values and detecting and addressing outliers
B. Performing transformations such as: Scaling (standardization/normalization/robust scaling), Encoding (one-hot encoding, label encoding, etc.; optional: feature engineering or domain-derived attributes

3. Exploratory Data Analysis (EDA):

Conduct a data-centric EDA with both numerical and visual analysis as follows:

A. Numeric analysis via descriptive statistics computation: mean, variance, standard deviation skewness, kurtosis and percentiles
B. Visual analysis by including visualizations such as histogram, boxplot, scatter plot, correlation and/or heatmap; try to discuss patterns, distributions, correlations, and any anomalies observed.

4. Model Selection & Training:

Choose any two model(s) [you can also choose any model(s) of your own choice] from the following, depending on your problem type and compare the methods:

1. Regression Models
2. Linear Regression
3. Polynomial Regression
4. Classification Models
5. Logistic Regression
6. Decision Tree
7. k-Nearest Neighbors (kNN)
8. Support Vector Machine (SVM)
9. Clustering Models

Student Name: **ID:** **Major:**

10. K-means
11. Hierarchical Clustering
12. Dimensionality Reduction
13. PCA
14. ICA

Train the selected model(s) using appropriate preprocessing, splitting strategy (train/validation/test), and model-fitting workflow. (Optional but encouraged: Hyperparameter tuning, pipelines, cross-validation).

5. Evaluation:

Report the correct evaluation metrics based on the chosen task type, any of the following:

1. Regression
2. MSE (Mean Squared Error)
3. RMSE (Root Mean Squared Error)
4. R^2 Score
5. Accuracy
6. Precision
7. Recall
8. F1-score
9. Cluster visualization (2D projection recommended)
10. Dimensionality Reduction
11. Explained variance ratio
12. 2D visualization of transformed components

That’s all, all the best.