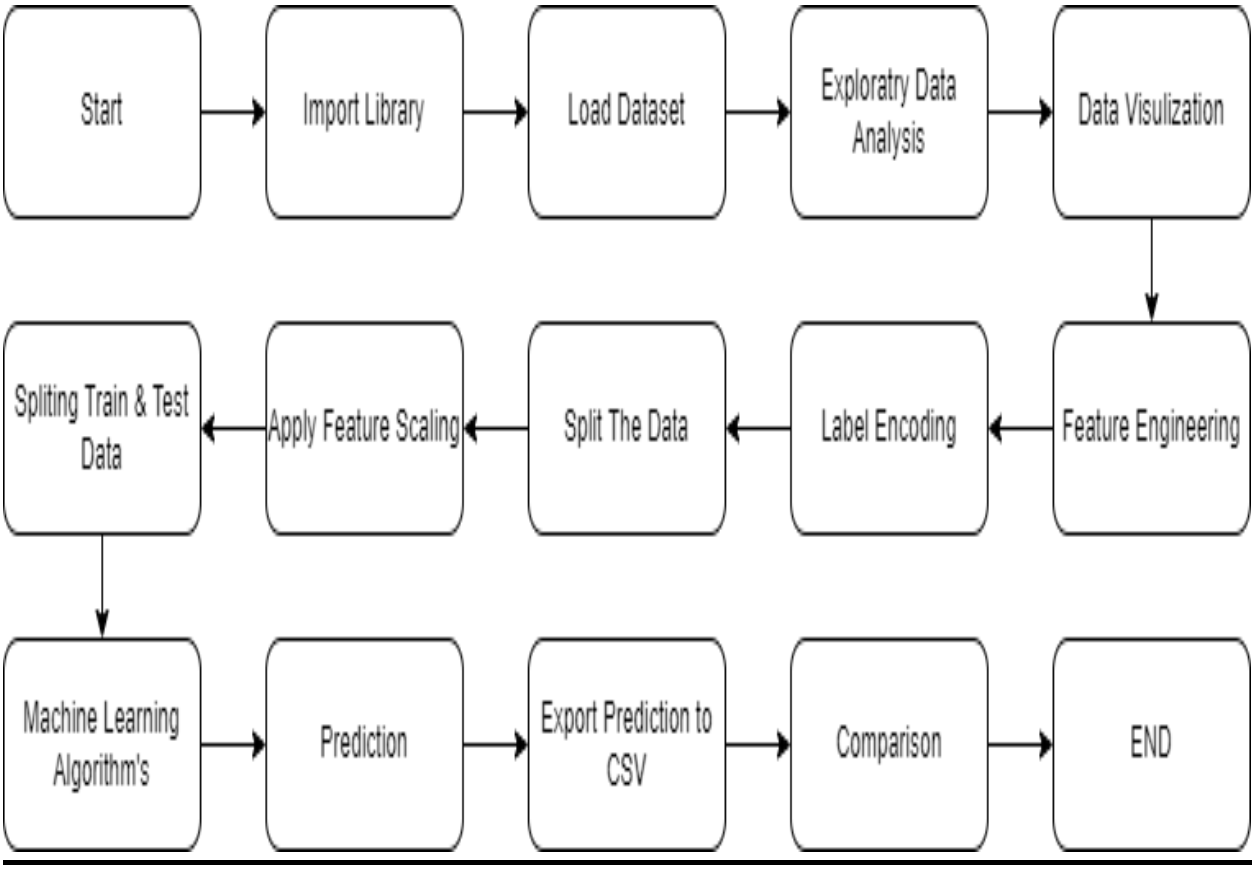
# Architecture

## Adult Census Income Prediction

# Problem Statement

To build a classification methodology to determine whether Adult Census Income Prediction.

# Architecture




# Data Description

I have collected Adult Census Income Prediction the dataset Kaggle.com. The Data has been extracted from the census bureau.

The data contains the following attributes:

**Features:**

1. Age: Continuous. It denotes the age person.
2. Workclass: What ta people do.
3. Fnlwgt: The nation is meant.
4. Education : Information about studies has been provided.
5. Education-num: Information about studies has been provided.
6. Marital-status: Information about marital has been provided.
7. Occupation: It has been said who is working in which profession.
8. Relationship: Information about relationship has been provided.
9. Race: The nation is meant.
10. Sex: Information about gander has been provided.
11. Capital-gain: It denotes the monitory gains by the person.
12. Capital-loss : It denotes the monitory loss by the person.
13. Hours-per-week: Working as many hours a week.
14. Country: People are citizens of any country.

# Target Label:

Salary 50k+ Plus or not.

# Data Validation:

In this step, I perform different sets of validation on the given set of training files.

1. Name Validation- I validate the name of the files based on the given name in the schema file. I have created a regex pattern as per the name given in the schema file to use for validation. After validating the pattern in the

name, we check for the length of data in the file name as well as the length of time in the file name. If all the values are as per requirement, we move such file to 'Good_Data_Folder' else I move sucj files to 'Bad_Data_Folder'.

2. Number of Columns- I validate the number of columns present in the files, and if it doesn't match with the value given in the schema file, then file is moved to 'Bad_Data_Folder.'

# Model Training

1)Data Export from DB – The data in a stored database is exported as a CSV file to be used for model training.

2)Data Preprocessing –

a)The dataset, the null values were replaced with.

b) Check for null values in the columns. If present, impute the null values using the categorical imputer.

c) Scale the numeric values using the standard scaler.

3) Model Selection – After the clusters have been created, we find the best model for each cluster. I am using seven algorithm's KNN, Navie-Bayes, Logistic Regression, Decision tree, Random Forest, SVM and Gradient Boosting Algorithm's. All the algorithm's are passed with the best parameters derived from GridSearch. All the models for every cluster are saved for use in prediction.