# Adult Census Income Prediction

## Objective:

Development of a predictive Adult Census Income Prediction for country wised people income rate. The model will determine whether 50k + Plus salary or not.

## Benefits:

Detection of the higher salary

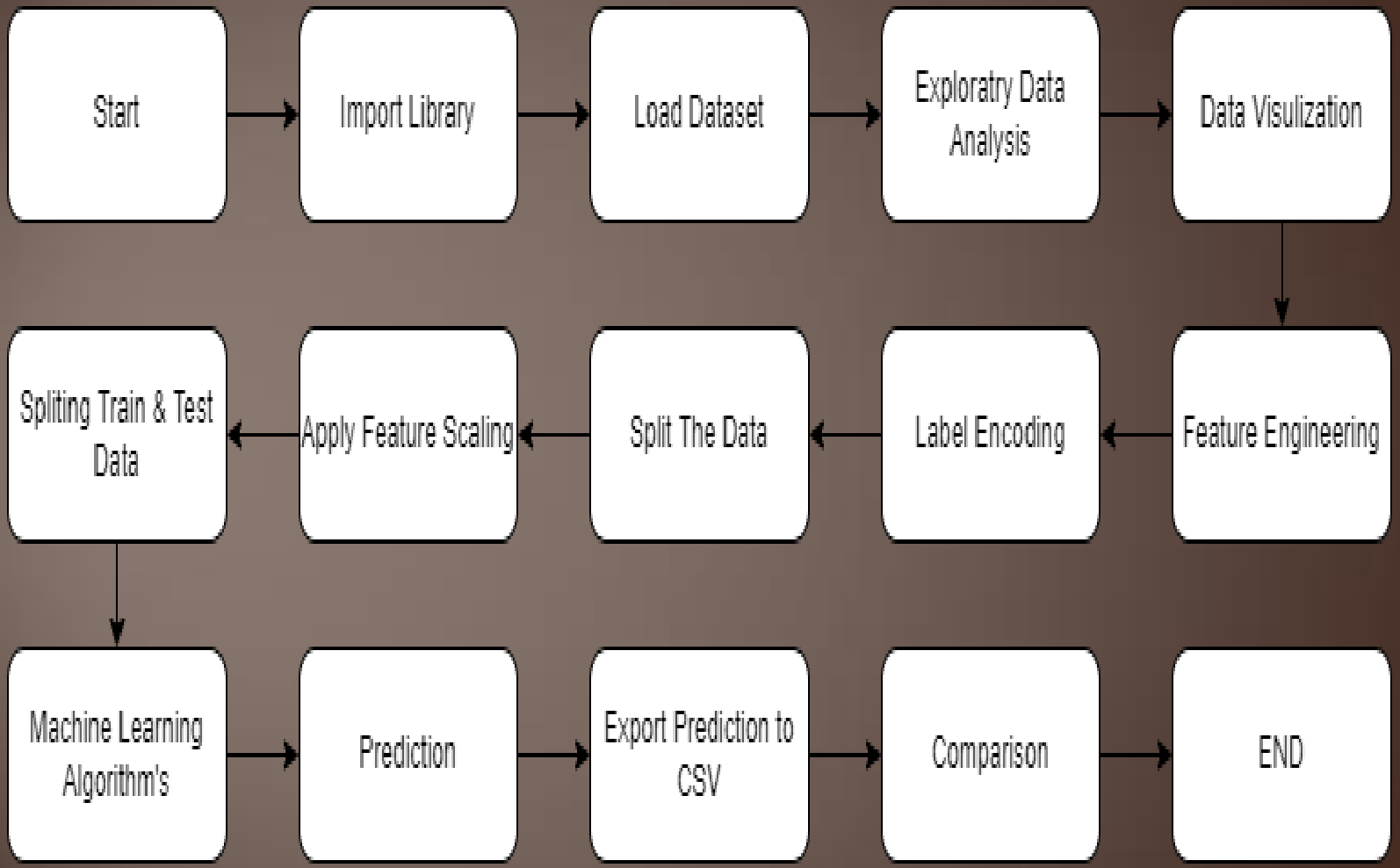
The reason for the higher salary

Will be salary if anyone join a new job

# Data Sharing Agreement:

1. Sample file name (adult.csv)
2. Number of Columns
3. Number of rows
4. Columns names
5. Columns data type

# Architecture



## Data Validation and Data Transformation:

**Name Validation** – Validation of file name as per the adult. I have create a regex pattern for validation. After it check for data format .

**Number of Columns** – Validation of number of columns present in the file.

**Name of Columns** – The name of the columns is validated and should be the same as given in the schema file.

**Data type Columns** – The data type of columns is given in the schema file. It is validated when I insert the files into database.

**Null Values in Columns** – If any of the columns in a file have all the values as NULL of missing, I discard such a file and move it .

## Model Training:

**Data Export from DB:**

The accumulated data from db is export in csv format for model training

**Data Preprocessing:**

1. Check for null values in the columns, if present impute the null values.
2. Encode the categorical values with numerical values.
3. Perform Standard Scalar to scale down the values.

## Model selection –

After the clusters are created, I find the best model for each cluster. By using 7 algorithm's KNN, Naïve Bayes, Logistic Regression, Decision Tree, Random Frost, Grident Boosting and SVM. For each cluster both the hyper tunned algorithm's are used. I calculate the ACIP score for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

# Prediction

1. The testing files are shared in the batch and I perform the same validation operations, data transformation and data insertion on them.
2. The accumulated data from db is exported in csv format for prediction.
3. I perform data per-processing techniques on it.
4. Once the prediction is done for all the clusters, The predictions are saved in csv format and shared.

# Q & A

Q1) What is the source of data ?

The data for training is collected by Kaggle.com. The source give me by the client.

Q2) What was the type of data ?

The data was the combination of numerical and Categorical values.

Q3) What is the complete flow you followed in this project ?

Refer slide 5th for better Understanding .

Q4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed ?

I am using different logs as per the steps that I follow in validation and modeling like file validation log. Data Insertion, Model Training log, prediction log etc.

Q6) What techniques were you using for data per-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables.
- Checking and changing distribution of continuous values
- Convert categorical data into numeric values.
- Scaling the data.

Q7) How training was done or what models were used ?

a. Before diving the data in training and validation set I performed clustring over fit to divied the data into cluster.

b. As per cluster the training and validation data were divided.

c. The scaling was performed over training and validation data

d. Algorithm's like KNN, Naïve Bayes, Logistic Regression, Decision Tree, Random Frost, Grident Boosting and SVM were used based on the recall final models used for each cluster and I save the model.

Q8) How Prediction was done?

The testing files are shared by the client. I perform the same life cycle till the data is clustered. Then on the basis of cluster number model is loaded and preform prediction. In the end I get the accumulated data of predictions.

Q9) What are the different stages of deployment?

a. When the model is ready I deploy it in Fire environment. Where SIT and UAT preformed over it.

b. Once We get sign off from fire I deploy in Earth and UAT is performed over it.

c. After getting the sign off from Earth I deploy in production.