# Adult Census Income Prediction

Umang Tank

- Objective
- Dataset
- Data Description
- Architecture
- Data analysis steps
- EDA
- Database connectivity
- Model Selection
- Prediction Result
- Deployment
- Question & Answer

# Objective

- Census data is aim to increase the awareness about how the income factor actually has an impact not only on the personal lives of people, but also an impact on the nation and its betterment.
- We will today have a look on the data extracted from the 1994 Census bureau database, and try to find insights about how different features have an impact on the income of an individual. Though the data is quite old, and the insights drawn cannot be directly used for derivation in the modern world, but it would surely help us to analyze what role different features play in predicting the income of an individual.

# Dataset

- The dataset provided to us contains many rows, and 13 different independent features. We aim to predict if a person earns more than 50k$ per year or not. Since the data predicts 2 values (>50K or <=50K), this clearly is a classification problem, and we will train the classification models to predict the desired outputs.

# Data Description

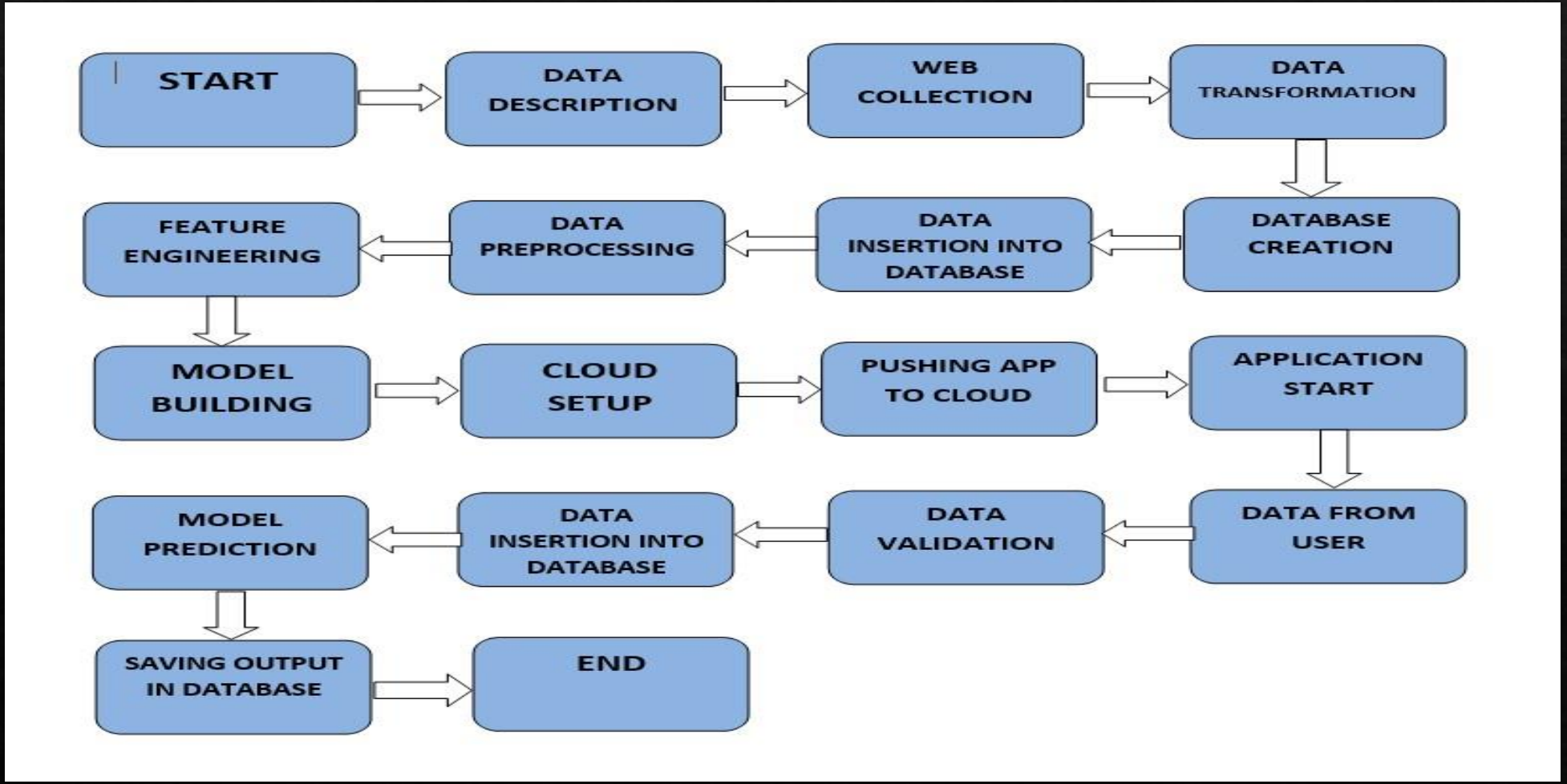
- **Age** — The age of an individual, this ranges from 17 to 90.
- **Workclass** — The class of work to which an individual belongs.
- **Education** — Highest level of education
- **Education_num** — Number of years for which education was taken
- **Marital_Status** — Represents the category assigned on the basis of marriage status of a person
- **Occupation** — Profession of a person

- **Relationship** — Relation of the person in his family
- **Race** — Origin background of a person
- **Sex** — Gender of a person
- **Capital_gain** — Capital gained by a person
- **Capital_loss** — Loss of capital for a person
- **Hours_per_week** — Number of hours for which an individual works per week
- **Native_Country** — Country to which a person belongs

## Output:

- **Income** — The target variable, which predicts if the income is higher or lower than 50K$.

# Architecture



# Data Analysis steps

## DATA COLLECTION

In step 1,  we collect data which is generally present in a database or on internet.

## DATA PREPROCESSING

In step 2, we preprocess the data which involves data cleaning by handling outliers, null values etc.

## EXPLORATORY DATA ANALYSIS

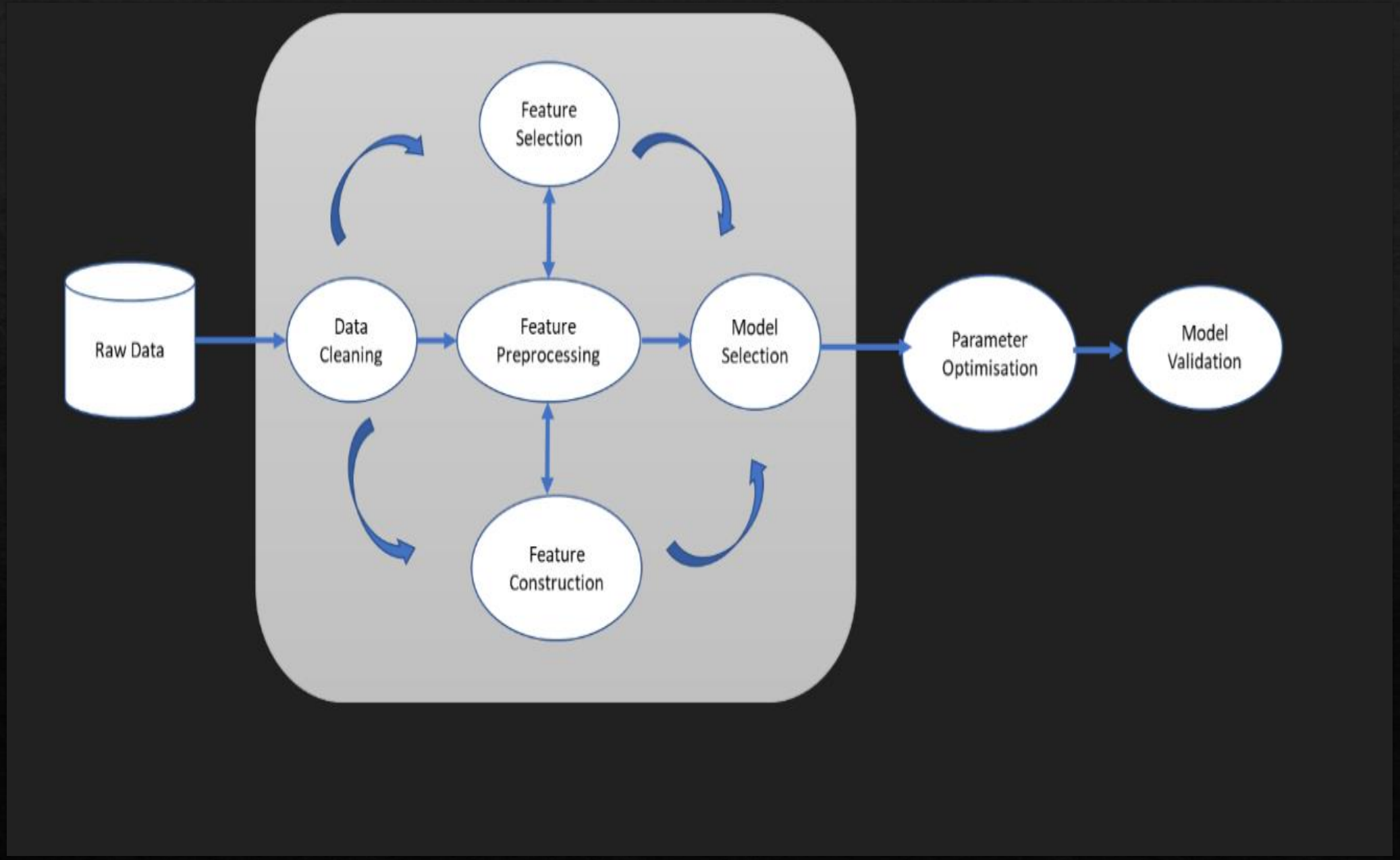
In step 3, we explore the data by performing univariate and bivariate analysis on the features.

## FEATURE SELECTION

In step 4, we use feature selection techniques to filter out the most important features to perform model creation

## MODEL CREATION AND EVALUATION

In step 5, we finally build models on our dataset and choose the model which gives the best accuracy.


Feature Selection
Raw Data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimisation
Model Validation
Feature Construction

# EDA

- Missing value count
- No of rows and columns(Shape)
- categorical/Numerical columns
- Correlation heat map
- Null value handling(impute null value)
- Skewness and log transformation

# Database

- Database creation and connection :
- Table Creation : we have to create a table, if Table is already created then we need to insert new data into database.
- Insertion : All valid input data are inserted into the tables

# Model Selection

- Evaluates classification models using Logistic Regression ,Random forest , Decision Tree, SVC & Ada Boost classifier.
- Compute metrics and generate graphs for model evaluation and importance analysis
- Finally, we fit the Ada boost classifier model with optimal tuning parameters on the entire dataset. We then could use this model to predict Income.

# Model Prediction Result on test dataset

| | model | Precision | recall | f1_score | accuracy |
|---|---|---|---|---|---|
| 0 | LogisticRegression | 0.727834 | 59.365738 | 65.393405 | 0.843738 |
| 1 | DecisionTreeClassifier | 0.634412 | 62.181387 | 62.804969 | 0.816835 |
| 2 | AdaBoostClassifier | 0.766728 | 62.151749 | 68.652807 | 0.858849 |
| 3 | RandomForestClassifier | 0.710831 | 63.604031 | 67.135930 | 0.845139 |
| 4 | SVC | 0.768444 | 51.244813 | 61.486486 | 0.840348 |

# Model Deployment

- The final model is deployed using on Heroku using Flask framework

# Question & answer

## Question 1. Explain about the Project and your day to day task :

The aim is to build models to determine the income level of the people in U.S. It is a binary classification problem to predict if an individual has an income higher than $50k/year.

As a data scientist I am involving in each an every phase of the project. My responsibility consisted of gathering the dataset ,labelling the data for the model, training the model on the prepared dataset , deploying the training model to the cloud, monitoring the deployed model for any issues.

## Question 2: What is the source and size of data ?

**Kaggle - https://www.kaggle.com/overload10/adult-census-dataset, Size of the data usually in MB.**

## Question 3: What was the type of data?

The data was the combination of numerical and Categorical values.

## Question 4: What is Precision and Recall ?

**Precision:** Out of all the points to be predicted Positive, How many of them are actully Positive.

**Recall:** Out of all the points that are lebelled positive, how many of them are predicted Positive.

## Question 5 : What techniques r you using for data pre-processing ?

1) Removing unwanted attributes.
2) Visualizing relation of independent variables with each other and with dependent variable.
3) Cleaning data and imputing if null values are present.
4) Convert Categorical data to numerical data.
5) Scaling the data.

## Question 6 : Does Your Dataset Show Normally Distributed Or Not? If Not Then Which Techniques You Will Use To Make It Normal?

No, These Data Set Does Not Show Normal Distribution Behavior. I Used Log Transformation Techniques To Make It Normally Distributes.

## Question 7 : Which Tool You Are Used For Implementation This Model?

1) IDE : VS Code
2) Cloud : Heroku
3) DataBase : MySql

## Question 8: In which technology you are most comfortable?

I have worked in machine learning and beginner in deep learning.

## Question 9 : What Is Accuracy ?

Accuracy Is One Metric For Evaluating Classification Models.

Accuracy = Number Of Correct Predictions /Total Number Of Prediction.

## Question 10 : How did you optimize your solution?

1) Model optimization depends on various factors
2) Train with better data or do data pre-processing in efficient way.
3) Increase the quantity of training data etc.
4) Try and use multithreaded approaches

## Question 11 : How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem

# Thank You