

Adult Census Income Prediction

Umang Tank

- 
- ❖ **Objective**
 - ❖ **Dataset**
 - ❖ **Data Description**
 - ❖ **Architecture**
 - ❖ **Data analysis steps**
 - ❖ **EDA**
 - ❖ **Database connectivity**
 - ❖ **Model Selection**
 - ❖ **Prediction Result**
 - ❖ **Deployment**

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 $\leq 50K$), this clearly is a classification problem, and we will train the classification models to predict the desired outputs.

Data Description

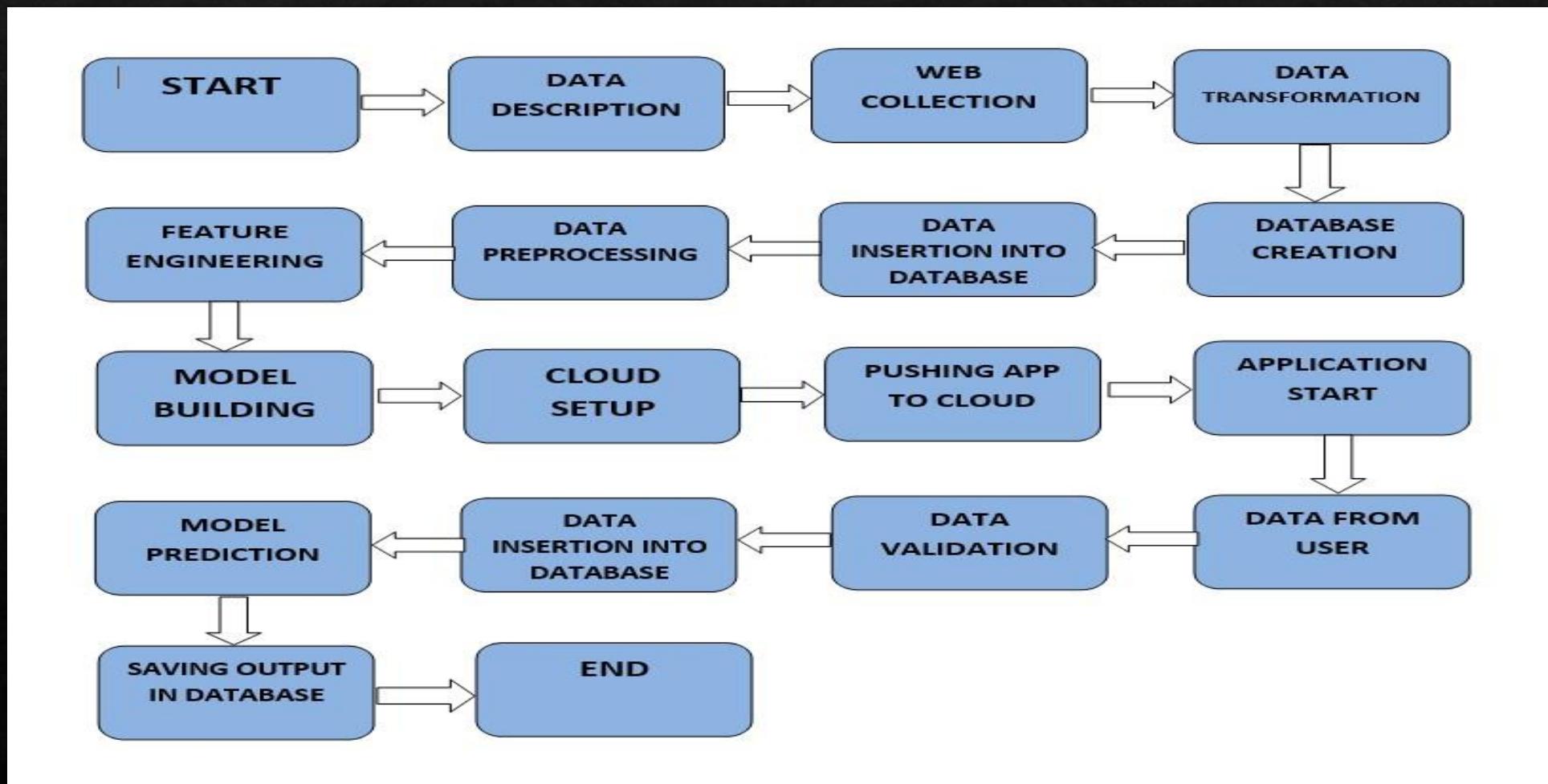
- ❖ **Age** — The age of an individual, this ranges from 17 to 90.
- ❖ **Workclass** — The class of work to which an individual belongs.
- ❖ **Education** — Highest level of education
- ❖ **Education_num** — Number of years for which education was taken
- ❖ **Marital_Status** — Represents the category 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.



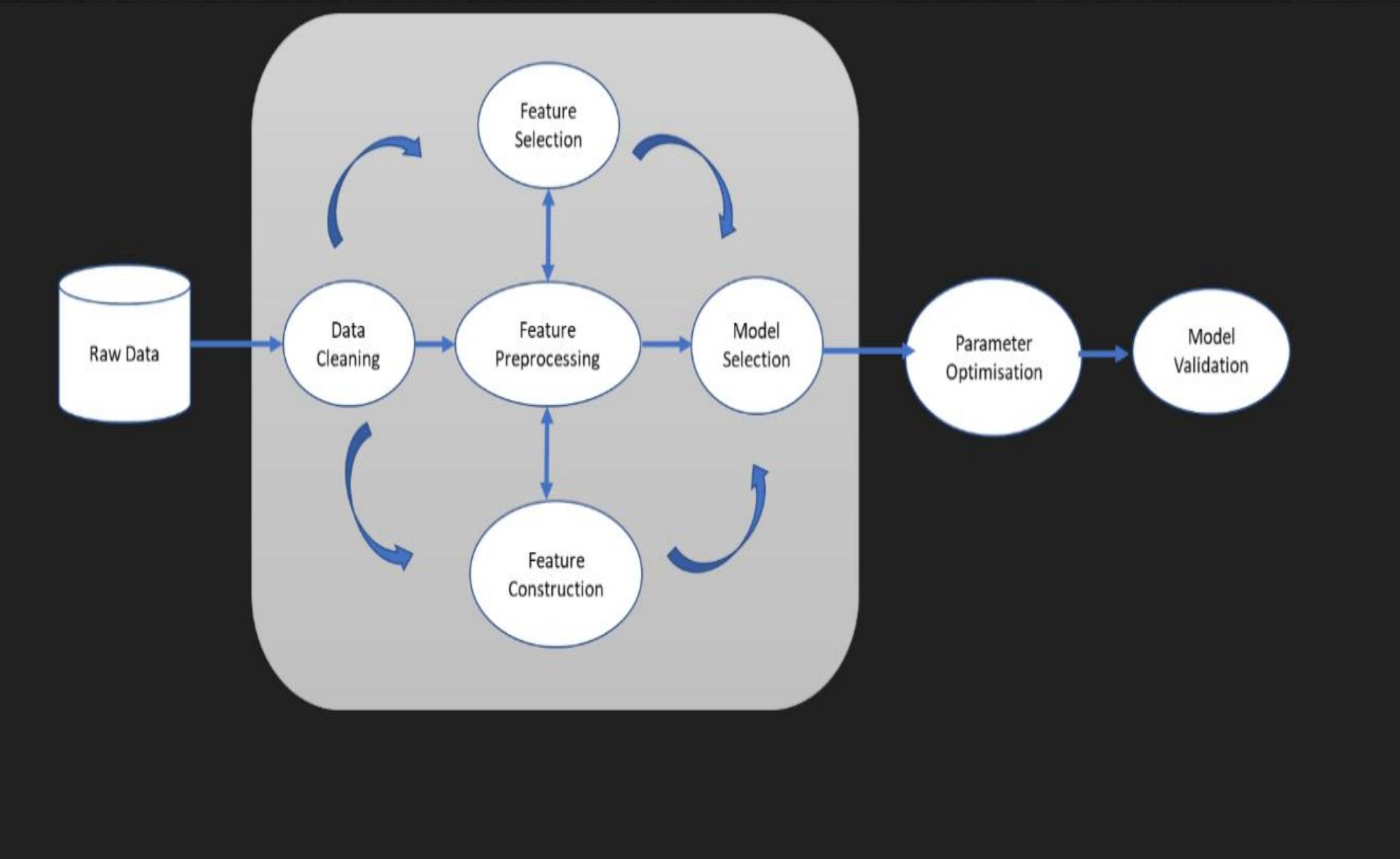
FEATURE SELECTION

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.



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



Thank You