CENSUS INCOME PREDICTION

Umang Tank

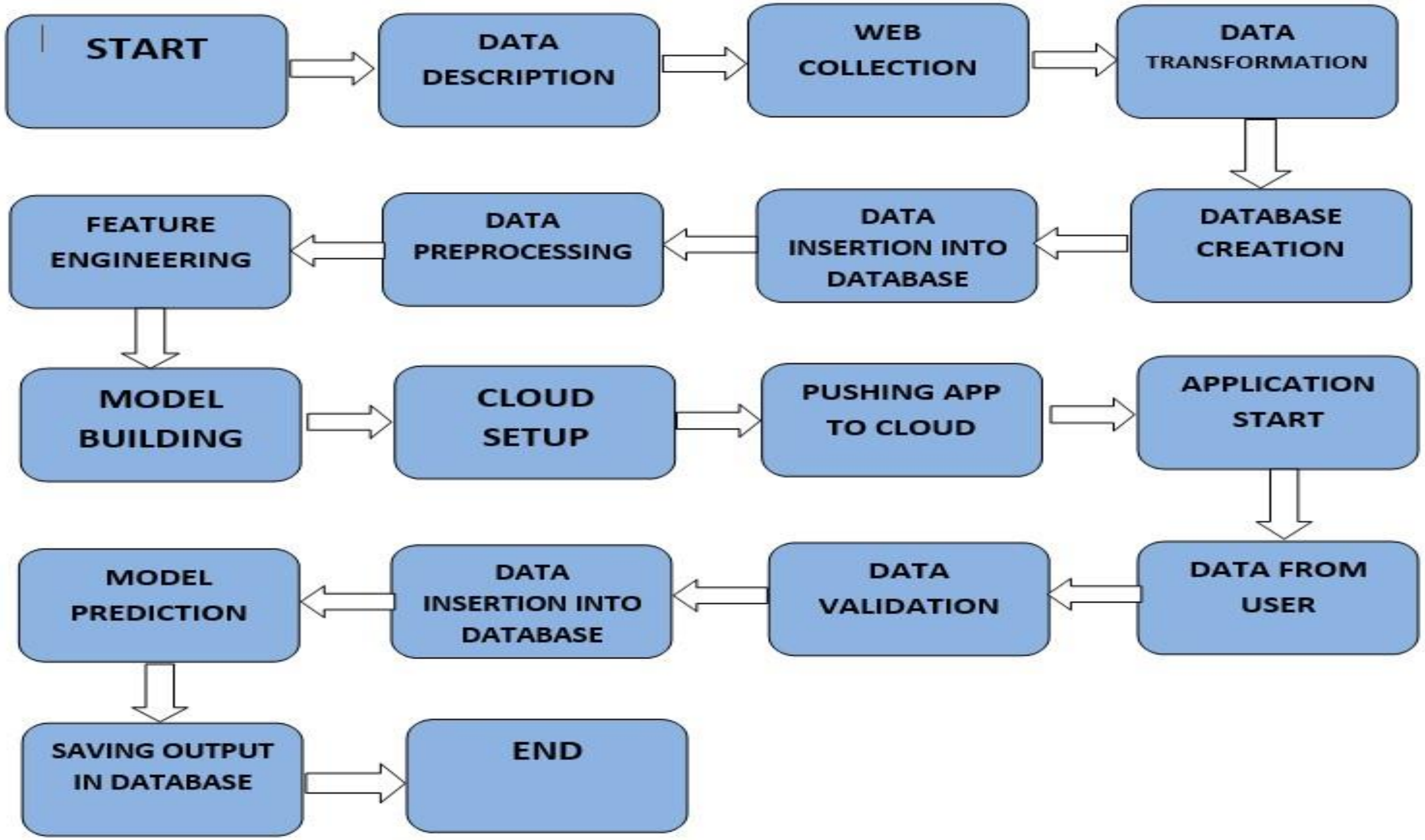
Introduction

Census data is 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.

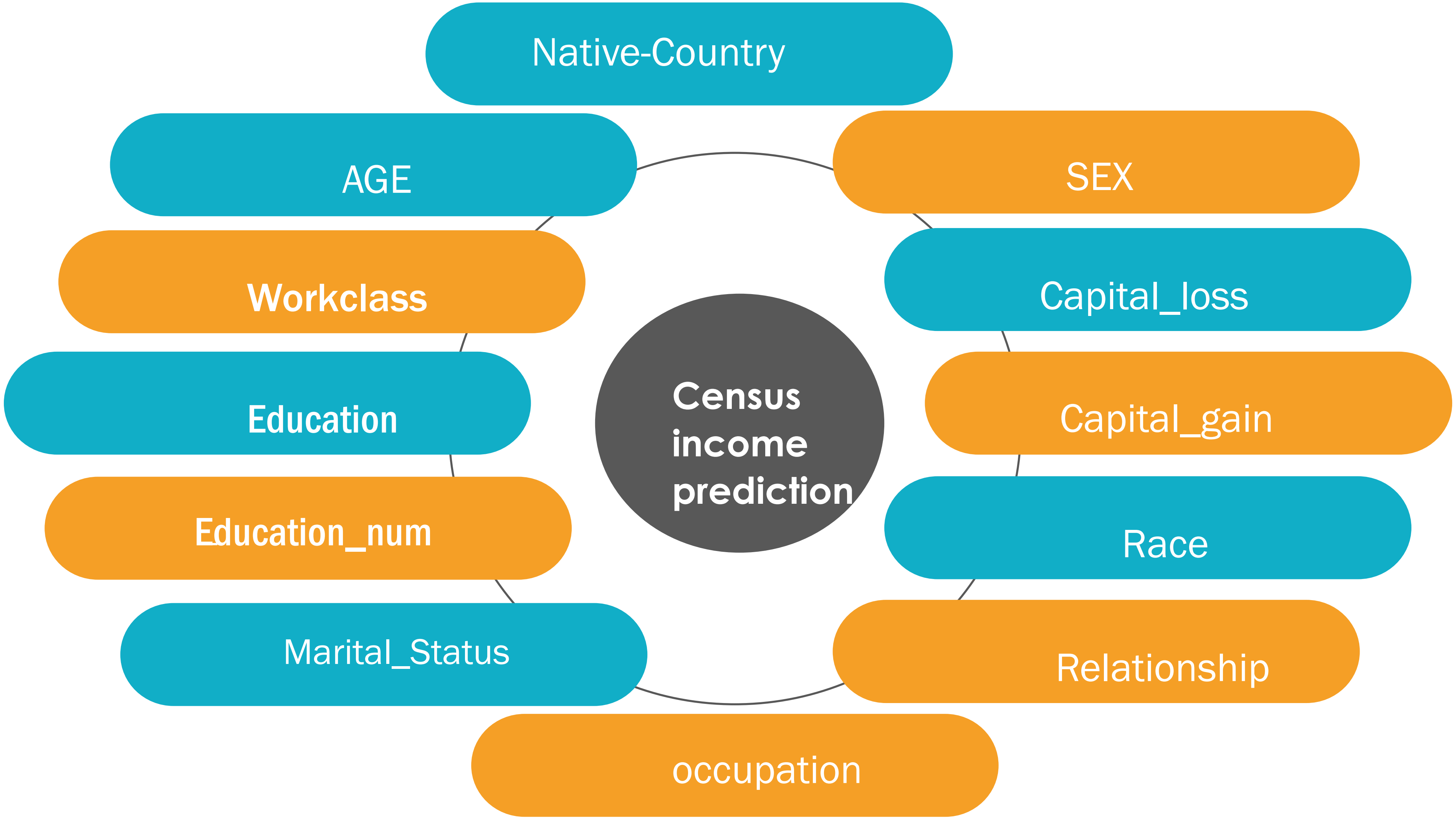
Objective

The dataset provided to us contains 32560 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.

Architecture



DATASET


Census income prediction

Native-Country

AGE

Workclass

Education

Education_num

Marital_Status

SEX

Capital_loss

Capital_gain

Race

Relationship

occupation

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.

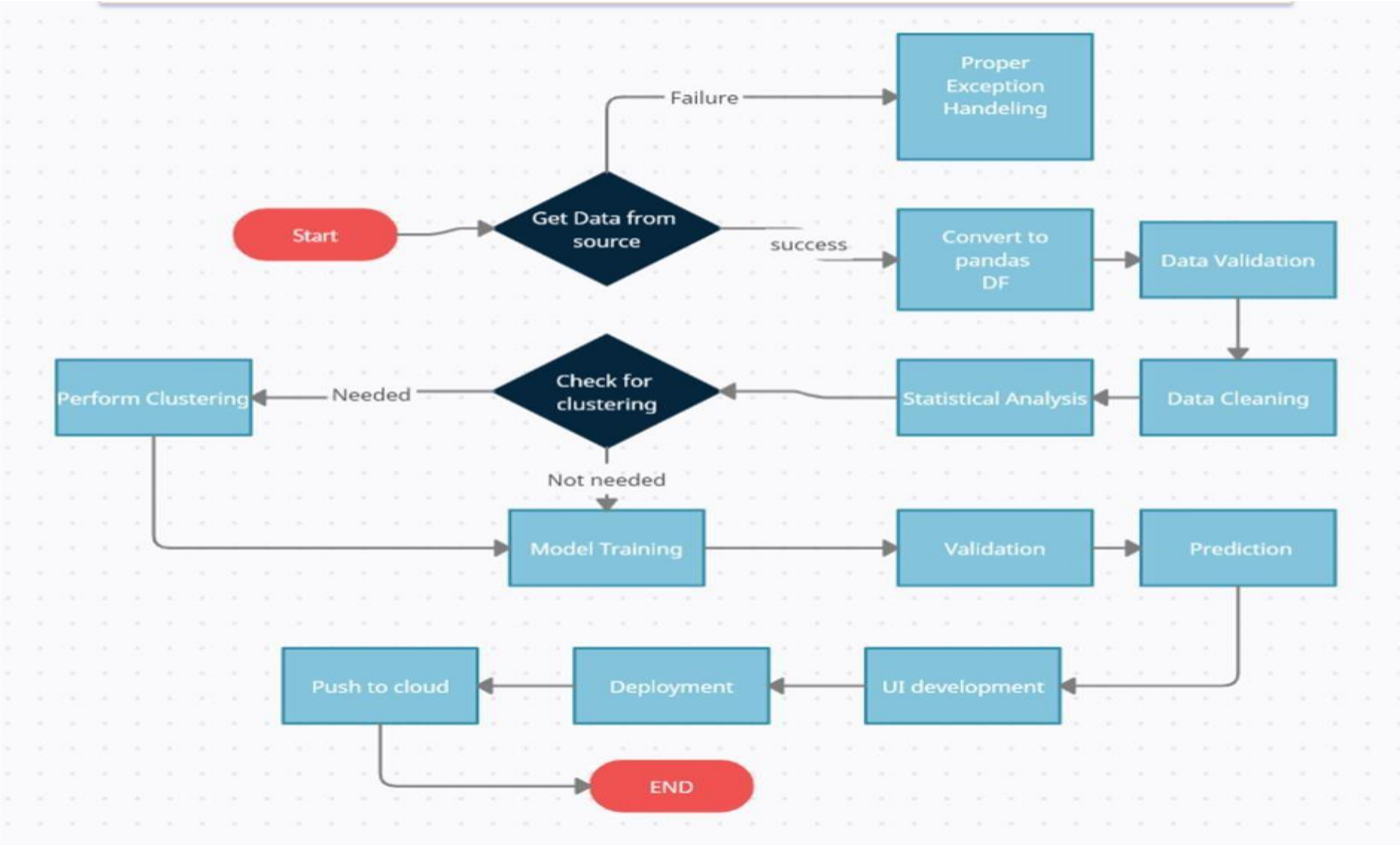
FEATURE SELECTION

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.

Model Training and Validation workflow



Model Training and Validation workflow

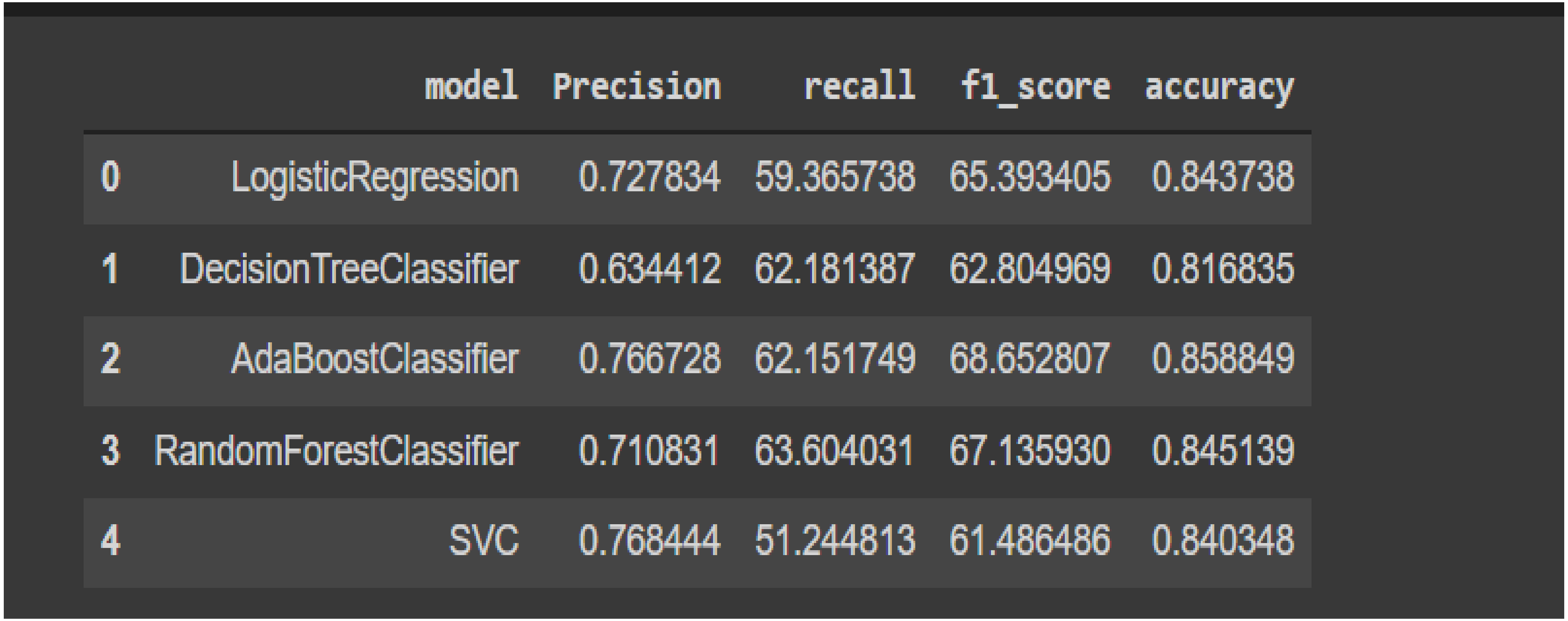
❑ Data Collection:

- Kaggle

❑Data Preprocessing

- Missing data
- Outliers
- Feature engineering
- Feature engineering

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

Database Connection and Deployment

Database Connection:

MySql

Deployment

Heroku

The final model is deployed using on Heroku using Flask framework

THANK YOU