Adult Census Income Prediction

Objective:

To develop a predictive model that monitors the adult income. Based on various factors. The model will determine whether the income is less than 50k or more than 50k.

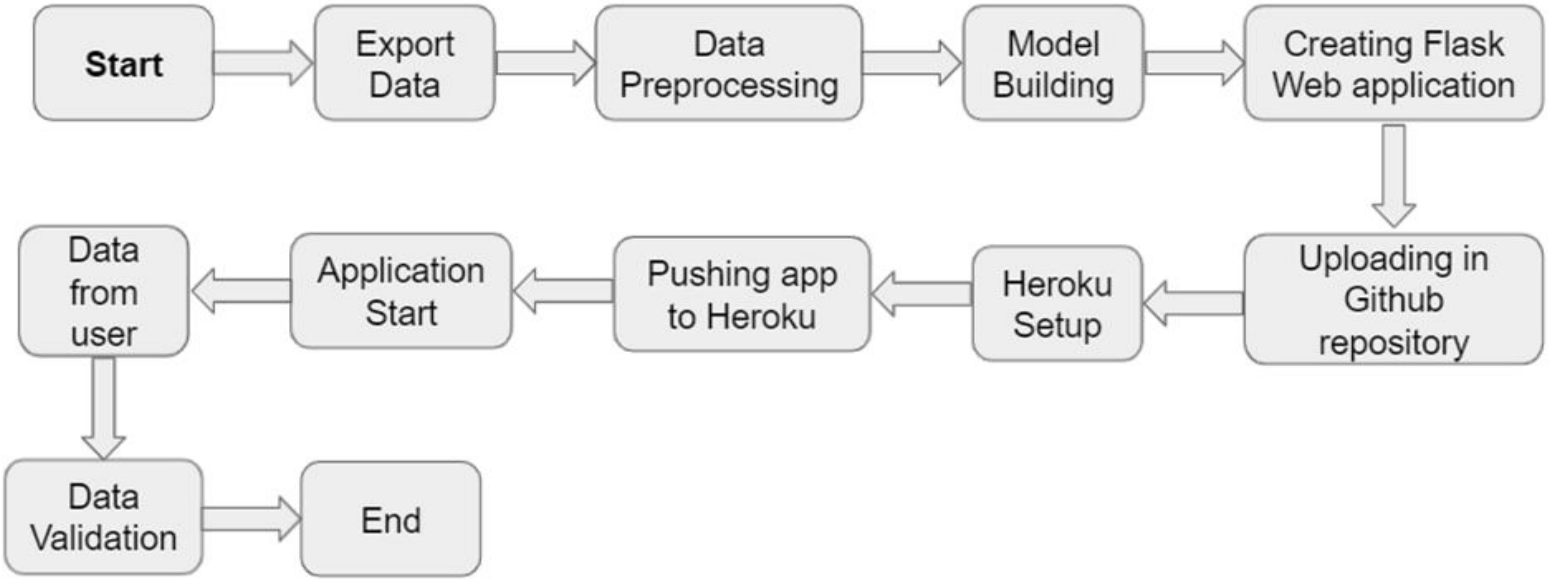
Benefits:

- Gives better insight to the adult income.
- Determines the salary based on factors such as qualification, age, occupation ,gender etc.
- Based on capital gain/loss,salary can be predicted as more than 50k or less than 50k.

Data Sharing Agreement

- Sample file name (adult.csv)
- Number of columns (15)
- Number of rows (32561)
- Column names (age, workclass, education-num, marital-status, occupation,relationship, race, sex, capital-gain, capital-loss,hours-per-week, country, salary)
- Column Data type (categorical ,numerical)

Architecture



Data Validation and Data Transformation

Number of Columns – Validation of number of columns present in the files, and if it doesn't match then the file is removed.

Name of Columns - The name of the columns is validated and should be the same as given in the dataset file. If not, then the file is removed..

Data type of columns - The data type of columns is given in the Numerical as well as Categorical file. It is validated when we insert the files into Database. If the datatype is wrong, then the file is removed.

Model Training

- ➔ Data Export from DB :
  - The accumulated data from db is exported in csv format for model training.
- ➔ Data Preprocessing:
  - Performing EDA to get insight of data like identifying distribution , outliers ,trend among data etc.
  - Check for null values in the columns. If present impute the null values.
  - Encode the categorical values with numeric values.
  - Perform Standard Scalar to scale down the values.

➔ Model Selection:

After data preprocessing is done,Randomised Search CV is used for hyper parameter tuning.By using Randomised Search CV, it takes random parameters to achieve a more reliable estimate of the data points.

After that, Decision Tree is used to train the data and predict the data.

Prediction:

- The testing files are shared in the batches and we perform the same Validation operation, data transformation and data insertion on them.
- The accumulated data from db is exported in csv format for prediction
- We perform data pre-processing techniques on it.
- Decision Tree model created during training is loaded and the testing data is predicted
- Once the prediction is done for all the clusters. The predictions are saved in csv format and shared.

Q & A:

Q1) What's the source of data?

The data for training is provided by the Ineuron.

Q 2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer slide 5th for better Understanding

Q4)What techniques were you using for data pre-processing?

Removing unwanted attributes

Visualizing relation of independent variables with each other and output variables

Checking and changing Distribution of continuous values

Removing outliers

Cleaning data and imputing if null values are present.

Converting categorical data into numeric values.

Scaling the data

Q 5) How training was done or what models were used?

Training was done using different models like XGBoost,Random Forest etc but we decided on Decision Tree since it gives higher accuracy.

Q6) What are the different stages of deployment?

When the model is ready we deploy it in Heroku.

Heroku provides certain link to access the web application by any user.