Detailed Project Report (DPR)

Adult Census Income Prediction

Written By	SUDHANSU GOUDA
Version	1.0
Date	25/02/2023

Document Change Control Record

Version	Date	Author	Comments

Reviews

Version	Date	Reviewer	Comments

- Approval Status:

Version	Review date	Reviewed By	Approved By	Comments

Index

1. Introduction

1.1 Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Adult Census Income Prediction, it has been discussed to predict the category of income of candidate and for understanding the effects of different factors on the salary. Taking various aspects of a dataset collected for Income Prediction, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to make decisions to improve decisions.

1.2 Machine Learning

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

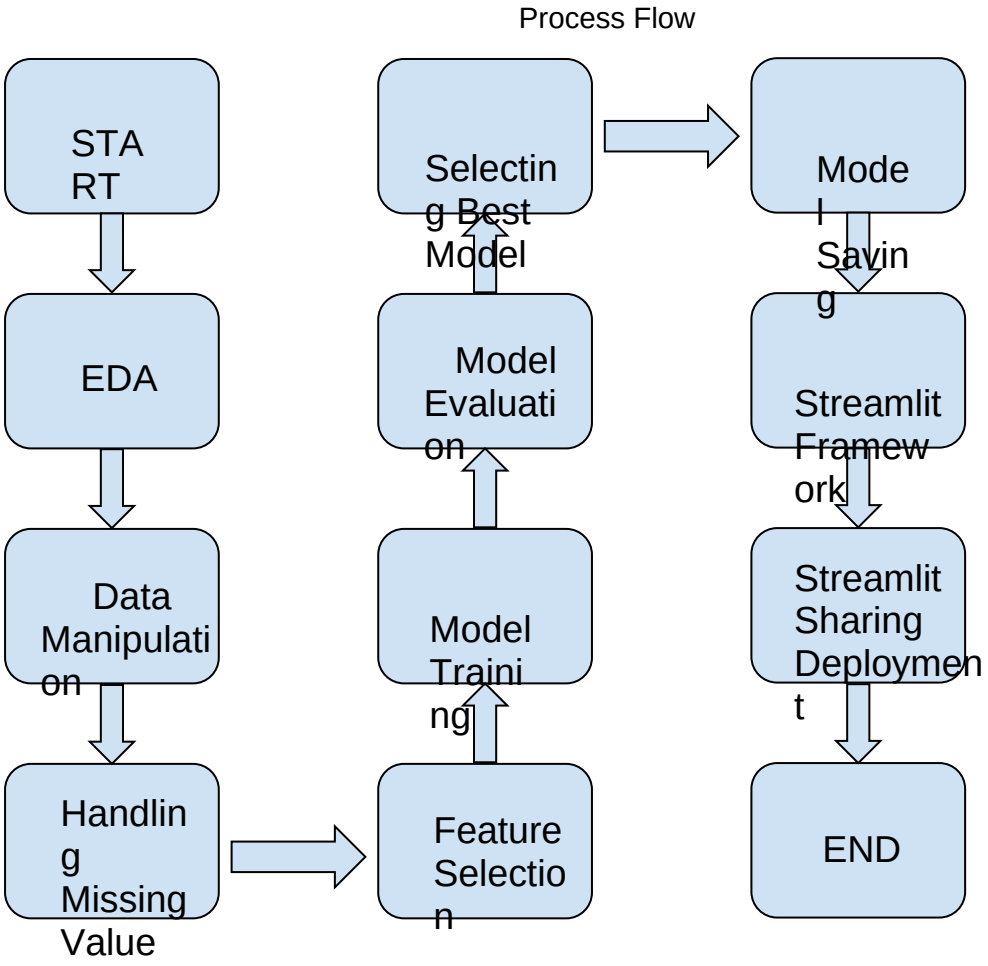
In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key.
ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way.

1.3 Problem Statement

The Goal is to predict whether a person has an income of more than 50K a year or not.
This is basically a binary classification problem where a person is classified into the
>50K group or <=50K group.

2. Architecture:

Following workflow was followed during the entire project.



2.1 Data gathering:

Data source: https://www.kaggle.com/datasets/overload10/adult-census-dataset

2.2 Raw Data Validation:

After data is loaded, various types of validation are required before we proceed further for any operation. Validations like, checking for complete missing values in any columns, etc. These are required because The attributes which contains these are of no use. It will not play role in contributing the sales of an item from respective outlets.
Like if any attribute is having zero standard deviation, it means that's all the values are same, its mean is zero. Which indicate that either the attribute increases or decreases that output will remain the same. Similarly, if any attribute is having full missing values, then there is no use of taking that attribute into an account for operation. It's unnecessary increasing the chances of dimensionality curse.

2.3 Data Transformation

Before sending the data into the database, data transformation is required so that data are converted into such a form with which it can easily be inserted into the database. Here, the 'capital-gain' and 'capital-loss' attributes contain a vast amount of missing values. So, they are removed.

2.4 New Feature Generation

We haven't derived a new category.

2.5 Data preprocessing

In data pre-processing all the processes required before sending the data for model building are performed. Like, here 'capital-gain' and 'capital-loss' the attributes are having some values equal to 0 no doubt both of these attributes are viable for prediction due to maximum entry being zero in the model they can't contribute much. So they have been removed from the dataset. In 'workclass' there were some fields '?' which ment null so the rows were removed. The column of 'hours-per-week' having numerical values was further categorised into brackets of 10 to 10 values. The column of 'marital-status' having string values with 7 different values were categorised into two categories.

2.6 Feature Engineering:

After preprocessing it was found that some of the attributes are not important to the item sales for the particular outlet. So those attributes are removed. There are some columns that needs to be dropped as they don't seem to help in our analysis.

2.7 Model building:

After doing all kinds of preprocessing operations mention above and performing model training and testing the accuracy we came to the conclusion that Random Forest model have the highest accuracy with 81.68% accuracy.

2.8 Model saving:

Model is saved using the pickle library in '. pkl' format.

2.9 Web App Setup:

After saving the model in .pkl file format we then create an app.py streamlit web app framework (Written in python) and then use requests to extract all the form selection selected by the user and then we predict the salary prediction by using the selected records by the user.

2.10 Git Hub:

The whole project directory will be pushed into the GitHub repository.

GitHub Project link: https://github.com/cursD15/Adult_Census_Income_Prediction

2.11 Deployment:

The cloud environment was set up and the project was deployed from GitHub into the Streamlit Sharing cloud platform.

WebApp link - https://cursd15-ineuron-project-app-te9xub.streamlit.app/

3. Data set description:

In this dataset there is data about people who work at different sector with different gender type and marital status which gives us a variety of combinations of data. Using all the observations it is inferred what role certain properties of an candidate play and how they affect their salary and lifestyle. The dataset looks like as follow:

```
In [4]: df.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	Ur St
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	Ur St
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	Ur St
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	Ur St
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cu

rkclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	country	salary
e-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
-emp-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
ate	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
ate	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
ate	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

The data set consists of various data types from integer to float to object as shown in Fig.

```
In [5]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  country         32561 non-null  object
 14  salary          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands pre-processing of data.
The dataset should therefore be explored as much as possible.
Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes.

```
In [6]: df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
age	32561.0	38.581647	13.640433	17.0	28.0	37.0	48.0	90.0
fnlwgt	32561.0	189778.366512	105549.977697	12285.0	117827.0	178356.0	237051.0	1484705.0
education-num	32561.0	10.080679	2.572720	1.0	9.0	10.0	12.0	16.0
capital-gain	32561.0	1077.648844	7385.292085	0.0	0.0	0.0	0.0	99999.0
capital-loss	32561.0	87.303830	402.960219	0.0	0.0	0.0	0.0	4356.0
hours-per-week	32561.0	40.437456	12.347429	1.0	40.0	40.0	45.0	99.0

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and model values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during the model building.

4. Implementation and Results

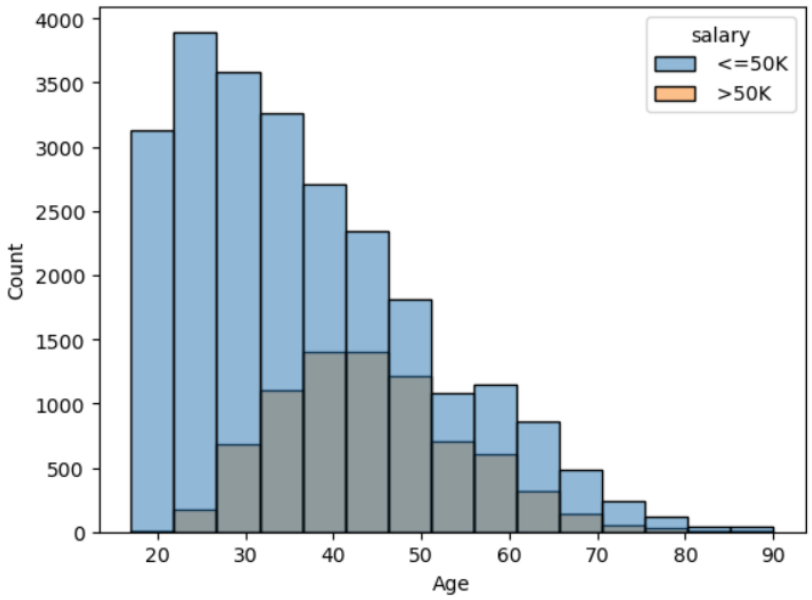
In this section, the programming language, libraries, implementation platform along with the data modeling and the observations and results obtained from it are discussed

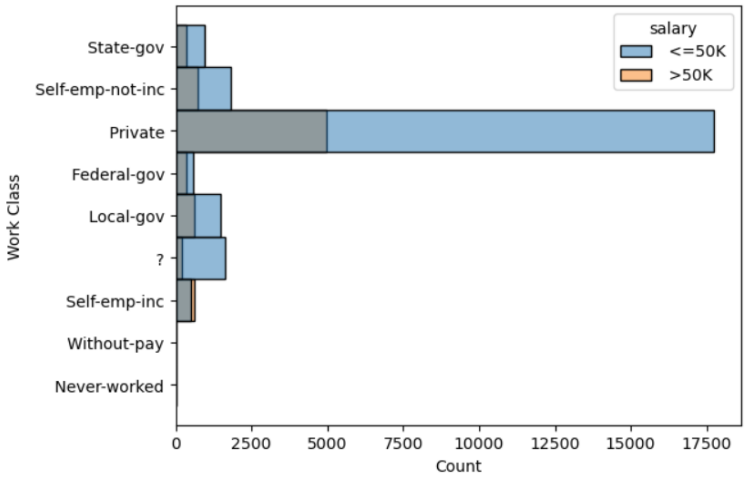
4.1 Implementation Platform and Language

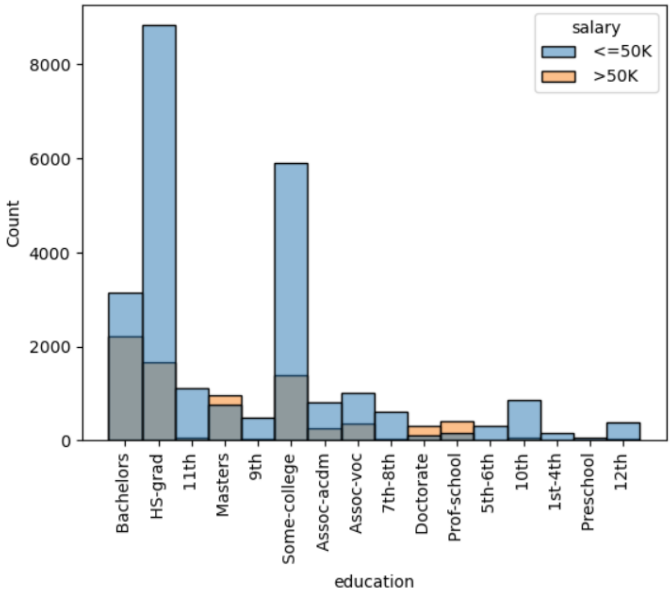
Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the 'batteries included language' for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

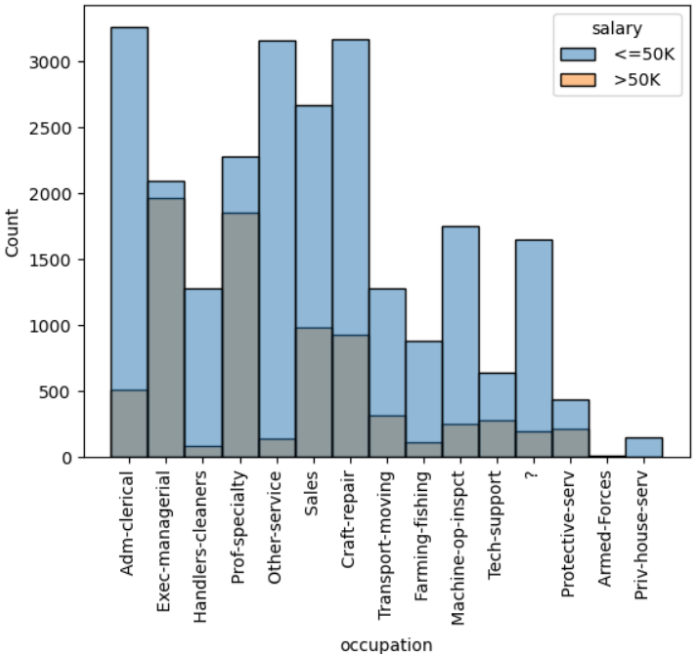
In this work, the Python libraries of NumPy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. Random forest classifier is used to solve tasks. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in 'literate programming', where human friendly code is punctuated within code blocks, has been used.

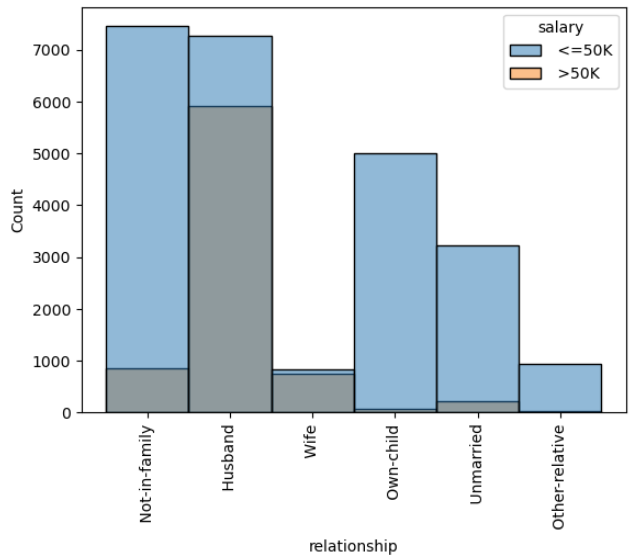
4.2 Observations

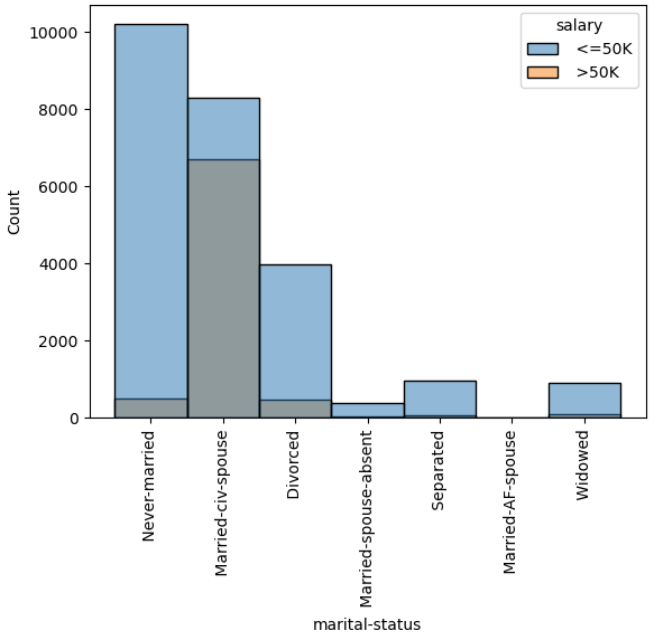


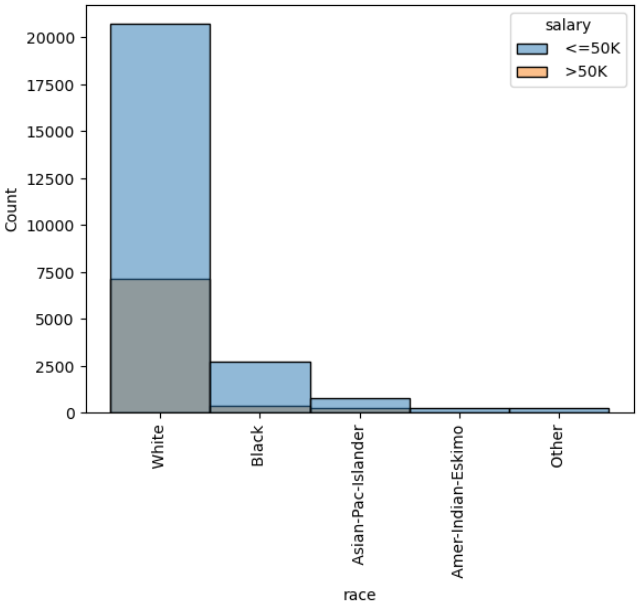

salary
<=50K
>50K
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
?
Self-emp-inc
Without-pay
Never-worked
Work Class
0
2500
5000
7500
10000
12500
15000
17500
Count

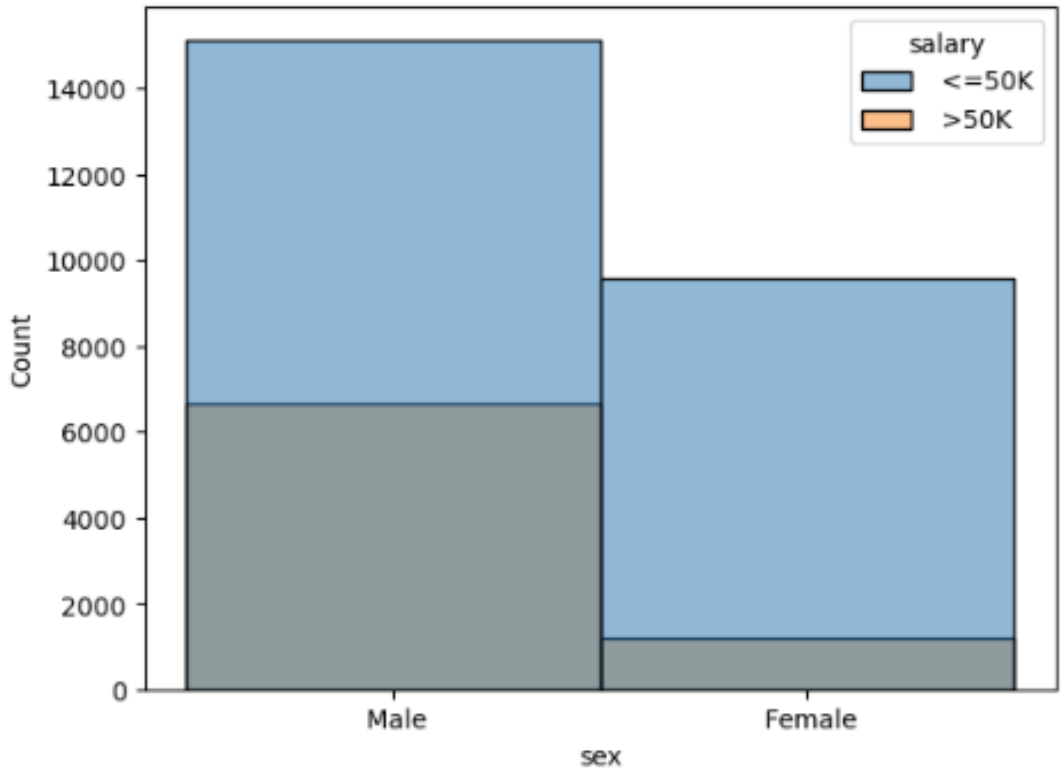

salary
<=50K
>50K
Count
0
2000
4000
6000
8000
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
education

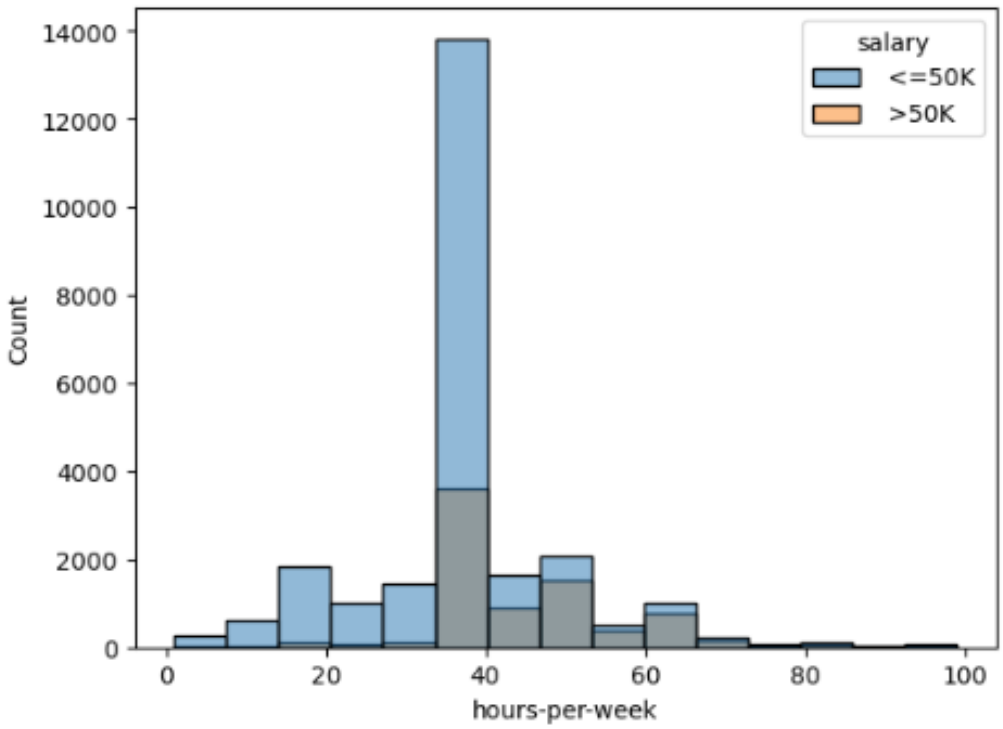

salary
<=50K
>50K
Count
0
500
1000
1500
2000
2500
3000
Adm-clerical
Exec-managerial
Handlers-cleaners
Prof-specialty
Other-service
Sales
Craft-repair
Transport-moving
Farming-fishing
Machine-op-inspct
Tech-support
?
Protective-serv
Armed-Forces
Priv-house-serv
occupation

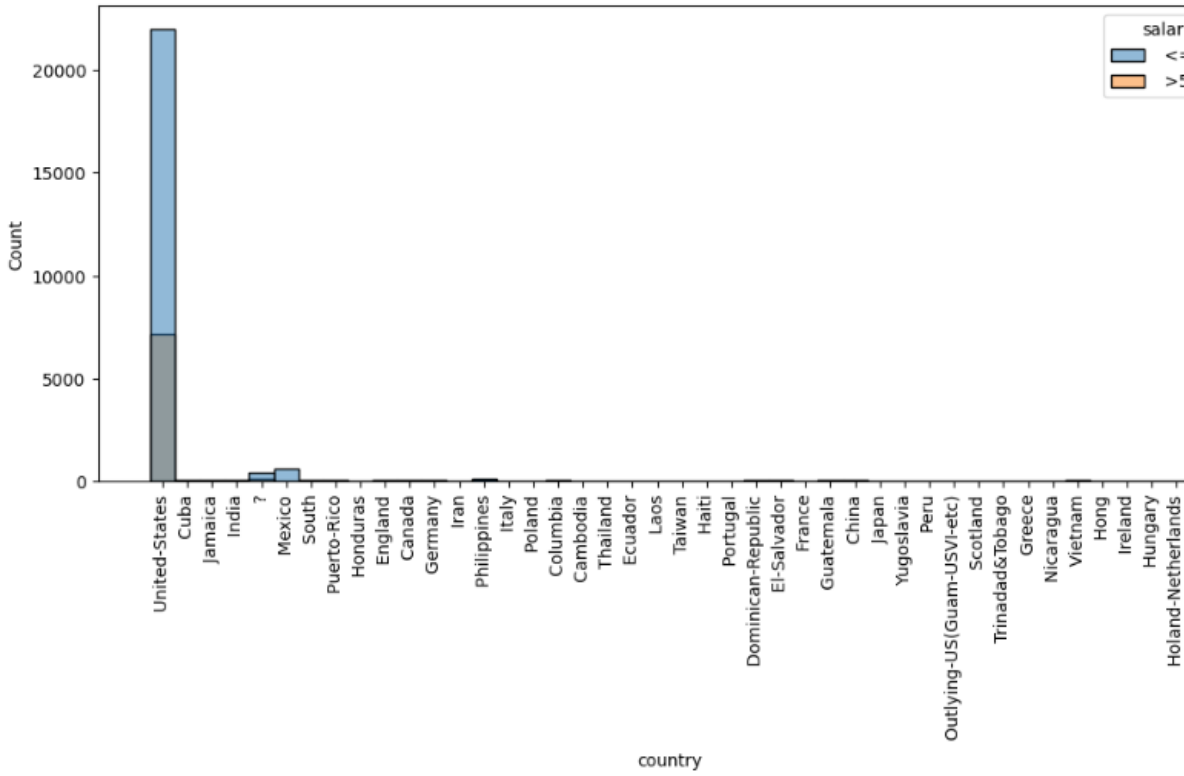

salary
<=50K
>50K
Count
0
1000
2000
3000
4000
5000
6000
7000
Not-in-family
Husband
Wife
Own-child
Unmarried
Other-relative
relationship


salary
<=50K
>50K
Count
0
2000
4000
6000
8000
10000
Never-married
Married-civ-spouse
Divorced
Married-spouse-absent
Separated
Married-AF-spouse
Widowed
marital-status


salary
<=50K
>50K
Count
0
2500
5000
7500
10000
12500
15000
17500
20000
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
race


salary
<=50K
>50K
14000
12000
10000
8000
6000
4000
2000
0
Count
Male
Female
sex


salary
<=50K
>50K
14000
12000
10000
8000
6000
4000
2000
0
Count
0
20
40
60
80
100
hours-per-week



We get the following insights from visualization:

-Aage group 0-20 there are not any individual of salary >50K, same occured with group <75 years.

-The majority work in the private sector... and '?' == NULL value.

-FnlWgt == final weight :- we were not able to find any pattern with final weight.

-Pree school to 12th == Extra... Mostly individulas with higher education have higher chance of having salary >50K.

-Married have a higher chance of having salary >50K as compared to unmarried.

-Prof-specialty Exec-managerial have a higher rate to having salary >50K as compared to others.

-We can see clearly that husband have a higher rate to having salary >50K.

-As compared in the ratio white and Asian race have higher rate of salary >50K but at the same time white race have much higher entries as compared to others.

-We can see clearly that male have a higher rate to having salary >50K simliar to husband and wife case.

-Distribution of capital-gain is not very linear same is with capital-loss.

-After simplifying hours-per-week we came to conclusion that salary increase from less than 40 hrs to 40 hrs then decrease.

-Country mostly contain US which is not very helpful.

-
-
-
-
-
-

- ## 4.4 Metrics for Data Modelling

For model selection we had used evaluation techniques which is score.

- ## 4.5 Prediction results

After doing all kinds of preprocessing operations mention above and performing scaling data is passed to Random Forest model It was found that it performs best Random Forest model have the highest accuracy with 81.68% accuracy. So Random forest performed well in this problem.

- ## 5. Conclusion

- In conclusion, adult data sense prediction is a promising technology with a broad range of potential applications in different fields. Its ability to analyze patterns in adult data can provide valuable insights and predictions that can enhance safety, mental health, education, law enforcement, and marketing. With continued advancements in the technology, its scope is likely to expand further, making it a valuable tool for various industries. Overall, adult data sense prediction has the potential to improve the lives of individuals and contribute to a safer and healthier society.

- ## 6. Future Scope

Adult data sense prediction is a technology that uses machine learning algorithms to analyze patterns in adult data. This technology can be applied to various fields to provide insights and predictions. For example, in online safety, it can help identify harmful content and prevent users from engaging with it. In mental health, it can identify individuals who are at risk of developing certain disorders and provide them with early intervention and support. In education, it can help teachers and administrators identify students who may be struggling with mental health issues or engaging in risky behaviors. In law enforcement, it can be used to identify individuals who may be involved in illegal activities. In marketing, it can help companies target their products and services to individuals who are most likely to engage with them. As the technology continues to evolve, the potential applications for adult data sense prediction are likely to expand, making it a valuable tool for various industries.

7. Q & A:

Q1) What's the source of data?

Ans. The data for training is provided by the client from:
https://www.kaggle.com/datasets/overload10/adult-census-dataset

Q 2) What was the type of data?

Ans. The data was the combination of numerical and Categorical values.

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

Ans. Refer the Architecture section for this.

Q 4) What techniques were you using for data pre-processing?

- Removing unwanted attributes
- Visualizing relation of independent variables with each other and output variables
- 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?

- Before diving the data in training and validation set, we performed clustering over fit to divide the data into clusters.
- As per cluster the training and validation data were divided.
- The scaling was performed over training and validation data.
- Different algorithms approach and Finalized algorithm is Random Forest.

Q 6) How Prediction was done?

Ans. The testing files are shared by the client. We pass its data to the best model which we have saved in pickle format and get the prediction.

Q 7) Where the model was deployed?

Ans. When the model is ready, we deploy it in Streamlit Sharing platform. This model is a web application where user can enter the data and these data gets extracted in the backend and user gets the prediction result.