[Date]

**Rutuja Bankar**

**(ENTC)**

**Adult Census Income Prediction**

**Abstract:**

The prominent inequality of wealth and income is a huge concern, especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improves the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for this purpose. Classification has been done to predict whether a person's yearly income in the US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollars category based on a certain set of attributes

**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 groups.

**Problem Solution:**

Machine learning can help businesses improve products or speed up manual and time-consuming processes. Managing an increasing number of online customer interactions has become a pain point for most businesses. The concept of demand forecasting is used in multiple industries, from retail and e-commerce to manufacturing and transportation.

**Technical Requirements:**

In this project, the requirements are to get personal income through various platforms. For that, in this project, we are going to use different technologies. Here are some requirements for this project.

* Model should be exposed through API or user interface, so that anyone can test the model.
* Cassandra database should be integrated into this project for any kind of user input.
* Model should be deployed on the cloud (Azure, AWS, GCP).

**Data Requirements:**

Data Requirement completely depends on our problem.

• For training and testing the model, we are using the adult census income dataset from the UCI machine learning repository.

• From the user we are taking the following input:

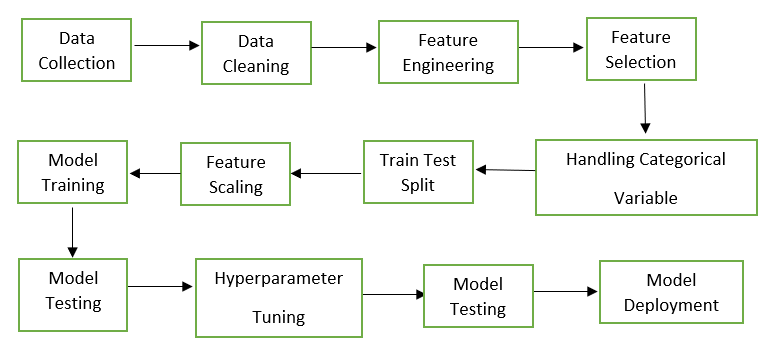
1. Age
2. Work for class
3. Education
4. Marital-Status
5. Occupation
6. Relationship
7. Race
8. Sex
9. Country
10. Capital Loss

**Tools used:**

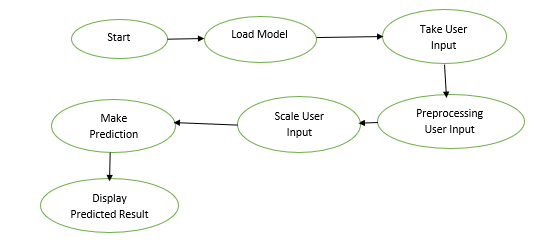
Python programming language and frameworks such as NumPy, Pandas,

TensorFlow and Jupyter Notebook are used to build the whole model. 

**Design Details:**

****

**Deployment Process:**



**KPIs (Key Performance Indicators):**

**Marketing:**

1. Website Traffic

2. Social Media Traffic

3. Conversation Rate on Call-To-Action Content

4. Blog Articles Published Per Month

5. Clickthrough Rates

**IT:**

1. Total System Downtime

2. Number of Tickets/Resolutions

3. Number of Developed Features

4. Count of Critical Bugs

5. Back-up Frequency

**Sales:**

1. Customer Lifetime Value (CLV)

2. Customer Acquisition Cost (CAC)

3. Average Dollar Value for New Contracts

4. Average Conversion Time

5. Number of Engaged Leads

**Staffing:**

1. Absenteeism Rate

2. of Overtime Hours Worked

3. Employee Satisfaction Number

4. Employee Turnover Rate

5. Number of Applicants

**Deployment:**



**Scope:**

This software system will be a model evaluation and keep the best model in the system to predict the outcomes.

**Constraints:**

We will only be predicting the model on the limited number of features; an additional number of features can increase the model’s robustness**.**

**Risks:**

The possibility of losing a prospective client if the system is unable to predict the right candidate who is eligible for the loans.

**Out of Scope:**

Delineate specific activities, capabilities, and items that are out of scope for the project

**Dataset:**

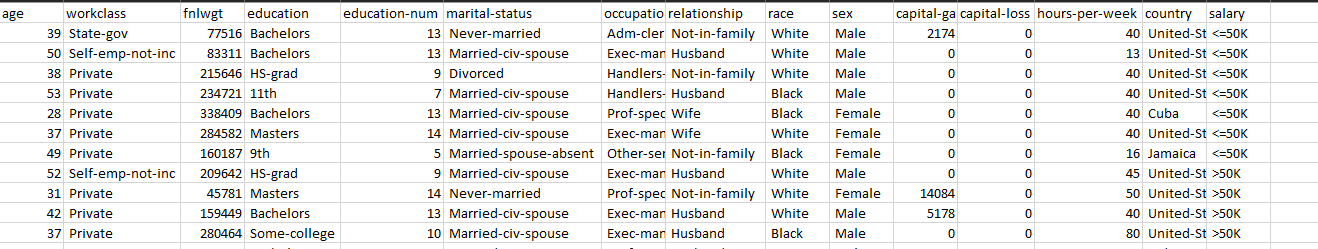
financial datasets, a machine learning model might be able to predict the behaviour of a given asset. That’s why the financial sector is doing everything in its power to create an effective ML model, as anything that can predict even reasonably well has the potential to generate millions of dollars. Machine learning is already predicting the behaviour of citizens, which is impacting the way policy makers are doing their jobs



**Adult Census dataset overview:**

Consists of a single rectangular dataset with 32562 records, 14 feature columns, and one target column. In which Age is a Numerical Discrete Variable, Work Class is a Categorical Nominal Variable, the flight is a Numerical Continuous Variable, Education is a Categorical Nominal Variable, Education Num is Numerical Discrete Variable, Marital Status is a Categorical Nominal Variable, Occupation is a Categorical Nominal Variable, Race is a Categorical Nominal Variable, Sex is a Categorial Binary Variable, Capital Gain and Loss is a Numerical Continuous Variable, Hours per week is a Numerical Discrete Variable, Country is a Categorical Nominal Data and Target Variable is a Categorical Binary Variable.

**Adult Data Table:**



**Test Cases:**

|  |  |  |
| --- | --- | --- |
| **Test Case Description** | **Pre-Requisite** | **Expected Result** |
| Verify whether the Application URL is Accessible to the user | Application URL Should be defined | The application URL should be Accessible to the user |
| |  | | --- | | Verify whether the Applicationloads completely for the userwhen the URL is accessed | | 1. Application URL is accessible 2. Application is deployed | The Application should load completely for the user when the URL is accessed |
| Verify whether the User is able to sign up for the application | Application is accessible | The User should be able to sign up For the application |
| Verify whether a user is able to successfully use the application | Made sure to check for the test cases from backend. | The user should be able to see successfully valid results |
| Verify whether user is able to see input fields on logging | 1. Application is accessible 2. User is able to log into the application | User should be able to edit input fields on logging |

**Conclusion:**Adult census income prediction web application predicts the person’s income is more than 50k or not. The accuracy of this application is good and it will predict 85-90% accurate results. I hope you have now gained some knowledge of what is Data Analysis, how to build Machine learning models, and how python helps us to build those models. This article may be the first step of your Machine Learning journey but, remember, "The secret of getting ahead is getting started.”