**Architecture Design**

**Adult Census Income Prediction**

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

**Document Control**

### **Change Record**:

| **Version** | **Date** | **Author** | **Comments** |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### **Approval Status:**

| **Version** | **Review**  **Date** | **Reviewed By** | **Approved By** | **Comments** |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

**Index**

| **Content** | **Page No.** |
| --- | --- |
| Abstract | 4 |
| 1. Introduction | 4 |
| 1.1 What is Architecture Design? | 4 |
| 1.2 Scope | 4 |
| 1.3 Constraints | 4 |
| 2. Technical Specification | 5 |
| 2.1 Dataset | 5 |
| 2.2 Logging | 6 |
| 2.3 New Feature Generation | 7 |
| 2.4 Deployment | 7 |
| 3. Technology Stack | 7 |
| 4. Proposed Solution | 8 |
| 5. Architecture | 8 |
| 5.1 Architecture Description | 8 |
| 6. User Input/Output Workflow | 11 |

**Abstract**

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted. Certain attributes of each product and store have been defined. The aim is to build a predictive model and find out whether the income of the person is greater than or less than $50,000. Using this model we will get insights as well as we get to determine the income group.

**1. Introduction**

**1.1 What is Architecture Design?**

The goal of Architecture Design (AD) or a low-level design document is to give the internal design of the actual program code for the ‘**Adult Census Income Prediction** ’. AD describes the class diagrams with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

**1.2 Scope**

Architecture Design (AD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

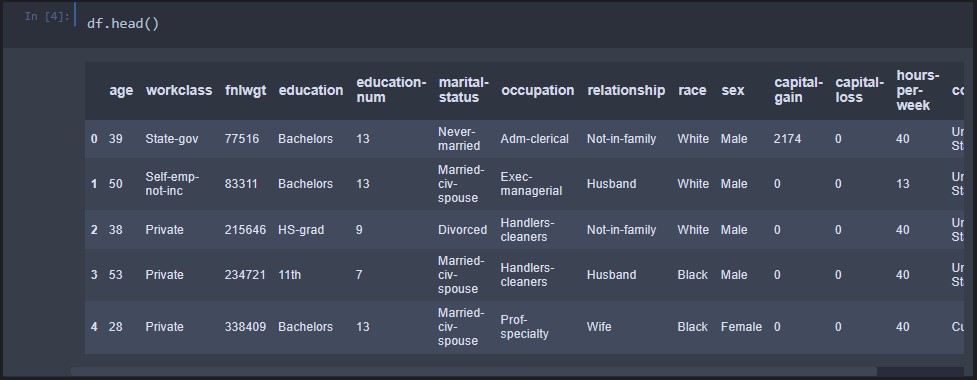
**1.3 Constraints**

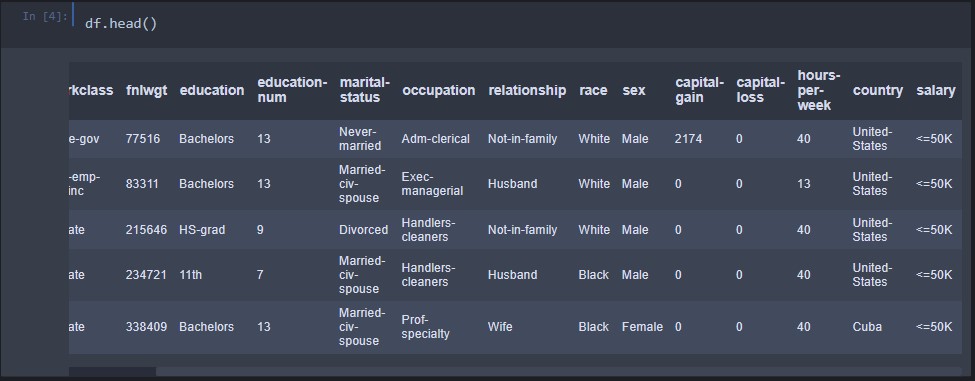
We only predict the income category based on occupation, workclass, education, gender, age, work hours, marital status about the candidate.

**2. Technical Specification**

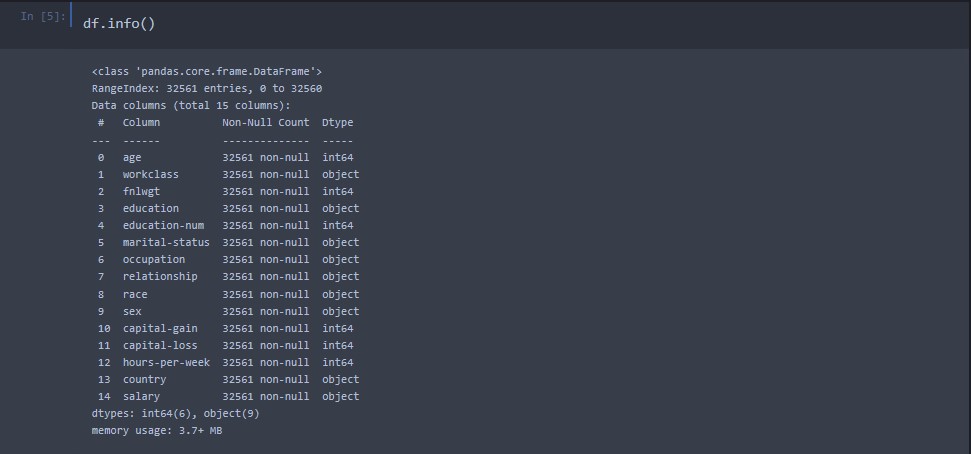
**2.1 Dataset**

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:





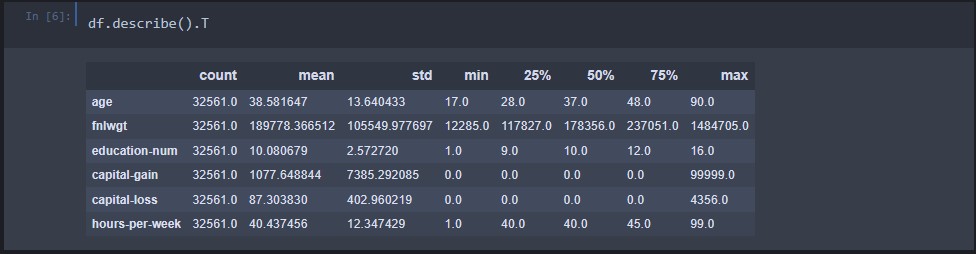
The data set consists of various data types from integer to floating to object as shown in Fig.



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.



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 their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tell 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, play 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 a one-hot encoding scheme during the model building.

**2.2 Logging**

We should be able to log every activity done by the user

* The system identifies at which step logging require.
* The system should be able to log each and every system flow.
* Developers can choose logging methods. Also, can choose database logging.
* The system should be not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

**2.3 Database**

The system needs to store every request into the database and we need to store it in such a way that it is easy to retain and look into the records.

The system should capture every data that any user gave and the prediction that has been made by that input.

**2.4 Deployment**

For the deployment process, we will be using Streamlit Sharing cloud to Deploy our model. Streamlit is used to make web apps as well as a cloud platform that lets companies build, deliver, monitor and scale apps.



**3. Technology Stack**

| Web App | Streamlit |
| --- | --- |
| Back End | Python |
| Deployment | Streamlit Sharing |

**4. Proposed Solution**

We will use Exploratory Data Analysis (EDA) to find the Key relations between different attributes and will use a ML algorithm to predict the future sales demand. We can tell the company what are all the challenges they may face, what are the brands or products which is sold the most & other such kind of things, this helps sales team to understand which product to sell & which product to promote & other such kind of things.

**5. Architecture**

Process Flow



**5.1 Data Gathering**

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

**5.2 Raw Data Validation**

After data is loaded, various types of validation are required. Validations like checking for zero standard deviation for all the columns, checking for complete missing values in any columns, etc. These are required because The attributes which contain these are of no use. It will not be much of use in determining of the salary category.

Like if any attribute is having zero standard deviation, it means that’s all the values are the same, its mean is zero. Missing data in the training data stands as a problem in front of us.

**5.3 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.

**5.4 New Feature Generation**

We haven’t derived a new category.

**5.5 Data Pre-processing**

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.

**5.6 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.

**5.7 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.

**5.8 Model Saving**

Model is saved using pickle library in ‘.pkl’ format.

**5.9 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.

**5.10 GitHub**

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

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

**5.11 Deployment**

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

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

**6. User Input / Output Workflow.**

