

|  |
| --- |
|  |
|  |
| November 24  iNeuron.ai  Authored by: Darshita Paghadal |



## **Census-Income Dataset Analysis** Income Prediction

**Abstract**

The economic well-being of a Nation is highly driven by the income of the residents.

Countless decisions in private and public sectors are based on Census data. Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities.

Not only the above, the census data is also used for post census population estimates and projections, economic and social science research, and many other such applications. Hence, the importance of this data and its correct predictions is very clear to us.

Data has always been the backbone of many important decisions. When an assumption is backed up by facts and numbers, the chances of incorrectness and bad decisions decrease.

# 1. Introduction

## **1.1 Why this Low-Level Design Document?**

The purpose of this Low-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a low level.

It will explain It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the higher management for its approval.

The main objective of the project is to predict if a person is getting income more than 50k or not based on the given dataset.

## **1.2 Scope**

This software system will be a Web application This system will be designed to predict the census income rate at earliest for better understanding and useful for countless decision in private and public financial sectors.

Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities.

Not only the above, the census data is also used for post census population estimates and projections, economic and social science research, and many other such applications. Hence, the importance of this data and its correct predictions is very clear to us.

# 2. Technical Specifications

**2.1 The Dataset**

The dataset provided to us contains 32560 rows, and 14 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.

Mentioned below are the details of the features provided to us, which we will be feeding to our classification model to train it.

1. Age — The age of an individual, this ranges from 17 to 90.
2. Workclass — The class of work to which an individual belongs.
3. Fnlwgt — The weight assigned to the combination of features (an estimate of how many people belong to this set of combination)
4. Education — Highest level of education
5. Education\_num — Number of years for which education was taken
6. Marital\_Status — Represents the category assigned on the basis of marriage status of a person
7. Occupation — Profession of a person
8. Relationship — Relation of the person in his family
9. Race — Origin background of a person
10. Sex — Gender of a person
11. Capital\_gain — Capital gained by a person
12. Capital\_loss — Loss of capital for a person
13. Hours\_per\_week — Number of hours for which an individual works per week
14. Native\_Country — Country to which a person belongs.

**2.2 Predicted Data:**

1. Income — The target variable, which predicts if the income is higher or lower than 50K$.

**2.3 Logging**

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

* The System identifies at what step logging required
* The System should be able to log each and every system flow.
* Developers can choose logging methods. You can choose database logging/ File logging as well.
* System should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

**2.4 Deployment**

1. AWS



**3. Technology Stack**

|  |  |
| --- | --- |
| Front End | Flask |
| Backend | Python, Sklearn |
| Data | CSV Files |
| Deployment | AWS |

## **4. Proposed Solution**

## The solution proposed here to predict data is machine learning based solution using different algorithms. Different algorithms are tried and experimented on given data set and find proposed solution. As there is low amount of data available, decision tree-based algorithm works better. In case of large amount of data, we need to find other solution.

## In this problem, we build various models like logistic regression, knn classifier, support vector classifier, decision tree classifier, random forest classifier and xgboost classifier.

## A hyperparameter tuned random DecisionTreeClassifier gives the highest accuracy score of 85.33

## **5. Model training/validation workflow**

Data Set

EDA

Feature Engineering

SVC

Kneighbors Classifier

Decision Tree Classifier

Logistic Regression

Select Percentile Encoder

One Hot Encoder

Model pkl File

Best Performer

Column Transformer

Categorical Pipeline

Standard Encoder

Simple Imputer Encoder

Numerical Pipeline

Pipelines

## **6. User Interface Workflow**

`

Prediction

API request

Prediction API

Data

Fetch Data

Casandra Database

Product Application

Predict Best Model

Training Pipeline