High Level Design (HLD)

*Adult Census Income Prediction*

Revision Number: 1.0

Last date of revision: 9/08/2021

Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Date Issued | Version | Description | Author |
| 9/08/2021 | 1 | Initial HLD | Devendra PITALIYA |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

Document Version Control...

Abstract........................................................................

1. Introduction.5

1 .1 Why this High-Level Design Document? ..... . . . . . .5

1 .2 Scope. . . . . .5

* 1. Definitions ... .6

2.General Description.6

2. 1 Product Perspective .. . . . . .6

2.2 Problem statement .. . . . . .6

#### 2.3 PROPOSED SOLUTION .. . . . . .6

2.4 Technical.7

2.5 Data Requirements .

2.6 Tools used...........................................................................................................7

 2.6. 1 Hardware Requirements.................................................................................7

2.7 Constraints……………………………………………………………………………………………………………………8

1. Design Details ………………………………………………………………………………………………………9
   1. ProcessFlow.................................................................................................9
      1. Model Training and Evaluation …………………………………………………….9
      2. Deployment Process…………………………………………………………………….. 9
   2. Event log 1 0
   3. Error Handling . . 10
   4. Performance.................................................................................................11
   5. Reusability................................................11
   6. Application Compatibility. 1 1
   7. Resource Utilization. . . .11
   8. Deployment ..... ..... ........................ .. . . . 11 2
2. Conclusion.. . . . 1 2

Abstract

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 improve 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 the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollar’s category based on a certain set of attributes. The Random Forest Classifier Model was deployed which clocked the highest accuracy of 82.00%, eventually breaking the benchmark accuracy of existing works.

# 1 Introduction

## 1.1 Why this High-Level Design Document?

The purpose of this High-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 high level.

The HLD will:

* Present all of the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project
* List and describe the non-functional attributes like:

 Security

 Reliability o Maintainability

 Portability

 Reusability

 Application compatibility o Resource utilization

 Serviceability

## 1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

### 1.3 Definitions

Term Description

|  |  |
| --- | --- |
| UGV  Database  IDE  A WS | Unmanned Ground Vehicle  Collection of all the information monitored by this system  Integrated Development Environment  Amazon Web Services |

# 2 General Description

## 2.1 Product Perspective

The Adult senses income prediction is a machine learning-based income prediction model which will help us to predict the annual income base on parameters.

### 2.2 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.

### 2.3 PROPOSED SOLUTION

The dataset consist of 8 categorical and 6 continuous attributes containing information on age, education, nationality, marital status, relationship status, occupation, work classification, gender, race, working hours per week, capital loss and capital gain as . The binomial label in the data set is the income level which predicts whether a person earns more than 50 Thousand Dollars per year or not based on the given set of attributes. I have used many algorithm to predict the output data (income). Applied different classification algorithms (Logistic Regression, RandomForestClassifier, Gaussian, DecisionTreeClassifier, SVC) after hyper-parameter-tuning got 83% accuracy with RandomForestClassifier algorithm

### 2.4 Technical Requirements

This document addresses the requirements for predictinn of adult income.

* Good system – min 4gb ram, 1 tb HDD
* There are some library that used for data exploration like, pandas, numpy, matplotlib
* For backend development Flask is used

##### 2.5 Data Requirements

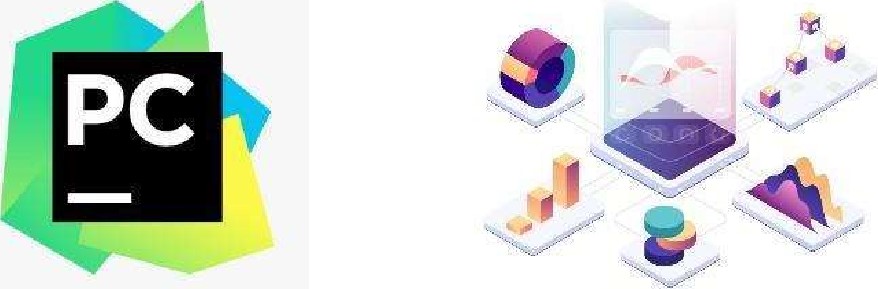
Data requirement completely depend on our problem statement.

We require huge amount of data for prediction which includes age, education, nationality, marital status, relationship status, occupation, work classification, gender, race, working hours per week, capital loss and capital gain

#### 2.6 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, TensorFlow, Keras and Roboflow are used to build the whole model.

 flask

PyCharm is used as IDE.

* For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
* AWS is used for deployment of the model.
* Tableau/Power Bl is used for dashboard creation.
* Front end development is done using HTML/CSS
*  Python flask is used for backend development.  GitHub is used as version control system.
* PC (check you are system supports:

##### 2.7 Constraints

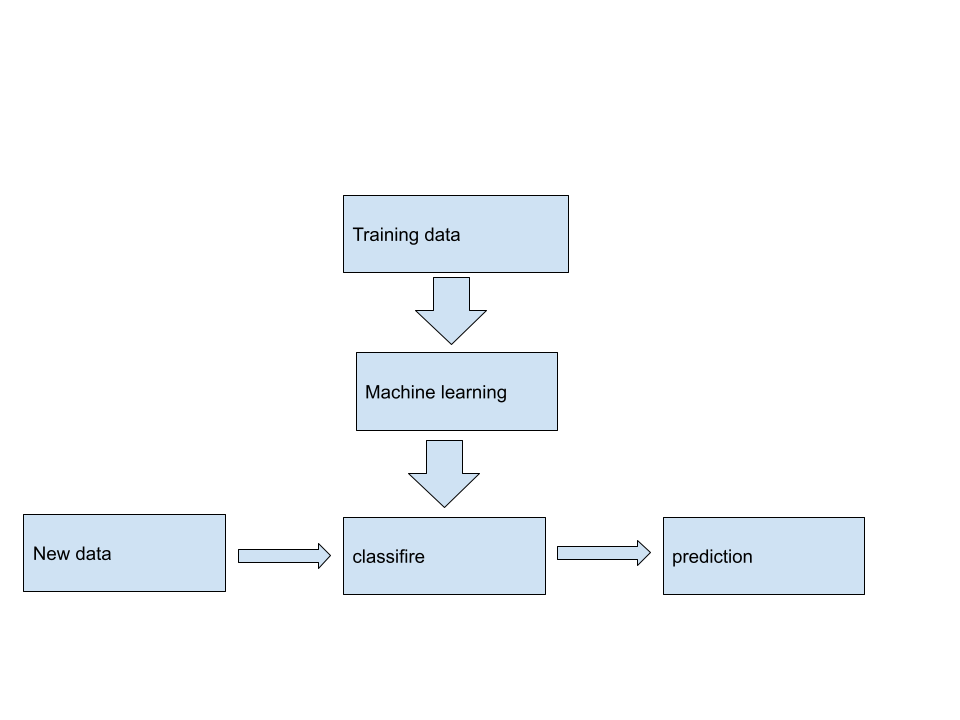
The adult senses income prediction model system must be user friendly, as automated as possible and users should not be required to know any of the workings.

1. Design Details

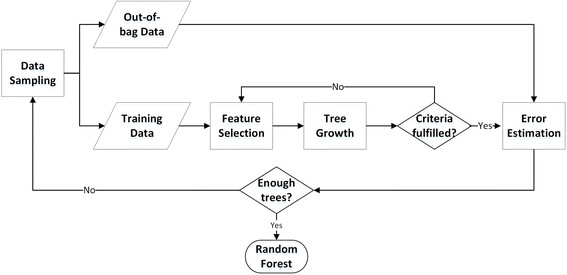
3.1 Process Flow

For identifying the income, we will use a machine learning base model. Below is the process flow diagram is as shown below.

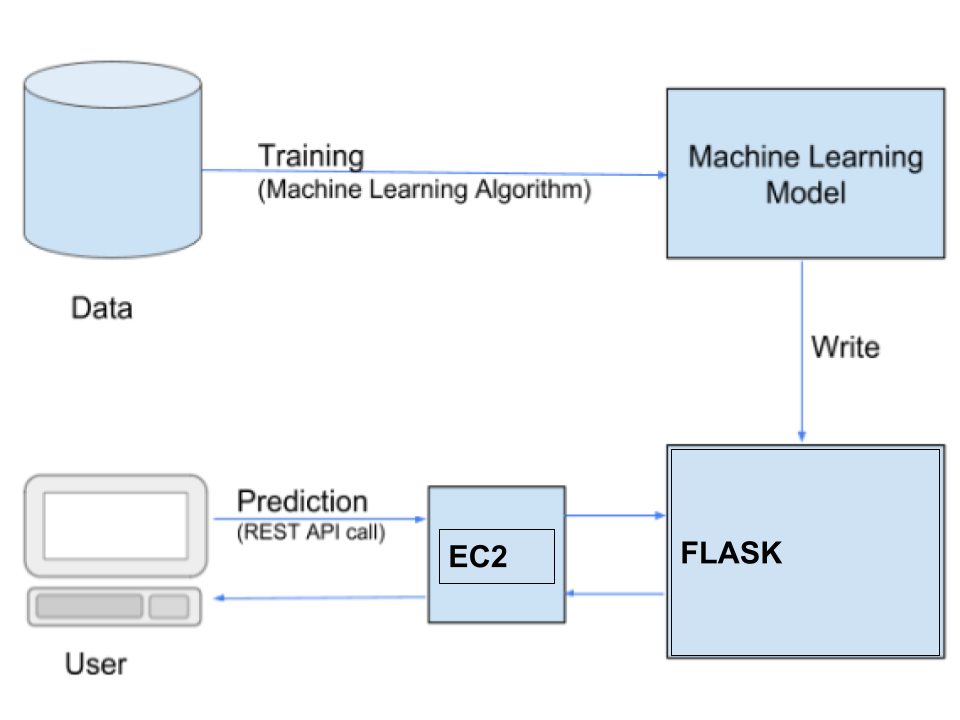
#### Proposed methodology



##### 3.1.1 Model Training and Evaluation



##### 3.1.2 Deployment Process



**3.3 Error Handling :-** Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

10

4.Performance

#### Financial Domain Problem whether the prediction task is to determine whether a person makes over $50K a year

#### 4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

#### 4.2 Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

#### 4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that

function is finished.

4.4 Deployment

EWS EC2



[This Photo](http://commons.wikimedia.org/wiki/file:aws_simple_icons_aws_cloud.svg) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

11

Conclusion

This paper proposed the application of Ensemble Learning Algorithm, Gradient Boosting Classifier with extensive Hyperparameter Tuning with Grid Search on Adult Census Data. Finally, the Validation Accuracy, so obtained, 82.00% which is, by the best of our knowledge, has been the highest ever numeric accuracy achieved by any Income Prediction Model so far. The future scope of this work involves achieving an over-all better set of results by using hybrid models with inclusion of Machine Learning and Deep Learning together, or by applying many other advanced pre-processing techniques without further depletion in the accuracy.

12

References

* + - 1. . <https://arxiv.org/ftp/arxiv/papers/1810/1810.10076.pdf>

13