# INTRODUCTION

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 Kaggle 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 Dollars category based on a certain set of attributes. The Gradient Boosting Classifier Model was deployed which clocked the highest accuracy of 88.16%, eventually breaking the benchmark accuracy of existing works.

* 1. **SCOPE**

Over the last two decades, humans have grown a lot of dependence on data and information in society and with this advent growth, technologies have evolved for their storage, analysis and processing on a huge scale. The fields of Data Mining and Machine Learning have not only exploited them for knowledge and discovery but also to explore certain hidden patterns and concepts which led to the prediction of future events, not easy to obtain. The problem of income inequality has been of great concern in the recent years. Making the poor better off does not seem to be the sole criteria to be in quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in the society. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals.

**1.2 Existing System**

The existing system uses Support Vector Machine (SVM) one of the main issues with this is that it need the data to be linearly separable. The system also does not provide enough preprocessing and visualization or Exploratory Data Analysis(EDA).

**Disadvantages of Existing System:**

The limitations of available systems are not sufficient to deal with the complex data. In this

section, we present some of the limitations that are present in the existing system.

* The model suffers from overfitting due to no generalization of data.
* The error on test data is high due to overfitting.
* The system also requires data extensive data preprocessing and Exploratory Data Analysis(EDA) inorder to perform feature engineering.

**1.3 Proposed System**

We aim to build other classification models like logistic regression, Naïve Bayes,

Decision Trees and others and also fine tune the parameters of the model. These models would be trained on a data set which will be engineered carefully after performing the feature engineering.

**Advantages:**

* Load and explore the dataset and generate ideas for data preparation and model selection.
* Systematically evaluate a suite of machine learning models with a robust test harness.
* Fit a final model and use it to predict class labels for specific cases.

1. **SYSTEM ANALYSIS**

It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components.

System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

Analysis specifies what the system should do.

**2.1 Functional Requirements**

For a successful project, the requirements must be clear. Especially the evaluation metric must be specified. In ML systems,

* 1. The quality of the resulting predictions can be considered a functional requirement
  2. Predictive power is of great significance.
  3. Training an ML model to go beyond a certain utility breakpoint turns into a functional requirement in practice.

**2.2 Performance Requirements**

Performance requirements define how well the system performs certain functions under specific conditions. Examples are speed of response, throughput, execution time and storage capacity. Selecting and interpreting performance measures appropriately is crucial for the acceptance of ML systems. The main requirements for good performance are maximum

1. accuracy and
2. precision

**2.3 Software requirements**

Operating System : Windows 7 , Windows 8, (or higher versions)

Language : Python 3.5 and other libraries likes numpy, pandas, matplotlib, seaborn and scikitlearn.

Mozilla Firefox(or any browser)

**2.4 Hardware requirements**

Processor : Pentium 3,Pentium 4 and higher

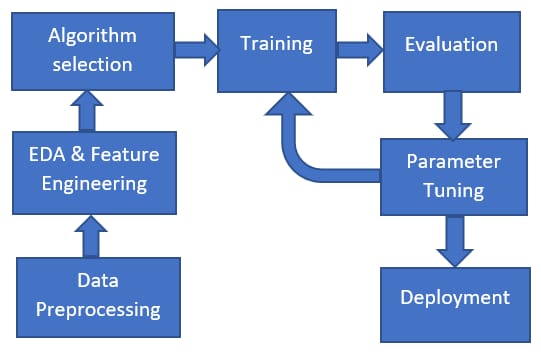
RAM : 2GB/4GB RAM and higher

Hard disk : 40GB and higher

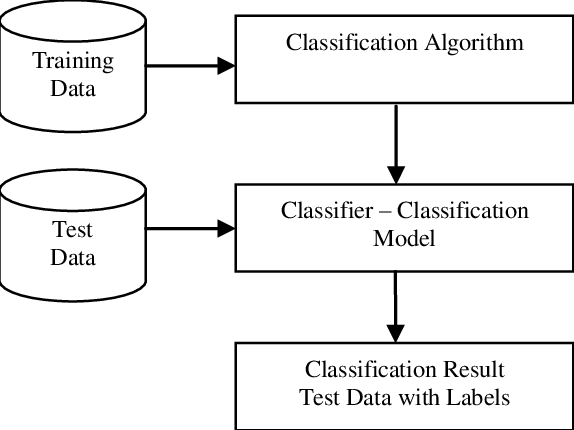
**3. SYSTEM DESIGN**

**3.1 UML Diagrams**

(Unified Modeling Language-UML)

****

**Process of trained model:**

****

**4. SYSTEM IMPLEMENTATION**

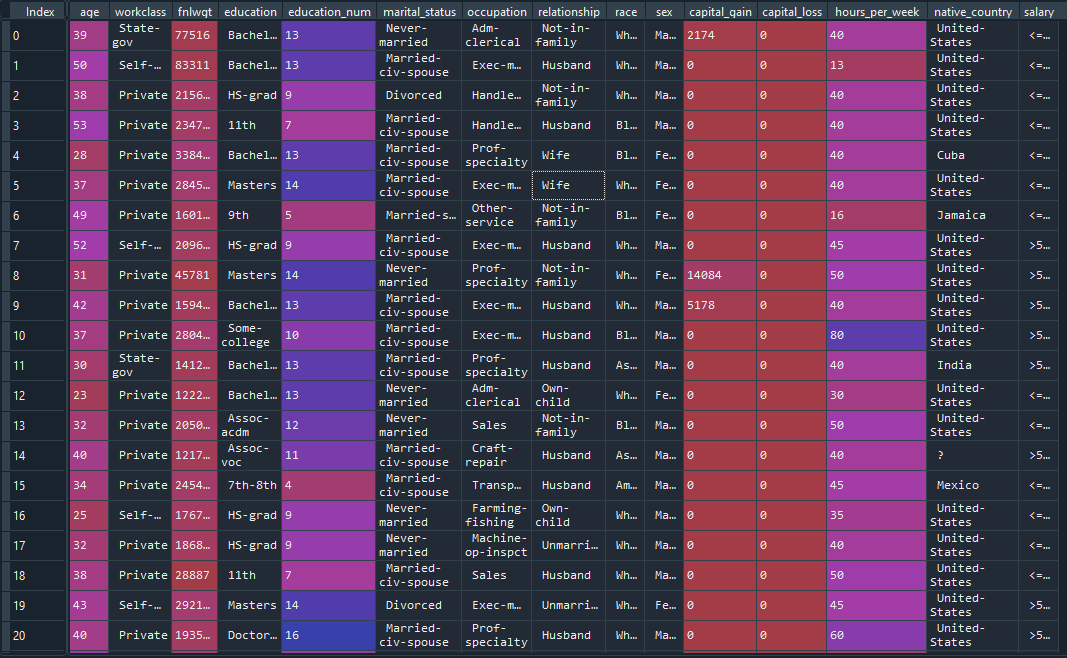
The implementation stage of any project is a true display of the defining moments that make a project a success or a failure. The implementation stage is defined as the system or system modifications being installed and made operational in a production environment. The phase is initiated after the system has been tested and accepted by the user. This phase continues until the system is operating in production in accordance with the defined user requirements.

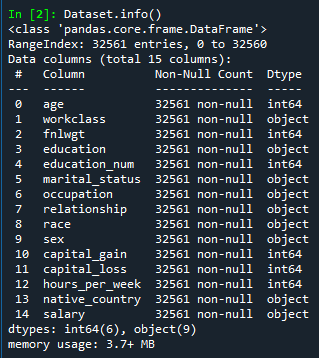
**4.1 The Dataset**

The data for our study was accessed from Kaggle Machine Learning Repository. The data set includes figures on 48,842 different records and 14 attributes for 42 nations. The 14 attributes 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. In this dataset, the target variable is Salary.

We aim to build classification models like Random Forrest, Decision Trees, K- Nearest Neighbors and others and also fine tune the parameters of the model. These models would be trained on a data set which will be engineered carefully after performing the feature engineering.

**4.2 DATA REPRESENTATION**





**4.3 Data Preprocessing**

Before processing the Adult Dataset, cleaning the data with certain preprocessing techniques becomes a necessity. This includes: 1) Handling Missing Values: The dataset contains certain set of missing values for categorical features, workclass, occupation, native-country which has been dealt with some algorithmic transformations applied to the data. The missing values are flexibly handled for every attribute by setting a default marker called ‘?’ and assigning a unique category for negating information loss. 2) Encoding of Categorical or Non-Numeric features: As all Categorical Features are non-numeric, encoding has been done in 2 stages: • Label Encoding: All categorical features are label encoded, where alphabetically each category is assigned numbers starting from 0. This is also done before running the Extra Trees Classifier Algorithm for efficient feature selection. • One-Hot Encoding: This involves splitting of different categorical features into its own categories where each and every category assumes a binary value i.e., 0 if it does not belong to that category and 1 if it belongs to that category. This is important for those categorical features where there exists no ordinal relationship in between them. One-Hot Encoding has been done for categorical features having more than 2 categories. Here, for all categorical features except sex attribute, all label encoded forms are transformed into One-Hot Encoded Forms. This is because sex attribute has only 2 categories i.e., male and female, which have been already represented in binary form in a single attribute and hence to avoid the curse of dimensionality, no One-Hot Encoding is done for sex attribute. 3) Shuffling: The whole dataset has been shuffled in a consistent way such that all the categories of different attributes remain included in Training Set and Validation Set. 4) Splitting: Now, the dataset is split into training and testing sets. With 80% of the data made available for training purposes and the rest 20% is used for testing.

.

**Importing Libraries**

Numpy, Matplotlib, and Pandas are required

****

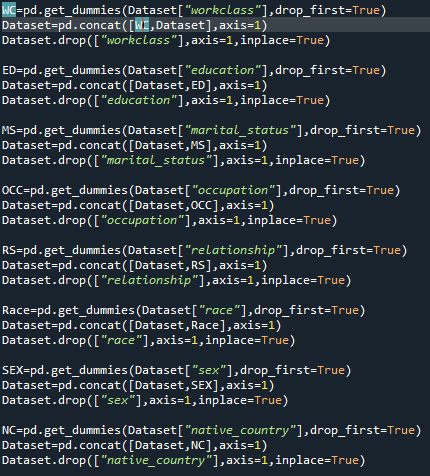
**Importing Dataset**

****

**Encoding Categorical Values**

Some features contain string values, which can’t be calculated in numpy library. Hence we need to convert them into numerical values.

Method: Label Encoding, One-Hot Encoding



**Encoding Independent values**

**Extracting features and labels**



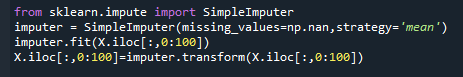
**Encoding Dependent variable**



**Taking care of missing values**

Some Features contain the missing values like NAN, ? or space.We need to Imputate missing value via feature mode.

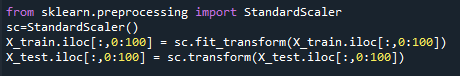
Method: Feature Mode, Median and Mean



**Splitting the dataset into training set and test set**



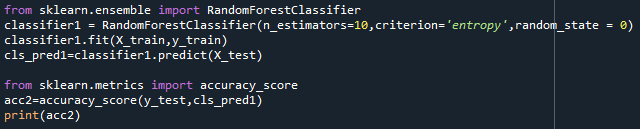
**Feature Scaling**



1. **Models**

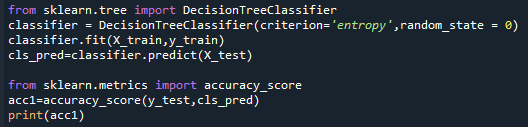
**Random Forest**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set.

****

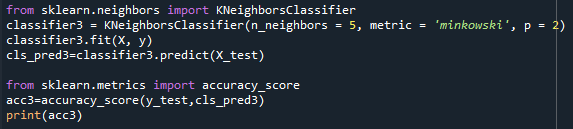
**Decision Tree**

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute

****

**K- Nearest Neighbors**

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

****

**6. Result**

**Accuracy :**

**Random Forest - 0.8233397312859885**

**Decision Tree - 0.8142034548944338**

**K- Nearest Neighbors - 0.769827255278311**

.

**7. Conclusion**

In the end, we got to know that Random Forest classifier is the most efficient of all.

And its accuracy is 0.82. Hence, we can predict the income level of adults upto an accuracy of eighty two percent.

**7.1 Future Scope**

The scope of Machine Learning is expanding across all fields such as banking and finance, information technology, media & entertainment, gaming, and the automotive industry. As the Machine Learning scope is very high, there are some of the areas where researchers are working toward revolutionizing the world for the future. They are

* 1. Automotive industry c) Quantum Computing
  2. Robotics d) Computer Vision

**8. Bibliography**

**[1]** <https://scikit-learn.org/stable/>

**[2]** <https://scikit-learn.org/stable/supervised_learning.html>