**ADULT CENSUS INCOME PREDICTION**

**PROJECT REPORT**

**Submitted by**

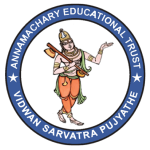
**LEGACY**

1. **A CHARAN SAI**
2. **K VENKATESWARA REDDY**
3. **V M UDHAYAKUMAR**
4. **K VINAY**
5. **A MAHESH**

***In partial fulfilment for the award of the Certificate***

**of**

**SUMMER INTERNSHIP PROGRAM**

**Department of Computer Science and Engineering**

**Annamacharya Institute of Technology and Sciences**

**Venkatapuram Village , Renigunta Mandal , Tirupati , Andhra Pradesh 517520**

**July 2019.**

### BONAFIDE CERTIFICATE

This is to certify that the project entitled ”**ADULT CENSUS INCOME PREDICTION**” submitted by **A. Charan Sai, K. Venkateswara Reddy, V.M. UdhayaKumar, K.Vinay, A.Mahesh** in partial fulfilment for the requirements for the award of internship certification in technologies of Machine learning and Deep learning is an authentic work carried out by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the project report has not been submitted to any other University/Institute for the award of any Degree or Diploma.

### Signature of Supervisor                                       Signature of Head of the Department

 Akshay Kumar Kothuri, M.Tech       Mrs.B.Rupa Devi M.Tech,Ph.D.,

AI & IoT Developer Assistant Professor & HOD,

SmartBridge Educational Services PVT LTD Department of CSE,

Hyderabad – 500096 AITS,Tirupati.

INDEX

1. Adult Census Income Prediction

1.1 Introduction

2. Review of Literature

2.1 Objective of Research

2.2 Problem Statement

3. Data Collection

4. Methodology

4.1 Exploring Data Analysis

4.1.1 Figures and Tables

4.1.2 Data Preparation and Cleaning

4.2 Statistical Techniques and Visualizations

4.2.1 Feature Study and Selection:

4.3 Data Modelling and Visualizations

5. Findings and Suggestions

6. Conclusion

ADULT CENSUS INCOME PREDICTION

USING RANDOM FOREST

1.INTRODUCTION

The prominent in equality of wealth and income is a huge concern especially in the United States. ... 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.

In this study Random Forest Classifier machine learning algorithm is applied to predict income levels of individuals based on attributes including age, education, marital status, gender, occupation, country and others. Income levels are defined as a binary variable 0 for income <=50K/year and 1 for >=50K/year.

2.LITERATURE REVIEW

2.1The Main Objectives of Adult census income:

* To predict the income of an Individual
* To know the individual economical status
* Based on there profits and loss they can get income in >50 or <50

2.2PROBLEM STATEMENT:

The data is acquired from UCI Machine Learning Repository and includes 32,561 individual data on 13 attributes based on 1994 census database. Random forest classifier is used since it gave better accuracy compared to Random forest classifier. The predictive accuracy of the model on test data is 85%. Important features prediction shows marital status, capital gain, education, age and hours per week are the top features which account for larger shares of the model accuracy. Using decision tree classifier also shows that these variables are the top 5 features in importance.

3.DATA COLLECTION

The data for the project was accessed from the UCI Machine Learning Repository (<https://www.kaggle.com/uciml/adult-census-income#adult.csv> ). The data is extracted by Barry Becker using 1994 census database. The data set includes figures on 32,561 observations and 13 attributes for 42 countries. The target variable in the data set is income level which shows whether a person earns more than 50K (K denotes thousands, 50K equals 50,000) per year or not based on a set of different features. There are 12 features containing information on education, gender, nationality, marital status, occupation, work classification, gender, race, work hours, capital loss and capital gain.

4. METHODOLOGY

4.1 Exploring Data Analysis:

* The Census Income dataset has 32,561 entries. Each entry contains the following information about an individual:
* **Age**: the age of an individual
* Integer greater than 0
* **Workclass**: general term to represent the employment status of an individual
* Private, Self­emp­not­inc, Self­emp­inc, Federal­gov, Local­gov, State­gov, Without­pay, Never­worked”.
* **FinalWeight**:  final weight. In other words, this is the number of people the census believes the entry represents.
* Integer greater than 0
* **Education​**: the highest level of education achieved by an individual.
* Bachelors, Some­college, 11th, HS­grad, Prof­school, Assoc­acdm, Assoc­voc, 9th, 7th­8th, 12th, Masters, 1st­4th, 10th, Doctorate, 5th­6th, Preschool
* **Education­Num**: the highest level of education achieved in numerical form.
* Integer greater than 0 ​
* **Marital­Status**:Marital status of an individual. Married­civ­spouse corresponds to a civilian spouse while
* Married­AF­spouse is a spouse in the Armed Forces for is set as ‘0’.
* Never-married, Separated, Widowed, Divorced is set as ‘1’
* **Sex**: To know the gender of the individual. If the gender is male then it set as ‘0’else set as the gender male as ‘1’.
* **Capital Gain**: The Capital Gain of an individual.
* Integer greater than 0 ​
* **Capital Loss**: The Capital Loss of an individual.
* Integer greater than 0 ​
* **Hours Per Week**: The working hours of an individual.
* Integer greater than 0 ​

4.1.1 Figures and Tables:

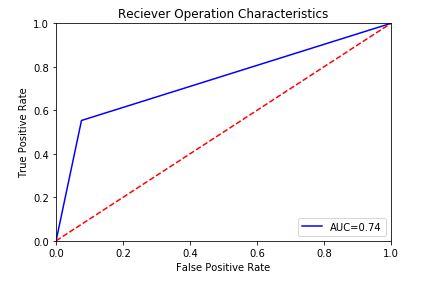


Fig1.1.1: ROC Curve

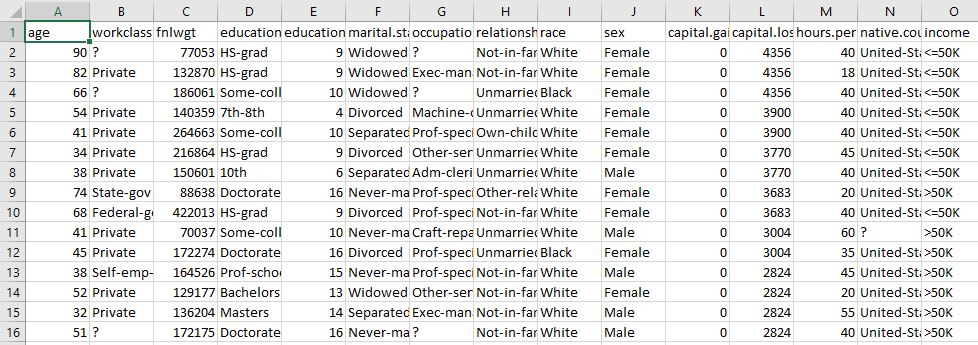


Table 4.1.1: Data Set

4.1.2 Data Preparation and Cleaning:

The following data preparation tasks are conducted to make the data suitable for running the machine learning model (Decision Tree classifier)

* Converting categorical (text) values into dummy variables: most of the variables are categorical (text) except capital gain, capital loss, hours per week, and years of education which are numeric. The categorical variables are transformed into dummy variables.
* Dropping unnecessary columns: a variable containing information on the sample weight of the individuals is dropped from since it is not required for the analysis like work class, education, Occupation, race, native country.
* Checking for null values and preparing separate features and target data frames: After doing all the above data cleaning steps, separates data frames for the feature and target data are generated.

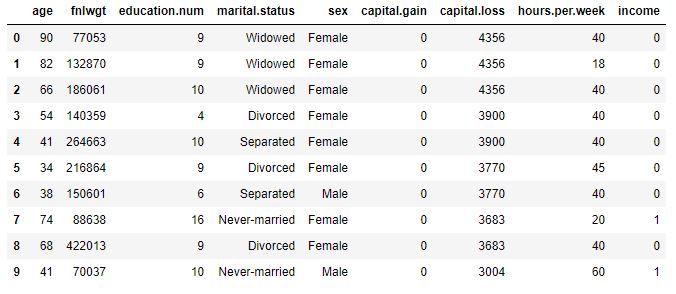


Table 4.1.2: Cropped Dataset

4.2 Statistical Techniques and Visualizations:

A supervised machine learning approach of Random Forest Classifier is used for the study. Random forest classifier is chosen due to two reasons. First since the outcome (target) variable is binary variable (income level >50K or not), using classification algorithms is better than regression algorithms. This is because the target having only values of 0 and 1, regression algorithms will perform less due to less variation in the target variable.

Secondly, random forest classifier is found to have better accuracy score compared to gaussian Random forest and Random forest Classifier gave accuracy score of 83% while ‘Gaussian NB’ gave accuracy of 74%. Random forest is preferred to decision tree since using results from many decision trees will avoid the overfitting problem associated with using a decision tree classifier. The results from the model show that both random forest and decision tree gave similar results. This is shown by the fact that the top 7 features in importance are the same in the two models.

The data for our study was accessed from the University of California Irvine (UCI) Machine Learning Repository [8]. It was actually extracted by Barry Becker using the 1994 census database. The data set includes figures on 48,842 different records and 14 attributes for 42 nations. The 14 attributes

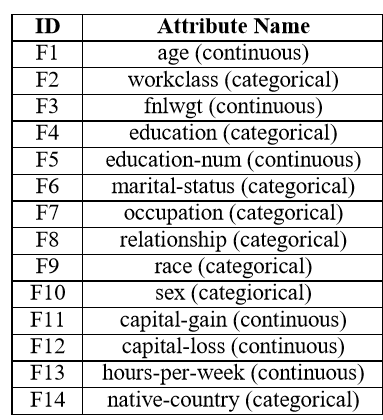


Table4.2.1: Features and extra trees Score

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 shown in Table 1. 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.

**4.2.1 Feature Study and Selection:**

Based on the scores of the Extra Tree Classifier for different attributes (as shown from Table 1) the most relevant features have been selected, that are going to be implemented in our model. A visual explanation of the Extra Trees Classifier or Extremely Randomized Trees is shown in Fig 1. As a result Features F9 (race) and F14 (native-country) have been eliminated as they have the least Extra Trees Classifier Scores.

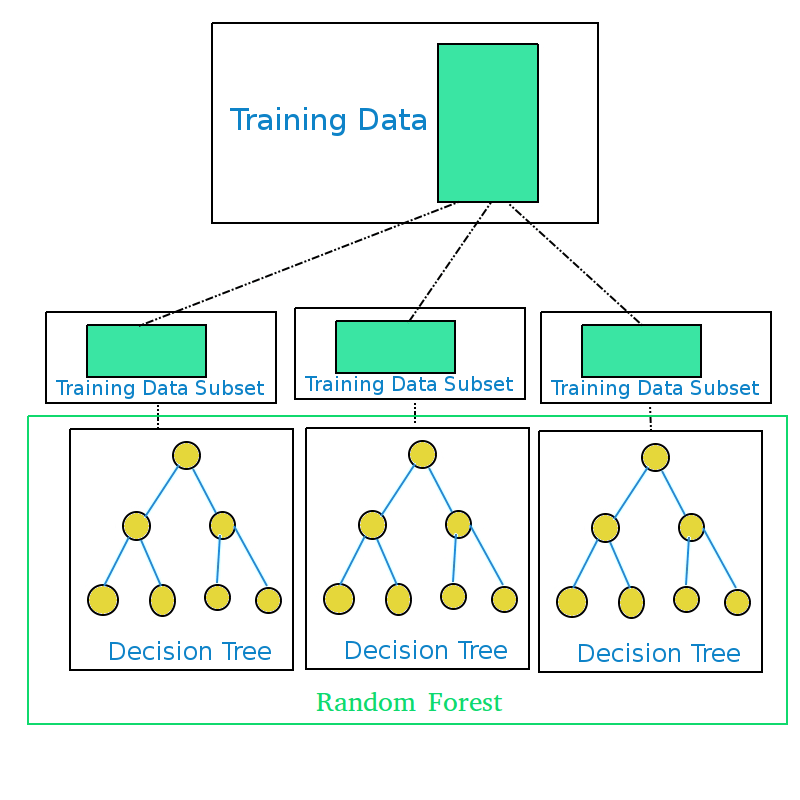


Fig4.2.1: Representation of Random Forest Model

A Correlation Matrix is shown in Fig 2, in the form of a Heat- Map showing Feature-to-Feature and Feature-to-Label Pearson Correlations where all the features are Continuous Variables.

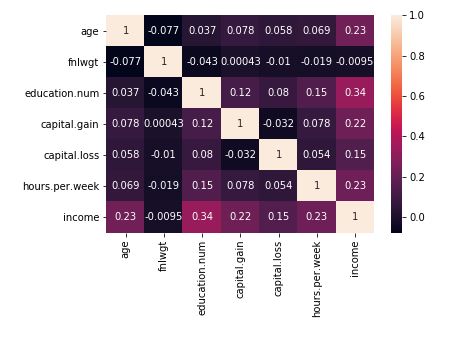


Fig 4.2.2. Heat-Map showing Feature-to-Feature and Feature-to-Label’s Pearson Correlation Coefficients

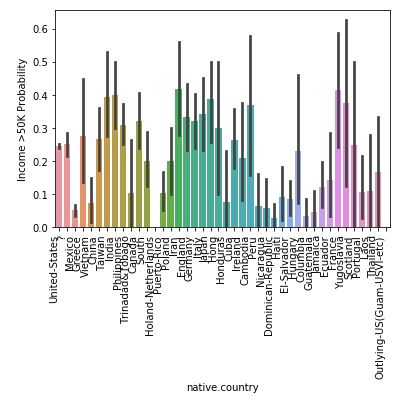


Fig4.2.3: Relation between Income and Native Country

Data Visualization has been done using Box and Whisker Plots of all continuous features to clearly understand the measures of their central tendencies.

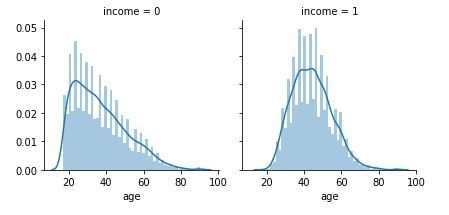


Fig4.2.4: Relation between Income and Age

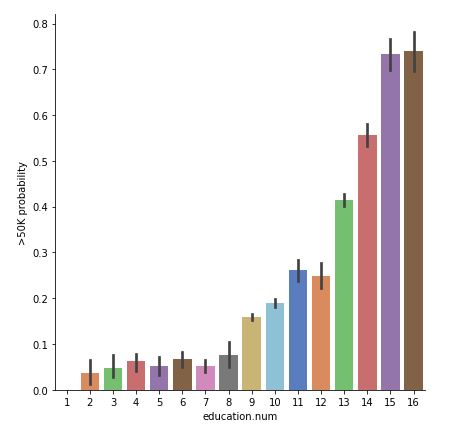


Fig4.2.5: Relation between Income and Educations

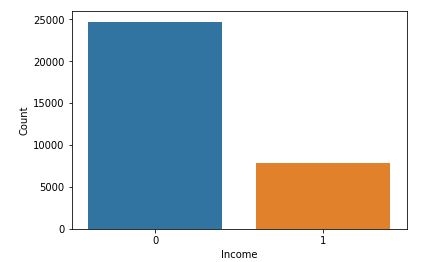


Fig4.2.6: Relation between Income above 50k and below 50k

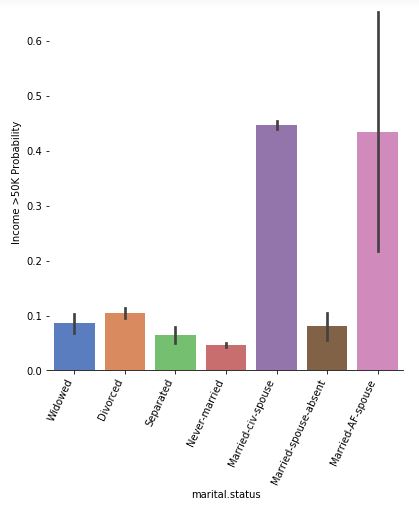


Fig4.2.7: Relation between Income and Marital Status

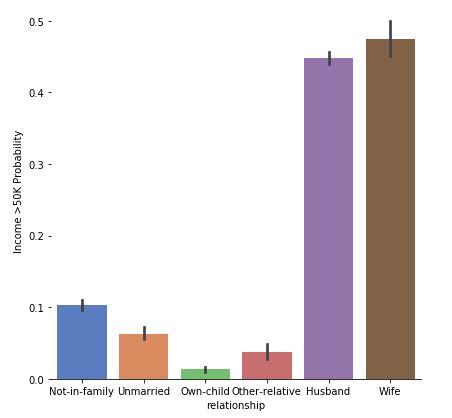


Fig4.2.8: Relation between Income and Relationship

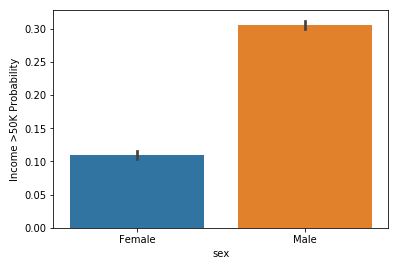


Fig4.2.9: Relation between Income and Sex

4.3 Data Modelling and Visualization:

Node-RED is a flow-based development tool for visual programming developed originally by IBM for wiring together hardware devices, APIs and online services as part of the Internet of Things. Node-RED provides a web browser-based flow editor, which can be used to create JavaScript functions. Elements of applications can be saved or shared for reuse. The runtime is built on Node.js. The flows created in Node-RED are stored using JSON. Since version 0.14 MQTT nodes can make properly configured TLS connections. In 2016, IBM contributed Node-RED as an open source JS Foundation project.

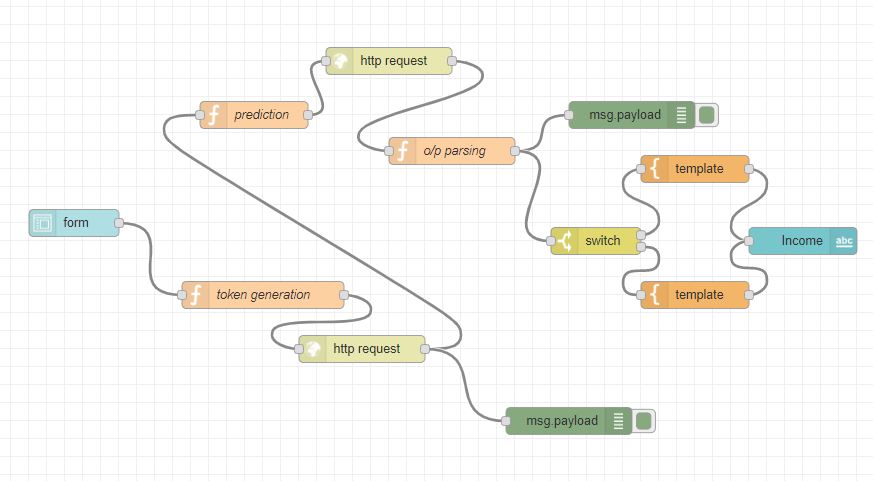


Fig 4.3.1: Node Red Flow

5. FINDINGS AND SUGGESSTIONS

Decision tree classifier is applied to the data to answer the following questions:

* What is income level of an individual with certain attributes?

Assessing whether an individual with certain attributes (age, education, sex, marital status and others) will earn higher or lower income levels. A higher income level is defined as an income level >50K/year while lower income level is defined as income <=50K/year. A function that predicts income level based on the fitted model and individual attribute is provided in the ‘ipython’ notebook.

* What are the key features determining income level?

Identifying what are the top 5 features explaining much of the difference between low- and high-income levels. Determining the key features can help in policy formulation by identifying the few factors that can give most of the gains in income.

6.Conclusion:

The conditional inference tree was more effective at classifying the Over Fifty column because the prediction of the dataset is more effective when based on a permutation of variables. The conditional inference tree seems to avoid the bias that decision trees can have due to the information gain of individual variables. Instead the condition inference tree selects the variable to split on based on the statistical significance of a permutation of variables. The applications of this prediction model are mostly towards understanding which qualities lead individuals to having an above average income. From a government standpoint, the information can be used to recognize which aspects of society need to have policy implementation in order to improve equity. For example, the graph demonstrating that more men get paid more than 50K when measured against women indicates that there may be an unequal professional environment between genders.