**PROFESSIONAL TRAINING REPORT**

**at**

**Sathyabama Institute of Science and Technology**

**(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the

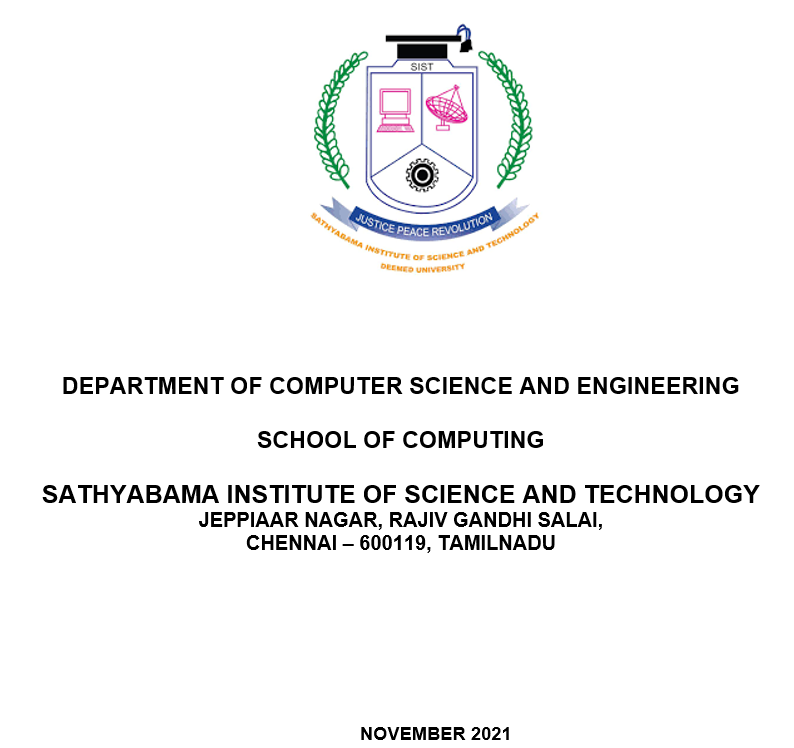
award of Bachelor of Engineering Degree in Computer

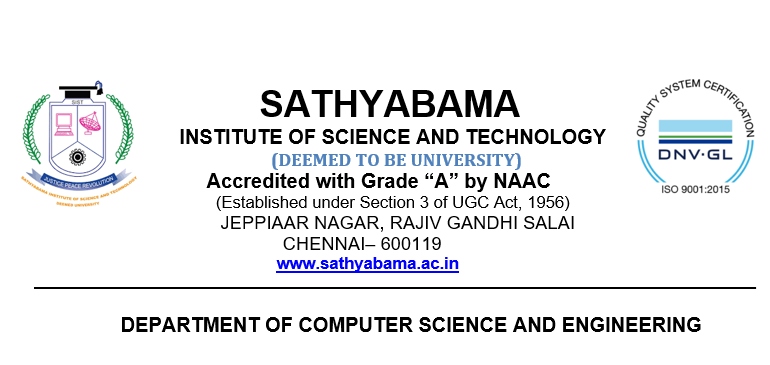
Science and Engineering

BY

SHAIK SHEERU ALI MOULALI

REG. NO. 39110932



 **BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **SHAIK SHEERU ALI MOULALI (Reg. No: 39110932)** who carried out the project entitled “**Adult Income Dataset based on Machine Learning**” under my supervision from June 2021 to November 2021.

## **Internal Guide**

## **Dr. J. Albert Mayan, M.E., Ph.D.,**

## 

## **Head of the Department**



## **Submitted for Viva voce Examination held on**

**Internal Examiner External Examiner**

**DECLARATION**

I SHAIK SHEERU ALI MOULALIhere by declare that the project report entitled “**Adult Income Dataset using MACHINE LEARNING”** done by me under the guidance of **Dr. J. Albert Mayan** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

DATE:

PLACE:

SIGNATURE OF THE

CANDIDATE

**ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. J. Albert Mayan, M.E., Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project

**TRAINING CERTIFICATE**



# **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 Gradient Boosting Classifier Model was deployed which clocked the highest accuracy of 88.16%, eventually breaking the benchmark accuracy of existing works

**TABLE OF CONTENTS**

**CHAPTER No**. **TITLE**  **PAGE No**

ABSTRACT 6

LIST OF FIGURES 9

LIST OF TABLES 10

LIST OF ABBREVIATIONS 11

**1.** **INTRODUCTION 12**

1.1. INTRODUCTION TO THE PROJECT 12

1.2. OUTLINE OF THE PROJECT 12

1.2.1 LIST OF ATTRIBUTES 13

1.3. LITERATURE REVIEW 14

**2. AIM AND SCOPE OF THE PRESENT INVESTIGATION** **16**

2.1. AIM 16

2.2. OBJECTIVES OF RESEARCH 16

2.3. PROPOSED SYSTEM 16

2.3.1. DATA COLLECTION 16

2.3.2. EXPLORING THE DATA SET 17

2.3.3. TESTING THE DATASET 17

**3.** **ALGORITHMS AND METHODS**  **21**

3.1. SYSTEM REQUIREMENTS 21

3.1.1. SOFTWARE REQUIREMENTS 21

3.2. ARCHITECTURE OF MODEL 21

3.3. EXPLANATION OF ALGORITHM 21

3.3.1. Data Preprocessing 21

3.3.2. Logistic Regression Algorithm 24

3.4. IMPLEMENTATION 25

**4.**  **RESULTS AND DISCUSSION** **27**

4.1. RESULTS 27

4.1.1. PERCENTAGE OF DATASET 27

4.2. ELUCIDATION OF FACTORS 27

**5.** **CONCLUSION AND FUTURE WORK** **30**

5.1. CONCLUSION 30

5.2. FUTURE ENHANCEMENTS 31

**REFERENCES 32**

**APPENDIX** **34**

A. SCREENSHOTS 34

B. SOURCE CODE 36

**LIST OF FIGURES**

**FIGURE NO FIGURE NAME PAGE No**

1.1 DOWNLOADED DATASET 17

2.1 DISTRIBUTION OF VALUE FEATURES 18

2.2 DISTRIBUTION OF TARGET VALUE (SALARY) 20

3.1 ARCHITECTURE OF THE SYSTEM 21

3.2 PSEUDOCODE (DATA PREPROCESSING) 23

3.4 PSEUDOCODE (LOGISTIC REGRESSION) 25

4.1 RESULT FOR COMPLETE DATASET 27

**LIST OF TABLES**

**TABLE NO. TABLE NAME PAGE NO.**

5.1 COMPARITIVE ANALYSIS 30

**LIST OF ABBREVIATIONS**

**ABBREVIATION EXPANSION**

PCA Principal Component Analysis

AUPRC Area Under Precision Recall Curve

AID Adult Income Dataset

AGA Artificial Genetic Algorithm

STA Streaming Transaction Data

LO Local Outlier Factor

TP True Positive

TN True Negative

FP False Positive

FN False Negative

**CHAPTER 1**

**INTRODUCTION**

1.1 INTRODUCTION TO THE PROJECT

The Adult Census Income data was extracted from the 1994 Census bureau database. The purpose of this study is to get the insight about the income of the persons from 1994 in a particular reign and can infer about the income inequality if there's any from the data set and to successfully predict whether a person make $50K a year or not.

This Adult Census data set contains the features about the people who earns more than $50K or less than $50K a year. Some features are education of the person, their nationality and age. In this project the Exploratory data analysis is done, and various insights are concluded from EDA. The feature engineering is done to prepare the data for the Machine Learning models and the feature importance bar graph is plotted for the perspicacity of the features, then the Classification models are trained, and the performances are compared with various performance metrics.

The classification algorithms that are used are Logistic regression, Random Forest Classifier, Gradient Boosting, Bernoulli Naive Bayes, and the Support vector Classifier. The Best performing model is finalized for the predictions.

1.2. OUTLINE OF THE PROJECT

The Adult Census data set is of shape (32561,15) i.e., It has 15 features and 32,561 entries. The features are age, work class, education, education.num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, fnlwgt and income. Income is the target Variable that is to be predicted that has two categorical values ">50K" and "<=50K". The education and education.num

are related as the education.num is the numerical conversion of the education column which consist of categorical values. The relationship features tells the relationship status about the person if they are Unmarried or married in various category. The race feature is used to tell the race of the person like black, white, Asian etc. The Age column has values between 16 and 100. The fnlwgt (final\_weight) feature is the weight on the Current Population (CPS) files are controlled to independent estimates of the civilian non institutional population of the US. These are prepared monthly by Population Division at the Census Bureau. Three sets of controls are used for this these are

1. A single cell estimate of the population 16+ for each state.

2. Controls for Hispanic Origin by age and sex.

3. Controls by Race, age, and sex.

Some of the currently used approaches to detection of such datasets are:

Artificial Neural Network, Fuzzy Logic, Genetic Algorithm, Logistic Regression,

Decision tree, Support Vector Machines, Bayesian Networks, Hidden Markov

Model, K-Nearest Neighbour, etc.,

1.2.1 List of Attributes :

>50K,<=50K.  
age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.  
education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.  
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  
sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran,Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad & Tobago, Peru, Hong, Holand-Netherlands.

1.3. LITERATURE REVIEW

Certain efforts using machine learning models have been made in the past by researchers for predicting income levels.

• Chockalingam et. al. [1] explored and analysed the Adult Dataset and used several Machine Learning Models like Logistic Regression, Stepwise Logistic Regression, Naive Bayes, Decision Trees, Extra Trees, k-Nearest Neighbor, SVM, Gradient Boosting and 6 configurations of Acti- vated Neural Network.

They also drew a comparative analysis of their predictive performance

• Bekena [2] implemented the Random Forest Classifier algorithm to predict income levels of individuals.

• Topiwalla [3] made the usage of complex algorithms like XGBOOST, Random Forest and stacking of models for prediction tasks including Logistic Stack on XGBOOST and SVM Stack on Logistic for scaling up the accuracy.

• Lazar [4] implemented Principal Component Analysis (PCA) and Support Vector Machine methods to generate and evaluate income prediction data based on the Current Population Survey provided by the U.S. Census Bureau.

• Deepajothi et. al. [5] tried to replicate Bayesian Networks, Decision Tree Induction, Lazy Classifier and Rule Based Learning Techniques for the Adult Dataset and presented a comparative analysis of the predictive performances

• Lemon et. al. [6] attempted to identify the important features in the data that could help to optimize the complexity of different machine learning models used in classification tasks.

• Haojun Zhu [7] attempted Logistic Regression as the Sta- tistical Modelling Tool and 4 different Machine Learning Techniques, Neural Network, Classification and Regression Tree, Random Forest, and Support Vector Machine for predicting Income Levels.

**CHAPTER 2**

AIM AND SCOPE OF PRESENT INVESTIGATION

2.1 AIM

The main aim of the project is to design and develop the dataset is to classify people earning <=50k or >50k based on several explanatory factors affecting the income of a person like Age, Occupation, Education, etc.

The methods we intend to use are:

* Binary Logistic Regression
* Decision Tree
* Random Forest

2.2 OBJECTIVES OF RESEARCH

➢ To study the given data.

➢ To apply data cleaning methods to remove unknown data from the given data set.

➢ To apply machine learning algorithms on the given data set to produce a model with maximum efficiency.

➢ To compare the different algorithms and draw conclusions from the developed model.

➢ Predict which model is suitable for detecting Adult Income Dataset to reduce several explanatory factors affecting the income of a person like Age, Occupation, Education, etc.

2.3. PROPOSED SYSTEM

2.3.1. DATA COLLECTION

Extraction was done by Barry Becker from the 1994 Census database. The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database. Prediction task is to determine whether a person makes over 50K a year.

The Census Income dataset has 48,842 entries. Each entry contains the following information about an individual:

Text

Description automatically generated

Fig 2.1: Downloaded dataset

2.3.2 EXPLORING THE DATASET

The Adult dataset is a widely used standard machine learning dataset, used to explore and demonstrate many machine learning algorithms, both generally and those designed specifically for imbalanced classification.

First, download the dataset and save it in your current working directory with the name “adult\_dataset.xlsx”

Download Adult Dataset (adult\_dataset.xlsx)

2.3.3. TESTING THE DATASET

Machine learning is breaking grounds in numerous fields including Finance. What if we could use Machine Learning models to identify incomes of individuals? I found just the right dataset for this, called Census Income Dataset. I used the information in the dataset to predict if someone would earn an income greater than $50K/yr.

I collected the data from the UCI Machine Learning repository, and then interpreted each feature individually. I used several Machine Learning models and concluded with an Accuracy of 84.27% and F1 Score of 0.65 with Gradient Boosting Classifier and maximum Area Under Curve of 0.90.

(# distribution of age, fnlwgt, education.num, capital-gain, capital-loss, hours-per-week and salary column )

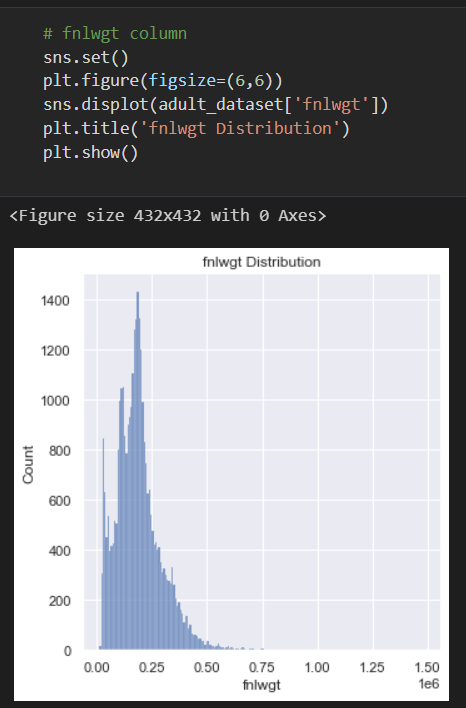


Fig 1: distribution of Age values

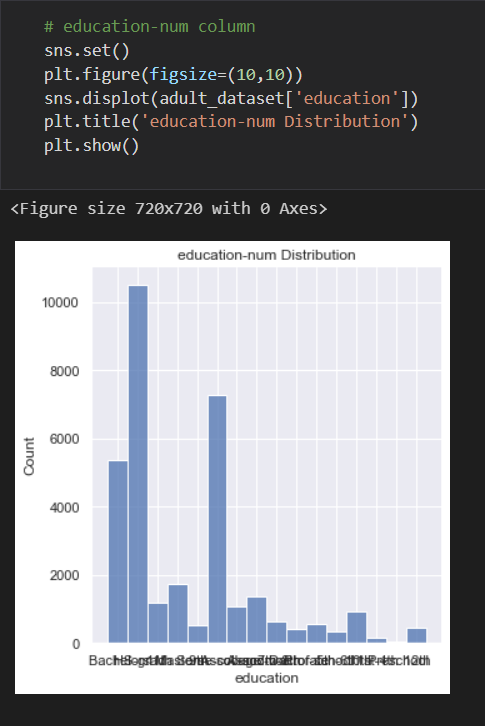


Fig 2: distribution of education-num class

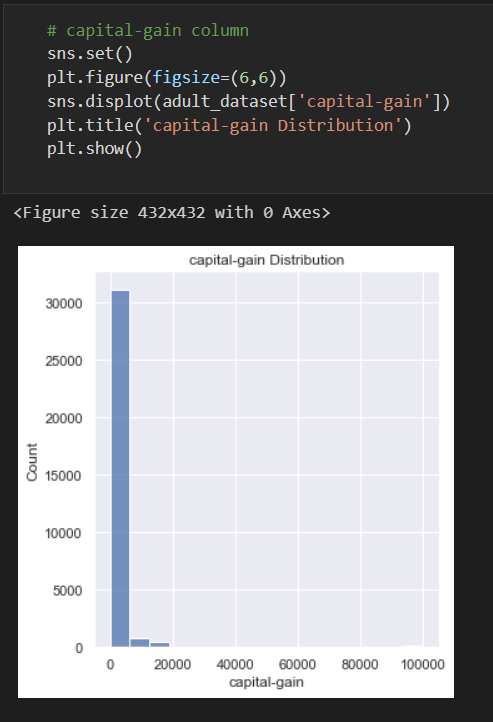
****

Fig 3: capital-gain column

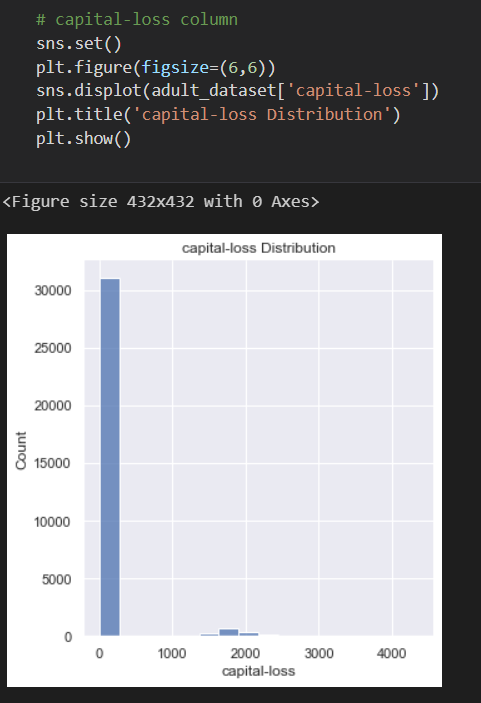


Fig 4: capital-loss column

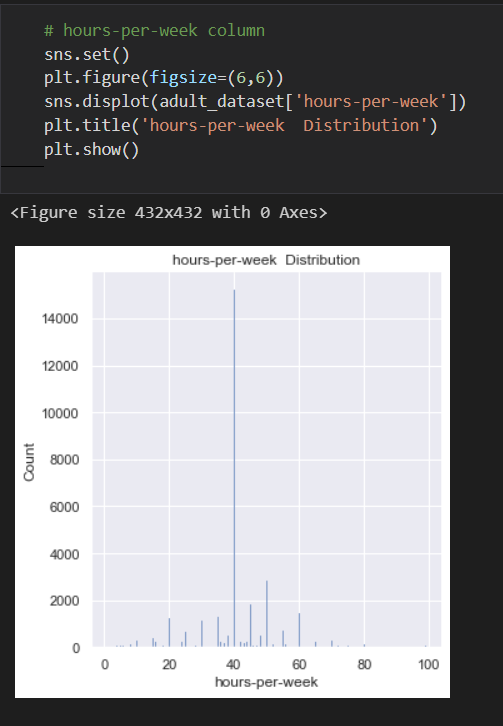


fig 5: hours-per-week column

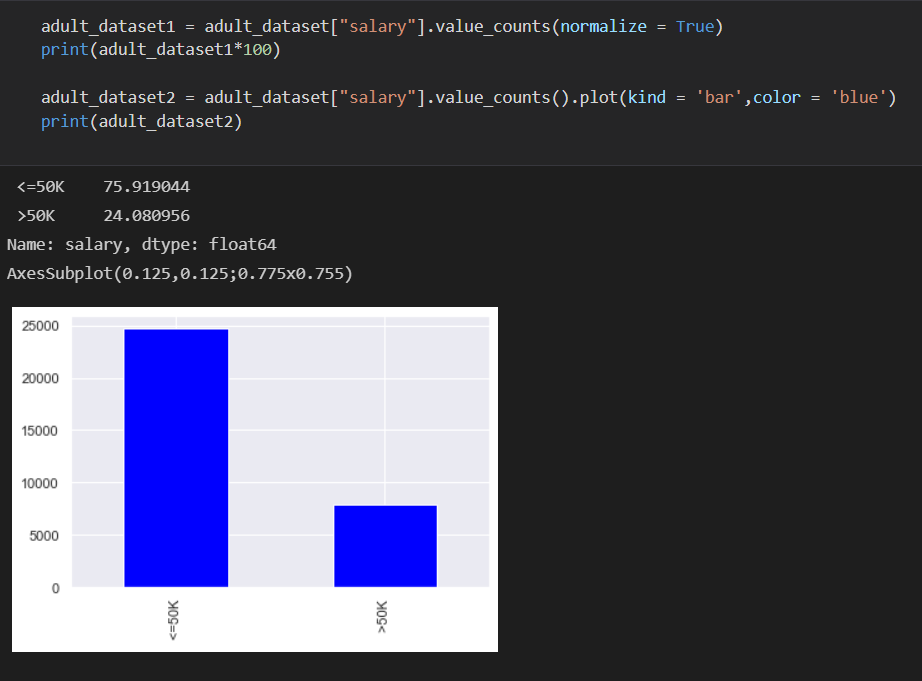


fig 6: salary

**CHAPTER 3**

ALGORITHMS AND METHODS

3.1.SYSTEM REQUIREMENTS

3.1.1 SOFTWARE REQUIREMENT

• Operating System - Windows OS

• Coding Language - Python

• Tool - Jupyter Notebook

• Libraries - pandas, matplotlib, sklearn

3.2. ARCHITECTURE OF MODEL

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities, called outliers. When looked at in detail on a larger scale along with real life elements, the full architecture diagram can be represented as follows:

**Diagram

Description automatically generated**

Fig 3.1 Architecture of the System

3.3 EXPLANATION OF ALGORITHM

The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together: This data is fit into a model and the following outlier detection modules are applied on it:

• Data Preprocessing

• Logistic Regression

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection.

This free and open-source Python library is built using NumPy, SciPy and matplotlib modules which provides a lot of simple and efficient tools which can be used for data analysis and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. We’ve used Jupyter Notebook platform to make a program in Python to demonstrate the approach that this paper suggests. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files.

3.3.1 DATA PREPROCESSING

Data preprocessing in the Adult Dataset, cleaning the data with certain preprocessing techniques becomes a necessity. This includes:

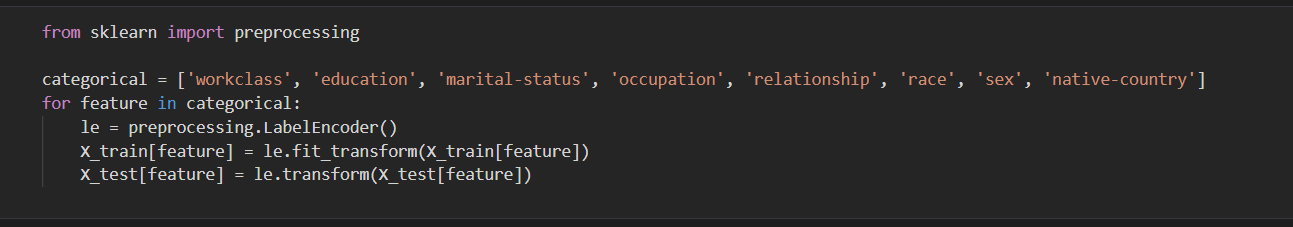
1) Handling Missing Values

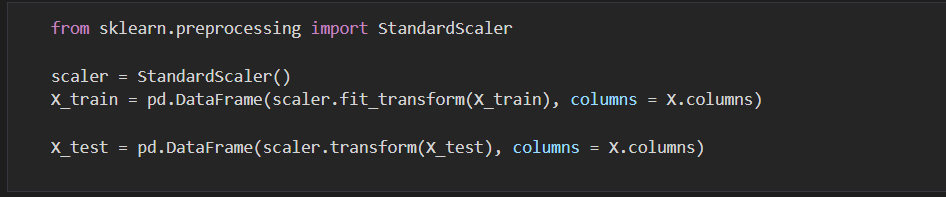
2) Encoding of Categorical or Non-Numeric features

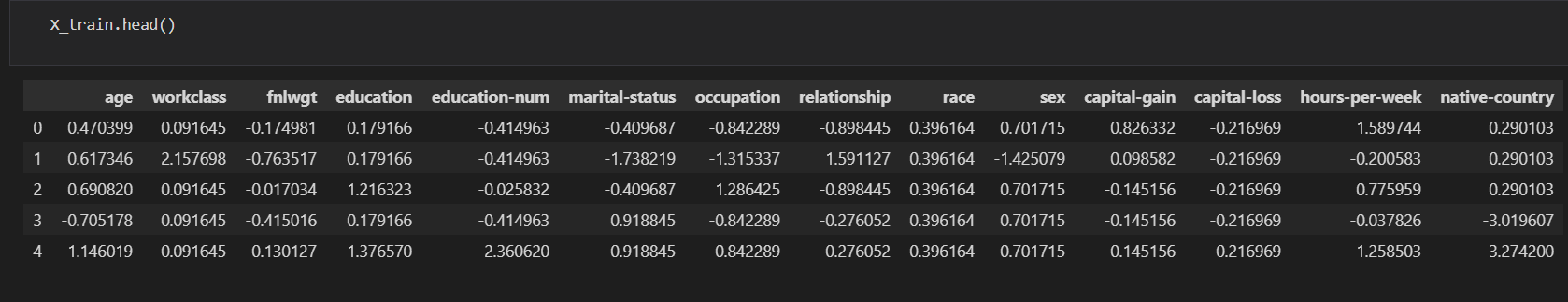
The pseudocode for this algorithm is written as:

Text

Description automatically generated







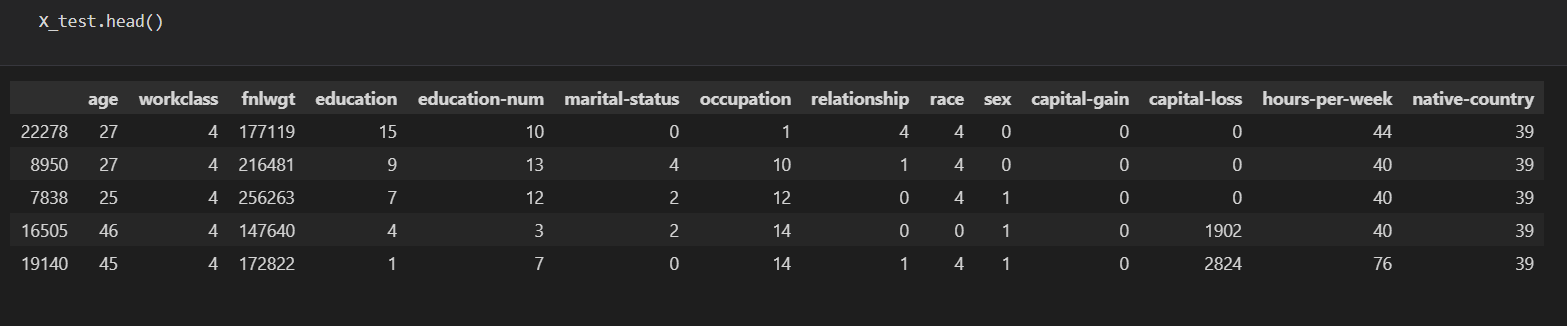


Fig 3.2: Pseudocode (DATA PREPROCESSING)

3.3.2 Logistic Regression

Logistic Regression is one of the easiest and most commonly used supervised Machine learning algorithms for categorical classification. The basic fundamental concepts of Logistic Regression are easy to understand and can be used as a baseline algorithm for any binary (0 or 1) classification problem. For this use-case, we are going to choose Logistic Regression as our classification model as it would be a good start for any beginner to start with a simple yet popular algorithm.

Logistic Regression is a Statistical predicting model that can predict either a ‘Yes’(1) or ‘No’(0). It is based on a Logit or Sigmoid function which ranges between 0 and 1. Now don’t get scared by hearing a Mathematical function. Let me explain it to you in simple terms.

A picture containing graphical user interface

Description automatically generated

If we observe here, the sigmoid function is an ‘S’ shaped curve that extends between values 0 and 1 which is plotted based on the output of the function when it is fed with a set of data. We can see the sigmoid function equation is 1 divided by 1 plus e raised to -z where is e is a mathematical constant called Euler’s Number.

A picture containing shape

Description automatically generated

So, the output of the equation should be something between 0 and 1. if the value of z goes to positive infinity value the predicted value will be 1 or if it goes to negative infinity it goes to 0. If you feed the sigmoid function with the numerical data

we have we will get a value which is either less than 0.5 i.e., classified as 0 or more than 0.5 which is classified as 1.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Fig 3.4: Pseudocode (Logistic Regression)

3.4. IMPLEMENTATION

Several machine learning techniques are employed to predict whether an observation makes greater than $50,000 a year or not. In a similar manner as logistic regression, the original data set is still split into a training set and a testing set. Models are trained on the training set and validated on the testing set.

First, a neural network (NN) with one hidden layer is trained. Although there are only 9 input variables, many of them are categorical and would result several dummy variables. Thus, the number of hidden nodes is chosen to be 40. Least square method is used to optimize the objective function. Maximum number of iterations are set to 5,000.

The prediction result of NN has an accuracy of 83.28%, and a misclassification rate of 16.72%.

Next, a classification and regression tree (CART) is grown on the training set. The CART model is obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. Although it is generally named CART, the tree grown in this case is actually a classification tree. The prediction result of CART has an accuracy of 82.94%, and a misclassification rate of 17.06%.

Random forest (RF) is another powerful machine learning tool. It improves predictive accuracy by generating a large number of bootstrapped trees. Final predicted outcome is attained by combining the results across all of the trees.

The prediction result of RF has an accuracy of 83.65%, and a misclassification rate of 16.35%.

Last, I use support vector machine (SVM) to predict the income level. SVM is a discriminative classifier that constructs a hyperplane in a high dimensional space used for classification.

The prediction result of SVM has an accuracy of 83.06%, and a misclassification rate of 16.94%.

**CHAPTER 4**

RESULTS AND DISCUSSION

4.1. RESULTS

Out of a total of 48,842 instances present in the dataset, 39,074 instances have been used for training while the rest 9,768 instances have been reserved for testing. After complete evaluation, the model performance are evaluated on the following metrics:

• The Training Accuracy describes the accuracy achieved on the Training Set.

From the model, a Training Accuracy of 88.73% is achieved.

• The Validation Accuracy describes the accuracy achieved on the Validation Set.

From the model, a Validation Accuracy of 88.16% is obtained.

• The Sensitivity or Recall is defined as the fraction of correctly identified positives.

4.1.1. PERCENTAGE OF DATASET

Calendar

Description automatically generated

Fig 4.1: Result for complete dataset

18

4.2. ELUCIDATION OF TERMS OF RESULT

• Accuracy: Accuracy is the percentage of correctly classified instances. It is one of the most widely used classification performance metrics.

Accuracy = Number of correct predictions /Total Number of predictions

(OR)

For binary classification models. The accuracy can be defined as:

Accuracy= TP+TN / TP+TN+FP+FN

## Confusion Matrix

The most common type of metric available to us is the confusion matrix, which is also called the confidence matrix. The confusion matrix is a matrix that looks like:

Table

Description automatically generated

What we can see from above is that the confusion matrix is a matrix between actual values vs predicted values. It is generally used for classification purposes, where it is necessary to predict the target as a 1 or 0. When we observe the actual value as absent, we give it a 0, and 1 otherwise. The same is done for predicted values as well. we can tell a lot of things from this matrix, such as:

Our confusion matrix looks like:

Text

Description automatically generated

19

• Precision and recall: Precision is the number of classified Positive or fraudulent instances that actually are positive instances.

Precision = Tp/(Tp+Fp)

• Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions. Recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall = Tp / (Tp + Fn)

• F1 score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

• Support: The support is the number of samples of the true response that lie in that class. Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn’t change between models but instead diagnoses the evaluation process.

**CHAPTER 5**

CONCLUSION & FUTURE ENHANCEMENTS

5.1 CONCLUSION

In this project we have presented several models and how to implement them on a classifier dataset, addressing several issues like feature engineering, data imbalance followed by model building and hyper-parameter tuning to ultimately boosting accuracy of models.

We could clearly see data is imbalanced but our prediction capabilities got diminished when we tried to oversample the minority, hence continued normally.

Ultimately the best performers were Logistic Regression, random forest and xgboost. Random Forest performed best when variable country wasn’t included. This is because the variable country is biased. XGBOOST on the other hand performed amazingly especially upon hyperparameter tuning. Stacking in this case didn’t scale accuracy much but was useful to see that it increased our AUC.

Table

Description automatically generated

Tab 5.1. comparative analysis

5.2 FUTURE ENHANCEMENTS

While we couldn’t reach out goal of 100% accuracy in adult income data detection, we did end up creating a system that can, with enough time and data, get very close to that goal. As with any such project, there is some room for improvement here.

The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it.

However, the output of these algorithms needs to be in the same format as the 22 others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased.

Finally! we successfully created a classification Machine Learning prediction model using Python and its powerful libraries which predicts whether a given adult’s income will be >50K or not.

**REFERENCES**

1]UCI ADULT DATASET Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

2] SMOTE: Synthetic Minority Over-sampling Technique -Nitesh V. Chawla - Journal of Artificial Intelligence Research 16 (2002) 321–357

3][www.kaggle.com](http://www.kaggle.com)

4][www.stackoverflow.com](http://www.stackoverflow.com)

5] Vidya Chockalingam, Sejal Shah and Ronit Shaw: “Income Classification using Adult Census Data”, https://cseweb.ucsd.edu/classes/wi17/cse258- a/reports/a120.pdf.

6] Sisay Menji Bekena: “Using decision tree classifier to predict income levels”, Munich Personal RePEc Archive 30th July 2017

7] Mohammed Topiwalla: “Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting”, University of SP Jain School of Global Management.

8] Alina Lazar: “Income Prediction via Support Vector Machine”, International Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.

9] S.Deepajothi and Dr. S. Selvarajan: “A Comparative Study of Classification Techniques On Adult Data Set”, International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October 2012.

10] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/ jmcauley/cse190/reports/sp15/048.pdf.

11] Haojun Zhu: “Predicting Earning Potential using the Adult Dataset”, https://rstudio-pubs-static.s3.amazonaws.com/235617 51e06fa6c43b47d1b6daca2523b2f9e4.html

12] https://archive.ics.uci.edu/ml/datasets/Adult

13] https://medium.com/mlreview/gradient-boosting-from-scratch1e317ae4587

**APPENDIX**

A. SCREEN SHOTS

Graphical user interface, text, application, email

Description automatically generated

Open the jupyter notebook platform

Graphical user interface, application, email

Description automatically generated

Upload the Dataset

Graphical user interface, text

Description automatically generated with medium confidence

Explore the Dataset

Graphical user interface

Description automatically generated with medium confidence

Determine no. of Adult Income Salaries in the Dataset

Table

Description automatically generated

Output (Adult Income Dataset using ML algorithms)

B. SOURCE CODE

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

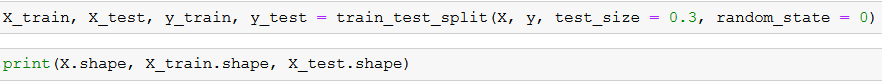
Description automatically generatedText

Description automatically generatedText

Description automatically generated

Text

Description automatically generatedGraphical user interface

Description automatically generated with low confidenceGraphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated with low confidenceText

Description automatically generatedText

Description automatically generated with medium confidence