|  |
| --- |
|  |
| Classification Modelling to Predict  Adult Income  Data Science Course 3 – Machine Learning  **Presented By Group 7**  Jing Tang Panthea Saffarzadeh  David Graham Ramila Mudarth  Maxim Smetanin Sivi Rakaj |
| |  |  |  | | --- | --- | --- | |  |  | [Course title] | |  |  |  | |  |  |  | |  |  |  | |  |  |  | |

Contents

[OBJECTIVE 1](#_Toc100258505)

[INTRODUCTION 1](#_Toc100258506)

[DATA PREPARATION 2](#_Toc100258507)

[DATA VISUALIZATION AND ANALYSIS (<<SIVI’s Graphs>>) 3](#_Toc100258508)

[MODELLING TECHNIQUES 6](#_Toc100258509)

[Data Transformation 6](#_Toc100258510)

[Ensemble Modelling 6](#_Toc100258511)

[Random Forest Classifier 7](#_Toc100258512)

[With Dimensionality Reduction 7](#_Toc100258513)

[TUNING MODEL 8](#_Toc100258514)

[Hyperparameter Tuning (PANTHEA) 8](#_Toc100258515)

[Grid Search CV 8](#_Toc100258516)

[Decision Tree Modelling 9](#_Toc100258517)

[Hard Voting 9](#_Toc100258518)

[Soft Voting 9](#_Toc100258519)

[Bagging 9](#_Toc100258520)

[ADA Boosting 9](#_Toc100258521)

[Gradient Boosting 9](#_Toc100258522)

[CONCLUSION 10](#_Toc100258523)

[Bibliography 11](#_Toc100258524)

# OBJECTIVE

In this project, we intended to evaluate different machine learning classification models to research the likelihood that an individual would earn an income of $50,000 a year annually or less given certain demographics and social features. The data set used is from a 1994 Census dataset (Cencus Income Data Set, n.d.).

We reviewed the data set to determine the data quality would suffice for our analysis. We however did have to perform additional task of data validation and process feature engineering to proceed with the analysis and modelling task.

# INTRODUCTION

There is an assumption made that demographics and socio economics status determines an individual’s annual income. Using US census data from 1994 we attempted to derive various predictive modelling to assess the possibility of an individual given their social status and demographics to have the ability to earn over 50K.

Through feature importance analysis we can determine if features education, gender, age or immigration status influenced an individual’s income. With in an community, these features can then be focused on to stabilize socio economic disparity. (Machine Learning Mastery , 2020)

# DATA PREPARATION

The dataset provides 14 input variables that are a mixture of categorical and numerical data types. The Target Data is the income level >50 K or <50K. Since there are only two options, this would be a binary classification.

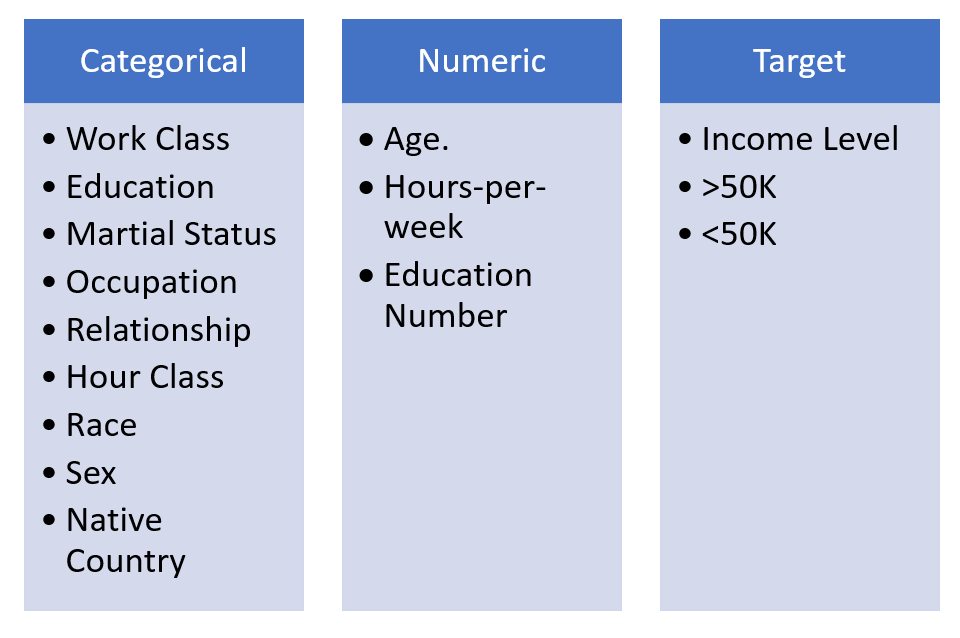


Figure 1 Data Classification

The data contains some irregular data “?” in Work class, Occupation and Native Country. We replaced replace the "?" values with Other. Further to that we created classification features >60 hours or <60 hours called Hour Class.

# DATA VISUALIZATION AND ANALYSIS

Let’s examine our dataset

Chart

Description automatically generated

Figure 2 Age distribution

# (<<SIVI’s Graphs>>)

Is the data balance. Analysis of the Target Variable and the observation Mildly Imbalanced



Overview of other feature columns:

Chart

Description automatically generatedChart, bar chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart, bar chart, waterfall chart

Description automatically generated

Pair plots of entire dataset:

Diagram

Description automatically generated

Correlation heatmap of dataset:

Chart, treemap chart

Description automatically generated

1. Prepare for training models

# MODELLING TECHNIQUES

## Data Transformation

Prior to modelling, the numeric and categorical data needed to be transformed.

Categorical Variables were transformed using One Code Encoder

Numerical Variables were transformed using MinMax Scaler

Why did we use Label Encoder??

The data was then classified into Train (80%) and Test (20%)

## Ensemble Modelling

The data set was subjected to the following modelling, thus achieving a model score as stated below

* Logistic Regression 83.86%
* Support Vector Machine 83.37%
* Neural Network 81.38%
* Random Forest Classifier 82.04%

## Random Forest Classifier

Data set was processed with estimators=100, the results were as follows

|  |  |
| --- | --- |
| Figure 3 - Confusion Matric | Figure 4 Classification Matrix |

### With Dimensionality Reduction

When PCA dimensionality reduction was applied, it reduced the data set to 1 dimension.

The Random Forest Classifier was applied again and results were shown below.

|  |  |
| --- | --- |
| Figure 5 - Confusion Matric | Figure 6 Classification Matrix With Dimensionality reduction |

# TUNING MODEL

### Hyperparameter Tuning (PANTHEA)

### Grid Search CV

# Decision Tree Modelling

### Hard Voting

### Soft Voting

### Bagging

### ADA Boosting

### Gradient Boosting

# CONCLUSION

|  |
| --- |
| We need to add a story here <<>>> |
| WHAT BIG AH AH MOMENT DO WE WANT TO STATE HERE What did we infer from our analysis? |
| More data here <>>> |

# Bibliography

*Cencus Income Data Set*. (n.d.). Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Census+Income

*Machine Learning Mastery* . (2020, October 27). Retrieved from Imbalanced Classification with the Adult Income Dataset: https://machinelearningmastery.com/imbalanced-classification-with-the-adult-income-dataset/