Machine Learning Assignment

Members

815108 N. Nonjoli

805494 D. Khumalo

1126619 O.N. Mekgwe

Contents

1 Introduction

1.1 What is Supervised Learning?

2 Dataset

2.1 Description

The aim of this dataset is to predict whether a person earns $50,000 per annum. This dataset has 14 variables, is multivariate and the area of focus is social.

Adult Data Set

Attribute	Values
age	Age of person
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwg	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 and Tobago, Peru, Hong, Holand-Netherland

48842 Datapoints

2.2 Terminology

Age	:	Age of person
Work Class	:	Class of work
Final Weight	:	Final weight of how much of the population it represents
Education	:	Education level
Education Number	:	Numeric education level
Occupation	:	Occupation of the person
Relationship	:	Type of relationship
Sex	:	Gender of the person
Capital Gain	:	Rise in value of an investment or real estate that gives it a higher worth than the purchase price
Capital Loss	:	Loss incured when an investment or real estate decreases in value
Hours	:	Average number of working hours per week
Native Country	:	Country of origin

2.3 Targets

2.4 Sample

2.5 What are we prediction?

3 Algorithms

3.1 Decision Tree

3.1.1 Description

Decision Trees are used to classify data, the classification can either be categorical or continuous. They are a type of Supervised Machine Learning. The tree can be described by decision nodes and leaves. The leaves describe the final outcomes, and the decision nodes are where the data is split[2].

3.1.2 How data was handled

The following was done to prepare the data:

- Headers were added and saved to a new file adult.csv
- Rows that had missing variables were removed from the data set.
- Redundant attributes/columns were removed, i.e: education-num

3.1.3 Reason

3.1.4 Performance

3.2 Naïve Bayes

3.2.1 Description

3.2.2 How data was handled

3.2.3 Reason

3.2.4 Performance

3.3 Linear Regression

3.3.1 Description

3.3.2 How data was handled

3.3.3 Reason

3.3.4 Performance

4 Results

4.1 Findings

4.1.1 Best Algorithm

4.1.2 Worst Algorithm

4.2 Recommendations

References

[1] Does thinking you look fat affect how much money you earn?
`https://www.timeslive.co.za/sunday-times/lifestyle/health-and-sex/2018-07-23-does-thinking-you-look-fat-affect-how-much-money-you-earn/`

[2] Decision Trees for Classification: A Machine Learning Algorithm
`https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html`