Machine Learning Assignment

UNIVERSITY OF THE WITWATERSRAND
JOHANNESBURG

Members

815108 N. Nonjoli
805494 D. Khumalo
1126619 O.N. Mekgwe

Contents

1 Introduction **1**
1.1 What is Supervised Learning? 1

2 Dataset **2**
2.1 Description 2
2.2 Targets 2
2.3 Sample 2
2.4 What are we prediction? 2

3 Algorithms **3**
3.1 Decision Tree 3
3.1.1 Description 3
3.1.2 How data was handled 3
3.1.3 Reason 3
3.1.4 Performance 3
3.2 Naïve Bayes 4
3.2.1 Description 4
3.2.2 How data was handled 4
3.2.3 Reason 4
3.2.4 Performance 4
3.3 Linear Regression 5
3.3.1 Description 5
3.3.2 How data was handled 5
3.3.3 Reason 5
3.3.4 Performance 5

4 Results **6**
4.1 Findings 6
4.1.1 Best Algorithm 6
4.1.2 Worst Algorithm 6
4.2 Recommendations 6

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	Description
age	continuous
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 Targets

2.3 Sample

2.4 What are we prediction?

3 Algorithms

3.1 Decision Tree

3.1.1 Description

3.1.2 How data was handled

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