Data Mining & Machine Learning assignment

Donal Crotty

Table of Contents

[Introduction 2](#_Toc499066033)

[Overview 2](#_Toc499066034)

[Data Exploration 3](#_Toc499066035)

[Linear Regression – Energy Efficiency 3](#_Toc499066036)

[Classification – Adult Census Income 3](#_Toc499066037)

[kNN – Adult Census Income 3](#_Toc499066038)

[Analyses 3](#_Toc499066039)

[Linear Regression – Energy Efficiency 3](#_Toc499066040)

[Classification – Adult Census Income 3](#_Toc499066041)

[kNN – Adult Census Income 4](#_Toc499066042)

[Evaluation 4](#_Toc499066043)

[Conclusion 4](#_Toc499066044)

[References 4](#_Toc499066045)

# Introduction

As part of the Data Mining and Machine Learning module we were given the task to select three data sets to carry out data mining techniques on. The data sets needed to be suitable for performing both regression and classification techniques on. We were required to perform three data mining techniques; Decision Trees for Classification, Linear Regression for Regression and K Nearest Neighbour (kNN) which can be used for both Classification or Regression.

When carrying out analyses on data, it is necessary to get to know the data you are dealing with prior to performing predictions etc., so you can get a feel as to whether the results you calculate are appropriate or not. In this case, it would be a requirement to explore the data; producing tables and graphs for example. It is also important to define both training and testing datasets. It is also required that you create a prediction model, apply that model to the test data and evaluate the results given.

# Overview

For the Regression analysis, I have chosen to use a dataset based on Energy Efficiency. The energy analysis is performed on 12 different building shapes simulated using an energy efficiency software package. Each building differs with by the attributes glazing area, the glazing area distribution, and the orientation, amongst other attributes. Various settings are simulated as functions of the previously mentioned attributes to obtain 768 building shapes, with each instance in the dataset registered as a building shape. The aim of the dataset is to use the stored attributes to predict a ‘Accurate quantitative estimation of energy performance of residential buildings.’ (Tsanas and Xifara, 2017). The two predicted attributes in this case are Cooling Load and Heating Load.

In Regression analysis, I will be performing the Linear Regression model on the aforementioned dataset. Regression analysis is used to predict the value of one variable called the dependent variable on the basis of other independent variables.

For Classification analysis, I have chosen to use a dataset based on Adult Census Income. The dataset consists of records which were extracted from census data under the following conditions: the age is greater than 16, the Annual Gross Income is less than 100,00, along with more conditions. The purpose of the data is to predict whether a person’s income is greater than, equal to or less than 50,000 per year. Some of the attributes within this dataset include, marital status, sex, occupancy, work class, age, hours per week etc.

In Classification analysis, I will be performing Decision Tree modelling on the above dataset. Classification analysis is used when given a collection of records, with each record containing a set of attributes, the aim is to then assign previously unseen records to a class as accurately as possible. Decision Tree is an all-purpose classifier that does well on most problems. A decision tree is essentially a flowchart of classification and is very useful for applications where the classification mechanism must be transparent for legal reasons for example.

For kNN analysis, I will be performing the method also on the Adult Census Income dataset. kNN is a non-parametric method. The input for the method consists of k’s closest training examples in the feature space. It is the output which decides whether kNN is to be used for Classification or Regression. In Classification, the output is a class membership, whereas in Regression, the output is the property value for the object.

# Data Exploration

Before performing the data mining techniques on the datasets, it is important to firstly explore the dataset we are working with. Exploring the data will allow us to understand the attributes and how important their roles are in making predictions.

## Linear Regression – Energy Efficiency

## Classification – Adult Census Income

## kNN – Adult Census Income

# Analyses

## Linear Regression – Energy Efficiency

## Classification – Adult Census Income

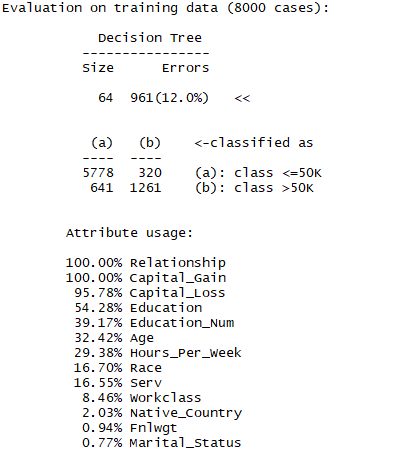


Figure 1 - Summary of Adult Income Decision Tree

The above screenshot represents the summary of the Adult Income Decision Tree Model. It shows that out of the 8000 instances in training data set, there were 961 wrongly classified instances. 320 instances were classified as less than or equal to 50k when they were greater than 50k. These are known as false negatives. 641 instances were classified as greater than 50k when they were less than 50k. These are known as false positives.

The screenshot above also shows the estimated relevance of each attribute to making the Wage prediction. For example, it predicts that Relationship and Capital Gain would be influential when making the prediction. On the other hand, it predicts that Fnlwgt and Marital Status would not be influential when predicting the Wage bracket.

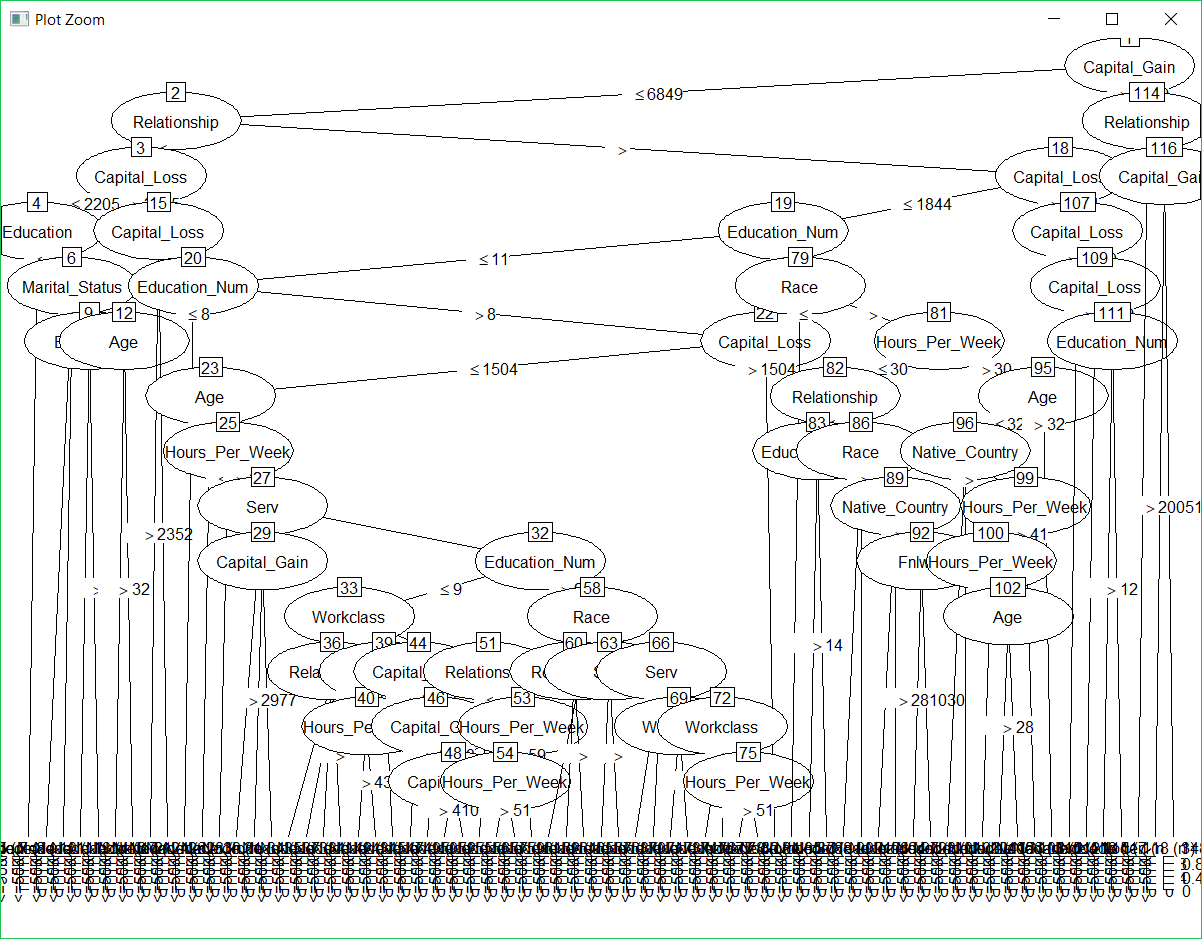


Figure 2 - Adult Census Income Decision Tree

The above screenshot shows the Decision Tree model plot for Adult Census Income. Due to how large the dataset is, the resulting decision tree is quite messy and difficult to understand. Ideally, it would be a lot easier to read the Decision Tree by using a smaller dataset which would result in a much smaller Decision Tree. However, by looking at the Decision Tree, we can evaluate a few attributes from it. Firstly, we can see that Capital\_Gain seems to be the root of the tree with Relationship being the first branch. We can see several splitting attributes, eg. Node 32 Education\_Num splits into the Serv and Workclass nodes. There are also cases of Multi-way splits within the tree. We can also notice evidence of both Induction and Deduction between the nodes within the Decision Tree.

## kNN – Adult Census Income

# Evaluation

# Conclusion

# References

Tsanas, A. and Xifara, A. (2017). *Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools*.