Predicting how hard you will work and how much you will make when you grow up

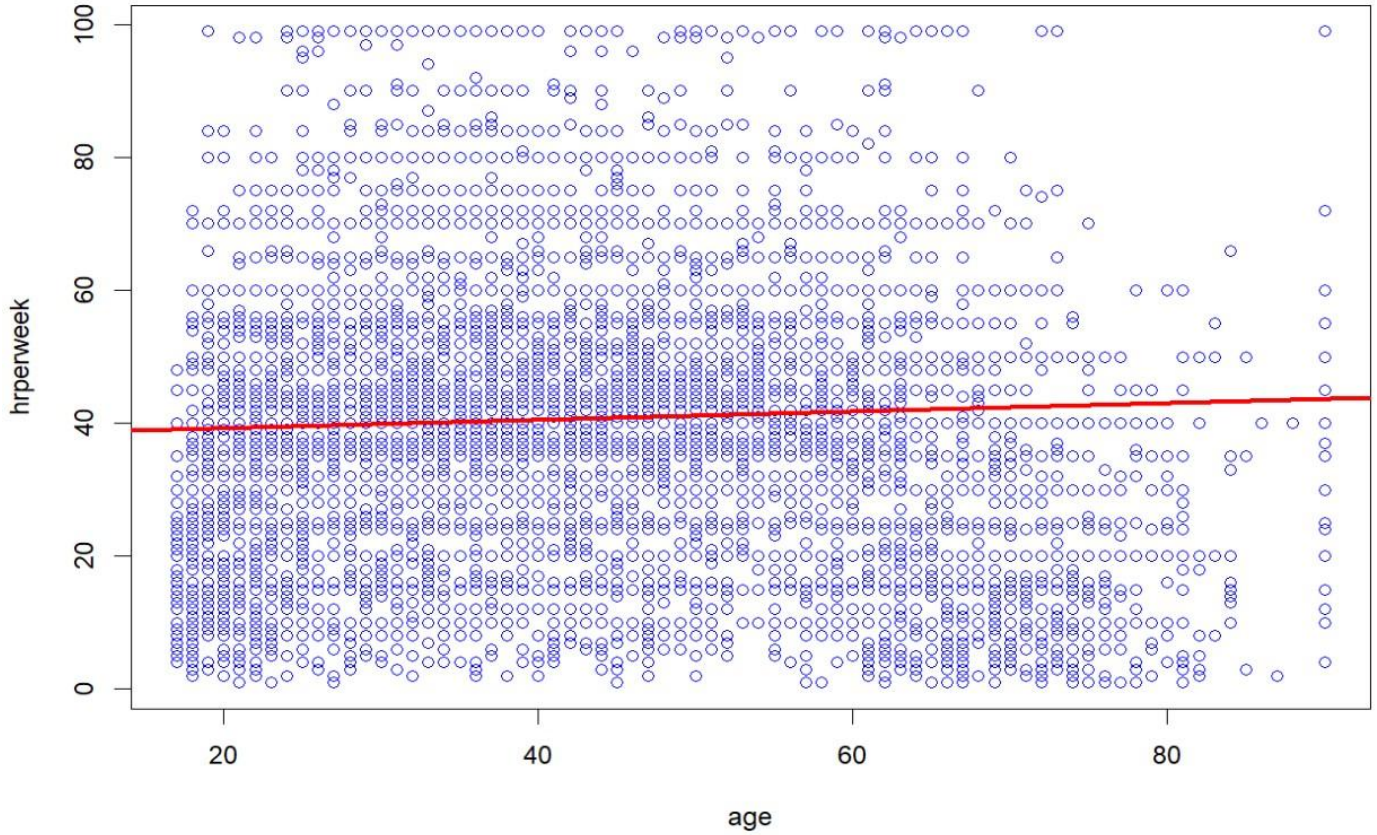
Hours Per Week

Our Hypothesis

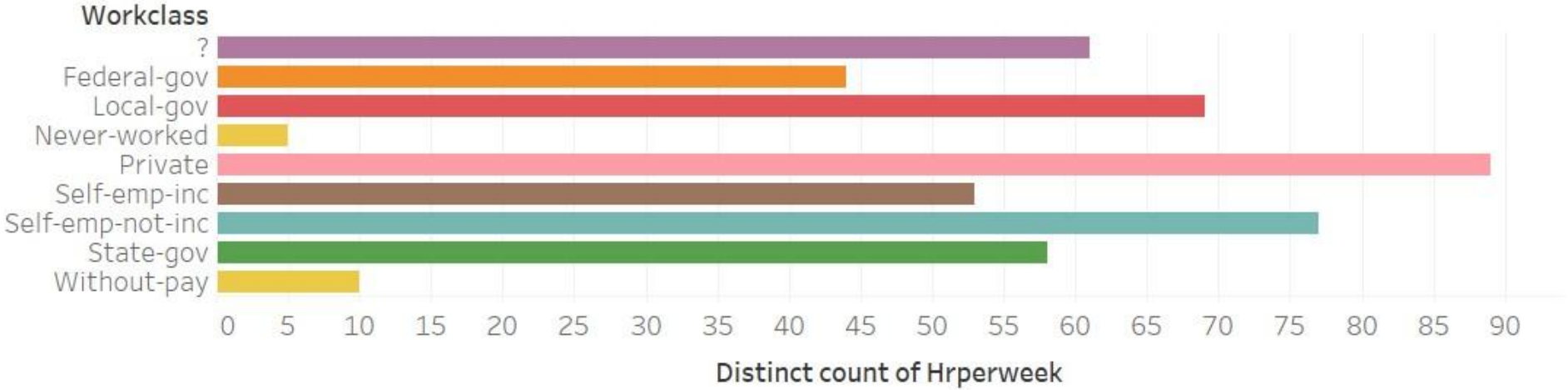
1. Age
2. Workclass
3. Education

These factors are hypothesised to be highly related to the hours worked per week.

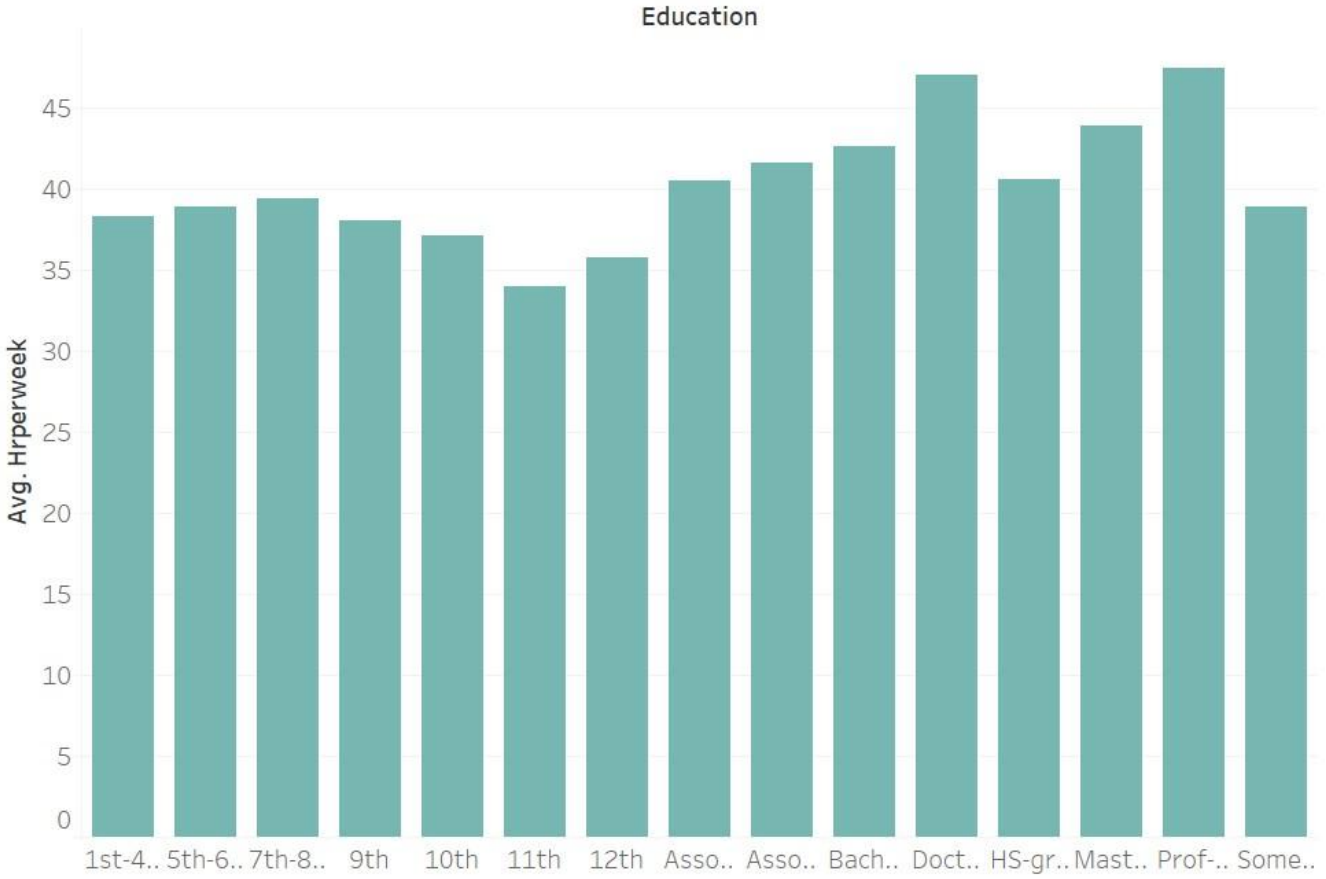
Age vs Hrs per week



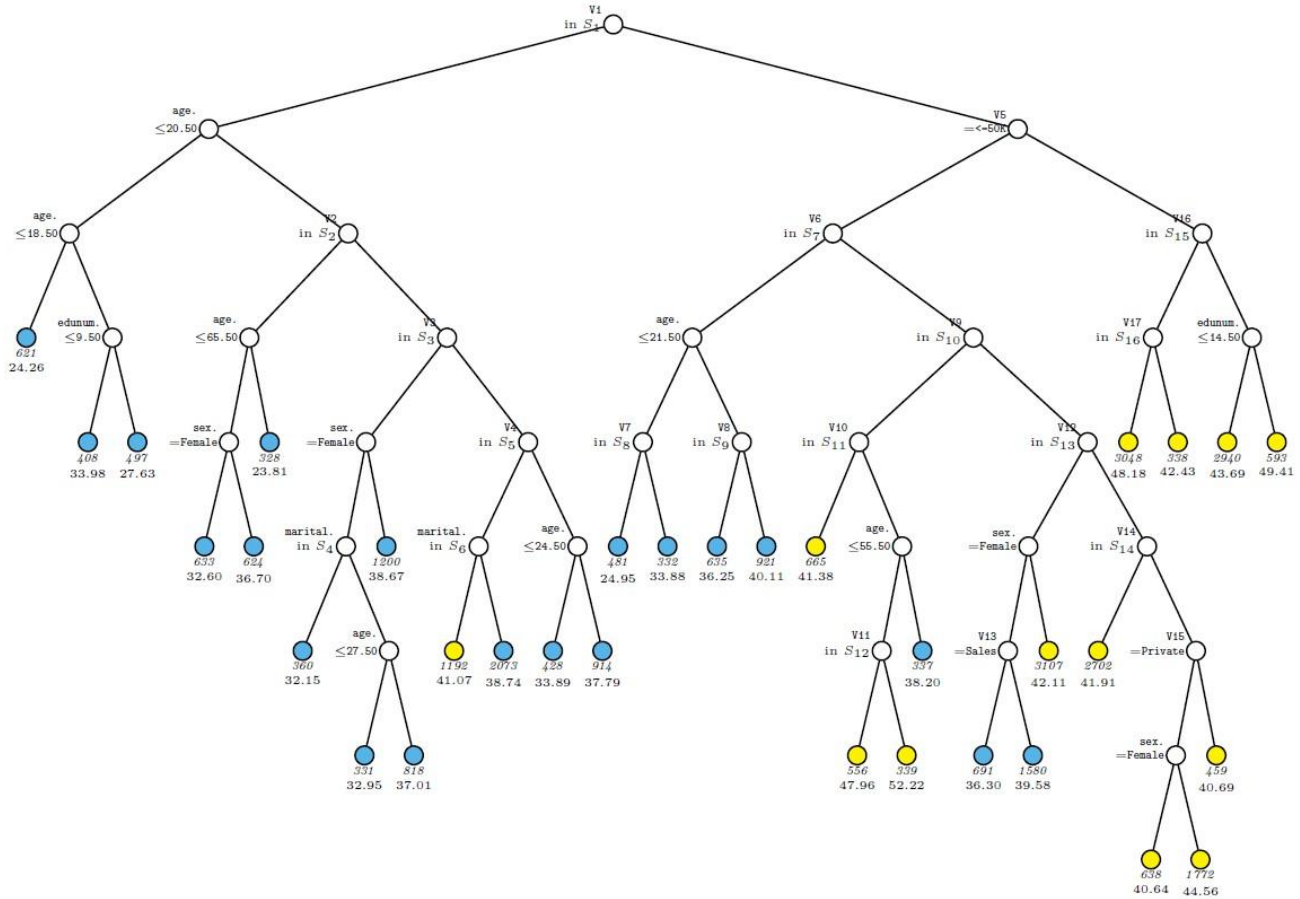
Workclass vs Hrs per week



Education vs Hrs per week



GUIDE – Decision Tree



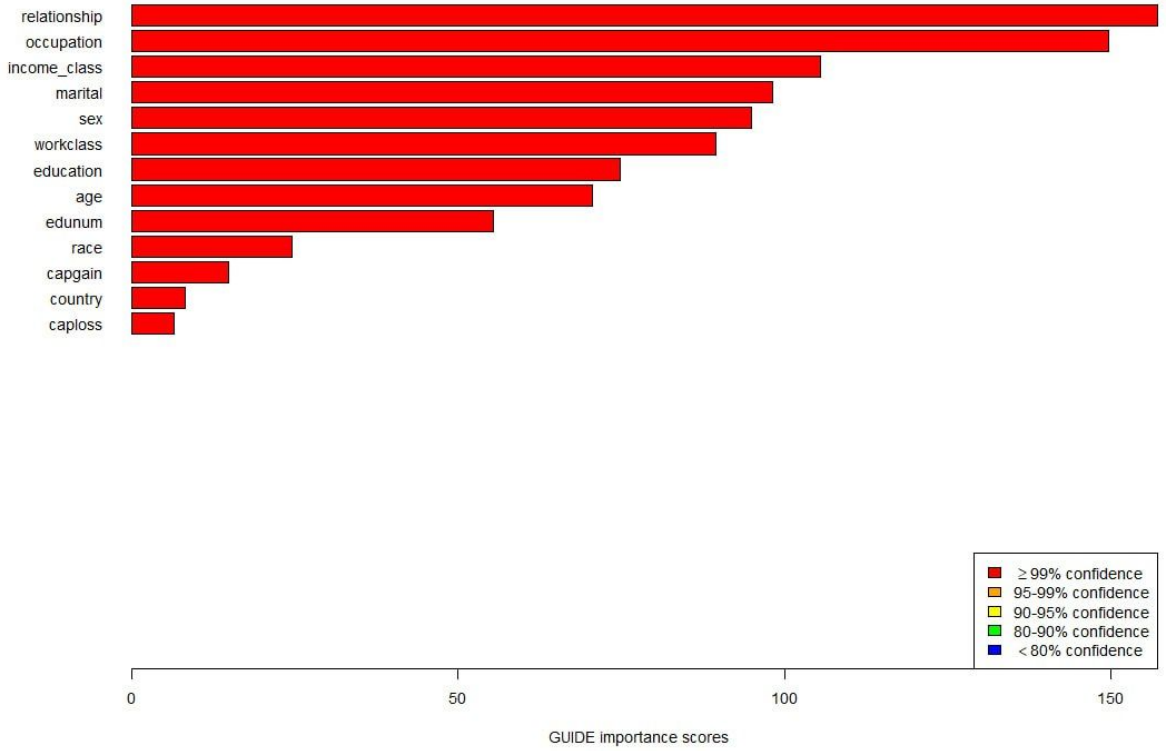
GUIDE

GUIDE v.41.2 0.250-SE piecewise-constant weighted least-squares regression tree for predicting hrperweek.

V1 - V3 = occupation., V4 = relationship, V5 = income.class, V6 = relationship, V7 - V8 = occupation, V9 = workclass, V10 - V14 = occupation, V15 = workclass, V16 = occupation, V17 = relationship.

Terminal nodes with means above and below value of 40.31 at root node are colored yellow and skyblue respectively. First best split variable is occupation while the second best split variable at root node is relationship.

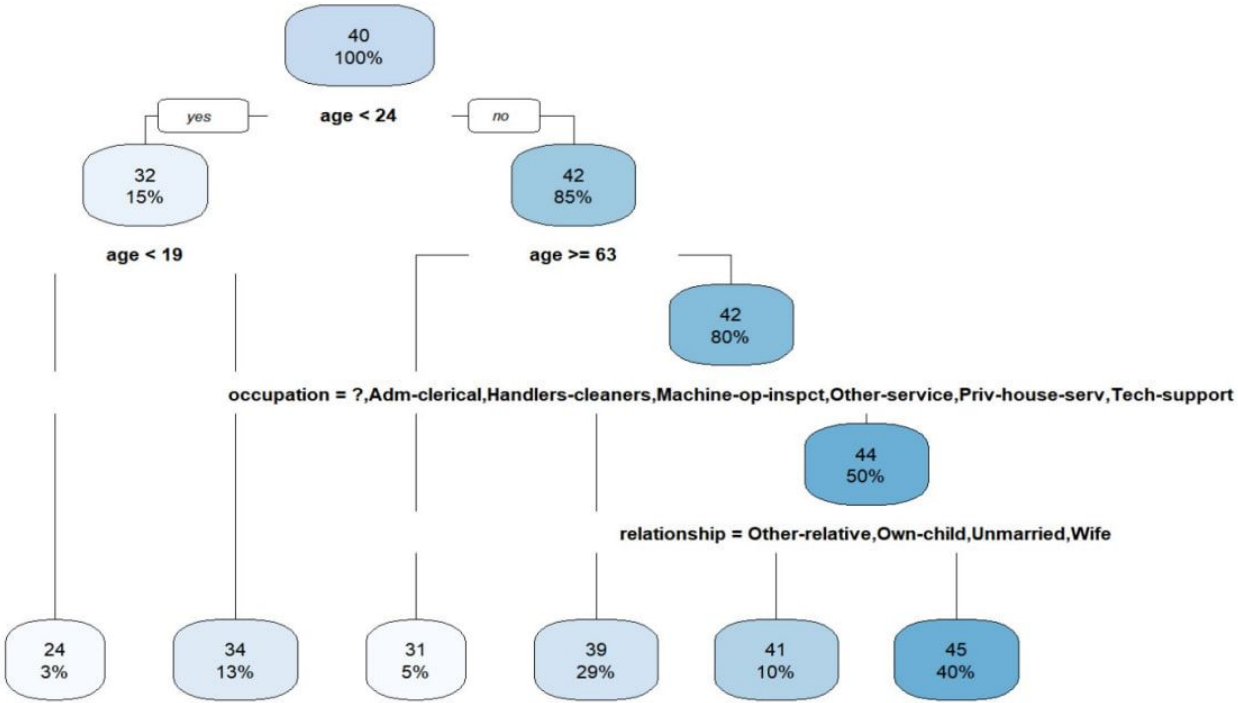
Importance scoring



Importance scoring

From the importance scoring, relationship, occupation, income-class and marital status are the 4 most important variables. This is evident from how they are the most frequently used in splitting the decision tree. However, we noticed that even though relationship is the most important, occupation was used for the very first split in the decision tree.

Rpart


40
100%
yes
age < 24
no
32
15%
42
85%
age < 19
age >= 63
42
80%
occupation = ?,Adm-clerical,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Tech-support
44
50%
relationship = Other-relative,Own-child,Unmarried,Wife
24
3%
34
13%
31
5%
39
29%
41
10%
45
40%