CSCI 620: Introduction to Big Data

Group 9

Data Mining Report on Adult Dataset

Akhil Karrothu
ak8367@rit.edu

Haohua Shen
hs6561@rit.edu

Nishith Agarwal
na9821@rit.edu

Saral Nyathawada
sn5409@rit.edu

Wei Zeng
wz9835@rit.edu

Zizhun Guo
zg2808@rit.ed

1. Introduction

Data Mining is the process of finding the previously unknown, valid and interesting patterns in data using various statistical and machine learning tools. Towards that end, in this project we will apply data mining technique called classification on the 'Adult' dataset from which can be found at this link: https://archive.ics.uci.edu/ml/datasets/Adult which is based on the 1994 census.

This dataset contains various attributes about a large population like income, age, education, marital status, gender, etc. Given the discrete nature of most of the attributes, we decided classification would be the best data mining technique that can be demonstrated on this dataset.

We will first clean-up the dataset by removing any missing values. Then we will analyze each attribute to see if it is important for our final goal or whether it can be modified to make better sense as a data point. Next, we will use different classification models available in R to create classifiers and compare their accuracy. We then make inferences on possible correlation between our class and various other attributes.

The various attribute names and their values are detailed as follows.

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: 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&Tobago, Peru, Hong, Holand-Netherlands.
income: <= 50K, >50K

2. Preprocessing

1. The variable marital_status has 7 levels. We found a pattern in this attribute and we created a new variable MaritalStatus which has only two variables Married which means living with wife/husband and alone.
2. Work_class has been simplified into 3 levels which are government, self and private.
3. Education has 14 levels and it has been reduced to 4 levels in the new attribute education_level
4. native_country has 42 levels which is a lot. We have simplified this extensively by creating a new attribute country_native which has only 2 levels developed and not developed. India and China have been included in developed as because of the highly educated immigrants with high paying jobs coming from these countries. Most of the values are United States.
5. Profession has been simplified into 2 levels “desk job” and “not desk job”
6. Capital_grain and capital_loss have been combined into 1 attribute capital_income which is basicall capital_gain – capital_loss. This has 3 levels “zero”,”negative” and “positive”.
7. The attributes fnlwgt, relationship which is analogous to maritalStatus, education_no which is analogous to education_level have been dropped to reduce complexity.
8. Missing values were omitted from the dataset.

Attribute	Value			
Income	<=50K (22654), >50K (7508)			
Gender	Male (20380), Female (9782)			
Marital Status	Alone (16076), Married (14086)			
Education Level	Assoc (2857), Below high school (3741), College (5044), Graduate (2002), HS-grad (16518)			
Native Country	Developed (28248), Developing (1914)			
Profession	Desk Job (13170), Not Desk (16992)			
Capital Income	Negative (1427), Positive (2538), Zero (26197)			
Hours worked per week	between_40_and_45(16606), between_45_and_60(5790), between_60_and_80(857), less_than_40(6714), more_than_80(195)			
Work Class	Government (4289), Private (22286), Self (3587)			
Age	Min	Max	Average	Median
	17	90	38.44	37

3. Initial Observations

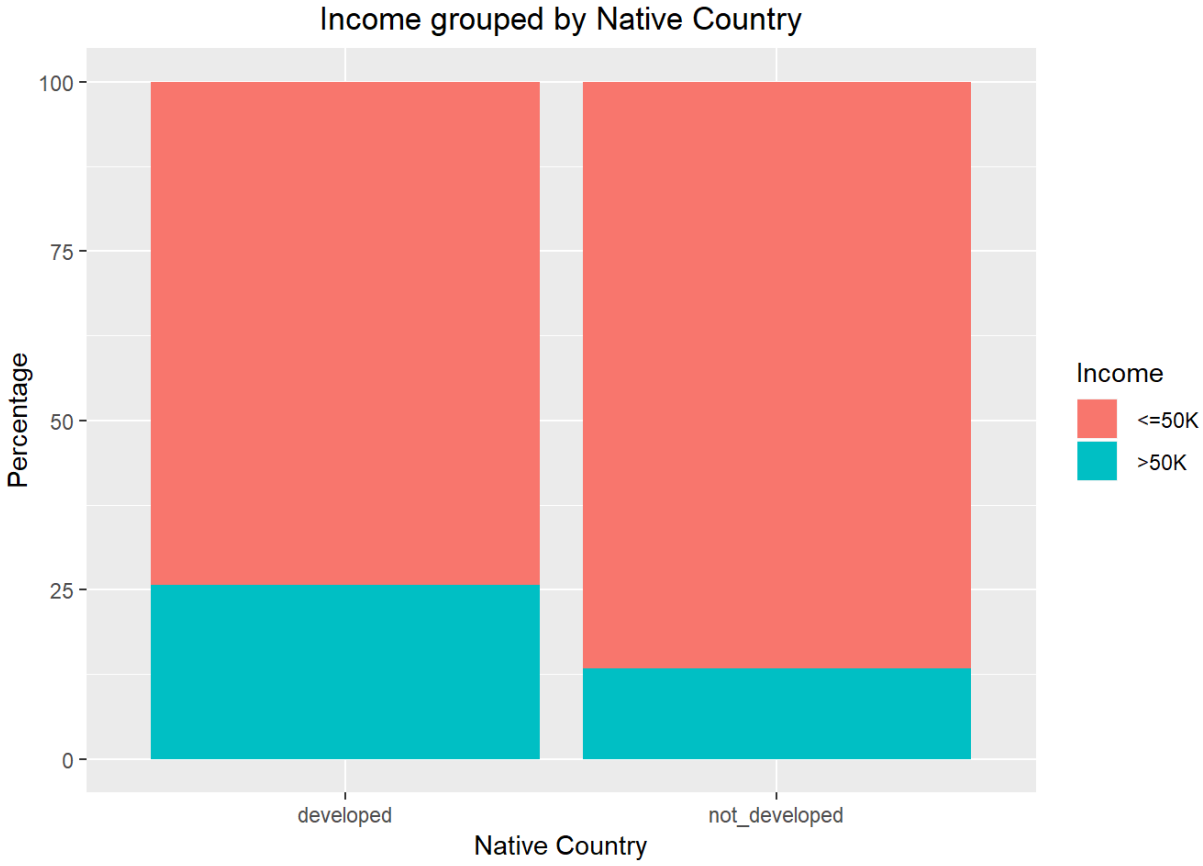
To understand and visualize the effect of a attribute on income level, we calculate the percentage of population earning <=50K and >50K and then plot it for each value of the attribute. For continues values we use line graph and for discrete values we use bar plots.

3.1 Age

An individual's income tends to increase as they get older and they get more experience. However, this should decline once they reach the retirement age as after retirement most people prefer not to work.

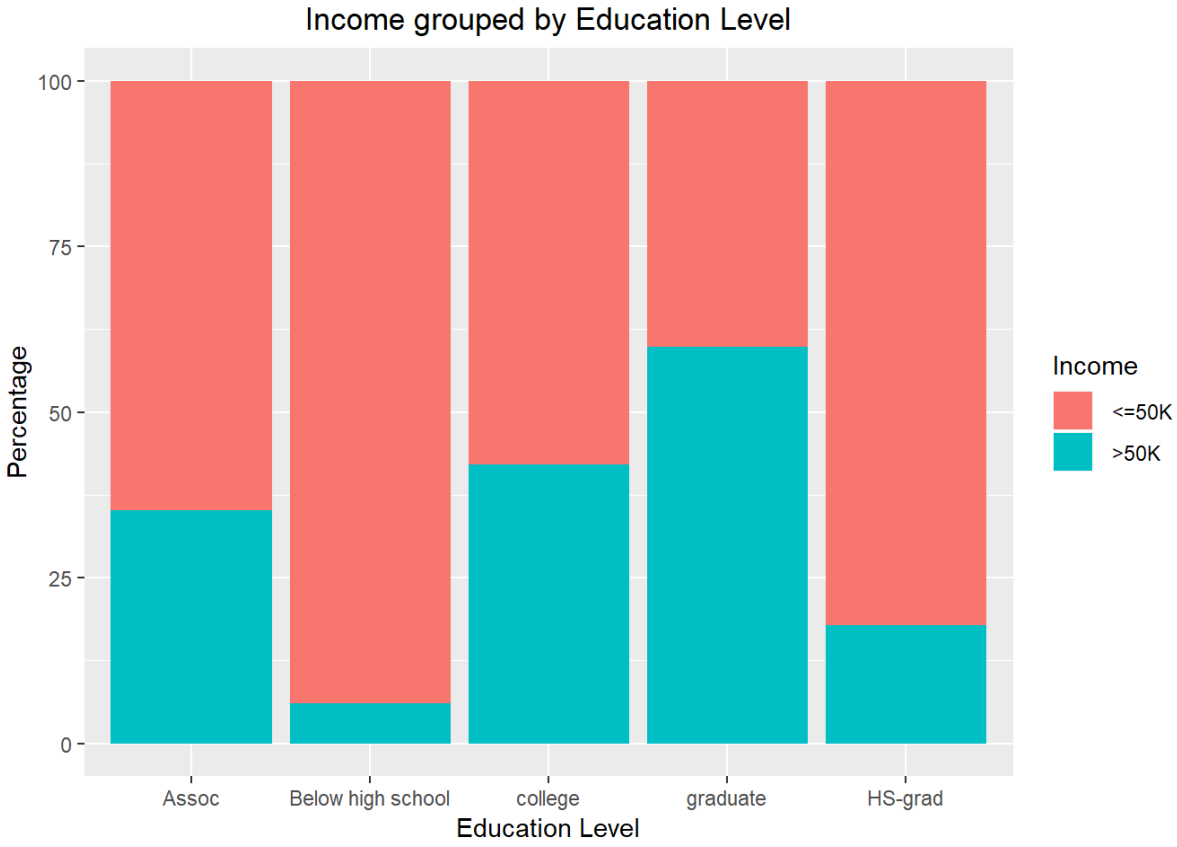
3.2 Native Country

People from richer developed countries have a higher expectation of income compared to people who come from developing countries. Therefore, it is expected that people whose native country is a developing country might settle for lower income jobs.



3.3 Education Level

Higher level of education allows an individual to pursue jobs that require more skill and specializations which in turn pay more. Therefore, higher level of education should correlate with better income.

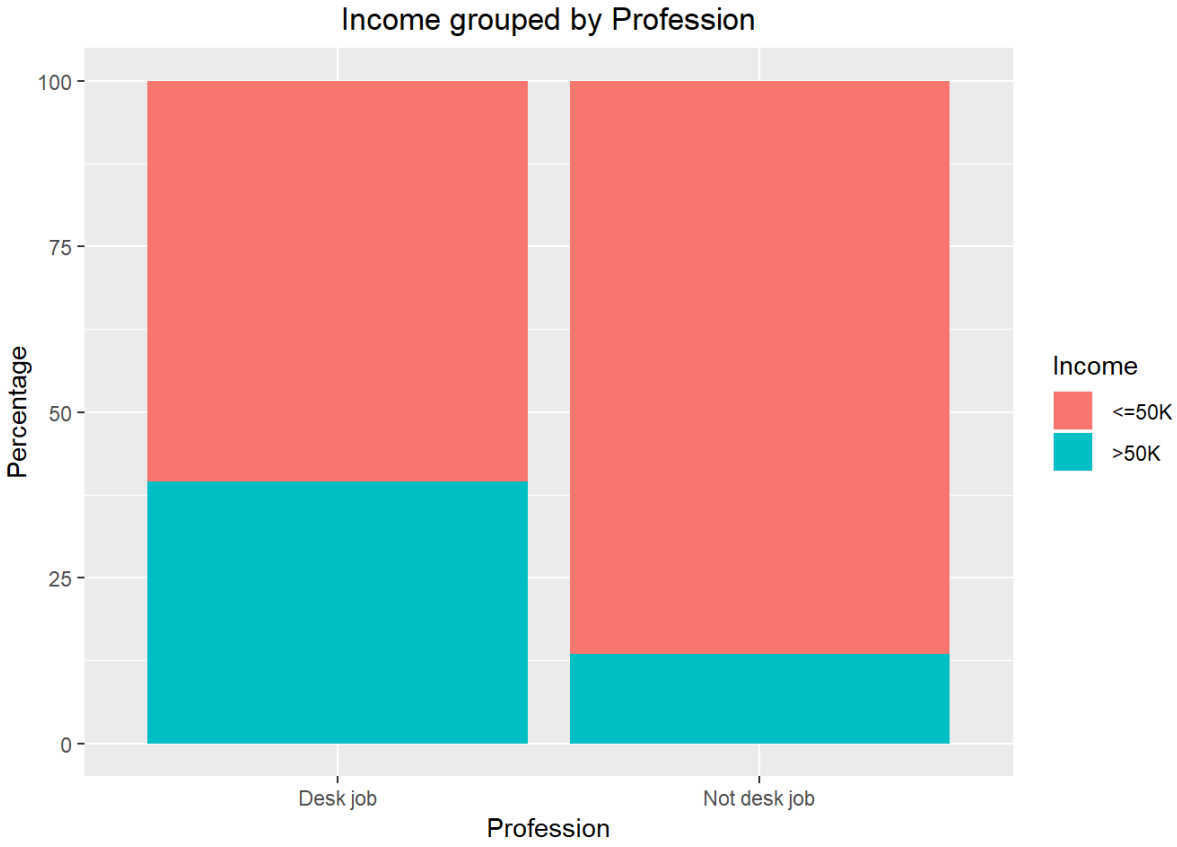


3.4 Marital Status

Bachelors tend to be younger in age which we have already seen is associated with lower level of income. Moreover, financial security of the family is an important factor in keeping a marriage going. Therefore, we predict that a married individual will on average have a higher level of income compared to a bachelor or separated individual.

3.5 **Profession**

Desk job tend to be specialized jobs which require higher level of education then non-desk jobs. As seen above, higher level of education leads to better pay. Therefore, desk jobs should also correlate to better paying jobs.

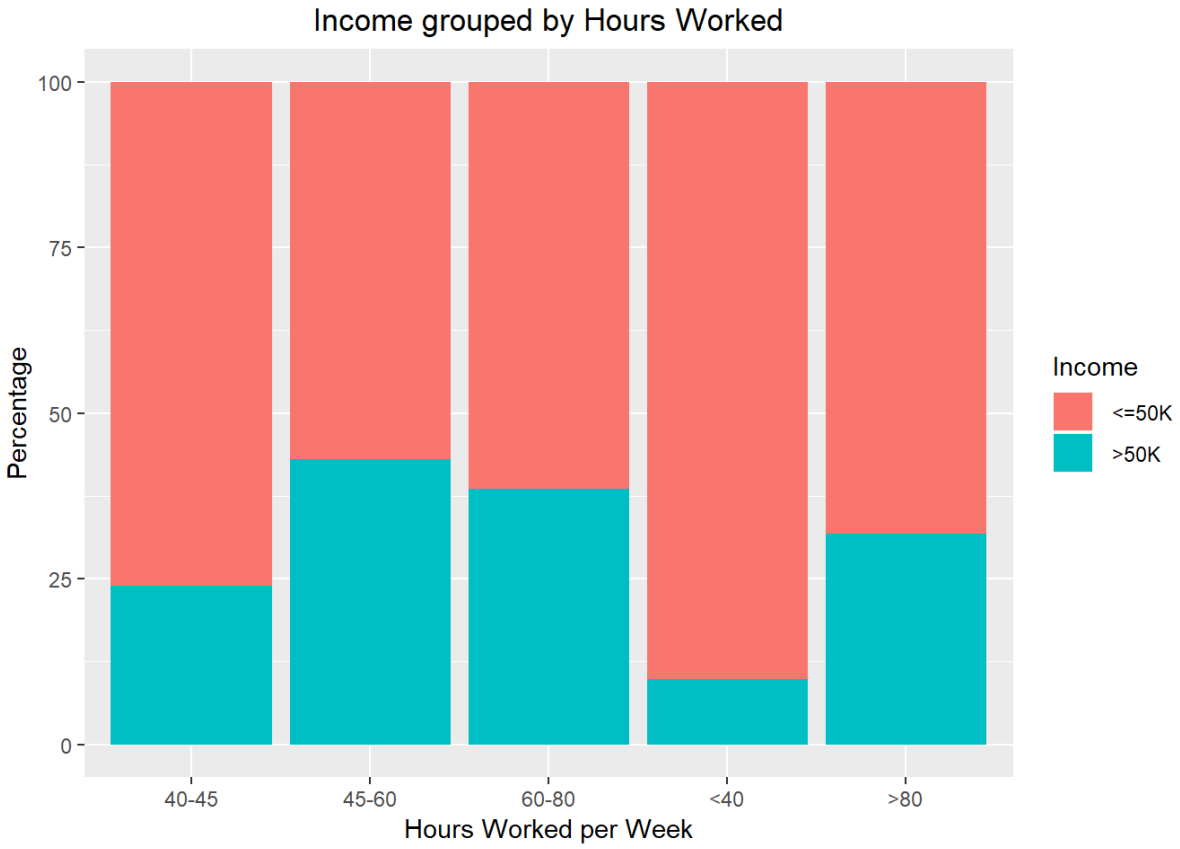


3.6 Gender

Unfortunately, gender pay gap still exists in our workforce especially in developing countries. Which in turn also leads to lower participation of women in workforce in these countries. Therefore, we predict that men on average will earn more than women.

3.7 Hours Worked

Although there will be some variation in income by number of hours worked and the an individual's income is directly proportional to the number of hours worked, we predict that it will not be a strong correlation as income depends more on the type of work done which governs the hourly pay rate which has a much higher effect on the income level.



4. Data Mining Task

To build a model which can predict whether the income of an individual exceeds $50K/yr based on census data available. This is a data mining task and not a data management task because we are looking for patterns in our data which are hidden, and which will help us predict outcomes on unknown and future data with reasonable consistency.

4.1 Classification

Classification is a predictive data mining technique and involves the process of finding a model for class attribute as a function of the values of other attributes given a collection of records. As we can see this fits our task perfectly as our goal is to create a model that can predict the income class of individual (less or greater than 50k) given their census data.

We have applied 4 types of data classification techniques: Decision tree classifier, K-nearest neighbor classifier, Naïve Bayes classifier and SVM classifier.

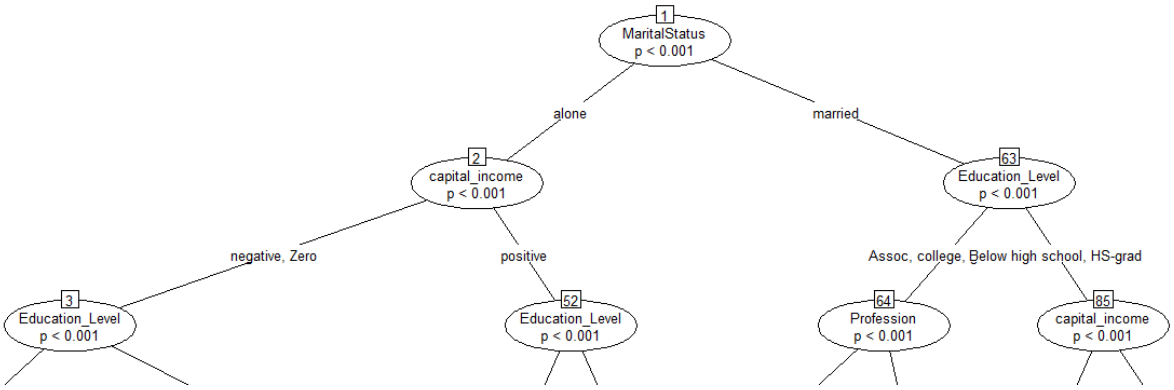
4.2 Test Design

Dataset is split in an 80-20 split. We use the larger split as our training dataset and the smaller split as our test dataset. We train our model using the training dataset and then test the accuracy of that model on test dataset. We do this make sure our model does not overfit the training dataset. As a learning step, we are also calculating the accuracy of each model by taking training data as test dataset to check which models are more prone to overfitting.

4.3 Decision Tree

We use partykit and rpart library to create two decision tree qualifiers.

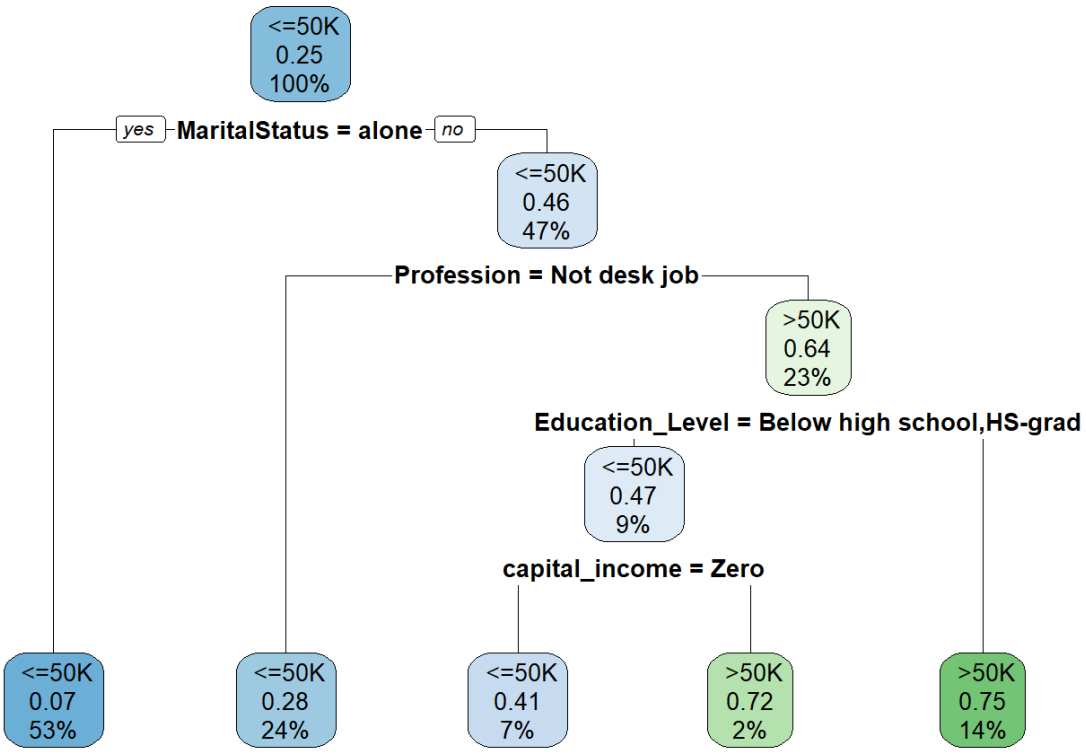
The resulting decision tree from partykit 69 terminal nodes, 68 inner nodes and a depth of 10. Below is a snippet of the top of the decision tree.



Shallowest nodes of the tree – Complete tree description in HTML file

This model gives us an accuracy of 83.43%.

Using rpart we get a much smaller tree which is shown below.



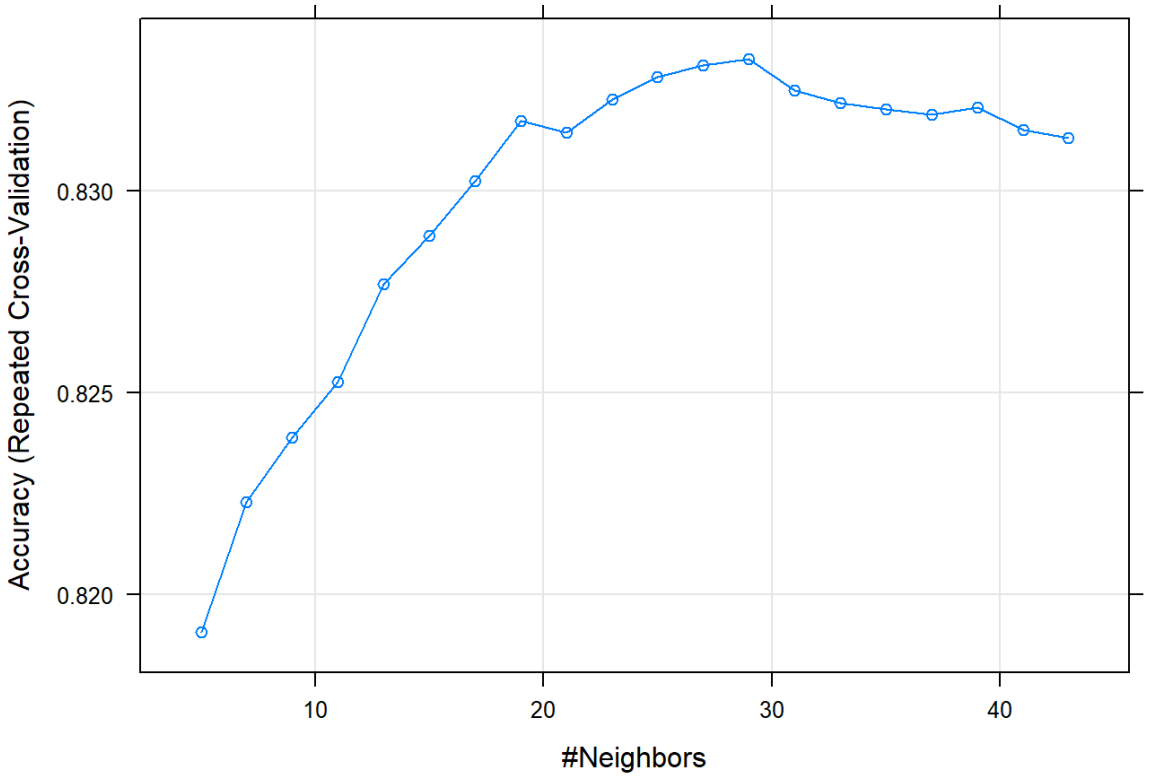
The accuracy of this model is 82.56%.

As we had observed earlier, Marital Status, education level, age and profession are among the attributes that result in the highest knowledge gain and reduction of entropy when used as a splitting attribute. Hence, they appear more frequently in the shallower nodes. While hours worked is amongst the smallest change in entropy when used as splitting attribute which is why it appears only in the deeper nodes.

4.4 K-nearest Neighbor

In our KNN model, the value of k is gradually increased, and accuracy is calculated at each step. Once we start seeing continues decline is accuracy, we stop and the k with highest accuracy is finalized. 10-fold cross-validation was used in this model.

K	Accuracy
5	0.819053
7	0.822291
9	0.823882
11	0.825266
13	0.827688
15	0.828878
17	0.830234
19	0.831714
21	0.831438
23	0.832254
25	0.832807
27	0.833098
29	0.833264
31	0.832489
33	0.832171
35	0.832019
37	0.83188
39	0.83206
41	0.831507
43	0.831299



The final value used for the model that gave the highest accuracy was k = 29. This model gives us an accuracy of 82.81% for our validation dataset.
This model also calculates the importance of each variable in classifying data.

Variable	Importance
Marital Status	100
Education	84.148
age	70.348
Profession	64.342
sex	42.78
Capital income	42.588
Education Level	23.14
race	8.877
Hours worked	5.535
Native Country	3.088
Work Classification	0

As predicated, marital status, education, age and profession are the most important attributes. Hours worked, race, native country and work classification are amongst the least important.

4.5 Naïve Bayes

Naïve Bayes classification utilizes the bayes rule of conditional probabilities to make predications about a dataset. By providing a training dataset, we first calculate the conditional probabilities of various attributes. Then we calculate the conditional probability of an observation given a class. The class with highest probability is then assigned to that observation.

Attribute		Probability	
		<=50K	>50K
Marital Status	Alone	0.66	0.14
	Married	0.34	0.86
Profession	Desk Job	0.35	0.69
	Not Desk Job	0.65	0.31
Native Country	Developed	0.93	0.97
	Not Developed	0.07	0.03
Education Level	Assoc	0.08	0.14
	Below High School	0.15	0.03
	College	0.13	0.28
	Graduate	0.04	0.16
	HS-grad	0.6	0.39
Capital Income	Negative	0.03	0.1
	Positive	0.04	0.21
	Zero	0.93	0.69
Hours per week	Between 40 and 45	0.56	0.53
	Between 45 and 60	0.14	0.33
	Between 60 and 80	0.02	0.04
	Less than 40	0.27	0.09
	More than 80	0.01	0.01
Race	Other	0.15	0.85
	White	0.09	0.91

Above are the probability distribution calculated by the model. By comparing the difference in conditional probability between values of attribute across classes, we can identify strong correlations.

As we have observed with other classifiers, Marital status, profession, capital income and education level show a clear distinction in probability across classes and are good for classification. While race, native country and hours worked per week show similar probabilities across different classes and are therefore bad discriminators.

This model gives us an accuracy of 82.31% for our validation dataset.

4.6 SVM Classifier

Since this is a binary classification task, we have opted to evaluate the Support Vector Machine classifier as well and compare its results. Given a dataset SVM computes a N-dimensional hyperplane which separates the data points by class (N being the number of attributes).

We have used library '1071' in R to create our SVM classifier.

This model gives us an accuracy of 83.75% for our validation dataset.

5. Model Comparison

Model		Accuracy on Validation Dataset (%)	Accuracy on Training Dataset (%)
Decision Tree	Partykit	83.43	83.63
	rpart	82.56	82.51
K-nearest neighbors		82.81	84.12
Naïve Bayes		82.31	82.51
SVM		83.75	83.62

Base on the above we can make the following observations:

1. SVM classifier has the best performance among the models used
2. K-nearest neighbors was the worst performing classifier
3. Decision Tree and naïve bayes classifiers were faster than SVM and KNN to train and validate
4. KNN took the longest time to train (this can be reduced if we take a good enough guess at the value of K at the start)
5. KNN classifier has the biggest variation in accuracy between validation and test dataset. This means that KNN has higher tendency to overfit to data.
6. SVM classifier on the other hand performs even better in validation set than training data

Given the above observations and performance measures, SVM is the best model for our dataset.

6. Conclusions

By and large the observations made during the data exploration were confirmed by the data mining components. Marital status, education, profession and age were the clearest indicators of an individual's income with marital status being the biggest discriminator.

Attributes like hours worked and race were proven to not be an indicator of income as predicted. Native country however surprisingly turned out to be not be a deciding factor.

Attributes like education level and gender were not as dominant as predicted during data exploration.

SVM classifier proved to be the best performing classifier based on the accuracy measure. Followed by decision tree, naïve bayes and knn classifier. KNN classifier has the lowest accuracy and biggest computational overhead of all the methods used. Naïve bayes and decision tree were the fastest models.

KNN classifier also proved prone to overfitting the data. SVM classifier performed best in this regard followed by decision tree and naïve bayes.

Rpart library tree was much simpler than the tree from partykit library at the cost of accuracy. However, the inferences about attribute important were same in both cases. Therefore, rpart can be used as a exploration method to fine tune the parameters for a more complex partykit tree.

Based on our observations we can conclude that if speed is important, decision tree classifier is the best classifier for this dataset. However, if higher accuracy is required then we can go for SVM classifier by sacrificing a little speed.

7. Workload Distribution

Akhil Karrothu:

1.**Preprocessing:**

a. Analyzed the attribute Occupation and reduced its complexity by creating a new attribute Profession with 2 levels in contrast to occupation which had 14 levels

b. Analyzed the attributes Capital_gain and Capital_income and combined them into a new attribute capital_income which has 3 levels.

c. Analyzed the attributes fnlwgt and relationship. Relationship was omitted because it is similar to marital_status and fnlwgt was also omitted to reduce complexity.

2.**Data Mining:**

a. Split the dataset such that training data contains 80% of the records and test data contains 20% of the records

b. Used the partykit library to build the decision tree with ctree function and plot it. Validated the model with the test data and which resulted in an accuracy of 83.43% and printed the confusion matrix.

c. Determined the value of K that gave the highest accuracy using 10-fold cross validation in the knn classifier, which turned out to be k=29. Built the model with k=29 which gave an accuracy of 82.81%

d. Compared and analyzed the different models used and the accuracy resulted when using this particular test data.

Haohua Shen:

1.Exploring dataset:

a. Load dataset,set attributes, check instance number

b. Check datatype and change some of the attributes to factors for preprocessing

c. Check level of each attributes.

2.Data cleaning:

a. Read dataset and set '?' values to NA

b. Drop all instance that have NA values.

3.Preprocessing:

a. Merge levels of hours_per_week attributes to 5 categories.

b. Summarize the native_country into regions.

c. Summarize the workclass attributes. Found that workclass level has no instance, then drop it.

Nishith Agarwal:

1. Data Exploration:

a. Performed initial exploration to identify important attributes.

b. Plotted the line graph for income distribution against age to verify assumptions.

c. Plotted bar graphs for various attributes like martial status, education level, profession, gender, etc. to find the distribution of income class so as to get a baseline for expected results of modeling.

2. R Script:

a. Collated work done by all team members into single R script.

b. Troubleshooting to make sure all the scripts worked on by various team members were compatible with each other so that they can be merged into a single R script.

3. Final Report:

a. Wrote the final report to illustrate all the work done during the project by the team.

b. Created comparison tables to compare the accuracy of various models used.

c. Analyzed the output of various models and translated them in form of tables, graphs or text so that they can be visualized easily.

d. Made final conclusions about the data set and the classification models based on the results obtained.

Saral Nyathawada:

1.Preprocessing:

a. Analyzed the attribute Marital_status, found a pattern and reduced it from 7 levels to 2 levels

b. Analyzed the attribute Work_class, and reduced it from 8 levels to 3 levels

c. Analyzed the attributes Education and education_no. Reduced Education from 14 levels to 4 levels and omitted education_no since it was trivial.

d. Analyzed native_country and noticed that a lot of the attributes were united states and reduced it to 2 levels from 42 levels.

2.Data Mining:

a. Split the dataset such that training data contains 80% of the records and test data contains 20% of the records

b. Used the rpart and rprt.plot library to build the decision tree and plot it. Validated the model with the test data and which resulted in an accuracy of 82.56% and printed the confusion matrix.

c. Used e1071 library to build the naive bayes model on the same training data and validated it which resulted in an accuracy of 82.31%

d. Used e1071 library to build and SVM model and validated it which resulted in an accuracy of 83.75%

Wei Zeng:

1.Preprocessing:

Group the variables "capital gain" and "capital_loss"

2.Analysis and visualization of the variables

a. Using box plot to visualize the distribution of capital_gain and found that Capital_gain is not a good variable for the predictive model since its extremely high number of zeros.

b. Explored the relationship between age and salary.

c. By visualizing the relationship between gender and income, found that gender is a good factor which greatly affect the salary of people.

Zizhun Guo:

1. Data Exploration:

a. Created the summary report for the dataset to get better understanding of underlying attributes.

b. Identified the attributes for pruning or simplification.

2. Presentation Preparation

a. Analyzed all the models, observations and conclusions made in the final report

b. Used learning from report and selected the most important aspects of the project to highlight them in the presentation

c. Created the presentation for the whole project which will be used as guideline for the demo video.