Data Mining Project

Anirudh Kumar Bansal
2014A7PS081P

Lakshit Bhutani
2014A7PS095P

November 21, 2016

Contents

1 Adult Data Set

This data was extracted from the census bureau database found at
http://www.census.gov/ftp/pub/DES/www/welcome.html

Donor Ronny Kohavi and Barry Becker,Data Mining and Visualization, Silicon Graphics.

1.1 Attribute Information

Income : *binary symmetric* - $> 50K, <= 50K$ US$

Workclass : *nominal* - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

Fnlwgt : *continuous*

Education : *ordinal* - Bachelors, Some-college, 11^{th}, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9^{th}, 7^{th}-8^{th}, 12^{th}, Masters, 1^{st}-4^{th}, 10^{th}, Doctorate, 5^{th}-6^{th}, Preschool

Education-num : *continuous*

Marital-status : *nominal* - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse

Occupation : *nominal* - 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 : *nominal* - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

Race : *nominal* - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

Sex : *binary symmetric* - Female, Male

Capital-gain : *continuous*

Capital-loss : *continuous*

Hours-per-week : *continuous*

Native-country : *nominal* - 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.

1.2 Data Preprocessing

- *Fnlwgt* (final weight) represents the number of people the census takers believe that observation represents. This attribute has been *removed* from the analysis as it has low predictive power (to predict income).

  The attribute *education* is also *removed* from the analysis as it is strongly correlated with the attribute *education-num*.

- To handle *null* values in attributes *native-country, occupation and workclass* two approaches were proposed:

  1. The attributes were replaced by the mode value as they were nominal i.e. native-country with 'United-States' , 'occupation' with 'Prof-specialty' and 'workclass' with 'Private'.
  2. The attributes were replaced by taking into consideration other attributes with strong association with these attributes. The procedure followed is -

     Native-country : The most frequent value was 'United-States' with support 89.74 %.

     Occupation : If *education-num* < 9 then occupation was set to 'Other-service' else it was set to 'Prof-specialty'.

     Workclass : If *occupation = Farming-fishing* then 'workclass' was set to 'Self-emp-inc' else if *occupation = Protective-serv* then 'workclass' was set to 'Local-gov' else if *occupation = Armed-forces* then 'workclass' was set to 'Federal-gov' else it was set to 'Private'.

     The *second approach* is adopted because it is more consistent with the original data.

- The attributes *education-num, marital-status, workclass, occupation and native-country* are discretized using domain knowledge.

  1. *Education-num* : The Education-num is discretized as -

Original value	Discretized value
$1-8$	1
$9-10$	2
$11-12$	3
$13-16$	$4-7$

2. *Marital-status* : The Marital-status is discretized as -

Original value	Discretized value
Married-civ-spouse Married-af-spouse Married-spouse-absent	Married
Never-married Seperated Widowed	Unmarried
Divorced	Divorced

3. *Workclass* : The Workclass is discretized as -

Original value	Discretized value
Never-Worked Without-pay	No-pay
Federal-gov Local-gov State-gov	Gov
Private	Private
Self-emp-inc	Self-emp-inc
Self-emp-not-inc	Self-emp-not-inc

4. *Occupation* : The occupation is discretized as -

Original value	Discretized value
Exec-manegerial Prof-specialty Tech-support	White-collar
Craft-repair Farming-fishing Handlers-cleaners Machine-op-inspect Transport-moving Adm-clerical	Blue-collar
Other-service Priv-house-serv	Services
Armed-Forces	Armed-Forces
Protective-serv	Protective-serv
Sales	Sales

5. *Native-country* : The countries are regrouped according to geography, former colonial powers and present economic conditions.

   British-Commonwealth : Canada, England, Ireland, Scotland.

   SE-Asia : Combodia, Laos, Phillipines, Vietnam.

   South-America : Columbia, Equador, El-Salvador.

   Latin-America : Dominican Republic, Haiti, Honduras, Mexico, Puerto-Rico, Guatemala, Jamaica, Nicaragua, Outlying-US, Trinadad & Tobago.

   Euro-1 : Holand-Netherland, Italy.

   Euro-2 : Greece, Hungary, Poland, Portugal, South, Yugoslavia.

   China : China, Hong, Taiwan.

   India : India.

   Others : Cuba, Iran, Japan.

- The attribute 'Hours-per-week' is divided into 5 bins of equal count.

Bin	Lower	Upper
1	$>= 1$	< 35
2	$>= 35$	< 40
3	$>= 40$	< 41
4	$>= 41$	< 48
5	$>= 48$	$<= 99$

- A new field *net_profit* is derived using the expression (*capital-gain* − *capital-loss*) and divided into 10 equal count bins.

Bin	Lower	Upper
1	$>= -4356$	< 0
2	$>= 0$	< 114
3	$>= 114$	< 401
4	$>= 401$	< 594
5	$>= 594$	$<= 14$
6	$>= 914$	< 991
7	$>= 991$	< 1055
8	$>= 1055$	< 1086
9	$>= 1086$	< 1111
10	$>= 1111$	$<= 99999$

1.3 Objective-1

To predict income of an unknown record based on the other attributes.

Methodology The data set is divided into training and testing data in the ratio 2 : 1.For classification,various classifiers like decision tree, k-NN and ensemble are used.
Decision tree(C5.0) and ensemble classifier are applied on the training data set to generate a predicting model for the testing data.
For k-NN classifier,the entire dataset(training and testing) is used and then the testing records are selected.
The classifiers are then compared based on its results for the testing data.

Note : In the following confusion matrices, the rows represent predicted values and the columns represent the actual values of income field.

1. Decision Tree

   - Simple Decision Tree
     - *Properties* : 20 fold cross validation
     - *Results* :

       Cross Validation : Mean : 86.4 and Standard Error : 0.2
       Number of Rules : 41 for $> 50K$ and 15 for $<= 50K$
       Confusion Matrix :

	<= 50K	>50K
<=50K	11194	1239
>50K	1241	2607

   - Expert Decision Tree
     - *Properties* : Pruning severity=50 and Minimum records per child branch=2
     - *Results* :

       Number of Rules : 87 for $> 50K$ and 15 for $<= 50K$
       Confusion Matrix :

	<= 50K	>50K
<=50K	8008	312
>50K	4427	3534

- Expert Decision Tree
  - *Properties* : Pruning severity=75 and Minimum records per child branch=2
  - *Results* :

    Number of Rules : 46 for $> 50K$ and 12 for $<= 50K$

    Confusion Matrix :

	<= 50K	>50K
<=50K	11233	1269
>50K	1202	2577

- Expert Decision Tree
  - *Properties* : Pruning severity=90 and Minimum records per child branch=5
  - *Results* :

    Number of Rules : 30 for $> 50K$ and 13 for $<= 50K$

    Confusion Matrix :

	<= 50K	>50K
<=50K	11494	1373
>50K	941	2473

2. K-NN algorithm

   The distance computations are done using *Euclidean metric* and prediction is made on the basis of the *mean* of the nearest neighbour values.

- *Properties* : K=2 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	12361	1680
>50K	74	2166

- *Properties* : K=3 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11688	1033
>50K	747	2813

- *Properties* : K=5 and number of folds in cross-validation=5,10(results are same for both the values)
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11627	1230
>50K	808	2616

- *Properties* : K=10 and number of folds in cross-validation=5,10(results are same for both the values)
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11778	1631
>50K	657	2215

3. Ensemble Classifier

   The five classifiers used for ensembling are :

   - *Decision Tree - Expert Mode* (Pruning severity = 75 and minimum records per child branch = 2)
   - *Decision Tree - Expert Mode* (Pruning severity = 90 and minimum records per child branch = 5)
   - *Simple Decision Tree* (cross validation = 20)
   - *k-NN* with k = 10 and 10 fold cross validation
   - *k-NN* with k = 5 and 10 fold cross validation

- *Properties* : *Ensemble Method - Voting* and if tie in voting, value is selected randomly.
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11816	1679
>50K	619	2167

- *Properties* : *Ensemble Method - Confidence weighed voting* and if tie in voting, value is selected randomly.
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11815	1679
>50K	620	2167

- *Properties* : *Ensemble Method - Confidence weighed voting* and if tie in voting, value of the classifier with highest confidence is selected.
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11778	1631
>50K	657	2215

- *Properties* : *Ensemble Method - Highest confidence wins*
- *Results* :

  Confusion Matrix :

	<= 50K	>50K
<=50K	11882	1756
>50K	553	2090

Overall Results

Model	Accuracy	Precision	Recall	F_1 score
Simple Decision Tree	84.767%	67.749%	67.784%	0.6777
Expert Decision Tree (50,2)	71.034%	44.391%	91.887%	0.5986
Expert Decision Tree (75,2)	84.822%	68.192%	67.004%	0.6759
Expert Decision Tree (90,5)	85.787%	72.437%	64.300%	0.6813
2-NN (10)	89.226%	96.696%	56.318%	0.7118
3-NN (10)	89.067%	79.017%	73.141%	0.7596
5-NN (5)	87.482%	76.401%	68.278%	0.7211
5-NN (10)	87.482%	76.401%	68.278%	0.7211
10-NN (5)	85.950%	77.124%	57.592%	0.6568
10-NN (10)	85.950%	77.124%	57.592%	0.6568
Ensemble(Voting,Random selection)	85.890%	77.781%	56.344%	0.6535
Ensemble(Confidence-voting,Random selection)	85.880%	77.754 %	56.344 %	0.6534
Ensemble(Confidence-voting,Highest confidence)	85.880%	77.754%	56.344 %	0.6534
Ensemble(Highest confidence wins)	85.820%	79.077%	54.342%	0.6442

Conclusions

- *3-NN with 10 fold cross validation* turns out to be the best predictor with *accuracy* of 89.067% and highest $F_1 score$ of 0.7596. Although 2-NN has a even higher accuracy of 89.226% but it has poor recall and hence will not be efficient in identifying people with income >50K therefore it is ruled out. K-NN can give complex shape decision boundaries as compared to linear boundaries of decision tree. Also since the dimensionality of the data was low, distance metric provided a good representation of actual similarity/dissimilarity of data points. These could be the reasons for better k-NN performance.

- Usually ensemble methods work well only when there is significant diversity among the models i.e. each model makes separate mistakes, but here since the accuracy of the ensembled classifiers are in the close range therefore ensemble method couldn't do better than the near average of all.

- Precision refers to the fraction of people earning $> 50K$ out of the total such people predicted by the model. Recall refers to the fraction of people predicted with income $> 50K$ out of the total such people. Since high value of both the measures is desired hence their harmonic mean i.e. F_1 score is chosen as the metric of evaluation along with overall accuracy of the model.

- Higher pruning of decision tree must have reduced over-fitting and this could be the reason why the generalization error of the model with higher severity of pruning was lower. Also more splits on each level of the tree prevents overtraining on noisy data.

1.4 Objective-2

To find the factors that affect income and how they affect income.

Methodology Both association rule mining and clustering are used to find the relationship between income and other attributes.
Apriori algorithm is used for association rule mining and K-means clustering algorithm is used for clustering.

Apriori:

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k−1. Then it prunes the candidates which have an infrequent sub pattern.

K-means clustering:

k-means is an unsupervised learning algorithms that is used for clustering of data points. The main idea is to define k centers, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is left the first step is completed and at this stage we need to re-calculate k new centroids of the clusters resulting from the previous step.This process is repeated till there are no more changes in the centroids.

1. **Occupation**:

Consequent	Antecedent	Support	Confidence
income>50K	occupation=White-collar	38.809%	38.618%
income>50K	occupation=Sales	11.268%	26.798%
income>50K	occupation=Blue-collar	42.303%	16.043%

People in blue-collar jobs are much more likely to be earning <50K compared to those in sales or in white-collar jobs. Hence the preprocessed grouping of occupations based on income has grounds based on actual data.

2. **Gender**:

Consequent	Antecedent	Support	Confidence
income>50K	gender=female	33.151%	10.925%
income>50K	gender=male	66.848%	30.376%

Males are about 3 times more likely to earn >50K than females. Therefore males are favoured over females for high income jobs.

3. **Worktime**:

Consequent	Antecedent	Support	Confidence
income>50K	hours-per-week=5	21.477%	41.754%
income>50K	hours-per-week=4	7.907%	35.266%
income>50K	hours-per-week=3	46.687%	21.269%
income>50K	hours-per-week=2	6.740%	15.643%

More the number of hours worked in a week, greater the probability of the individual earning >50K. Hence *the longer you work, the more you earn.*

4. **Native-country**:

Consequent	Antecedent	Support	Confidence
income>50K	native-country=India	0.309%	41.059%
income>50K	native-country=Euro-1	0.294%	34.722%
income>50K	native-country=British-commonwealth	0.751%	33.787%
income>50K	native-country=Other	0.591%	30.449%
income>50K	native-country=China	1.021%	27.454%
income>50K	native-country=United-States	91.497%	24.422%
income>50K	native-country=SE-Asia	0.884%	23.842%
income>50K	native-country=Euro-2	0.737%	22.500%

Some immigrants earn more than the native US citizens with *Indians having the maximum proportion of people earning >50K* whereas Indians make up relatively a small proportion of the immigrant population. Euro-1 countries were preprocessed as the European countries with healthy economic status. People of such countries are also earning more than the average US citizen. China forms a relatively large proportion of immigrant population but are almost similar in income distribution to the average US citizen.

5. **Marital-status**:

Cluster number	Cluster size	Income	Marital-status
1	24167	$<= 50K$ - 100%	Never-married : 63.7% Divorced : 36.3%
2	12988	$<= 50K$ - 100%	Married : 100%
3	11687	$> 50K$ - 100%	Married : 86% Divorced : 7.68% Un-married : 6.32%

Average silhouette coefficient = 0.7

High percentage of people who are either Un-marrried or Divorced are likely to earn $<= 50K$ whereas a significant proportion of married people have income $> 50K$, though some also have income $<= 50K$.

6. **Education-num**:

Cluster number	Cluster size	Income	Average Education-num
1	28138	$<= 50K$ - 100%	1.79
2	9017	$<= 50K$ - 100%	3.93
3	5867	$> 50K$ - 100%	2.10
4	4772	$> 50K$ - 100%	4.31
5	1048	$> 50K$ - 100%	6.41

Average silhouette coefficient = 0.8

While people with *less education are less likely to earn >=50K and vice versa*, the positive relationship between years of education and income does not hold for people with medium level of education, although people with high level of education will most certainly earn $> 50K$.

7. **Age**:

Cluster number	Cluster size	Income	Average age
1	37155	$<= 50K$ - 100%	36.87
2	7118	$> 50K$ - 100%	37.44
3	4569	$> 50K$ - 100%	54.92

Average silhouette coefficient = 0.7

Younger people tend to earn less than their older counterparts. Since high income jobs require experience hence this is a reasonable trend.

Conclusions

- An *old married male with a white collar job and high education working for relatively longer durations* is more likely to earn high.
- The data is *extremely skewed with respect to native-country* and hence generalizing trends is prone to error. With the information available if a person is an Indian or a European his probability of earning >50K is higher than an average US citizen.
- *Capital-gain doesn't seems to affect the income of a person.*

1.5 Objective-3

Understanding the social aspects of the American society

Methodology Association rule mining and clustering algorithms are used to find interesting patterns about the social life of people.
Apriori algorithm is used for association rule mining whereas Two step, K-means and Kohonen algorithms are used for clustering. The best clustering is selected based on Silhouette coefficient.

Two Step algorithm : Two Step Cluster is used to uncover patterns in the set of input fields. TwoStep Cluster is a two-step clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time.

Kohonen algorithm : Kohonen networks are a type of neural network that perform clustering. The basic units are neurons, and they are organized into two layers: the input layer and the output layer. All of the input neurons are connected to all of the output neurons, and these connections have strengths, or weights, associated with them. During training, each unit competes with all of the others to "win" each record. The output neuron with the strongest response is said to be the winner and is the answer for that input.

Silhouette Coefficient : The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

For each data point i, $a(i)$ is defined to be the average dissimilarity of the point i with all other points within the same cluster.Let $b(i)$ be the lowest average dissimilarity of i to any other cluster, of which i is not a member (i.e dissimilarity with its nearest neighbouring cluster).

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\ 0 & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & a(i) > b(i) \end{cases}$$

- Marital stability
  - Effect of education

Cluster number	Cluster size	Marital status	Average age	Average education-num
1	23004	Married - 100%	43.26	2.72
2	16117	Unmarried - 100%	28.13	2.43
3	9681	Divorced - 100%	45.16	2.41

As the *education level increases*, people tend to have better marital relationships as compared to people with lower level of education.

- Racism and Sexism
  - Race affecting income

Consequent	Antecedent	Support	Confidence
income>50K	native-country = Latin-America and race = Black	0.493%	10.373%
income>50K	native-country = Other and race = Black	0.018%	11.111%
income>50K	race = Black	9.592 %	12.081%
income>50K	native-country = Unites-States and race = Black	9.010%	12.156%
income>50K	native-country = India and race = Black	0.004%	50.000%

In general,*blacks of almost all native countries are disadvantaged* in terms of proportion of people earning >50K with India being a notable exception (though the support is very small). Infact, half of the blacks in India earn above 50K, more than the proportion of Indians as a whole (41.059%).

  - Gender compared with Race in affecting income

Consequent	Antecedent	Support	Confidence
income>50K	Sex = Male and race = White	58.832%	31.546%
income>50K	Sex = Female and race = White	26.671 %	11.836%
income>50K	Sex = Male and race = Black	4.866%	18.258%
income>50K	Sex = Female and race = Black	4.725%	5.719%
income>50K	Sex = Male and race = Asian-Pac-islander	2.051%	33.932%
income>50K	Sex = Female and race = Asian-Pac-Islander	1.058%	13.346%
income>50K	race = Asian-Pac-Islander	3.110%	26.925%

Gender plays a more powerful role than race in affecting income. For example, while the white race does better than black race in terms of income, less proportion of white females earn above 50K than black males.While the Asian-Pac-Islander males do better than white males,their female counterparts do worse than black males-implying that racial difference is overshadowed by gender divide.

- Work-style
  - Age affecting workclass

Cluster number	Cluster size	Average age	Workclass
1	36722	37.14	Private = 100%
2	6549	41.22	Govt = 100%
3	3862	45.33	Self-emp-not-inc = 100%
4	1709	45.91	Self-emp-inc = 99.2%

Average silhouette coefficient = 0.6

Younger people tend to work in private sector while older people opt for either government jobs or are self employed. One possible explanation is that the private sector is more willing to give young adults an early career job where as govt or self employed jobs require more experience.

  - Education and Occupation

Cluster number	Cluster size	Average education-num	Occupation
1	20662	2.08	Blue Collar = 100%
2	15879	3.47	White Collar = 100%
3	5813	1.83	Services = 99.8%
4	5505	2.57	Sales = 100%
5	983	2.42	White Collar = 100%

Highly educated people are more likely to be working in White Collar jobs whereas less educated people are more likely to be working in Services. People who are moderately educated find employment in Blue Collar jobs or Sales.

2 Student Data Set

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

2.1 Attribute Information

Sex : *binary symmetric*: female or male - student's sex

Age : *numeric*: from 15 to 22 - student's age

School : *binary symmetric* : *Gabriel Pereira or Mousinho da Silveira* - student's school

Address : *binary symmetric*: urban or rural - student's home address type

Pstatus : *binary symmetric* : living together or apart - parent's cohabitation status

Medu : *numeric*: from 0 to 4 - mother's education

Mjob : *nominal*: teacher, health, services, at_home or other - mother's job

Fedu : *numeric*: from 0 to 4 - father's education

Fjob : *nominal*: teacher, health, services, at_home or other - father's job

Famsize :*binary symmetric*: >= 3 or < 3 - family size

Famrel : *numeric*: from 1 - very bad to 5 - excellent - quality of family relationships

Reason : *nominal*: close to 'home', school 'reputation', 'course' preference or 'other' - reason to choose this school

Traveltime : *numeric*: 1 - < 15 min, 2 - 15 to 30 min, 3 - 30 min to 1 hour or 4 - > 1 hour - home to school travel time

Studytime : *numeric*: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours or 4 –>10 hours - weekly study time

Failures : *numeric* : (n if 1 <= n < 3, else 4) - number of past class failures

Schoolsup : *binary symmetric*: yes or no - extra educational school support

Famsup : *binary symmetric*: yes or no - family educational support

Activities : *binary*: yes or no - extra-curricular activities

Paidclass : *binary symmetric*: yes or no - extra paid classes

Internet : *binary symmetric*: yes or no - Internet access at home

Nursery : *binary symmetric*: yes or no - attended nursery school

Higher : *binary symmetric*: yes or no - wants to take higher education

Romantic : *binary symmetric*: yes or no - with a romantic relationship

Freetime : *numeric*: from 1 - very low to 5 - very high - free time after school

Goout : *numeric*: from 1 - very low to 5 - very high - going out with friends

Walc : *numeric*: from 1 - very low to 5 - very high - weekend alcohol consumption

Dalc : *numeric*: from 1 - very low to 5 - very high - workday alcohol consumption

Health : *numeric*: from 1 - very bad to 5 - very good - current health status

Absences : *numeric*: from 0 to 93 - number of school absences

G1 : *numeric*: from 0 to 20 - first period grade

G2 : *numeric*: from 0 to 20 - second period grade

G3 : *numeric*: from 0 to 20 - third period grade

2.2 Data Preprocessing

- The target class is alcohol consumption and since the data set has weekend as well as workday alcohol consumption, new attribute *alc* is derived with the weighted average of both *Walc* and *Dalc*.

$$alc = \frac{2 \cdot Walc + 5 \cdot Dalc}{7}$$

  This attribute *alc* is discretized as follows :

```
if(alc > 2)
   alc = 1
else
   alc = 0
```

- From the three attributes G_1, G_2 *and* G_3 representing grades at three different times, a single attribute *Grade* is derived which is the arithmetic mean of the three.

$$Grade = \frac{G_1 + G_2 + G_3}{3}$$

- The attribute *Medu and Fedu* is discretized into three bins as follows.

Original Value	Discretized Value
0	0
1 - 2	1
3 - 4	2

  The value 0 represents 'no education', 1 represents 'uptil high school' and 2 represents 'higher education'.

- The attribute *Famrel, Freetime, Goout and Health* is binarized as follows.

Original Value	Binarized Value
1 - 3	0
4 - 5	1

- The attribute *absences* is binarized as follows.

Original Value	Binarized Value
0 - 10	0
11 - 32	1

  The value 0 represents 'moderate absence' which was found to be frequent where as value 1 represents 'high absence' which was rare.

- The attribute *Grade* is discretized as follows.

Original Value	Discretized Value
0 - 7	0
8 - 14	1
15 - 20	2

- The attributes *Famsize and Reason* are removed from the analysis as they were found to have low predictive power as far as alcohol consumption is concerned.

2.3 Objective-1

To predict whether a student consumes heavy amount of alcohol based on the other attributes.

Methodology Since the data set has only 649 records, hold out method is not used. Instead the entire data set is used for model building and prediction in case of k-NN. For both decision tree(C5.0) and ensemble method, the dataset was sampled randomly and one of the sample was used for model building whereas the other was used for testing.

Note : In the following confusion matrices, the rows represent predicted values and the columns represent the actual values of alcohol consumption field.

1. Decision Tree

   - Simple Decision Tree
     - *Properties* : 20 fold cross validation
     - *Results* :

       Cross Validation : Mean : 83.5 and Standard Error : 1.6

       Number of Rules : 2 for alc=1 and 3 for alc=0

       Confusion Matrix :

	alc=0	alc=1
alc=0	249	33
alc=1	18	21

   - Expert Decision Tree
     - *Properties* : Pruning severity=50 and Minimum records per child branch=2 with 10 fold cross validation
     - *Results* :

       Cross Validation : Mean : 83.4 and Standard Error : 1.2

       Number of Rules : 4 for alc=1 and 3 for alc=0

       Confusion Matrix :

	alc=0	alc=1
alc=0	230	24
alc=1	37	30

- Expert Decision Tree
  - *Properties* : Pruning severity=75 and Minimum records per child branch=2 with 10 fold cross validation
  - *Results* :

    Cross Validation : Mean : 81.4 and Standard Error : 1.4

    Number of Rules : 2 for alc=1 and 1 for alc=0

    Confusion Matrix :

	alc=0	alc=1
alc=0	250	33
alc=1	17	21

- Expert Decision Tree
  - *Properties* : Pruning severity=90 and Minimum records per child branch=5 with 10 fold cross validation
  - *Results* :

    Cross Validation : Mean : 83.8 and Standard Error : 1.8

    Number of Rules : 1 for alc=1 and 3 for alc=0

    Confusion Matrix :

	alc=0	alc=1
alc=0	260	40
alc=1	7	14

2. K-NN algorithm

   The distance computations are done using *Euclidean metric* and prediction is made on the basis of the *mean* of the nearest neighbour values.

- *Properties* : K=2 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	532	77
alc=1	0	40

- *Properties* : K=3 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	510	55
alc=1	22	62

- *Properties* : K=5 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	516	78
alc=1	16	30

- *Properties* : K=10 and 10 folds of cross-validation
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	528	103
alc=1	4	14

3. Ensemble Classifier

   The five classifiers used for ensembling are :

   - *k-NN* with k = 2 and 10 fold cross validation
   - *k-NN* with k = 3 and 10 fold cross validation
   - *k-NN* with k = 5 and 10 fold cross validation
   - *Decision Tree - Expert Mode* (Pruning severity = 90 and minimum records per child branch = 5)
   - *Simple Decision Tree* (cross validation = 20)

- *Properties* : *Ensemble Method - Voting* and if tie in voting, value is selected randomly.
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	265	37
alc=1	2	17

- *Properties* : *Ensemble Method - Confidence weighed voting* and if tie in voting, value is selected randomly.
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	265	37
alc=1	2	17

- *Properties* : *Ensemble Method - Highest confidence wins*
- *Results* :

  Confusion Matrix :

	alc=0	alc=1
alc=0	266	45
alc=1	1	9

Overall Results

Model	Accuracy	Precision	Recall	F_1 score
Simple Decision Tree (20)	84.112%	53.846%	38.888%	0.4516
Expert Decision Tree (50,2)	80.996%	44.776%	55.555%	0.4959
Expert Decision Tree (75,2)	84.423%	55.263%	38.888%	0.4565
Expert Decision Tree (90,5)	85.358%	66.666%	25.925%	0.3733
2-NN (10)	88.135%	100.000%	34.189%	0.5095
3-NN (10)	88.135%	73.809%	52.991%	0.6169
5-NN (10)	85.516%	70.909%	33.333%	0.4534
10-NN (10)	83.513%	77.777%	11.965%	0.2074
Ensemble(Voting,Random selection)	87.850%	89.473%	31.481%	0.4657
Ensemble(Confidence-voting,Random selection)	87.850%	89.473%	31.481%	0.4657
Ensemble(Highest confidence wins)	85.669%	90.000%	16.666%	0.2812

Conclusions

- *3-NN with 10 fold cross validation* turns out to be the best predictor with highest *accuracy* of 88.135% and highest F_1 score of 0.6169. 2-NN has the same accuracy as 3-NN and 100.00% precision but it suffers with a poor recall value reflected in its modest F_1 score.
- With *increase in severity of pruning of the decision tree*, both accuracy and precision increased but recall decreased meaning that a greater proportion of students predicted by the model to be heavy alcohol consumers were actually so but it failed to identify many heavy alcohol consumers. This may have happened as the highly pruned trees under-fitted data and were unable to consider all the factors responsible for high alcohol consumption.
- In k-NN increasing the value of k beyond 3, worsened the model both in terms of accuracy and F_1 score. This may have happened due to presence of noise points in the neighbourhood.

2.4 Objective-2

To find the factors that affect alcohol consumption and how they affect alcohol consumption.

Methodology Association rule mining is used to find the relationship between student alcohol consumption and other attributes.Apriori algorithm is used for association rule mining

1. **Sex**:

Consequent	Antecedent	Support	Confidence
alc=1	sex=M	40.986%	32.707%
alc=1	sex=F	59.014%	7.833%

Males are more than 4 times likely to be heavy drinkers than females.

2. **Goout**:

Consequent	Antecedent	Support	Confidence
alc=1	goout=1	38.675%	32.271%
alc=1	goout=0	61.325%	9.045%

Students who frequently go out with friends are much more likely to be heavy drinkers than those who do not.This suggests peer pressure and social circle may play a role in alcohol consumption habits.

3. **Absences**:

Consequent	Antecedent	Support	Confidence
alc=1	absences=1	7.550%	40.816%
alc=1	absences=0	92.450%	16.167%

Students who frequently missed school were about *2.5* times more likely than regular students to be heavy drinkers.

4. **Failures**:

Consequent	Antecedent	Support	Confidence
alc=1	failures=0	84.592%	16.211%
alc=1	failures=1	10.786%	27.143%
alc=1	failures=2	2.465%	31.250%
alc=1	failures=3	2.157%	28.571%

Those with higher number of school failures tend to drink more.However,beyond a certain point there is no direct increase in alcohol consumption with the number of failures.

5. **Guardian**:

Consequent	Antecedent	Support	Confidence
alc=1	guardian=other	6.317%	31.707%
alc=1	guardian=father	23.575%	18.301%
alc=1	guardian=mother	70.108%	16.703%

As can be seen from the table above,the role of a guardian is very important in influencing the child's alcohol habits.Students with a non-parent guardian were about twice as likely to be heavy drinkers than those having a parent as their guardian.

6. **Famsup**:

Consequent	Antecedent	Support	Confidence
alc=1	famsup=no	38.675%	21.912%
alc=1	famsup=yes	61.325%	15.578%

Students with family educational support tend to be light drinkers when compared to their counterparts lacking educational support.

7. **Medu**:

Consequent	Antecedent	Support	Confidence
alc=1	medu=0	0.924%	33.333%
alc=1	medu=2	48.382%	18.790%
alc=1	medu=1	50.693%	17.021%

Students who have uneducated mothers are about twice as likely than those with educated mothers to be heavy drinkers.However,Fedu (Fathers education) is not a good predictor of drinking habits,suggesting that mothers have a greater role in influencing drinking habits.

8. **Mjob**:

Consequent	Antecedent	Support	Confidence
alc=1	mjob=services	20.955%	21.324%
alc=1	mjob=teacher	11.094%	19.444%
alc=1	mjob=athome	20.801%	17.778%
alc=1	mjob=other	39.753%	17.442%
alc=1	mjob=health	7.396%	10.417%

These observations again show the importance of a mother in upbringing of the child.Those with mothers working in the health sector were less likely to be heavy drinkers.

9. **Nursery**:

Consequent	Antecedent	Support	Confidence
alc=1	nursery=no	19.723%	23.438%
alc=1	nursery=yes	80.277%	16.699%

Students who have not been to nursery are more likely to be heavy drinkers,indicating that receiving early education is important in controlling alcohol habits.

Conclusions

- Some of the factors that make it more likely for a student to be a heavy drinker are:- is a male, goes out with friends frequently, is frequently absent from school, has suffered from past failures, has a non-parent guardian, lacks family educational support, has uneducated mother not working in health sector and has not been to nursery.
- This data can also be used to find ways by which student alcohol consumption can be controlled.These are:
  - Counsel students on combating peer pressure due to friends.
  - Monitor students who miss school regularly.
  - Provide academic counselling to students who have past academic failures.
  - Pay attention to the issue of family educational support and mother's educational status.
  - Ensure that every student gets access to nursery education at the minimum.

2.5 Objective-3

To look into the factors that influence grades and how students can improve their grades.

Methodology Association rule mining and clustering algorithm were used to find logical relations between various attributes and their combinations with grades. Apriori algorithm is used for association rule mining and k-means algorithm was used for clustering.As apriori algorithm works on a presence-absence table, 3 new fields: *grade=0*, *grade=1* and *grade=2* are created to be able to apply this algorithm.

1. **School**:

Consequent	Antecedent	Support	Confidence
grade=0	school=GP	65.177%	2.364%
grade=0	school=MS	34.823%	13.717%
grade=1	school=GP	65.177%	78.960%
grade=1	school=MS	34.823%	73.894%
grade=2	school=GP	65.177%	18.676%
grade=2	school=MS	34.823%	12.389%

GP school students perform better than MS school students.

2. **Medu**:

Cluster number	Cluster size	Average grade	Mother education
1	311	grade=1 - 90%	medu=1 - 96.46%
2	221	grade=1 - 100%	medu=2 - 100%
3	76	grade=2 - 100%	medu=2 - 100 %
4	41	grade=0 - 100%	medu=1 - 60% medu=2 - 40 %

Average silhouette coefficient = 0.9

A student getting the maximum grade of 2 is likely to have a highly educated mother though not all children of highly educated mothers get the maximum grade.Hence mother's education plays a key role in child's performance.

3. **Fjob**:

Consequent	Antecedent	Support	Confidence
grade=0	fjob=at_home	6.471%	9.524%
grade=1	fjob=at_home	6.471%	83.330%
grade=0	fjob=services	27.889%	6.630%
grade=1	fjob=services	27.889%	77.348%
grade=2	fjob=services	27.889%	16.022%
grade=0	fjob=teacher	5.547%	5.556%
grade=1	fjob=teacher	5.547%	55.556%
grade=2	fjob=teacher	5.547%	38.889%
grade=0	fjob=health	3.544%	4.348%
grade=1	fjob=health	3.544%	65.217%
grade=2	fjob=health	3.544%	30.435%

Students who have fathers in the teaching or health sector perform better than other students highlighting the father's role in guiding the child.A similar trend was observed with mother's job also.

4. **Studytime**:

Consequent	Antecedent	Support	Confidence
grade=1	studytime=1	32.666%	83.019%
grade=1	studytime=2	46.995%	75.082%
grade=1	studytime=3	14.946%	73.196%
grade=1	studytime=4	5.393%	71.429%
grade=0	studytime=1	32.666%	9.434%
grade=0	studytime=2	46.995%	6.230%
grade=2	studytime=2	46.995%	18.689%
grade=2	studytime=3	14.946%	25.773%
grade=2	studytime=4	5.393%	25.714%

With an increase in studytime,the student's performance becomes better - less proportion of students with lower grades and greater proportion of students with higher grades.

5. **Goout**

Cluster number	Cluster size	Grade	Average goout
1	501	grade=1 - 100%	goout = 0.39
2	107	grade=2- 100%	goout = 0.33
3	41	grade=0 - 100%	goout = 0.54

Average silhouette coefficient = 0.6

Students generally lose out on their grades, if they spend more time going out with friends. Since study time increases the grades students should keep in control the time they spend going out.

6. **Absence**

Consequent	Antecedent	Support	Confidence
grade=0	absences=0	92.450%	6.167%
grade=0	absences=1	7.550%	8.163%
grade=1	absences=0	92.450%	76.333%
grade=1	absences=1	7.550%	87.755%
grade=2	absences=0	92.450%	17.500%
grade=2	absences=1	7.550%	4.082%

Students attending classes regularly are less likely to get the lowest grade and are more likely to get the highest grade.

Conclusion

- A student of *Gabriel Pereira* having a *highly educated mother* and parents *teaching or in health related work* who spend *more time studying than going out* and *attends school regularly* is highly likely to *perform well and earn good grades.*
- However there is no conclusive indication that attributes like *Internet , family relations, travel time, romantic relationship, school educational support and family support* affect student's performance at school.

3 Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere gratitude to all of them.

We are highly indebted to *Prof. Poonam Goyal* for her guidance and constant supervision as well as for providing necessary information regarding the project.

We are thankful to out laboratory instructor *Ms. Rupal Bhargava* for helping us acquaint with the new technology IBM SPSS Modeler and providing us with hands-on experience on the same.