



University of Technology, Sydney

**32513**

## **Data Mining Algorithms**

### **Assignment Two**

**Date Analytics Project**

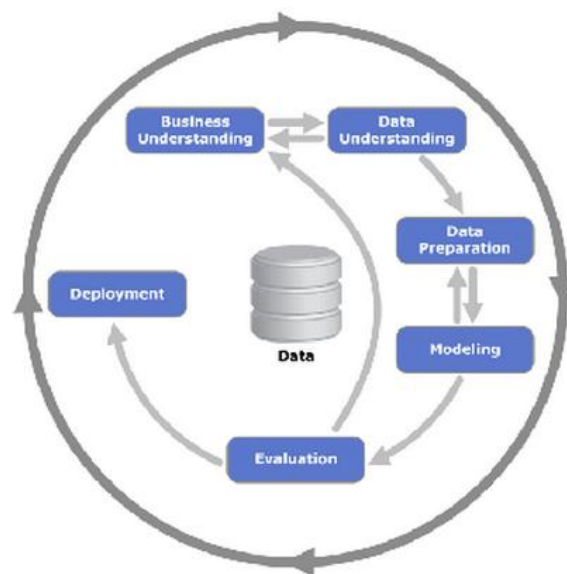
# Content

Introduction .....	3
Business understanding .....	4
Data understanding .....	5
Data collection.....	5
Data description .....	5
Data exploration .....	9
Data preparation.....	12
Discretization.....	12
Missing value .....	13
Modelling .....	15
Naïve Bayes Classifier .....	15
K-Nearest Neighbors .....	19
Decision tree .....	21
Introduction of decision tree:.....	21
Limit of decision tree: .....	21
Data processing .....	22
Modeling:.....	23
Results: .....	24
Evaluation .....	25
Summary .....	25

# Introduction

In generally, company's operational data stored in the data warehouse. Data warehouse is a huge collection of business data. The date is time-variant and non-volatile. Facing such large number of data, how to find out the value is the main challenge in data mining process. Data mining is a critical part in decision making support system. Through the results of data mining, managers could find some insights from the historical business data.

Data mining process are always based on CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process for data mining. There are six steps in this process life cycle which are business understanding, data understanding, data preparation, modeling, evaluation and development. The most important aspects in these steps is data. The following diagram indicates the life cycle of CRISP-DM:



CRISP-DM life cycle

This paper will follow CRISP-DM process to build a classification based on different data mining algorithms to predict the salary attribute value based on input attributes.

# Business understanding

The main business activity of company A is providing financial products to its customers. In order to provide more suitable products to target customers, the CEO wants to know the target customers' level of income. But the salary information is private so it is hard to collect detailed data about customers' salary. Therefore, the CEO want his data analysis team to develop a classification to predict the customers' salary level. The training data is collect from the third part company and the salary level is divided into two parts: salary > 50K and salary <=50K.

In this project, the objective is based on current attributes: age, employment class, education level, education year, marital status, occupation, relationship state, race, sex and work house to predict if the salary over 50K or not. The main activity is that the group of analysts create a classification model for the business objective. The project team will uses different data mining algorithms to create the classifications and according to the predicted salary to determine which data mining algorithm is the most fit in this situation. In terms of the above business analysis, the business objective and the goal of data mining can be identified as following:

Business objective: Develop the best classification to classify salary attribute based on the given data.

Data mining goal: the data mining goal is using the data mining algorithms learned in this semester to create the best classification. Experiencing the whole data mining process in this project.

# Data understanding

## Data collection

The initial data is collected from third party company. As a result, thirteen attributes were collect, there are 33601 records in the initial dataset and there are 4481 missing value in it.

## Data description

After data collection, we need to know the detail information about each attribute, the following table describes the data type and provides the example value for each attributes.

Attribute Name	Attribute Type	Example Value
ID	Ordinal	1,2,3,4,5
Age	Interval	23,45,78
Employment class	Nominal	Private, Local-gov
Education level	Ordinal	1,2,3
Education years	Interval	4,5,6,7
Marital status	Nominal	Never-married, Divorced
Occupation	Nominal	Craft-repair, Exec-managerial
Relationship status	Nominal	Husband, Own-child
Race	Nominal	White, Black
Sex	Nominal	Male, Female
Work hours per week	Interval	70,40,35
Native country	Nominal	United-States, Jamaica
Salary	Ordinal	<=50K,>50K

According to this table, there are three types of attributes which are ordinal, interval and nominal. Ordinal means the values have rank order to describe a kind of degree.

Interval means the values have a fixed size of interval between data points and nominal just some kinds of meaningful string such as name.

After understanding the type of attributes, it is necessary to understand the mean of each attribute and its values.

ID: it is an automatically generated value by system to distinguish one record from others.

Age: it describes the effect of time on a person.

Employment class: it defines the person's working type or where he work at. There are eight value for this attribute.

1. Private: the person work at a private company
2. Sel-emp-not-int: the person has his own unincorporated business and work for it.
3. Sel-emp-int: the person has his own incorporated business and work for it.
4. Federa-gov: the person work at federal government.
5. Loc-gov: the person work at local government.
6. State-gov: the person work at state government.
7. Without-pay: the person work by voluntarily and get no pay.
8. Never-work: the person has no work.

Education level: it describe the highest level of education that achieved by the person, the following is the mean of each values:

Preschool: the person's education level is below 1<sup>st</sup> grade

1<sup>st</sup> - 4<sup>th</sup>: the person's education level is 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> or 4<sup>th</sup> grade

5<sup>th</sup> - 6<sup>th</sup>: the person's education level is 5<sup>th</sup> grade or 6<sup>th</sup> grade

7<sup>th</sup> - 8<sup>th</sup>: the person's education level is 7<sup>th</sup> grade or 8<sup>th</sup> grade

9<sup>th</sup>: the person's education level is 9<sup>th</sup> grade

10<sup>th</sup>: the person's education level is 10<sup>th</sup> grade

11<sup>th</sup>: the person's education level is 11<sup>th</sup> grade

12<sup>th</sup>: the person's education level is 12<sup>th</sup> grade

HS-grad: the person's education level is High school

Some-college: the person graduates from the college which not provides degree

Assoc-voc: Associate degree of vocational program

Assoc-acdm: Associate degree of Academic program

Bachelors: the degree of bachelors

Master: the degree of master

Prof-school: Professional school degree

Doctorate: Doctorate's degree

Education years: It describes how many years that the person spends on education.

Marital status: this attribute define the person's marital status, there are 7 possible values for this attribute:

Married-civ-spouse: it means the person married with a civilian spouse

Never-married: it means the person never married

Divorced: it means the person has married but divorced

Separated: the couple do not live together and in the process for divorce

Widowed: the person is widowed

Married-spouse-absent: the person married with a civilian spouse who is absent

Married-AF-spouse: the person married with Armed Force spouse

Occupation: it defines the person's job, if persons have more than one job, the value will be the job which they work mostly, the values for this attribute is listed as below:

Craft-repair: the person's job belong to manufacture or repair

Prof-specialty: the person has professional specialty certification

Exec-managerial: the person work as a executive, administrator or manager

Adm-clerical: the job is support administrator such as clerical

Sales: the job belong to retail industry

Other-service: the job belong to service industry

Machine-op-inspct: the person work as machine operator or machine inspectors

Transport-moving: the person work in transport industry

Handlers-cleaners: the person work as handlers or equipment cleaner

Farming-fishing: the person work in farming, forestry or fishing industry

Tech-support: the person's work is supporting technicians

Protective-serv: the person's work is protecting something

Priv-house-serv: household service

Armed-Forces: the person work in military area

Relationship status: this attribute describe the current relationship of this person with other householders, the possible values are listed as below:

Husband: he is the husband of a family

Wife: she is the wife of a family

Own-child: the person is child of a family

Not-in-family: the person not the member of a family

Other-relation: the person is relate to the family

Unmarried: the person is unmarried but is a householder

Race: it define the human race of this person, there are only two values in this attribute which are white and black.

Sex: it define the person's gender, just like race, there are only two values: male and female for this attribute.

Work House per week: it describes how many houses the person spends on work during a week.

Native country: it define the person's nationality

Salary: it describes the person's income and this attribute only has two values:  $\leq 50K$  and  $> 50K$

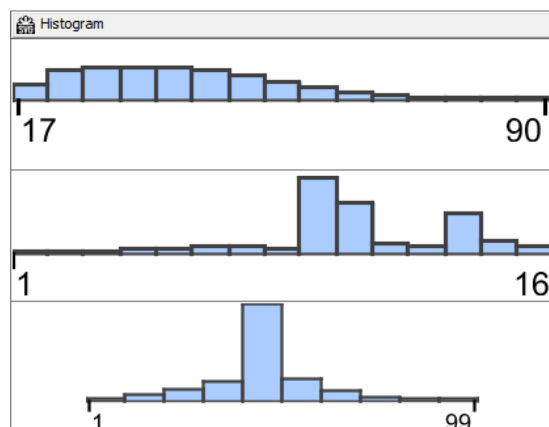


# Data exploration

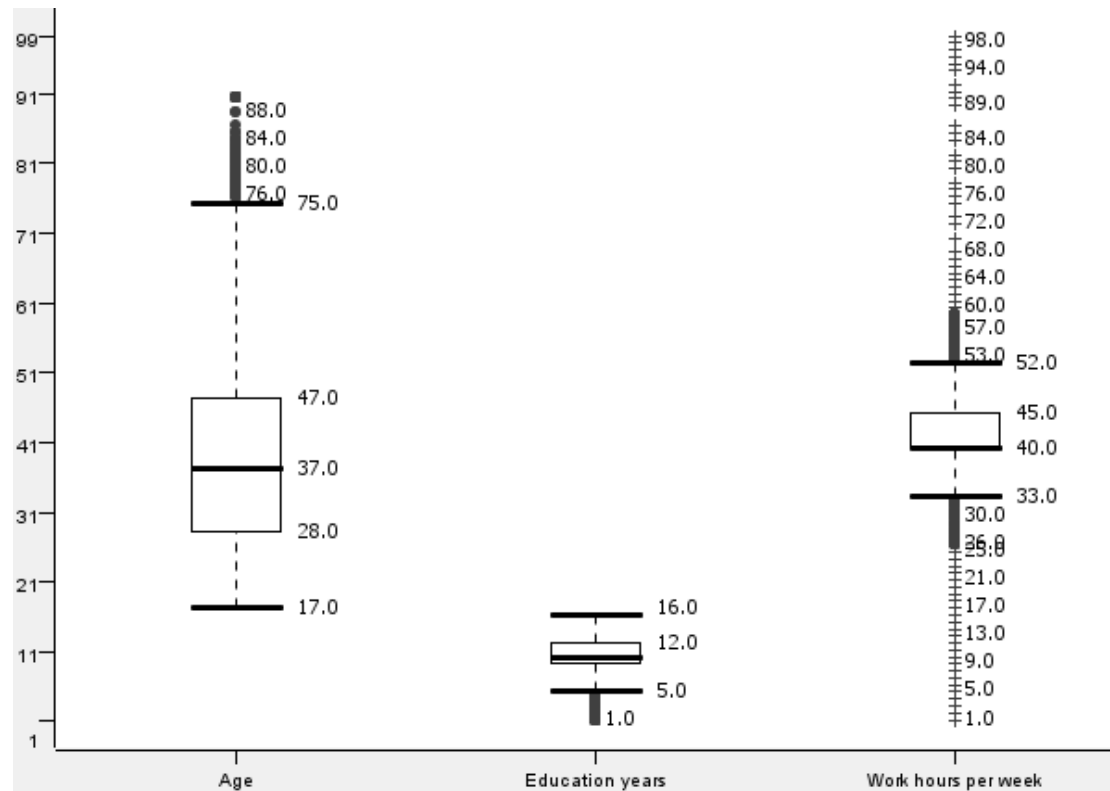
The data exploration will use statistic nod of KNIME. The first step is use CSV reader read in the dataset, then connect these two nods, execute the statistic nod. The statistic nod shows the data information about the interval attributes. Show as below:

Row ID	S Column	D Min	D Max	D Mean	D Std. deviation	D Variance	D Skewness	D Kurtosis
Age	Age	17	90	38.361	13.156	173.073	0.555	-0.101
Education years	Education years	1	16	10.104	2.543	6.464	-0.311	0.662
Work hours p...	Work hours per week	1	99	40.963	11.9	141.616	0.316	3.174

According to this table, the age of oldest person is 90 and for the youngest person is 17. The longest work hours is 99 and the min value is 1 house. The range for education year is between 1 year and 16 years. The following diagram is the histogram of these three attributes

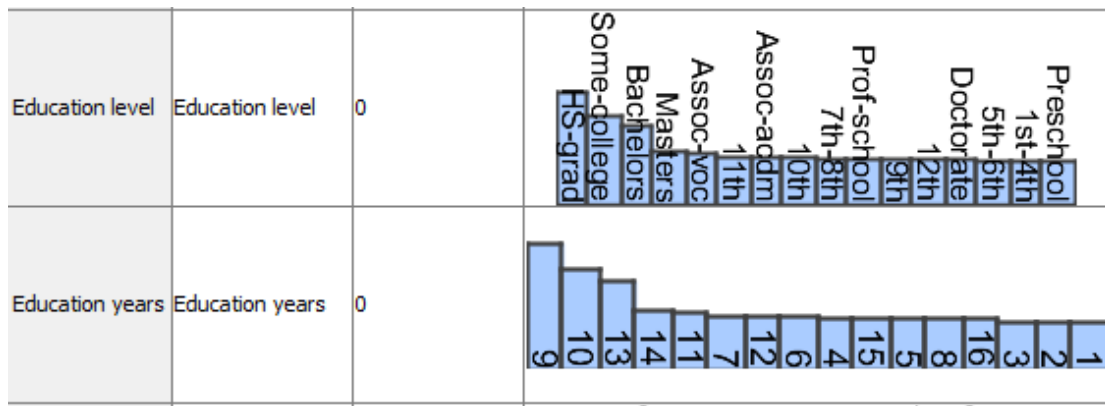


According to this histogram, compare with other two attributes, the distribution of age is uniform. The majority of value of education year are locate in the three particular range. The distribution of working hours is locate in one particular range. Based on the box plot, we can see more detailed information about these three attributes:



In terms of the box plot, half value of age are located between 47 and 28. The half of value of work hours are located between 45 and 50. Compare with other attributes, work hours have more abnormal value,

On the other hand, the histogram of education year and education level is very similar, show as below:



So there is relationship between these two attributes, according this histogram, we find the following relationship:

Education level	Education year
Preschool	1
1st-4th	2
5th-6th	3
7th-8th	4
9th	5
10th	6
11th	7
12th	8
HS-grad	9
Some-college	10
Assoc-voc	11
Assoc-acdm	12
Bachelors	13
Masters	14
Prof-school	15
Doctorate	16

# Data preparation

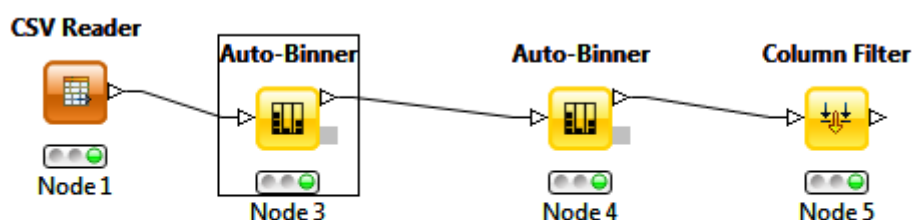
The main purpose of data preparation is making the raw data to be suitable to some data mining algorithms. In order to make the process as simple as possible, the useless attribute or common attribute need to be removed. In this case, the data of ID is not related to this project so it is removed before the modelling process. The education years will indicate the education level and these two attributes are classify the same thing so we just need to keep one of them. In this case, we keep the education level.

## Discretization

Discretise continue value will improve the performance of machine learning.

Discretization is a process which make quantitative value to be qualitative value. So this process could reduce the number of value for the interval attribute. Because the education year and ID are removed so there are only two attributes: work hours and age need to be discretised. The discretization will use bins and 5 bins for age, 10 bins for working hours.

We use two auto-binner nodes to discretise these two attributes and use column filter to see the result, the configuration show as below:



The CSV reader will load the dataset, the first will discretise the age attribute and the second one discretise the working hours attribute. The part of result of discretion is show as below:

Row ID	Age	Work hours per week	Age [Binned]	Work hours per week [Binned]
0	35	35	Bin 2	Bin 4
1	28	40	Bin 1	Bin 4
2	66	40	Bin 4	Bin 4
5	27	30	Bin 1	Bin 3
6	57	40	Bin 3	Bin 4
7	24	38	Bin 1	Bin 4
8	28	70	Bin 1	Bin 8
9	23	50	Bin 1	Bin 5
10	31	40	Bin 1	Bin 4
11	23	20	Bin 1	Bin 2
12	31	40	Bin 1	Bin 4
16	34	60	Bin 2	Bin 7
18	39	40	Bin 2	Bin 4
20	24	40	Bin 1	Bin 4
21	36	40	Bin 2	Bin 4
22	32	40	Bin 2	Bin 4
23	35	40	Bin 2	Bin 4
24	36	40	Bin 2	Bin 4
25	23	50	Bin 1	Bin 5
26	47	40	Bin 3	Bin 4
27	35	40	Bin 2	Bin 4
28	44	40	Bin 2	Bin 4
29	32	50	Bin 2	Bin 5
31	63	10	Bin 4	Bin 1

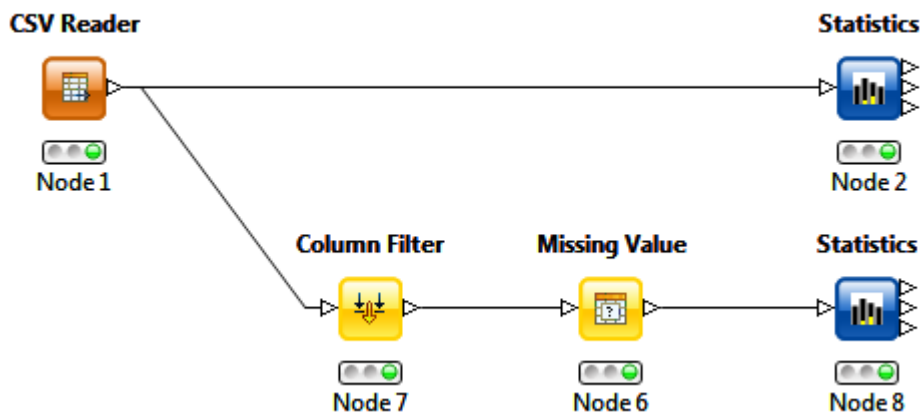
### Missing value

When we get the dataset, we find there are 4481 missing values. These missing values distribute in 12 attributes. In order to avoid the negative effect of missing values, in the data preparation process, we need to deal with these missing values.

There are some methods for dealing with missing value:

1. Using the label of “missing value” to fill the blank.
2. Removing the record which has missing value.
3. Use the most frequent value to fill the blank.

In this case, we will use the third methods to deal with missing values. We use missing value nod to do this process and use statistic nod to compare the results, the configuration show as below:



According to the comparison between two statistics, there are 1932 missing value in employment class and all of them are assign with “private”, in the occupation attribute, there are 1938 missing value and they are assign with “Prof-specialty”, there are 611 missing value in native country attribution and all of them assign with “United-States”.

# Modelling

In this part, we will use several different kinds of algorithms to develop classifier and use confusion matrix to view the result. According to the accuracy of the result to determine which one is the most fit this case.

## Naïve Bayes Classifier

Naïve Bayes classifier is a simple probabilistic classifier which is based on Probability theory. In this process, the model will calculate the possibility for each attributes. For example, there is an attribute called X,  $P(X)$  means the possibility of X is true. Then there is an evidence called E.  $P(X|E)$  means the possibility of the case that evidence E is true and X is true. This is the basic principle of Naïve Bayes algorithm. This algorithm uses supervised learning method to deal with the case involving categorical valued attributes and continuous valued attributes. Supervised learning method is one kind of machine learning task which used to deal with labeled training data. In this case, the platform we use to perform this modeling is KNIME and the primary functional node is Naïve Bayes Learners and Naïve Bayes Predictors.

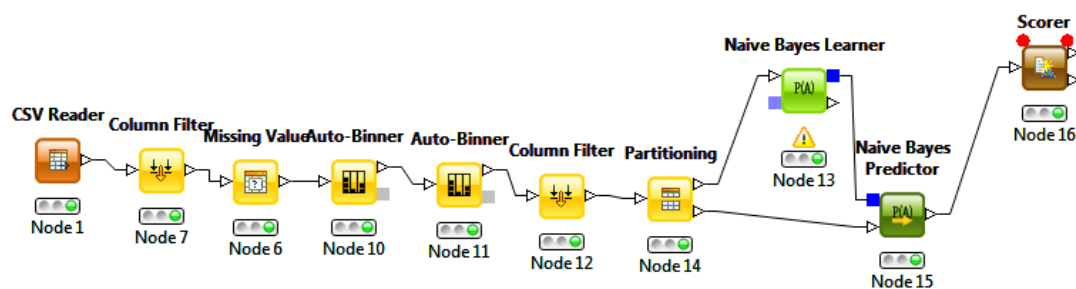
First we use Column Filter node, Auto-Binner node and Missing Value node to deal with the raw dataset for data preparation. Column Filter node will remove education year attribute which is overlapped with education level attribute. Auto-Binner node will reduce the number of value for age attributes and working hours attribute. The value of age will be assigned into 5 bins and the value of working hours attribute will be assigned into 10 bins. This process also called discretise. The principle for binning is showed as below:

Age Range	Bin No.
17-31	Bin 1
32-46	Bin 2
47-60	Bin 3
61-75	Bin 4

76-90	Bin 5
Working Horse	
1 h - 10 h	Bin 1
11 h - 20 h	Bin 2
21 h - 30 h	Bin 3
31 h - 40 h	Bin 4
41 h - 50 h	Bin 5
51 h - 59 h	Bin 6
60 h - 68 h	Bin 7
70 h - 78 h	Bin 8
80 h - 89 h	Bin 9
90 h - 99 h	Bin 10

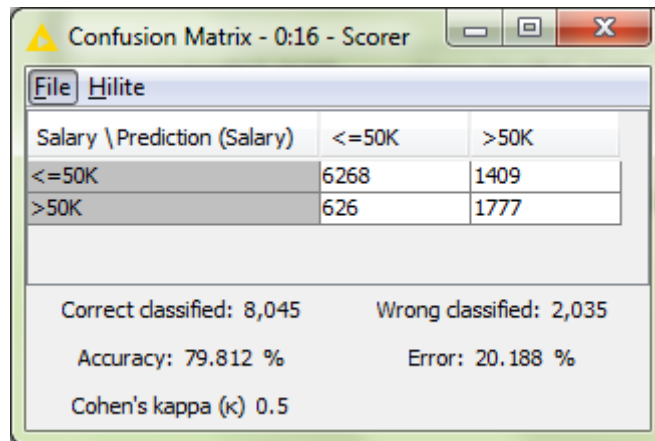
The Missing Value nod will assign the missing value with the most frequent value in the same attributes. After data preparation, the new dataset will be loaded into Partitioning nod, this nod will split dataset into two group randomly. One for Naïve Bayer Learner and another for Naïve Bayer Predictor nod. The whole dataset splitting percentage for learner nod and predictor nod is 70% and 30%. Then we use Scorer nod to see the result.

The configuration is showed as below:



The result will be presented in confusion matrix by Scorer nod. Confusion matrix is a table which describes the performance of classifier.

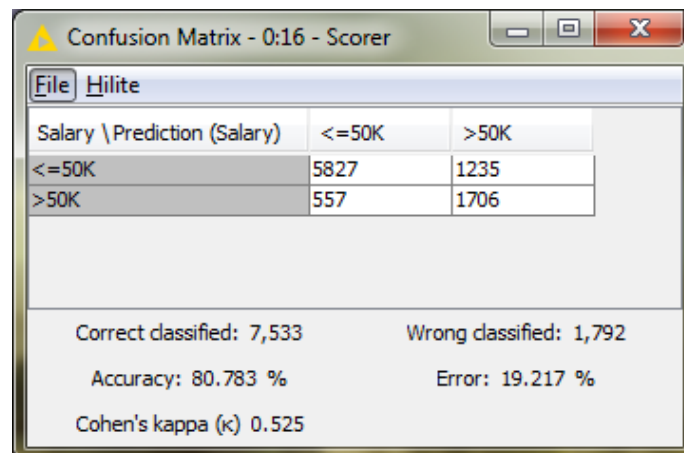




Salary \ Prediction (Salary)	<=50K	>50K
<=50K	6268	1409
>50K	626	1777

Correct classified: 8,045      Wrong classified: 2,035  
 Accuracy: 79.812 %      Error: 20.188 %  
 Cohen's kappa ( $\kappa$ ) 0.5

According to the confusion matrix, the classifier predicts that the salary > 50K is 3186 and salary <= 50K is 6894. But in the reality, the salary > 50K is 2403 and salary <=50 K is 7677. As a result, the accuracy of prediction is 79.812%. According to this accuracy, we think the classifier is not good enough, so we change the process for deal with the missing value, we delete the records which have missing value and executed again, the new confusion matrix is showed below:

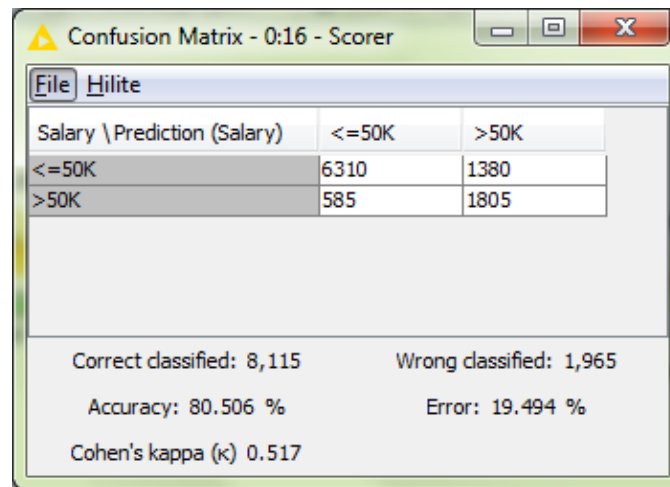


Salary \ Prediction (Salary)	<=50K	>50K
<=50K	5827	1235
>50K	557	1706

Correct classified: 7,533      Wrong classified: 1,792  
 Accuracy: 80.783 %      Error: 19.217 %  
 Cohen's kappa ( $\kappa$ ) 0.525

The new accuracy is 80.783% which is a little higher than the previous one. Therefore, the process of dealing with the missing value will have a little effect on the result. Using most frequent value to deal with the missing value will make the effect of frequent value on the result more significant. Finally, it affect the accuracy of classifier.

We also try the third option, do nothing to the missing values and the new result is showed as below:



According to the three results, the best way is remove the records which contain the missing values, which could make the result be more accurate. The next step we try the different configuration on dealing with the overlapping attributions. The results is showed as below:

condition	remove education level	remove education year	keep both
accuracy	79.882	80	80.172

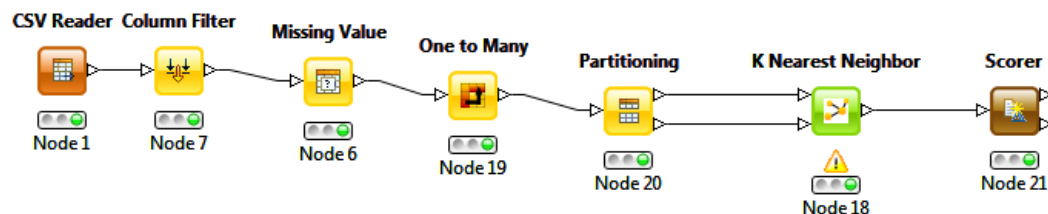
In terms of this table, although education year and education level are overlapping attributes, they are still can make contribution to increase the accuracy of the predicted result.

## K-Nearest Neighbors

K-Nearest neighbors is a simple machine learning algorithm, it is a lazy algorithm which mean it not use any kinds of training data sets to generate classifier. There is no requirement on prior knowledge for data distribution. It classifiers object is based on the closest data samples in the training data set. For example, in our case, if we want to predict a person's salary and the attributions is like: education level is master degree, gender is male, marital status is never-married, race is white, working hours is 60 hours per week. The classifier will look for the records which have the same attributes. If most of their salary is >50K, the predicted result is >50K.

Here we still use KINME to build this classifier. Because K-Nearest Neighbors algorithm only can deal with numeric values, the nominal attributes such as employment class, race and sex need to be transform to scalar. We use "one to many" nod to do this process. Also, the education level and education year indicates the same thing, they are overlapped, so we remove education level in building this classifier.

The model is showed like below:



The CSV reader input the training data set, column filter will remove education level attribute, missing value nod will use the most frequent value to replace the missing values, partitioning nod will divide the training dataset into two parts for classifier building and we use scorer nod to see the result. In this algorithms, K value means how many data sample we selected for analysis. Because there are 33601 records in the training dataset, so for the first time, the K value is 1000, The result is showed below:

Salary \ Class [kNN]	<=50K	>50K
<=50K	6840	234
>50K	1525	726

Correct classified: 7,566      Wrong classified: 1,759  
 Accuracy: 81.137 %      Error: 18.863 %  
 Cohen's kappa ( $\kappa$ ) 0.36

According to the confusion matrix, the classifier predicts the salary > 50K is 960 and in the fact there are 2251 salary >50 K. The classifier predicts the salary <=50K is 8365 and in the fact there are 7074 record are <=50K. The accuracy is 81.137%. The accuracy is unacceptable, so we try different way for deal with missing value and overlap attributes in the same K value to see the result.

Dealing with missing value:

condition	do nothing	use most frequent value	delete the record with missing value
accuracy	80.783	81.577	80.204

Dealing with overlapping attributions

condition	remove education level	remove education year	keep both
accuracy	80.204	77.684	81.308

We also try different K value to find out the highest accuracy. The result is showed below:

K value	300	400	500	600	700	800	900	1000
Accuracy	85.37	84.03	83.46	83.12	82.46	82.05	81.36	81.13

Because K-nearest Neighbors method use the most similar dataset sample to predict target value, using most frequent value to replace the missing value will increase the similar values and results in the increase of accuracy. According to this result, the accuracy is decreasing with the increase of K value. In generally speaking, the number of sample is bigger, the result should be more accurate. So the K-nearest Neighbors is not the suitable method to deal with this problem. Although, this algorithm is simple and easy to implement but its performance is poor when deal with large training dataset.

## **Decision tree**

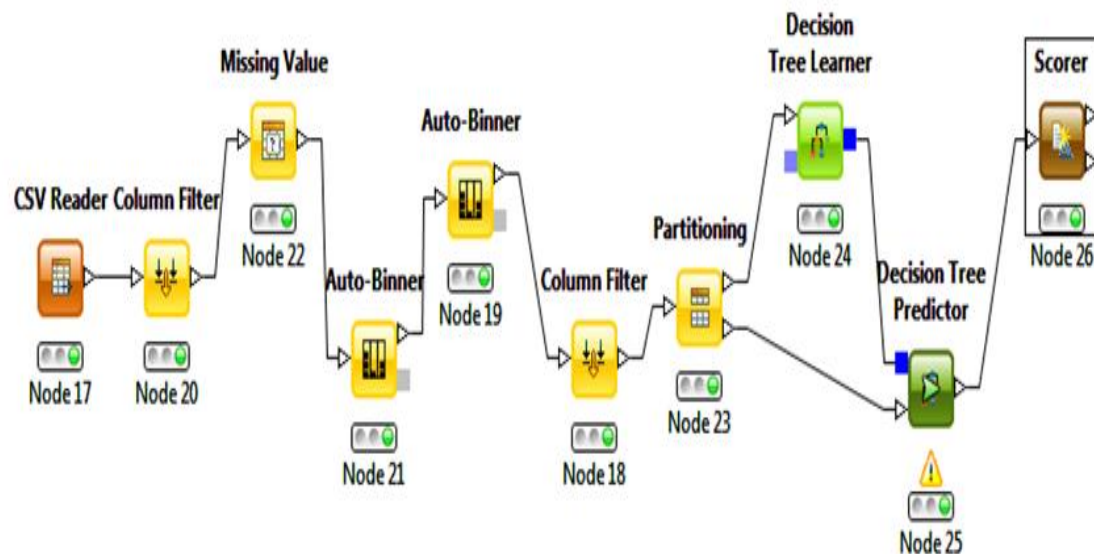
Introduction of decision tree:

In data mining, decision tree is an imitation of the tree structure to form a kind of practice of the method of inductive reasoning. This is a discrete value function approximation method is robust noise data. It showed different characteristics in its structure, each internal node test in typically an attribute; each edge represents the test results; the leaf nodes typically represent classes or class distributions, and the top node is the root node. Decision tree is divided into classification and regression trees. Discrete variables makeup of the classification tree and continuous variables build the regression tree.

Limit of decision tree:

Though the decision tree is widely used, but it also has many shortcomings. First the lacking of scalability, due to its, limited memory, so it would be very difficult to deal with large training set. Second, if we need to deal with some of the continuous variable, or a large data collection, we need to change the algorithm at same time. To a certain extent can raise the cost of data mining at the same time reduces the accuracy of data classification. Especially in a lot of categories of the data, we need a lot of time to build a complete decision tree, which in a certain extent is a waste of human resources and increase the operation cost.

## Data processing



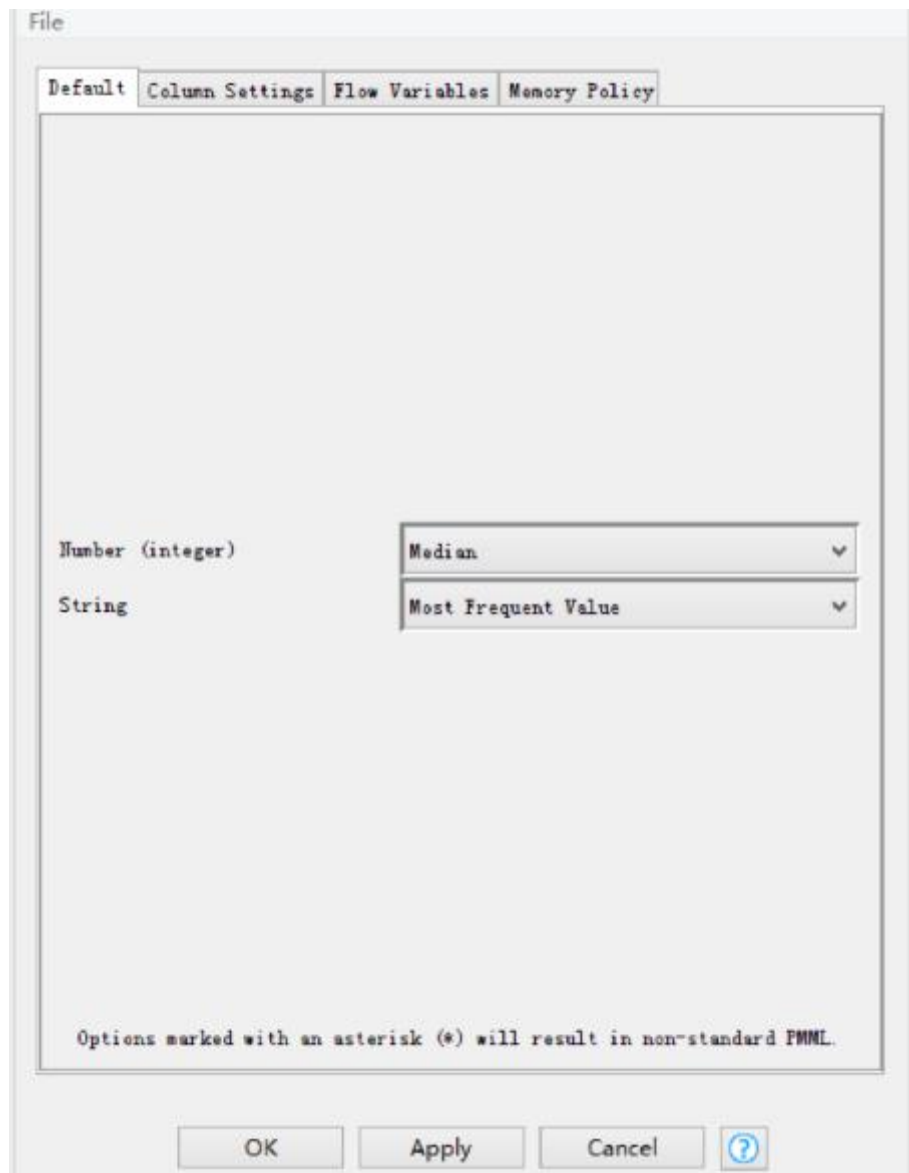
So as the planning and classification according to the data before, we designed a more reasonable decision tree structure of data to us for processing. As shown in the figure, we adopted in decision tree, a binning technology of different kinds of data planning in the bins. Because we think according to the definition of decision tree, this kind of algorithm for data classification is one of the best treatments of French. At the same time, in processing data, we can change the type of data at the same time; finally the result is along with the change.

In this processing, we focus on the maintaining performance not the accuracy.

As you can see in the diagram, we used 2 different bin values to select the useful data in different attributes. By testing the same box and different values, we found the following 2 values have the very high accuracy than others.

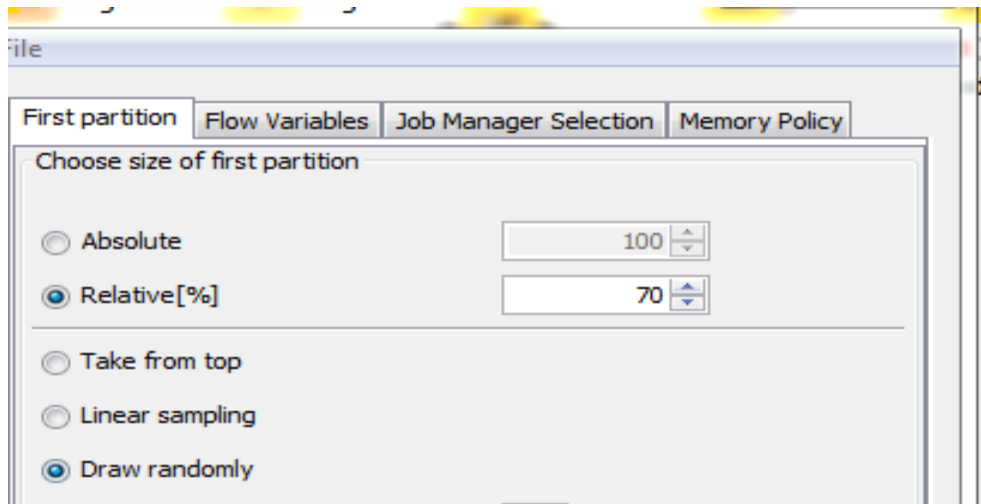
Attribute	Number of bins
Age	5
Work hours	10

There are some other attributes which we didn't put in the bins such as the capital loss, because we put the column filter to select the useful data. After that, we put the missing value to avoid the missing data attributes. If we got some missing values, we can still get the chance to predict that.



Modeling:

Following the diagram, in the data partition part, we spread the data as 30% and 70% as the testing set and the training set which is going to the model learning and model predicting the class.



Results:

In we randomly divided the training data into 65% and 35% data set, the result as shown below.

Confusion Matrix:

Confusion matrix is a table that is used to describe the performance of a classification model or a classifier. In the below shown matrix:

Confusion Matrix - 0:26 - Scorer		
File Hilite		
Salary \ Pr...	<=50K	>50K
<=50K	6941	554
>50K	845	1529
Correct classified: 8,470		Wrong classified: 1,399
Accuracy: 85.824 %		Error: 14.176 %
Cohen's kappa (κ) 0.586		

1. There are two kinds of predicted class: < = 50K and > 50K
2. The classifier made 9686 predictions.



3. Classifier to predict > 50 k, and the < = 50 k, 2083 times and 7786 times.
4. But in reality, there are 3219 people' salary > 50 k and 7495 people < = 50 k.

In addition, through the confusion matrix, the accuracy of the classifier is 85.824%, the error rate of 14.176%. And the cohen's kappa is 0.586

## Evaluation

The table below is the results of the model we used as a summary

classifier	Training set accuracy	Training set error
Naïve Bayes Classifier	79.812%	20.188%
K-Nearest Neighbors	81.137%	18.863%
Decision tree	85.824%	14.176%

Because this time our goal is to choose a better classifier, based on the study, we do the precision of decision tree is one of the best, it's not hard to see from the above results, seems to logistic regression should provide good performance with our own training set, but we lost a large number of testing data in the process of pre-processing of training, it looks like is no effective test data set

Based on two other models, it is easy to see that the decision tree is one of the best choice, then the K-Nearest Neighbors, finally the Naïve Bayes Classifier. Due to time constraints, we can only make three models, there are also a lot of other model may be better than this. Based on our research, however, the decision tree is our final selection model.

And the characteristics of the decision tree are brighter, it can use less data preparation, reduced to some extent this data preparation time. At the same time, it can also be combined with other decision-making techniques, such as PMML technology.

## Summary

This paper is based on the CRISP – DM structure for a population of salary analysis, in the article we use three different classifier contrast, the results are conform to the established before data analysis demand. In the early stage of the data preparation

phase, we have very good planning for different data, and use different classifiers were screened to meet the requirements of our data. Then we record for each classifier, the data processing and the result of that we make it easier to find what kind of classifier is suitable for us.

In the process of experiment, we used the three models, different results are obtained, and by contrast, the decision tree is clearly a more suitable for our model. Because of the characteristics of the decision tree, we can save a lot of time, and at the same time improves the accuracy. Relative to other, it can handle a large number of attributes and estimate the importance of prediction variables, but it can also provide higher results even if there is a large database. And we can change the attributes, to improve performance.

Therefore, the use of in the future, we can apply this model to a number of similar cases based on population statistics, such as the quality of life satisfaction, or customer satisfaction.