



Census Information Exploration

Sidra Anam*

Saurabh Gupta

M.Tech Scholar, CSE Department, Pranveer Singh Institute
of Technology, Kanpur, India

Ass. Prof, CSE Department, Pranveer Singh Institute of
Technology, Kanpur, India

Abstract- Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, neural networks and genetic algorithms [1]. In this paper we have done numerical analysis by taking a “Census Income” dataset [2]. It is a real data of a particular area whose work is to gather all the information regarding the age, workclass, education, education number, marital status, occupation, relation, sex, capital gain, capital loss, hours per week and salary etc. For this we gathered different samples from a particular area of United States. We also inserted some records to make it useful. We found this data in as much as dirty form, that even we can’t apply cleaning tools such as ETL. For upcoming this we manually cleaned it and made it in a form so that we can apply tools to it. This paper is concentrated on the analysis and prediction of income that whether income exceeds \$50K/yr based on this census data.

Keywords: Census Income dataset, ETL tool, PSW Modeler.

I. INTRODUCTION

The whole process of income analysis and prediction consists of some useful and important steps. These are data selection, manually cleaning, cleaning of data by using advance tool, and apply data mining technique to evaluate a result. For cleaning of data we have used ETL tool. For the purpose of data mining we have used PASW Modeler tool [3]. The classification of final result is done by using C5.0 tool. All these tools help to mine this large census dataset and conclude to a result.

II. PROJECT STEPS

The attribute information of census income dataset is as follows:

- Age: continuous.
- Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- Education: Bachelor, 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.
- Sex: Female, Male.
- Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- Capital-gain: continuous.
- Capital-loss: continuous.
- Hours-per-week: continuous.
- Salary: - >50k, <=50k.

A. Step 1: Data selection:

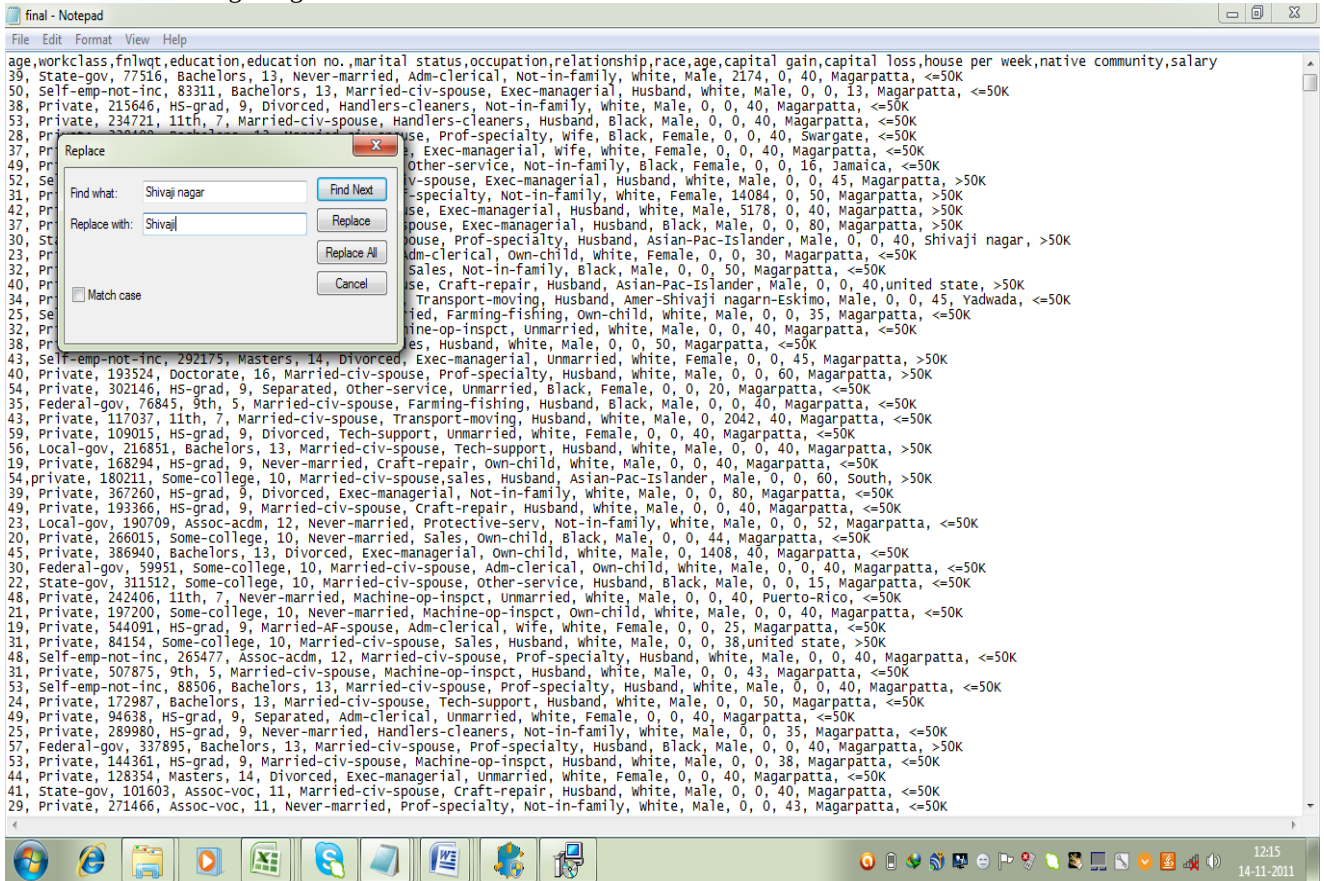
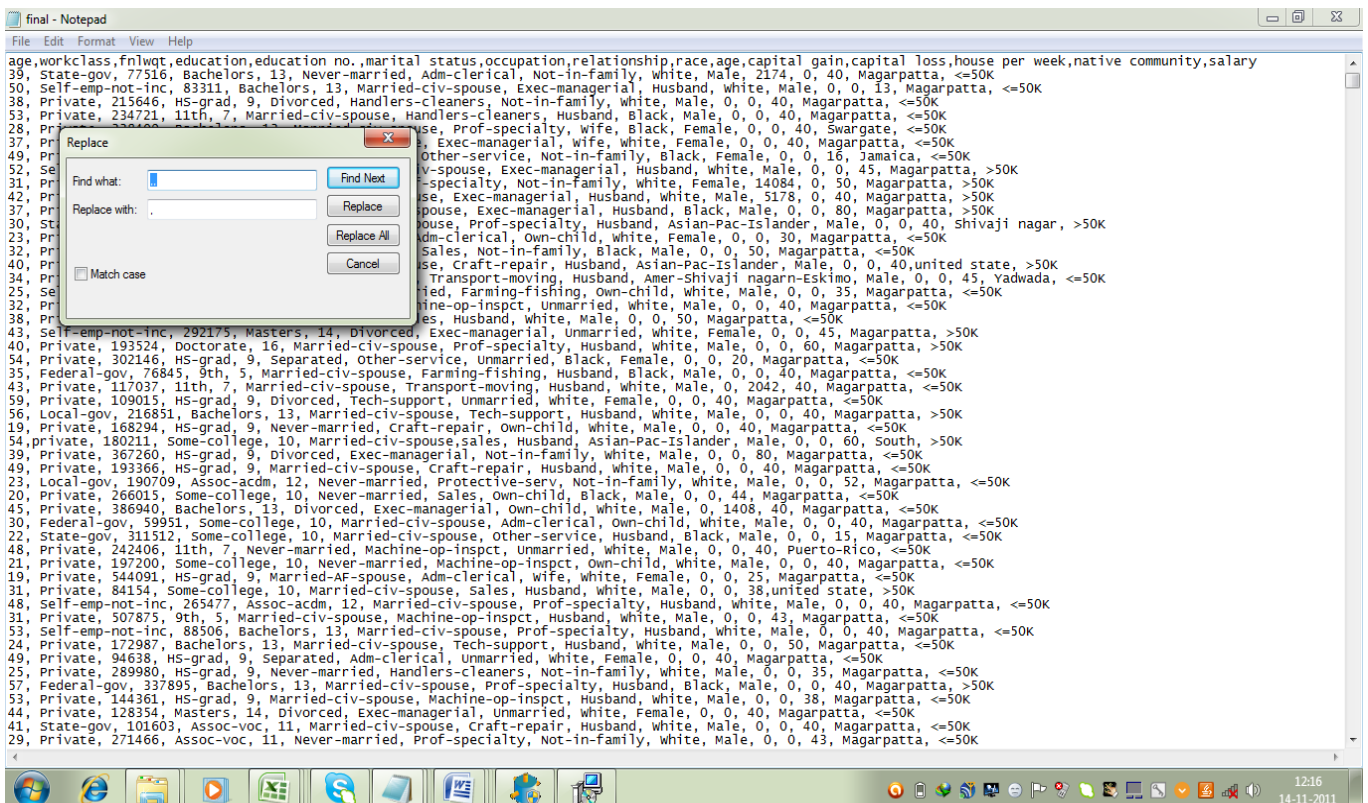
The stage is concerned with selecting the data that are relevant to some criteria. We apply further process in it for mining results.

Raw Data

File Edit Format View Help
age,workclass,fnlwot,education,education no.,marital status,occupation,relationship,race,age,capital gain,capital loss,house per week,native community,salary
39, State-gov, 77526, bachelors, 13, Never-married, Adm-clerical, Not-in-family, white, Male, 2174, 0, 40, Magarpatta, <=50K
50, Self-emp-not-inc, 83131, bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, white, Male, 0, 0, 40, Magarpatta, <=50K
50, Private, 213646, HS-grad, 9, Divorced, handlers-cleaners, Not-in-family, white, Male, 0, 0, 40, Magarpatta, <=50K
51, Private, 234723, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, black, Male, 0, 0, 40, Magarpatta, <=50K
28, Private, 338409, bachelors, 13, Married-civ-spouse, Prof-specialty, wife, black, Female, 0, 0, 40, Swargate, <=50K
37, Private, 284282, Masters, 14, Married-civ-spouse, Exec-managerial, wife, white, Female, 0, 0, 40, Magarpatta, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 208642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, white, Male, 0, 0, 45, Magarpatta, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, white, Female, 14084, 0, 50, Magarpatta, >50K
42, Private, 158449, bachelors, 13, Married-civ-spouse, Exec-managerial, husband, white, Male, 5178, 0, 40, Magarpatta, >50K
37, Private, 280464, some-college, 10, Married-civ-spouse, Exec-managerial, Husband, black, Male, 0, 0, 80, Magarpatta, >50K
30, State-gov, 141297, bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, male, 0, 0, 40, shivaji nagar, >50K
23, Private, 122272, bachelors, 13, Never-married, Adm-clerical, Own-child, white, female, 0, 0, 30, Magarpatta, <=50K
32, Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, white, Male, 0, 0, 50, Magarpatta, <=50K
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, united state, >50K
34, Private, 135242, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Shivaji nagar-Eskimo, Male, 0, 0, 45, Yadwada, <=50K
25, Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, white, Male, 0, 0, 35, Magarpatta, <=50K
32, Private, 160919, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, white, Male, 0, 0, 40, Magarpatta, <=50K
28, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, white, Male, 0, 0, 50, Magarpatta, <=50K
43, Self-emp-not-inc, 292175, Masters, 14, Divorced, Exec-managerial, Unmarried, white, Female, 0, 0, 45, Magarpatta, >50K
35, Private, 139324, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, white, Male, 0, 0, 60, Magarpatta, >50K
54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, black, Female, 0, 0, 20, Magarpatta, <=50K
53, Federal-gov, 109013, HS-grad, 9, Divorced-civ-spouse, Farming-fishing, Husband, black, Male, 0, 0, 40, Magarpatta, <=50K
43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, white, Male, 0, 2042, 40, Magarpatta, <=50K
59, Private, 109013, HS-grad, 9, Divorced, Tech-support, Unmarried, white, Female, 0, 0, 40, Magarpatta, <=50K
56, Local-gov, 216851, bachelors, 13, Married-civ-spouse, Tech-support, Husband, white, Male, 0, 0, 40, Magarpatta, >50K
19, Private, 180211, some-college, 10, Married-civ-spouse, Sales, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, >50K
54, Private, 180211, some-college, 10, Married-civ-spouse, Sales, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, >50K
39, Private, 190709, HS-grad, 9, Divorced, Exec-managerial, Not-in-family, white, Male, 0, 0, 40, Magarpatta, <=50K
49, Private, 193166, HS-grad, 9, Married-civ-spouse, Craft-repair, Husband, white, Male, 0, 0, 40, Magarpatta, <=50K
23, Local-gov, 190709, Assoc-acdm, 12, Never-married, Protective-serv, Not-in-family, white, Male, 0, 0, 52, Magarpatta, <=50K
20, Private, 266015, some-college, 10, Never-married, Sales, Own-child, black, Male, 0, 0, 44, Magarpatta, <=50K
48, Private, 186940, Assoc-acdm, 12, Never-married, Protective-serv, Not-in-family, white, Male, 0, 0, 40, Magarpatta, <=50K
30, Federal-gov, 59931, some-college, 10, Married-civ-spouse, Adm-clerical, Own-child, white, Male, 0, 0, 40, Magarpatta, <=50K
27, State-gov, 111512, some-college, 10, Married-civ-spouse, Other-service, Husband, black, Male, 0, 0, 40, Magarpatta, <=50K
48, Private, 242406, 11th, 7, Never-married, Machine-op-inspct, Unmarried, white, Male, 0, 0, 40, Puerto-Rico, <=50K
21, Private, 8444, HS-grad, 9, Separated, Adm-clerical, Unmarried, white, Female, 0, 0, 40, Magarpatta, <=50K
19, Private, 544091, HS-grad, 9, Married-AF-spouse, Adm-clerical, wife, white, Female, 0, 0, 25, Magarpatta, <=50K
31, Private, 8444, HS-grad, 9, Separated, Adm-clerical, Unmarried, white, Female, 0, 0, 40, Magarpatta, <=50K
48, Self-emp-not-inc, 265477, Assoc-acdm, 12, Married-civ-spouse, Prof-specialty, Husband, white, Male, 0, 0, 40, Magarpatta, <=50K
31, Private, 507875, 9th, 5, Married-civ-spouse, Machine-op-inspct, Husband, white, Male, 0, 0, 43, Magarpatta, <=50K
51, Self-emp-not-inc, 88506, bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, white, Male, 0, 0, 40, Magarpatta, <=50K
24, Private, 172987, bachelors, 13, Married-civ-spouse, Tech-support, Husband, white, Male, 0, 0, 50, Magarpatta, <=50K
49, Private, 172987, HS-grad, 9, Separated, Adm-clerical, Unmarried, white, Female, 0, 0, 40, Magarpatta, <=50K
25, Private, 289980, HS-grad, 9, Never-married, Handlers-cleaners, Not-in-family, white, Male, 0, 0, 35, Magarpatta, <=50K
57, Federal-gov, 174993, bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, black, Male, 0, 0, 40, Magarpatta, <=50K
51, Private, 144161, HS-grad, 9, Married-civ-spouse, Machine-op-inspct, Husband, white, Male, 0, 0, 38, Magarpatta, <=50K
44, Private, 128154, Masters, 14, Divorced, Exec-managerial, Unmarried, white, Female, 0, 0, 40, Magarpatta, <=50K
41, State-gov, 101603, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, white, Male, 0, 0, 40, Magarpatta, <=50K
29, Private, 271466, Assoc-voc, 11, Never-married, Prof-specialty, Not-in-family, white, Male, 0, 0, 43, Magarpatta, <=50K

B. Step 2: Manually cleaning:

Our data is in text format which is very dirty so before applying the cleaning process we have to manually clean the data. So for this following things are done:

a. Replace some data by other which go in another column because of extra space:**b. Removing extra commas from data:**

C. Step 3: Cleaning of data by using Advance ETL tool

a. Extraction:

Read the data

The screenshot shows the 'final.Lats - Advanced ETL Processor' window. The 'Data Flow Diagram' tab is active, displaying a workflow: Reader (highlighted with a red box) → Validator → Transformer → Writer. Below the diagram, the 'Reader' and 'Writer' properties are shown. The 'Reader' property is set to 'C:\Users\Indra\Desktop\final\dm\try\final.txt'. The 'Writer' property is empty. Below the properties, a data table is displayed with 15 columns: AGE, WORKCLASS, FILWOT, EDUCATION, EDUCATION, MARITAL, STATUS, OCCUPATION, RELATIONSHIP, RACE, AGE, CAPITAL, CAPITAL, HOUSE, PERCENT, NATIVE, COMMUNITY, and SALARY. The table contains 19 rows of data.

AGE	WORKCLASS	FILWOT	EDUCATION	EDUCATION	MARITAL	STATUS	OCCUPATION	RELATIONSHIP	RACE	AGE	CAPITAL	CAPITAL	HOUSE	PERCENT	NATIVE	COMMUNITY	SALARY
1	workclass	education	education	marital	s	occupati	relations	race	age	capital	q	capital	house	per	native	c	salary
2	39	State-go	Bachelor	13	Never-ma	Adm-cle	Not-in-f	White	Male	2174	0	40	Maqarpa	<=50K			
3	50	Self-emp	83311	Bachelors	13	Married	Exec-man	Husband	White	Male	0	0	13	Maqarpatta	<=50K		
4	38	Private	HS-grad	9	Divorced	Handler	Not-in-f	White	Male	0	0	40	Maqarpa	<=50K			
5	53	Private	234721	11th	7	Married	Handlers	Husband	Black	Male	0	0	40	Maqarpatta	<=50K		
6	28	Private	338409	Bachelors	13	Married	Prof-spe	Wife	Black	Female	0	0	40	Swagrate	<=50K		
7	37	Private	Masters	14	Married	Exec-ma	Wife	White	Female	0	0	40	Maqarpa	<=50K			
8	49	Private	160187	9th	5	Married	Other-se	Not-in-f	Black	Female	0	0	16	Jamaica	<=50K		
9	52	Self-emp	209642	HS-grad	9	Married	Exec-man	Husband	White	Male	0	0	45	Maqarpatta	>50K		
10	31	Private	Masters	14	Never-ma	Prof-sp	Not-in-f	White	Female	14084	0	50	Maqarpa	>50K			
11	42	Private	Bachelor	13	Married	Exec-ma	Husband	White	Male	5178	0	40	Maqarpa	>50K			
12	37	Private	Some-col	10	Married	Exec-ma	Husband	Black	Male	0	0	80	Maqarpa	>50K			
13	30	State-go	141297	Bachelors	13	Married	Prof-spe	Husband	Asian-P	Male	0	0	40	Shivai naqar	>50K		
14	23	Private	122272	Bachelors	13	Never-m	Adm-cler	Own-chil	White	Female	0	0	30	Maqarpatta	<=50K		
15	32	Private	205019	Assoc-acd	12	Never-m	Sales	Not-in-f	Black	Male	0	0	50	Maqarpatta	<=50K		
16	40	Private	Assoc-vo	11	Married	Craft-r	Husband	Asian-Pa	Male	0	0	40	united s	>50K			
17	34	Private	7th-8th	4	Married	Transpo	Husband	Amer-Shi	Male	0	0	45	Yadvada	<=50K			
18	25	Self-emp	176756	HS-grad	9	Never-m	Farmin	Own-chil	White	Male	0	0	35	Maqarpatta	<=50K		
19	32	Private	186824	HS-grad	9	Never-m	Machine	Unmarrie	White	Male	0	0	40	Maqarpatta	<=50K		

Set Reader Property

The screenshot shows the 'final.Lats - Advanced ETL Processor' window with the 'Data Reader Properties' dialog box open. The 'General' tab is selected. The 'Source File is:' section has 'Delimited' selected. The 'File Encoding' is set to 'Auto'. The 'Delimiter' is set to ','. The 'Text Qualifier' is set to '"'. The 'Line Terminators' section has '(CR)+(LF)', '(EOF)', '(LF)', and '(LF)+(CR)' listed. The 'Lines to Skip' section has 'Header' set to 1 and 'Footer' set to 0. The 'Use Mask to find files to load' and 'Search in Subdirectories' checkboxes are unchecked. The 'OK' button is highlighted.

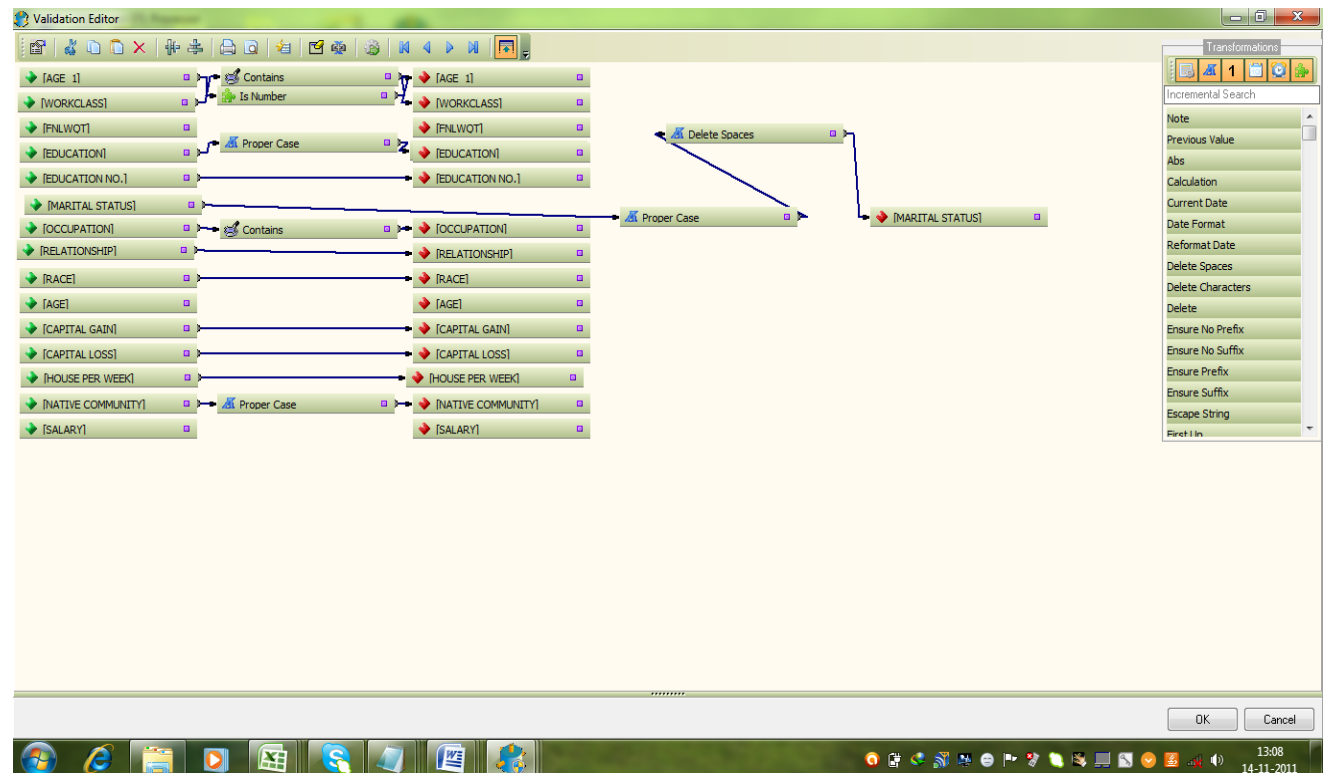
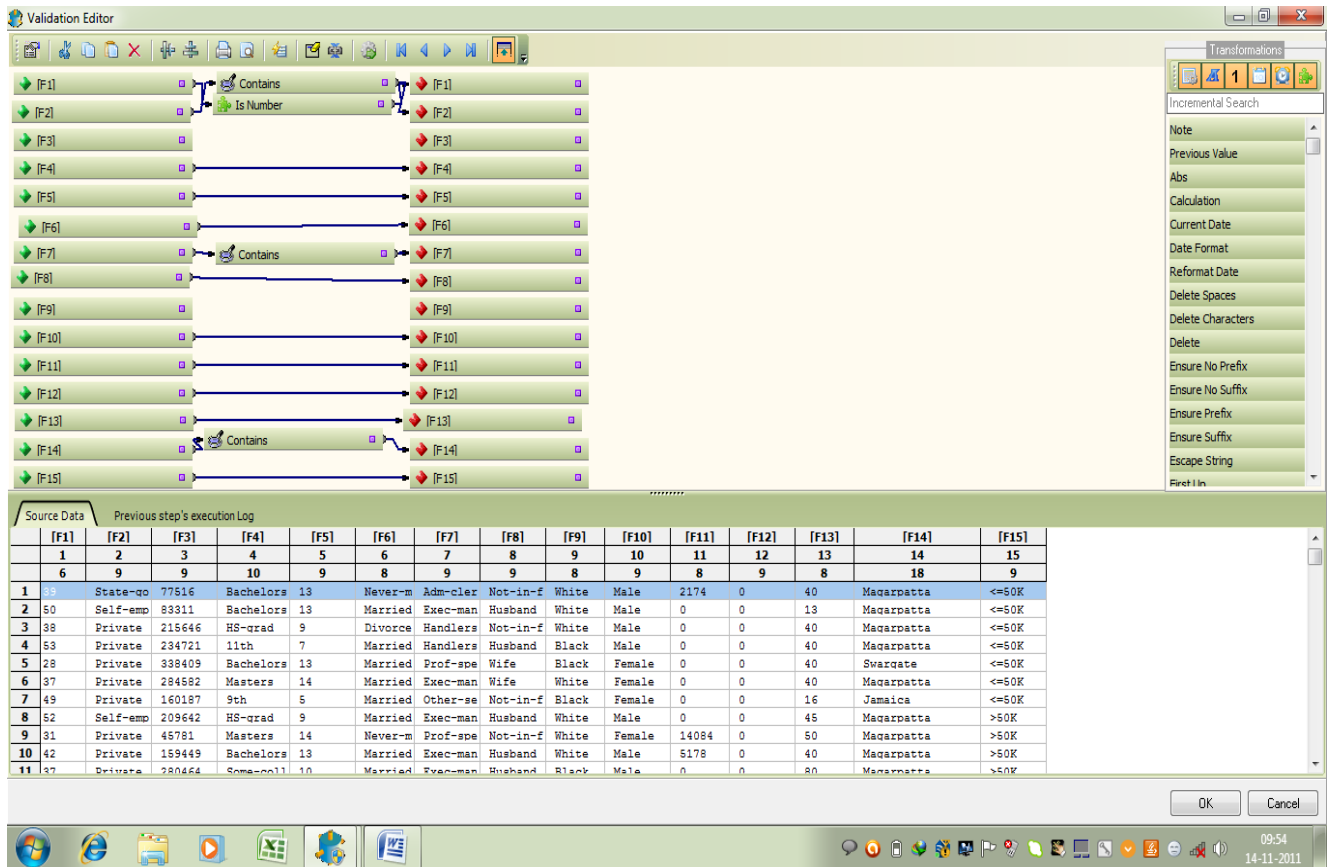
b. Validation:

Now we have applied the ETL tool for the cleaning process for the cleaning the data for:

- Convert age field which contains? Value, set it to correct value.

- Convert workclass field which contains? Value, set it to correct value.

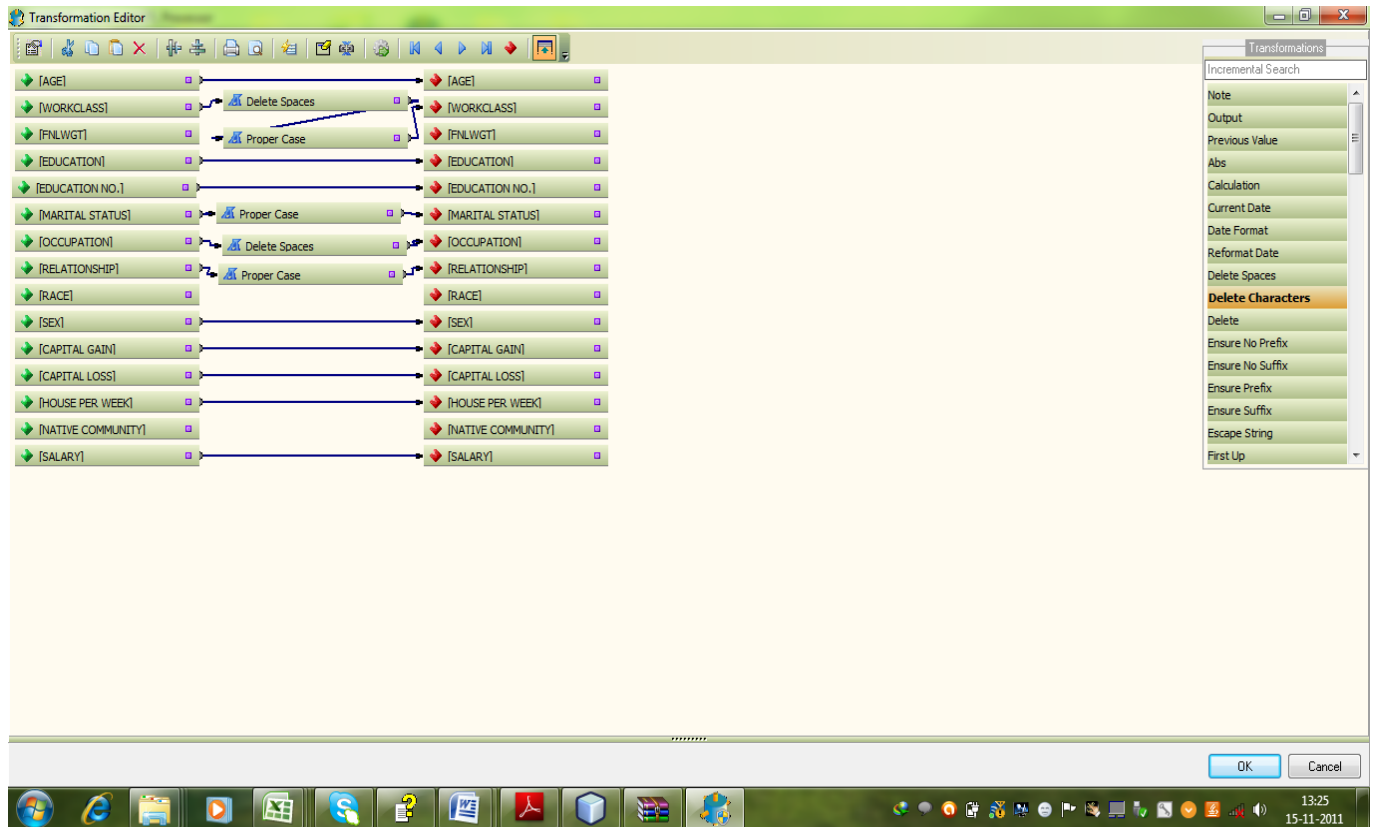
- Set occupation field to particular value.



c. Transformation of data:

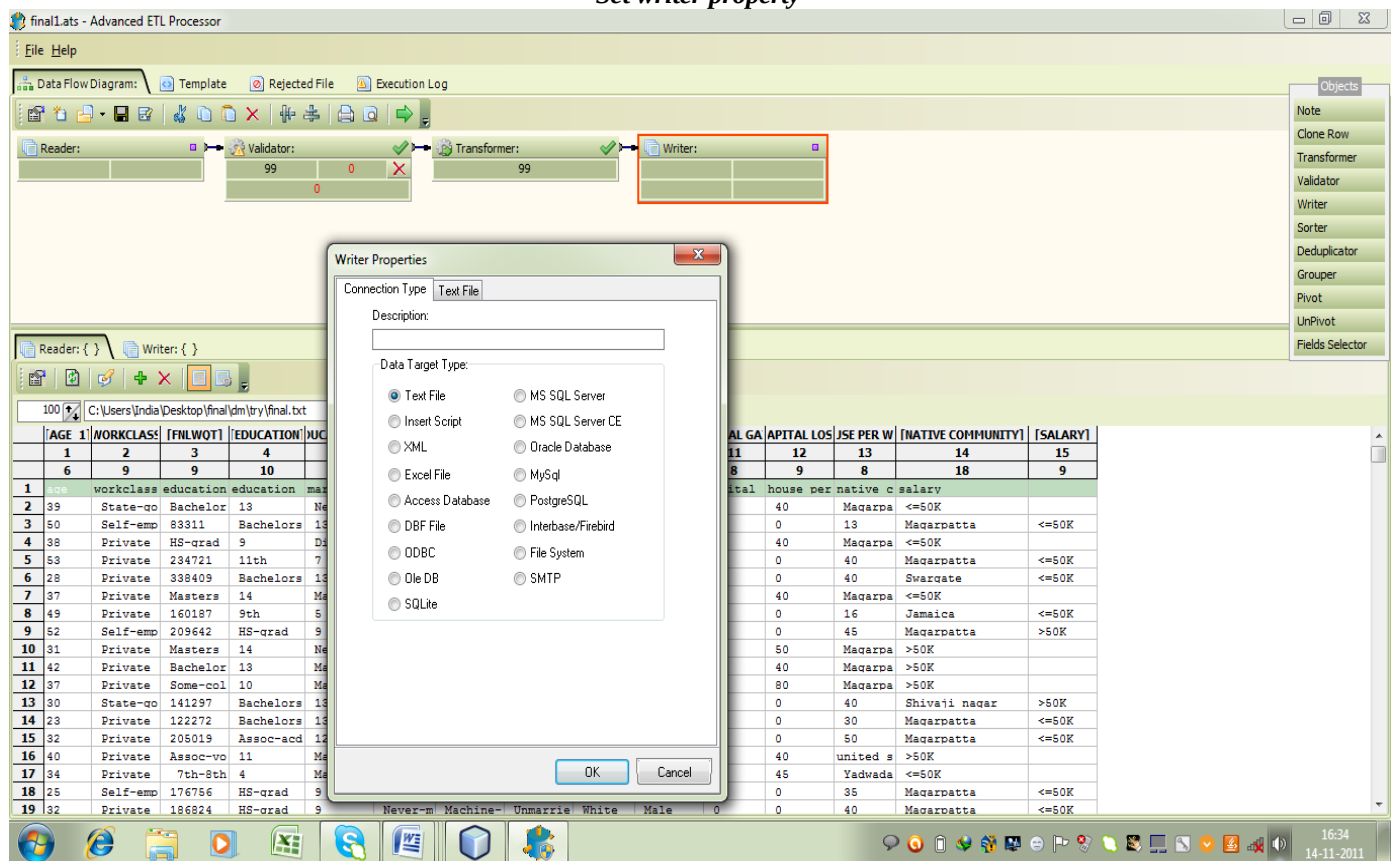
The data is not merely transferred across, but transformed in order to be suitable for

the task of data mining. In this stage, the data is made usable and navigable.



d. Loading:

Set writer property



Load file into its destination

The screenshot shows the 'finalLats - Advanced ETL Processor' application. The 'Data Flow Diagram' tab is active, displaying a workflow with a Reader, Validator, Transformer, and Writer. The Reader is connected to a file 'C:\Users\Indra\Desktop\final\dm\try\final.txt'. The Validator and Transformer are connected to the Reader. The Writer is connected to the Transformer. The Writer's output is shown in a table with columns 'AGE', 'WORKCLASS', 'EDUCATION', 'EDUCATION', and 'MARITAL'. The Writer's status is '0' and '0'.

A 'Processing Data' dialog box is open, showing the source file 'C:\Users\Indra\Desktop\final\dm\try\final.txt' and the progress of the transformation. The progress bar is at 68%. The dialog box also shows a list of records being processed, including columns like 'Line No', 'Type', 'Description', 'Map Name', 'Author', 'Version', 'Description', 'Preparing Writers...', 'Writer: {} Is Ready', and 'All Writers are Ready'.

Execution Log

The screenshot shows the 'finalLats - Advanced ETL Processor' application with the 'Execution Log' tab active. The log displays a list of records with columns 'Line No', 'Type', 'Time', 'Record', 'Object', and 'Description'. The log shows the progress of the transformation, including the start and end of the transformation, the preparation of writers, and the completion of the transformation. The log also shows the status of the writer, including 'Nothing is Mapped', 'Starting Transformation...', 'Map Name:', 'Author:', 'Version:', 'Description:', 'Preparing Writers...', 'Nothing is Mapped', 'Starting Transformation...', 'Map Name:', 'Author:', 'Version:', 'Description:', 'Preparing Writers...', 'Target text file name is blank', 'Starting Transformation...', 'Map Name:', 'Author:', 'Version:', 'Description:', 'Preparing Writers...', 'Writer: {} Is Ready', and 'All Writers are Ready'.

Downloaded from <http://ajph.org/> on November 10, 2015

[Convert it into excel file](#)

final - Microsoft Excel

Step 4: Apply data mining technique to evaluate a

Data mining the extraction of hidden predictive

...and the other is the fact that the ...

experts may miss because it lies outside their expectations [10]. There are various technique available on data mining in this project use a clustering and classification described one by one later.

a. Data mining techniques:

There are many data mining tools and techniques. Here, we have used **PASW Modeler** tool. PASW Modeler is a data mining workbench that enables you to quickly develop predictive models using business expertise and deploy them into business operations to improve decision making [5]. Designed around the industry-standard CRISP-DM model, PASW Modeler supports the entire data mining process, from data to better business results. The benefits of using this tool are as follows:

- a) **Accessible and Relevant:** Increase analyst productivity with simplified data mining that actively engages and seamlessly guides them through the analytics process. PASW Modeler's easy-to-learn visual workflow interface requires no programming skills, reducing the learning curve and making the power of analytics accessible to expert and novice alike.
- b) **Predictive Analytics:** Gain a quick competitive advantage with the best models of future behaviour. With time saving features like automated modeling,

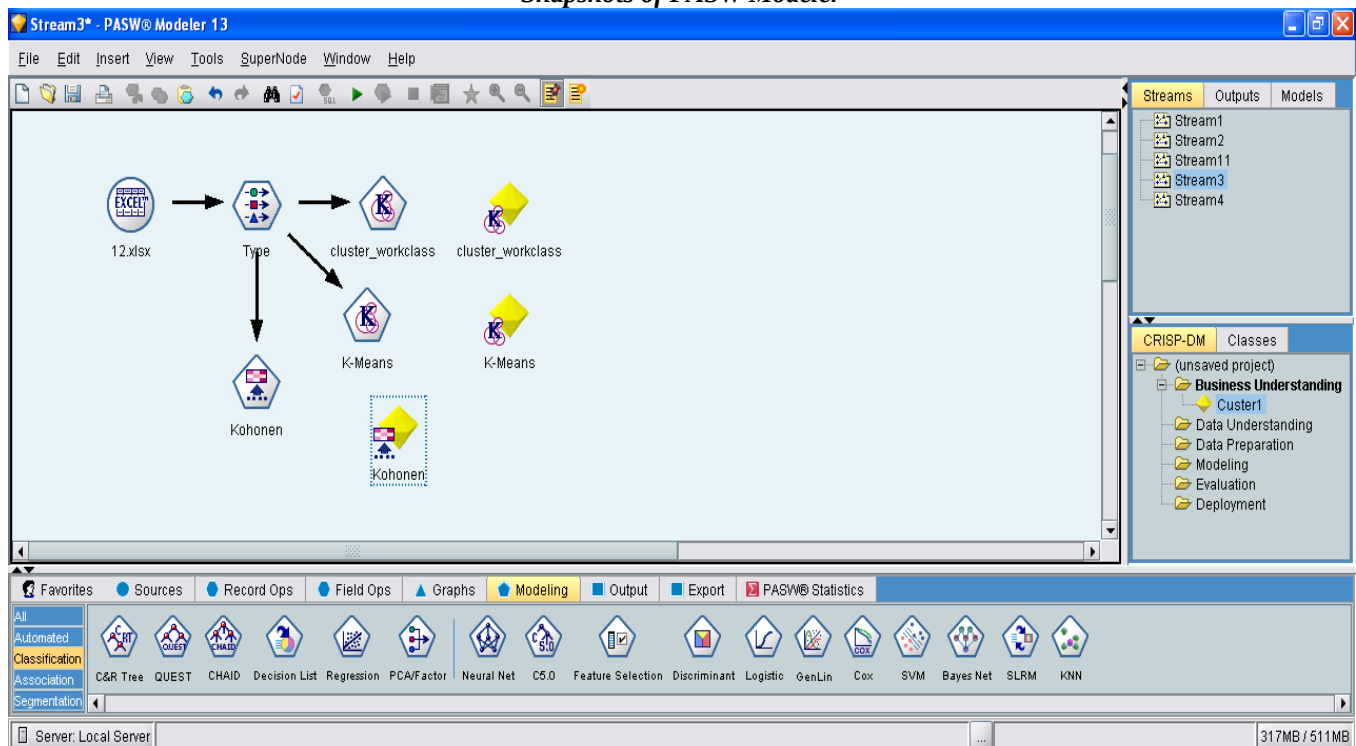
advanced data preparation, and ensemble modeling, PASW Modeler can give you quick solutions and improved returns.

- c) **Adaptable:** Reduce costs and maximize your technology with data mining that leverages your existing infrastructure. PASW Modeler's open and scalable architecture allows many procedures to take place within a core database, including access to embedded algorithms. This can help maximize your database for enhanced performance and speed.

(a). Clustering:

Clustering is the process of grouping physical or abstract objects into classes of similar objects [6]. Snapshots of the PASW Modeler activity of clustering in which cluster on the basics of sex ratio of male and female and another task is to cluster on the basics of education in which calculate the number of people 9th, 10th, 11th, bachelor, master or related to other education term. And also cluster on the basics of workclass how many employees are in government job and how many in private jobs in which we apply the Kohonen for clustering of sex, k-mean for clustering the education and K-mean for clustering the workplace.

Snapshots of PASW Modeler

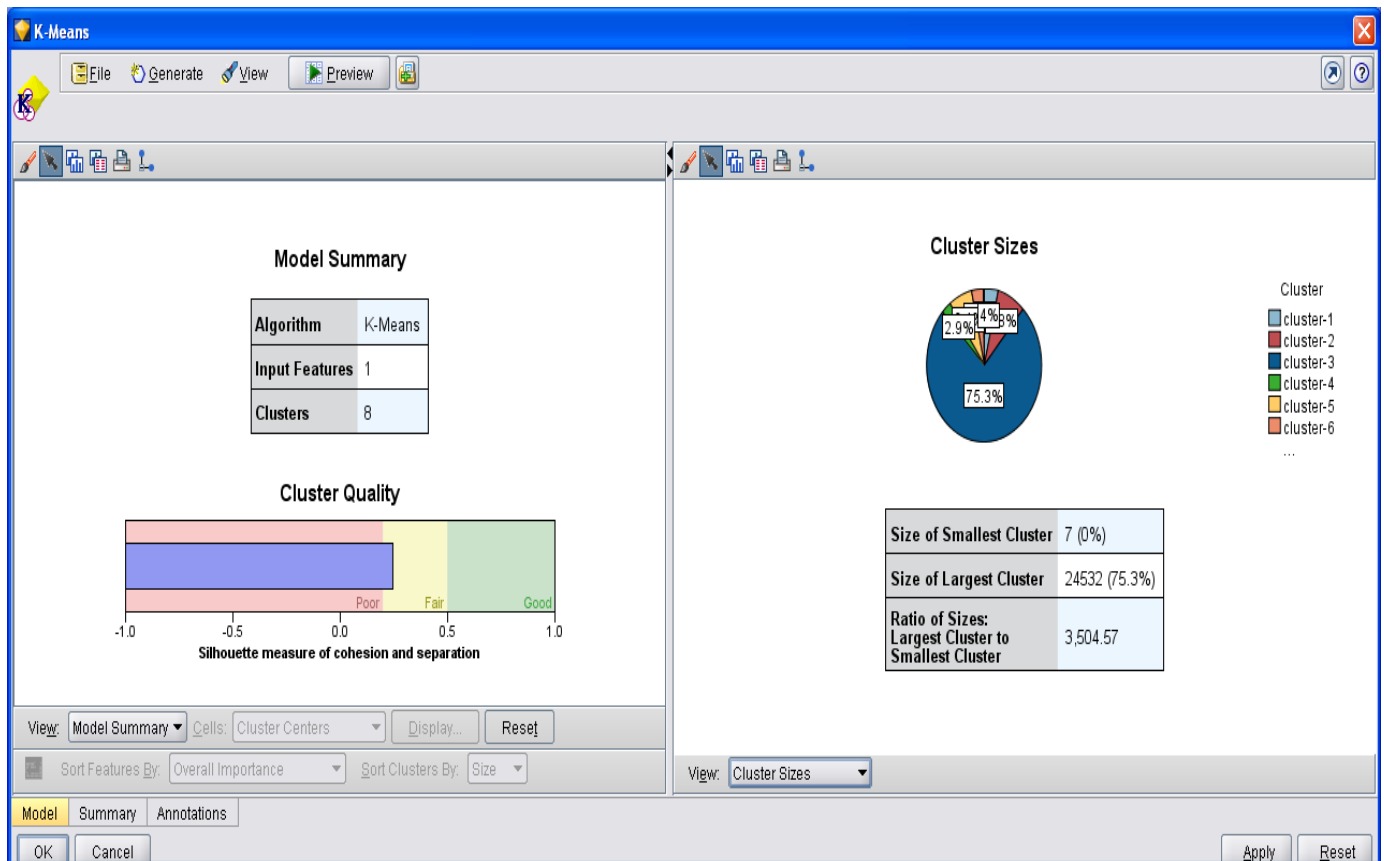


i. k-means algorithm:

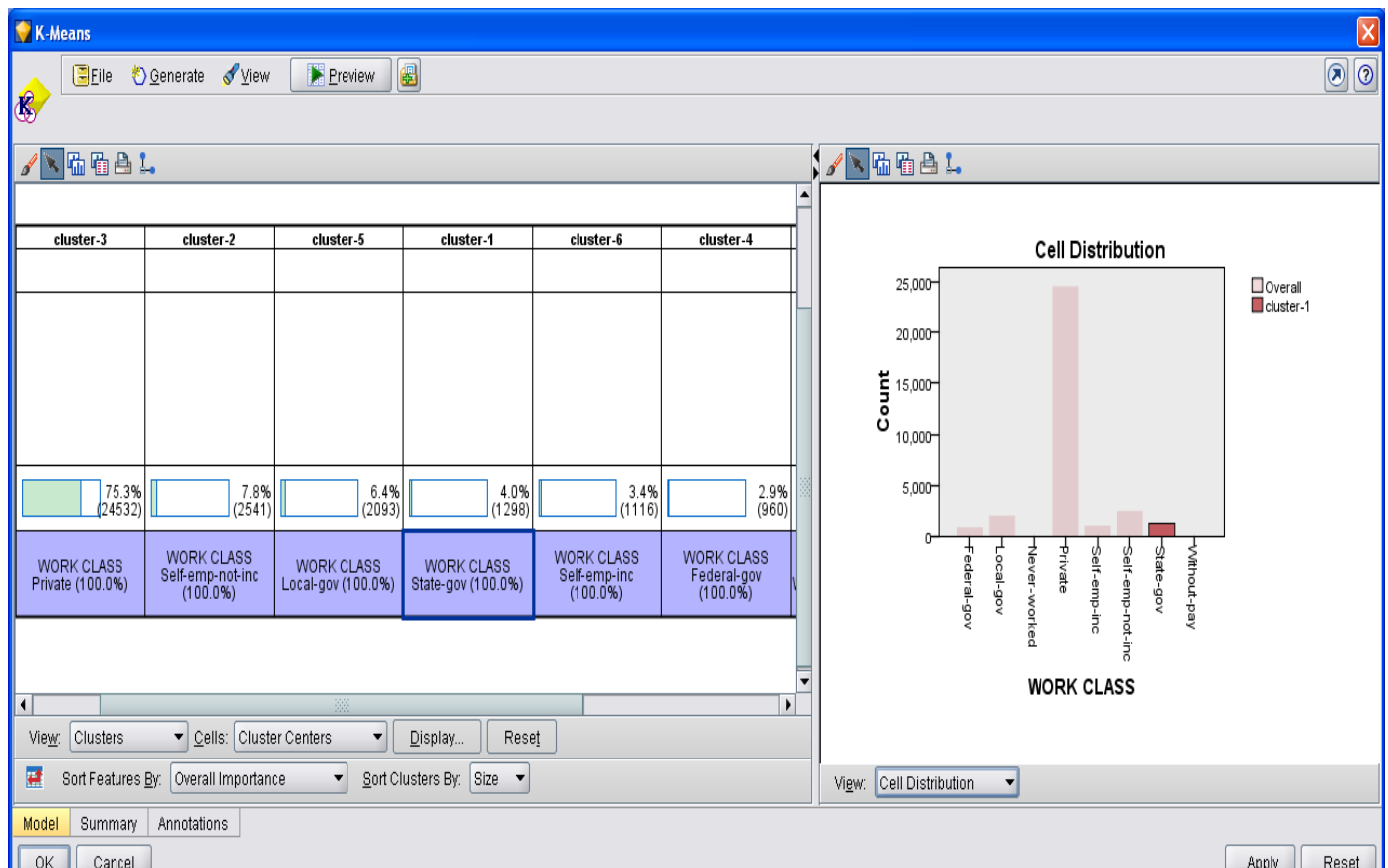
The k-means method has been shown to be effective in producing good clustering results for many practical applications. In statistics and machine learning, **k-means clustering** is a method of cluster analysis which aims to partition n observations into k clusters in which each

observation belongs to the cluster with the nearest mean [7]. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data as well as in the iterative refinement approach employed by both algorithms.

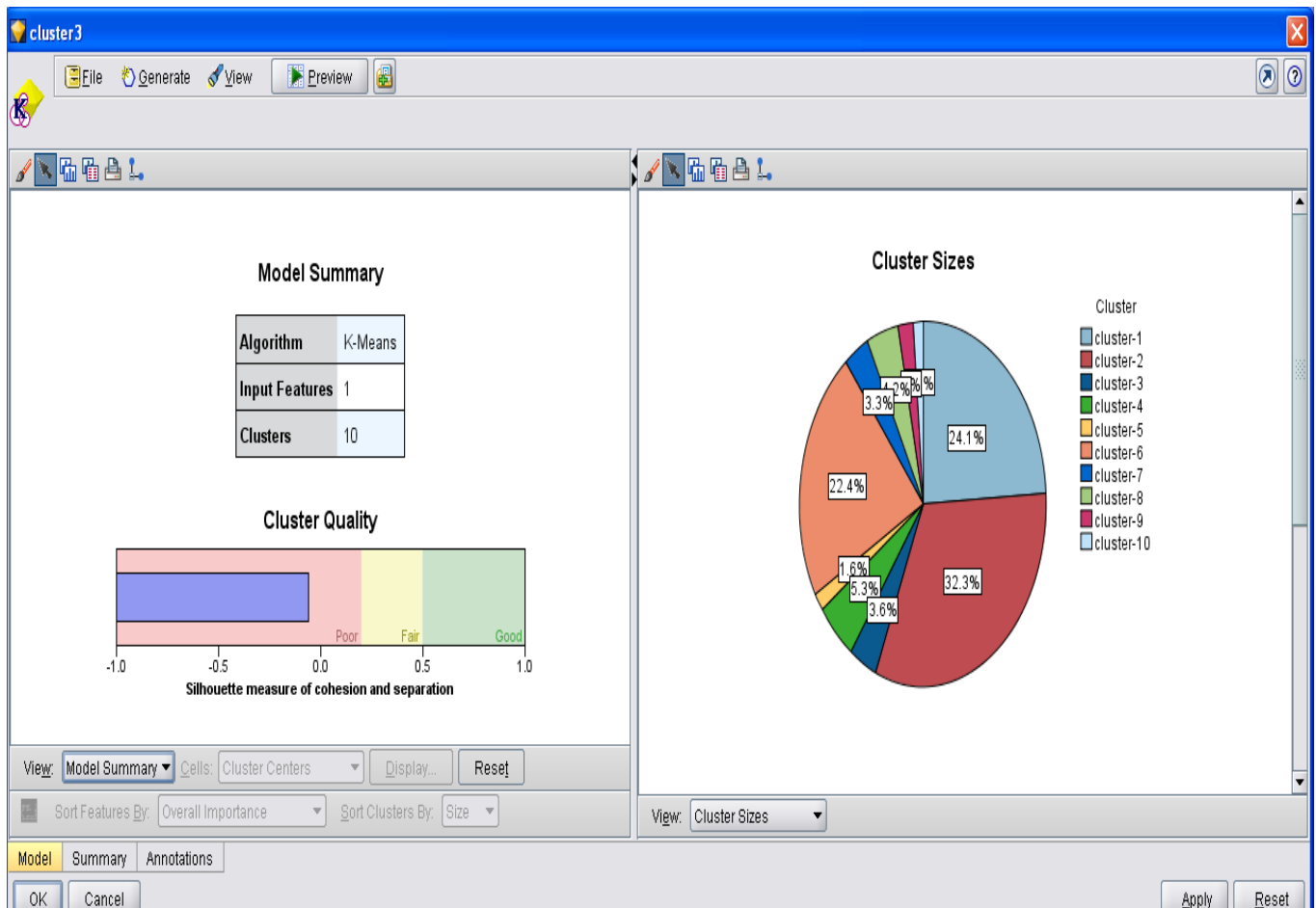
Overall result of k-means algorithm on the basis of workplace



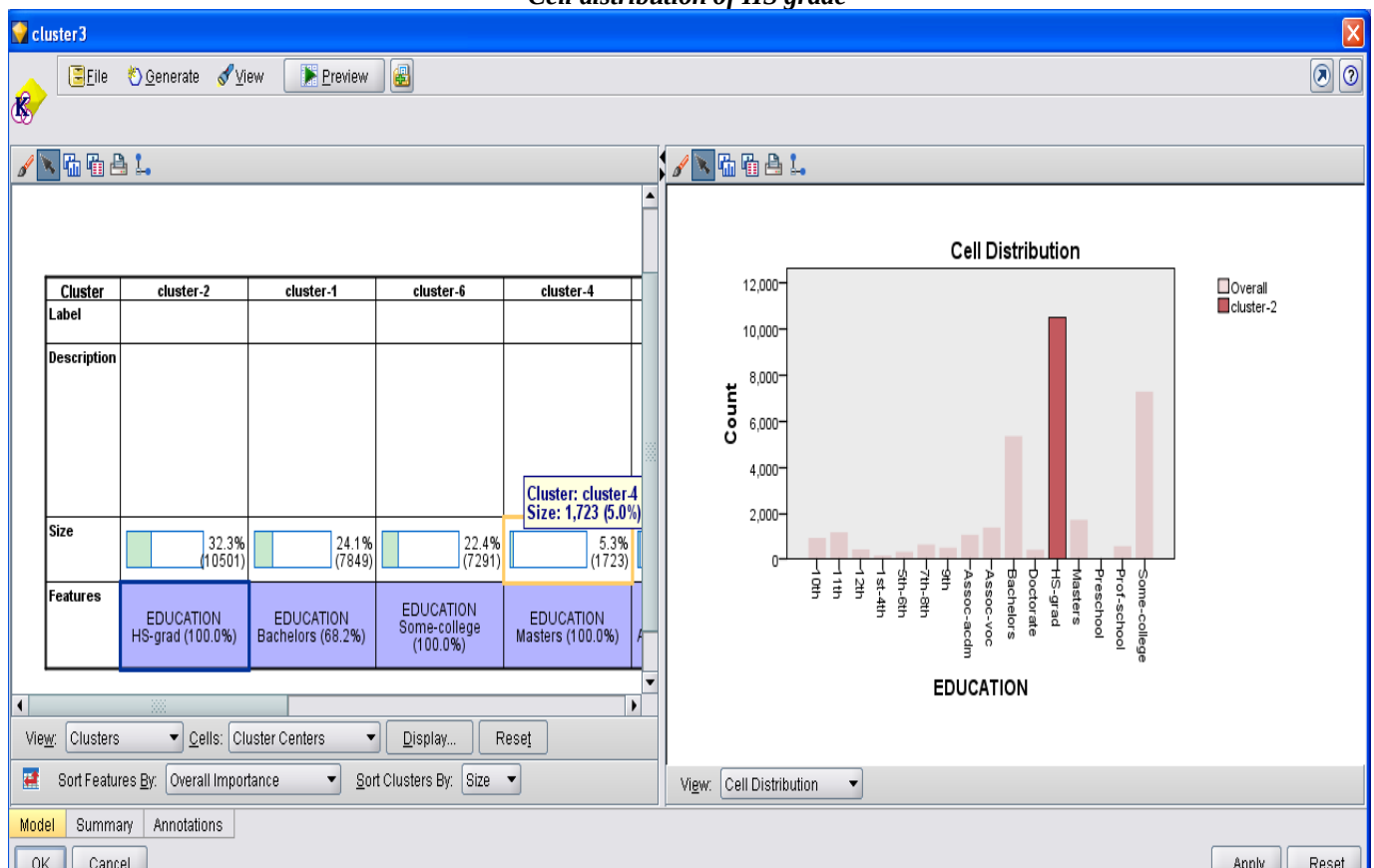
Cell distribution of work class



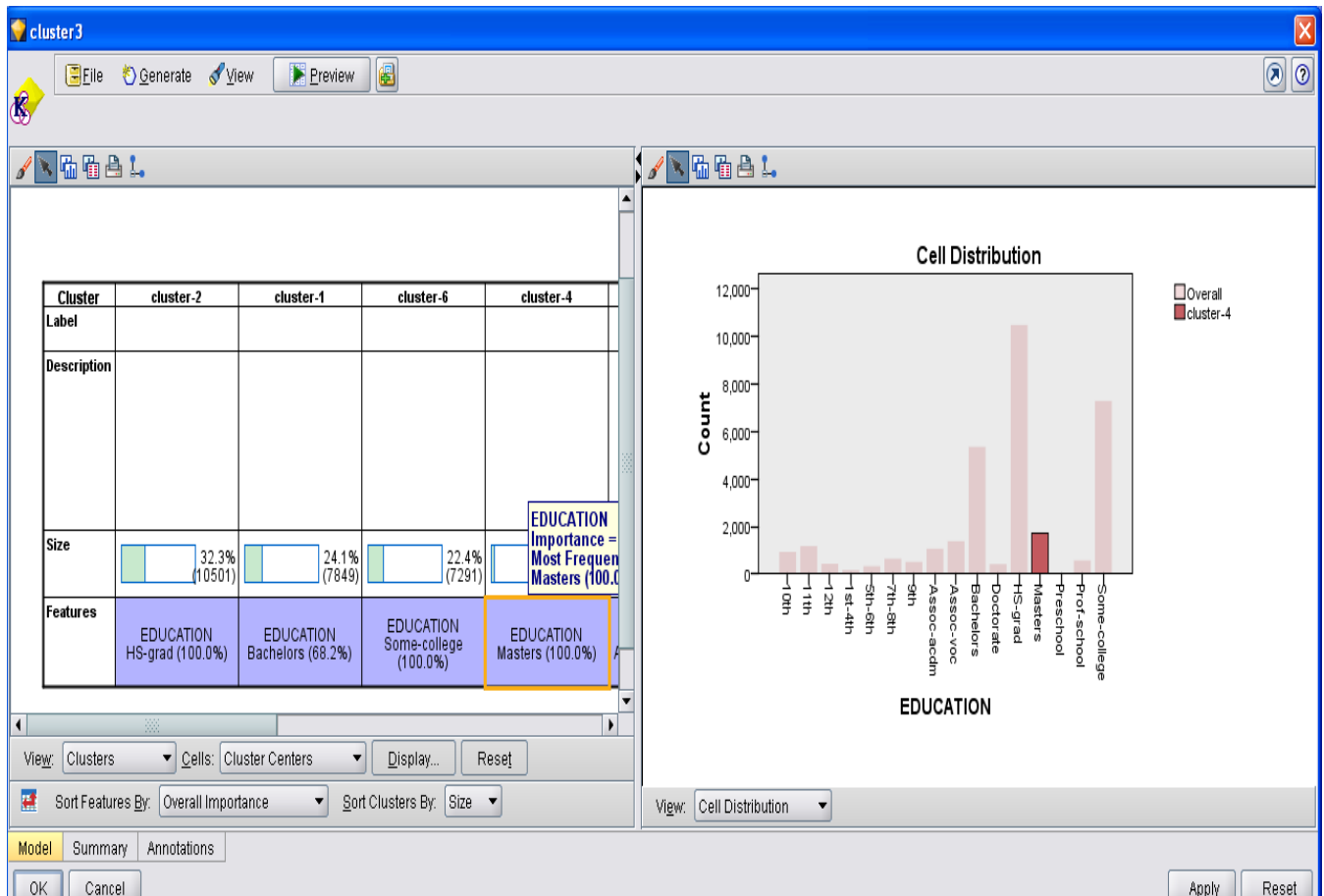
Overall result of k-means algorithm on the basis of education clustering



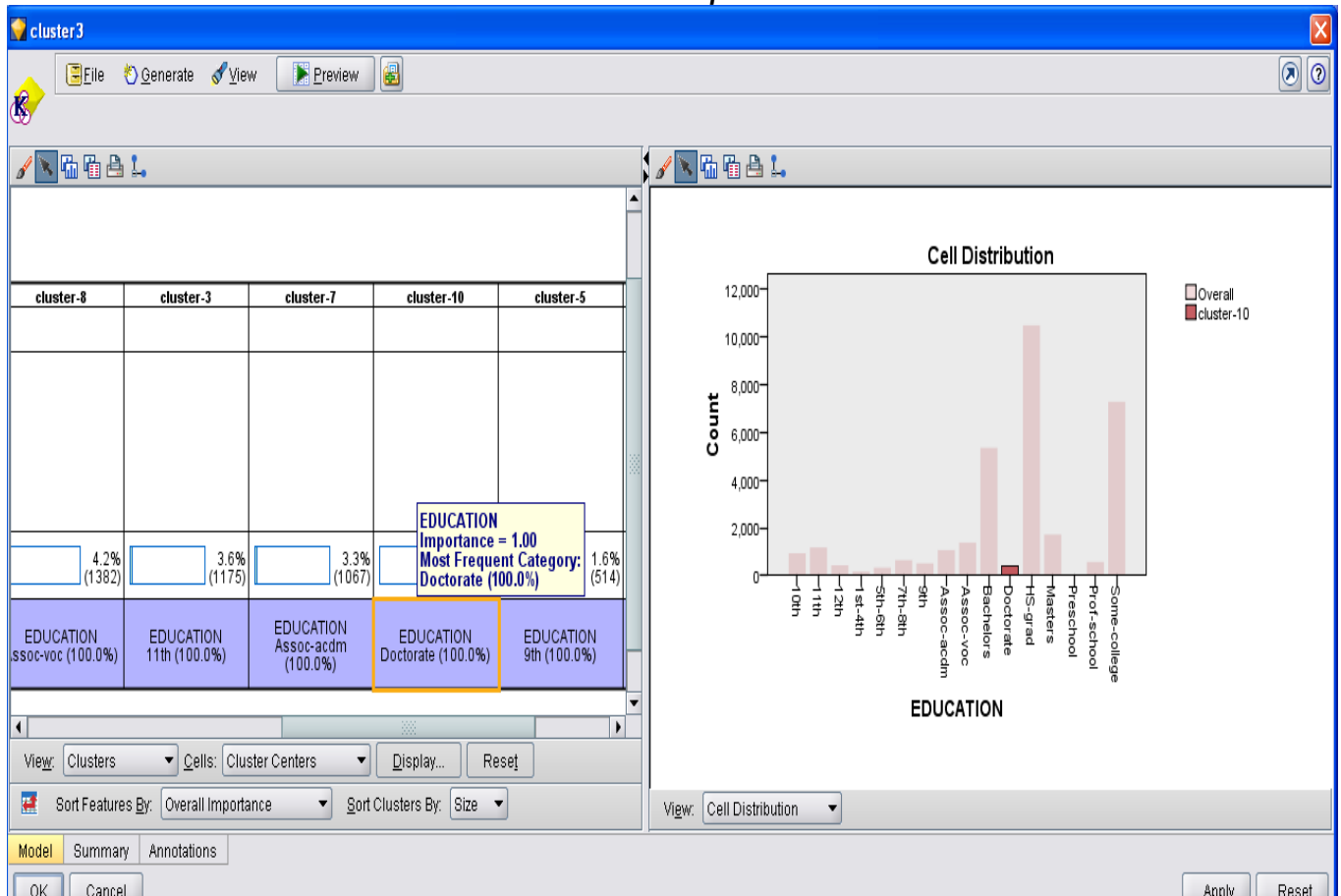
Cell distribution of HS grade

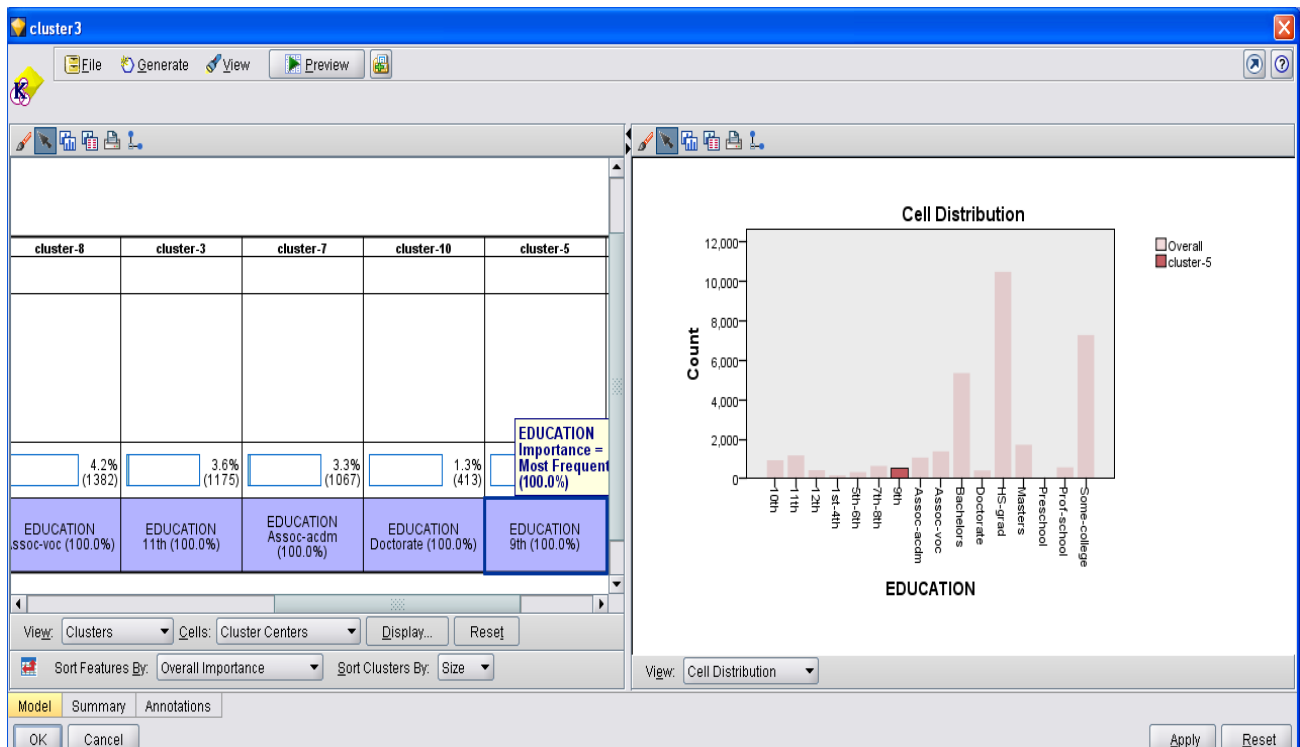


Cell distribution of masters



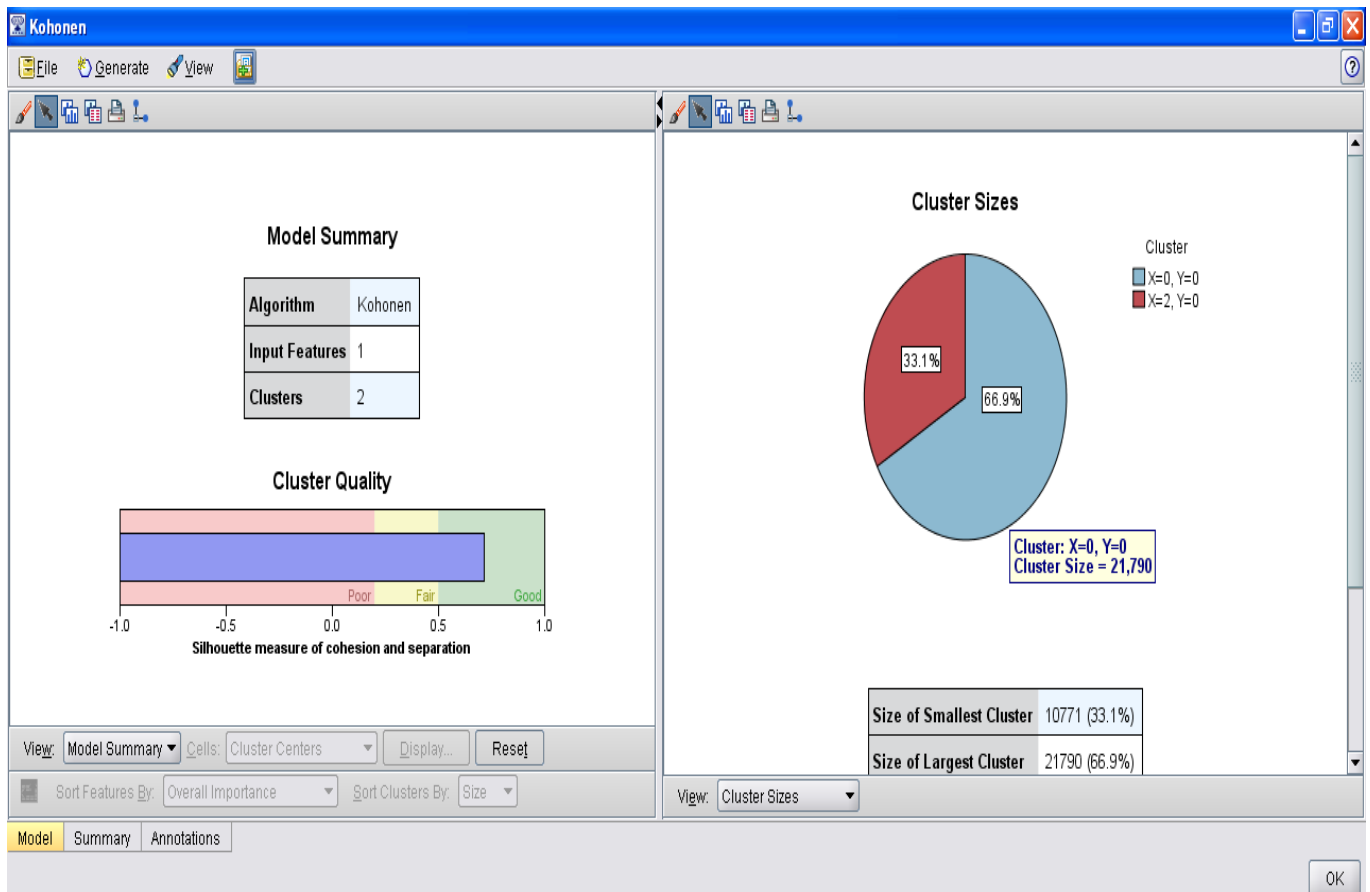
Cell distribution of doctorate

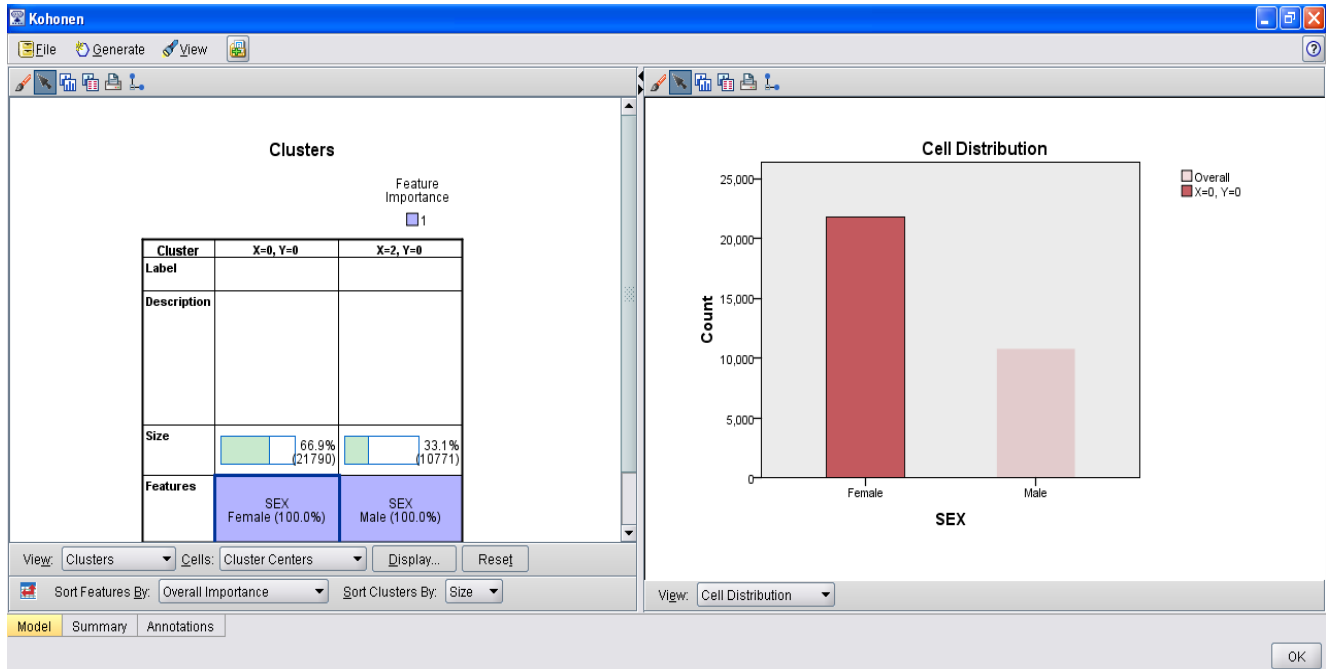
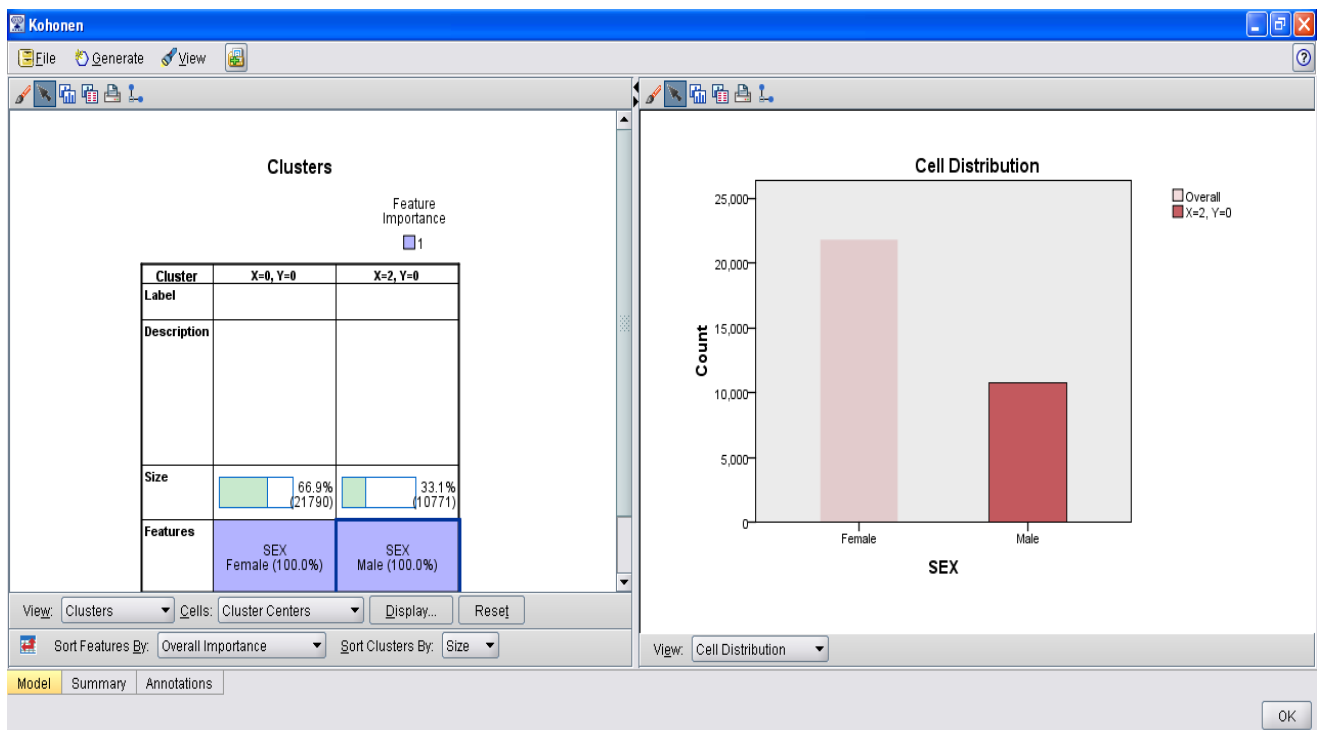


Cell distribution of 9th classii. **Kohonen:**

Kohonen networks [8] are a type of neural network that perform clustering, also known as a knet or a self-organizing map. This type of network can be used to cluster the dataset into distinct groups when you don't know what those groups are at the beginning. Records are grouped so that records

within a group or cluster tend to be similar to each other, and records in different groups are dissimilar. It is a type of unsupervised learning. Overall result after applying Kohonen-Here we get a sex ratio in which how many male and females are there in the particular region.

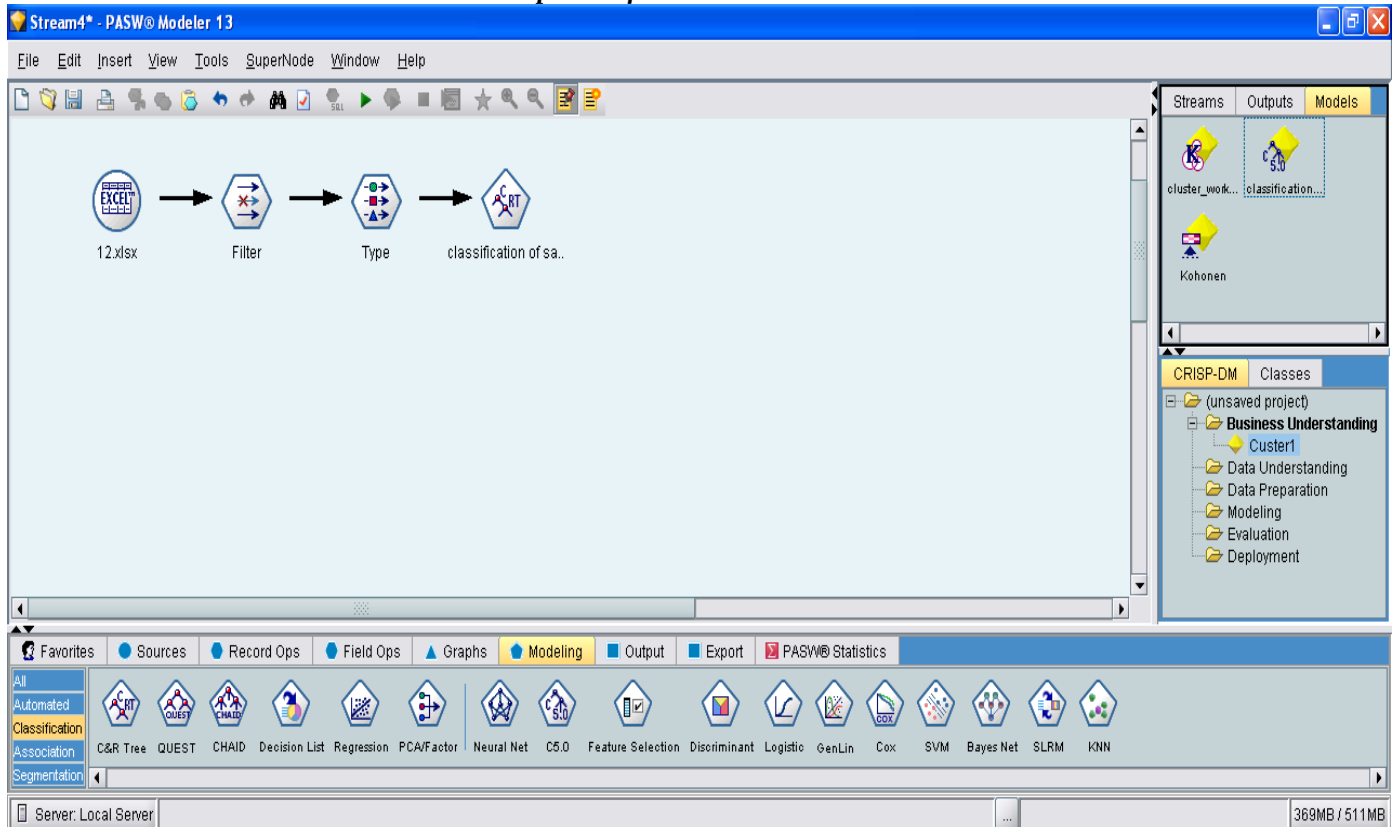


Female cell distribution**Male cell distribution****(b). Classification:**

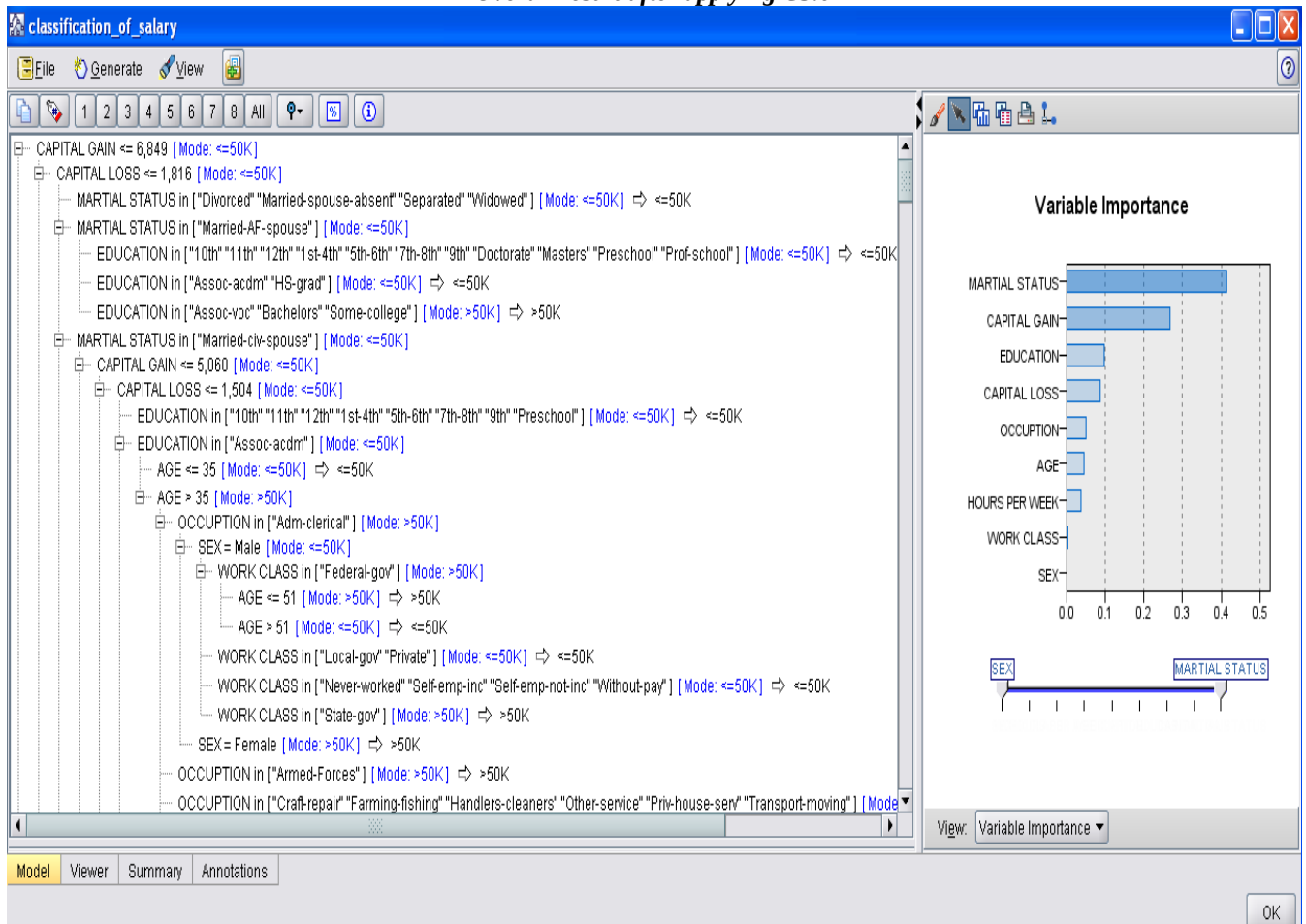
In Data Mining one of the most common tasks is to build models for the prediction of the class of an object on the basis of its attributes [9]. To predicate whether an income exceeds greater than 50K or less than 50K. Apply C5.0 for classification. This node uses the C5.0 algorithm to build either a decision tree or a rule set. A C5.0 model works by splitting the sample based on the field that

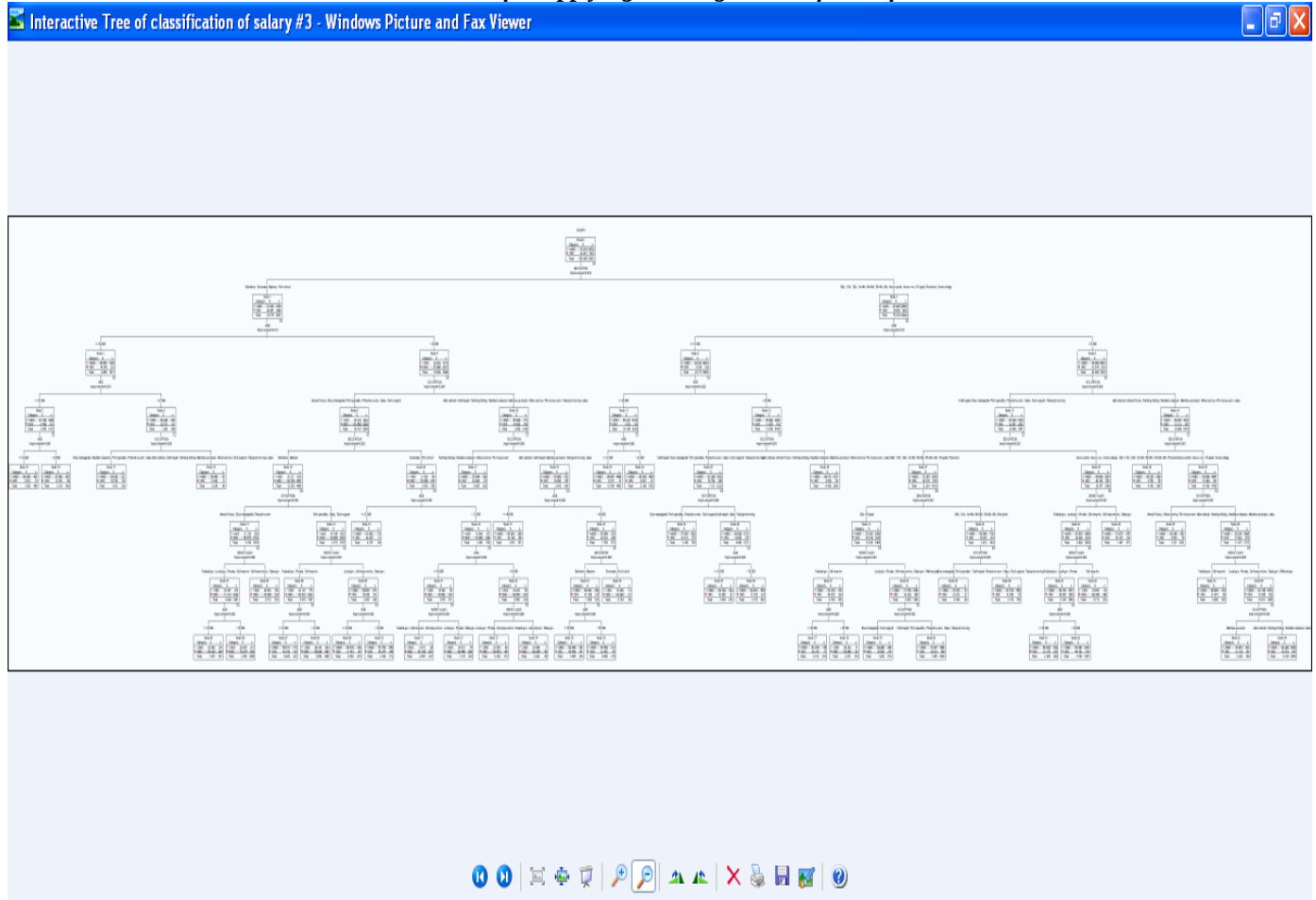
provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.

Snapshots of C5.0 in PASW Modeler



Overall result after applying C5.0



Decision tree after applying C5.0 algorithm of classification**III. INTERPRETATION AND CONCLUSION**

Now-a-days most enterprises are actively collecting and storing data in large databases. They have recognized the potential value of these data as an information source for making business decisions. The increasing demand for better decision support is answered by an extending availability of knowledge discovery, and data mining is one step at the core of the knowledge discovery process. We have also applied the process of data mining for the analysis and prediction of income of a particular area. After applying the tools we have analyzed following things:

- In clustering on workplace (Private, government, self employee etc). Cluster formed is five. Clustering quality fair.
- In clustering on sex (Male, Female). Number of Cluster formed is two. Clustering quality good.
- In clustering on education (9th, 10th, 12th, bachelors, masters etc). Number of Cluster formed is ten. Clustering quality good.
- In Classification Technique, predication on salary whether salary greater than 50K or less than 50K.
- Census information exploration system is useful for future analysis or predication of salary.

IV. REFERENCES

- Joyce Jackson, "Data Mining: A Conceptual Overview", Management Science Department, University of South Carolina.
- <http://census.ire.org/data/bulkdata.html>
- <http://www.spss.com.hk/software/modeling/modeler/>
- <http://www.obgyn.cam.ac.uk/cam-only/statsbook/stdatmin.html>
- IBM® SPSS® Modeler Professional, "Make Better Decisions Through Predictive Intelligence", IBM Software, Business Analytics.
- Jerzy Stefanowski, "Data Mining - Clustering", Institute of Computing Sciences Poznan University of Technology, Poznan, Poland, Lecture 7, SE Master Course, 2008/2009.
- Ed Berthold, "Intelligent Data Analysis", ISBN 9783540430605.
- S. Oja, E. Kaski, Kohonen Maps. Elsevier Science, 2003.
- J. Ross Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- Chid Apte, "Data Mining: Concepts and Techniques", Second Edition, University of Illinois at Urbana-Champaign.