

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Census Information Exploration

Sidra Anam* M.Tech Scholar, CSE Department, Pranveer Singh Institute of Technology, Kanpur, India Saurabh Gupta 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:

- a. Age: continuous.
- b. Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- c. 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.

- d. Education-num: continuous.
- e. Marital-status: Married-civ-spouse, Divorced, Nevermarried, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- f. Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlerscleaners, Machine-op-inspct, Adm-clerical, Farmingfishing, Transport-moving, Priv-house-serv, Protectiveserv, Armed-Forces.
- g. Sex: Female, Male.
- h. Relationship: Wife, Own-child,
- Husband, Not-in-family, Other-relative, Unmarried.
- i. Capital-gain: continuous.
- j. Capital-loss: continuous.
- k. Hours-per-week: continuous.
- l. 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.



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:



a.

Extraction:

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

									R	lead t	he dat	a					
킩 fi	nal1.ats -	Advanced ET	L Processor														
i <u>E</u> il	e <u>H</u> elp																
-	Data Flow	Diagram	🔁 Template	Deject	ad Fila 🔊	Execution I											Objects
1000						Execution	og										Objects
	1 10 🖻	- 🖬 🖼	🗳 🛄 🕻	-#P 🗙 🗋	🏯 😂 🛛	1 🖓 🚽											Clone Row
	Reader:		•	🖄 Validator:		🖌 🔶	B Transform	ner:	H	• 📄 Writer:		٥					Transformer
				0	0	\times		0									Validator
					0												Writer
																	Sorter
																	Deduplicator
																	Grouper
																	Pivot
																	UnPivot
R	Reader: {	} 🖻 Wri	ter: { }														Fields Selector
	- 1 (B)	2 LA															
1 100		91-							_								
	100	C: \Users \India	\Desktop\final	\dm\try\final.tx	t			5	4								
	AGE 1	NORKCLASS	[FNLWQT]	EDUCATION	UCATION N	UTAL STA	CCUPATION	ELATIONSH	RACE	[AGE]	VPITAL GA	APITAL LOS	JSE PER W	[NATIVE COMMUNITY]	[SALARY]		<u>^</u>
	6	9	9	10	9	8	9	9	8	9	8	9	8	18	9		
1	aqe	workclass	education	education	marital s	occupati	relations	race	age	capital o	a capital	house per	native c	salary			
2	39	State-go	Bachelor	13	Never-ma	Adm-cle	Not-in-f	White	Male	2174	0	40	Magarpa	<=50K	4-5 OV		
4	38	Private	HS-grad	9	Divorced	Handler	Not-in-f	White	Male	0	0	40	Magarpa	<=50K	~-30K		
5	53	Private	234721	11th	7	Married	Handlers	Husband	Black	Male	0	0	40	Magarpatta	<=50K		
6	28	Private	338409	Bachelors	13	Married	Prof-spe	Wife	Black	Female	0	0	40	Swarqate	<=50K		
8	37	Private	Masters 160187	14 9th	Married-	Exec-ma Married	Other-se	Not-in-f	Black	Female	0	40	Magarpa 16	Jamaica	<=50K		
9	52	Self-emp	209642	HS-grad	9	Married	Exec-man	Husband	White	Male	0	0	45	Magarpatta	>50K		
10	31	Private	Masters	14	Never-ma	Prof-sp	Not-in-f	White	Female	14084	0	50	Magarpa	>50K			
11	42	Private	Bachelor	13	Married-	Exec-ma	Husband	White	Male	5178	0	40	Magarpa	>50K			
13	30	State-go	141297	Bachelors	13	Married	Prof-spe	Husband	Asian-P	Male	0	0	40	Shivaji nagar	>50K		
14	23	Private	122272	Bachelors	13	Never-m	Adm-cler	Own-chil	White	Female	0	0	30	Magarpatta	<=50K		
15	32	Private	205019	Assoc-acd	12 Manual and	Never-m	Sales	Not-in-f	Black	Male	0	0	50	Magarpatta	<=50K		
17	34	Private	7th-8th	4	Married-	Transpo	Husband	Amer-Shi	Male	0	0	45	Yadwada	<=50K			
18	25	Self-emp	176756	HS-grad	9	Never-m	Farming-	Own-chil	White	Male	0	0	35	Magarpatta	<=50K		
19	32	Private	186824	HS-grad	9	Never-m	Machine-	Unmarrie	White	Male	0	0	40	Magarpatta	<=50K		
) (8											\$	0 🛈 🐓 🕅 🐺	9 🏴 📎	🔍 💐 🛄 🖄 😒 🦉 🐗 🗘) 16:30 14-11-2011
									Set 1	Reade	r Prop	oerty					
🌔 f	inal1.ats -	Advanced E	TL Processor														
Ele Help																	
	Data Flow	Diagram:	Template	e 🙋 Reject	ed File 🔉	Execution L	.og										Objects
	P * d	-		ր×∣վե	±181	a 📥 🗎	(R.)	Deedee D						×			Note
				State 1			Data	Reader Prop	perties		-						Clone Row
	Reader:			Validator:	0	×	Ge Ge	neral Text	Data Read	ling Restriction	ns Rejected	File					Transformer



b. Validation:

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

- a) Convert age field which contains? Value, set it to correct value.
- b) Convert workclass field which contains? Value, set it to correct value.
- c) Set occupation field to particular value.

🌔 Va	lidation l	Editor																X
) 🚺 🗙	₩ \$	🔒 🖬 ⁄a	1 🗹 🏟	🚳 🛛 🕅											Transformations	
• •	F1]		o ha	🔹 🛋 Contains		- he-	🔶 (F1)										Б 🔏 1 📋 🥘	
• I	F2]		,J	Is Number		_ •∦	♦ [F2]										Incremental Search	
•	'F3]						🔶 [F3]										Note	^
•	F4]		• >				♦ [F4]										Previous Value	
•	F51		• >			,	(F5)										Abs	
	[E6]						(F6)										Current Date	
•	F71		•	• 🛋 Contains		•	 IF71 										Date Format	
	-8]		•	Sy contains		-	(F8)										Reformat Date	
ا (F91		•				♦ [F9]										Delete Spaces	
•	F10]		• >				♦ [F10]										Delete Characters	
•	F11]		• >				♦ [F11]										Ensure No Prefix	
ب	F12]		• >				🔶 [F12]										Ensure No Suffix	
•	[F13]		• >				🔶 [F13]										Ensure Prefix	
•	F14]		2 .	🥩 Contains		• h	🔶 [F14]										Ensure Suffix	
ا ﴿	F15]		• >				🔶 [F15]		•								Escape String	•
1		1	a abcali	aution from													,	
/ 501	urce Data	[F2]	s step's exe [F3]	F41	[F5]	[F6]	[F7]	[F8]	[F9]	[F10]	[F11]	[F12]	[F13]	[F14]	[F15]			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
1	6 39	9 State-go	9 77516	10 Bachelors	9 13	8 Never-m	9 Adm-cler	9 Not-in-f	8 White	9 Male	8 2174	9	40	18 Magarpatta	9 <=50K	J		
2	50	Self-emp	83311	Bachelors	13	Married	Exec-man	Husband	White	Male	0	0	13	Magarpatta	<=50K			
4	38 53	Private	234721	HS-grad 11th	7	Married	Handlers Handlers	Not-in-f Husband	Black	Male	0	0	40	Magarpatta Magarpatta	<=50K			
5	28	Private	338409	Bachelors	13	Married	Prof-spe	Wife	Black	Female	0	0	40	Swarqate	<=50K			
6	37	Private	284582	Masters 9th	14	Married	Exec-man Other-se	Wife Not-in-f	White	Female	0	0	40	Magarpatta Jamaica	<=50K			
8	52	Self-emp	209642	HS-grad	9	Married	Exec-man	Husband	White	Male	0	0	45	Magarpatta	>50K			
9	31	Private	45781	Masters	14	Never-m	Prof-spe	Not-in-f	White	Female	14084	0	50	Magarpatta	>50K			
10	37	Drivate	280464	Some-coll	10	Married	Exec-man Evec-man	Husband	Rlack	Male	0	0	80	Magarpatta	>50K	-		-
																	OK Can	cel
			a p										-		D D D			
	7 8	6		2 🛤									2	2 🗿 🛯 🗞 🕺 👘	PP 🖏 🕛	🔊 🛄 🖄 💙 🖺	😅 📲 🗣 14-11-20	011
() 11	and an	F 10			_													Y
V	alidation	Editor	1.00.0			[23.] M	4 N N										La Caracteria	^
	Г 6 <u>5</u>		9P					•										
*	[AGE 1]	4001		 Contains Is Number 		- J	 AGE 1 Duopuci 	100]	•								Incremental Search	
-	TENII MOT	1						ASSI			_						Note	^
-	FOLICAT			🗕 🔏 Proper Ca	se	• 'z				*	A Delete Sp	aces		}			Previous Value	
•	[EDUCAT	ION NO.1	•)-					ON NO.1									Abs	
	MARITA	AL STATUS]	• •									\sim					Current Date	
•	[OCCUPA	TION]	• >-	🛥 🥩 Contains		•			۰	- 🔏 Prope	er Case	G	• ►	🖙 🔶 [MARITAL STATUS	5] 🛛		Date Format	
٠	[RELATIO	NSHIP]	• >-				🔶 [RELATIO	NSHIP]	٠								Reformat Date	
	[RACE]		• >				🔶 [RACE]		٥								Delete Spaces Delete Characters	
-	[AGE]		٥				🔶 [AGE]		٥								Delete	
	[CAPITAL	GAIN]	•			•	ICAPITAL	GAIN]	۵								Ensure No Prefix	
	[CAPITAL	LOSS]	• •			•		LOSS]									Ensure No Suffix Ensure Prefix	
*	[HOUSE F	PER WEEK]				•	(HOUSE PE)	R WEEK]	•								Ensure Suffix	
*		COMMUNITY		 A Proper Cas 	je			COMMUNITY									Escape String	
	[SALART]						V [SALAKT]										First I In	
																	OK Canc	;el
6										-	_						13:08	
	1					4								🧿 🗄 😪 💐 😽	😑 🕛 🐉	🔍 💐 🛄 🔊 오	14-11-20	11

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.

🖬 🕹 🗅 🗂 🗙	+ ♣ 🔒 🗋 🖆 🤘	e 🚳 N 4 P N 🔶 🗖 🖡		Transformat
[AGE]	•)			Incremental Search
	n 👉 🔏 Delete Spaces			Note
				Output
	🖷 🖛 📶 Proper Case			Previous Value
		EDUCATION		Abs
IEDUCATION NO.]		IEDUCATION NO.]		Current Date
[MARITAL STATUS]	Proper Case	Imarital Status]		Date Format
[OCCUPATION]	😐 🏎 🔏 Delete Spaces			Reformat Date
[RELATIONSHIP]	Proper Case	🛛 🔎 🔶 [RELATIONSHIP]	0	Delete Spaces
[RACE]	•	♦ [RACE]		Delete Character
[SEX]	•••	► 🔶 [SEX]		Delete
🔶 [CAPITAL GAIN]	•	[CAPITAL GAIN]		Ensure No Prefix
[CAPITAL LOSS]	• •	(CAPITAL LOSS)		Ensure No Suffix
[HOUSE PER WEEK]	•)	- 🔶 [HOUSE PER WEEK]		Ensure Prefix
[NATIVE COMMUNITY]		[NATIVE COMMUNITY]		Ensure Suria
[SALARY]		SALARY]		First Up
				OK

d. Loading:

	Set writ	ter property				
👏 final1.ats - Advanced ETL Processor						3
<u>F</u> ile <u>H</u> elp						
Data Flow Diagram: O Template Rejected	File 🙆 Execution Log				Objects	
					Note	
					Clone Row	
📄 Reader: 🔍 🛏 🐼 Validator:	🗸 🛏 💮 Transformer: 🗸 🛏 🕞 Wr	iter:			Transformer	
99	0 X 99				Validator	
					Writer	
					Sorter	
		~			Deduplicator	
	Writer Properties				Grouper	
	Connection Type Text File				Birst	
	Description				Pivot	
					UnPivot	
Reader: { }					Fields Selector	
😭 🕅 🧭 🕂 🗙 🔲 🔜	Data Larget Type:					
	Text File MS SQL Server					
100 C: \Users \India \Desktop \final \dm \try \final.txt	Insert Script					
AGE 1 NORKCLASS [FNLWOT] [EDUCATION]	CALC ON A CONTRACT OF CONTRACT.	AL GA APITAL LOS JSE P	ER W [NATIVE COMMUNITY]	[SALARY]		-
			3 14	15		
1 age workclass education education m	Excel File My5qi	ital house per nati	ve c salary			
2 39 State-go Bachelor 13	Access Database O PostgreSQL	40 Mag	arpa <=50K			
3 50 Self-emp 83311 Bachelors	13 OBF File OInterbase/Firebird	0 13	Magarpatta	<=50K		
4 38 Private HS-grad 9	ODBC File System	40 Mag	arpa <=50K			
5 53 Private 234721 11th		0 40	Magarpatta	<=50K		
7 37 Drivate Masters 14	Va Ole DB SMIP	40 Mag	arna <=50K	~-30K		
8 49 Private 160187 9th	SQLite	0 16	Jamaica	<=50K		
9 52 Self-emp 209642 HS-grad		0 45	Magarpatta	>50K		
10 31 Private Masters 14	Ne	50 Mag	arpa >50K			
11 42 Private Bachelor 13	14	40 Mag	arpa >50K			
12 37 Private Some-col 10	14	80 Mag	arpa >50K			
13 30 State-go 141297 Bachelors	13	0 40	Shivaji nagar	>50K		
14 23 Private 1222/2 Bachelors	13	0 30	Magarpatta	<=50K		
16 40 Drivate Assoc-up 11	4a	40 vnit	nagarpatta	~-50K		
17 34 Private 7th-8th 4	OK Car	cel 45 Yad	wada <=50K			
18 25 Self-emp 176756 HS-grad		0 35	Magarpatta	<=50K		
19 32 Private 186824 HS-grad	Never-m Machine- Unmarrie White Male	0 0 40	Magarpatta	<=50K		-
🚱 🧭 🚞 🖸 🔠			ଚ 🧿 🛈 🐓 💱 🛤	😑 🖿 繁 (🔪 💐 🛄 😒 😕 🌌 📣 16:34 14-11-2011	ľ

Load file into its destination

🐌 finalLats - Advanced ETL Processor																
Eile Hel	p															
🚠 Data P	low Diagram:	🙆 Template	🙆 Rejecte	ed File 🛛 🔊	Execution L	og										Objects
1 😭 🏠	-	l 🐇 🗅 🕻) × (4+)	* 🔒 🖸												Note
Pead			- Validatori			Transform	ar:		Writer							Clone Row
216	n n		22000	0	×	g nansioni	22000		2219	9	0					Transformer
	~ ~		22000	0	<u> </u>		22000		0		0					Validator
		1		(-				_			V	D		Writer
					Processing	Data							~			Sorter
																Deduplicator
					Source: Started:	C:\Users\L 14-11-201	ndia (Desktop (1 16:34:36	final (dm (try (hnal.txt							Grouper
					Time:	00:00:05										Pivot
				_	Time Lef	t: 00:00:02										UnPivot
Reade	er: { } 🛛 📄 Wr	iter: { }			Records	: 21600										Fields Selector
	हे। 🥜 👍	X			L											
100		1Deelsteel Geel	i 🗾		Line No	Type D	escription						^			
100	C: Users undia	Desktop (nnai	lam lary (nnai, tx)	t	1	↓ Inf St	tarting Transfo	rmation					=		1	
AG	2 1 NORKCLAS	IFNLWQ11	EDUCATION 4		3	⊋int M Inf Δι	ap Name: uthor:							ISALARY	-	<u> </u>
6	9	9	10	9	4	Inf Ve	ersion:							9		
1 age	workclass	education	education	marital s	5	🔍 Inf D	escription:									
2 39	State-go	Bachelor	13 Rachelers	Never-ma	6	Inf Pr	eparing Writer	rs						Z=EOW		
4 38	Private	HS-grad	9	Divorced	7	Ų Inf W	riter: {}Writer	{ } Is Read loady	у					-SUK		
5 53	Private	234721	11th	7			I WITTELS DIE N	cauv	6	8%				<=50K		
6 28	Private	338409	Bachelors	13	Close	once finished		fret						<=50K	_	
8 49	Private	Masters 160187	14 9th	Married-		once misineu	j 🔄 Cical log	y misc	Car	ncei				<=50K	_	
9 52	Self-emp	209642	HS-grad	9	Married	Exec-man	Husband	White	Male	0	0	45	Magarpatta	>50K		
10 31	Private	Masters	14	Never-ma	Prof-sp	Not-in-f	White	Female	14084	0	50	Magarpa	>50K			
11 42	Private	Bachelor	13	Married-	Exec-ma	Husband	White	Male	5178	0	40	Magarpa	>50K			
13 30	State-go	141297	Bachelors	13	Married	Prof-spe	Husband	Asian-P	Male	0	0	40	Shivaji nagar	>50K		
14 23	Private	122272	Bachelors	13	Never-m	Adm-cler	Own-chil	White	Female	0	0	30	Magarpatta	<=50K		
15 32	Private	205019	Assoc-acd	12	Never-m	Sales	Not-in-f	Black	Male	0	0	50	Magarpatta	<=50K	_	
16 40	Private	Assoc-vo	11	Married-	Craft-r	Husband	Asian-Pa	Male	0	0	40	united s	>50K		-	
18 25	Self-emp	176756	HS-grad	Married-	Never-m	Farming-	Amer-Shi Own-chil	White	Male	0	45	1adwada 35	<=SUK Magarpatta	<=50K	-	
19 32	Private	186824	HS-grad	9	Never-m	Machine-	Unmarrie	White	Male	0	0	40	Magarpatta	<=50K	1	-
(0			6								Ş	9 🧿 🗎 🐓 💱 🔛	9 🖻 🗞	🔪 🖺 📃 😒 🖉 🚜	●) 16:34 ●) 14-11-2011

Execution Log

🌔 final1.at:	s - Advanced	ETL Processor					x
<u>F</u> ile <u>H</u> elp)						
📥 Data Flo	ow Diagram:	O Template	Rejected	d File	Execution Log		
		"C:\Program Files (v	86)\DB Soft	ware Lahr	ratory/Advanced ETI /Data Transfr 🦚		
	🖻 🎐 💡	C. Programmics (X	00) 00 3010				
Line No T	ype	Time	Record	Object	Description		1
172 🗵	Error	12-11-2011 19:01		Rea	Could not find any files to read		
173 🔍	Inform	12-11-2011 19:01		Rea	Path: C:\Program Files (x86)\DB Software Laboratory\		
174 🔍	Inform	12-11-2011 19:01		Rea	Mask:		
175 🔍	Inform	12-11-2011 19:01		Writ	Wrote : 0 Line(s)		
176 🔍	Inform	12-11-2011 19:01			Transformation Completed		
177 🔍	Inform	12-11-2011 21:33			Starting Transformation		
178 🔍	Inform	12-11-2011 21:33			Map Name:		:
179 🔍	Inform	12-11-2011 21:33			Author:		
180 🔍	Inform	12-11-2011 21:33			Version:		
181 🔍	Inform	12-11-2011 21:33			Description:		
182 🔍	Inform	12-11-2011 21:33			Preparing Writers		
183 🗵	Error	12-11-2011 21:33		Writ	Nothing is Mapped		
184 🔍	Inform	12-11-2011 21:51			Starting Transformation		
185 🔍	Inform	12-11-2011 21:51			Map Name:		
186 🔍	Inform	12-11-2011 21:51			Author:		
187 🔍	Inform	12-11-2011 21:51			Version:		
188 🔍	Inform	12-11-2011 21:51			Description:		
189 🤇	Inform	12-11-2011 21:51			Preparing Writers		
190 🙆	Error	12-11-2011 21:51		Writ	Nothing is Mapped		
191	Inform	12-11-2011 21:51			Starting Transformation		
192 🥥	Inform	12-11-2011 21:51			Map Name:		
193 🥥	Inform	12-11-2011 21:51			Author:		
194 🦲	Inform	12-11-2011 21:51			Version:		
195	Inform	12-11-2011 21:51			Description:		
196	Inform	12-11-2011 21:51			Preparing Writers		
197 🕼	Frror	12-11-2011 21:51		Writ	Target text file name is blank		
198	Inform	12-11-2011 21:54			Starting Transformation		
199	Inform	12-11-2011 21:54			Map Name:		
200	Inform	12-11-2011 21:54			Author:		
201	Inform	12-11-2011 21-54			Version:		
202	Inform	12-11-2011 21:54			Description:		
203	Inform.	12-11-2011 21:54			Preparing Writers		
204	Inform.	12-11-2011 21:54		Writ.	Writer { } Is Ready		
205	Inform	12-11-2011 21-54			All Writers are Ready		•
•					m		Þ.
	Ø		X			♀ Q î � � ऄ ₽ ► ♡ < \$ \$ 9657 14.11-20	

Cleaned data in text format

	nai - Note	pau		
File	Edit F	ormat View	Help	
295,347,507,436,554,554,547,545,443,577,507,507,507,507,507,507,507,507,507	Privat Privat Federa Privat Self Privat Pr	e: 271466 e: 22275; e: 222956 e: 22275; 1: -00v, 252 mp-inc, 1: e: 237993 e: 216666 e: 56352; e: 147372 e: 147372 e: 148146 e: 59496; e: 59496; e: 149810 e: 59496; e: 30496; e: 30496; e: 3439511 e: 343951 e: 345	<pre>, Assoc-voc. 11, Never-married, Prof-specialty, Wot-in-family, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Married-civ-spouse, Feer-managerial, Wife, Orther, Female, 0, 0, 0, 40, Magarpatta, <50K Never-married, other-service, own-child, White, Female, 0, 0, 30, united State, <-S0K Never-married, other-service, own-child, White, Male, 0, 0, 50, Magarpatta, >50K Some-college, 10, Married-civ-spouse, Fech-support, Husband, White, Male, 0, 0, 40, Magarpatta, >50K Some-college, 10, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, Magarpatta, >50K Some-college, 10, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0, 0, 40, Magarpatta, >50K Some-college, 10, Married-civ-spouse, Machine-op-inspct, Husband, White, Male, 0, 0, 40, Magarpatta, >50K Some-college, 10, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, >50K Network, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, <50K Network, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, <50K Network, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, <50K Network, Married-civ-spouse, Sates, Husband, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Divorced, Kadm-Clerical, Not-in-family, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Never-married, Sates, Own-child, White, Male, 0, 0, 40, Magarpatta, <50K Some-college, 10, Married-Civ-spouse, Prof-specialty, Other-relative, White, M</pre>	ř
6	9 (6	🗧 💽 🔄 🏹 📖 🔿 🖉 🖉 🖉 🖉 🖉	15:23 14-11-2011

Convert it into excel file

6		17 (1 -)	÷					final - I	Microsoft I	Excel	and the set							- 0	x
-		lome Insert	Page L	ayout Form	ulas Data	Review View												- 10	σx
	Paste	Cut Copy Format Painter	Calibri B I	• 11 • <u>U</u> • ⊞ • 3			📑 Wrap Text 🔤 Merge & Cer	General	• • • • • • • • •	• •.0 •.0 Fo	onditional Form rmatting ~ as Tal	nat Cell Die + Styles +	Insert	Delete Format	Σ AutoSum Fill ▼ Clear ▼	Sort & F	Find &		
	Clip	board 6	<u>ا</u> ر	Font	6	Alignme	nt	Nui Nui	mber	G	Styles			Cells	E	diting			
		A1	• (0	<i>f</i> _∞ age															*
	A	В	С	D	E	F		G		Н	1	J		К	L	М	N	0	
	L age	workclass	education	education no	. marital status	occupation		relationship	:	Sex	capital gain	capital lo	55	house per week	salary				
1	2	39 State-gov	Bachelor	s 13	Never-married	Adm-clerical		Not-in-family		Male	217	4	0	40	<=50K				
4	3	50 Self-emp	Bachelor	s 13	Married-civ-spo	Exec-manager	rial	Husband		Male		0	0	13	<=50K				
4	1	38 Private	HS-grad	9	Divorced	Handlers-clea	iners	Not-in-family		Male		0	0	40	<=50K				
1	5	53 Private	11th	7	Married-civ-spo	Handlers-clea	iners	Husband		Male		0	0	40	<=50K				
(5	28 Private	Bachelors	s 13	Married-civ-spo	Prof-specialty	/	Wife		Female		0	0	40	<=50K				
1	7	37 Private	Masters	14	Married-civ-spo	Exec-manager	rial	Wife		Female		0	0	40	<=50K				_
1	3	49 Private	9th	5	Married-spouse	Other-service		Not-in-family		Female		0	0	16	i <=50K				_
1	Ð	52 Self-emp	HS-grad	9	Married-civ-spo	Exec-manager	rial	Husband		Male		0	0	45	>50K				
1	0	31 Private	Masters	14	Never-married	Prof-specialty	/	Not-in-family		Female	1408	4	0	50	>50K				
1	1	42 Private	Bachelor	s 13	Married-civ-spo	Exec-manager	rial	Husband		Male	517	8	0	40	>50K				_
1	2	37 Private	Some-co	I 10	Married-civ-spo	Exec-manager	rial	Husband		Male		0	0	80	>50K				
1	3	30 State-gov	Bachelors	s 13	Married-civ-spo	Prof-specialty	/	Husband		Male		0	0	40	>50K				
1	4	23 Private	Bachelor	s 13	Never-married	Adm-clerical		Own-child		Female		0	0	30	<=50K				_
1	5	32 Private	Assoc-acc	: 12	Never-married	Sales		Not-in-family		Male		0	0	50	<=50K				
1	6	40 Private	Assoc-vo	<u>(</u> 11	Married-civ-spo	Craft-repair		Husband		Male		0	0	40	>50K				
1	.7	34 Private	7th-8th	4	Married-civ-spo	Transport-mo	ving	Husband		Male		0	0	45	<=50K				
1	8	25 Self-emp	HS-grad	9	Never-married	Farming-fishi	ng	Own-child		Male		0	0	35	<=50K				
1	9	32 Private	HS-grad	9	Never-married	Machine-op-i	nspct	Unmarried		Male		0	0	40	<=50K				
2	0	38 Private	11th	7	Married-civ-spo	Sales		Husband		Male		0	0	50	<=50K				
2	1	43 Self-emp	Masters	14	Divorced	Exec-manager	rial	Unmarried		Female		0	0	45	>50K				
2	2	40 Private	Doctorate	16	Married-civ-spo	Prof-specialty	/	Husband		Male		0	0	60	>50K				
2	3	54 Private	HS-grad	9	Separated	Other-service	•	Unmarried		Female		0	0	20	<=50K				
2	4	35 Federal-g	9th	5	Married-civ-spo	Farming-fishi	ng	Husband		Male		0	0	40	<=50K				
2	5	43 Private	11th	7	Married-civ-spo	Transport-mo	ving	Husband		Male		0	2042	40	<=50K				-
I	< → >I	Sheet1 Sh	neet2 🖉 Sh	eet3 🖉 💭 🦯							14			Ш		_]	
F	leady														Œ		0% 🕞		-+
(?		1											¥ 💀 😑 🕨 🕅	🐚 🕸 💻	S 📀	👩 🙀 (» 15:3	0

D. Step 4: Apply data mining technique to evaluate a result:

Data mining, the extraction of hidden predictive information from large databases, is powerful new technology with great potential to help companies focus on important information in their the most data warehouses. The process of data mining consists of three stages. The initial exploration, model building or pattern identification with validation or verification, and is

concluded with deployment (i.e., the application of the model to new data in order to generate predictions) [4]. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that 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 [/]. 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.

🙀 K-Means 20 📑 Eile 👋 Generate 💰 View 🖹 <u>P</u>review 🔒 ĸ, 1 📉 🖬 🖷 📇 🛴 i 🖌 📉 🖬 🖷 📇 🐛 Cluster Sizes Model Summary Cluster cluster-1 Algorithm K-Means cluster-2 cluster-3 Input Features 1 ciuster-4 75.39 cluster-5 Clusters 8 **Cluster Quality** Size of Smallest Cluster 7 (0%) Size of Largest Cluster 24532 (75.3%) Ratio of Sizes: Largest Cluster to Smallest Cluster -0.5 -1.0 0.0 0.5 1.0 3,504.57 Silhouette measure of cohesion and separation Display. View: Model Summary ▼ Cells: Cluster Centers Reset Sort Clusters By: Size View: Cluster Sizes -Summary Annotations Model 0K Cancel Apply <u>R</u>eset

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



💡 K-Means															X
) <u>G</u> enerate 🛛 🔏 <u>V</u> iew	Preview	8												0
╱ ╱╲┺┺┺┆	-					1	6 6 4	! _							
					•										
cluster-3	cluster-2	cluster-5	cluster-1	cluster-6	cluster-4					Cell [Distri	bution			
							25,000							Overall	
							20,000-							Cluster-1	
							t 15 000-								
							Cou								
							- 10,000-								
75.3% (24532)	7.8% (2541)	6.4% (2093)	4.0% (1298)	3.4% (1116)	2.9%		5,000-						_		
WORK CLASS Private (100.0%)	WORK CLASS Self-emp-not-inc (100.0%)	WORK CLASS Local-gov (100.0%)	WORK CLASS State-gov (100.0%)	WORK CLASS Self-emp-inc (100.0%)	WORK CLASS Federal-gov (100.0%)		0-¥	-Federal-go	Local-gov	Private	Self-emp-i	Self-emp-	-Without-pa	1	
								20	rkea		inc	not-inc	ау		
									V	VORK	CLA	SS			
Vie <u>w</u> : Clusters	▼ <u>C</u> ells: Clust	er Centers 👻	Display Rese	2 <u>1</u>											
🦉 Sort Features	<u>By:</u> Overall Importan	ce 🔹 Sort C	lusters By: Size 💌			Vi <u>e</u> w	: Cell Distrib	ution	•						
Model Summary	Annotations														
OK Cancel														Apply	<u>R</u> eset



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



Sidra Anamet al, International Journal of Advanced Research In Computer Science, 5 (3), March-April, 2014,119-133

Cell	distribution	of masters
$\mathcal{L}\mathcal{L}\mathcal{U}\mathcal{U}$	awantoutout	of musicis



Cell distribution of 9th class



ii. 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.



Sidra Anamet al, International Journal of Advanced Research In Computer Science, 5 (3), March-April, 2014,119-133



Female 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 lowestlevel splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned.





Overall result after applying C5.0



Decision tree after applying C5.0 algorithm of classification

Linteractive Tree of classification of salary #3 - Windows Picture and Fax Viewer



0 🛛 🖾 🐳 🧊 🖉 🖉 🖉 🖉 🖉 🖉

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:

- a. In clustering on workplace (Private, government, self employee etc). Cluster formed is five. Clustering quality fair.
- b. In clustering on sex (Male, Female). Number of Cluster formed is two. Clustering quality good.
- c. In clustering on education (9th, 10th, 12th, bachelors, masters etc). Number of Cluster formed is ten. Clustering quality good.
- d. In Classification Technique, predication on salary whether salary greater than 50K or less than 50K.
- e. 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.
- [2]. http://census.ire.org/data/bulkdata.html
- [3]. http://www.spss.com.hk/software/modeling/modeler/
- [4]. http://www.obgyn.cam.ac.uk/camonly/statsbook/stdatmin.html
- [5]. IBM® SPSS® Modeler Professional, "Make Better Decisions Through Predictive Intelligence", IBM Software, Business Analytics.
- [6]. Jerzy Stefanowski, "Data Mining Clustering", Institute of Computing Sciences Poznan University of Technology, Poznan, Poland, Lecture 7, SE Master Course, 2008/2009.
- [7]. Ed Berthold, "Intelligent Data Analysis", ISBN 9783540430605.
- [8]. S. Oja, E. Kaski, Kohonen Maps. Elsevier Science, 2003.
- [9]. J. Ross Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
- [10]. Chid Apte, "Data Mining: Concepts and Techniques", Second Editionx, University of Illinois at Urbana-Champaign.