# Exploring Classification & Clustering Techniques for Predictive Analytics

### Manisha, Pawan Lathwal

Department of Computer Science, Department of Computer Science, B.M.I.E.T, Sonipat (Haryana)

ABSRACT: Diseases have been increased tremendously. Eight person from ten have diabeties, so the data in the hospital is increasing day by day. A huge amount of data has to be handled and for that we are using classification and clustering techniques which classifies the data and forms the clusters as well as we developed a tool for Apriori which predicts the strong rules. The tool is developed using .NET framework . For Clustering and Classification WEKA tool is used for extracting the results.

Keywords: Classification, Clustering, Predictive Analytics.

#### I. INTRODUCTION

Data Mining is the extraction of the knowledge from large amount data. of Data Mining is a detailed process of analysing large amounts of data and picking out the relevant information. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. There are many types of Data Mining like Clustering, Classification, and Association Rules Mining. Classification in the data mining is to classify the data into similar and dissimilar classes. Clustering is the technique which is used for finding the clusters. Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, "if an item set is not frequent, any of its superset is never frequent".

#### II. CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Here we use the diabetic data of the hospital in CSV format. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

ZeroR predicts the mean (for numeric class) or mode (for nominal class). Here ZeroR predicts class value and to build a model WEKA took 0 seconds.

		NSERT PAGE LA	OUT FOR	MULAS D	ATA REVIE	V VEW								-		Marisha Has	ja <sup>1</sup>
۵۵ هر	ut opy •	Calibri	_		≣ ∛≁			General		ļ ļ	1		(	∑ AutoSu III *	-	ň	
Paste	ury ormat Painter	BI∐-	- <u>b</u> -	<u>A</u> -≡≡	日住住	🗄 Merge &	Center +	<b>ş</b> .%,	€0_00 Co 00_+0 Co	nditional Forma matting * Tabl	tas Cell	insert Dele	e Formal	t dear*	Sort & Filter* 1		
Clipb		G Fa		6	Align		6		101	natung * Tabi Styles	, silip.	Cel			Editing	elect -	
E2		X √ fr	[0.10]														
	B terpatient r		_	F weight	G				L					Q R		T.	1
		n race gend 7 Caucasian Fema		veight 2	admission dis 6	charge aom 25	ission timi 1	e in npayer 1?	coc medical_s Pediatrics	num_lab_nur 41	n_procnu 0	m_meonumo 1	er_cnun 0	nder_enumb 0	er_idiag_1 0 250.8		diag ?
		9 Caucasian Fema	e (10-20)	?	1	1	7	3?	?	59	0	18	0	0		6 250.01	
		5 AfricanAn Fema		?	1	1	7	2 ?	?	11	5	13	2	0	1 6		V27
		6 Caucasian Male		?	1	1	7	2?	?	44	1	16	0	0		8 250.43	
		7 Caucasian Male	(40-50)	?	1	1	7	1?	?	51	0	8	0	0	0 1		
		1 Caucasian Male 9 Caucasian Male	[50-60) [60-70)		2	1	2	3?	?	31	6	16 21	0	0	0 4		V45
		8 Caucasian Male			1	1	7	5 ?	?	73	0	12	0	0	0 4		
		3 Caucasian Fema			2	1	4	13 ?	?	68	2	28	0	0	0 35		
		9 Caucasian Fema			3	3	4	12 ?	InternalM		3	18	0	0	0 4		
		2 AfricanAn Fema			1	1	7	9?	?	47	2	17	0	0	0 250		
		1 African An Male		?	2	1	4	7?	?	62	0	11	0	0	0 1		
		5 Caucasian Ferna 7 Caucasian Male		?	1	3	7	7?	Family/Ge Family/Ge		0	15	0	1	0 4	8 250.43 8 411	
		2 Caucasian Male 1 AfricanAn Fema		?	3	0	2	10 7	Pamily/Ge ?	49	5	31	0	0	0 4		
17 735	78 8632881	9 AfricanAn Male	[60-70]	?	1	3	7	12 ?	?	75	5	13	0	0	0 9		
18 770	76 9251935	2 AfricanAn Male	[50-60]	?	1	1	7	4?	?	45	4	17	0	0	0 4	0 411	
		8 Caucasian Fema		?	1	1	7	3 ?	Cardiolog		0	11	0	0	0 60		
	82 1.07E+0	8 AfricanAn Male	[70-80)	?	1	1	7	5 ?	?	35	5	23	0	0	0 40	2 425	
21																	
22																	
23 ( ) READY		etic data - Sample	÷		0 e		] (;			: 4	ę	<b>6</b> 4		E C C C C C C C C C C C C C C C C C C C	• • •	() 팅 (	+ 1 0541
READY	arch the w		•		D e		]	Q .	0		ę	Q 4		[111]	5	이 된 <sub>1</sub>	+ 1 0541 108-20
23 READY READY Webs Exp Proprocess	arch the w				D ê		] (;	9.4	0		ę	ø 🛛			5	¢) € "	+ 1 0541 108-20
23 READY READY Webs Exp Programme these for	arch the we liver Slaufy thu	eb and Windows			D e		] (;				ę	ø 🛛		111	5	() E (	+ 1 0541 108-20
23 READY READY Webs Exp Proprocess	arch the we liver Slaufy thu	eb and Windows			0 6		] (],				<u>e</u>	ø 4			5	<01 € 02	+ 1 0541 108-20
23 55107 55107 Selor Selor Selor Choose Test options	arch the we liver Claufy Clau Zerofi	eb and Windows Kr   Associate   Solo			• •		] (;				Ŗ	đ		1111	5	(이 팅 )	+ 11 05:41 1-03-20
23 READY READY Webs Exp Proprocess Choose Text options () Use Yai	arch the w Granty Clau Zeroit Seroit	eb and Windows for Associate Solid	tathlutar 1 estiv arput	ada	D e		] [-					<i>a</i>			5	() 팀 <sub>()</sub>	+ 11 05:41 1-03-20
23 55407	arch the we loca Cooxify Close ZeroR Magnet Historic	eb and Windows for Associate Selection Gen	tathéotar 1 estiv arput - Classif	ada	fell mainin		] (				<b>e</b>	đ 🛛		III	5	¢) € œ	+ 11 05:41 1-03-20
Second Control Co	arch the we loca Zoroft Zoroft Hertset	eb and Windows ter Resource Selo Set 2 2	tathéotov V entire arpat – Classet f ieck predis	ander er molel ( ta class v	foll commissions for	(7 APE)	9 6				<b>P</b>	đ			5	4) B	+ 1 0541 108-20
KEADY KEADY KEADY KEADY KeAD KeA	arch the we lover Davify Clau Davify Clau Davify Clau Davie Davie Davie Histopet Histopet Histopet Histopet	eb and Windows for Resource Selon for Resource Selon for II For III For IIII For IIIII For IIII For IIII For IIII For IIII For IIIIII For IIIIII For IIIIII For IIIIIIIIIII For IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	tathéotov V entire arpat – Classet f ieck predis	ander er molel ( ta class v	fell mainin	(7 APE)	3 6				<b>e</b>	đ 🛛			5	4) <b>E</b> <sub>2</sub>	+ 1 0541 108-20
KEADY KEADY KEADY KEADY KeAD KeA	arch the we loca Zoroft Zoroft Hertset	eb and Windows ter Associate Solor Set	tathlatav 1 estiv arpat – Classif irok predis ne takon 1 = Evalual.	under er model ( ms class v o beild mo	fell mainin Ales: Ver fell: D secon	(7 APE)	] (;				<b>e</b>	đ 4			5	(小民 <sub>)</sub>	+ 10 0541 +03-20
KEADY KEADY KEADY KEADY KeAD KeA	love Love Love Clove Love Clove Love Clove Love Clove Here See	eb and Windows ter Associate Solor Set	tatidatar 1 enterarpat - Classif ensk predis me takun 1	under er model ( ms class v o beild mo	fell mainin Ales: Ver fell: D secon	(7 APE)	9 6,					4 <b>4</b>			5	4) 60 80 90	+ 11 05:41 1-03-20
SECOV	arch the we bree Courly Courly Courl Courly Courl Parait Here are with Mice policies arMad	eb and Windows ter Resource Sele Sec	ente arpai - Classif - Cla	er nodel ( ta class v o build mo to co looi mo	foll rowinin Alos: Ver del: O secon opili	ng Aret) Ida	3 6	112			<b>e</b>	4 <b>X</b>			5	1 (1) 10 <sub>0</sub>	+ 11 05:41 1-03-20
EEDOF EEDOF EEDOF EEDOF EEDOF EEDOF EEDOF EEDOF Crosses Crosse	arch the we love Courly Courly Courl Courly Courl Arrow Resources Not solves	eb and Windows for Associate Selar for 15 12 2 for 15 12 2 for 15 12 2 for 15 12 2 for 15 12 1 for 15 12 12 12 12 12 12 12 12 12 12 12 12 12	tathéotor V entre arpat — Classif fr ans takan to — Evaluation — Evaluation — Summary — Summary	er model ( na class v o build mo on on boil ma savified P Classified P	foll rowinin Alos: Ver del: O secon opili	(7 APE)	3 6		1 1		<b>e</b>	<u>6</u> 4			5	• 小 弓 <sub>兆</sub>	+ 11 05:41 1-03-20
223 45207 4520	arch the will David Churk County Churk County Churk County Churk Marchador National Churk Marchador National National Churk Marchador National Churk Marchador Nation Nation Nation Nation Nation Na	eb and Windows for Associate Sela- for III - 2 Sec	telletary v nele arget enk predi- net somet = Doulout- ters predi- net somety = Summery ope mati- sometry ()	er model ( ta class v o build mo en tori ta loss fassified P classified P	foll trainin altas: Tas del: D secon opili natances Instances	17 Art)					<u>ę</u>	<u>6</u> 4			5	<b>ا</b> (ب) الح ( ال	+ 11 05:41 1-03-20
22 23 24 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	ter Associate Selon ter Associate Selon ter Selon te	tableta and entre angut — Classif — Classif — Classif — Classif — Summary —	ander er nofel ( ta class v o build mo an on levi classified D Classified cite a erori pand error	foll monitoin alos: Ves del: D secon opili ==== tetatores Tostances	ng met) dr 6 0.1 0.11 0.12			1 1		<b>ę</b>	6 4			5	4) E <sub>10</sub>	+ 11 05:41 1-03-20
223 SECT S	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	Set     Approximate     Set     Set <td< td=""><td>tathéotov v enteranyat – Clevat fr enter predis – Statusty &amp; Clevat – Statusty &amp; Clevan – Statusty &amp; Statusty &amp; Clevan – Statu</td><td>er nodel ( to chain a contraction of the chain of the cha</td><td>fell restricts line: Wes split sectores for sectores</td><td>17 Art)</td><td></td><td></td><td>4 4</td><td></td><td><b>e</b></td><td><u>4</u> 4</td><td></td><td></td><td>5</td><td></td><td>+ 11 05:41 1-03-20</td></td<>	tathéotov v enteranyat – Clevat fr enter predis – Statusty & Clevat – Statusty & Clevan – Statusty & Statusty & Clevan – Statu	er nodel ( to chain a contraction of the chain of the cha	fell restricts line: Wes split sectores for sectores	17 Art)			4 4		<b>e</b>	<u>4</u> 4			5		+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	ter Associate Solor	tathésta angat — Classif fi merk pendit merk pendit merk pendit merk pendit merk pendit merk pendit merk pendit merk pendit	ander er nofel ( ta class v o build mo an on levi classified D Classified cite a erori pand error	full mainte line: De spill	rg set) dr 6 0 1 0.1 0.1 0.1 001					<b>e</b>	<u>a</u> 4			5	4) (15) <sub>(2</sub> )	+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windows for Associate Selection for A	talbéau v Antonio articles and a second articles and a second articles and a second articles are a summary of a second articles and a second articles and a second articles are a second at a second articles and a second articles are associated at a second at	er model ( ta class v o build mo en co looi do Classified D Classified Classified Classified car parts error plute error plute error plute error plute error	full mainte line: De spill	eg ast) dz 6 0 1. 1. 0.2 100 100 0 0 0 0 0 0 0			1 1		<b>ę</b>	6 <mark>4</mark>			5	4) ( 15) <sub>(2</sub> )	+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windows for Associate Selection for A	talbéau v Antonio articles and a second articles and a second articles and a second articles are a summary of a second articles and a second articles and a second articles are a second at a second articles and a second articles are associated at a second at	er noiel ( na class v o build mo an on loui Classified D Classified class e coor pared error pared error pared error of lantes	full mainfail alter Tes split === split === formers for prove troop	rg ment) da 0.1 0.1 100 100 8	332 332 1	110 E	1	xii 👔	<b>e</b>	<u>6</u> 4			5		+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windows for Associate Selection for A	talbéau v Antonio articles and a second articles and a second articles and a second articles are a summary of a second articles and a second articles and a second articles are a second at a second articles and a second articles are associated at a second at	er noiel ( na class v o build mo an on loui Classified D Classified class e coor pared error pared error pared error of lantes	full mainfail alter Tes split === split === formers for prove troop	rg ment) da 0.1 0.1 100 100 8	332 332 1	110 E	1 1 1	xii 👔	<u>ę</u>	6 4			5		+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windows to Remove bio Sec. 2 V Remove bio Remove	tabbato v entre argut not cargat an taken ( = Dyalout entre argut an taken ( = Dyalout entre argut an taken ( = Dyalout entre argut	er nodel ( tr nodel ( tr class v tr class v tr class v tr class v classified T classified T classified tr classified tr classifi	full mainten line: Tes del: O seron spili	rg Away) da 1 0.11 100 100 100 100 100 100 100 100	300 300 1 8 8 8 8 8 1	105 C	1 1 1 2 2 2	<u>х</u> і (1)	<u>ę</u>	6 4			5		+ 11 05:41 1-03-20
Web Ep Web Ep Web Ep Course Course Course Course O Los 4 Portholide Santheling Santheling Santheling	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windows to Remove bio Sec. 2 V Remove bio Remove	talbéau v Antonio articles and a second articles and a second articles and a second articles are a summary of a second articles and a second articles and a second articles are a second at a second articles and a second articles are associated at a second at	er nodel ( tr nodel ( tr class v tr class v tr class v tr class v classified T classified T classified tr classified tr classifi	Foll moving the second	g set) dz 0.1 0.1 100 0 0 9 8 Precision	333 333 1 1 1 0	the t	4 1			<u>a</u>			5	1 () 12 <sub>(2</sub>	+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	eb and Windowski blob	enter arpat enter arpat enter arpat en taken ( = Bvalast) = Bvalast en taken ( = Bvalast) en taken ( = Bvalast) enter arbat par enter enter arbat par enter enter arbat enter	er nodel ( tr nodel ( tr class v tr class v tr class v tr class v classified T classified T classified tr classified tr classifi	fell restricts los: Tes idi: 0 accor upili recor recor f9 forte 0 0 0 0	rg Away) da 1 0.11 100 100 100 100 100 100 100 100	300 300 1 8 8 8 8 8 1	105 C	1 1 1 2 2 2			<u>8</u> 4			5		+ 11 05:41 1-03-20
22 23 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	se Renark blog	Latitude of the second	er model ( ta class w en to build mo en to losi on losi ta class fiel class f	full controls dist: Descur opil.t toror toror top top top top top top top top top top	rg Away) da 1 0.11 100 100 100 100 100 100 100 100	300 300 1 8 8 8 8 8 1	105 C	1 1 1 2 2 2			8 X			5		+ 11 05:41 1-03-20
22 23 24 25 25 25 25 25 25 25 25 25 25	erch the win here Courly Courly Courl Courly Courl Courly Course Marca	se Renark blog	Latitude of the second	er model ( ta class v ta class v ta class v ta class v ta class v ta class field ta classifiel T classifiel T classifiel T classifiel v ta correct ta	full controls dist: Descur opil.t toror toror top top top top top top top top top top	rg Away) da 1 0.11 100 100 100 100 100 100 100 100	300 300 1 8 8 8 8 8 1	105 C	1 1 1 2 2 2			đ <mark>4</mark>			5		+ 11 05:41 1-03-20

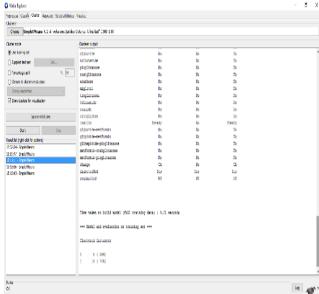
### Manisha et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 3, Issue 2, June 2016, pp. 76-78

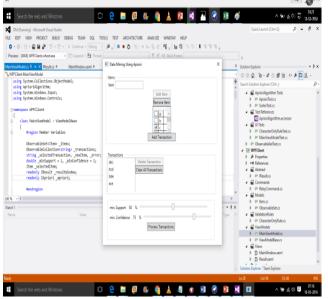
- a) ZeroR: It is the simplest classification method which relies on the target and ignores all predictors.
- b) This classifier simply predicts the majority category (Class).
- c) TP Rate: Rate of true positives (instances correctly classified as a given class).
- d) FP Rate: Rate of false positives (instances falsely classified as a given class).

The Confusion Matrix for classification.

#### III. CLUSTERING

Clustering technique find groups of items that are similar. It divides a data set so that records with similar content are in the same group, and groups are as different as possible from each other. Cluster data using the k means algorithm. We can use either the Euclidean distance (default) or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean.





This model took 0.01 second to build.Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "gender").

#### IV. APRIORI TOOL

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

 $F_l = (Frequent Itemsets);$ 

for  $(k=1; F_k \neq \emptyset; k++)$  do begin

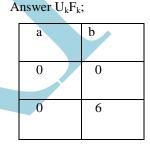
 $C_{k+1}$ =Apriori\_gen ( $F_k$ ); // New Candidates for all transactions  $t \in Database$  do begin

 $C'_t$  = subset ( $C_{k+1}$ , t); // Candidates Contained in *t* for all candidates  $c \in C'_t$  do

c.count++; end

 $F_{k+1} = \{C \in C_{k+1} c.count >= minimum support\}$ 

end end

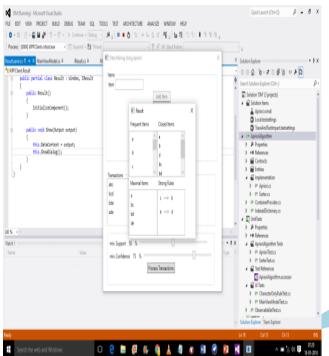


The tool was developed on visual studio. We are using Visual Studio .NET tool, and the coding is done in C Sharp (C#). For Front End WPS (Windows Presentation Foundation) is used. WPS is a powerful framework for building Windows applications.WPF employs XAML, an XML-based language, to define and link various interface elements.

The main window has options to add and delete transactions as well as we can set minimum support and minimum confidence on this page. If process transaction is clicked after selecting transaction, the result window will appear.

The result window will calculate strong rules by frequent items. It will find all frequent itemsets and get frequent items, Items whose occurrence in database is greater than or equal to the min.support threshold. Get frequent itemsets:Generate candidates from frequent items. Then prune the results to find the frequent itemsets and will generate strong association rules from frequent itemsets, rules which satisfy the min.support and min.confidence threshold.

## Manisha et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 3, Issue 2, June 2016, pp. 76-78



We took min sup 50% and min confi 75% after taking transactions as abc, bcd, bde, ace and from this strong rules are  $c \rightarrow b$ ,  $e \rightarrow d$ .

#### V. CONCLUSION

The paper shows the clustering & classification rule techniques for predictive analysis and evaluates the performance of association rule mining algorithms in the context of database partitioning. The focuses on Apriori and sampling algorithms for frequent item sets mining when the data is partitioned into a number of given segments. Apriori algorithm scans the database multiple number of times for counting the support for the item sets. The strong rules can be predicted by using this tool. We shows the clustering & classification of data on WEKA and it classified the data perfectly & shows the confusion matrix.

#### VI. REFERENCES

[1]. Association Rule Learning – Wikipedia, the free encyclopedia.

- [2]. Chen, H., Zhan, Y., Li, Y., (2010), "The application of Decision Tree in Chinese email Classification". In the proceeding of 9th International Conference on machine Learning and Cybernetics, 2010, pp. 305-308.
- [3]. Feng Yucai, "Association Rules Incremental Updating Algorithm", Journal of Software, Sept., 1998
- [4]. Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers.
- [5]. Execution of APRIORIAlgorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.
- [6]. Sonam Narwal, Comparison the Various Clustering and Classification Algorithms of WEKA Tools, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013.
- [7]. [7] Nimrat Kaur Sidhu, Rajneet Kaur "Clustering In Data Mining "International Journal of Computer Trends and Technology (IJCTT),V4(4):710-714 April Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.
- [8]. Huang, A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining
- [9]. [9] Shraddha, A Literature Review on Data Mining and its Techniques, Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X
- [10]. Antonie, Application of Data Mining Techniques for Medical Image Classification, SIGKDD conference, 2001
- [11]. Baluni, A comparative study of various approaches to explore factors for vehicle collision, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 1, Issue 3, September – October 2012.
- [12]. R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules," In Proc. of VLDB '94, pp. 487-499, Santiago, Chile, Sept. 1994.