RESEARCHERID
THOMSON REUTERS

[Patel, 4(11): November 2017]
DOI- 10.5281/zenodo.1059587

ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

GENDER WISE STUDENT PERFORMANCE ANALYSIS USING CLUSTERING TECHNIQUE OF DATA MINING

Bhavesh Patel*[1] & Jyotindra Dharwa[2]
*[1&2]Assistant Prfessor, MCA, AMPICS, Ganpat University

ABSTRACT

Data mining in educational field is a major application of data mining, it use machine learning to learn from data by studying algorithms and their constructions. In Data mining, clustering is the task of grouping aset of objects in such a way that objects in the same group are more similar to each other than to those in other groups. There are too many algorithms for clustering technique but k-mean algorithm is easy to interpret and understand so, in this paper clustering K-mean algorithm is discussed and applied on students data sets to find the gender wise performance in theory and practical subjects of computer science course. K-mean algorithm is implemented on student's dataset using WEKA tool.

Keywords: *Data mining, machine learning, clustering, WEKA, K-mean algorithm.*

I. INTRODUCTION

Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets [1]. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities [2].

Performance evaluation is one of the bases to monitor the progression of student performance in higher Institution of learning. Base on this critical issue, grouping of students into different categories according to their performance has become a complicated task. With the help of data mining methods, such as clustering algorithm, it is possible to discover the key characteristics from the students' performance and possibly use those characteristics for future prediction. There have been some promising results from applying k-means clustering algorithm with the Euclidean distance measure, where the distance is computed by finding the square of the distance between each scores, summing the squares and finding the square root of the sum [3].

This paper presents k-means clustering algorithm for finding the gender wise performance in theory and practical subject of the computer science course.

This research paper is organized in the following sections. Section I describes the introduction. Section II describes Literature survey. Section III deals with research objective and model. Section IV discusses used methodology in research work. Section V represent the implementation of model. Section VI described the conclusion and future work of research work.

ISSN 2348 – 8034
Impact Factor- 4.022

II. LITERATURE SURVEY

Kabakchieva applied classification techniques on the data collected from University of National and World Economy (UNWE) system. The work focused on predicting students‟ performance at enrolment stage by finding patterns from data related to students personal and pre-university information.[4]

Quadril andKalyankar applied classification and prediction techniques. These techniques helped them to build student drop out model and predict each student‟s academic performance by measuring student‟s Cumulative Grade Point Average (CGPA). [5]

Kovacic applied classification and feature selection techniques to identify the influencing factors at the enrolment stage. The data was collected from the Student Management System of open Polytechnic of New Zealand. The research showed that identification of students at risk even before they start their study could help the university to take proper actions to improve the academic success.[6]

Thames Valley University developed the Mining Course Management Systems (MCMS) project to improve student retention strategies by analysing student behaviour and early identification of those at risk. In this project, association, clustering, classification and prediction techniques were applied on collected data from current university information systems (library, student administration, online learning system, online resource system, online test system and so on) and integrated into data warehouse. The project could build models to predict each single student performance and their behaviour. [7]

Delavari applied classification and prediction as main techniques to educational data. Through these techniques, she could discover pattern of successful student in a subject as well as student success rate for individual lecturer and predict the rate. The knowledge gained from these models could help institution for better decision making in setting up new strategies or improving the current strategies to increase the student success rate. [8]

III. RESEARCH OBJECTIVE AND MODEL

Research Objective
Based on the above literature survey we have observed that there is not any model is defined by researcher for Gender wise student performance in theory and practical subject. Model 1: (male, female student performance in theory and practical subject)

So in the proposed work we have implemented the following model using clustering k-mean algorithm:

Model
Gender Wise Result Analysis for all the Theory and Practical Subjects of the course.

IV. METHODOLOGY

Development of k-mean clustering algorithm
Given a dataset of n data points $x_1, x_2, \ldots, x_n$ such that each data point is in $\mathbf{R^d}$, the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points $\{m_j\}$ (j=1, 2, …, k) in $\mathbf{R^d}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\min_j d^2(x_i, m_j)\right] \quad (1)$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between x_i and m_j. The points $\{m_j\}$ (j=1, 2, …,k) are known as cluster centroids. The problem in Eq.(1) is to find k cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized.

The *k*-means algorithm provides an easy method to implement approximate solution to Eq.(1). The reasons for the popularity of *k*-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data.

The *k*-means algorithm can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function in Eq.(1). The *k*-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids. The problem of finding the global minimum is NP-complete. The *k*-means algorithm updates cluster centroids till local minimum is found. Fig.1 shows the generalized pseudocodes of *k*-means algorithm; and traditional k-means algorithm is presented in fig. 2 respectively.

Before the *k*-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say *l*, where the positive integer *l* is known as the number of *k*-means iterations. The precise value of *l* varies depending on the initial starting cluster centroids even on the same dataset. So the computational time complexity of the algorithm is $O(nkl)$, where *n* is the total number of objects in the dataset, *k* is the required number of clusters we identified and *l* is the number of iterations, $k \leq n$, $l \leq n$ [9].

Steps to implement the K-means clustering is as per the following

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values
Step 2: Initialize the first K clusters

Take first k instances orTake Random sampling of k elements
Step 3: Calculate the arithmetic means of each cluster formed in the dataset.
Step 4: K-means assigns each record in the dataset to only one of the initial clusters
V. Each record is assigned to the nearest cluster using a
Each record is assigned to the nearest cluster using ameasure of distance (e.g Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

Algorithms for K-means clustering is as per the following

MSE = largenumber;
Select initial cluster centroids $\{m_j\}_j$
$K = 1$;

Do
OldMSE= *MSE*;
*MSE*1 = 0;

For $j = 1$ to k
$m_j = 0; n_j = 0$;
Endfor

For $i = 1$ to n
For $j = 1$ to k

Compute squared Euclidean distance $d^2(x_i, m_j)$;

Endfor

ISSN 2348 – 8034
Impact Factor- 4.022

Find the closest centroid m_j to x_i;
$m_j = m_j + x_i; n_j = n_j + 1;$
$MSE1 = MSE1 + d^2(x_i, m_j);$

Endfor

For $j = 1$ to k
$n_j = \max(n_j, 1); m_j = m_j / n_j;$
Endfor

$MSE = MSE1;$
while ($MSE < OldMSE$)

V. IMPLEMENTATION

Implementation K-Means is used to analyse the gender wise performance in theory and practical subject of computer science course. We have used WEKA tool for performing analysis on these selected data.

Total we have collected 3558 students' dataset where we have applied the k-means clustering algorithm to implement the model.

To perform the gender wise analysis in theory and practical subjects we have considered the result of following theory subjects and practical subjects as a parameters.

Theory Subject List	Practical Subject List
Communication skill	C Programming Language
Information Technology	Microsoft Office
Basic of Computer Network	Digital Electronics
Communication skill-II	Advance concepts of C programming
Computer Organization	Web programming
Environment Disaster Management	Data and File Structure
Open Source Technology	Core Java programming
Management Information System	Database Management System
Networking – I	DotNet
Career & Personality Development	Advance Data base Management System
Software Testing	Computer Graphics
	Advance Dotnet
	Operating System

Screen shot for implementation of clustering K-means algorithm using WEKA tool is as per the following:

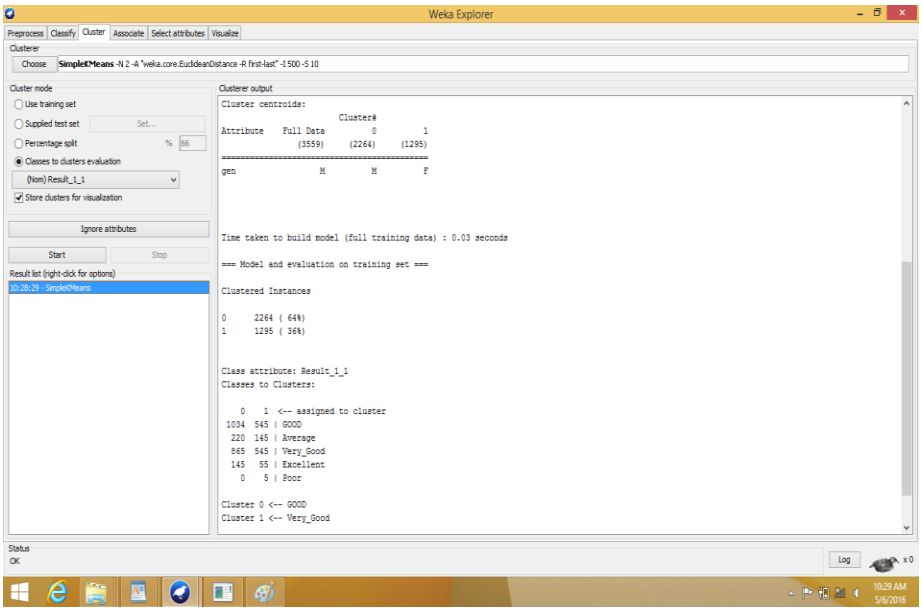


Figure 1: Gender wise analysis for subject 1 of semester 1 in WEKA tool using K-mean algorithm

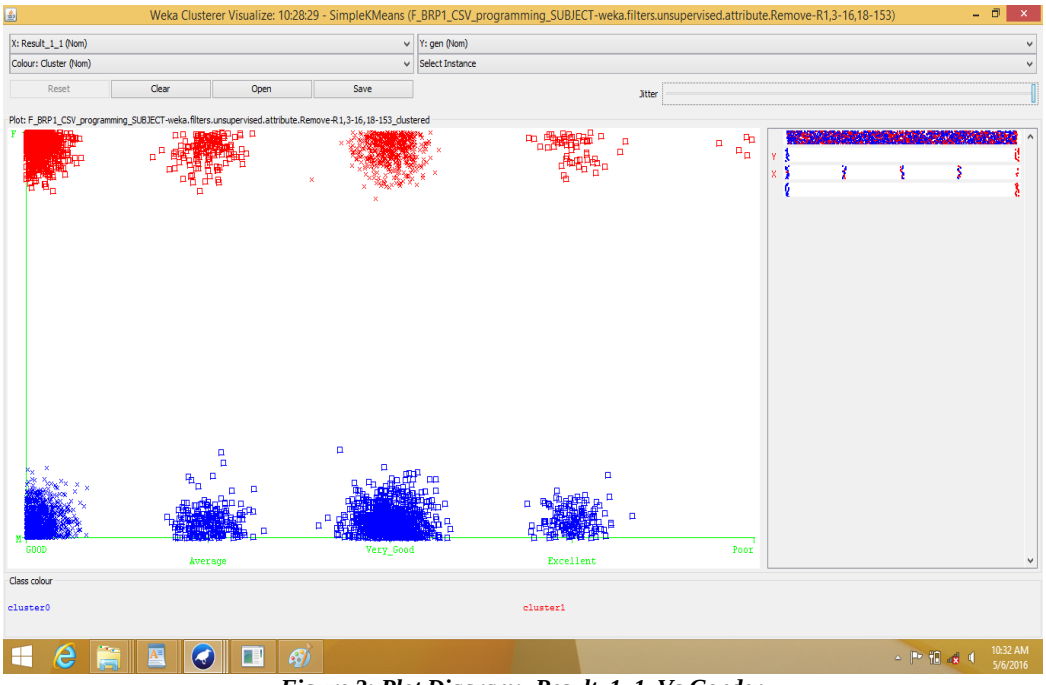


Figure 2: Plot Diagram: Result_1_1 Vs Gender

ISSN 2348 – 8034
Impact Factor- 4.022

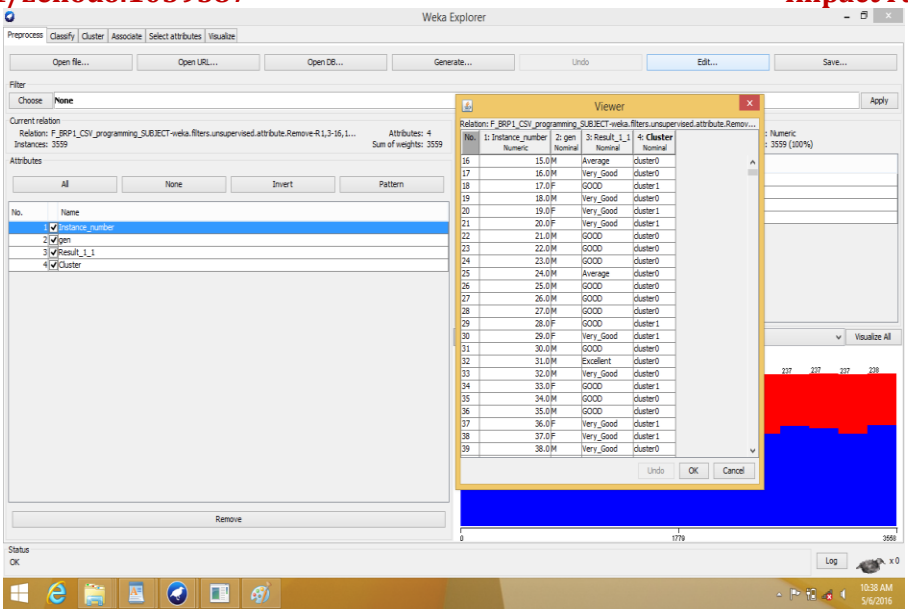
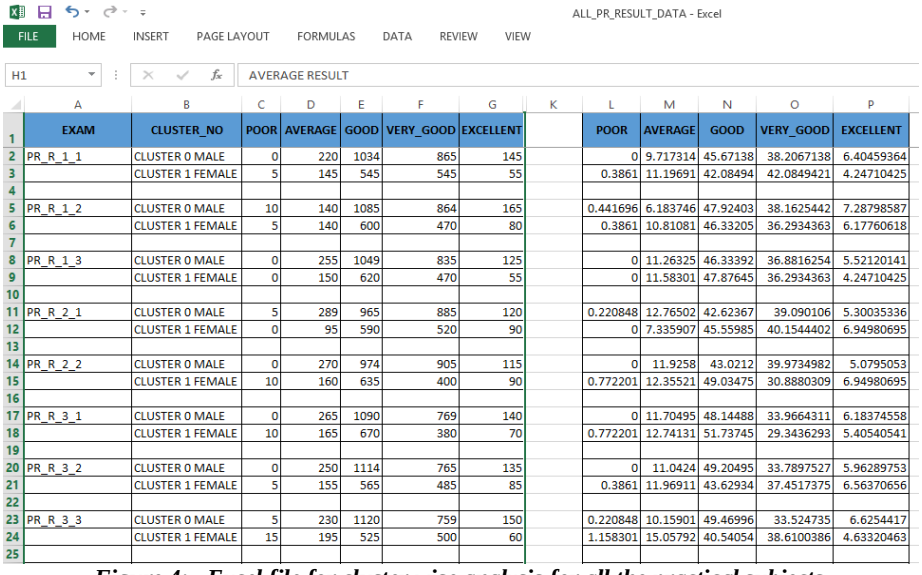


Figure 3: Store data in cluster 0 and 1, 0- male and 1- Female

Note

Same gender vs result analysis process will be repeat for all the theory and practical and save the result buffer data and cluster visualize data for the future use. make the separate excel sheet for all the theory subject result vs gender and all the practical subject vs gender. by completing the above task we got the following result in excel format.



	EXAM	CLUSTER_NO	POOR	AVERAGE	GOOD	VERY_GOOD	EXCELLENT		POOR	AVERAGE	GOOD	VERY_GOOD	EXCELLENT
2	PR_R_1_1	CLUSTER 0 MALE	0	220	1034	865	145		0	9.717314	45.67138	38.2067138	6.40459364
3		CLUSTER 1 FEMALE	5	145	545	545	55		0.3861	11.19691	42.08494	42.0849421	4.24710425
4													
5	PR_R_1_2	CLUSTER 0 MALE	10	140	1085	864	165		0.441696	6.183746	47.92403	38.1625442	7.28798587
6		CLUSTER 1 FEMALE	5	140	600	470	80		0.3861	10.81081	46.33205	36.2934363	6.17760618
7													
8	PR_R_1_3	CLUSTER 0 MALE	0	255	1049	835	125		0	11.26325	46.33392	36.8816254	5.52120141
9		CLUSTER 1 FEMALE	0	150	620	470	55		0	11.58301	47.87645	36.2934363	4.24710425
10													
11	PR_R_2_1	CLUSTER 0 MALE	5	289	965	885	120		0.220848	12.76502	42.62367	39.090106	5.30035336
12		CLUSTER 1 FEMALE	0	95	590	520	90		0	7.335907	45.55985	40.1544402	6.94980695
13													
14	PR_R_2_2	CLUSTER 0 MALE	0	270	974	905	115		0	11.9258	43.0212	39.9734982	5.0795053
15		CLUSTER 1 FEMALE	10	160	635	400	90		0.772201	12.35521	49.03475	30.8880309	6.94980695
16													
17	PR_R_3_1	CLUSTER 0 MALE	0	265	1090	769	140		0	11.70495	48.14488	33.9664311	6.18374558
18		CLUSTER 1 FEMALE	10	165	670	380	70		0.772201	12.74131	51.73745	29.3436293	5.40540541
19													
20	PR_R_3_2	CLUSTER 0 MALE	0	250	1114	765	135		0	11.0424	49.20495	33.7897527	5.96289753
21		CLUSTER 1 FEMALE	5	155	565	485	85		0.3861	11.96911	43.62934	37.4517375	6.56370656
22													
23	PR_R_3_3	CLUSTER 0 MALE	5	230	1120	759	150		0.220848	10.15901	49.46996	33.524735	6.6254417
24		CLUSTER 1 FEMALE	15	195	525	500	60		1.158301	15.05792	40.54054	38.6100386	4.63320463
25													

Figure 4: Excel file for cluster wise analysis for all the practical subjects

ISSN 2348 – 8034
Impact Factor- 4.022

Table 2: Genderwise mean value for Practical subjects

GENDER	POOR	AVERAGE	GOOD	VERYGOOD	EXCELLENT
MALE	0.141973751	10.69851085	46.50113579	36.96365472	5.694724886
FEMALE	0.413678985	12.10700496	45.61500276	36.48648649	5.322669608

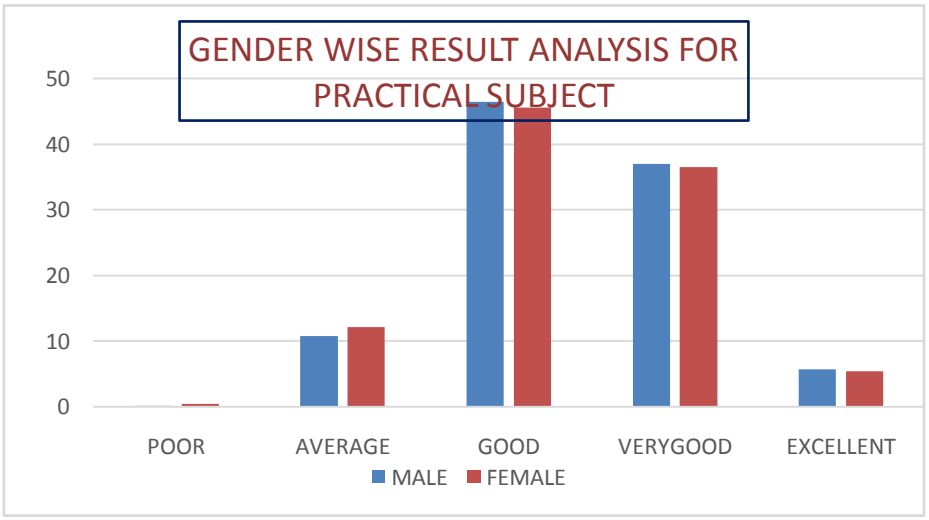
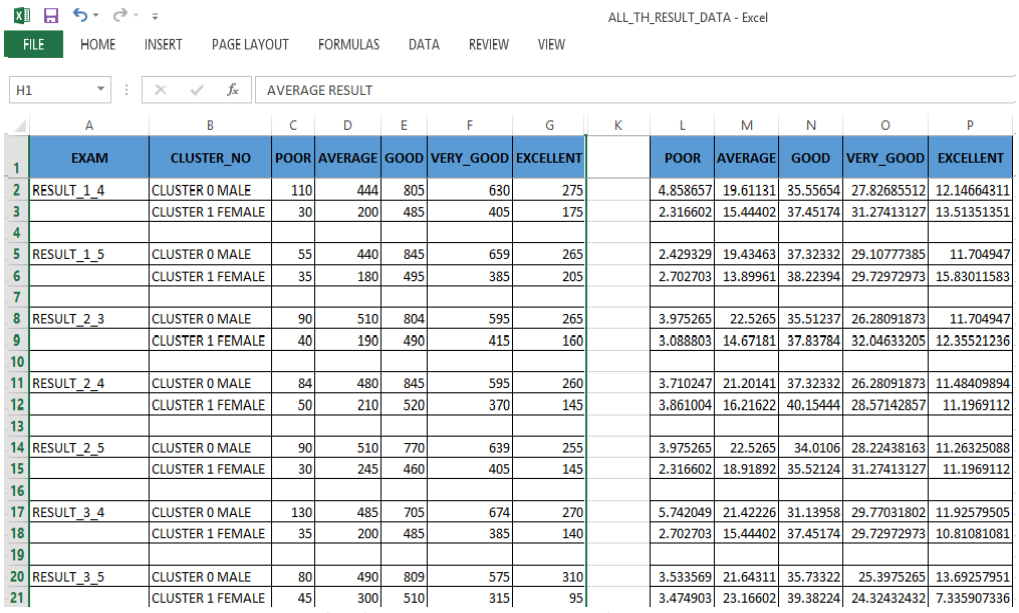


Figure 5: Genderwise mean value for PRACTICAL subjects



	A	B	C	D	E	F	G	K	L	M	N	O	P
1	EXAM	CLUSTER_NO	POOR	AVERAGE	GOOD	VERY_GOOD	EXCELLENT		POOR	AVERAGE	GOOD	VERY_GOOD	EXCELLENT
2	RESULT_1_4	CLUSTER 0 MALE	110	444	805	630	275		4.858657	19.61131	35.55654	27.82685512	12.14664311
3		CLUSTER 1 FEMALE	30	200	485	405	175		2.316602	15.44402	37.45174	31.27413127	13.51351351
4													
5	RESULT_1_5	CLUSTER 0 MALE	55	440	845	659	265		2.429329	19.43463	37.32332	29.10777385	11.704947
6		CLUSTER 1 FEMALE	35	180	495	385	205		2.702703	13.89961	38.22394	29.72972973	15.83011583
7													
8	RESULT_2_3	CLUSTER 0 MALE	90	510	804	595	265		3.975265	22.5265	35.51237	26.28091873	11.704947
9		CLUSTER 1 FEMALE	40	190	490	415	160		3.088803	14.67181	37.83784	32.04633205	12.35521236
10													
11	RESULT_2_4	CLUSTER 0 MALE	84	480	845	595	260		3.710247	21.20141	37.32332	26.28091873	11.48409894
12		CLUSTER 1 FEMALE	50	210	520	370	145		3.861004	16.21622	40.15444	28.57142857	11.1969112
13													
14	RESULT_2_5	CLUSTER 0 MALE	90	510	770	639	255		3.975265	22.5265	34.0106	28.22438163	11.26325088
15		CLUSTER 1 FEMALE	30	245	460	405	145		2.316602	18.91892	35.52124	31.27413127	11.1969112
16													
17	RESULT_3_4	CLUSTER 0 MALE	130	485	705	674	270		5.742049	21.42226	31.13958	29.77031802	11.92579505
18		CLUSTER 1 FEMALE	35	200	485	385	140		2.702703	15.44402	37.45174	29.72972973	10.81081081
19													
20	RESULT_3_5	CLUSTER 0 MALE	80	490	809	575	310		3.533569	21.64311	35.73322	25.3975265	13.69257951
21		CLUSTER 1 FEMALE	45	300	510	315	95		3.474903	23.16602	39.38224	24.32432432	7.335907336

Figure 6: Excel file for cluster wise analysis for all the theory subjects:

ISSN 2348 – 8034
Impact Factor- 4.022

Table 3: Gender wise mean value for all theory subjects

GENDER	POOR	AVERAGE	GOOD	VERYGOOD	EXCELLENT
MALE	3.931095406	21.00465789	35.82958561	27.38917443	11.44394475
FEMALE	3.123903124	17.02351702	38.08353808	28.92242892	12.81151281

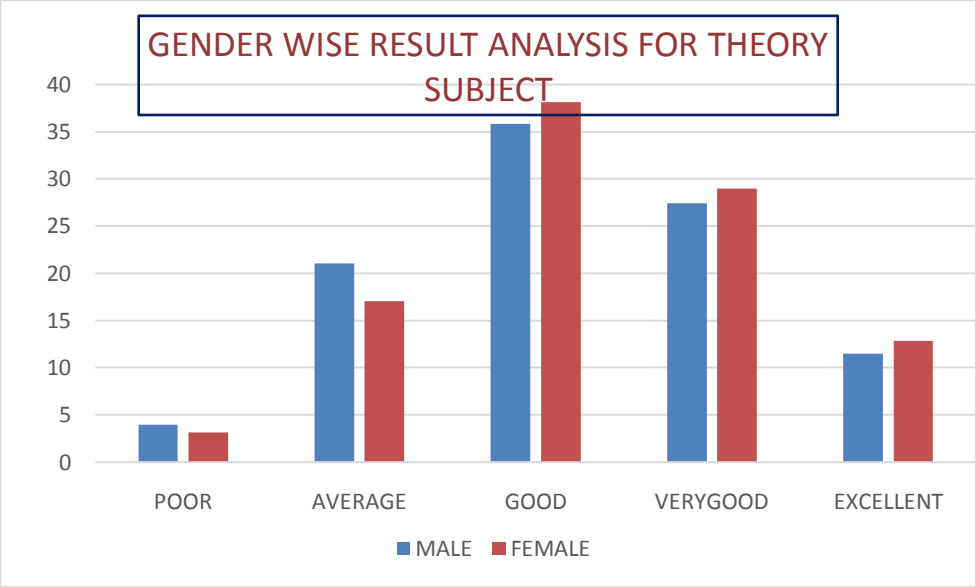


Figure 7: chart for gender wise analysis of all theory subjects

ISSN 2348 – 8034
Impact Factor- 4.022

Result of Implementation

Model	Analyzed Outcome
Gender Wise Result Analysis for all the Practical Subject of the course.	**Assumption:** **For the practical subject our assumption is that male students can get good score as they are more powerful in logic and reasoning compare to girls** **Result of analysis:** But above myths is not proved by above analysis. This analysis seems that female students' also good performer in programming subjects as per the directed in the chart, there is not any major variation in the result of male performer and female performer.
Gender Wise Result Analysis for all the Theory Subject of the course.	**For the theory subjects our assumption is that female students are good in craming compare to male students:** this myth is proved by analysis as directed by the given chart and mean table, performance ratio of female students is higher in the category of good, very good and excellent then male students' performer.

VI. CONCLUSION

In this research paper we have applied clustering data mining technique. There are too many clustering algorithms are available in data mining but we have used clustering algorithm K-means to implement the model as it is very simple and easy to interpret. To perform the analysis, we have collected total 3558 student recordset with the gender wise result of all the theory and practical subjects of computer science course from the various higher educational institute. We have applied WEKA data mining tool to implement the clustering K-means algorithm for getting the gender wise performance in theory and practical subjects. As an analysis outcome, we proved that there is not any major variation in the performance of boys and girls particularly in practical subject. This analysis also proved that girls are performed well as compared to boys students in the theoretical subjects. So, if boys students are tried to improve the performance in theory subjects then definitely higher educational institute improve the result of educational institution. This result is also helpful to the mentors and management of educational institute to take the decision in right direction. Further, we would like to extend this work by implementing the stream wise performance of students using clustering algorithms.

VII. ACKNOWLEDGEMENT

We are grateful to my collegue and friend for his valuable guidance and support to improve my research work. We also thankful to our associate professor Dr. Jyotindra N. Dharwa sir for guiding us regarding the statistical concepts. We thanks to all the helping hands those have directly or indirectly support to us.

REFERENCES

1. *Connolly T., C. Begg and A. Strachan (1999) Database Systems: A Practical Approach to Design, Implementation, and Management (3rd Ed.). Harlow: Addison-Wesley.687*
2. *Alaa el-Halees (2009) Mining Students Data to Analyze e-Learning Behavior: A Case Study.*
3. *Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006*
4. *D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernetics and Information technologies, vol.13, pp.61-72, March 2013.*
5. *M.N. Quadril, and N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Global Journal of Computer Science and Technology, vol.2, pp.2-5, April 2010.*

6. *Z.J, Kovačić, "Early Prediction of Student Success: Mining Students Enrolment Data", Proceedings of Informing Science & IT Education Conference (ISITE), June 2010.*
7. *S. Oussena, "Mining Course Management Systems (MCMS)", Final Report, Joint Information Systems Committee, Jun 2010.*
8. *N. Delavarani, M.R. Beikzadeh , S. Phon-Amnuaisuk "Data Mining Application in Higher Learning Institutions" , Informatics in Education, vol. 7, pp. 31–54, 2008.*
9. *Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006*