

Cluster analysis of socio-economic factors and academic performance of school students

Kapila Devi, Saroj Ratnoo

Department of Computer Science and Engineering, Gurur Jambheshwar University of Science and Technology, Hisar, India

Article Info

Article history:

Received Jan 18, 2023

Revised Apr 30, 2023

Accepted May 6, 2023

Keywords:

Clustering analysis
Educational data mining
Partition around medoid
Socioeconomics status
Students' academic
performance

ABSTRACT

The objective of the paper is to examine the academic performance of students' vis-a-vis socio-economic factors using clustering analysis. The grades obtained in the 10th class are taken as the measure of academic performance the variables such gender, caste, parental education and occupation are considered as the socio-economic indicators. Three clustering algorithms are employed. The K-medoid performs better in the validation process to form the groupings based on intra-cluster homogeneity and inter-cluster heterogeneity. The clustering analysis results in two interesting groups of the students. One of the clusters is dominated by the students of general category and the other one by the scheduled caste category. Next, the appropriate statistical tests are applied to determine the factors that significantly differ in the two clusters. Cluster analysis shows that caste, parents' education and occupation, and family income are the differentiating factors between the two groups. However, we are unable to establish significant difference between the academic performance of the two groups of students at a 5% significance. The research carried out in this paper may be beneficial for making policies to bridge the gap in the educational attainment of the students from deprived sections of society.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kapila Devi
Department of Computer Science and Engineering
Gurur Jambheshwar University of Science and Technology
Hisar, India
Email: kapila628@gjust.org

1. INTRODUCTION

Education is considered one of the most enabling factor to empower individuals and nations all over the world. However, the socio economic factors (SES) such as gender, poverty, family status, and area of residence play influential role in shaping students' academic success [1]–[4]. Several researchers have studied the association between SES and academic achievement. Considine and Zappalà [5], investigated the extent of the impact of socioeconomic, individual and family related factors on academic performance of financially disadvantaged students in Australia. The authors found that ethnicity, gender, absence from school, parents' educational status, type of housing, and student age were all statistically significant variables for the academic performance. Malecki and Demaray [6] also established moderate and significant relationships between social support scores and grade point average (GPA) scores. The study suggested that social support could restrain the relationship between academic performance and poverty. Islam and Khan [7] showed a correlation between socio-economic status and academic attainment of 12th class School students. The study found a significant difference in the academic success of different SES groups. Recently, a research has investigated the significance of self-concept in relation to SES and school academic accomplishment. This research is based on

345 students at junior high school in China. The results confirmed that self-concept and family SES were significantly correlated with the performance of students in mathematics and Chinese [8].

In Indian scenario, gender and caste are unique features that leads to the educational deprivation [9]. Bhagavatheeswaran [10] looked at the barriers and enablers to scheduled caste (SC) and scheduled tribe (ST) adolescent girls entering and completing the secondary education in Northern Karnataka. They identified the caste barriers for the entry and retention of the SC/ST girls in secondary education, and highlighted the need for working to change the belief and expectation around gender norm as well as improving the quality of education. John *et al.* [11] examined the association of caste categories (general, other backward caste (OBC) and SC/ST) with the academic achievement of Indian adolescents belonging to the same school environment. The research investigated the psychological measures of life satisfaction and self-esteem, gender, age and family income of X and XI grade students. The study confirmed a significantly positive relationship between the caste categories and academic performance.

Most of the earlier studies regarding attainment of educational outcomes follow the traditional statistical approaches on smaller amounts of data. Educational data mining (EDM) is the relatively new domain that employs machine learning techniques to discover novel and interesting trends from educational data. EDM has been successfully applied for prediction, association rule mining, clustering analysis and outlier detection [1], [12]. The predictive and clustering analytics are the two major tools of educational data mining. The predictive analytics as its name suggests, is a supervised learning technique that can predict a response variable based on the values of predictors. There exist many studies for predicting the students' academic performance based on SES [13]–[16]. Notwithstanding to predictive modelling, clustering analysis is an unsupervised machine learning technique that discover natural groupings in the data based on intra cluster homogeneity and inter cluster heterogeneity of the data instances. The clustering algorithms minimises the intra cluster homogeneity and maximises the inter cluster heterogeneity, i.e., the data instances within a cluster are very similar to each other whereas data instances falling in different clusters are dissimilar to each other. Several clustering algorithms have been applied on the educational data [17]–[21].

K-means is one of the most popular partitioning clustering algorithms that has been broadly applied in the domain of educational data mining [22]. Prachuabsupakij and Chiengpongpan [21] applied k-means to investigate the relationship between personal data and students' performance in the information technology programme. The authors found five clusters in the student's data. Furthermore, the mother and father's education, income, and gender were the significant variables that influenced the group formations. Another study performed k-means clustering to find the number of clusters using student GPA [23]. Bharara *et al.* [24] performed disposition analysis using K-means clustering. The study concluded that students' interaction with learning management system and parents' involvement in students' education were the significant features that affected the students' academic performance.

K-means clustering algorithms do not work proper in the presence of mixed type of attributes and outliers. It may also converge to sub-optimal solutions [25]. We have applied K-medoid, hierarchical, model based and fuzzy clustering algorithms on Jawahar Navodaya Vidyalaya (JNV) students' data on socio economic indicators and academic performance. JNVs are Indian government initiative to provide the state-of-the-art education to the talented students from deprived and rural background. It will be interesting to investigate whether these schools have fulfilled the aim of narrowing the gap between the academic performance of the student groups belonging to different socio-economic status. We didn't come across any study to answer the above question. This study is a novel contribution to the domain of EDM.

In this paper, we perform a statistical as well as clustering analysis of socio-economic factors and academic performance of JNV students of five successive batches of the admission years 2006-07, 2007-08, 2008-09, 2009-10 and 2010-11 at Khunga Kothi, Jind district, Haryana (India). We collected socio-economic data related to location of residence (rural/urban), gender (male/female), educational and occupational status of parents and, family size and income of the students. The grades achieved in tenth class of central board of secondary education (CBSE) results is taken as the indicator of academic performance. K-medoid clustering algorithm is used to divide the students in homogenous groups. The two groups are described using descriptive statistics and then inter-group differences are analysed by employing suitable statistical tests. The aim of this research is to explore if the JNV system is able to cover-up the gap in academic performance of students originating from differential socio-economic status. Towards this aim, we intend to answer the following two research questions:

- Does there exist a significant difference in the academic performance of the groups of JNV students belonging to differential socio-economic status?
- Which socio-economic indicators are significantly different in the groups?

The result of this research show that there are two groups of students in the JNV. One group consists of students belonging to general category with their parents involved in agriculture. This group have relatively high income. The second group comprises of scheduled caste students with labourer parents who have low income. The research also shows that the female students' academic performance is significantly better than

their male counterparts. The academic performance of the two groups is not found to be significantly different at 5% level of significance, however, the students from the higher SES group perform better at a significance level of 10%. This shows that the JNV schools have been successful in reducing the academic performance gap to some extent.

The organization of the paper is as shown in: section 2 describes the research methodology used in this research. Section 3 presents the results and compare these to other relevant studies. Section 4 concludes the research and points to its limitations and future direction.

2. RESEARCH METHOD

This section describes the experimental design. Initially, the section includes statistical summary of the data. Next, it describes data pre-processing steps, application of clustering algorithms, statistical tests and tools used.

2.1. About Jawahar Navodaya Vidyalayas (JNVs)

To make the quality education accessible to the children of socially and financially deprived backgrounds predominantly from rural area, the national policy on education 1986 envisioned opening of residential schools called JNVs. Education at JNV is fully supported by government and is completely free for girls, especially abled and students belonging to SC/ST categories and students from families whose income is below poverty line. The admissions to JNVs are made to class VI through Jawahar Navodaya selection test conducted by CBSE. The seats in JNVs are reserved for children belonging to SC/ST categories in ratio of their population in the respective district of schools' location, but not less than the national average of 15%. Further, 75% seats are reserved for children belonging to rural areas. JNVs follow CBSE curriculum like most of the schools in urban centres. According to the survey conducted by the development evaluation advisory committee, the JNV schools have been able to help in bringing excellence among rural talents in education and learning life skills. The pass percentage has been significantly higher in class X board examinations compared to the other CBSE schools [26].

2.2. Data collection and descriptive summary

There are 576 JNVs in India spread across 27 states and 7 Union Territories. The state of Haryana has 21 JNVs. The data is collected from Jawahar Navodaya Vidyalaya Khunga Kothi, Jind, district random selected out of 21 JNVs located in Haryana (India). The data from the JNV include attributes related to socio-economic status and as well as academic performance of the students. The socio-economic data come from the enrolment forms which are collected from the students at the admission time in VI class while academic performance data is taken in the form of tenth class grades from CBSE results. The data sample pertains to 257 students for five consecutive admission years 2006-07, 2007-08, 2008-09, 2009-10 and 2010-11. The detail of strength of student taken is displayed in Table 1.

Table 1. Data count

Sr. No	Admission year	Student strength	10 th class year	Student strength
1	2006-07	40	2010-11	30
2	2007-08	65	2011-12	52
3	2008-09	75	2012-13	59
4	2009-10	75	2013-14	61
5	2010-11	70	2014-15	55

The attribute details with their respective distribution in the dataset is shown in Table 2 and Table 3. The gender distribution of the students shows a bias towards male students despite that one third seats are reserved for girls. With respect to caste distribution, there are 33.9% students belong to SC, 10.5% students belong to OBC, and 55.6% students belong to the general category. The representation of SC students is above their population ratio in national (16.2%), Haryana (20.17%) and Jind District (23.16%) as per the 2011 census [26]. The sample data has 84% of students from rural area as compared to only a 16% from urban areas. Since agriculture is the main occupation of the rural population in Haryana, therefore, 45% of fathers' occupation is agriculture followed by 33% of the fathers being labours. Only 12% fathers are from service sector. As most of the women in the rural area are unemployed, 76% mothers are housewives. The Jind district is relatively backward in literacy in Haryana as male and female literacy rates of the district are 80.81% and 60.76% compared to the male female literacy rates of 84.06% and 65.94% of the state of Haryana which is higher than the national male female literacy rates of 82.14% and 65.46% respectively [27]. The male female literacy rates (78% and 61%) of the sample data reflects

the same. Two of the variables (number of family members and grade point obtained by the students in tenth class on ten-point scale) are discrete variables whereas income is treated as numeric variable. The five-point summaries of these variables are also given in Table 3. The median income is only 20,000 and the median family members is 5. The median grade point is 8 which is impressive indeed.

Table 2. Data description (categorical variables)

Attributes	Description	Codes (categories)	Frequency distribution	Percent (%)		
Gender	Male or female	Male (M)	164	64		
		Female (F)	93	36		
Caste	Caste of the students	Scheduled caste (SC)	87	33.9		
		Other backward caste (OBC)	27	10.5		
		General (GEN)	143	55.6		
Residence	Urban or rural	Rural (RL)	217	84		
		Urban (UR)	40	16		
Father_edu	Father education	Illiterate (ILTR)	30	12		
		Primary (PMRY)	14	05		
		Middle (MDDL)	44	17		
		Matric (MTRC)	121	47		
		Secondary (SECD)	33	13		
		Graduate (GRDTA)	15	06		
		Postgraduate (PGDT)	1	0.6		
		Father_occup.	Father occupation	Agriculture (AGR)	116	45
Mother_edu.	Mother education	Labourer (LBR)		89	35	
		Service (SRV)		38	15	
		Self employed (SEMP)		14	05	
		Illiterate (ILTR)	102	39		
Mother_occup.	Mother occupation	Primary (PMRY)	41	16		
		Middle (MDDL)	45	18		
		Matric (MTRC)	55	21		
		Secondary (SECD)	10	4		
		Graduate (GRDTA)	4	2		
		Postgraduate (PGDT)	0	0		
		Mother_occup.	Mother occupation	Housewife (HW)	195	76
				Labourer (LBR)	23	9
Anganwadi (AWADI)	06			2		
Agriculture (AGR)	33			13		

Table 3. Data description (numerical variables)

Attributes	Description	Type	Five point summary
Income	The income of the family	Numeric	Min. Q1 Median. Q3. Max 1000 12000 20000 40000 300000
Family members	Number of family members	Discrete	Min. Q1 Median. Q3. Max 1 5 5 6 14
Grades	Overall grade obtained in tenth class	Discrete	Min. Q1 Median. Q3. Max 5 8 8 9 10

2.3. Data pre-processing

Data pre-processing is a crucial step of data mining process that converts the raw data into useful and efficient data forms. The raw data is never in a form ready to be mined. The raw data may contain missing values and redundant attributes. It may also require some data transformation to make the data mining process effective. We pre-processed the data in the following ways.

The raw data contained some missing values. We replaced the missing values for categorical and discrete variables with their mode values. For continuous variable income, the missing values were replaced with the average income of the category of the fathers' occupation. For example, if we found a missing value for the family income variable of a father involved in agriculture then it was replaced with the average family income of all the remaining fathers involved in agriculture. The average incomes of the different categories of father occupation are Table 4.

Not all the attributes are equally relevant for any data mining task. Some attributes do not contribute any important information, but these may prove to be noisy and influence the performance of data mining algorithm adversely. During data selection invaluable attributes such as roll no, names of the students and their parents were removed.

This step is applied to transform the data in an appropriate form that is useful for the data mining process. In this step, the four attributes regarding parents' education and occupation were discretised into various categories given in Table 2. We applied normalization to three attributes (family income, grade point and number

of family members) to scale the data values between -1 and 1. This was done to avoid undue weightage given to the attributes with larger range values while computing similarity of examples for clustering algorithm.

Table 4. Average income (Rs) of different categories of father occupation

Occupation	Average income
AGR	28,402
LBR	19,192
SEMP	34,500
SVR	1,22,155

2.4. Applying clustering algorithms and determining the number of clusters

Here, we have applied centroid-based, hierarchical and model-based clustering algorithms. We have also validated the performance of the three algorithms and determined the number of clusters (K) using silhouette width and dunn index. The K-Mediod algorithm stands winner out of the three.

2.4.1. K-medoid

It is also called partition around medoid (PAM). K-Medoid algorithm uses one representative object per cluster as the reference point for clustering in place of the mean value of objects contained in a cluster as is the case for the K-means algorithm. To group n objects into K clusters, the K-medoid algorithm initially selects K representative objects randomly. Then it keeps replacing the representative objects by other non-representative objects until the quality of clusters keeps improving. The algorithm iterates until it finds the truly centrally located objects (medoids) for K clusters. Since the data involve categorical as well as numeric attributes, we have used 'gower' dissimilarity for computing the distance matrix.

2.4.2. Diana

Divise analysis, it is a topdown hierarchical clustering algorithms. It begins by assuming all the observations in a single cluster and then successively partitions the cluster until each observation ends up in a single cluster. At each step, the cluster tree grows by selecting the cluster with the largest diameter for split. The diameter of a cluster is the largest divergence between any of its observations. Subsequently, the selected cluster is split into two most heterogeneous clusters.

2.4.3. Model based clustering

In model-based clustering, the main idea is to consider that data are coming from a mixture of underlying probability distributions. The most accepted model based clustering approach is the Gaussian mixture model (GMM), where each observation is assumed to come from one of the K (number of clusters) multivariate normal distributions [28]. We selected the best clustering algorithm out of the above three based on connectivity, Dunn index and Silhouette width as shown in Table 5. Based on the below table, out of the clustering algorithm, PAM stands out with the lowest connectivity, highest Dunn index and maximum silhouette width. The optimal number of clusters across all the algorithms is 2.

Table 5. Cluster validation

Nclust		2	3	4	5	6
Clustering technique						
Diana	Connectivity	36.42	64.12	72.38	73.16	78.19
	Dunn	0.24	0.09	0.10	0.10	0.11
	Silhouette	0.34	0.33	0.29	0.27	0.25
PAM	Connectivity	23.74*	38.30	62.33	80.56	91.62
	Dunn	0.26*	0.09	0.25	0.10	0.11
	Silhouette	0.35*	0.31	0.31	0.30	0.27
Model based clustering	Connectivity	33.61	85.35	72.58	70.22	87.98
	Dunn	0.16	0.10	0.11	0.10	0.11
	Silhouette	0.35	0.30	0.27	0.23	0.22

2.5. Characterizing the clusters

To answer the research question of this study, we need to find out what are the significant variables that discriminate the two clusters in the best way. In other words, we need to look at if the distribution of values in the two clusters is significantly different or not. To do this, we used Chi-squared test of independence for categorical variables. The Chi-squared test is not reliable if the expected frequency value in any of the cells of

contingency table is less than 5. In such cases, we used fisher exact test. We used t-test to validate if there is a significant difference in the number of family members, family income and the grades of the students falling in the two clusters.

2.6. Tools used

R and RStudio were used to implement the various clustering algorithm and for applying statistical tests. R and RStudio are widely used open-source platforms for statistical computing and data analytics. R provides a wide variety of inbuilt functions for hypothesis testing based on classical statistical tests, linear and non-linear predictive modelling, and clustering analysis. We used ‘clValid’ package for carrying out the cluster validation.

3. RESULTS AND DISCUSSION

This section describes the results of the cluster analysis. The selected clustering algorithm PAM divides the students into two clusters as suggested by cluster validation process. The clusters fomred uncovers valuable and interesting insights. Tables 6-8 give the intra and inter cluster frequency distributions of various variables for the clusters.

Table 6. Cluster analysis: intra-cluster frequency distribution of categorical variables

Attributes	Sub-attributes	Cluster 1		Cluster2	
		Frequency	Percentage (%)	Frequency	Percentage (%)
Gender	M	67	45	26	24
	F	83	55	81	76
Caste	GEN	138	92	5	5
	OBC	11	7	16	15
	SC	1	1	86	80
Residence	RL	131	66	68	64
	UR	19	48	21	36
Father_edu	ILTR	13	8.6	17	15.9
	PMRY	7	4.6	7	6.5
	MDDL	24	16	20	18.7
	MTRC	75	50	46	43
	SECD	20	13.3	13	12.1
	GRDTA	10	6.6	4	3.7
	PGDT	1	0.6	0	0
	AGR	113	75.3	3	2.8
Father_occp	EX	7	4.6	3	2.8
	LBR	1	0.6	85	79.4
	SEM	7	4.6	7	6.5
	SVR	22	14.7	9	8.4
	ILTR	60	40	42	39.3
Mother_edu	PMRY	18	12	23	21.5
	MDDL	21	14	24	22.4
	MTRC	41	27.3	14	13.1
	SECD	7	4.6	3	2.8
	GRDTA	3	2	1	0.9
	AGR	33	22	0	0
Mother_occp	AWADI	1	0.6	3	2.8
	EX	1	0.6	1	0.9
	HW	115	76.6	80	74.8
	LBR	0	0	23	21.5

Cluster analysis reveals that most of the students belonging to the general category (97%) fall in cluster 1 whereas all the students belonging to the schedule cast category (99%) except one fall in cluster 2. The OBC students are distributed in both the clusters, i.e., 41% in cluster 1 and 59% in cluster 2. This shows that more of the OBC students are socio-economically closer to the SC students.

Henceforth, cluster one is dominated by the students of general category and the second cluster contains students of SC category. The major proportion (72%) of the girl students fall in cluster one. This means that girls from SC category are underrepresented in the school. Fathers belonging to general category appear to be slightly more educated than the fathers belonging to the SC category. There exist 91 % and 84 % literate fathers in the general and SC category clusters respectively which is above the male literacy rate of 80 % of Jind district. Fathers belonging to the first cluster are involved in agriculture whereas in the second cluster, fathers are labourers. Although the proportions of illiterate mothers in the clusters are similar, the proportion of educated mothers of general category students increases beyond middle school. There are 60 % literate

mothers within both the clusters. The literacy rate for the mothers is only at par with women literacy rate of Jind district. The study indicates that educated fathers are availing the opportunity to send their children to JNV. There is a clear distinction between the occupation of SC and non-SC mothers. SC mothers are labourers and there is lesser proportion of housewives in this category.

Intra-cluster five number summaries for the number of family members, family income and grades obtained by the students are given in the Table 7. The five number summaries for number of family members and grades do not vary a lot between the two groups of SC and non-SC students. The first quartile, median and third quartile of income is lower for the second cluster dominated by SC population of students. The maximum income in both the clusters is far away from the median income. This is due to the presence of few fathers/mothers in government services.

Table 7. Cluster analysis: intra-cluster frequency distribution of numerical variables

Attributes	Cluster 1	Cluster 2
Income	Min. Q1 Median. Q3. Max 1000 15000 24500 48000 240000	Min. Q1 Median. Q3. Max 3600 10000 15477 30000 300000
Family members	Min. Q1 Median. Q3. Max 1 4 5 6 14	Min. Q1 Median. Q3. Max 3 5 5 6 9
Grades	Min. Q1 Median. Q3. Max 6 8 8 9 10	Min. Q1 Median. Q3. Max 5 8 8 9 10

Table 8. Cluster analysis: inter-cluster frequency distribution of variables

Attributes	Sub-attributes	Cluster 1		Cluster2	
		Frequency	Percentage (%)	Frequency	Percentage (%)
Gender	M	67	72	26	28
	F	83	51	81	49
Caste	GEN	138	97	5	3
	OBC	11	41	16	59
	SC	1	1	86	99
Residence	RL	131	66	68	34
	UR	19	48	21	52
Father_edu	ILTR	13	43	17	57
	PMRY	7	50	7	50
	MDDL	24	55	20	45
	MTRC	75	62	46	38
	SECD	20	61	13	39
	GRDTA	10	71	4	29
	PGDT	1	100	0	0
	Father_occup	AGR	113	97	3
	EX	7	70	3	30
	LBR	1	1	85	99
	SEM	7	50	7	50
	SVR	22	71	9	29
Mother_edu	ILTR	60	59	42	41
	PMRY	18	44	23	56
	MDDL	21	47	24	53
	MTRC	41	63	14	37
	SECD	7	70	3	30
	GRDTA	3	75	1	25
Mother_occup	AGR	33	100	0	0
	AWADI	1	25	3	75
	EX	1	50	1	50
	HW	115	59	80	41
	LBR	0	0	23	100

Further, we have statistically validated the attributes that play a significant role in discriminating the two clusters. Pearson's chi-squared test of independence and fisher exact test are applied for categorical variables. For numeric and integer variables, t test is applied to validate if there exists a significant difference between the mean values of the variables for the two clusters. Table 9 shows the test statistics for the various variables. It shows that gender, caste, father occupation, mother education, and mother occupation and family income play significant role in differentiating the two clusters. It is worth noting that there is no significant difference in the fathers' education in the two clusters but there is significant difference in their occupation.

Table 9. Statistical test results

Attribute name	Statistical test applied	X-Squared/t value	df value	p-value
Gender	Chi-squared	10.35	1	0.0013
Caste	Chi-squared	208.5	3	<2.2e-16
Residence	Chi-squared	1.803	1	0.1794
Father education	Fisher exact test	-	-	0.44
Father occupation	Fisher exact test	-	-	<2.2e-16
Mother education	Fisher exact test	-	-	0.019
Mother occupation	Fisher exact test	-	-	2.622e-16
Family members	t-test	1.0495	250.36	0.295
Family income	t-test	2.6876	247.98	0.003842
grades	t-test	1.6316	235.31	0.05205*

Further, it is clear from the Table 7 that the family income of the students belonging to the general category is significantly higher. The attribute ‘family member’ has no significant role in the cluster formation. Moreover, no significant difference is found in the academic performance of students falling in the two clusters at a significance level of 0.05. With the above findings, we can characterize the two clusters based on socio economic status of the students as shown in Figure 1.

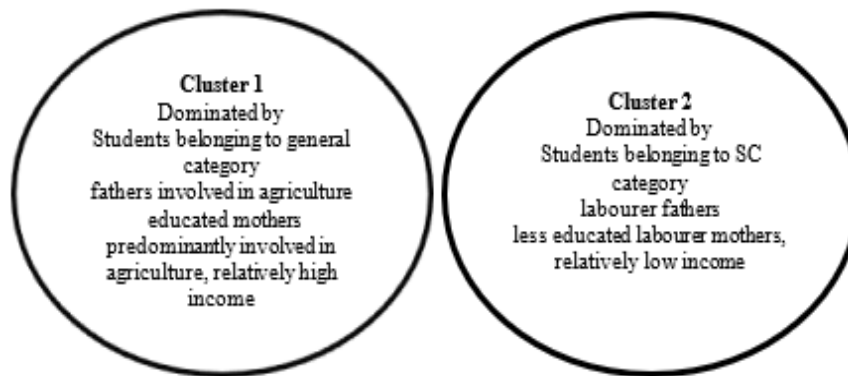


Figure 1. Characterizing clusters of the students

Here follows the discussion on the results. We set out to examine the two research questions by applying clustering analysis:

- Does there exist a significant difference in the academic performance of the groups of JNV students belonging to differential socio-economic status?
- Which socio-economic indicators are significantly different between the groups?

In our study, the null hypothesis states that there exists no significant difference in the academic performance of the students’ groups based on socio-economic status. The null hypothesis could not be rejected at a significance level of 5%, which is inconsistent with the inference drawn from many of the existing studies which shows a correlation between socio-economic status and academic performance. It has been noticed that ethnic minorities from low socio-economic status are consistently disadvantaged than other category in terms of educational outcomes [5], [9], [29]. Bhagavatheeswaran [10] also identified the caste as a barrier for the entry and retention of the SC/ST girls in secondary education. However, the result of this study depict that the academic performance of the students belong to SC and Non-SC group is significantly different only at 10 % significance level. This indicates that the JNV school may have helped SC students to enhance their academic performance, however, additional support is required to further boost the academic performance of the students belonging to the lower SES.

Regarding second question- caste, educational and occupational status of family, and family income indicators divide the students at JNV into two separate groups, i.e., students belonging to the general category and students belonging to the SC category. The results also lend support to [30] study that participation of parents in educational processes not only directly impacts the academic performance, but also indirectly influences the learning attitudes and behaviours of their children. Another important dimension is that there are significantly less females compared to males in the SC group.

4. CONCLUSION AND FUTURE SCOPE

Many studies have concluded that SES has bearing on student academic performance. The adverse socioeconomic circumstances have detrimental effect on the academic attainment of students. In this paper, we have carried out clustering analysis of the JNV students' data. The clustering algorithm divided the students into two groups based on differential socio-economic factors. Majority of general category students (from better-off family backgrounds) and schedule caste students (from lower status family backgrounds) fall in different clusters. Statistical tests showed significant difference in the socio-economic status of the two groups. However, no significant difference was noticed in the academic performance at 5% significance level. According to this study, the JNV school has been able to cover up the gap in academic performance of the students, who come from marginal sections of society to some extent. The residential environment and financial support by the government may have proved to be beneficial in bridging the gap. We have worked on data sample from a single school, and the result is significant. Since the academic structure, reservation policy and financial support of the JNV schools is same all over India, it is likely that findings for the other JNV school may be similar. However, it would be fascinating to re-test the study with the 12th class result of the same students. We will also try to repeat the study for larger samples of JNV schools at zonal and national level. Such studies will permit to explore the regional differences, if any. It would also be interesting to perform a similar study on other schools affiliated to CBSE and compare it to the one carried out here. These studies will prove to be beneficial for making policies to bridge the gap in the educational attainment of deprived sections of society.




REFERENCES

- [1] M. Broer, Y. Bai, and F. Fonseca, "A review of the literature on socioeconomic status and educational achievement," in *IEA Research for Education*, vol. 5, 2019, pp. 7–17.
- [2] K. Devi, S. Ratnoo, and A. Bajaj, "Impact of socio-economic factors on students' academic performance: A case study of Jawahar Navodaya Vidyalaya," in *Lecture Notes in Networks and Systems*, vol. 419 LNNS, 2022, pp. 774–785.
- [3] S. R. Sirin, "Socioeconomic status and academic achievement: A meta-analytic review of research," *Review of Educational Research*, vol. 75, no. 3, pp. 417–453, Sep. 2005, doi: 10.3102/00346543075003417.
- [4] S. Thomson, "Achievement at school and socioeconomic background—an educational perspective," *npj Science of Learning*, vol. 3, no. 1, p. 5, Mar. 2018, doi: 10.1038/s41539-018-0022-0.
- [5] G. Considine and G. Zappalà, "The influence of social and economic disadvantage in the academic performance of school students in Australia," *Journal of Sociology*, vol. 38, no. 2, pp. 129–148, Jun. 2002, doi: 10.1177/144078302128756543.
- [6] C. K. Malecki and M. K. Demaray, "Social support as a buffer in the relationship between socioeconomic status and academic performance," *School Psychology Quarterly*, vol. 21, no. 4, pp. 375–395, 2006, doi: 10.1037/h0084129.
- [7] M. R. Islam and Z. N. Khan, "Impact of socio-economic status on academic achievement among the senior secondary school students," *Educational Quest- An International Journal of Education and Applied Social Sciences*, vol. 8, no. 3, p. 643, 2017, doi: 10.5958/2230-7311.2017.00117.9.
- [8] S. Li, Q. Xu, and R. Xia, "Relationship between SES and academic achievement of junior high school students in China: The mediating effect of self-concept," *Frontiers in Psychology*, vol. 10, Jan. 2020, doi: 10.3389/fpsyg.2019.02513.
- [9] C. P. S. Chauhan, "Education and caste in India," *Asia Pacific Journal of Education*, vol. 28, no. 3, pp. 217–234, Sep. 2008, doi: 10.1080/02188790802267332.
- [10] L. Bhagavatheswaran *et al.*, "The barriers and enablers to education among scheduled caste and scheduled tribe adolescent girls in northern Karnataka, South India: A qualitative study," *International Journal of Educational Development*, vol. 49, pp. 262–270, Jul. 2016, doi: 10.1016/j.ijedudev.2016.04.004.
- [11] R. K. John, B. Xavier, A. H. Waldmeier, A. H. Meyer, and J. Gaab, "The governmental ranking of class and the academic performance of Indian adolescents," *PLoS ONE*, vol. 15, no. 11 November, p. e0241483, Nov. 2020, doi: 10.1371/journal.pone.0241483.
- [12] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.
- [13] T. Thiele, A. Singleton, D. Pope, and D. Stanistreet, "Predicting students' academic performance based on school and socio-demographic characteristics," *Studies in Higher Education*, vol. 41, no. 8, pp. 1424–1446, Aug. 2016, doi: 10.1080/03075079.2014.974528.
- [14] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 3, Dec. 2020, doi: 10.1186/s41239-020-0177-7.
- [15] F. Jauhari and A. A. Supianto, "Building student's performance decision tree classifier using boosting algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1298–1304, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1298-1304.
- [16] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.
- [17] A. Dutt, "Clustering algorithms applied in educational data mining," *International Journal of Information and Electronics Engineering*, 2015, doi: 10.7763/ijee.2015.v5.513.
- [18] J. Ha, M. Kambe, and J. Pe, *Data Mining*. Elsevier, 2012.
- [19] D. Kerr and G. K. W. K. Chung, "Identifying key features of student performance in educational video games and simulations through cluster analysis," *JEDM - Journal of Educational Data Mining*, vol. 4, no. 1, pp. 144–182, 2012, [Online]. Available: <http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/25>.
- [20] M. Khalil and M. Ebner, "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories," *Journal of Computing in Higher Education*, vol. 29, no. 1, pp. 114–132, Apr. 2017, doi: 10.1007/s12528-016-9126-9.




- [21] W. Prachuabsupakij and S. Chiengpongpan, "Cluster analysis of personal data towards student's graduation in information technology program," *ACM International Conference Proceeding Series*, pp. 76–80, 2020, doi: 10.1145/3396743.3396792.
- [22] O. T. Omolewa, A. T. Oladele, A. A. Adeyinka, and O. R. Oluwaseun, "Prediction of student's academic performance using k-means clustering and multiple linear regressions," *Journal of Engineering and Applied Sciences*, vol. 14, no. 22, pp. 8254–8260, 2019.
- [23] S. Wijayanti, Azahari, and R. Andrea, "K-Means cluster analysis for students graduation (case study: STMIK widya cipta dharma)," in *ACM International Conference Proceeding Series*, Jun. 2017, vol. Part F129684, pp. 20–23, doi: 10.1145/3108421.3108430.
- [24] S. Bharara, S. Sabitha, and A. Bansal, "Application of learning analytics using clustering data Mining for Students' disposition analysis," *Education and Information Technologies*, vol. 23, no. 2, pp. 957–984, Mar. 2018, doi: 10.1007/s10639-017-9645-7.
- [25] N. R. Shamsuddin and N. I. Mahat, "Comparison between k-Means and k-Medoids for mixed variables clustering," in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, Singapore: Springer Singapore, 2019, pp. 303–308.
- [26] N. Aayog, "Evaluation Study on Navodaya Vidyalaya Smiti (NVS)," Programme Evaluation Organisation, Government of India, 2015.
- [27] Registrar General & Census Commissioner of India, "Census of India 2011 Provisional Population Totals Paper 1 of 2011 : Goa," 2011. [Online]. Available: http://www.censusindia.gov.in/2011-prov-results/prov_data_products_goa.html.
- [28] C. Fraley and Adrian E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 1, pp. 611–631, 2002, [Online]. Available: <http://www.jstor.org/stable/3085676>.
- [29] N. Berger and J. Archer, "School socio-economic status and student socio-academic achievement goals in upper secondary contexts," *Social Psychology of Education*, vol. 19, no. 1, pp. 175–194, Mar. 2016, doi: 10.1007/s11218-015-9324-8.
- [30] Z. Li and Z. Qiu, "How does family background affect children's educational achievement? Evidence from Contemporary China," *Journal of Chinese Sociology*, vol. 5, no. 1, p. 13, Dec. 2018, doi: 10.1186/s40711-018-0083-8.

BIOGRAPHIES OF AUTHORS



Kapila Devi    is an assistant professor of computer science in the directorate of distance education, Guru Jambheshwar University of Science and Technology, Hisar, India. She is also a research scholar in the department of computer science and engineering at the same university. She has approximately ten years of teaching experience. Her research and teaching interest include data structure, artificial intelligence, data mining, and machine learning. She can be contacted at email: ckapila628@gjst.org.



Saroj Ratnoo    received M.Sc. in computing science from Birkbeck College, University of London, UK in 1994. She joined the department of computer science and engineering, (GJUS&T), Hisar in 1996 as an assistant professor. She completed her doctorate degree from the school of computer and systems sciences, Jawaharlal Nehru University, New Delhi, India in 2010. She continues to work in the department of computer science and engineering, (GJUS&T), Hisar, India as professor. Her research interests include nature-inspired algorithms, data mining, and machine learning. She has published more than 30 papers. She can be contacted at email: ratnoo.saroj@gmail.com.