Analysis of Educational attainment of Students using Anova Test

[1] Dr. HemaMalini B H, [2]Dr. L Suresh, [3] Dr. Suma V, [4]Shankar M. M.

[1] *Dept of CSE, BMS Institute of Technology and Management, Bengaluru, INDIA*

[2] *Cambridge Institute of Technology, Bengaluru, INDIA*

[3] *Dayananda Sagar College of Engineering, Bengaluru, INDIA*

[4]*CARES, Bangalore, Bengaluru, INDIA*

ABSTRACT

Educational Data Mining (EDM) is a booming trend. It shows the various techniques and strategies that affect the performance of students. For the present research work, the data is taken from the engineering students of a reputed institution in Karnataka. The dataset consists of the details collected from the institutional repository. One of the statistical technique- Anova test is applied on the pre-processed data from the dataset, consisting of 1186 records. The two tests- namely descriptive F value and T-test is applied. The calculation contains the probability distribution too. The results obtained show that the father's occupation affects the performance of the student. A significant contribution can be made to society by predicting the parameters that affect the performance of students thus improving the results. However, the current research shows that, mother's occupation does not affect significantly.

Keywords: Anova, Descriptive Statistics, Multiple comparisons, f-value, variance, degree of freedom, p-value, significant, T test.

INTRODUCTION

Educational Data Mining (EDM) is an upcoming trend in the field of Data Mining. In present days, in the era of ICT, where all the stakeholders are using the technology, to solve their day-to-day problems, the educational institutions are also not left behind in the race. To sustain in the business market, they must assess the performance of the students, who are their stake holders. The institutions must check the factors affecting the students' performance at an earlier stage.

In view of this, a research has been taken up in the field of Educational Data Mining (EDM). The required data for research is collected from the repository of a reputed institution in Karnataka affiliated to a premiere university. The raw data thus collected is pre-processed. The pre-processed data is used for analysis. An effort is made to identify the parameters that affect the performance of the students the most.

Rohit Ahlawat et. al have researched on the enrolment trends in Indian academia and the features that affect the trends using k-means clustering method. This technique was applied on the dataset of the company and the crime dataset. The main objective of their work was to investigate the total enrolment, enrolment of foreign national students, community-wise enrolment, and enrolment due to crime rate. The inferences drawn were in north eastern states, the enrolment for higher education was very less, and it was high in south Indian states like

Kerala, Karnataka, Tamilnad, Andhra Pradesh and Maharashtra. Reasons could be many like good infrastructure, population, presence of good companies and crime rate [2].

Ting Zeng has developed a five-side educational data mining structure (5S-EDMF) to investigate the performance of learners of college and to recommend the learning resources accordingly. During data collection, the researchers observed about the approaches and behavior of the students online and offline. The analysis gave them a supplementary tool to reach and support students. The model developed was field-tested. After investigating the collected statistics, the researchers determined that their structure provided an effective learning atmosphere; they concluded their model was a good predictor for performance of students [5].

P. Nithya et. al have done an extensive survey on Educational Data Mining (EDM). The authors have discussed the various goals of EDM, the methods of Educational Data Mining. Also, the applications of EDM are discussed which includes Analysis and Visualization of Data, Enrolment Management, Forecasting Students Outlining, Forecasting Student Performance, Combining Students, Planning and scheduling, Organization of Syllabus and Identifying Cheating in Online Exams [6].

Lot of work has happened in educational data mining to predict the performance of undergraduates and improve their results. Many methods have been proposed by the researchers. The detailed study is shown.

LITERATURE REVIEW

Asraful Alam Pathan et. al, present a mining model grounded on decision tree (DT) for assessing the students C programming skills. In supervised learning, DT is a rule set. It is data mining algorithms that uses top-down recursive technique with divide and conquer [1]. The researchers have composed data from 70 undergraduates of Structured Programming Language (SLP) course. They have produced two datasets, consisting of behavioural and previous educational attributes of the student. Using the dataset, two decision trees are constructed which could categorise the students as three categories-Good, Mediocre and Poor. The intention is to concentrate on the poor performers. The decision tree model proposed could appropriately classify 87% students.

Kittinan Limsathitwong et. al have developed a Dropout Prediction System (DPS) using Cross-Industry Standard Process for Data Mining (CRISP-DM) [3]. The student drops out had a major impact on the education of students in Thailand. The researchers have given an extensive investigation on the methods that could be used on the dataset. The main goal was to reduce the dropout rate in Thailand. Their present work was a case study of the students pursuing their first year and second year of graduation. The prediction prototypes were established to apply the Decision Tree algorithms.

Musa Wakil Bara et. al have done an investigation to study the presentation of Self-Organizing Map Clustering Method to investigate the e-learning accomplishments of the students to classify the clusters of students [4]. The data file obtained shows their activities from the log file generated. Using the SOM technique, the researchers created three clusters and the learning behavior of students were analyzed in each of the cluster. The relation between the learning behavior and their academic performance was examined. The outcome concluded

that, students who are in both ‘High Learners’ and ‘Very Active’ category (Cluster1) appeared as the best in the success of their course.

Amjad Abu Saa have tried to determine relationships between personal and social factors of the students, their performance in education in earlier semester by means of data mining techniques. The data collection was done through survey conducted in regular classes and using Google Forms [7]. The dataset consisted of 270 records. On the data so collected, four decision tree algorithms were applied. Naïve Bayes algorithm was also applied, and inferences were drawn. It was concluded that the performance of undergraduates not only depend on educational background but there were various social aspects affecting their performance.

A. Dinesh Kumar et. al have studied the prediction techniques in data mining in detail [8]. The paper reviews the data mining tools and prediction algorithms used in educational data mining. It also projects at identifying better prediction algorithms and new data mining tools to be applied to foresee the performance of undergraduates. This in turn would help the instructor and institution to increase their education level.

Hilal Almarabeh et. al have studied the performance of undergraduates by using various data mining skills and using WEKA tool [9]. Five different classifiers were used namely Bayesian Network, NaiveBayes, J48, ID3 and Neural Network. The data set comprised of 225 occurrences and also10 attributes. It was concluded that Bayesian network performed better than the other four classifiers. It was concluded that Bayesian Network gave the best accuracy.

D. Fatima et. al have taken up a survey on EDM. The topics like student retention and attrition, individual recommender schemes within education were studied. Also, an attempt is made to study the ways to use data mining for course management system data analysis. In present literature, the gaps were identified and prospects for additional investigation were offered [10].

Pooja Thakar et. al have done an extensive survey on educational data mining and have presented a paper covering the work done during 2002 to 2014 in the field of educational data mining and its opportunity in upcoming days [11].

Wongpanya Nuankaew et. al investigated to determine the concealed academic association between educational attainment and higher educational organizations. Their paper offered a method which merged two methods namely decision tree and attribute selection to recognize the sets of attributes. The intention was to choose the prominent attributes that might affect the academic performance. Various data mining techniques were applied to investigate the association between the academic attainment and higher educational organizations in Thailand. For the study, the dataset was taken from seven schools in Thailand. Their present paper proposed a elementary theoretical model with decision tree technique [12].

Sadiq Hussain et. al gathered data from Assam in India from three colleges. The data consisted of socio-economic, demographic and academic statistics of 300 undergraduates with 24 attributes. Four classification techniques namely the J48, PART, Random Forest and Bayes Network Classifiers were applied. WEKA tool was the data mining tool adapted for implementation. The accuracy and classification inaccuracies were studied, and it was concluded that the Random Forest Classification technique was the suitable algorithm for the

given dataset. Also, it was observed that using WEKA, the Apriori algorithm was used to the data set to discover certain finest rules [13].

Ajay Kumar Pal et. al have investigated on BCA students data sets. Their works define the usage of data mining methods to increase the effectiveness of academic performance in the educational establishments. The researchers have presented a real-world investigation piloted in VBS Purvanchal University in India. Their technique has facilitated to recognize the students who require exceptional guiding by the educator which provides education with high quality [14].

Arpit Bansal et. al have suggested a method for amendment in K-Means Clustering Algorithm. The researchers have determined that large datasets are divided into small data collections that are called clusters by a method known as clustering. In the proposed technique, the K-Means clustering eliminates the two major disadvantages of K-Means clustering, which are improving the precision level and the calculation interval spent in grouping the dataset. For small datasets, the calculation time and accuracy level might not make a big difference, but when huge dataset is considered with trillions of records, slight scattering in precision level makes a big difference and could lead to a dreadful situation, if not controlled appropriately [15].

HemaMalini B. H et. al have applied statistical methods and machine learning techniques to assess the performance of students and check the parameters that affect their performance the most [17, 18, 19].

RESEARCH METHODOLOGY

In the related works considered so far, to study the performance of students, the scholars have made efforts to apply machine learning algorithms to obtain the predictive models. In the present work an effort is made using the statistical method to assess the student performance. The Anova technique is applied on the given data set. An attempt is made to discover the variable that contribute the most for student performance.

The data collection is performed from a premiere engineering institution in South India. The institution has automated most of its administrative works. At the time of admission of students, the institution collects the preliminary information like Pre university marks, education of parents, occupation, annual income etc. in the institutional repository. Hence, the admission data could be obtained from the institutional repository. The data was collected as an Excel file in .csv format. Initially five years admission data was collected from the institutional repository. Later, the data was filtered. The data collection process concentrated on admission related data of the students from two batches 2017-18 and 2018-19 academic years. It is shown in Fig 1.

The students chosen for the study are from the core branches of engineering like Computer Science and Engineering, Mechanical Engineering, Telecommunication Engineering, Information Science and Engineering, Civil Engineering, Electrical and Electronics Engineering and Electronics and Communication Engineering.

The data so collected must be pre-processed. Out of this data, only the recent two batches data had to be filtered and collected. The senior student data was removed during pre-processing. The study concentrated on recent two batches of students. Initially around 3000

records were collected. After filtering the left-out students and senior students, the final data set was reduced to 1186 records. The students who might have opted change of college during second and third rounds, the students who might have left the college were also removed. The reason is that their university marks would not be available for these students. They would not appear for examinations from the institution where they have already opted a change. The students who did not appear for university examinations also were eliminated from the dataset, since outcome variable is mandatory. Since the university results are used as an outcome variable, the results are necessary to measure the performance of the students.

CASE STUDY

The data for investigation is collected from a prestigious engineering college in Karnataka affiliated to a popular university.

The good ranking students seek admission in this institution. The major admissions to this institution happen through the CET (Common Entrance Test), COMED-K (Engineering and Dental Colleges of Karnataka undergraduate Entrance Test) and management quotas. The CET rank is based on the score on 12th Grade also. The CET rank is announced taking into consideration the PUC (Pre-University marks) also. 50% weightage is given to PUC marks and 50% weightage is given to CET score. The CET rank is considered for seat allotment in engineering stream. These are the government seats.

The institution offers engineering course in the following disciplines: Computer Science and Engineering (CSE), Electronics & Communication Engineering (ECE), Information Science & Engineering (ISE), Telecommunication Engineering (TCE), Electrical and Electronics Engineering (EEE), Mechanical Engineering (ME) and Civil Engineering (CIV).

The engineering course is deliberated by the university for four years (a total of eight semesters). Semester schemes are executed by all colleges under this university. Each academic year will have Odd and Even semesters.

After seat allotment process is completed, the students get admission in the college opted. At the time of admission, admission number is assigned to every student. This is a unique number for every student. After the completion of admission process, the student credentials are submitted to the university through student resume in a specified format given by the university. University assigns a unique USN (University Seat Number) to every candidate. This USN is used throughout the four years of engineering for all academic purposes. Till the generation of USN, the institution uses admission number to identify the students. In the institutional repository, the admission number is saved. This admission number is taken as primary key in the present study till the generation of USN. The dataset was prepared before the generation of USN. So, the USN is mapped with the admission number. After receiving the USN, it becomes the primary key. The files with and without USNs were merged to form a resultant data file. The admission number is local to the institution. The USN is global in the University. The pre-processed data file with all the input variables is ready for processing.

The outcome variable would be the university results. After preparing the input file, the next process was to wait for the university results. After the announcements of results on the university website, all the subject marks were downloaded and gross marks of every candidate was calculated and used as an outcome variable. The gross marks are used as outcome variable for the analysis.

Data filtering was done even after getting the university results. If the gross marks of all 6 subjects were a single digit marks or zero, it was assumed as false data. These entries were filtered off from the pre-processed data file. Some students had admission number but were without an USN. It was taken as change of college case and such entries were filtered off. The data file was saved as .CSV file.

R Programming language is used for implementation.

The figure 1 shows the sample data file. It consists of the following attributes: Branch, Blood Group, Gender, Reservation, Annual Income, Selection, Pre_University_Marks, Pre-university_Board, Lodging, father's profession, mother's profession, Pre_University Percentage and Pre_University_Results.

	A	D	F	G	H	I	J	K	L	M	N	O	P
1	Branch	Selection	Gender	BloodGroup	Reservation	Board_PUC	Marks_PUC	Accomodation	Fr.Occupation	MR.Occupation	AnnualIncome	Percent_PUC	Results
2	CSE	cet	M	O+	GM	State	298	Day Scholar	PRIVATE	Professionals	6000000	90.03	550
3	CSE	cet	M	O+	GM	State	298	Day Scholar	PRIVATE	Professionals	6000000	90.03	550
4	CSE	comedk	M	O+	GM	State	270	Day Scholar	PRIVATE	PRIVATE	1500000	88	556
5	CSE	comedk	M	O+	GM	CBSE	280	Hostel	BUSINESS	HomeMaker	400000	86.6	527
6	CSE	cet	M	AB+	SC/ST	State	212	Day Scholar	BUSINESS	PRIVATE	80000	75.8	444
7	MECH	management	M	O+	OBC	State	193	Day Scholar	BUSINESS	HomeMaker	600000	76.5	526
8	CSE	cet	M	O+	GM	State	280	Day Scholar	PRIVATE	HomeMaker	800000	90	526
9	CSE	cet	F	O+	GM	State	288	Hostel	PRIVATE	PRIVATE	800000	95	567
10	CSE	cet	M	O+	GM	State	274	Day Scholar	PRIVATE	GOVT SERVICE	3500000	83.33	571
11	CSE	management	M	O+	GM	CBSE	214	Hostel	GOVT SERVICE	BUSINESS	1000000	76.2	416
12	CSE	cet	M	B+	GM	CBSE	281	Day Scholar	PRIVATE	HomeMaker	700000	93	545
13	CSE	cet	F	B+	OBC	State	290	Day Scholar	BUSINESS	HomeMaker	480000	93.5	593
14	CSE	comedk	M	A+	OBC	State	259	Hostel	BUSINESS	HomeMaker	200000	83.6	423
15	CSE	management	F	A+	J&K	State	255	Day Scholar	BUSINESS	HomeMaker	200000	81.2	544
16	CSE	comedk	F	O+	GM	State	298	Hostel	PRIVATE	HomeMaker	600000	97.7	558
17	CSE	management	F	O+	GM	State	266	Day Scholar	BUSINESS	PRIVATE	600000	88.66	442
18	CSE	comedk	M	AB+	GM	CBSE	277	Day Scholar	PRIVATE	HomeMaker	1000000	92.33	560
19	CSE	comedk	M	A+	GM	CBSE	215	Day Scholar	Professionals	Professionals	1000000	79.8	560
20	CSE	comedk	M	O+	OBC	State	183	Day Scholar	PRIVATE	HomeMaker	518228	61.33	374
21	CSE	comedk	M	O+	GM	CBSE	259	Day Scholar	PRIVATE	GOVT SERVICE	5000000	89.2	463
22	CSE	comedk	M	O-	OBC	CBSE	266	Hostel	PRIVATE	HomeMaker	500000	88.67	508
23	CSE	management	F	A+	GM	State	249	Day Scholar	BUSINESS	HomeMaker	600000	86.16	550

CleanedDataWithAllResultsNew

Figure 1: Sample Data File

ALGORITHM

Phase 1: Raw data is collected from the institutional repository.
Phase 2: Data cleaning is done by removing the students from higher semesters and gathering data of 2017-18 batch students and 2018-19 batch students.
Phase 3: Obtaining the unique students admission number from institutional repository.

Phase 4: The admission number is mapped to the USN.

Phase 5: The university website provide the university results in specific time.

Phase 6: Apply Anova technique on the available dataset.

Phase 7: Tabulate all results.

Phase 8: Generate results, investigate the influence of different input variables on the performance of undergraduates.

Descriptive Statistics, ANOVA and Multiple comparison analysis

"Anova is a statistical technique applied to examine if the means of two or more groups are substantially different from one another". This technique is preferred for analysis in this work since out of two groups of datasets, one group has categorical value and other group has

continuous value. “In statistics, a categorical variable is one that takes on a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property”. For example – In the given dataset, Branch, Selection, Reservation, Accommodation, Board_PUC, Mother Occupation Father Occupation, are all the categorical values. “A continuous distribution is one in which data can take on any value within a specified range”. In the current data set, Percent_PUC and Results are the continuous values.

Here, p-value is Probability value. To test the significant difference amongst these branches, Anova test is applied. To check whether Anova is significant or not p value has to be referred.

f value indicates Fischer ratio which shows the difference between explained variance and unexplained variance. When Anova is more than 3 or 4, p value should be less than 0.05. To claim the test is significant, p value should be <0.05.

Table 1: Descriptive - Branch

	Mean score
CIV	5.9
CSE	6.8
ECE	7.0
EEE	6.7
ISE	6.8
MECH	6.1
TCE	6.1

CIV – Civil Engineering
CSE – Computer Science & Engineering
ECE – Electronic & Communication Engineering
EEE – Electrical & Electronics Engineering
ISE – Information Science & Engineering
ME – Mechanical Engineering
TCE - Telecommunication Engineering

Inference1:

Table 1 shows the descriptive branch analysis. From the tabulated mean score, it is evident that on an average, ECE branch stands the highest, second highest is CSE and ISE branches and the lowest being MECH and TCE branches. Result using the CGPA showed ECE, CSE and ISE students have obtained higher score on average, significant difference is established using one-way Anova test, f=15, degree of freedom, df (6,1129), P<0.001. Since p value is <0.01, the difference between branch on result CGPA is statistically significant.

Table 2: Descriptive – Selection

	Mean score
cet	7.2
comedk	7.1
management	5.8
pio	5.6

cet – Common Entrance Test

comedk - Consortium of Medical, Engineering and Dental Colleges of Karnataka

pio – People of Indian Origin

Inference 2:

Table 2 shows that students from CET secured highest result CGPA while compared to others on an average. Significant difference is established using one way anova test, f=125, df (3, 1132), P<0.001. Since p value is <0.01, the difference between selection criteria on result CGPA is statistically significant.

Table 3 shows that the F value is higher than 3. So, Anova is significant.

Table 3: ANOVA

term	df	sumsq	meansq	F value	P value
Branch	6	149	24.9	15	0.000
Residuals	1129	1837	1.6		

df – Degree of freedom: It is a statistical computation to say the technique belongs to which family.
sumsq – sumsquare
meansq – meansquare

Table 4: Multiple comparison by Branch

	diff	lwr	upr	p adj
CSE-CIV	0.91	0.46	1.35	0.00
ECE-CIV	1.17	0.71	1.63	0.00
EEE-CIV	0.85	0.30	1.39	0.00
ISE-CIV	0.90	0.42	1.38	0.00
MECH-CSE	-0.69	-1.10	-0.27	0.00
TCE-CSE	-0.66	-1.27	-0.05	0.03
MECH-ECE	-0.95	-1.38	-0.52	0.00
TCE-ECE	-0.92	-1.54	-0.30	0.00
MECH-EEE	-0.63	-1.15	-0.11	0.01
MECH-ISE	-0.68	-1.13	-0.23	0.00
TCE-ISE	-0.65	-1.29	-0.02	0.04

diff – difference
lwr – lower
upr - upper

Inference 3:

Table 4 shows the multiple comparisons by Branch. Here, ignore all negative values and values lower than 3. Anova bothers about the significant difference. Once it is clear that Anova is significant, if the combination is significant has to be checked. Here ECE-CIV, the difference is 1.17. It is a significant difference. Any p<0.05, it is significant.

Table 5: ANOVA by Selection

	df	sumsq	meansq	F value	P value
Selection	3	495	165.0	125	.000
Residuals	1132	1491	1.3		

df – Degree of freedom:
sumsq – sumsquare
meansq – meansquare

Table 5 shows the selection category. It is clearly seen that F value is higher than 3. So, Anova is significant.

Table 6: Multiple comparison by Selection

	diff	lwr	upr	p adj
management-cet	-1.36	-1.6	-1.15	0.00
pio-cet	-1.54	-1.9	-1.13	0.00
management-comedk	-1.29	-1.5	-1.04	0.00
pio-comedk	-1.47	-1.9	-1.04	0.00

cet – Common Entrance Test
comedk - Consortium of Medical, Engineering and Dental Colleges of Karnataka
pio – People of Indian Origin

diff – difference
lwr – lower
upr - upper

In Table 6, using multiple comparison by selection, all combinations are significant since $p<0.05$. In the lwr and upr, if both values are either positive or negative, then Anova is significant. But, if one is positive and another is negative, then, Anova is not significant.

Descriptive F value and T test

Table 7: ANOVA

Mean diff	Female	Male	T value	P value
0.63	7.1	6.5	8.2	0.00

Table 7 shows that since there are only two groups, Male and Female, T test is applied. If more than two groups were present, Anova test would have been applied. Female secured more CGPA than Male on average, $t=8.2$, $P<0.01$, and it is statistically, a significant difference.

Inference 4:

Female students secured more CGPA than Male students on average.

Table 8: Descriptive - Board PUC

	Mean score
CBSE	6.8
ICSE	6.8
Others	5.4
State	6.6

CBSE- Central Board of Secondary Education

ICSE - Indian Certificate of Secondary Education

Inference 5:

Table 8 indicates that the students who have come from ICSE and CBSE have the highest mean score. It means, highest number of students has come from ICSE and CBSE boards are

performing better. Remaining students are from state boards and other boards. Students who have come from others boards have the lowest mean score.

Table 9: ANOVA by Board PUC

term	df	sumsq	meansq	F value	P value
Board_PUC	3	28	9.3	5.4	0.00
Residuals	1132	1958	1.7		

df – Degree of freedom:
sumsq – sumsquare
meansq – meansquare

Table 9 indicates that F value is higher. Also, it is shown that $p < 0.05$. So, p is significant. It means that there is a significant difference in the marks scored amongst students from different boards.

Table 10: Multiple comparison by Board_PUC

	diff	lwr	upr	p adj
Others-CBSE	-1.40	-2.44	-0.37	0.00
State-CBSE	-0.21	-0.43	0.02	0.08
Others-ICSE	-1.39	-2.58	-0.21	0.01
State-Others	1.20	0.17	2.23	0.01

diff – difference
lwr – lower
upr - upper

Table 10 indicates the multiple comparisons by the PUC board. The following combinations are significant since $p < 0.05$: Others-CBSE, Others-ICSE, State-Others. It means that when the combination of Others-CBSE, Others-ICSE, State-Others is taken, there is a notable difference in the marks scored.

Table 11: Descriptive Accommodation

	Mean score
Day Scholar	6.8
Hostel	6.5
PG	6.7

Inference 6:

Table 11 indicates that the maximum numbers of students in the dataset are day scholars, second stands the PG and rest are staying in hostel.

Table 12: ANOVA by Accommodation

Term	df	sumsq	meansq	F value	P value
Accommodation	2	17	8.4	4.9	0.01
Residuals	1133	1969	1.7		

df – Degree of freedom:
sumsq – sumsquare
meansq – meansquare

Hostel-Day Scholar meandiff= -0.26, lwr= -0.45 upr=-0.06 value= 0.01

Table 12 indicates that the p value is significant since $p < 0.05$. It means that there is a notable difference in scores of students residing in hostels, as day scholars and in paying guest accommodations.

Table 13: Descriptive – Father Occupation

	Mean score
Business	6.5
Defense	7.1
Farmer	6.6
Government Service	6.6
Others	6.6
Private	6.8
Professionals	6.8

Inference 7:

Table 13 shows that the students participated in the dataset, the highest number of the students are with their parents working in defense. Second highest mean score is for parents working in private and working as professionals. The remaining parents are either working in Government Service or farmers or other jobs.

Table 14: ANOVA

term	df	sumsq	meansq	F value	P value
Father Occupation	6	28	4.6	2.7	0.01
Residuals	1129	1958	1.7		

df – Degree of freedom:
sumsq – sumsquare
meansq – meansquare

Defense -BUSINESS meandiff= 0.6, lwr= 0.12 upr=0.67 pvalue= 0.03

Since the difference is not so significant, the F value and P value are not considered for residuals.

Table 14 shows the P value is significant. The occupation of the parent has an impact on the performance of the student.

ANOVA test was applied on the given dataset. The mean values of the attributes are computed. Multiple comparisons by Branch, Descriptive – Selection are obtained. ANOVA by Selection, Multiple comparisons by Selection are tabulated. Descriptive F value and T test are computed. Descriptive – Reservation, ANOVA by Reservation and Multiple comparisons by Reservation are also done. Results are tabulated. Descriptive - Board PUC, ANOVA by Board PUC and Multiple comparisons by Board_PUC values are tabulated. Descriptive Accommodation, ANOVA by Accommodation are obtained. Descriptive – Father Occupation is tabulated. The ANOVA values are computed and tabulated for F value and P value.

The following inferences are drawn from the study:

1. On an average, ECE branch stands the highest, second highest is CSE and ISE branches and the lowest being MECH and TCE branches. Result using the CGPA showed ECE, CSE and ISE students have obtained higher score on average.
2. Students from CET secured highest result CGPA while compared to others on an average.
3. Compared to all other branches, there is a significant difference between ECE-CIV branches.
4. Female students secured more CGPA than Male students on average.
5. The students who have come from ICSE and CBSE have the highest mean score. It means, highest number of students has come from ICSE and CBSE boards are performing better. Remaining students are from state boards and other boards.
6. The maximum numbers of students in the dataset are day scholars, second largest count resides in the Paying Guest (PG) and rest are staying in hostel.
7. The highest numbers of the students are with their parents working in defense. Second highest mean score is for parents working in private and working as professionals. The remaining parents are either working in Government Service or farmers or other jobs.

CONCLUSION

In the area of Educational Data Mining, assessing the parameters that affect the performance of any engineering student is very substantial. The engineering colleges need to predict the performance of students and need to know the parameters affecting their performance. So, an effort is made by applying the statistical technique to study the parameters which affect the performance of a student.

The work was carried out on huge collection of students' data. The data set consisted of 1186 collected from a premiere engineering institution of Karnataka, affiliated to a premiere university. It was concluded that mother's occupation does not contribute to outcome variables. But, indeed the performance of student depends on father's occupation. The student's

performance is excellent if father is in defense. The students from ICSE or CBSE background, perform better than the students from State board or any other boards.

REFERENCES

1. Asraful Alam Pathan, Mehedi Hasan, Md. Ferdous Ahmed, and Dewan Md. Farid, (2014), "Educational Data Mining: A Mining Model for Developing Students' Programming Skills", IEEE, 978-1-4799-6399-7. https://ieeexplore.ieee.org/document/7083552
2. Rohit Ahlawat, Sushil Sahay, Sai Sabitha, Abhay Bansal, (2016), "Analysis of factors affecting enrollment pattern in Indian universities using k-means clustering", International Conference on Information Technology (InCITe) - The Next Generation ITSummit, IEEE, 321-326. https://ieeexplore.ieee.org/abstract/document/7857639
3. Kittinan Limsathitwong, Kanda Tiwatthanont, Tanasin Yatsungnoen, (2018), "Dropout Prediction System to Reduce Discontinue Study Rate of Information Technology Students", 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, IEEE, 110-114. https://ieeexplore.ieee.org/document/8391176
4. Musa Wakil Bara, Nor Bahiah Ahmad, Mohammed Maina Modu, Hamisu Alhaji Ali, (2018), "Self-Organizing Map Clustering Method for the Analysis of E-Learning Activities", IEEE. https://ieeexplore.ieee.org/document/8363155
5. Ting Zeng, (2017), "The Research and Practice of a Five-sided Educational Data Mining Framework", IEEE. 1050-1053. https://ieeexplore.ieee.org/document/8122514
6. P. Nithya, B. Umamaheswari, A. Umadevi, (2016), "A Survey on Educational Data Mining in Field of Education", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, 69-78. http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-1-69-78.pdf
7. Amjad Abu Saa, (2016), "Educational Data Mining & Students' Performance Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 212-220. https://thesai.org/Downloads/Volume7No5/Paper_31-Educational_Data_Mining_Students_Performance_Prediction.pdf
8. A.Dinesh Kumar, R.Pandi Selvam, K.Sathesh Kumar, (2018), "Review on Prediction Algorithms in Educational Data Mining", International Journal of Pure and Applied Mathematics, Volume 118 No. 8 531-537,531-538. https://acadpubl.eu/jsi/2018-118-7-9/articles/8/77.pdf
9. Hilal Almarabeh, (2017), "Analysis of Students' Performance by Using Different Data Mining Classifiers", International Journal of Modern Education and Computer Science, 9-15. http://www.mecs-press.org/ijmecs/ijmecs-v9-n8/v9n8-2.html
10. D. Fatima, Dr.Sameen Fatima, Dr.A.V.Krishna Prasad, (2015), "A Survey on Research work in Educational Data Mining", IOSR Journal of Computer Engineering (IOSR-JCE), 2278-0661, Volume 17, Issue 2, Ver. II, PP 43-49. http://www.iosrjournals.org/iosr-jce/papers/Vol17-issue2/Version-2/G017224349.pdf
11. Pooja Thakar, Anil Mehta, Manisha, (2015), "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Applications (0975 – 8887), Volume 110 – No. 15, 60-68. https://arxiv.org/ftp/arxiv/papers/1509/1509.05176.pdf
12. Wongpanya Nuankaew, Pratya Nuankaew, Sittichai Bussaman, Passakorn Tanasirathum, (2017), "Hidden Academic Relationship between Academic Achievement and Higher Education Institutions", IEEE. https://ieeexplore.ieee.org/document/7904982

13. Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwib, Najoua Ribata, (2018), "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 9, No. 2, pp. 447~459, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v9.i2. pp 447-459. http://ijeecs.iaescore.com/index.php/IJEECS/article/view/9746
14. Ajay Kumar Pal, Saurabh Pal, (2013), "Analysis and Mining of Educational Data for Predicting the Performance of Students", International Journal of Electronics Communication and Computer Engineering, Volume 4, Issue 5, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209,1560-1565. https://pdfs.semanticscholar.org/1e43/2b8683f8b003349d1fa865ec861791d5b405.pdf
15. Arpit Bansal, Mayur Sharma, Shalini Goel, (2017), "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 157 – No 6, 35-40.https://pdfs.semanticscholar.org/f0c5/a423fcdbe7d5faa52d58b26a9bfd2ca36587.pdf
16. Dorina Kabakchieva, (2012), "Student Performance Prediction by Using Data Mining Classification Algorithms", International Journal of Computer Science and Management Research, Vol 1 Issue 4, ISSN 2278-733X, 686-690. https://www.researchgate.net/profile/Dorina_Kabakchieva/publication/272178031_Student_Performance_Prediction_by_Using_Data_Mining_Classification_Algorithms/links/5a151ffda6fdccd697bc01ea/Student-Performance-Prediction-by-Using-Data-Mining-Classification-Algorithms.pdf
17. HemaMalini B. H., L. Suresh. (2018). "Assessment of Performance of Engineering Students using Educational Data Mining". Journal of Emerging Technologies and Innovative Research. 5(6). 16-20. http://www.jetir.org/view?paper=JETIRC006003
18. HemaMalini B. H., L. Suresh, Mayank Kushal. (2019), "Comprehensive Analysis of Students' Performance by applying Machine Learning Techniques", Smart Innovation, Systems and Technologies, Smart Intelligent Computing and Applications, Proceedings of the Third International Conference on Smart Computing and Informatics, 978-981-32-9689-3. https://link.springer.com/chapter/10.1007/978-981-32-9690-9_60
19. HemaMalini B. H., L. Suresh. (2018), "Data Mining in Higher Education System and the Quality of Faculty Affecting Students Academic Performance: A Systematic Review", International Journal of Innovations & Advancement in Computer Science, 2347 – 8616 Volume 7, Issue 3. http://academicscience.co.in/admin/resources/project/paper/f201803081520493538.pdf