Correlation between Absence, Interest in the Field and Grades in an Organization using Regression Model

Ahan Chatterjee, Trisha Sinha

Abstract: ***"Who questions much, shall learn much, and retain much."-Francis Bacon English Philosopher This quotation by Francis Bacon conveys that people, who question more, learn more in comparison to their peers and it is quite natural that one has to be present in the class to bring forth his or her question. In today's world of advancing technology, absence from class is a major challenge not only in West Bengal, India but across the globe. The growth in this trend of being absent in class eventually leads to poor grade. In this paper, we aim to find a relation between absence, interest of student in subject, staying in hostel or travelling regular to college and the final grade of the student. Also, it will be our aim to find out how the above mentioned factors affect the grade of that particular student. A sample of 86 students is taken into account to predict the outcome. The data regarding the interest in each subject is collected from the individual students to calculate while the attendance and grades are collected from the college authority. The statistical models used in findings are students t-test, Pearson's correlation and regression model. Hypothesis generated for each independent variable affecting the target variable Grade and through statistical calculations, conclusions are drawn whether or not they really affect or it is simply a myth. This study is beneficial for both, the college authority as well as the students to create awareness among them regarding the drawbacks of not attending the classes and not creating interest in the subject.***

Keywords: Descriptive Statistical Analysis, Linear Regression model, Multivariable Linear Regression Model, Statistical Analysis

I. INTRODUCTION

In the modern world of advancing technology, a major challenge in a university is the increasing trend of absenteeism among the students. Absenteeism is defined as the habit of one to fail to present himself or herself in his or her respective class or in an event without a reasonable excuse in support and the term absentee is used to describe somebody who does this act frequently. It is seen over various research and blind observation that failing to attend classes lead to poor grades scored by those students, whereas on the other hand students whose attendance scale is higher than their respective peers, score more marks. The Grade of a student not only depends on attendance, but the interest of that student on the subjects.

A student with high interest in a subject will invest more time behind that particular subject whereas low interest leads to negligence of subjects which will eventually lead to low grades. The travelling time consumed by a student is also a factor determining his grade. A student who lives in the university or college hostel does not require any travelling, whereas a student who travels for an hour or so has lower time for studying after reaching home. The exhaustion and tiredness caused by this travelling adversely affects the student and leads to lowering of concentration in the student which ultimately results in lower grades than his peers residing in the hostel.

All the assumptions that have been framed are framed as hypothesis and proper statistical model is used to prove those and relations have been proposed between them that how they will vary with the change of the each and every factor.

In this paper we used various statistical tests to conclude on the relations, and we have also used regression model to predict the *OLS* equation to give the overview how grade vary on each factor.

Pearson Correlation factor is used as a parameter to check how closely the factors are related to each other, if related or not also. Many other factors also influence the grades like self study time, other engagements etc. these factors are not taken into account for this paper.

This paper proposes an equation which shows how a student's grade varies over factors like attendance, interest in subject and travelling time of a student.

II. LITERATURE REVIEW

Various previous research proven that successful student has better attendance track record and a better grade also. Research work of Ahmat and Zahari showed that there is a negative correlation between absence and grades (r = -0.611), which proves more absence leads to poor grade. [1] Horn and Jansen also stated that academic performance have a positive correlation in between them (r = 0.716) [2]. Senior research also suggests similar results, it also shows a positive correlation between attendance and grades (r = 0.768). [3] Research work of Zhang and Wang also suggests similar results through linear regression model that a positive correlation between the desired variables. [4]
Statistical analysis done by Leon has also strong positive correlation of around (0.814) which also evident of our assumption. [5]
Research work by Bashir also suggests similar outcomes. [6]
Similarly research work carried out by Narula and Nagar shows a strong 76.8 % correlation between attendance and grades. [7]

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Ahan Chatterjee, B.Tech Student, Department of Computer Science and Engineering Specialization in Data Analytics, The Neotia University B.Tech Student, Department of Robotics, The Neotia University
Trisha Sinha, B.Tech Student, Department of Computer Science and Engineering Specialization in Data Analytics, The Neotia University B.Tech Student, Department of Robotics, The Neotia University

All this research work suggests there is a strong positive correlation between attendance and grades.

III. PROPOSED WORK

Our paper aims to bring a model which can evaluate the impact of each factor on grade such as attendance, interest in subject, and travelling time. In the previous research works the main spotlight was given to only one to one variation between attendance and grade. But here more independent variables are introduced (attendance, interest in subject, and travelling time) and one dependent variable (grade). This model will give closer view of the grade variation with more parameters taken into account.

IV. LINEAR REGRESSION MODEL

Regression model or analysis is one of the statistical methods to study the relationship between two or more variable in a dataset. The regression model comes under the supervised learning part as the dependent and independent values are known and numeric in nature. It's generally used to find effect of one variable on other. For example, the effect of price inflation of vegetables upon low crop yields. Here generally we use the statistical significance to assess the significance of the model of the estimated relationships i.e. degree of confidence.

The model has a dependent variable y, such as linear function of one independent variable x and an error u. It's represented as,

$$y = a + bx + u \qquad (i)$$

The term with error u, is assumed to have a mean value of zero, a constant variance, and to be uncorrelated with itself across observations.

The estimation of to find the linear regression, we need to determine the coefficients, a and b the unknown parameters. The estimated equation will have the form of,

$$y' = a' + b'x \qquad (ii)$$

The error that is calculated with each pair of data value as,

$$u' = y - y' = y - (a' - b'x) \qquad (iii)$$

The scatter and plotted diagram of general linear regression line is shown in the below figure.

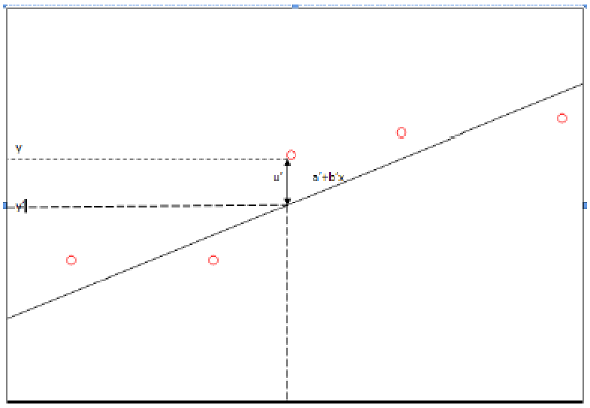

Fig 1: Regression Residual

Source: Created by the Author

The most basic method to calculate the coefficients, a' and b' is the method of *Ordinary Least Square (OLS);* here the values of a' and b' are chosen such that Sum of Squared Residuals (SSR) are minimized. SSR is expressed as,

$$SSR = \sum u'^2 = \sum (y - y')^2$$
$$= \sum (y - a' - b'x)^2 \qquad (iv)$$

We need to minimize the SSR and for it we take calculus approach, partial derivatives of both a' and b' and equate them to zero. This will generate two equations, and solving them will give the formula.

$$\frac{\partial SSR}{\partial a'} = -2\sum (y - a' - b'x) = 0 \qquad (v)$$

$$\frac{\partial SSR}{\partial b'} = -2\sum x(y - a' - b'x) = 0 \qquad (vi)$$

Solving these two equations we get,

$$-2\sum (y - a' - b'x) = 0$$
$$\sum y - na' - b'\sum x$$

$$a' = y'' - b'x'' \qquad (vii)$$

And calculating and substituting the required values we get,

$$b' = \frac{\sum xy - y''\sum x}{\sum x^2 - x''\sum x} \qquad (viii)$$

Equation number *(vii) and (viii)* is used to find the coefficients of the regression equation a' and b'.

1. Methodology

This research work is carried off in The Neotia University, across 86 students. The students are selected in random basis. The interests in subjects are recorded from each and every student through goggle form, from where we have computed the interest. The attendance and the student is hosteller or day scholar has been concluded through the information from college. Various other factors also influence the grade such as communication, teaching method, and self learning etc.

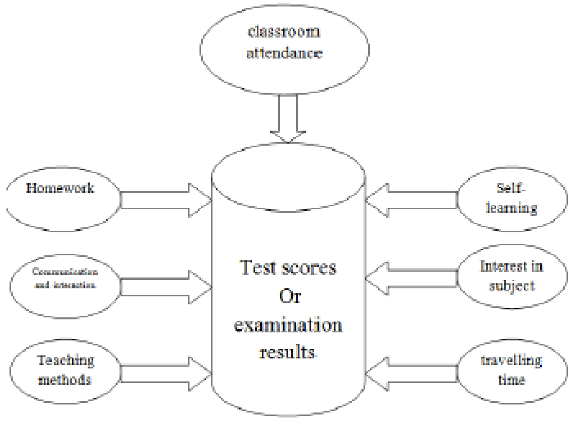

Fig 2: Test scores relevant factors

Source: Created by the Author

Important data which are involved in this study are grades (CGPA), attendance, interest, hosteller or not.

The hypothesis generated for the first case, in between CGPA and total attendance.

Null Hypothesis = H_o = Attendance doesn't affect Grades.

Alternate Hypothesis = H_a = Attendance affect Grades.

Two hypotheses are being generated, and we will now go statistical independent T-Test to check whether null or alternate hypothesis stands true. The independent T-test is used to determine whether there are any significant effect between attendance and academic performance (CGPA).

The T-test has been carried out in Jupyter Notebook and the p value comes as

$$Pvalue = 0.0158$$

Thus the $Pvalue < 0.05$ thus in this case we can safely reject our Null Hypothesis and the alternate hypothesis stands and we can safely conclude that Attendance affect Grades.

The frequency of attendance varied from 20% to some cases of 95%. To visualize that frequency we have plotted the histogram to visualize the frequency.

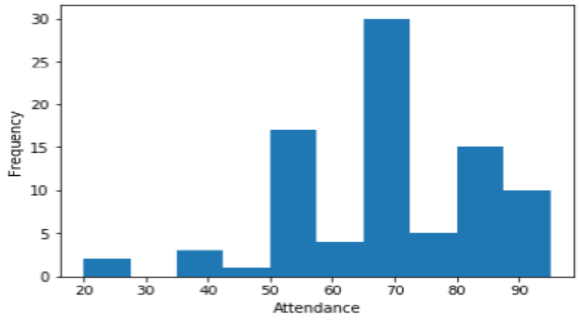

Fig 1: Histogram of Attendance Source: Created by the Author

Through this histogram we can see the most students have their attendance in the range of 65% - 75%.

Similarly, for the next variable interest in subject we will use similar hypothesis generation and conclude with the independent T-test. The hypothesis generated for the second case, in between CGPA and interest in subject. Null Hypothesis = H_o = Interest in subject doesn't affect Grades.

Alternate Hypothesis = H_a = Interest in subject affect Grades.

Two hypotheses are being generated, and we will now go statistical independent T-Test to check whether null or alternate hypothesis stands true. The independent T-test is used to determine whether there are any significant effect between attendance and academic performance (CGPA).

The T-test has been carried out in Jupyter Notebook and the p value comes as

$$Pvalue = 0.08678$$

Thus the $Pvalue < 0.05$ thus in this case we can safely reject our Null Hypothesis and the alternate hypothesis stands and we can safely conclude that Interest in subject affect Grades.

We have collected the data of interest from the students in the scale of 0-5 that is 0 means no interest at all and on the other hand 5 means high interest in the subject. We have plotted a histrogram for this also to visualize the frequency distribution of interst among students.

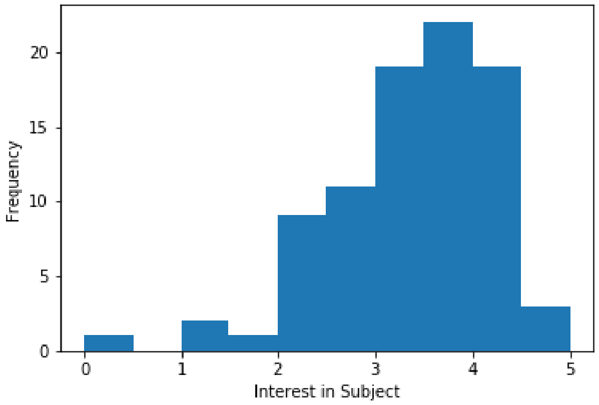

Fig 2: Histogram of Interest in subject Source: Created by the Author

Similarly for the variable hosteller or not we have hot encoded the dataset, as being a hosteller is a yes or no, and that is a categorical value and we can apply T-test only in case of numerical data, thus we have applied one hot encoding method in Jupyter Notebook to convert the categorical variable into numerical variable.

Similarly we taken hypothesis,

The hypothesis generated for the second case, in between CGPA and interest in subject.

Null Hypothesis = H_o = Being Hosteller doesn't affect Grades.

Alternate Hypothesis = H_a = Being Hosteller affect Grades.

The T-test has been carried out in Jupyter Notebook and the p value comes as

$$Pvalue = 0.016867$$

Thus the $Pvalue < 0.05$ thus in this case we can safely reject our Null Hypothesis and the alternate hypothesis stands and we can safely conclude Being Hosteller affect Grades. In this case we can't plot histogram as this is a categorical variable, thus we plot density curve to check the effect of distance travelling over grades.

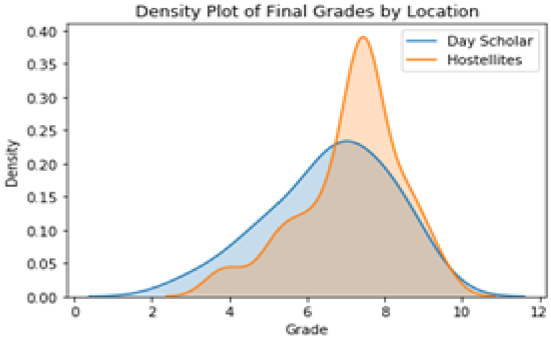

Fig 3: Density Plot of Final Grade by Location Source: Created by Author

In the T-Test we confirmed that all factors which we took into account affects the grade. In the next statistical test we would find how strongly these factors are correlated with the grade (CGPA). Pearson's Correlation test is used to find the correlation among them.

Attendance and interest in subject are numeric data and in Jupyter notebook correlation is easily calculated among them, but in case of being a hosteller or not this is a categorical variable thus we use one hot encoding to convert this into numerical data type and to find the correlation (r).

From the Pearson's Correlation Test:

Independent Variable	Dependent Variable	Pearson Correlation	Type of Correlation
Attendance	Grade (CGPA)	0.448	Positive Correlation
Interest in Subject	Grade (CGPA)	0.227	Positive Correlation
Travelling Time (Day Scholar)	Grade (CGPA)	-0.184	Negative Correlation
No travelling Time (Hosteller)	Grade (CGPA)	0.184	Positive Correlation

Table 1: Correlation among the variablesSource: Created by the Author

- There is a positive correlation among attendance and dependent variable, and it's a pretty strong correlation shows with the increase in attendance will lead to higher grade.
- Similarly there is also positive correlation between Interest in Subject and Grade, here the correlation is a bit weak but it also shows which the increase of one in a subject respective grade will increase.
- Here we see there is a negative relation in between travelling time and grade; this proves that will the increase of daily travelling time grades tends to decrease. Here student gets lower time in home to study and tiredness also plays a key role in that.
- Again there is a positive correlation between no travelling time and just being a hosteller yield better result.

To visualize the correlation among the variables, we plot a pairplot using the seaborn library in Jupyter Notebook, to visualize how the independent and dependent variables are correlated among each other. In the pairplot is creates a matrix and plots the relationship among each pair of the column. Through the diagonal it also draws a univariate analysis.

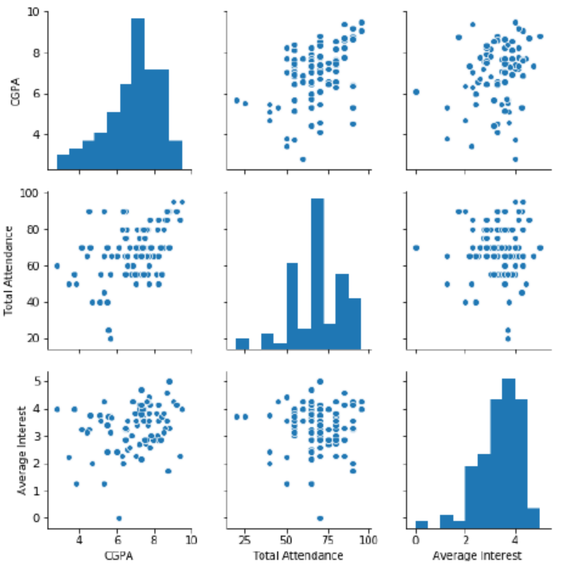

Fig 4: Pairplot of all variables Source: Created by Author

In further data visualization we used the distplot() function in the seaborn library to draw the central tendency curve over the histogram of attendance, grade (CGPA), and interest in subject.

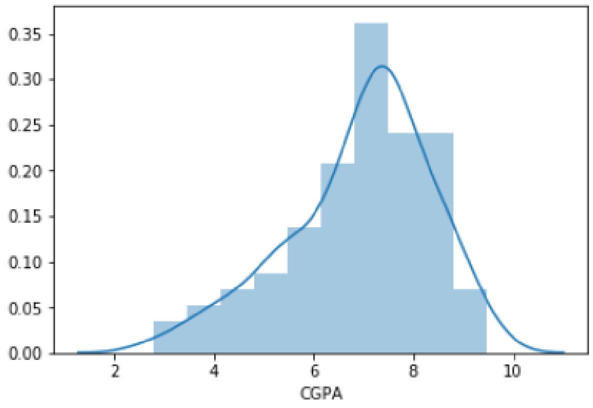

Fig 5: Tendency curve over histogram in CGPA

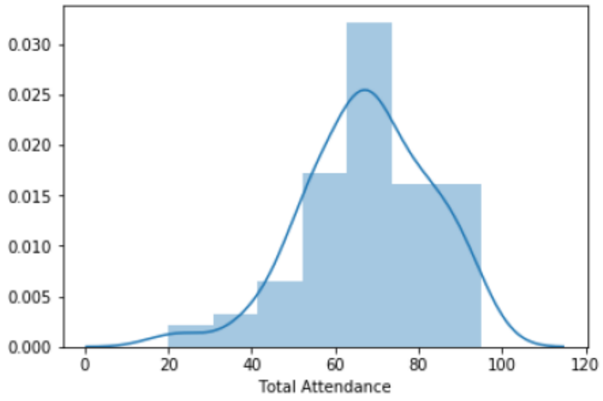

Fig 6: Tendency curve over histogram in Total Attendance

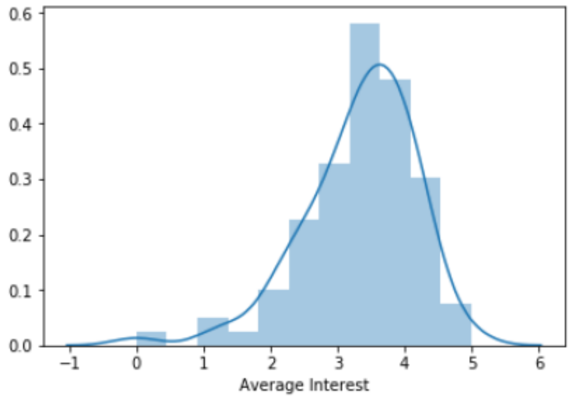

Fig 7: Tendency curve over histogram in Interest in subject Source: Created by Author

In modern days the result section carries a new factor called the percentile, percentile is defined as how many candidates scored below your marks that is called your percentile. For example if your percentile is 98 then, 98% of the appearing candidates in the exam has lower marks than yours. Thus a visualization of the percentile of marks is necessary in nowadays result analysis. Thus we have plotted a percentile curve which shows how the percentile is changing over the marks scored in the collected dataset.

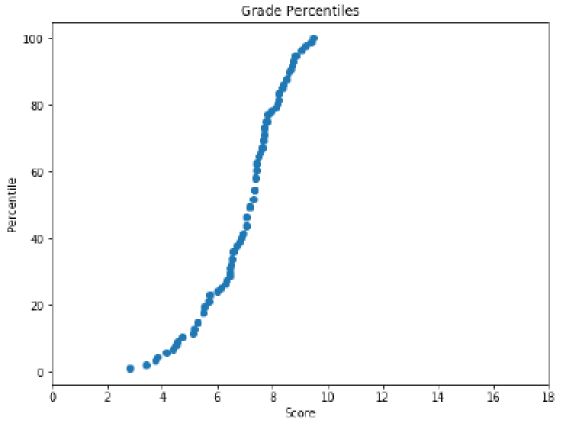

Fig 8: Percentile Curve Source: Created by the author

A line relation is also being plotted between the attendance and grade (CGPA) and it turns out as;

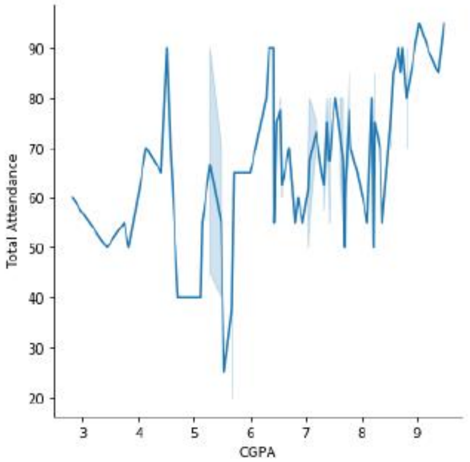

Fig 9: Line relation between attendance and CGPA Source: Created by Author

The median of the CGPA in the dataset is calculated and a curve is plotted showing how much peers is scoring above the median and how many of them are below. The graph in the upper part signifies the better student in comparison to others present in the dataset.

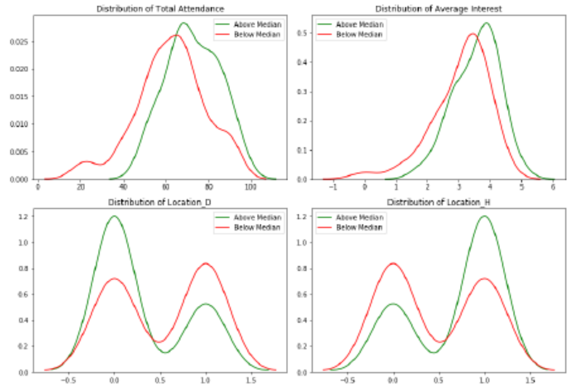

Fig 10: Graph plotted above and below median CGPA for other independent variables

Source: Created by the Author

Here in the dataset the median came out to be 7.19 thus students above and below median are plotted. The Location_D signifies the day scholars where the students travel to the college or university, and the Location_H signifies students who stay in the hostel in college or university campus where travel time is nil.

After all the statistical model and data visualization, the final linear equation varies as;

$$\begin{aligned} \boldsymbol{Grade} = \mathbf{3.05} + \mathbf{0.04} * \boldsymbol{Total\ Attendance} \\ + \mathbf{0.36} * \boldsymbol{Average\ Interest} \\ + -\mathbf{0.28} * \boldsymbol{Location\_D} + \mathbf{0.28} \\ * \boldsymbol{Location\_H'} \end{aligned}$$

This equation is obtained from the linear regression model which has been implied on the following dataset.

- The coefficient of the attendance is +0.04 which signifies that when a candidate is present for unit (1) class the final grade (CGPA) will be increased by an average of 0.04%.
- The coefficient of the Interest in subject is +0.36 which signifies when the interest of a candidate increase by a unit (1) in a subject it's grade (CGPA) will be increased by an average of 0.36%
- The coefficient of the day scholar that is those who travel to college everyday is -0.28 which signifies when the travelling of a candidate increase by a unit (1) it's grade (CGPA) will be decreased by an average of 0.28%.
- The coefficient of the hostellers that is those who stay in college everyday is +0.28 which signifies staying in hostel of a candidate increase by a unit (1) it's grade (CGPA) will be increased by an average of 0.28%.

V. CONCLUSION AND FUTURE SCOPE:

The study has been done extensively with lots of data visualization and statistical modeling has been used in the paper. The correlation and T-test showed that there is a relation among the taken independent variable and the coefficients of the linear regression showed by how much grade will vary with a unit change of the independent factors. The research area can be expanded by including various other factors into consideration like parent’s education, family status etc. If we include more factors like this more accurate results could be achieved. In the further research we will try to include those factors to give more accurate variation of grade.

Compliance with Ethical Standards

Funding: No such funding was received for this research project.

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

1. Nurhafiaz Ahmad, Ahmad- Zia, Siti Asmah Mohamed, Hasfazilah Ahmad, Mohd Fazmi Zahari, The impact of class absenteeism on students academic performance using regression model, 2012
2. Bolivar A. Senior, Correlation between Absences and Final Grades in a college course
3. Chengcheng Zang, Fei Wang, Research on correlation analysis between test score and classroom attendance based on linear regression model, 2010
4. Costas Leon, The relation between Absences and Grades: A Statistical Analysis, 2018
5. Suleiman Obeidat, The importance of class attendance and cumulative GPA for academic success in industrial Engineering Classes, 2016
6. Meenakshi Narula, Pankaj Nagar, Relationship between student’s performance and class attendance in a programming language course, 2013

BIOGRAPHIES OF AUTHOR

Ahan Chatterjee is a young and dynamic engineering student at The Neotia University, in the department of Computer Science and Data Analytics. He has been a part of research internship at CSIR-CDRI , GoOffer Hyperlocal Pvt. Ltd. as a Research & Development Intern. He also worked as Research Analyst Intern at Research Guruji after his completion of first year itself. He has 1 accepted and 1 published research article to his name. He is a member of International Association of Engineers, Hong Kong. He wants to pursue his career in research based domain. His field of interest is Data Analytics, Machine Learning and Artificial Intelligence.

Trisha Sinha completed her initial studies from a convent school and later joined The Neotia University to pursue her career in Robotics Engineering. She has performed fairly well all through her school life and is in the 2nd year of her B.Tech Programme. She has a strong grasp over C and C++ language, along with some handy skill in Arduino programming and working with various sensors. She has interest in the fields of Automation, Deep Learning, Machine