Factors Affecting Student's Academic Performance

Wan Alia Izzati Binti Wan Abdul Razak[1], Siti Nur Adawiyah Binti Khairul Akmal[2], Nur Fatinhieyah Binti Azizan[3], Sharyar Wani[4], Abdul Qayoom Hamal[5], Abdul Hafeez Muhammad[6]

[1234]Department of Computer Science, Kulliyyah of Information & Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia
[5]Faculty of Information & Communication Technology, HELP University, Kuala Lumpur, Malaysia
[6]Dept. of Computer Science, Bahria University, Lahore, Pakistan
wanaliaizzati@gmail.com, adawiyahakmal@gmail.com, fatinhieyah@gmail.com. sharyarwani@iium.edu.my, abdul.qayoom@help.edu.my, ahafeez.bulc@bahria.edu.pk

Abstract— The educational system measures students' performance using varied assessment methods such as quizzes, tests, examinations and assignments. Academic performance along with skill sets is strongly linked with positive outcomes upon completion of studies in the ever-demanding job market or pursuance of higher studies. Various factors can affect students' academic performance such as mental issues, working status, time spent on gadgets and study duration. This research studies the role of various factors in order to understand their role in the student's academic performance. Furthermore, various machine learning algorithms are implemented for the prediction process for a student's performance range. The results present that various internal and external pressure on a student is a major contributor to the state of academic performance among other factors discussed in this paper. Light Gradient Boosting Machine performs the best compared to other algorithms in the multi-class classification problem with an overall accuracy and F1-score of 0.87 and 0.86, respectively.

Keywords— Academic Performance, Student, Self Confidence, Family Pressure, Light Gradient Boosting Machine, Gradient Boosting Classifier, K Neighbors Classifier

I. Introduction

Academic performance is measured by assessing a student's ability in a variety of academic areas. Classroom performance, graduation rates, standardised test results and students' past semester CGPA/GPA are commonly used by teachers and education administrators to assess student accomplishment. Most post-secondary institutions currently employ the grade point average (GPA) as an indicator of their students' academic success. The GPA is considered to provide clearer picture of an individual's or a group of students' relative level of performance.

There are many factors that can affect academic performance of students such as the current condition of their mental health, working status, duration of the study time, spending time on social media per day, etc. Statistically, about 41.2% and 44.3% of students claim that their academic performance was affected by their emotions and mental health respectively [1]. Time management skills also have an impact on students' achievements. Proper time management becomes an important aspect in education especially at higher levels of study [2]. Students who are working or spending too much time on social media might have lower GPA compared to full-time students. However, these cannot be generalized as students have different capabilities and ways to study.

It becomes essential to investigate factors affecting student performance in order to understand the challenges faced by the students. The determination of these factors will help the students to work further in their lacking areas to improve their academic performance. Therefore, this research investigates various factors that contributes to the student's academic performance. Based on these factors, this research implements various prediction algorithms to determine students' academic performance in the upcoming semesters.

II. Related Work

In a bid to support students and instructors alike, machine learning has been the focus of many researches for analysis and prediction of voluminous and complex educational data. A hybrid technique involving principal component analysis with various machine learning models has been proposed in [3]. Hybrid Naïve Bayes outperforms other models with an accuracy above 0.90. Overall, the results indicate the hybrid approaches outperform the baselines for all the models rendering the hybrid approach quite affective for prediction.

Academic performance in higher institutions of learning is mostly based on quizzes, assignments, lab exams, mid, and final exams. Beyond this other information such as participation in the class, accessing instructor resources, etc for individual topics also tend to be key indicators for prediction of student performance. A constant analysis of these components tends to aid the students and teachers to plan strategies to enhance the quality of teaching and learning. Decision tree, Support Vector Machine, and Naive Bayes have been implemented to predict the student's academic performance on a dataset containing information about 400+ students [4]. Using standard metrics of measurement Naïve Bayes scored a 77% accuracy while trailing behind Decision Tree with a F1-Score of 0.73. Based on a descriptive qualitative analysis in [5], age, parental income and number of study hours were found to have a significant role in improving the student performance of graduate students. Therefore, it can be concluded that the non-conventional factors i.e. other than scores in assessment methods can be used as predicting variables for early detection of academic performance.

While excessive use of electronic gadgets may not directly impact the academic performance of students, they tend to impact physical, emotional and mental well-being which in turn become deterrents to decent academic performance. In a quantitative study conducted by [6] on a sample of 240 students in age range of 12-16 years, it was found that almost 70% students already developed the habit of using smart devices before sleep. A 59% student population reported morning headaches while 53% students reported lack of focus in the classes or other tasks. These findings were found to be significantly correlated to the use of smartphones leading to obesity, sleep disorders, aggressive behaviour among others.

In a predictive analytic task for academic performance of public schools in Brazil, Gradient Boosting Machine was used on two datasets – one containing variables prior to the start of school year and the second two months into the academic year. Results indicate that grades and number of absences as the most significant indicators to the end of year academic performance. Other important indicators include neighbourhood, school and age. GBM presented a high efficiency for the prediction process using these variables with ROC. Scores greater than 0.90 in all cases [7]. Financial conditions, learning motivation and gender were found to be significantly correlated to academic performance in a computer science course in Nigerian College for Education. Using Decision Trees for the prediction process, it was found that 67% students tend to pass based on the level of the aforementioned factors. Gender was found to have a effect in the course performance, whereby the higher likelihood was related to male students [8]. Similarly in [9], using K-means clustering and multiple linear regression, it was observed that in semester test, quiz and assignment scores were key indicators of academic performance in the final exams. Chronic stress in medical students has been reported to affect their academic performance leading to depression and suicides. In an Ethiopian medical student setting, it was found that more than half of the medical students were stressed. The main source of stress in this cross-sectional study was academic stress. The high percentage of stress was found to be significantly associated with smoking, alcohol consumption and khat chewing [10]. Age and academic performance was found to be negatively corelated among a study of more than 1300 students by [11]. Table 1 provides some more details pertaining to tools/techniques of the aforementioned works.

III. Experimental Setup

A. *Dataset*

The dataset used in this research was collected by surveying United States students from different levels of study - high school, college and master. It is publicly available on Kaggle as "Effects of Depression and Anxiety on Academic Performance Among the Students" [12]. The recorded data consists of 18 variables and information about 352 students. GPA is chosen as a dependent variable for this study.

B. *Data Pre-processing*

After checking and processing for missing values, duplicate data, outliers etc. the GPA column was categorized into low, medium and high categories. Low category indicates GPA lower than 2.49, medium level ranges between 2.50 and 3.49 while the high level is a representative of GPA more than 3.50.

C. *Correlation*

Correlation describes the relationship between two quantitative variables. A correlation test is used to determine the strength of each feature in relation to the intended output and the correlation coefficient of each feature is calculated.

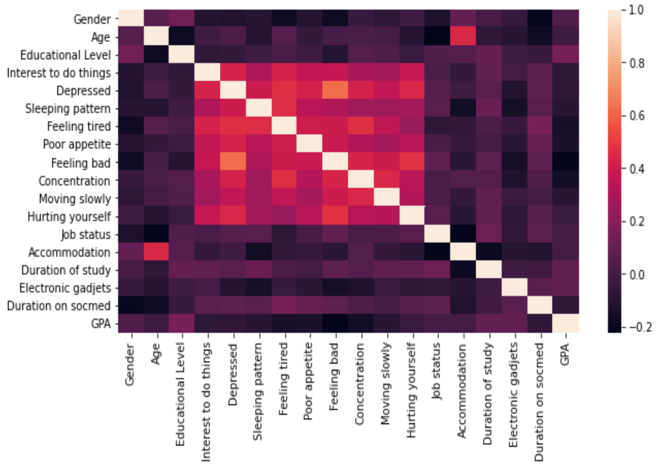

Fig. 1 Correlogram for all variables in the dataset

Fig 1 is a correlogram of the dataset, adapted from the heatmap function. Correlation coefficients for all pairings of variables are shown in the correlogram (with more vivid colours for more severe correlations), while correlations that are not substantially different from 0 are represented by a white box. Correlations that are positive are shown in blue while those that are negative are shown in red. The colour intensity is related to the correlation coefficient, darker the boxes greater the correlation (the closer to -1 or 1).

D. *Data Visualization*

The graphical depiction of information and data is known as data visualisation. Data visualisation is the process of converting information into a visual representation, such as a map or graph, in order to make data simpler to comprehend and extract insights from. Data visualization's major objective is to make it simpler to see patterns, trends, and outliers in huge data sets. Fig 2- Fig 17 provides some valuable insights about the data using various visualizations. Fig 10 until Fig 18 specifically visualize the ratings of attributes related to health issues among students.

IV. Exploratory Analytics

This section is divided into two parts – the first section deals with some exploratory analysis about the mental health of university students. The second section presents comparative performance of the prediction models used in this research.

A. *Exploratory Analysis:*

The exploratory analysis presents the following conclusions:

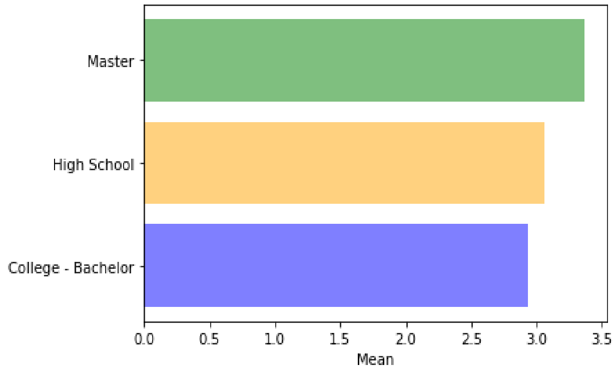

Fig. 19 Average GPA of each education level

The average GPA of master students has the highest value which is 3.37 followed by high school at around 3.06 and the lowest average GPA is reported by bachelor students bearing a value of 2.93 (see Figure 19).

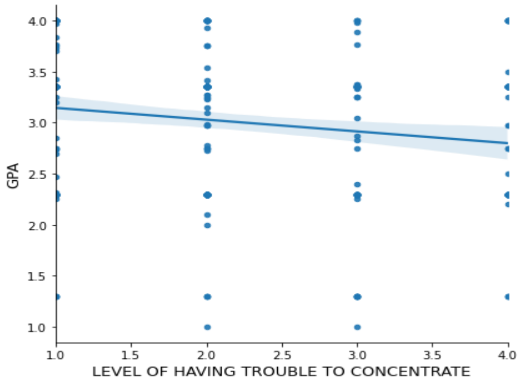

Fig. 20 Relationship between the level of having trouble to concentrate and GPA

Fig 20 is based on two attributes that contain the rating of students about their concentration issues, such as reading the newspaper or watching television and the student's GPA. The plot shows that higher the rating of concentration issues, the lower the GPA.

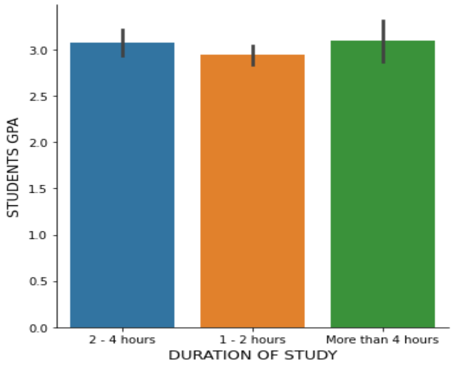

Fig. 21 Relationship between the duration of study and GPA

Students spending more time to study report higher GPA's as seen in Fig 21. Spending more time most helps to increase understanding of the lessons, which in turn helps to perform well in exams leading to a better overall score.

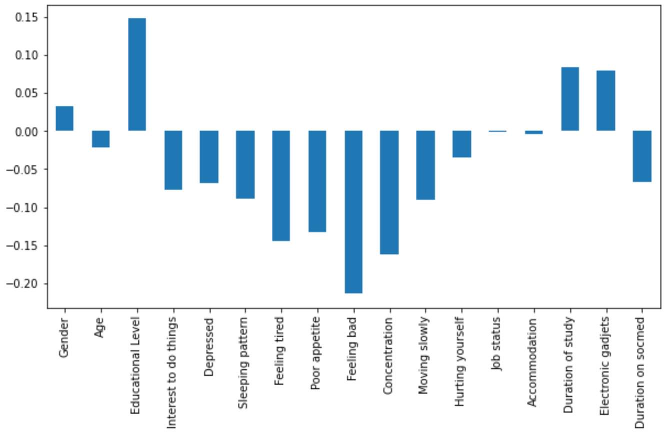

Fig. 22Correlation Bar Chart

The correlation bar chart in Fig 22 indicates highest value of negative correlation for feeling bad. It indicates that among all the factors, the most likely to affect students' academic performance is students feeling bad about themselves. The pressure of proving oneself or meeting family expectations tends to play a critical role towards the performance. When students feel bad about themselves, their motivation to study decreases and can affect their academic performance. The second factor that has a major impact on student's performance is trouble concentrating on things, followed by students feeling tired or having little energy to do anything.

B. *Predictive Performance:*

The predictive performance has been presented in Table II, Fig 23-Fig 25.

TABLE II
PREDICTIVE PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score	AUC
lightGBM	0.87	0.87	0.85	0.86	0.94
GBC	0.69	0.69	0.67	0.68	0.84
KNN	0.56	0.57	0.53	0.56	0.74

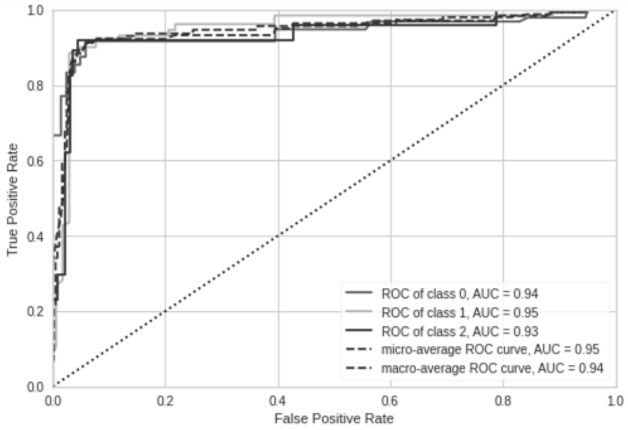

Fig 23. ROC Curves for lightGBM

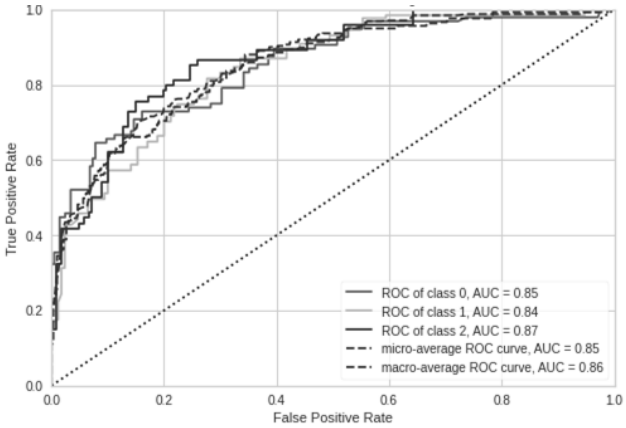

Fig 24. ROC Curves for GBC

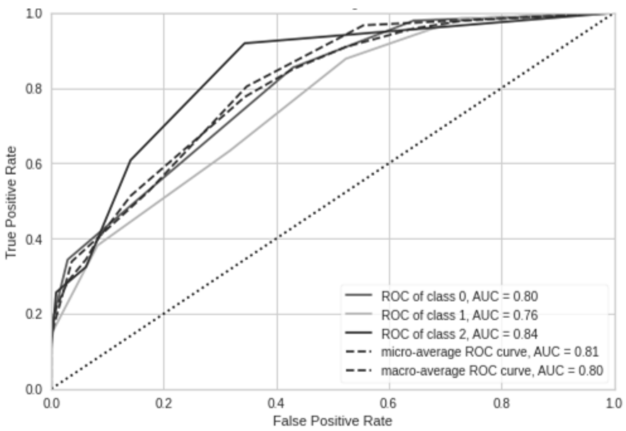

Fig 25. ROC Curves for KNN

The overall spredictive performance reveals that Light Gradient Boosting Machine (lightGBM) outperform Gradient Boosting Classifier (GBM) and K Neighbors Classifier (KNN) classifier for the multi-class classification in this research. lightGBM scores highest in each category with an accuracy of 87% and harmonic mean of 86%. GBC and KNN lag behind by 18% and 31% for accuracy while the difference in harmonic mean is lesser by 2 and 3 percent compared to the accuracy respectively.

V. CONCLUSION

This paper identified the factors that contribute significantly to a student's academic performance based on the survey taken from the students. The correlation test revealed that lack of self-confidence due to pressure of proving oneself or for the family is the most significant reason for low academic performance. This affects concentration and quality study time which are significant for one to perform well in various academic activities.

The paper explored three predictive modelling techniques – lightGBM, GBM and KNN for predicting academic performance of students in a multi-class classification setting. The performance metrics deemed lightGBM as the most suitable algorithm overall for this prediction problem using the current dataset. Other models should be explored in future with more data and further data pre-processing techniques to achieve better predictive performance. The data size is considerably small consisting 352 rows which might result in biased results as it is unrepresentative of the population. Therefore, it is recommended to explore the domain problem with more data in future.

REFERENCES

[1] J. Moreira de Sousa, C. A. Moreira, and D. Telles-Correia, "Anxiety, depression and academic performance: A study amongst Portuguese medical students versus non-medical students," *Acta Med. Port.*, vol. 31, no. 9, pp. 454–462, Sep. 2018, doi: 10.20344/amp.9996.

[2] A. Florence Aduke, “Time Management and Students Academic Performance in Higher Institutions, Nigeria — A Case Study of Ekiti State,” *Int. Res. Educ.*, vol. 3, no. 2, p. 1, Apr. 2015, doi: 10.5296/ire.v3i2.7126.

[3] P. Sokkhey and T. Okazaki, “Hybrid machine learning algorithms for predicting academic performance,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 32–41, 2020, doi: 10.14569/ijacsa.2020.0110104.

[4] A. Sai Saketh, “Student’s Academic Performance Prediction Using Machine Learning Approach,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 9s, pp. 6731–6737, Jun. 2020, Accessed: Jun. 29, 2021. [Online]. Available: https://www.kaggle.com.

[5] A. Abaidoo, “Factors contributing to academic performance of students in a Junior High School,” *Univ. Educ. (Distance Learn.*, p. 99, 2018, Accessed: Jun. 29, 2021. [Online]. Available: https://www.grin.com/document/450284.

[6] A. M. Hegde, P. Suman, M. Unais, and C. Jeyakumar, “Effect of Electronic Gadgets on the Behaviour, Academic Performance and Overall Health of School Going Children-A Descriptive Study,” *J. Adv. Med. Dent. Sci. Res. |Vol*, vol. 1, 2019, doi: 10.21276/jamdsr.

[7] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil,” *J. Bus. Res.*, vol. 94, pp. 335–343, Jan. 2019, doi: 10.1016/j.jbusres.2018.02.012.

[8] K. David Kolo, S. A. Adepoju, and J. Kolo Alhassan, “A Decision Tree Approach for Predicting Students Academic Performance,” *Int. J. Educ. Manag. Eng.*, vol. 5, no. 5, pp. 12–19, Oct. 2015, doi: 10.5815/ijeme.2015.05.02.

[9] O. Tinuke Omolewa, A. Taye Oladele, A. Adekanmi Adeyinka, and O. Roseline Oluwaseun, “Prediction of Student’s Academic Performance using k-Means Clustering and Multiple Linear Regressions,” *J. Eng. Appl. Sci.*, vol. 14, no. 22, pp. 8254–8260, Oct. 2019, doi: 10.36478/jeasci.2019.8254.8260.

[10] L. Melaku, A. Mossie, and A. Negash, “Stress among Medical Students and Its Association with Substance Use and Academic Performance,” *J. Biomed. Educ.*, vol. 2015, pp. 1–9, Dec. 2015, doi: 10.1155/2015/149509.

[11] M. Jose, P. S. Kurian, and V. Biju, “Progression analysis of students in a higher education institution using big data open source predictive modeling tool,” in *2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016*, Apr. 2016, pp. 113–117, doi: 10.1109/ICBDSC.2016.7460352.

[12] “Depression and Academic performance of students | Kaggle.” https://www.kaggle.com/kanerudolph/depression-and-academic-performance-of-students (accessed Jun. 29, 2021).

Appendix

TABLE I
SUMMARY OF RELATED WORK

Reference	Tools/Techniques	Results
[3]	Researchers proposed hybrid models where they combined the Principle Component Analysis (PCA) to baseline models of Random Forest, Decision Tree, Naive Bayes and Support Vector Machine, and 10-CV method to improve the classification performance	The proposed hybrid models boost the accuracy of support vector machine using radial basis function kernel (SVMRBF) from 75.01% to 83.88%, Naive Bayes from 35.79% to 86.27%, Decision Tree from 78.42% to 98.32%, and Random Forest from 80.06% to 98.92%.
[4]	- Decision tree - Support Vector Machine - Naive Bayes	Naive Bayes classification algorithm performs better.
[5]	- descriptive analysis - linear model of graduate student performance was designed. - Simple random sampling technique. - close ended questionnaires.	Age, parental income and study hour have significant role in improving the student performance of a graduate student.
[6]	- descriptive analysis - questionnaire survey	Although use of electronic gadgets is not the sole leading cause to health problems, they do affect significantly to mental and physical health disorders
[7]	- predictive analysis. - Classification models based on the Gradient Boosting Machine (GBM). - DS-I to train the classification model I (CM-I) - DSII to train the classification model II (CM-II)	CM-I, in 2015 The training data – ROC curve 0.967168. 2016 the curve 0.950594, 2015–2016,0.943076. CM-II, 2015, 2016 and 2015–2016, the training data ROC curve 0.990906, 0.986791 and 0.989298, respectively.
[8]	- predictive analysis - decision tree approach	66.8% of the students were predicted to pass while 33.2% were predicted to fail. Gender factor, 55.10% of the male students are predicted to pass while 59.03% of the female is predicted.
[9]	- k-means clustering - Multiple Linear Regression (MLR)	The result shows the intercept B0 = -1.830012, B1 = 0.153268 and B2 = 0.98686. The R^2 = 0.822 which shows the model fits well and the error obtained is minimal R values are between 0-100.
[10]	- A cross sectional survey - Logistic regression analysis	The prevalence of stress was high during the initial three years of study. Stress was significantly (p = 0.001) but negatively (r = −0.273) correlated with academic achievement.
[11]	- predictive analysis, model.	Negative correlation exists between age and the academic performance

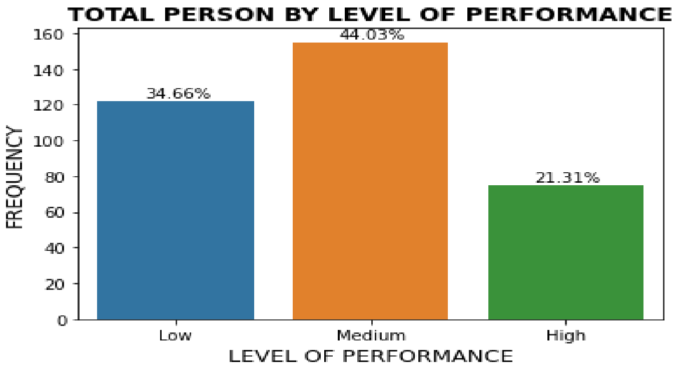

Fig. 2 Level of performance vs. Students' frequency

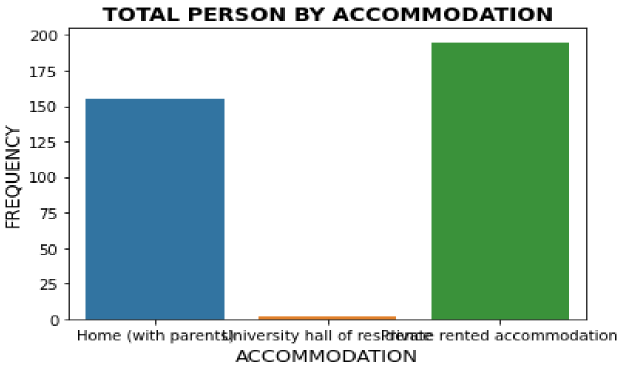

Fig 3 Types of accommodation vs. Students' frequency

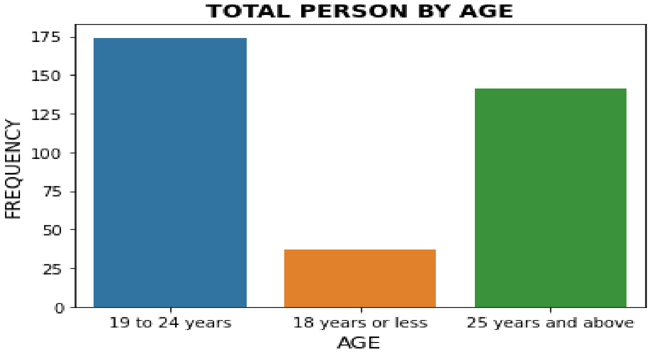

Fig. 4 Age range vs. Students' frequency

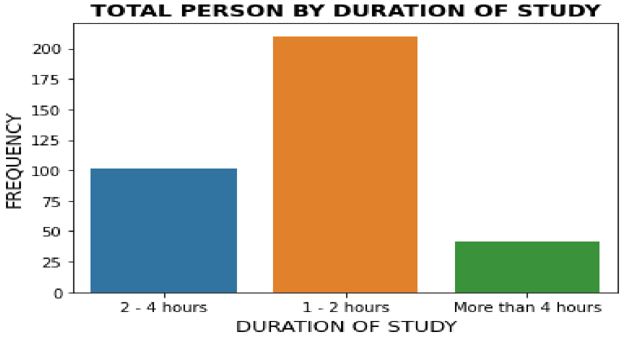

Fig.5 Duration of study vs. Students' frequency

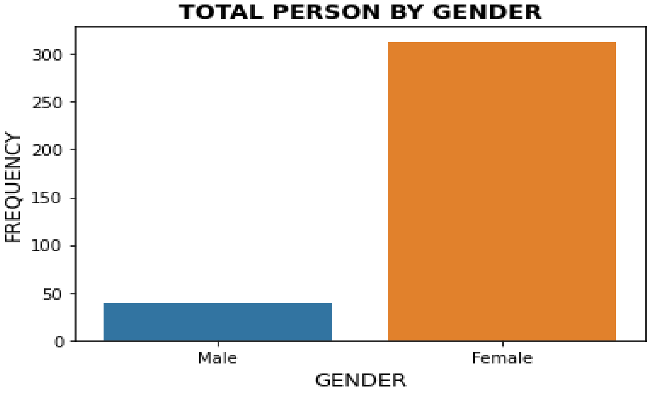

Fig.6 Gender vs. Students' frequency

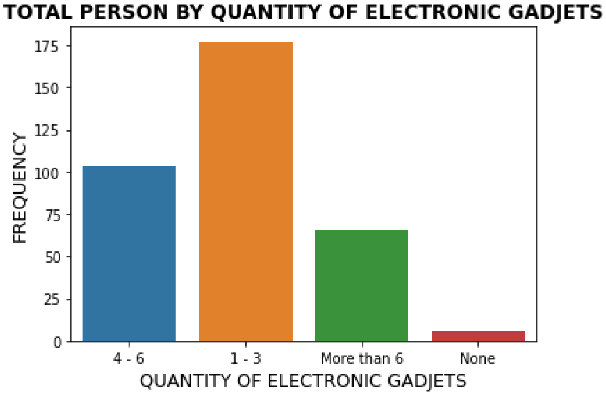

Fig. 7 Quantity of electric gadgets vs. students' frequency

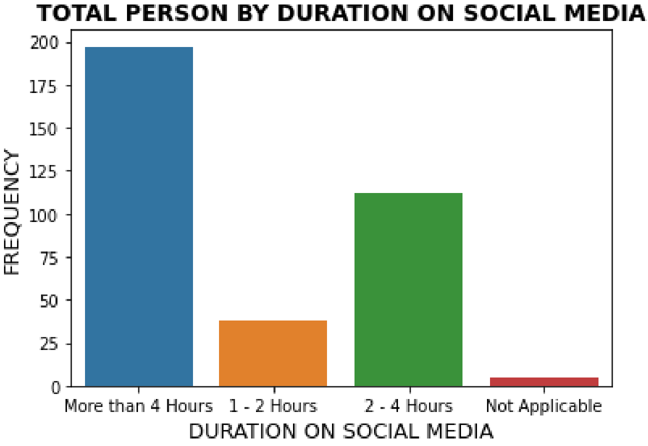

Fig.8 Duration spent on social media vs. Students' frequency

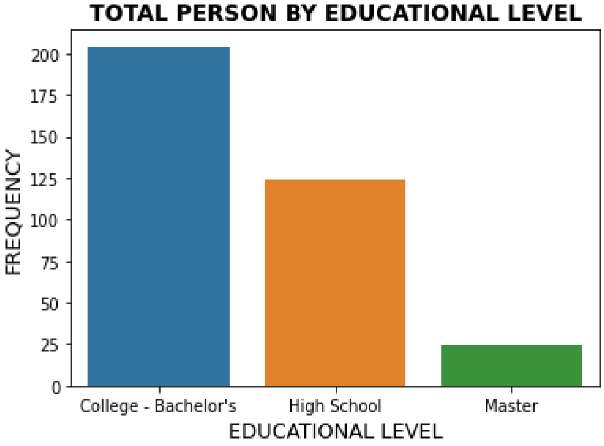

Fig. 9 Education level vs. Students' frequency

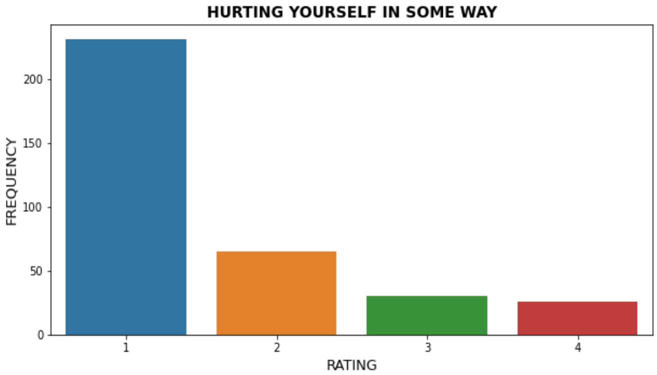

Fig. 10 Attribute: Hurting Oneself

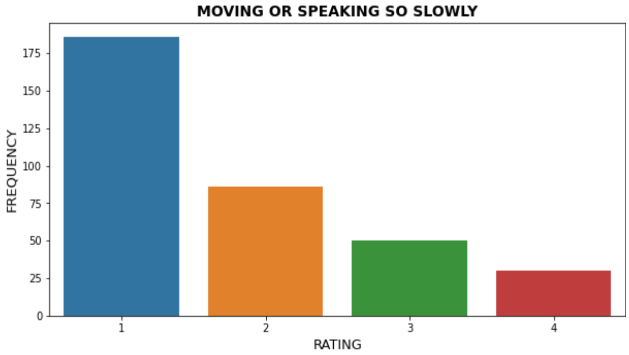

Fig. 11 Attribute: Moving or Speaking Slowly

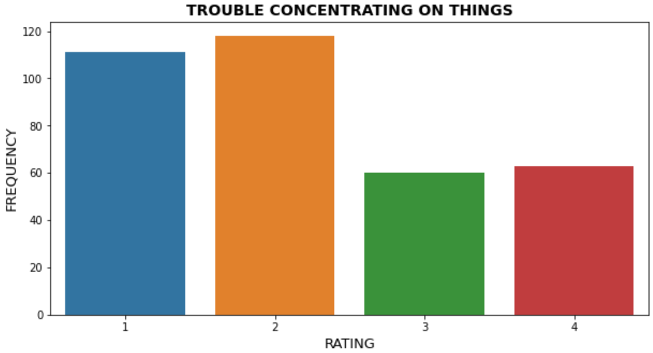

Fig. 12 Attribute: Concentration Issues

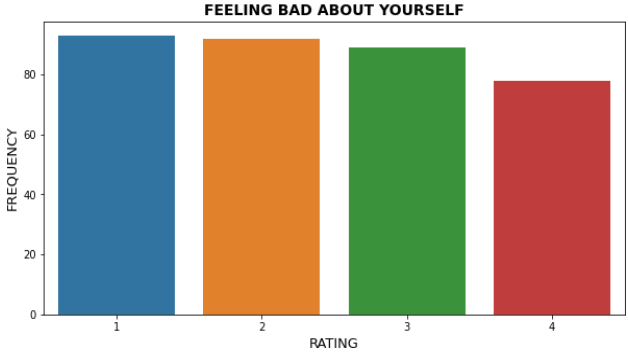

Fig. 13 Attribute: Self Worth

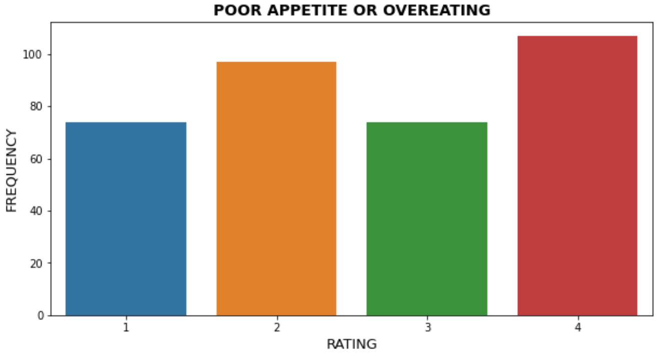

Fig. 14 Attribute: Eating Issues

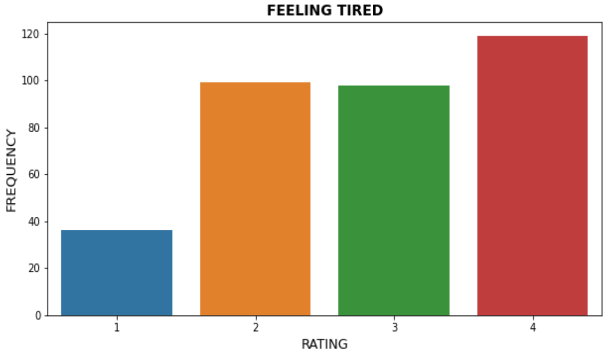

Fig. 15 Attribute: Exhaustion

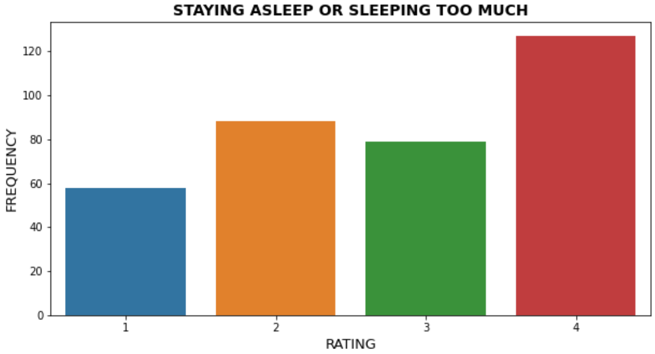

Fig. 16 Attribute: Sleeping Pattern

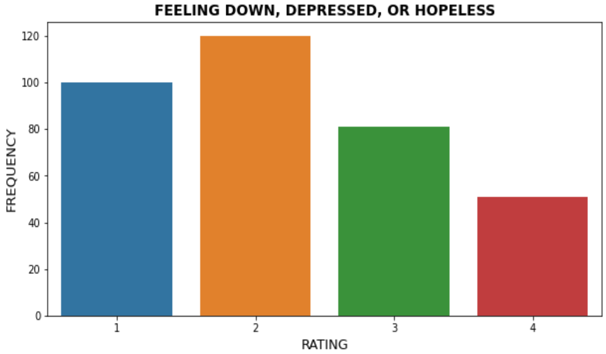

Fig. 17 Attribute: Emotional State

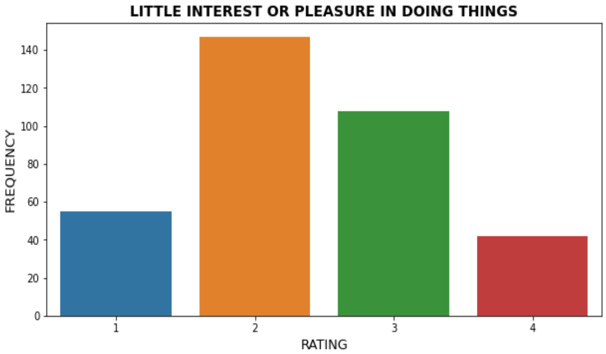

Fig. 18 Attribute: Motivation