**ANL 488 FINAL REPORT**

**Identifying Socioeconomic Factors that Affect the Academic Performance of Tertiary Students in SUSS**

**Logo

Description automatically generated**

**Submitted by**

**Meghna Ramanan**

**SCHOOL OF BUSINESS Singapore University of Social Sciences**

**Presented to Singapore University of Social Sciences in partial fulfillment of the requirements for the  
Degree of Bachelor of Science  
in Business Analytics**

**2023**

Abstract

This study aims to understand the relationship between various socioeconomic factors and how they affect the academic performance of tertiary students in SUSS. This was for the purpose of school administrations to identify weaker students faster to provide assistance and additional resources in order to raise graduation rates. Data from Near East University in Cyprus is used to train the model and tested on data collected from SUSS students. A wide range of socioeconomic factors were used to predict the target variable, the student’s cumulative GPA. 4 predictive models – CART, CHAID, Neural Network and Random Forest – were generated but showed very weak accuracy rates. Since the models were not able to perform well with the given datasets, further research was done to understand why this was the case. The primary issue that was identified was that the SUSS dataset that was used to test the models had a very small range of cumulative GPAs. Hence, the dataset implied that the varying range of socioeconomic factors do not affect academic performance as most students received the same cumulative GPA band. It was concluded that further studies have to be conducted using larger pools of students so as to get bigger datasets with wider range of GPAs. By doing so, the models would be able to better predict the cumulative GPAs of students.

table of contents

[Introduction 4](#_Toc150164630)

[Business Problem 4](#_Toc150164631)

[Business Analytics problem 4](#_Toc150164632)

[Literature Review 2](#_Toc150164633)

[Article 1 2](#_Toc150164634)

[Article 2 2](#_Toc150164635)

[Article 3 3](#_Toc150164636)

[Data Understanding and Preparation 4](#_Toc150164637)

[Data Understanding 4](#_Toc150164638)

[Data preparation 9](#_Toc150164639)

[Data Collection 9](#_Toc150164640)

[Modelling and Evaluation 11](#_Toc150164641)

[Test Case 11](#_Toc150164642)

[Cyprus and SUSS data models 12](#_Toc150164643)

[Analysis of results 13](#_Toc150164644)

[Recommendation and Conclusion 16](#_Toc150164645)

[References 17](#_Toc150164646)

[Annex A 18](#_Toc150164647)

# Introduction

## Business Problem

Education is a crucial part of the lives of young people who are aiming to find a place in the world. Obtaining an education is the fastest way to achieve social mobility to gain greater stability (Tan, 2018). However, even to achieve an education, socioeconomic factors play a part in whether students would be able to enter into universities and how they would perform there (McLaughlin & Sheridan, 2016). With universities wanting to maintain or boost their graduation rates, it is especially important to them that they are aware of students who may struggle to graduate in time. Hence, it is crucial that they are able to identify these students ahead of time so that they are able to provide more resources to such students.

## Business Analytics problem

This project is designed to identify socioeconomic factors that have close correlations with how students are performing in university in terms of their cumulative GPA. The aim of this project is to create a predictive model that would be able to highlight the relationship with certain background factors that students have that would indicate that they may fall behind in university.

The models generated from this study would be evaluated based on its accuracy rates, as well as its recall rate. Since the study aims for students to be picked out before they struggle and fall through the cracks in school, it is more imperative for the study to pick out the students who are weak even if it does pick out students who do not need help, incorrectly. Using these two measures, the most effective model will be chosen.

# Literature Review

## Article 1

Li et. al., (2020) conducted research on the impact of SES on the academic achievement of high school students, as well as the mediating effect of self-concept between the relationship between SES and academic achievement. This research was conducted on middle school students in China.

The methodology used to conduct this research is descriptive statistics, as well as hypothesis testing. Mediation analysis was also conducted to identify the correlation between SES and self-concept. SES is calculated by using a mediation model from 5 variables (each parents’ education level, each parents’ occupation, and the annual house income)

The results showed that there is heavy correlation between SES and the academic achievements of the students. It was also noted that self-concept and SES were also rather highly correlated and self-concept did have a mediating effect on the impact SES had on academic achievement. However, while using self-concept to mediate the impact, it must be noted that while the direct impact of SES on academic achievement did fall slightly while still remaining significant. The impact self-concept had on academic achievement is also significant.

These results indicate that there is strong evidence to back the proposed topic of research, to better understand the impact socioeconomic background has on a student’s performance in academics. It also highlighted another point that can be taken note as a gap in this project, which is the psychological aspect of a student, as measured by the self-concept measure, which would also have an impact on a students’ performance.

## Article 2

Chytrý et al. (2022) conducted a research project on the effects of SES on university students’ performance during the pandemic period. This article places emphasis on the variables that are impacted by the socioeconomic status of the students’ families and are exacerbated by the pandemic. For example, the access to internet and having infrastructure in place at home to conduct at home learning is a variable that is affected by SES and this actively impacts the academic performance of the student.

This research paper uses hypothesis testing to test the effects of variables on student performance. The focus on the analysis of variance using the F test and the least significant difference is used as the primary method of analysis.

## Article 3

Marks et al. (2001) conducted a research project to better understand the impact of socioeconomic factor as well as the impact of the school that students attend on their performance academically. This research took into consideration ethnicity, indigenous status, gender and many other such factors. The analysis of the data was done through a few different methods, such as basic statistical analysis using box and whisker plots to analyse the spread of the ENTER scores that students attain and estimating the range of means to a 95% confidence interval. Additionally, regression analysis was also carried out to understand the correlations between the factors themselves, as well as the impact that the variables have on the academic performance of the students.

The results of the research paper show that socioeconomic backgrounds do have an impact on the academic performance of students. One of the factors that the paper considers is the type of work that the parents of the students do. When mapped out on a box and whisker plot, it shows that the students with parents who do white collar jobs and are high ranking executives score on the higher end of the scale while students who have parents who do more blue collar jobs have a wide range of scores spanning almost the entire scale.

# Data Understanding and Preparation

## Data Understanding

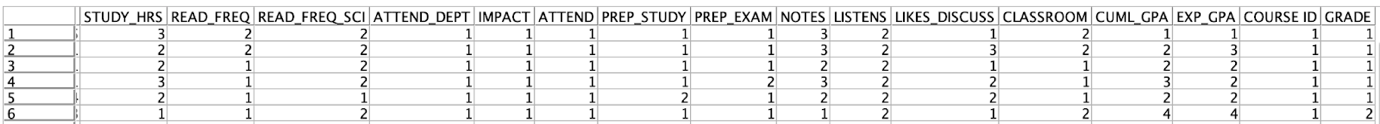
To undertake this project, the data that was collected from Near East University in Cyprus in 2019 will be used as a basis. While it would be ideal to conduct this project with data that is collected from students of SUSS, due to the time consuming nature of data collection through surveys and the sheer amount of data that would need to be collected to generate high accuracy models, the next best alternative was chosen.

Using this dataset still served as an adequate replacement for collecting data as Cyprus is a first world country. With a GDP per capita of over USD 30,000, Cyprus is a well-developed nation with a strong economy (World Bank, 2023). Since its development is close enough to Singapore’s, Additionally, considering that Cyprus’ strong education system is similar to Singapore’s, it would be possible to use the data collected from both locations to compare and study (Cyprus Profile, 2022). Lastly, since the data from Cyprus was collected relatively recently, it can still be used in comparison with data collected in SUSS at this time.

This dataset was deemed fit to be used in the project due to how recently the data was collected. Since it has only been 5 years since the data was collected, there are fewer issues such as outdated system of education and differences in difficulty of tertiary courses offered. Secondly, Cyprus us a well-developed nation and is considered a first world country with a GDP per capita of over USD 30,000. Lastly, Cyprus has a similarly structure education system compared to Singapore and hence, allows us to more easily compare the countries. Having considered these criteria, the dataset from Cyprus was used in this project.

Hence, the data from Cyprus is used as a training set for the models that would be generated. These models would be tested using data that would be collected from a small set of data that would be collected, approximately 40 records, from SUSS students.

The dataset from Cyprus consists of 33 fields, as shown below.   
  

*Fig. 3.1: Snapshot of fields in Dataset*

The table below shows a description of the fields of data, as well as the values that each field can take.

*Table 3.2: Data Description*

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Field Name** | **Description** | **Values it can contain** |
| 1 | STUDENTID | Student ID number | STUDENT1, STUDENT2, STUDENT3, etc. |
| 2 | AGE | Age of student | 1: 18-21 2: 22-25 3: above 26 |
| 3 | GENDER | Gender of student | 1: female 2: male |
| 4 | HS\_TYPE | Type of High school that student attended | 1: private  2: state  3: other |
| 5 | SCHOLARSHIP | If the student received a scholarship, and how much of the tuition is covered under the scholarship | 1: None 2: 25% 3: 50% 4: 75% 5: Full |
| 6 | WORK | If the student works | 1: Yes 2: No |
| 7 | ACTIVITY | Does the student have regular artistic or sports activity | 1: Yes 2: No |
| 8 | PARTNER | Does the student have a partner | 1: Yes 2: No |
| 9 | SALARY | Monthly salary | 1: USD 135-200 2: USD 201-270  3: USD 271-340 4: USD 341-410 5: above 410 |
| 10 | TRANSPORT | Mode of Transport to University | 1: Bus 2: Private car/taxi 3: bicycle 4: Other |
| 11 | LIVING | Accommodation type in Cyprus | 1: rental 2: dormitory 3: with family 4: Other |
| 12 | MOTHER\_EDU | Mother’s highest educational qualification | 1: primary school 2: secondary school 3: high school 4: university 5: MSc. 6: Ph.D. |
| 13 | FATHER\_EDU | Father’s highest educational qualification | 1: primary school 2: secondary school 3: high school 4: university 5: MSc. 6: Ph.D. |
| 14 | #\_SIBLINGS | Number of siblings | Ordinal data field - takes any integer value |
| 15 | PARENTS | Parents’ marriage status and presence in student’s life | 1: married 2: divorced 3: died - one of them or both |
| 16 | MOTHER\_JOB | Mother’s occupation | 1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other |
| 17 | FATHER\_JOB | Father’s occupation | 1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other |
| 18 | STUDY\_HRS | Number of hours spent studying weekly | 1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours |
| 19 | READ\_FREQ | Frequency for reading non-scientific journals/books | 1: None, 2: Sometimes, 3: Often |
| 20 | READ\_FREQ\_SCI | Frequency of reading scientific journals/books | 1: None, 2: Sometimes, 3: Often |
| 21 | ATTEND\_DEPT | Attendance to the seminars/conferences related to the department | 1: Yes 2: No |
| 22 | IMPACT | Impact of your projects/activities on your success | 1: positive 2: negative 3: neutral |
| 23 | ATTEND | Attendance to classes | 1: always 2: sometimes 3: never |
| 24 | PREP\_STUDY | Who the student prepares for midterm exams with | 1: alone 2: with friends 3: not applicable |
| 25 | PREP\_EXAM | When the student prepares for midterm exams | 1: closest date to the exam 2: regularly during the semester 3: never |
| 26 | NOTES | Taking notes in classes | 1: never 2: sometimes 3: always |
| 27 | LISTENS | Listening in classes | 1: never 2: sometimes  3: always |
| 28 | LIKES\_DISCUSS | Student feels that discussion improves their interest and success in the course | 1: never 2: sometimes 3: always |
| 29 | CLASSROOM | Flip-classroom | 1: not useful 2: useful 3: not applicable |
| 30 | CUML\_GPA | Cumulative grade point average in the last semester (/4.00) | 1: <2.00 2: 2.00-2.49 3: 2.50-2.99 4: 3.00-3.49 5: above 3.49 |
| 31 | EXP\_GPA | Expected Cumulative grade point average in the graduation (/4.00) | 1: <2.00 2: 2.00-2.49 3: 2.50-2.99 4: 3.00-3.49 5: above 3.49 |
| 32 | COURSE\_ID | Course Type | Nominal data field - values 1 to 9 |
| 33 | GRADE | Final Grade | 0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA |

## Data preparation Based on the types of values each field holds, the data types of each field needs to be changed as shown in Fig 3.2. Additionally, the fields “CUML\_GPA”, “EXP\_GPA” and “GRADE” have the role of target as these are the fields we would be using to verify if the model is able to accurately predict students’ academic performance. There are 3 targets in this dataset so as to allow for greater variation in the modelling process, where more models can be explored by using different targets in the modelling process. This would enable the best possible model with the highest accuracy to be obtained.

Additionally, since this dataset does not have any missing values or anomalous values, the data does not need to be cleaned.

## Data Collection

In order to conduct data collection, an online survey was sent out to SUSS students. The goal of the study was to get at least 40 responses, so that the Cyprus study would be used as the training data and make up 80% of the total dataset, and the SUSS dataset would make the remaining 20% testing dataset.

The survey was sent out to the student body targeting those in their 3rd and 4th year of study as it would have a more representative of the impact that socioeconomic factors have on their cumulative GPA as opposed to the GPAs of first and second year students. The impact of their socioeconomic background and the study habits of students in their last years of study would have a more consistent impact on their grades and GPAs. Students from various majors were also used in this study so that they results were not skewed due to the responses collected from only one major.

In the Cyprus study, they used Engineering and biomedical students only. However, since we do not have equivalent degrees in SUSS, the ideal was to proceed with data collection was to have a wide variety of majors represented in the dataset so that the choice of major does not impact the results of the study directly.

The survey followed the format of questions asked in the Cyprus dataset almost exactly, so that the datasets would be comparable. Some questions had to be tweaked slightly, such as the question asking respondents about their monthly salary. Since the field contains records using USD, in order to ask a similar question to the SUSS students to fill up the same field, some calculations needed to be performed to convert the options presented in the Cyprus data set to SGD.

In the Cyprus dataset, the income levels were split into 5 different ranges, as shown below:

* 1: USD 135-200
* 2: USD 201-270
* 3: USD 271-340
* 4: USD 341-410
* 5: above USD 410

Based on information database Expatistan (2023), in order to maintain the same standard of living as in Cyprus, it is 52% more expensive in Singapore. Using this ratio as the comparison rate and by converting the USD into SGD amounts, the ranges of the options were converted as shown below:

* 1: SGD 380 - 560
* 2: SGD 561 - 760
* 3: SGD 761 - 950
* 4: SGD 951 - 1150
* 5: above SGD 1151

# Modelling and Evaluation

## Test Case

In order to first identify the relationship between socioeconomic factors and the academic performance of students in Cyprus, we conduct a test case using the Cyprus dataset. In this scenario, we use the Cyprus dataset to train and test the models to identify the relationship between socioeconomic factors and academic performance in the students in Cyprus.

A picture containing text, font, screenshot, number

Description automatically generated

*Fig 3.1.1: CART model test case results*

A picture containing text, font, screenshot, number

Description automatically generated

*Fig 3.1.2: Neural Network model test case results*

A picture containing text, font, screenshot, number

Description automatically generated

*Fig 3.1.3: CHAID model test case results*

A picture containing text, font, screenshot, number

Description automatically generated

*Fig 3.1.4: SVM model test case results*

A picture containing text, font, screenshot, number

Description automatically generated

*Fig 3.1.5: Random Forest model test case results*

The above figures show the accuracy rates produced by each of the models. The neural network model has the lowest accuracy rate while the SVM model has the highest accuracy rate. However, in the case of the SVM model, since it has an accuracy rate of 100%, it can be inferred that the model is overfit based on the dataset. This indicates that the model has fit itself to the training data too well, such that it would not be able to classify unseen data well, thus defeating the purpose of the model (IBM, n.d.). Thus, the model that has the highest accuracy would be the Random Forest model.

Based on the results of the Random Forest model, it is possible that using the socioeconomic indicators of a student, the models are able to predict the range in which the student’s cumulative GPA falls under. This shows that there is some relationship between socioeconomic indicators and students’ cumulative GPA.

## Predictive Models

Based on the test cases, since the SVM model has been overfit, it would not be able to predict the results of the unseen data accurately and hence has been removed as a possible model to deploy. In order to conduct this project, the models have to be generated using the Cyprus data as the training data and the SUSS data as the testing data. The figures below show the results of the CHAID, CART, Neural Network and Random Forest models generated using the training dataset and the testing dataset.

A screenshot of a table

Description automatically generated

*Fig 3.2.1: CHAID model results*

A screenshot of a document

Description automatically generated

*Fig 3.2.2: CART model results*

A screenshot of a document

Description automatically generated

*Fig 3.2.3: Neural Network model results*

A table with numbers and a number

Description automatically generated with medium confidence

*Fig 3.2.4: Random Forest model results*

Based on the results of the 4 models, it can be seen that the results of all 4 models are unable to reach high percentages of accuracy. Among the 4 models, the CART model has the highest accuracy but is still not useful to be deployed due to its high rate of inaccurate predictions.

However, in the context of identifying students who may need help in their education journey based on their socioeconomic status, it is more important that those who need help do not get falsely identified and fall through the cracks. Hence, using the recall percentage, which refers to the total number of True Positives divided by the sum of True Positives and False Negatives, each model can be evaluated.

In order to use the recall formula, the confusion matrix needs to be read as a binary. The number of students who are likely to need additional help would be those who have a cumulative GPA in band 1 (below 2.00) or band 2 (below 2.50).

However, using the formula on such a small sample size proves to be futile as all 4 models have the same recall percentage of 50%. This recall method would prove to be much more useful with a bigger dataset with a greater range of cumulative GPAs. Therefore, due to the ineffectiveness of all 4 models, we are forced to conclude that none of these models would function well if deployed to identify weak students.

## Analysis of results

From the results of the models generated, it is apparent that the models are unable to accurately predict the cumulative GPAs of the students based on the socioeconomic factors. If deployed, it is unlikely that these models would likely not be able to provide actionable results for college administrations to work with to deploy additional support and resources to select students.

One probable reason for this would be that socioeconomic factors are not a good predicting factor of academic performance of students in SUSS. While previous research from various parts of the world have shown the strong link between socioeconomic background of students and their performance in school, this may not be the case in SUSS. From the data collected from SUSS students, it can be seen that the students are from varying backgrounds.

For example, from Figure 3.3.1, it can be seen that the students who have been surveyed are from varying age groups, indicating that they are in various stages in their lives and that each individual’s priorities at that stage in life would be different, and their focus on education and learning varies based on their age (Capotosto, n.d.).

A pie chart with numbers and a few percentages

Description automatically generated

*Fig 3.3.1: Age distribution of respondents*

As can be seen in Figure 3.3.2, the salary or allowance of students also largely varies, indicating that the economic status of students ranges vastly.

A pie chart with numbers and a number of percentages

Description automatically generated

*Fig 3.3.2: Income distribution of respondents*

To narrow in on their social status, both parents’ education level would be a good gauge of the type of family the students have grown up in and an indirect indicator of their economic status. From Figures 3.3.3 and 3.3.4, it can be seen that students come from very different backgrounds with parents who are extremely educated and some with parents who have not received much education.

A colorful pie chart with numbers

Description automatically generated

*Fig 3.3.3: Mother’s education level distribution of respondents*

*A colorful pie chart with white text

Description automatically generated*

*Fig 3.3.4: Father’s education level distribution of respondents*

Students have also followed varying educational routines, as can be seen from the differing amount of time spent on studying per week in Figure 3.3.5, and whether they engage with scientific literature in their own time in Figure 3.3.6.

A colorful pie chart with numbers

Description automatically generated

*Fig 3.3.5: Hours studying per week distribution of respondents*

*A pie chart with numbers and a few percentages

Description automatically generated*

*Fig 3.3.6: Reading scientific literature distribution of respondents*

Even though the students surveyed have come from vastly different socioeconomic backgrounds, the results of their cumulative GPAs do not have a wide range. Figure 3.3.7 shows the distribution of cumulative GPAs among the respondents of the survey.

A green circle with red and orange circles

Description automatically generated

*Fig 3.3.7: Cumulative GPA distribution of respondents*

As can be seen from the chart, the distribution in cumulative GPAs is concentrated in 2 bands, the 3.00 to 3.49 band and the 3.50 to 4.49 band. Due to the lack of distribution in the GPAs and the small dataset, the models were unable to predict GPAs accurately. The lack of distribution in GPAs implies that the differing socioeconomic backgrounds and habits the students follow have little to no effect on their academic performance.

While previous research around the world has indicated the link between socioeconomic factors and academic performance, this result shows that that is not necessarily the case in SUSS. This could be due to a myriad of reasons such as the importance placed on education by the country and its government and the equalising forces put in place during the early years of education, such as local schools following consistent syllabi and having the same resources, ensuring that each student receives the same quality of education no matter which school they are in. In addition, policies have been put in place to aid students from disadvantaged backgrounds to access additional support and resources so as to level the playing field.

# Recommendation and Conclusion

While this study was not successful in identifying the factors that influence and affect student performance highly, it has provided some insight into how SUSS students have been able to even the playing field such that despite their varied backgrounds, their cumulative GPAs do not vary much . This study has shown that the impact of socioeconomic factors on education varies based on location as well, as can be seen by the difference in impact of factors between the students in Cyprus and in SUSS.

Further studies can be conducted to understand the extent of the impacts of these factors. Since the data collected from SUSS students was a very small sample, additional works can be done using larger datasets consisting of only SUSS students. This dataset can be used to train and test the model so that the model is more representative of SUSS students’ data, providing a clearer understanding of the impacts of various factors on academic performance of SUSS students. In doing so, a wider range of cumulative GPAs would be obtained, allowing the impact of socioeconomic factors to be identified.

Additionally, to verify if socioeconomic factors do have an impact on students’ academic performance, other methods such as non-linear regression and analysis of variance can be conducted. This would allow us to understand the exact extent each factor has an impact on academic performance.

# References

Education. Cyprus Profile - Cyprus’ Trade &amp; Investment Business Portal. (n.d.). https://www.cyprusprofile.com/sectors/education-research#:~:text=Top%2DClass%20Education&amp;text=Despite%20its%20small%20size%2C%20Cyprus,in%20the%20last%20five%20years.

Expatistan. (2023, November 6). Cost of living comparison cyprus vs Singapore. Expatistan, cost of living comparisons. https://www.expatistan.com/cost-of-living/country/comparison/singapore/cyprus

GDP per capita, PPP (current international $) - cyprus. World Bank Open Data. (n.d.). https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?locations=CY

Li, S., Xu, Q., & Xia, R. (2020). Relationship between SES and academic achievement of junior high school students in China: The mediating effect of self-concept. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02513

Marks, G., McMillan, J., & Hillman, K. (2001). Tertiary entrance performance : the role of student background and school factors. *Lsay Research Reports*.

McLaughlin, K. A., & Sheridan, M. A. (2016). Beyond cumulative risk: a dimensional approach to childhood adversity. Current Directions in Psychological Science, 25, 239-245. doi:10.1177/0963721416655883

Ramzai, J. (2020, June 20). Simple guide for top 2 types of decision trees- CHAID &amp; Cart. Medium. https://towardsdatascience.com/clearly-explained-top-2-types-of-decision-trees-chaid-cart-8695e441e73e

Singapore Urban Poverty Line: Education inequality. Singapore Urban Poverty Line | Education Inequality. (n.d.). https://www.the-best-of-you.com/urban-poverty/education-inequality

Tan, T. (2018, November 1). *Singapore system has done better than other countries in improving social mobility: MSF Paper*. The Straits Times. Retrieved February 13, 2023, from https://www.straitstimes.com/singapore/singapore-system-has-performed-better-than-others-in-improving-social-mobility-but-will

What is overfitting?. IBM. (n.d.). https://www.ibm.com/topics/overfitting

Young adults. National College Transition Network. (2018, March 28). https://www.collegetransition.org/student-supporters/research-briefs/young-adults/

# Annex A

Survey sent out to SUSS Students

A screenshot of a phone

Description automatically generated

A screenshot of a phone

Description automatically generated A screenshot of a survey

Description automatically generatedA screenshot of a survey

Description automatically generated A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated