Classifying Student Grades Based on Socioeconomic Background

Ahmad Zaidan bin Adnan
(1718733)
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
zaidan.adnan@live.iium.edu.my

Muhamad Safwan bin Rahumatulla
(1722491)
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
safwan.rahumatulla@live.iium.edu.my

Muhammad Nur Aqmal bin Khatiman
(1712719)
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
aqmal.khatiman@live.iium.edu.my

Muhammad Salikin bin Ismail
(1711083)
Department of Computer Science
International Islamic University Malaysia
Selangor, Malaysia
salikin.ismail@live.iium.edu.my

ABSTRACT

In this paper, a Decision Tree algorithm was implemented to predict the student performance based on their socioeconomics status. The analysis was conducted by using a dataset from supplementary materials of (Sulaiman, Akhir, Hussain, Jamin & Ramli, Data on the impact of socioeconomic status on academic achievement among students in Malaysian public universities, 2020). Before predicting the performance, we did some data pre-processing which is treating the missing values and outliers and proceeded with Exploratory Data Analysis (EDA) and feature selection. Orange was used to develop the model. For the performance evaluation, several metrics in ordinal classification are applied namely: Classification Accuracy(CA), F1 Score, Precision and Recall). As for the result, after the model was used to predict values for the test set, the predicted values were compared with the actual values and the performance measures were calculated.

Keywords—Decision Tree, Orange, Academic Performance, Socioeconomic

INTRODUCTION

Poverty is a common phrase that has been used when describing the current issues of the world. Poverty is the cause of many problems occurring in the world today economically, socially and politically. According to Anggraeni et. al [1], poverty is the failure of a person to fulfil essential needs, such as food, clothes, jobs, health and shelter. The causes that relegate an individual into the levels of poverty (different countries provide different indicators for the levels) are various whether it is happening internally or externally. Among the causes are unemployment, low wages, discrimination, lack of education, government benefits, high cost of living as well as inflation [2]. In 2015, 193 world leaders alongside the United Nations have decided to adapt the eradication of poverty among society as one of the 17 goals for the Sustainable Development Goals 2030 (SDG 2030) since poverty is affecting many lives around the world and hinders them from opportunities that should be benefited by everyone throughout the society [3]. Globally, about 500 million people or 6.5 % of the global population earn $ 1.90 or less per day which is the poverty line for global poverty as a whole [4]. In Peninsular Malaysia, however, 0.7 % of the household population live in poverty according to Vaziri et.al [5]. Since poverty is the primary cause for many issues today, it is important to bring the community together in order to discuss possible solutions towards this main problem for the benefits of the future generation.

Poverty, without any doubt, can influence many elements in a human life. One such element is academic performance of students. With limited resources, students who live in poverty have difficulties in getting better education opportunities. In order to address this issue, a study must be conducted to find the correlation between household income and academic performance of students in Malaysia, which is the research problem for this project. For the project, we will use the available resources to execute the process including open source data and machine learning methods to develop a model for the prediction of poverty and its influence towards academic performance. By using such methods, we would be able to analyse the impact of these economic factors towards the success or decline in academic results. The dataset that will be used are focused on the socioeconomic status and the academic achievements of students from a Malaysian public university. The data that is included in the dataset is the students' GPA, their parents' income level, socioeconomic status and more. In order to forecast the effects of income levels towards academic achievement, we will develop decision tree models.

BACKGROUND

According to Li, Xu and Deng [6], decision trees are one of the powerful methods commonly used in various fields, such as machine learning, image analysis, and identification of patterns. Decision trees are strong due to improved classification comprehensibility in terms of extraction from feature-based samples. Moreover, decision trees have not only been shown to be effective in many areas, but also have fewer parameters. In the context of constructing decision trees, two key rules are considered. One is the stopping criterion for deciding when tree growth can be stopped and leaf nodes are produced. The other is how class labels in leaf nodes can be allocated. The first rule suggests that since all samples belong to the same class, the growth of the tree should be ended. The second rule stresses the value of setting a threshold. We choose decision tree algorithms because these algorithms are the easiest to read and interpret without requiring statistical knowledge. In other words, we can have an insight on the probabilities of the specific decision tree in the view of a new audience.

RESEARCH QUESTIONS

There are several questions that are going to be addressed in this project. The research questions are as follows:

1. What effect does poverty have on the academic performance of students ?
(Diagnostic)
2. Is the number of coursework will help students perform well in their studies ?
(Diagnostic)
3. Which ethnicity and gender has the highest achievements ?
(Descriptive)
4. Is gender is a factor that can affect performance of a student in his or her study ?
(Diagnostic)
5. What is the common CGPA of a student that comes from a high income family ?
(Predictive)

RESEARCH OBJECTIVES

There are several objectives that are targeted to be achieved in this project. The objectives are as follows:

1. To produce an insight on the correlation between poverty and students' academic performance in Malaysia.
2. To analyse how students' academic performance are affected by the socioeconomic status of their family.
3. To predict the academic performance of students who are affected by poverty in the future using machine learning models.

RESEARCH SIGNIFICANCE

This project aimed will give several impacts to the community such as:

- The prediction model will provide an insight to the education institution to make an early action to help the needed student financially.
- The educational institution can identify the students who need extra support.
- Trend of the academic performance can be observed from the predictive model to improve and enhance student performance in the future.
- The educational institution can get an idea to make an initiative to boost the student performance.

LITERATURE REVIEWS / RELATED WORKS

There are several studies and researches in the domain of student academic performance and its factors published in the last few years. Most papers analyse the factors that influence the performance and make use of several machine learning models to get an insight of the trend and pattern. 10 papers of related works are listed as shown in table 1.

There are numerous ways to evaluate student performance. Every paper has distinct variables to represent good and bad performance of the students. However, the GPA of the student has been used by several studies to be a target variable to predict student's performance. As implemented by [7][8], where dependent variables such as past grades have been used to predict the GPA.

Some studies only classify the student performance by high-performing and poor-performing. This binary classification has been used in the paper by [9]. Students were initially classified into a specific group based on semester grade in the first year and observed the changes to the group based on the performance in four years.

Several Machine learning models have been used to predict student performance. In this domain, the Naive Bayes, Regression, Decision tree and Artificial Neural Network are the most common. Those machine learning models were great in dealing with classification problems. Most of the studies listed in the table make use of this algorithm to predict and classify the student performance based on various variables such as social background, GPA, grade and likelihood to complete graduation. There are various attributes and factors that have been elicited to see its impact on students. The common attributes and factors observed in the papers are past performance, student current grade, course difficulties and family background.

Therefore, in our project we use CGPA as an indicator to represent the performance of the student. The CGPA above 3.0 can be considered as a well-performing student and above 3.5 are excellent, while below 3.0 are considered as poor-performing.

The ANN model is one of the most popular machine learning algorithms among the listed papers. The model changes weights of the connected neuron to predict the correct target label for input. The ANN backpropagation technique allowed the model to be more resistant on noisy datasets and perform well in predicting patterns that have not been trained before. Fully connected multilayer feed forward ANN have an input layer, several hidden layers and the output layer. The papers in [10][11] have implemented this model in their paper by defining the model with one input layer, two hidden layers and one out layer. The activation function used for hidden units was the Rectifier Linear Unit.

The decision tree model has an internal node that represents a test on attribute and tree branch represents the test outcome. Leaf nodes represent the target attribute, grades and the root node represented by the upper first node in the tree. The paper published by [12] used this model to classify the student fail or pass on final grade based on several variables.

The logistic regression model has been used in several research to describe the relation between several independent variables such as class performance, attendance, assignment, lab work and many more [13]. The regression model was able to describe the probability of students to fail which is always a value between 1 and 0. The implementation can be observed in the paper by [14][15] where GPA was predicted trend and pattern can be obtained by analyzing influences of various courses.

On the other hand, the Naive Bayes classification model was used by several papers. The algorithm can be considered as the simplest variation of the Bayesian network which assumes every attribute independent from other target attributes.

No.	Year	Authors	Research Problem / Application	Main techniques	Results	Future works
1	2019	Hussein Altabrawee, Osama Ali, Samir Qaisar Ajmi	Identify student who need extra support and taking appropriate actions using machine learning model	Decision Tree, Naive Bayes, Artificial Neural Network(ANN), Logistic Regression	ANN model recorded the highest accuracy among others. However, the Logistic Regression model got higher precision compared to ANN. The models show that not every attribute has an impact on predicting the achievement of the students.	
2	2015	Havan Agrawal, Harshil Mavani	Identify high-risk students and its features which affect the performance of students.	Artificial Neural Network (ANN)	The model recorded accuracy rate at 70.48% and the analysis shows that past performance greatly influences student's performance.	
3	2019	Diego Buenano-Fernandez, David Gil, Sergio Lujan-Mora	Predict final grades of students based on their historical performance.	Decision Tree	The decision tree model has recorded 96.5% precision on forecasting future grades of the students.	
4	2017	Junshuai Feng	Predict student performance in the context of Educational Data Mining (EDM) and identify the relationship between the attributes.	Decision Tree Artificial Neural Network (ANN)	Both models perform well in making correlation between the variables. ANN recorded 97.9% accuracy compared to 94.1% by Decision Tree classifier.	Increase the complexity and use more meaningful data. Apply some realistic educational data.
5	2019	Lubna Mahmoud Abu Zohair	Proving efficiency of Support Vector Machine (SVM) and learning discriminant analysis model in predicting student's performance.	Support Vector Machine (SVM)	The prediction accuracy for the SVM model is high. The main key indicators for predicting student grade were defined.	
6	2017	Jie Xu, Kyeong Ho Moon and Mihaela van der Schaar	Build a system to keep track of students' academic performance and predict their final GPA and likelihood to graduate by observing their current progress.	Linear Regression, Logistic Regression, Random Forest, K-Nearest Neighbour(KNN), Ensemble-based Progressive	The Random Forest model performs the best, while KNN performs the worst in most cases. Predictive power recorded by courses in the same department are more compared to prerequisite	The performance prediction to elective courses can be extended. Recommend courses to

				Prediction (EPP)	courses.	the student using the prediction results.
7	2017	Ermiyas Birihanu Belachew, Feidu Akmel Gobena	Using machine learning techniques to analyze data in the student management information system.	Naive Bayes, Artificial Neural Network (ANN), Support Vector Machine (SVM)	The Naive Bayes model got the highest accuracy compared to other models at 95.7%.	Use more data to gain better understanding.
8	2018	Agoritsa Polyzou, George Karypis	Analysis on poorly performing students by formulating the problem. Gain insight on the factors that contribute to the performance.	Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting	The Gradient Boosting model was the best performing model in terms of AUC and F1 score. While, the decision tree model recorded the lowest performance. The difficulty of the course has more influence on the students to drop the course compared to the students' issues.	
9	2017	Raheela Asif, Agathe Merceron, Syed Abbas Ali, Najmi Ghani Haider	Predicting a student's academic achievement by observing social variables like age, sex, marital status, nationality and many more. Analyse low and high achieving students.	Decision Tree, Bayesian Information Criterion (BIC) modification of K-means algorithm, Random Forest 1-Nearest Neighbour	Naive Bayes reported to be the best classifier followed by the Random Forest model. The two classes of student, high-performing student and low-achieving student tend to remain in the same class during the four years.	Refine heatmaps to extract indicators of low and high performance without machine learning algorithms.
10	2016	Ahmad Mueen, Bassam Zafar, Umar Manzoor	Analyze and predict student's academic performance based on academic record and forum participation.	Naive Bayes, Artificial Neural Network (ANN), Decision Tree.	The predictive power, accuracy rate and precision of Naive Bayes was the highest among others. Lack of participants in on-line discussion forums was reported as the main factor that influenced the poor performance of the students. The high-performing students were active in forum discussion.	

METHODOLOGY

1. What effect does poverty have on the academic performance of students ?
(Predictive)

Upon inspection, the dataset contains features that are irrelevant to the research problem. Therefore, feature extraction needs to be done. In determining what features to keep and what to remove, a research was done to determine which socioeconomic factor affects academic performance.

For academic performance, CGPA was chosen as the target variable over GPA. This is because CGPA was more stable in determining the students performance during their time in the university over GPA, which is based on one semester's result.

Then, socioeconomic factors that contribute to poverty need to be chosen so that a comparison can be done between academic excellence and students who struggle financially to those who are not. It is noted that the standard for poverty differs from country to country. In Malaysia, poverty or low-income class is defined to those who are in category B40. B40 or Below 40 is defined as those who have household income of less than RM4,000. Therefore, household income is chosen to be a factor in the project. To add to that, there are also other characteristics of B40. Based on the research [20], dependency on only one income, parent's education level that is below SPM and dependency on one income only is considered to be characteristic of B40. As such, we have included those features in the project. The final feature and description are listed in the table 2:

Variable	**Factors**	**Description**
Dependent or Target variable	CGPA range	Range: ● 2: CGPA between 2-2.5 ● 3: CGPA between 2.5-3 ● 4: CGPA between 3-3.5 ● 5: CGPA between 3.5-4
Independent	Household with one working parent	● 1: Household have only one working parent ● 0: Household have both working parent
	Father SPM Education	● 0: No SPM qualification ● 1: Have SPM qualification or above
	Mother SPM Education	● 0: No SPM qualification ● 1: Have SPM qualification or above
	Household income category	● 1: Household income less than RM4,000 ● 2: Household income more than RM4,000 ● 3: Household income more than RM10,000

2. Is the number of coursework will help students perform well in their studies ?
(Diagnostic)

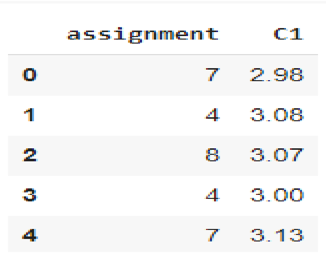
For this research question, the raw dataset from the survey in University Malaysia Terengganu will be used.

```
df = pd.read_csv("data.csv")
```

```
df.head(5)
```

Since the raw dataset contains many unused columns, the unused columns will be dropped. Only the column "assignment" and "C1" will be used in this analysis.

```
df.drop(df.columns.difference(['assignment','C1']), 1, inplace=True)
```

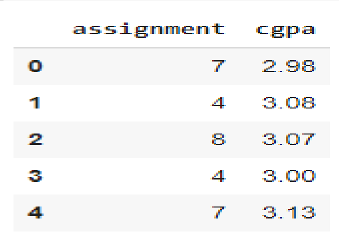


	assignment	C1
0	7	2.98
1	4	3.08
2	8	3.07
3	4	3.00
4	7	3.13

Figure 1

The name of column "C1" does indicate anything that shows the performance of students. Therefore, the column "C1" will be renamed to "cgpa" for better understanding of the whole dataset.

```
df = df.rename(columns={"C1": "cgpa"})
```



	assignment	cgpa
0	7	2.98
1	4	3.08
2	8	3.07
3	4	3.00
4	7	3.13

Figure 2

Duplicated rows will interrupt the validity of the analysis. As a solution, duplicated rows will be removed. The first step is to identify the number of duplicated rows.

```
duplicate_rows_df = df[df.duplicated()]
```

```
print("number of duplicate rows: ", duplicate_rows_df.shape)
```

```
number of duplicate rows:  (78, 2)
```

From the result, there are 78 duplicated rows in the dataset. Therefore, it is needed to drop the duplicated rows.

```
df = df.drop_duplicates()
```

```
assignment    333
cgpa          333
dtype: int64
```

The above values is the number of rows before dropping duplicate rows.

```
assignment    255
cgpa          255
dtype: int64
```

The above values is the number of rows after dropping duplicate rows.

Finally, in order to test the results, scatterplot will be developed to show the correlation between the rate of spending for courseworks and the CGPA of students.

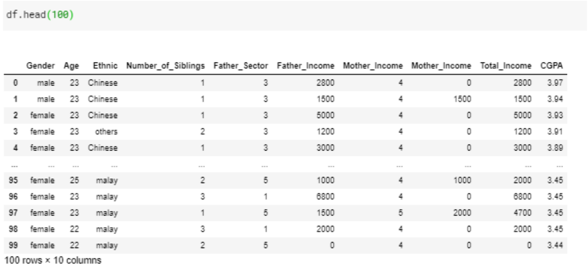
```
fig, ax = plt.subplots(figsize=(10,6))
ax.scatter(df['assignment'], df['cgpa'])
ax.set_xlabel('assignment')
ax.set_ylabel('cgpa')
plt.show()
```

3. Which ethnicity and gender has the highest achievements ?
(Descriptive)

The steps to solve this question is as below;

Once the data is read, we renamed all the columns

```
df.rename(
   columns={
      "gender": "Gender",
        "age": "Age",
        "number.of.siblings" : "Number_of_Siblings",
        "father.sector"      : "Father_Sector",
        "father.income" : "Father_Income",
        "ethnic" : "Ethnic",
        "mother.sector" : "Mother_Income",
        "mother.income" :  "Mother_Income",
        "total.income"  : "Total_Income",
        "cgpa" : "CGPA"
    },
    inplace=True
)
```



```
df.head(100)
```

	Gender	Age	Ethnic	Number_of_Siblings	Father_Sector	Father_Income	Mother_Income	Mother_Income	Total_Income	CGPA
0	male	23	Chinese	1	3	2800	4	0	2800	3.97
1	male	23	Chinese	1	3	1500	4	1500	1500	3.94
2	female	23	Chinese	1	3	5000	4	0	5000	3.93
3	female	23	others	2	3	1200	4	0	1200	3.91
4	female	23	Chinese	1	3	3000	4	0	3000	3.89
...	...	...	...	...	...	...	...	...	...	...
95	female	25	malay	2	5	1000	4	1000	2000	3.45
96	female	23	malay	3	1	6800	4	0	6800	3.45
97	female	23	malay	1	5	1500	5	2000	4700	3.45
98	female	22	malay	3	1	2000	4	0	2000	3.45
99	female	22	malay	2	5	0	4	0	0	3.44

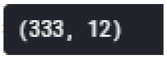
100 rows × 10 columns

Figure 3

All the variables are grouped into categorical and numerical respectively.

```
numerical = [
        'Age','Number_of_Siblings',   'Father_Sector',
'Father_Income', 'Mother_Income', 'Mother_Income',
'Total_Income','CGPA'
]
categorical = [
  'Gender', 'Ethnic'
]

df = df[numerical + categorical]
Df.shape
```



```
(333, 12)
```

Figure 4

We found out that we have 333 rows of numerical data and 12 rows of categorical data.

After this step, we use a boxplot to identify which gender and ethnicity have higher achievements in term of their cgpa

```
fig, ax = plt.subplots(3, 3, figsize=(15, 10))
for var, subplot in zip(categorical, ax.flatten()):
        sns.boxplot(x=var,  y='CGPA',  data=df,
ax=subplot)
```

We have also performed this step to further support our findings

i) Gender

```
        grouped_df   =   df.groupby(['Gender',
df.index]).agg({'CGPA': 'sum'})
        print(grouped_df)
```

ii) Ethnicity

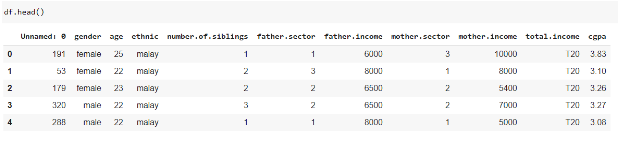
```
        grouped_df   =   df.groupby(['Ethnic',
df.index]).agg({'CGPA': 'sum'})
        print(grouped_df)
```

4. Is gender is a factor that can affect performance of a student in his or her study ?
(Diagnostic)

To address this answer, we will use the numpy package. The strategy to solve this question is to find the correlation between CGPA range and gender column. Firstly, the columns will be converted into category types. Then, we will use the numpy package to find covariance between the two variables. Also, to strengthen our hypothesis, we create the correlation matrix to find the relationship between gender, CGPA and CGPA range.

5. What is the common CGPA of a student that comes from a high income family ?
(Diagnostic)

In Malaysia, the social classes consist of three parts, B40, M40, and T20. For the sake of simplicity, we have decided to assign these classes with monthly income below RM 4,000, below RM 10,000 and lastly above RM 10,000 respectively. This method has been done using Excel.



```
df.head()
```

	Unnamed: 0	gender	age	ethnic	number.of.siblings	father.sector	father.income	mother.sector	mother.income	total.income	cgpa
0	191	female	25	malay	1	1	6000	3	10000	T20	3.83
1	53	female	22	malay	2	3	8000	1	8000	T20	3.10
2	179	female	23	malay	2	2	6500	2	5400	T20	3.26
3	320	male	22	malay	3	2	6500	2	7000	T20	3.27
4	288	male	22	malay	1	1	8000	1	5000	T20	3.08

Figure 5

By using python, we can find the proportion of the social classes with the income category of the students.

```
b = sns.countplot(x=df['total.income'])
b.axes.set_title('Division  of  social
class', fontsize = 20)
b.set_xlabel('Social Class', fontsize =
20)
b.set_ylabel('Count', fontsize = 20)
plt.show()
```

Next step we need to compare the overall cgpa by plotting another two graphs using seaborn package which are violin and swarm plot. The code is as follows.

```
plt.figure(figsize=(10,6))
sns.violinplot(x=df['total.income'],y=d
f['cgpa'],split=True)
sns.despine(left=True)
plt.show()

sns.swarmplot(x=df['total.income'],y=df
['cgpa'])
plt.show()
```

The final results are shown in the Results section.

RESULTS

The results for all proposed research questions are as follows:

1. What effect does poverty have on the academic performance of students ?
(Predictive)

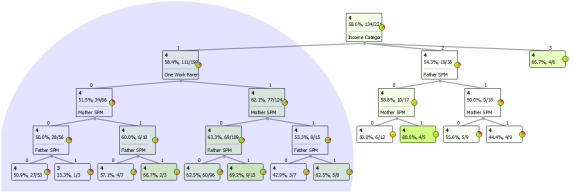


Figure 6

The figure shows the predicted CGPA range of B40 and non-B40 to be in the same range, which is4 (CGPA of 3.0-3.5). In particular, the highlighted blue area, which is B40 group is predicted that their CGPA to be mostly in range 4 and only one is predicted to be in range 3 (CGPA 2.5-3.0). In conclusion, the model stated that there is no difference in academic performance between those in poverty or not in poverty.

2. Is the number of coursework will help students perform well in their studies ?
(Diagnostic)

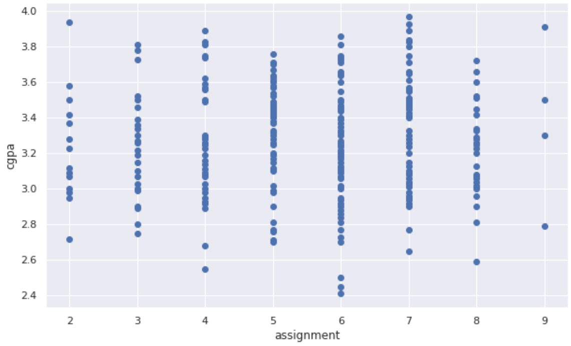


Figure 7

The figure shows the scatterplot of the rate of spending for courseworks and the CGPA of students. From the scatterplot, it can be seen that the rate of spending for courseworks does not affect the CGPA of students. The highest rate has students with low CGPA. The lowest rate also has students with high CGPA. Therefore, the number of courseworks (the rate of spending for courseworks) will not help students perform well in their studies.

3. Which ethnicity and gender has the highest achievements ?
(Descriptive)

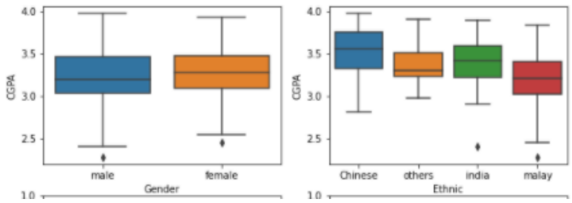
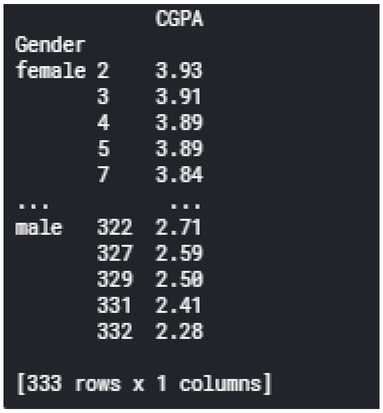


Figure 8

We found out that female students have higher achievements than male students. This is further proved with the analysis below

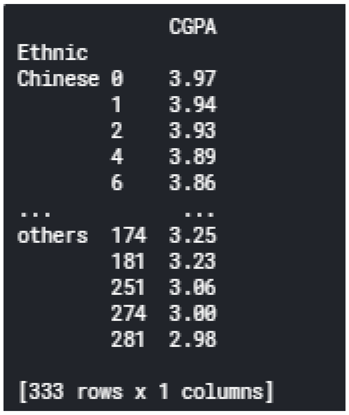


```
                CGPA
Gender
female 2        3.93
       3        3.91
       4        3.89
       5        3.89
       7        3.84
...              ...
male   322      2.71
       327      2.59
       329      2.50
       331      2.41
       332      2.28

[333 rows x 1 columns]
```

Figure 9

Whereas for ethnicity we found out that , chinese students have the highest achievement in terms of cgpa. This is followed by indians, others and malay students respectively. This further proved with the analysis below

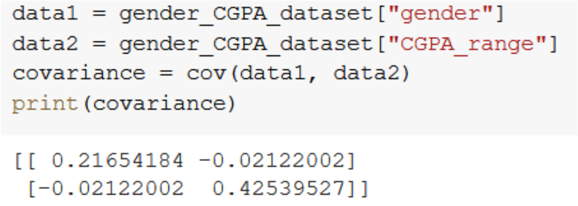


```
                CGPA
Ethnic
Chinese 0       3.97
        1       3.94
        2       3.93
        4       3.89
        6       3.86
...              ...
others  174     3.25
        181     3.23
        251     3.06
        274     3.00
        281     2.98

[333 rows x 1 columns]
```

Figure 10

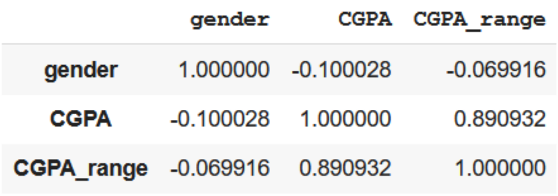
4. Is gender is a factor that can affect performance of a student in his or her study ?
(Diagnostic)



```
data1 = gender_CGPA_dataset["gender"]
data2 = gender_CGPA_dataset["CGPA_range"]
covariance = cov(data1, data2)
print(covariance)
```

```
[[ 0.21654184 -0.02122002]
 [-0.02122002  0.42539527]]
```

Figure 11 : The snapshot result of covariance value between column gender and CGPA range.



	gender	CGPA	CGPA_range
gender	1.000000	-0.100028	-0.069916
CGPA	-0.100028	1.000000	0.890932
CGPA_range	-0.069916	0.890932	1.000000

Figure 11 : The snapshot result of correlation matrix between column gender, CGPA and CGPA range.

From the snapshots above, it clearly indicates that there is no relationship between gender and CGPA of a student. Thus, gender does not affect the performance of students in the study.

5. What is the common CGPA of a student that comes from a high income family ?
(Diagnostic)

The figure below the proportion of the social classes of the dataset. The majority of them come from low income families followed by middle and upper class. This unbalanced sample makes it more challenging to find the real effects of income given to the performance of the students because the gap exists is very large.

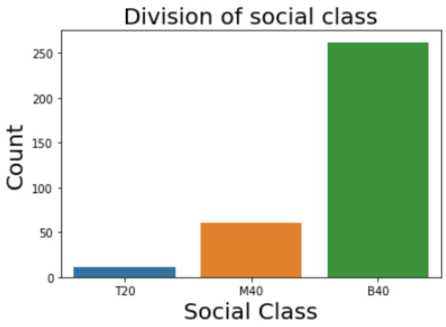


Figure 12: Division of social class

For the violin plot, we can see the distribution brings the upper income families the highlight.

.

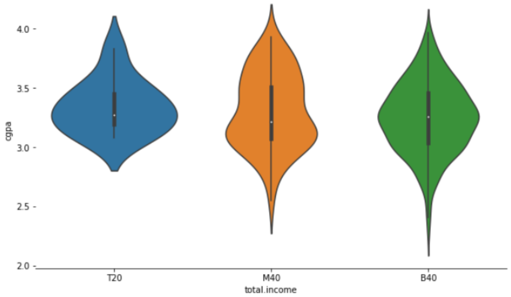


Figure 13

Swarm plot shows a more detailed figure.

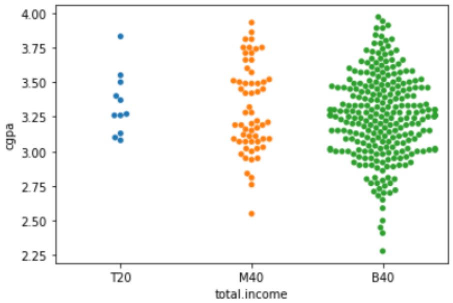


Figure 14

From these plots, we can see the income class does not have a clear correlation with the CGPA. The B40 group has almost achieved a normal distribution. All of the T20 students on the other hand got above average cgpa while M40 is much the same with B40.

DISCUSSION

Overall, the result is not satisfactory. However, as the data science process is already repeated three times and each time using a different algorithm, the team settled on the decision tree model. The reason is that as it has the best accuracy among the tested algorithms and the visualization can justify the performance.

There are many reasons for the low performance score of the model. The first is that it is suspected that there is truly no correlation between poverty and academic performance. It is hard to find evidence for this reason as Exploratory Data Analysis (EDA) cannot be done on categorical data. But, on the earlier modeling using logistic regression and using continuous data, there is no correlation or relationship between the factors chosen and academic performance. As such, this can be one of the reasons. Second, the data might not be enough and this might cause underfitting. Third and lastly, there might be a mistake in the process somewhere. The model parameter might be wrong for the problem or there is a mistake during data cleaning. Care was taken to avoid careless mistakes, but the team's lack of experience can cause problems.

Finally, there is to add that, Orange as a Machine Learning tool is highly limited to what it can do. In creating the decision tree model, there is no option to choose how features are selected when creating the decision tree. Thus, there is no method to select Gini Impurity or Information Gain when creating the decision tree model. There is also a lack of documentation on the tool itself, which can be confusing when trying to learn the tool. As such, as a student, it is highly advisable to avoid using Orange for learning purposes.

Apart from our data science process, the factors that play a vital role in determining one’s CGPA are very vague and inconsistent. Some previous researchers that did the study have found a lot of different variables including gender, family income, school involvement and the list goes on. It is quite hard to jump into conclusion on what really influences a student’s performance from time to time. To demonstrate using our findings,, let us take gender as the independent variable. We can see that the percentage of female students is higher compared to male in terms of the most excellent CGPA (above 3.50) which is 69.3% to 30.6%. Generally speaking, this particular finding shows that female students are more intelligent in academic i.e. CGPA compared to male students. And to support this particular finding, a study [16] found in most of the departments, female college students always outperformed their male counterparts and the reasons behind it include better class attendance and stronger motivation. Another study conducted by [17] also upholds this point by stating that female students have higher perception of involvement with schools, thus resulting in higher academic performance compared to men. Unfortunately, this statement is not always true, a study [18] conducted on students from the matriculation programme of Universiti Kebangsaan Malaysia (UKM) conveys that the majority of unsuccessful students are female. There are actually a lot more studies that contradicted our findings thus making it very nondeterministic or perhaps requiring more advanced algorithms, a bigger scale dataset and experienced people in this sector in order to reveal the key contributors to one's academic performance. However, in our opinion, the parental income is the main reason. From the basic thoughts alone, if one is struggling to even fulfill their needs, then it will be quite troubling for them to focus on the other aspects of their life, especially education. An interesting critical review [19] has done a very good explanation on correlating the wealth inequality with the consequences that it brought to educational achievement in which they stated that the underprivileged students will be more likely to averse the student loans thus lower the chances to even complete their studies in college because they are forced to jump into workplace in order to support their family. With respect to college graduates or college completion rates, [20] finds that the majority of them were coming from higher wealth backgrounds while their counterparts were being left behind. In conclusion, the efforts of tightening the gap of education between the rich and the unfortunates are very crucial in addition to this information era where everyone should have the similar opportunities to be successful. In other words, the dreams of making education accessible to everyone where one's background does not matter need to be realized in the future.

FUTURE WORK

Many different adaptations, tests, and experiments have been left for the future due to lack of time. In this work, we have only considered 4 independent variables to examine the CGPA outcomes which yield a not so good model. Thus, future works must include a variety of logical factors other than socioeconomic alone. The possibilities can include personal problems, family issues, mental health, time management, curricular involvement and the list goes on. However, a variety of variables is not enough, future works should also apply the appropriate sampling techniques that are ultimately bias-free. Furthermore, the class of the education needs to be segregated into different categories such as primary, secondary or bachelor's degree because every class has their own uniqueness in benchmarking the performance of the students. For example, the examination system in university majorly consisted of carry marks which will be accumulated throughout the semester meanwhile, academic performance in secondary schools will only be manifested via final exam at the end of the year. Also, this could lead to a more systematic approach to solve this puzzle. In addition, due to the non-linear relationship of family income with CGPA, future works must consider building a neural network model which is very well-known for its extensive application in predictive analytics. Last but not least, with regards to the tools that we used which is Orange and Excel, an extension for the near future is the use of multiple tools that are popular in doing statistical analysis. For example, Statistical Package for Social Sciences (SPSS), R and MATLAB.

REFERENCES

[1] E. Yunaeti Anggraeni et al., "Poverty level grouping using SAW method", International Journal of Engineering & Technology, vol. 7, no. 227, p. 218, 2018. Available: 10.14419/ijet.v7i2.27.11948.

[2] W. Wilson and R. Taub, Poverty and Welfare in America: Examining the Facts, 11th ed. Santa Barbara, California: ABC-CLIO, 2019, p. 26.

[3] Cuaresma, J. Crespo, W. Fengler, H. Kharas, K. Bekhtiar, M. Brottrager and M. Hofer, "Will the Sustainable Development Goals be fulfilled? Assessing present and future global poverty.", Palgrave Communications, vol. 4, no. 1, pp. 1-8, 2018.

[4] Hoy, Christopher and A. Sumner, "Growth with adjectives: Global poverty and inequality after the pandemic", Center for Global Development Working Paper, 2020.

[5] Vaziri, Mehrdad, M. Acheampong, J. Downs and M. Majid, "Poverty as a function of space: understanding the spatial configuration of poverty in Malaysia for Sustainable Development Goal number one", GeoJournal, vol. 84, no. 5, pp. 1317-1336, 2019.

[6] Nor Fatimah Che Sulaiman, Noor Haslina Mohamad Akhir, Nor Ermawati Hussain, Rahaya Md Jamin and Nur Hafizah Ramli, "Data on the impact of socioeconomic status on academic achievement among students in Malaysian public universities", Data in brief, vol. 31, no. 106018, 2020.

[7] J. Feng and S. Jha, *Predicting students' academic performance with decision trees and neural networks*.

[8] L. Abu Zohair, "Prediction of Student's performance by modelling small dataset size",

International Journal of Educational Technology in Higher Education, vol. 16, no. 1, 2019.

[9] R. Asif, A. Merceron, S. Ali and N. Haider, "Analyzing undergraduate students' performance using educational data mining", *Computers & Education*, vol. 113, pp. 177-194, 2017.

[10] H. Altabrawee, O. Ali and S. Ajmi, "Predicting Students' Performance Using Machine Learning Techniques", *JOURNAL OF UNIVERSITY OF BABYLON for Pure and Applied Sciences*, vol. 27, no. 1, pp. 194-205, 2019.

[11] Havan Agrawal and Harshil Mavani, "Student Performance Prediction using Machine Learning", *International Journal of Engineering Research and*, vol. 4, no. 03, 2015.

[12] D. Buenaño-Fernández, D. Gil and S. Luján-Mora, "Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study", *Sustainability*, vol. 11, no. 10, p. 2833, 2019.

[13] J. Xu, K. Moon and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742-753, 2017.

[14] E. Belachew and F. Gobena, "Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 2, pp. 46-50, 2017.

[15] A. Polyzou and G. Karypis, "Feature extraction for classifying students based on their academic performance", *11th International Conference on Educational Data Mining*, pp. 356 - 362, 2017. [Accessed 8 January 2021].

[16] Hanita, M, Y. and Azman, N., "Academic achievement among male and female students: the role of learning support and students' engagement", Malaysian Journal of Learning and Instruction (MJLI), vol. 15, no. 2, pp. 257-287, 2018.

[17] Arof. Razali, "Rural students and academic performance: a case study of program Matrikulasi Universiti Kebangsaan Malaysia", Unpublished PhD dissertation, Cornell University, 1985.

[18] Rauscher, E. and Elliott III, W. (2014) The Effect of Wealth Inequality on Higher Education Outcomes: A Critical Review. Sociology Mind, 4, 282-297. doi: 10.4236/sm.2014.44029.

[19] Pfeffer, F. T. (2018). *Growing Wealth Gaps in Education. Demography, 55(3), 1033–1068.* doi:10.1007/s13524-018-0666-7

[20] S. Van den Berg, Poverty and Education, 10th ed. Paris/Brussels: International Institute for Educational Planning/International Academy of Education, 2008.