Predicting Poverty and Its influence in Academic Performance in Malaysia

Wan Huzaifah bin Wan Azhar, Muhammad Salikin bin Ismail, Ahmad Zahin Fikri Bin Rozlee, Ahmad Zaidan bin Adnan. Department of Computer Science, Kulliyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia.
wanhuz96@gmail.com, salikinbaik@gmail.com, zahinfikri128@gmail.com, zaidan3.adnan@gmail.com

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

1. 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 [5]. In Peninsular Malaysia, however, 0.7 % of the household population live in poverty according to Vaziri et.al [6]. 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.

1.1. The Background

According to Li, Xu and Deng [8], 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.

1.2. Problem Statement

Poverty is an issue that should not be overlooked by all authorities. The negative impacts it has brought to the country is quite significant especially in the education sector. One effect of poverty is it decreases the potential to learn and chances to enrol in schools [25]. Another reason why we should focus on this issue is because the sentiment and stereotypes of the community towards low performing students are very bad. Moreover,those students are also being perceived as lazy and underperformed by their teachers. In this project, we want to visualize the dependencies of the income of one household with the student's academic performance in a developing country which is Malaysia.

1.3. Research Hypothesis

Poverty, whether originating from family backgrounds or national economic issues, have created many difficulties in terms of education necessities, technology or financial issues towards affected students. Our hypothesis is that these difficulties have affected the students' academic performance negatively or in simple terms, the decline in students' academic grades.

1.4. Research Objectives

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

1.5. Research Significance

- 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.

2. 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 [9][10], 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 [11]. 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 [12][13] 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 [14] 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 [15]. 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 [16][17] 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. Research in [18] reported that this model outperformed other models in predicting final grade based on given GPA, assignment grade, participation rate, attendance and test average grade. However, the implementation for the rest of the studies recorded that this model got the worst performance in terms of precision and accuracy.

The ROC, mean square error, accuracy and precision metrics are the most popular evaluation metrics to assess the model. Most papers compare the performance of the models using these metrics to get the best algorithm for prediction.

No.	Year	Authors	Research Problem / Application	Main techniques	Results	Future works
1	2019	Hussein	Identify student who need	Decision Tree,	ANN model recorded the	

		Altabrawee, Osama Ali, Samir Qaisar Ajmi	extra support and taking appropriate actions using machine learning model	Naive Bayes, Artificial Neural Network(ANN), Logistic Regression	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 Prediction (EPP)	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 courses.	The performance prediction to elective courses can be extended. Recommend courses to the student using the prediction results.
7	2017	Ermiyas Birihanu Belachew, Feidu Akmel Gobena	Using machine learning techniques to analyze data in the student	Naive Bayes, Artificial Neural Network (ANN),	The Naive Bayes model got the highest accuracy compared to other models at	Use more data to gain better

			management information system.	Support Vector Machine (SVM)	95.7%.	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.	

Table 1: List of papers with similar project

3. METHODOLOGY

3.1 Data Description

The dataset is taken from supplementary materials of "The impact of socioeconomic status on academic achievement among students in Malaysian public universities"[19] . It is based on a survey conducted with final year social science students of Universiti Malaysia Terengganu. The students are from various states in Malaysia with different ethics. The datasets consist of one file in Microsoft Excel (xlsx) format as well as its description in Microsoft Word format (docx). The data contains various information of students in Universiti Malaysia Terengganu, which includes student's GPA, parent's income level, socioeconomic status and others. The survey was done on 965 students, however it was reduced to a sample size of 333 students through random sampling.

3.2 Tools Used

In this project, Orange was chosen as the data analytics and data modeling tool. Orange is a GUI based software for data mining and analysis. First of all, the main reason for choosing Orange over R is that the team want to expand existing toolsets and learn other tools. Secondly, while R is a powerful tool for data analytics and modeling, it decided that it is an unnecessarily powerful tool for our purpose. As such, Orange was chosen as it is much easier and faster for classification modeling such as decision tree or random forest. Finally, Orange is also simple to use as it is a UI based tool compared to command-line tools of R.

Microsoft Excel is also used to clean the dataset. Although the traditional data cleaning process is to use R, Excel works for our purpose as the dataset is small. To add to that, Excel also has its formula function to create new data based on previous data. For example, a new data Total_Income can be created using formula "=SUM(Mother_Income, Father_Income)" and applies to all 333 entries. Thus, it works for our purpose.

3.2 Machine Learning Algorithm

In this project, a classifier algorithm is needed to predict academic performance of B40 and non-B40 class. Therefore, the team has chosen Decision Tree as the main modeling algorithm. It was chosen as it can classify or predict academic performance based on given socioeconomic factors. It was also chosen because the variables in the dataset are mostly categorical and not continuous. Thus, it is suitable for this project.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on a feature (e.g. whether a coin flip comes up heads or tails). Each branch represents the outcome of the test, and each leaf node represents a class label, which is the final decision taken after computing all features. The paths from root to leaf represent classification rules.

A decision tree generated by choosing features as a node of the tree. The tree starts with root node, which is the first, topmost node and ends with leaf node, which is the node without an out branch. The node is chosen each step by how much information the feature has. The feature with the most information will be selected as the root node and the process is repeated until there is no more feature. Lastly, based on the feature, a decision will be made based on the majority of the vote in the feature. Figure 1 below shows an example of feature selection. If the weather is cloudy and the person is hungry, a "Walk" decision will be made if the majority of the data in the dataset votes that cloudy weather and hungry person is to take the bus, such as 25/30 votes is for "Walk" while 5/30 is to "Bus".

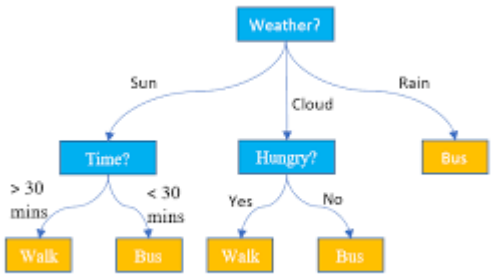

Figure 1: Example of decision Tree

Figure 1 shows an example of a decision tree. The blue node represents features and the yellow node represents the final decision. If the weather is cloudy and the person is hungry, then the final decision is to walk to get to their destination.

3.3 Training and Testing datasets

The dataset is splitted into train and test dataset. The split between train dataset and test dataset is 70% train dataset and 30% test dataset based on the 70/30 rule. There is no reason for the chosen rule. But, as there is no definitive number to split the dataset in Machine Learning, the 70/30 rule was chosen. It is also a rule of thumb that works for certain people.

3.4 Modeling Process

Overall, the modeling process is based on the standard Data Science Process. First, raw data is collected. Then the data will be processed. During data processing, data wrangling and feature extraction will be done. After that, the dataset will be cleaned to remove formatting issues. After that, modeling using Machine Learning algorithm will be developed using the cleaned dataset. Finally, data visualization will be generated to justify the result.

3.4.1 Data Collection

A dataset is needed to explain socioeconomic status of a student and their GPA during their university year. In this case, data collection was done by searching through search engine Google and finding relevant dataset based on the key terms that are related to the topic. Fortunately, there is already a group of researchers in Malaysia that collect the data for their research, which can be used for our research purpose. Their research and their supplementary dataset was found on ScienceDirect.com under the title of “Data on the impact of socioeconomic status on academic achievement among students in Malaysian public universities". As such, raw data is collected.

3.4.2 Data Processing

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

Table 2: Chosen features based on socioeconomic factors and academic performances

In the dataset, there are no features for "household with one working parent". Therefore, data wrangling needs to be done to create it. The feature was created by specifying a formula in Microsoft Excel. The formula stated that, if there is no income for only mother or only father, then there is only one working parent and the value is 1. Else, the value is 0, which stated that both parents are working. Feature of parent SPM education is also reduced to two categories, which is 2 for having SPM qualification and above and 1 for not having SPM qualification.

There are also missing values in the dataset. In the dataset, the household income category and level of education is missing some entries. For both of the features, mode of the row was chosen as a method to fill the missing value.

3.4.3 Data Cleaning

Data cleaning was mostly done to remove formatting issues and remove irrelevant features. It is further cleaned to make it able to read smoothly into Orange. As the dataset is almost clean, only simple cleaning is required such as removing whitespace. There are other columns in the dataset that contain whitespace and this can cause trouble for Orange as it will detect the whitespace as different features.

3.4.4 Modeling

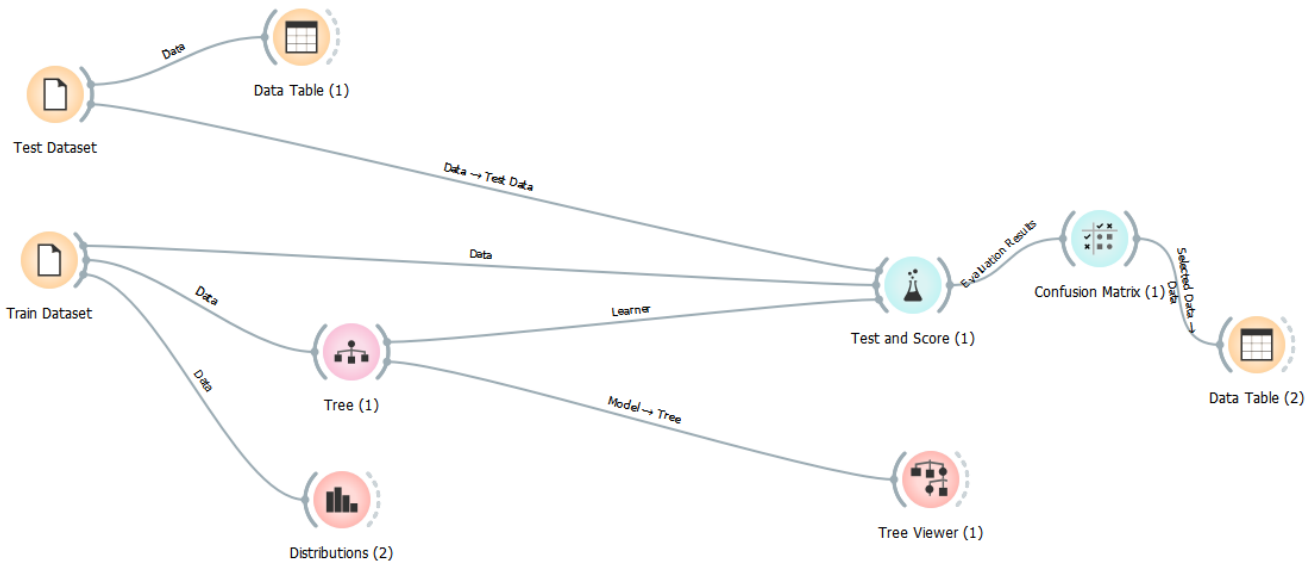

Figure 2: Modeling in Orange

Figure 2 described the whole process of data modeling in Orange. First, the train and test dataset was loaded into Orange. Then, the train data is inputted into the Tree node to train the decision tree model. Then, the Tree is outputted into the Test and Score node and Tree Viewer. The Tree Viewer node is used to visualize the tree while Test and Score is used to measure performance of the decision tree. The Test Dataset and Train Dataset was inputted into Test and Score to measure performance. The result is displayed in the Tree and Score and it also outputs Confusion Matrix. The result will be discussed in Chapter 4.

3.5 Performance measurement

As the project used a classification algorithm, the metrics chosen must be able to accurately measure the model performance of a classifier. For the project, few metrics were chosen to analyse the model. The metric is classification accuracy, precision, recall and F1 score. Classification accuracy refers to how accurate the model is, precision is the ratio that the model correctly predicts positive observations to the total predicted positive observations, recall is the ratio of correctly predicted positive observations to the all observations in actual class and F1 is the weighted average score of Precision and Recall. The formula below described how to calculate the metric using confusion matrix:

Accuracy = TP+TN/TP+FP+FN+TN
Precision = TP/TP+FP
Recall = TP/TP+FN
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4. RESULTS

4.1 Data Visualization

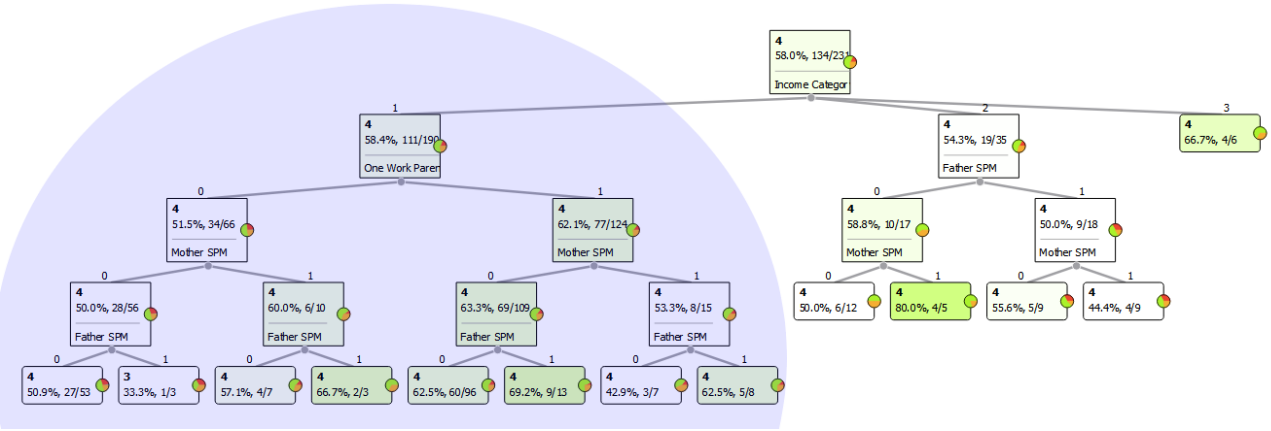

Figure 3: Tree viewer of the model

Figure 3 is the decision tree as visualized by the Tree Viewer in Orange. 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.

4.2 Model Accuracy

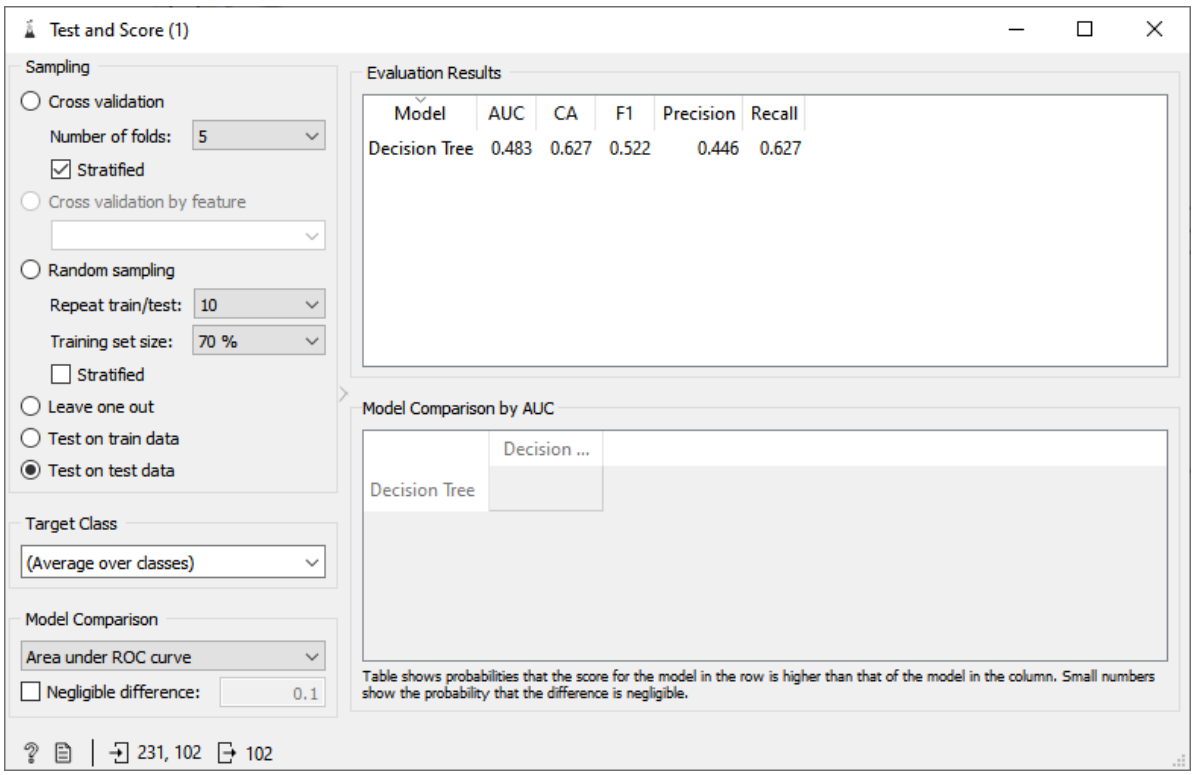

Figure 4: Test and Score result by Orange

Figure 4 shows the score of the model by applying the test dataset on the model. Classification Accuracy

(CA) is 0.627 or 62.7%, F1 score is 0.522, Precision and Recall both are 0.446 and 0.627. The result shows that the accuracy for this model is around 62.7%, which is not the best but also not a very good model as accuracy score can be misleading. F1 score especially shows that the model is barely acceptable as the average value of F1 score is 0.5. Precision of 0.446 indicates a low positive rate but recall of 0.677 tells us that 67.7% of predicted observations are correct. All in all, the low F1 score means that the model cannot be reliably used.

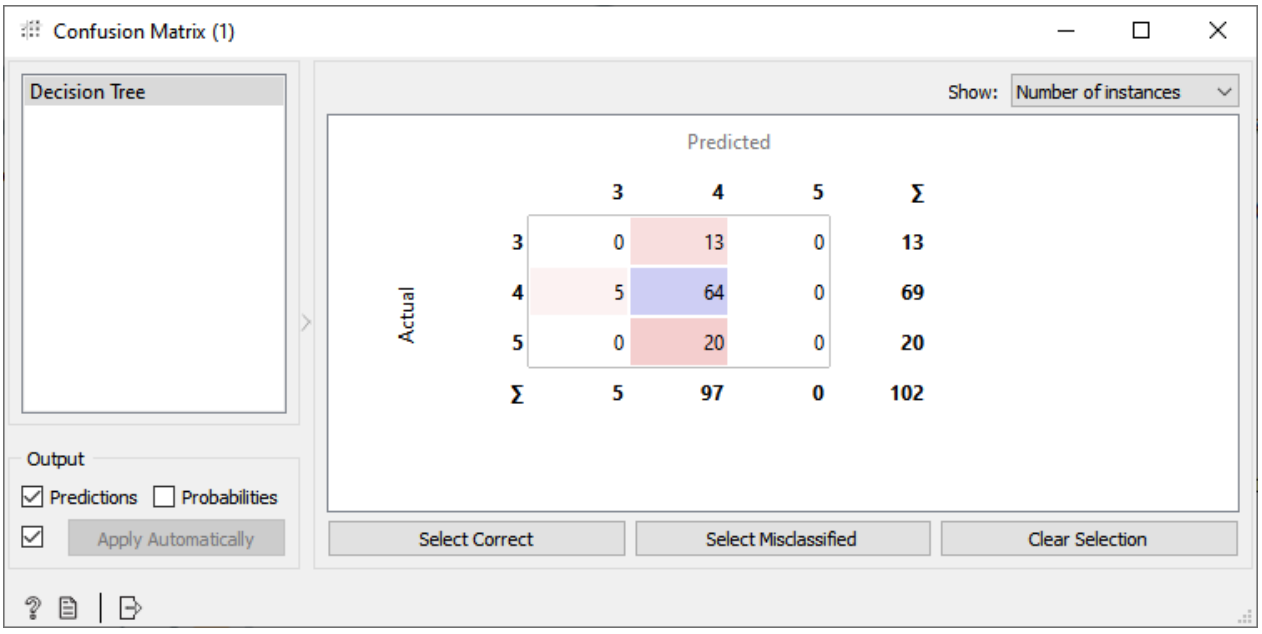

Figure 5: Confusion matrix of Test and Score in Orange

Confusion matrix of the test score also stated that the model is not a good model. In particular, only 64 out of 102 predicted observations are correct, which is only slightly above average. The confusion matrix also specifies that the model did not predict any CGPA in range 5 (3.5-4.0) and low number of predictions in range 2 (2.5-3). This can happen if the dataset itself is not correctly distributed, as range 4 (3.0-3.5) has the highest number of rows.

5. DISCUSSION

5.1 Result analysis

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.

5.2 Academical Research

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 [21] 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 [22] 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 [23] 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 [24] 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, [25] 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.

6. FUTURE WORKS

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.

7. 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]

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

[6] 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.

[7] M. Li, H. Xu and Y. Deng, "Evidential decision tree based on belief entropy", Entropy, vol. 21, no. 9, p. 897, 2019.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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].

[18] A. Mueen, B. Zafar and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", *International Journal of Modern Education and Computer Science*, vol. 8, no. 11, pp. 36-42, 2016.

[19] Siwar Chamhuri, Mohd Khairi Ismail, Nurul Alias and Siti Zalikha, "Kumpulan Isi Rumah Berpendapatan 40 Peratus Terendah (B40) di Malaysia: Mengenal Pasti Trend, Ciri, Isu dan Cabaran", no. 2, pp. 33–50,

2019.
[20] Dayioğlu, M., Türüt-Aşik, S. Gender differences in academic performance in a large public university in Turkey. High Educ 53, 255–277 (2007). https://doi.org/10.1007/s10734-005-2464-6
[21] 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.
[22] Arof. Razali, "Rural students and academic performance: a case study of program Matrikulasi Universiti Kebangsaan Malaysia", Unpublished PhD dissertation, Cornell University, 1985.
[23] 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.
[24] Pfeffer, F. T. (2018). *Growing Wealth Gaps in Education. Demography, 55(3), 1033–1068.* doi:10.1007/s13524-018-0666-7
[25] S. Van den Berg, Poverty and Education, 10th ed. Paris/Brussels: International Institute for Educational Planning/International Academy of Education, 2008.