An empirical study on the first semester of the MSc in Data Science and Management

Giulia Di Martino, Martina Bozzi

January 31, 2022

Contents

1 Introduction

Data regarding 26 students attending the MSc in Data Science and Management at LUISS Guido Carli has been collected through a unified survey which expired at the beginning of January. The objective of this empirical analysis is the study of the student's performance, and feedbacks at the end of the first semester of the academic year 2021/22. In a first stance, by carrying out a data visualization analysis, specific insights concerning the background, results and improvements of the students have been graphically displayed. Then, a CART tree model was built in order to inspect the feedbacks provided by the students. Finally, with the aim of investigating the "sentiment" behind those feedbacks, supervised and unsupervised sentiment analysis have been carried out.

2 Data description

The dataset is composed of 26 observations, and 25 variables (18 categorical variables, and 7 numeric). The 26 observations represent the number of students attending the master program. From the dataset, the 25 relevant variables used in the following analysis have been extracted. The survey has been conducted anonymously: it gathers data on the students' performance based on the evaluation of 3 exams out 4 (the grade concerning the Internet and Network Economics exam was not available at the closing date of the survey), and on the feedbacks provided by the students on the attended courses.

3 Data analysis

The first step of the analysis focused on the inspection of the dataset. The group of students who participated to the survey counts 14 males and 12 females (plot1). Through data manipulation, it has emerged that the first 10 students with the highest mean grade (out of 3 exams) are 7 females and 3 males, having an average grade of 29,25. Contrary, the 5 students with the lowest exam performance are 4 males, and 1 female with an average grade of 23,6.
Moreover, it appears that, on average, the 26 students:

- have passed 2.85 exams out of 3;
- have an age of 22.46 years;
- have dedicated to sport activities 1.65 hours per week;
- considered to have a programming level prior to the beginning of the master course in Data Science and Management of 2.65/10. This index has reached 4.81/10 at the end of the first semester;
- have studied 3.77 hours per week.

In detail, the females who took part to the survey have on average an age of 21.83 years, passed 3.00 exams out of 3 by studying 4.08 hours per week. On the other hand, the males have an average age of 23.00 years, passed 2.71 exams out of 3 by studying 3.50 hours per week.
Furthermore, it was deducted that the 5 students out of 26 who got a scholarship passed all the 3 exams, studying 4 hours per week, and have an average age of 22.20 years. Whereas, the students

who did not get a scholarship registered overall a slightly lower performance.
Focusing on the parental education level, it became clear that the students having both their parents with a university degree tend on average to perform better.

3.1 Data visualization

Throughout the second part of the empirical research a data visualization analysis has been performed. The latter was useful to obtain deeper insights on the data previously described. The first graph (plot2) shows the average age (in percentage terms) of the students under analysis: 23.08% are 21 years old, 42.31% are 22, 23.08% are 23, 7.69% are 25 and 3.85% are 28.

Analysing the students' achievements, and considerations regarding the first semester of the course provided additional insights. As a matter of fact, the second plot (plot3) outlines the total number of exams passed at the end of the first semester: 21 students have passed 3 exams, whereas 5 students have passed 2 exams. Based on the data gathered, only 19.23% of the students got a scholarship (plot4) for the current academic year (2021/22). Moreover, it emerged that, according to 57.69% of the students (plot5) the most difficult course of the semester was Advanced Statistics. Likewise, Internet and Network Economics was claimed by 19.23% of them to be the most tough subject. A bar plot (plot6) was exploited to inspect the students' study preferences: it resulted that 16 of them favoured individual studying, while the remaining 10 have preferred to study with a group of friends.
Furthermore, for what concerns study organization (plot7), 12 students indicated to "have a plan for every day", 9 of them argued "to speedrun tasks when time is over", and only 5 students considered themselves "to study whenever they feel guilty". During the first semester of classes, due to the amount of study required, 5 students retained to not be able to balance their private life (plot8). On the other hand, the remaining 21 students found themselves able to manage the two tasks. Indeed, 9 students did not practice sport activities on a weekly basis, whereas 15 of them managed to exercise two or three hours per week (plot9).

Focusing on the students' academic background, it became apparent that 61.54% of them obtained their bachelor degree in LUISS Guido Carli. A share of students (11.54%) got their bachelor in Sapienza University of Rome, and 7.69% in Roma Tre. The remaining part came from other universities: UniFi, UniVe, Unibo and Unina (plot10). Moreover, in relation to passed abroad experiences, 15 students took part to an Erasmus' program (plot 11).

The inspection of students' performance enabled the visualization of their improvements throughout the first semester. The latter has been carried out by taking into consideration 4 main predictors: the mean grade (out of 3 exams), the bachelor field, the programming level at the beginning of the courses, and the programming level after attending the classes. A first comparison has been made between the students' mean grades, and their programming level before and after taking the courses. From the scatter plot, it is evident that the majority of the students enrolled to this master initially considered themselves to have a low programming level. At the same time, those who reached a high mean grade at the end of the semester are also the ones who improved the most their programming skills (plot12, plot13). At this point, the relation between the student's different bachelor fields and their programming level has been analysed in detail. The programming level's variable scale is out of 10 points.

- The number of students who got their bachelor in Economics is 7. Among them, 3 students declared to have an initial programming level of 2/10, 2 students of 1/10, 1 student of 3/10 and 1 student of 6/10. At the end of the courses, the same students reported that their programming level has increased as follows: 2 students believe to have a current programming level of 3/10, 2 students of 5/10, 1 student of 4/10, 1 student of 6/10, and 1 student of 7/10 (plot 14, plot15);

- The students who obtained their bachelor in Economics and Management are 8. Among them, 3 students stated to have an initial programming level of 1/10, 2 students of 2/10, 2 students of 5/10 and 1 student of 3/10. At the end of the semester, the same students improved their programming level as follows: currently, 4 students possess a programming level of 6/10, 2 students of 3/10, 1 student of 4/10, 1 student of 5/10. (plot16, plot17);

- The students who have a background in Management and Computer Science are 2. Among them, 1 student had an initial programming level of 5/10, and the other one of 6/10. At the end of the courses, the same students reported that their programming level has increased reaching 7/10 (plot18, plot19);

- The bachelor in Economics and Business has been obtained by 2 students. Among them, 1 student argued to have a starting programming level of 1/10, and the other one of 7/10. Those same students noted that their programming level has reached 3/10 for the first one, and 7/10 for the second student (plot20, plot21);

- The students who got their bachelor in Political Science sum up to 5. Among them, 3 students had a programming starting level of 1/10, 1 student of 2/10 and 1 student of 5/10. The same students reported an increase in their programming level: currently, 2 students consider to have programming level of 5/10, 1 student of 3/10, 1 student of 4/10, 1 student of 6/10 (plot22, plot23);

- The student graduated in Communication is 1. The latter had an initial programming level of 2/10, which has remained unchanged at the end of the semester (plot24, plot25);

- One student has a background in Business Administration. The programming level initially amounted to 1/10. The latter reached 4/10 at the end of the courses (plot26, plot27);

This analysis implies that, on average, the students with a background in Business Administration, and Political Science experienced a greater programming level improvement. Contrary, a lower improvement has been recorded by the students who took a bachelor in Communication, and Economics and Business.

3.2 Sentiment analysis

With the aim of inspecting the students' level of satisfaction concerning the first semester of the master program, a sentiment analysis was conducted. The latter has been divided in two parts: in a first stance a CART model was built, afterwards a supervised and unsupervised sentiment analysis was run.

3.2.1 CART tree model

In order to carry out the CART tree model analysis, two relevant variables have been extracted from the original dataset: "course feedback", and "programming level after". The first variable consists of brief feedbacks on the courses provided by each student. Whereas, the second one was used as a threshold in order to evaluate the level of satisfaction of the students concerning the programming skills acquired during the courses.
Then, an additional variable "NEGATIVE" has been added to the dataset: "NEGATIVE" is TRUE when the programming level of each student takes a value equal to 4 or lower. This means that those students haven't reached a sufficient programming level at the end of the first semester, suggesting a negative feedback as outcome. On the other hand, "NEGATIVE" takes the connotation FALSE, when the programming level of each student takes a value equal to 5 or higher (out of 10), suggesting a positive feedback as a result.
Consequently, text preprocessing was conducted on the dataset. In a first stance, a corpus containing each feedback was created with the aim of analysing them individually. At this point, it was possible to perform data cleaning:

- the corpus was transformed to lowercase;
- punctuation was removed;
- stemming was performed;
- english standard stopwords were removed;
- a set of misleading words was discarded from the corpus;

Once the corpus was cleaned, it was ready to be inspected. A Document term matrix (DTM) was created with the aim of finding the words' frequency. Looking at the output, there are 26 documents (so a DTM made of 26 lines) with 60 terms. Regarding sparsity, the non-sparse terms are 89 and, the sparse ones are 1471. In detail, the sparsity is 94% (1471 sparse terms (89/1471), and by computing 1471/(1471+89) = 0.94). In order to lower the sparsity, a threshold of 0.92 was set. In this way, the tm package drops the terms that are very infrequent, so only the terms with a sparsity greater than 0.92 have been removed. Consequently, the DTM showed a sparsity of 85% with 26 lines, and 8 terms.
The dataset was later split into train and test set as follows: 80% for the train set, and 20% for the test set. Subsequently, the model was fitted using "NEGATIVE" as dependent variable, and the training set as data. After having fitted the model, the following CART tree was generated (plot28). A CART output is a decision tree where each fork is a split in a predictor variable, and each end node contains a prediction for the dependent variable. In this case, three stemmed words (satisfi, studi, challeng) remained after having fitted the model, and therefore are considered to be the most relevant ones. From the CART tree it can be deducted that:

1. if the students are satisfied, the output is "FALSE", meaning that they provided a positive feedback on the first semester of the courses;
2. the students who are not satisfied, blame the amount of study required by the courses;
3. the students who do not blame the amount of study, think that the courses of the first semester were too challenging. Indeed, the final outcome of these three words combined is "TRUE", meaning that the students gave negative feeedbacks on the courses.

Finally, the test set was used to make predictions. According to the accuracy table (representing how much a model can be trusted), only one feedback of the test set was mispredicted by the model.

3.2.2 Supervised sentiment analysis

For the second part of the research, a supervised sentiment analysis was run using labeled data. The latter provides an accurate outcome regarding the "sentiment" of the feedbacks provided by the students. The following steps were performed throughout the analysis:

1. data cleaning was executed, and a DTM was created;
2. the dataset was split into train and test set. The latter are divided in two parts: one part (x train and x test) which has been converted into a dataframe, and contains the words derived from data cleaning; whereas the other set (sentiment train and sentiment test) consists of the labelled sentiment (happy/sad);
3. the proportions were computed, resulting that 57% of the feedbacks were labeled as "happy", and 43% were labeled as "sad";
4. the matrix of features of train and test set were then converted into a dataframe. Regarding the set consisting of labelled sentiment, the features "happy" or "sad" were transformed into numbers;
5. at this point, the classifier was fitted using train data by exploiting the function "svm", and the features of the train data have been employed to fit the model. The latter are being classified using the words contained in the x train set, applying an optimal linear separating hyperplane.

Finally, the sentiment values ("happy and sad") were predicted using the variables of the test set, together with parameters obtained from the training set. The model, as a matter of fact, does not know the sentiment of test data, since it has never been used to tune the model. The new function for the predictions contains the previous fitted model with training data as dependent variable, and as newdata the test set (with the unseen words). In order to understand the results of the previous passage, a confusion matrix was built. Out of the five feedbacks contained in the test set, all of them were correctly interpreted resulting into an accuracy of one. Concluding, in the test set appear 3 positive feedbacks, and 2 negative ones.

3.2.3 Unsupervised sentiment analysis

For the last part of the research, an unsupervised sentiment analysis was carried out. In this case, data is not labelled and the sentiment is extracted adopting dictionaries. The latter do not have a polarity score, hence the sentiment index is given by the number of positive and negative words. The dictionaries employed in this section of the analysis are:

1. Harvard-IV dictionary (DictionaryGI);
2. Henry's Financial dictionary (DictionaryHE);
3. Loughran-McDonald Financial dictionary (DictionaryLM).

In a second stance, a corpus was created with the tm package to inspect the data. Following, both the sentiment analysis and text preprocessing were carried out in a single function. Consequently, the sentiment behind the students' feedbacks was analysed using only the Harvard-IV dictionary, due to the fact that the other two types of dictionaries led to misleading outcomes given their financial nature (since they provided as a result an overall neutral sentiment). Contrary, according to the Harvard-IV dictionary, out of the 26 feedbacks contained in the corpus:

- 5 were found to be strongly negative;
- 5 were slightly negative, with an average negativity index of 0.3/1;
- 4 were found to be strongly positive;
- 7 were slightly positive, with an average positivity index of 0.4/1;
- 5 were found to be neutral.

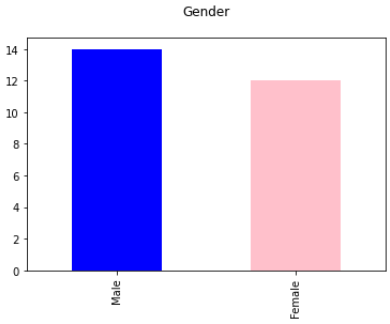
Concluding, the sentiment direction confirms that the sentiment goes from negative to positive.

4 Conclusion

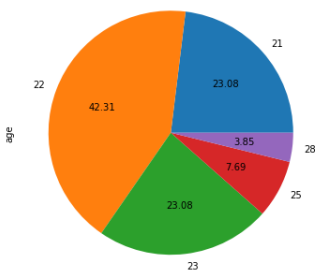
Concluding, the empirical analysis brought up a series of variegated results. The first part focused on the students' performance: the latter was evaluated taking in consideration several variables such as their academic background, the hours they dedicated to free time activities, and the hours spent studying. From the latter it emerged that, on average, out of the 26 students who took part to the unified survey, the females are younger and have a higher mean grade compared to the males. Moreover, there are 5 students who obtained a scholarship, who passed 3/3 exams of the first semester, obtaining the highest mean grade. Later on, a graphical representation of the comparison between the students' programming level, and their academic background was displayed. On average, the students with a background in Business Administration, and Political Science experienced a greater programming level improvement.
Subsequently, throughout the second part of the analysis three models were built, based on the students' personal opinions on the courses. With the CART tree model an insight about the reasons behind the students' negative feedbacks has been gained. Indeed, the latter showed that the students' discontent derived mainly from the amount of study required, and the challenges encountered during each course.
Finally, the supervised, and unsupervised sentiment analysis were carried out. From the first one, it can be deducted that the proportion of students who provided a positive feedback is 60%, whereas 40% gave a negative one.
The unsupervised model required the usage of three different dictionaries, among which just one was considered to be accurate for the interpretation of the results. The latter confirmed that the proportion of feedbacks provided by the students was overall balanced between positive and negative; with an inclination towards the negative ones. At the same time, the sentiment direction has underlined that the sentiment fluctuates from negative to positive, confirming the results obtained in the supervised sentiment analysis.

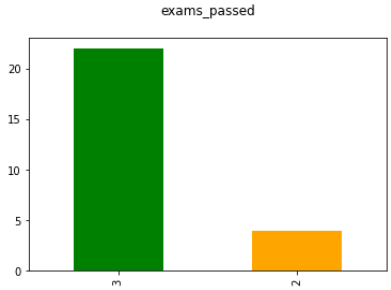
5 Appendix

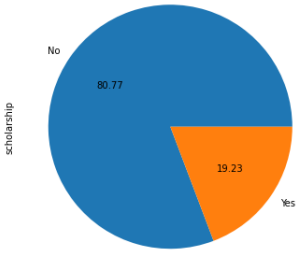
5.1 Data analysis

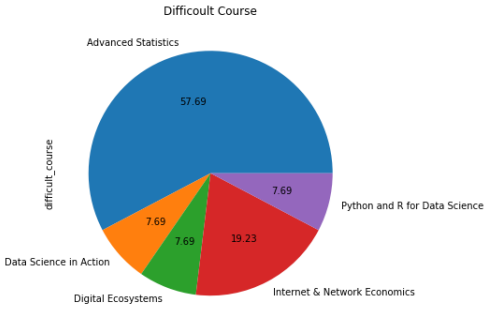


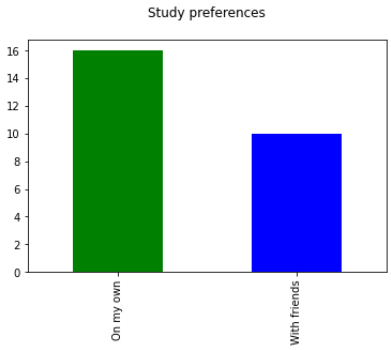
5.2 Data visualisation

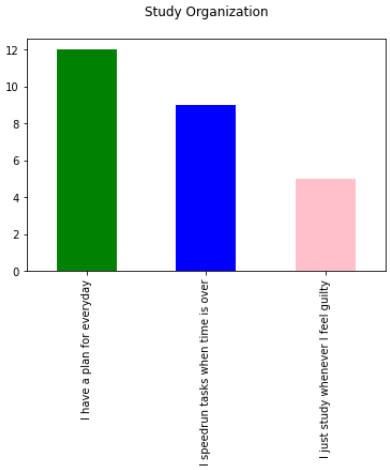


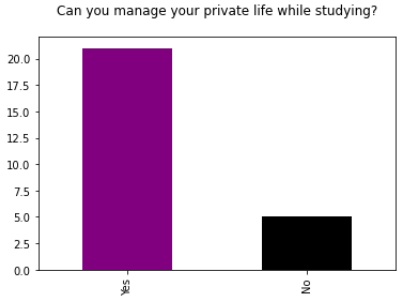


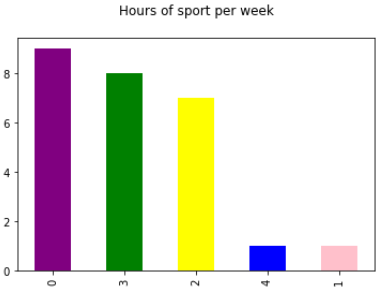


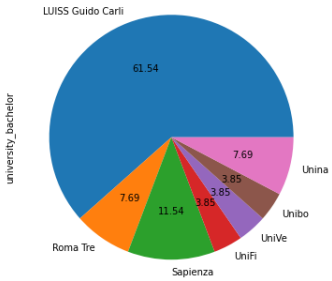


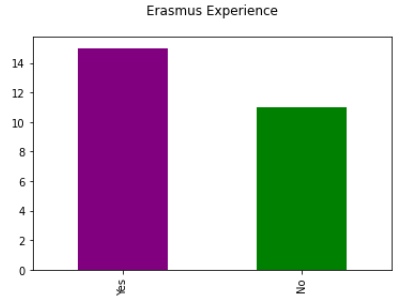


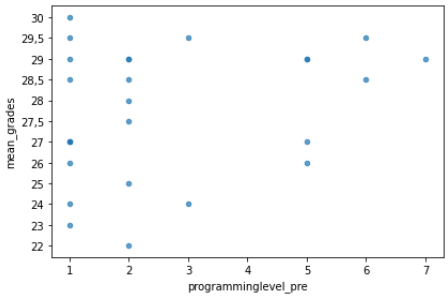

Study Organization
I have a plan for everyday
I speedrun tasks when time is over
I just study whenever I feel guilty

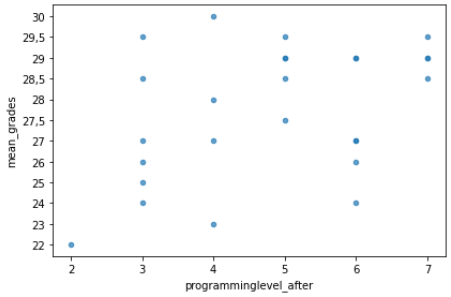

Can you manage your private life while studying?
Yes
No

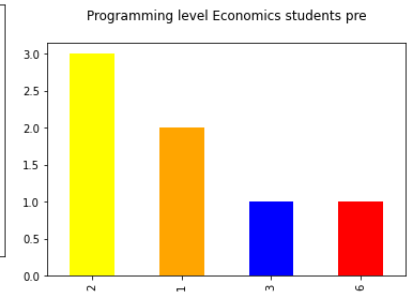

Hours of sport per week

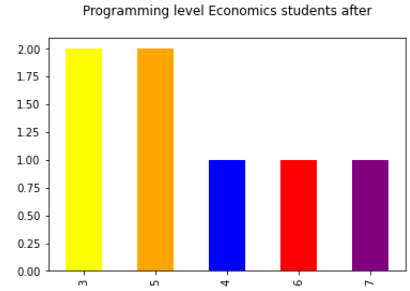

LUISS Guido Carli
61.54
university_bachelor
Roma Tre
7.69
Sapienza
11.54
UniFi
3.85
UniVe
3.85
Unibo
3.85
Unina
7.69

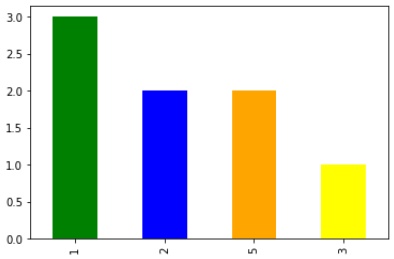

Erasmus Experience
Yes
No

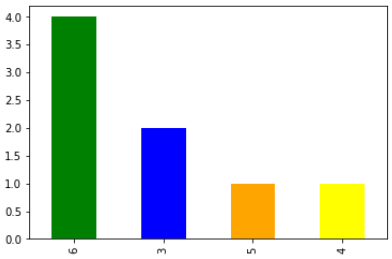

mean_grades
programminglevel_pre

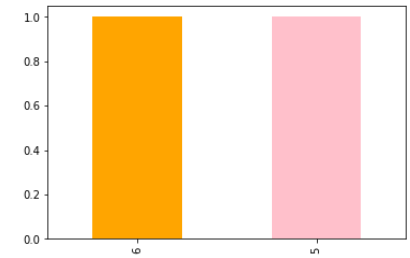

mean_grades
programminglevel_after

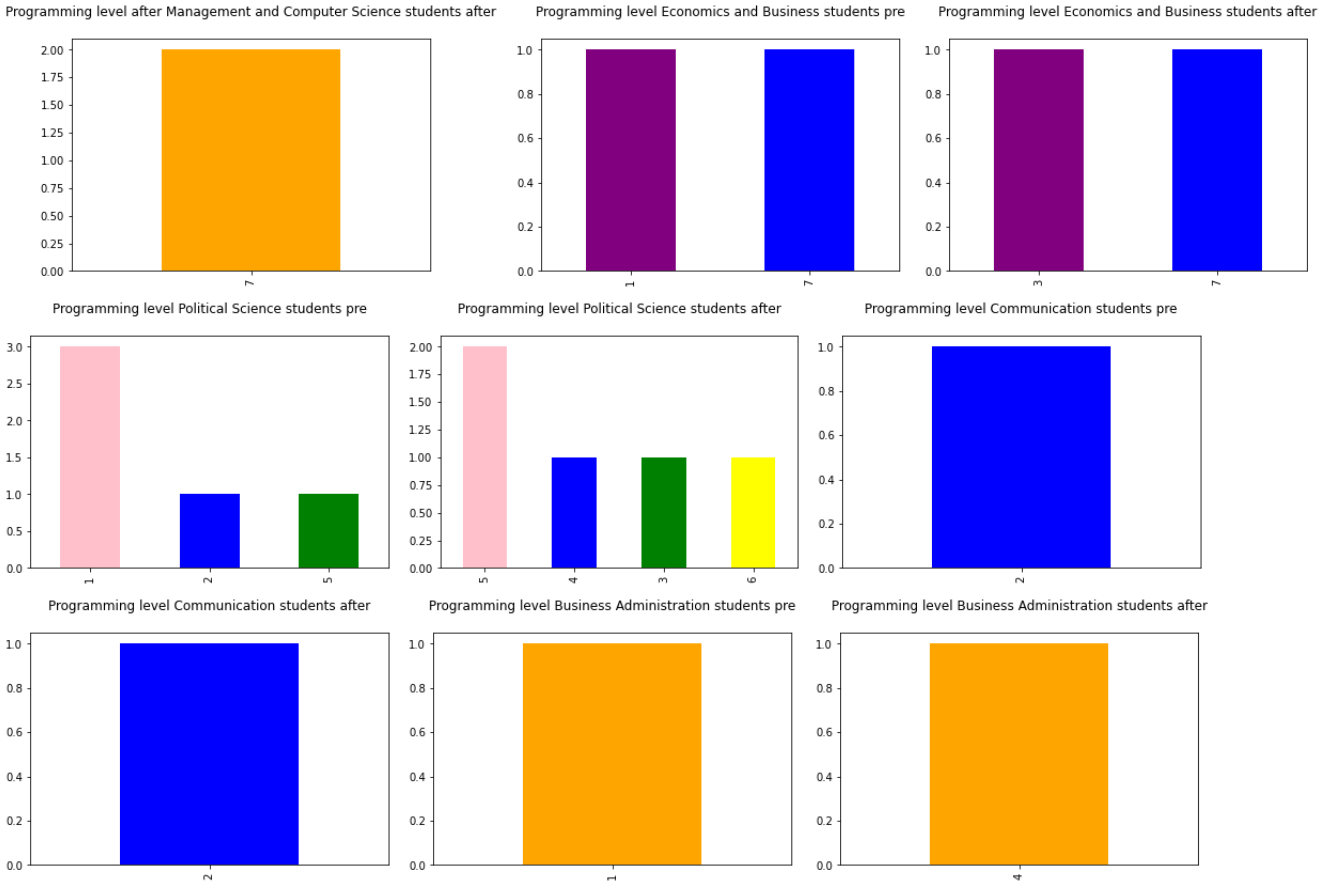

Programming level Economics students pre


Programming level Economics students after


Programming level Economics and Management students pre


Programming level Economics and Management students after


Programming level Management and Computer Science students pre


Programming level after Management and Computer Science students after
Programming level Economics and Business students pre
Programming level Economics and Business students after
Programming level Political Science students pre
Programming level Political Science students after
Programming level Communication students pre
Programming level Communication students after
Programming level Business Administration students pre
Programming level Business Administration students after

5.3 CART tree model