A logistic model to predict if preliminary year students will get into college.

Issue:

The project involves using a logistic regression model to analyze a dataset consisting of information about students who are about to complete their preliminary year. The aim is to predict whether these students will be successful or unsuccessful in gaining admission to college. To achieve this, the model will take into account 33 different factors that may contribute to the final outcome.

However, not all of these factors are expected to have the same level of significance in determining whether a student will be successful or not. Therefore, part of the project will involve identifying which factors have the greatest impact on the outcome, and which factors are less important. By doing so, we can gain a better understanding of the key factors that contribute to a student's success or failure in gaining admission to college.

Ultimately, the goal of the project is to develop a logistic model that accurately predicts the likelihood of a student being successful in gaining admission to college, based on the available data. By identifying the most significant factors, we can create a model that is more effective at predicting outcomes, which can be used to inform decision-making and support students in their academic pursuits.

Findings:

The findings suggest that the logistic model performed relatively well in predicting the success or failure of students who are doing a preliminary year and whether they will be admitted to college. The accuracy of the model is high at 0.86363, indicating that it correctly classified a large proportion of the cases in the dataset.

The precision of the model is also high at 0.8462, indicating that of the students predicted to be successful, 84.62% actually did gain admission to college. The recall value of 0.9167 suggests that of the students who actually succeeded in gaining admission to college, 91.67% were correctly identified by the model.

The F1-score, which is a measure of the overall performance of the model, is 0.88, indicating that the model is able to balance both precision and recall.

Overall, the findings suggest that the logistic model is a useful tool for predicting the success or failure of students in gaining admission to college, and that the identified factors are significant in determining the outcome. However, further analysis may be needed to identify any potential limitations or areas for improvement in the model.

Discussions:

However, there are some limitations and areas for further discussion. One limitation is that the model may not be generalizable to other datasets or populations. The dataset used to build the model may have specific characteristics that are not representative of other populations or circumstances, which may affect the accuracy and reliability of the model.

Another limitation is that the model may not capture all of the relevant factors that contribute to a student's success or failure in gaining admission to college. While the model considered 33 different factors, there may be other important variables that were not included in the analysis. Additionally, some of the factors considered may have been collinear, meaning that they were highly correlated with each other and may have artificially inflated their importance in the model.

Further discussion could also involve exploring ways to improve the model's performance. This could include experimenting with different machine learning algorithms, refining the variables considered in the model, and increasing the sample size of the dataset to improve the accuracy and generalizability of the model.

In conclusion, while the findings of the logistic model are promising, there are limitations and opportunities for further discussion and refinement to improve its accuracy and usefulness in predicting the success or failure of students in gaining admission to college.

Appendix A: Method

Logistic regression is used on a dataset to predict whether students completing a preliminary year will be admitted to college. First, the dataset is imported using the pd.read_excel() function and information about the dataset is obtained using data.info(). The data is then cleaned by dropping unnecessary columns, filling empty values with the mode of the respective factor, and converting object values to int or float values using label encoder. The cleaned data is stored in a new

variable. The dataset is divided into training and testing sets using an 80:20 ratio. The training data is used to train the logistic regression model, which is then tested on the test data to evaluate its performance. The accuracy of the model is calculated by comparing predicted values with actual values. The logistic model also provides precision, recall, and F1-score values. The coefficients of various vectors in the new data are obtained to identify positively and negatively correlated factors.

Appendix C: Results

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print('Accuracy:', accuracy)
print("Precision: ", precision)
print("Recall: ", recall)
print("F1-score: ", f1)
```

```
Accuracy: 0.8636363636363636
Precision:  0.8461538461538461
Recall:  0.9166666666666666
F1-score:  0.8799999999999999
```

The coefficients are

	feature	coefficient
21	Completed Community Service Requirement? (1=ye...	1.216846
24	Number of Workshops Attended	0.922985
29	Completed Connect? (1=yes, 0=no)	0.918969
2	Federal Ethnic Group	0.915511
23	Number of Peer Mentor Meetings Attended	0.588715
5	Attended Orientation? (1=yes, 0=no)	0.412935
20	Completed Campus Event Requirement? (1=yes, 0=no)	0.401207
25	F17 GPA	0.299432
9	Completed Summer Bridge? (2=completed all, 1=c...	0.275527
6	Attended Experience Day? (1=yes, 0=no)	0.247443
28	Number of Credits Earned	0.216230
8	Athlete? (1=yes, 0=no)	0.121882
14	Receptivity to Academic Assistance (percentile...	0.072988
22	Number of Faculty Advisor Meetings Attended	0.059056
10	Dropout Proneness (percentile score before sta...	0.053809
16	Receptivity to Social Engagement (percentile s...	0.050450
17	Receptivity to Career Guidance ((percentile sc...	0.038631
15	Receptivity to Personal Counseling (percentile...	0.012981
1	SAT Score	0.001108
18	Receptivity to Financial Guidance (percentile ...	-0.003163
12	Educational Stress (percentile score before st...	-0.007368
19	Desire to Transfer (percentile score before st...	-0.009066
27	CUM GPA	-0.040823
4	Pell Grant Eligible? (1=yes, 0=no)	-0.044285
11	Predicted Academic Difficulty (percentile scor...	-0.064198
13	Receptivity to Institutional Help (percentile ...	-0.122403
26	S18 GPA	-0.225562
3	Gender	-0.248215
7	Resident/Commuter (1=resident, 0=commuter)	-0.278693
0	High School GPA	-0.307349