Project Update : Predicting Students' End-of-Term Performances

Github: https://github.com/nmahanloo/CST_383_Final_Project YouTube: https://youtu.be/OQAxYBdoEaY PowerPoint: https://csumb0-my.sharepoint.com/:p:/g/personal/chzavala_csumb_edu/EajFKM3TZQdLk_S5A05x8UEBey1qxO7UAIuU_Lt4LdG1mg?rtime=Lds3VgON3Eg

Team members

1) Saulloa@csumb.edu 2) nmahanloo@csumb.edu 3) chzavala@csumb.edu 4) doj@csumb.edu

Introduction

Our project focuses on utilizing machine learning to predict students' end-of-term performances. By gathering comprehensive data encompassing personal details, study habits, and family support, we aim to construct a predictive model capable of accurately forecasting final grades. Currently, we are in the process of meticulously refining our dataset to identify key variables. As our work progresses, we anticipate uncovering discernible patterns that will enhance the efficacy of our predictive model, potentially offering valuable insights for educators to better support student success.

Choice of Dataset

The datasets we've chosen play a vital role in training our model to predict students' GPAs accurately. They capture key aspects of a student's academic journey, from demographics to study habits and external factors. We omitted certain columns that didn't align with our research goals or lacked relevance. By focusing on these essential features, we aim to create a robust model that provides valuable insights for educators and stakeholders alike.

The list with the provided numbering system:

1. Student Age (1: 18-21, 2: 22-25, 3: above 26)
2. Graduated high-school type: (1: private, 2: state, 3: other)
3. Scholarship type: (1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full)
4. Additional work: (1: Yes, 2: No)
5. Regular artistic or sports activity: (1: Yes, 2: No)
6. Do you have a partner: (1: Yes, 2: No)
7. Total salary if available (1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)
8. Accommodation type in Cyprus: (1: rental, 2: dormitory, 3: with family, 4: Other)
9. Weekly study hours: (1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)
10. Attendance to the seminars/conferences related to the department: (1: Yes, 2: No)
11. Impact of your projects/activities on your success: (1: positive, 2: negative, 3: neutral)
12. Attendance to classes (1: always, 2: sometimes, 3: never)
13. Preparation to midterm exams 1: (1: alone, 2: with friends, 3: not applicable)
14. Preparation to midterm exams 2: (1: closest date to the exam, 2: regularly during the semester, 3: never)
15. Taking notes in classes: (1: never, 2: sometimes, 3: always)
16. Listening in classes: (1: never, 2: sometimes, 3: always)
17. Discussion improves my interest and success in the course: (1: never, 2: sometimes, 3: always)
18. Expected Cumulative grade point average in the graduation (/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
19. OUTPUT Grade (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)

What it is you are going to predict

The goal of our project is to use the information from the dataset to develop a forecast that will give you an accurate prediction of the end-of-term performance of the student.

What features you plan to use as predictors

To build an effective predictive model, we'll be using specific features as predictors. These include the student's Grade Point Average (GPA), the number of hours they dedicate to studying each week, their class attendance, how often they take notes during lectures, and their unique STUDENT ID. By considering these key factors, we hope to gain a deeper insight into students' academic behaviors and improve the accuracy of our predictions.

Preliminary work on data preparation

- Data cleaning: Identify missing values and remove unnecessary rows and columns.
- ---“StandardScaler” and/or “MinMaxScaler” for normalization from “sklearn.preprocessing”
- Scaling numeric features, since the dataset varies in size

- Scaling categorical features, we plan on adding value to the non-numeric features

Preliminary work on data exploration and visualization

Presenting graphs below based on the dataset, highlighting the correlation with GPA as the primary focus.

```
!pip install pandas numpy matplotlib seaborn
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (2.0.3)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (1.25.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-packages (3.7.1)
Requirement already satisfied: seaborn in /usr/local/lib/python3.10/dist-packages (0.13.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2024.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.2.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (4.53.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (1.4.5)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (24.0)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib) (3.1.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

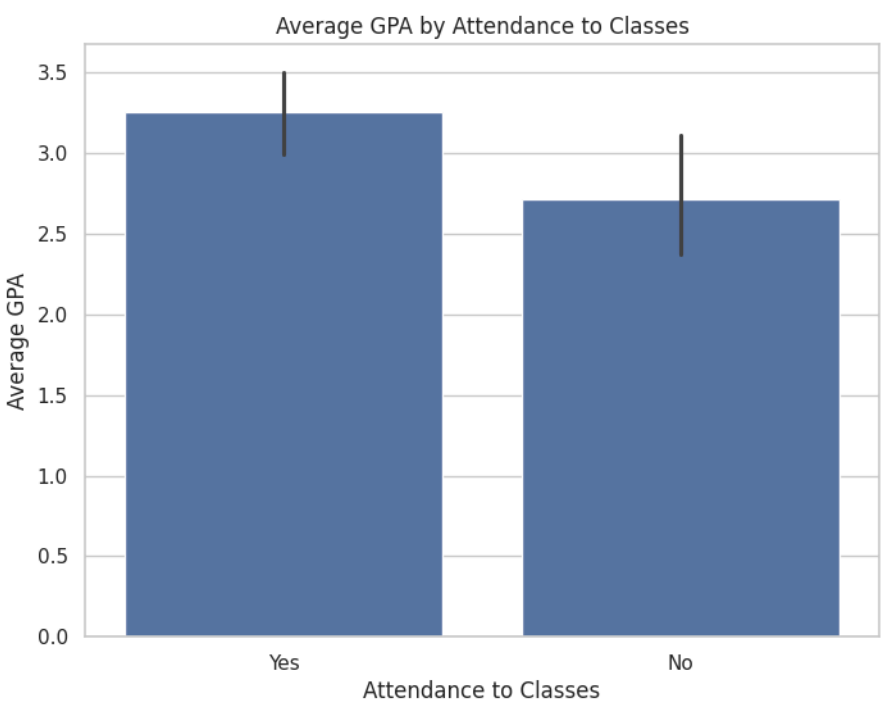
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```

```
# Configure visualizations
sns.set(style='whitegrid')

url = 'https://raw.githubusercontent.com/kd65541/CST383_project/main/archive/StudentsPerformance_with_headers.csv'
df = pd.read_csv(url)
```

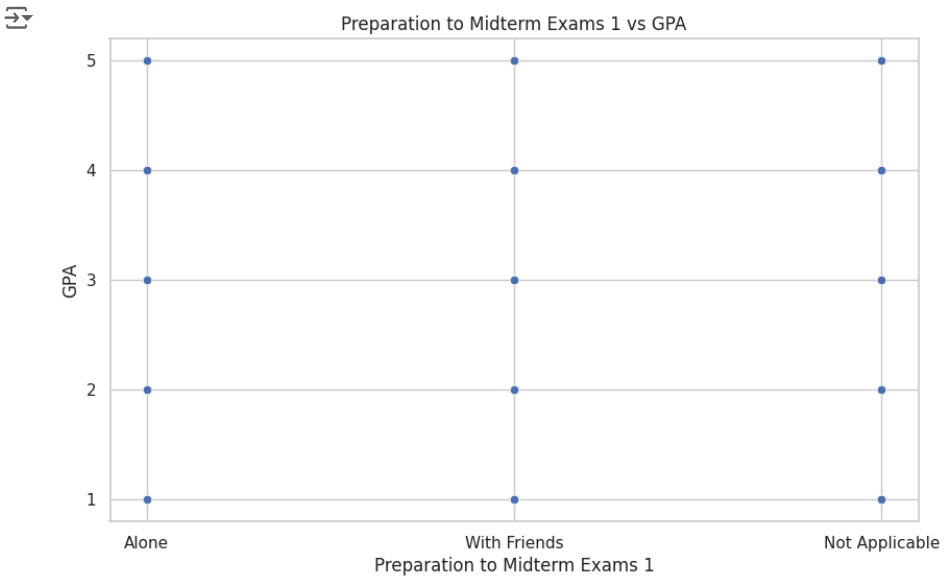
Double-click (or enter) to edit

```
# Bar graph of Weekly study hours vs GPA
plt.figure(figsize=(8, 5))
sns.scatterplot(data=df, x='Weekly study hours', y='Cumulative grade point average in the last semester (/4.00)')
plt.title('Weekly Study Hours vs GPA')
plt.xlabel('Weekly Study Hours')
plt.ylabel('GPA')

# Set x-axis ticks to display integer values only
plt.xticks(range(df['Weekly study hours'].astype(int).min(), df['Weekly study hours'].astype(int).max() + 1))

# Set y-axis ticks to display integer values only
plt.yticks(range(int(df['Cumulative grade point average in the last semester (/4.00)'].min()), int(df['Cumulative grade point average in the

plt.show()
```

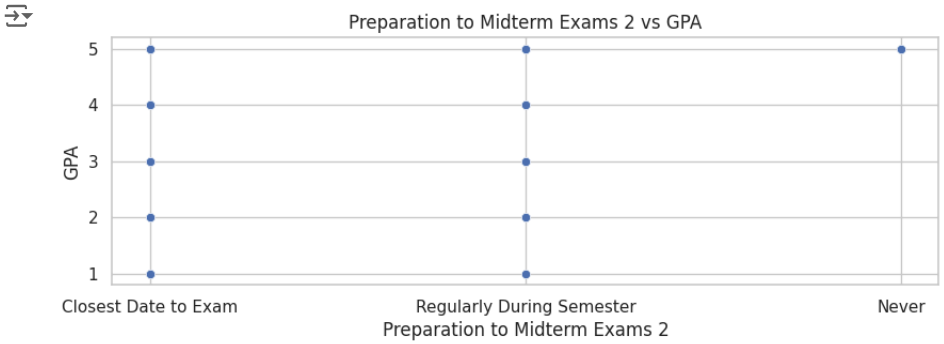

```
# Bar plot of average GPA vs attendance to classes
df['Attendance to classes'] = df['Attendance to classes'].replace({1: 'Yes', 2: 'No'})

plt.figure(figsize=(8, 6))
sns.barplot(data=df, x='Attendance to classes', y='Cumulative grade point average in the last semester (/4.00)')
plt.title('Average GPA by Attendance to Classes')
plt.xlabel('Attendance to Classes')
plt.ylabel('Average GPA')
plt.show()
```



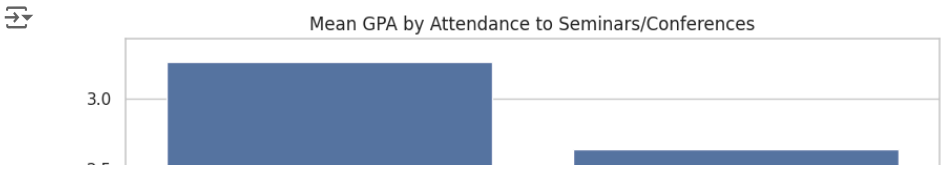
```
# Scatter plot of Preparation to midterm exams 1 vs GPA
# Define the mapping of numerical values to categorical labels
preparation_labels = {
    1: 'Alone',
    2: 'With Friends',
    3: 'Not Applicable'
}

# Map numerical values to categorical labels in the DataFrame
df['Preparation to midterm exams 1'] = df['Preparation to midterm exams 1'].map(preparation_labels)

# Scatter plot of Preparation to midterm exams 1 vs GPA with categorical labels
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Preparation to midterm exams 1', y='Cumulative grade point average in the last semester (/4.00)')
plt.title('Preparation to Midterm Exams 1 vs GPA')
plt.xlabel('Preparation to Midterm Exams 1')
plt.ylabel('GPA')

# Set y-axis ticks to display integer values only
plt.yticks(range(int(df['Cumulative grade point average in the last semester (/4.00)'].min()), int(df['Cumulative grade point average in the

plt.show()
```



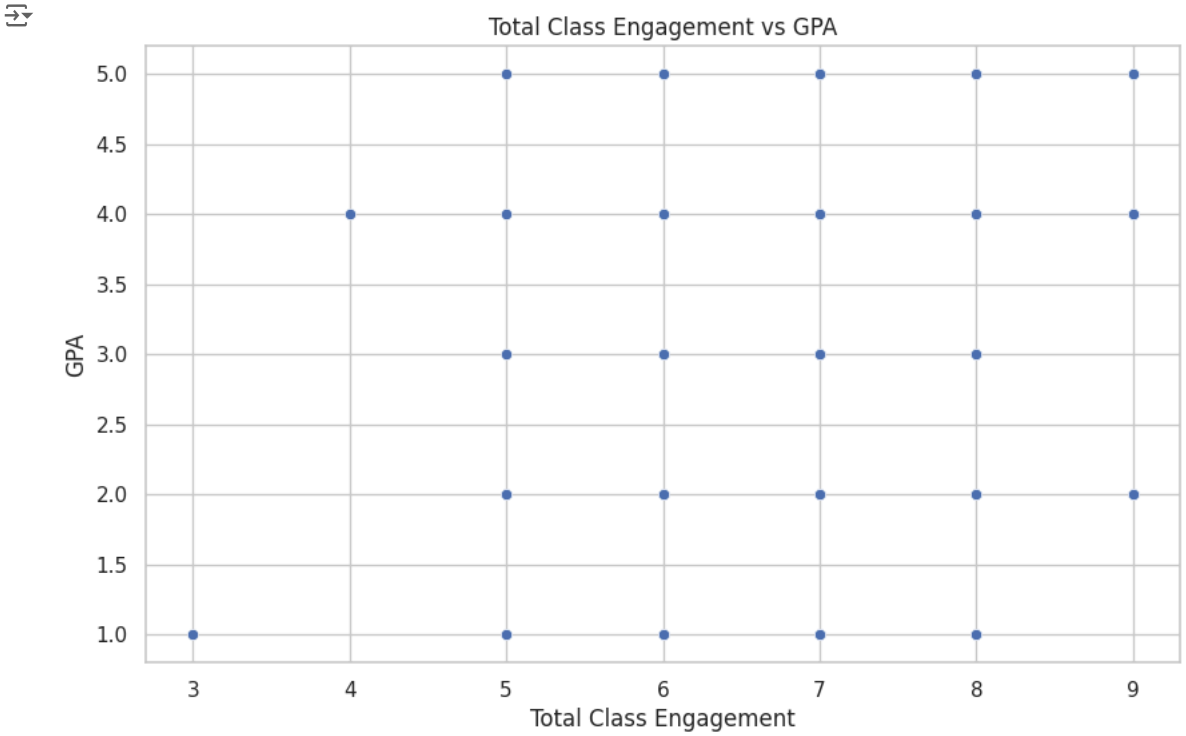
```
# Scatter plot of Preparation to midterm exams 2 vs GPA
# Define the mapping of numerical values to categorical labels
preparation_labels_2 = {
    1: 'Closest Date to Exam',
    2: 'Regularly During Semester',
    3: 'Never'
}

# Map numerical values to categorical labels in the DataFrame
df['Preparation to midterm exams 2'] = df['Preparation to midterm exams 2'].map(preparation_labels_2)

# Scatter plot of Preparation to midterm exams 2 vs GPA with categorical labels
plt.figure(figsize=(10, 3))
sns.scatterplot(data=df, x='Preparation to midterm exams 2', y='Cumulative grade point average in the last semester (/4.00)')
plt.title('Preparation to Midterm Exams 2 vs GPA')
plt.xlabel('Preparation to Midterm Exams 2')
plt.ylabel('GPA')

# Set y-axis ticks to display integer values only
plt.yticks(range(int(df['Cumulative grade point average in the last semester (/4.00)'].min()), int(df['Cumulative grade point average in the

plt.show()
```




```
# Group the DataFrame by 'Attendance to seminars/conferences related to the department' and calculate the mean GPA for each group
gpa_by_attendance = df.groupby('Attendance to the seminars/conferences related to the department')['Cumulative grade point average in the last

# Plotting the clustered bar chart
plt.figure(figsize=(10, 6))
sns.barplot(data=gpa_by_attendance, x='Attendance to the seminars/conferences related to the department', y='Cumulative grade point average i
plt.title('Mean GPA by Attendance to Seminars/Conferences')
plt.xlabel('Attendance to Seminars/Conferences')
plt.ylabel('Mean GPA')
plt.show()
```




```
# Create a new column "Total Class Engagement" by summing up the values from the three columns
df['Total Class Engagement'] = df['Taking notes in classes'] + df['Listening in classes'] + df['Discussion improves my interest and success i

# Scatter plot of Total Class Engagement vs GPA
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Total Class Engagement', y='Cumulative grade point average in the last semester (/4.00)')
plt.title('Total Class Engagement vs GPA')
plt.xlabel('Total Class Engagement')
plt.ylabel('GPA')
plt.show()
```



Preliminary work on machine learning to make predictions

- 80% of the dataset will be used for training the model
- 20% of the dataset will be reserved for testing.

- **Introduction**

  - **Why was the project undertaken?**
    The purpose of undertaking this project was to investigate the effects of various factors on students' performance and academic success at the end of an academic term to develop an accurate prediction model using data science.

  - **What was the research question, the tested hypothesis or the purpose of the research?**
    To develop an adequate prediction model, which factors based on the available data on the dataset are effective in increasing students' performance and achieving academic success at the end of a semester? Some hypotheses, such as the positive effects of more hours of weekly study, more attendance and participation in classes (seminars/conferences), and proper preparation for mid-semester exams, on increasing students' GPA and success at the end of the semester.

- **Selection of Data**

  - **What is the source of the dataset? Characteristics of data?**
    The CSV dataset used in this project has been presented by Joakim Arvidsson on the Kaggle website, a well-known platform for data science and machine learning datasets. This dataset provides valuable insights into the factors that influence students' educational performance at the end of the academic term. Collected from approximately 145 students, the dataset comprises a rich array of variables that capture various aspects of students' demographics, academic activities, and personal circumstances.

  - **Any munging or feature engineering?**
    The primary aim of the dataset is to serve as a comprehensive resource for conducting a case study on the efficient factors that impact students' academic outcomes. This can include exploring how different variables interact and contribute to the overall performance of students in their studies. We identified the missing values and removed unnecessary rows and columns. We also added value to some non-numeric features.

- **Methods**

  - **What materials/APIs/tools were used or who was included in answering the research question?**

We used Python to program our project code, the numpy and pandas for data management and operation, the matplotlib and seaborn libraries for platting, and the sklearn library for training set and regression.

- **Results**

  - **What answer was found to the research question; what did the study find? Was the tested hypothesis true? Any visualizations?**
    Yes. The results and plots proved that a certain amount of weekly study hours helps to increase the students' GPA. Less than that will result negatively, and spending more weekly hours will not be helpful either. Based on the sample in this dataset, this specific amount of study should be at least two hours per week. On the other hand, these results proved that the attendance of students to their classes, conferences/seminars and their total engagements in classes have direct correlations with their GPA levels.

- **Discussion**

  - **What might the answer imply and why does it matter? How does it fit in with what other researchers have found? What are the perspectives for future research? Survey about the tools investigated for this assignment.**
    The results of this research are consistent with similar research by others in this field. The tools used in this research were helpful and efficient in finding these results and can be useful in performing future projects and research.

- **Summary**

  - **Most important findings.**
    The project aimed to explore the factors affecting students' academic performance and success to develop an accurate prediction model using data science. The primary research question focused on identifying which factors in the dataset effectively enhance students' performance and academic success by the semester's end, with hypotheses including the positive impact of weekly study hours, class attendance, participation in seminars/conferences, and mid-semester exam preparation on students' GPA. The dataset, sourced from Kaggle and provided by Joakim Arvidsson, contained data from approximately 145 students covering demographics, academic activities, and personal circumstances. The findings revealed that weekly study hours significantly impact GPA, with at least two hours per week being beneficial, and that attendance in classes and participation in seminars/conferences directly correlate with

higher GPA levels. These results align with existing research, emphasizing the importance of regular study and class engagement.