Project 4: Analysis on Students' Performance

Team 9 - Rachel Schoen, Andrea Wu, Mahind Rao, Mahalel Peter

Agenda

01

Problem Statement

02

Data Description & Preparation

03

Exploratory Analysis

04

Database Design

05

Model Designs & Evaluations

06

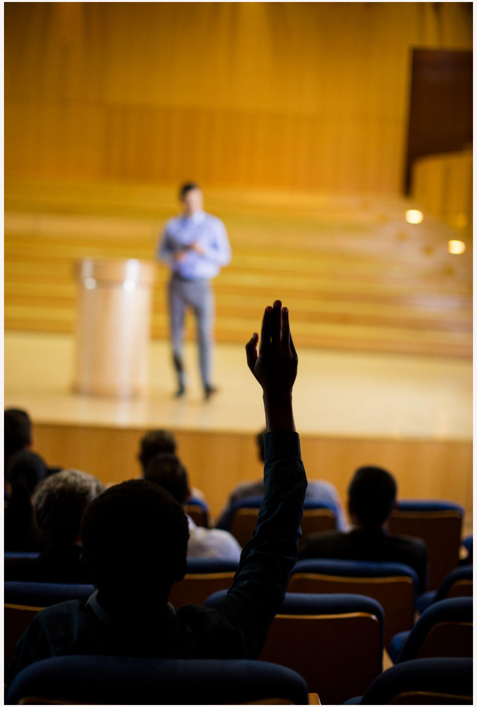
**Limitat &
Future
Considerations**

Problem Statement

What are the key factors that influence the success of a student?

- How do students with additional work responsibilities fare academically compared to those without?
- How does the type of accommodation affect students' study habits and performance?
- How does discussion and group work influence students' interest and success in their courses?

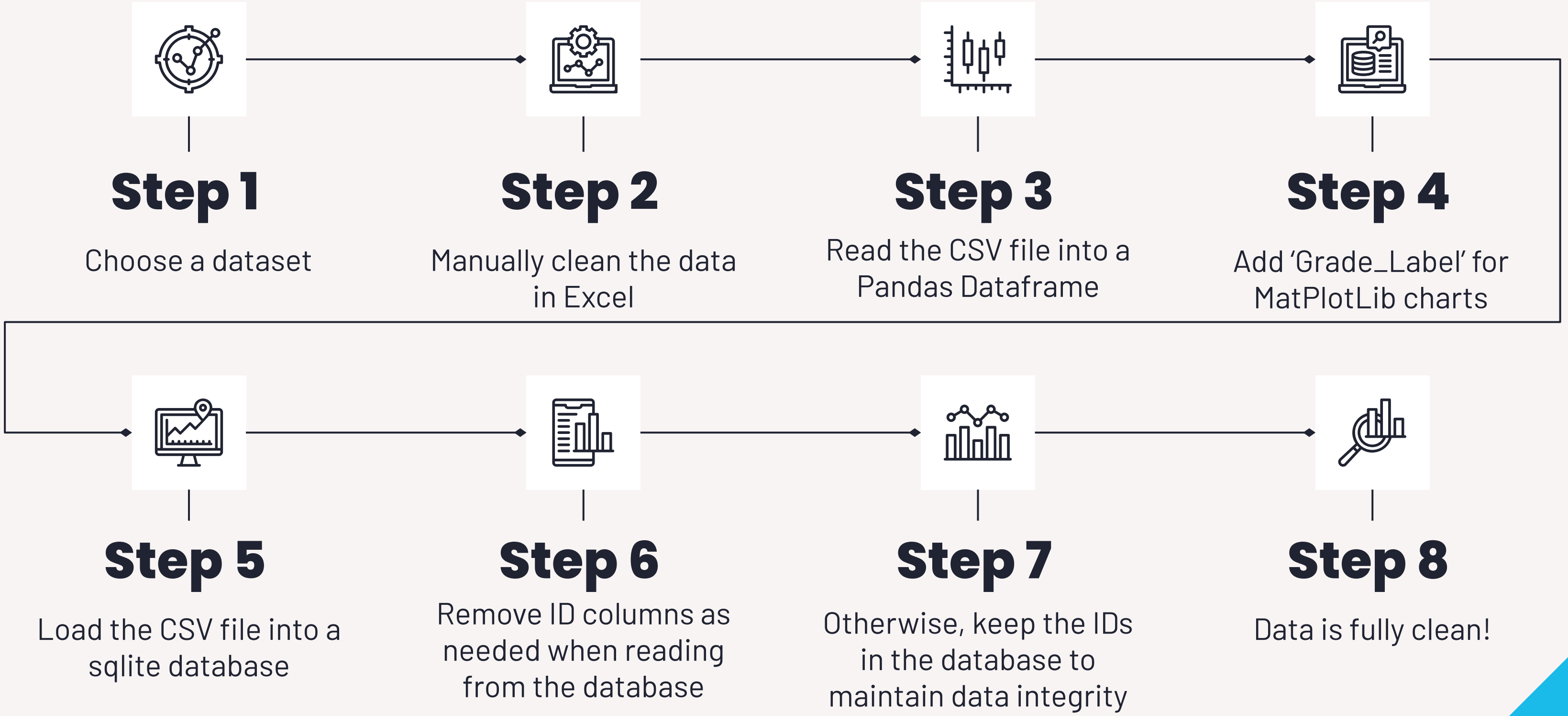
And most importantly:
- **Can we predict a student's academic performance (e.g., grade point average or success in courses) based on their personal background, study habits, and extracurricular activities?**



Data Description

- **Dataset: Students performance**
- **Format: CSV file**
- **Data Source: Dataset from Kaggle**
  https://www.kaggle.com/datasets/joebeachcapital/students-performance/data
- **Data Structure**
  - Data of 145 students
  - 33 Columns
  - Majority is categorical data

Data Preparation



Step 1

Choose a dataset

Step 2

Manually clean the data in Excel

Step 3

Read the CSV file into a Pandas Dataframe

Step 4

Add 'Grade_Label' for MatPlotLib charts

Step 5

Load the CSV file into a sqlite database

Step 6

Remove ID columns as needed when reading from the database

Step 7

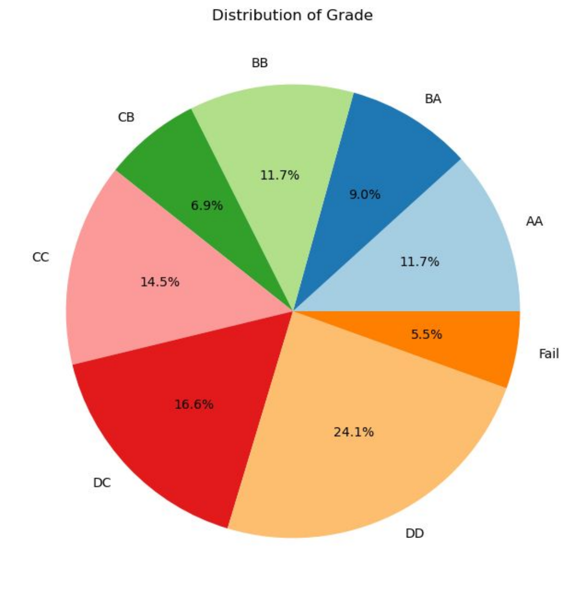
Otherwise, keep the IDs in the database to maintain data integrity

Step 8

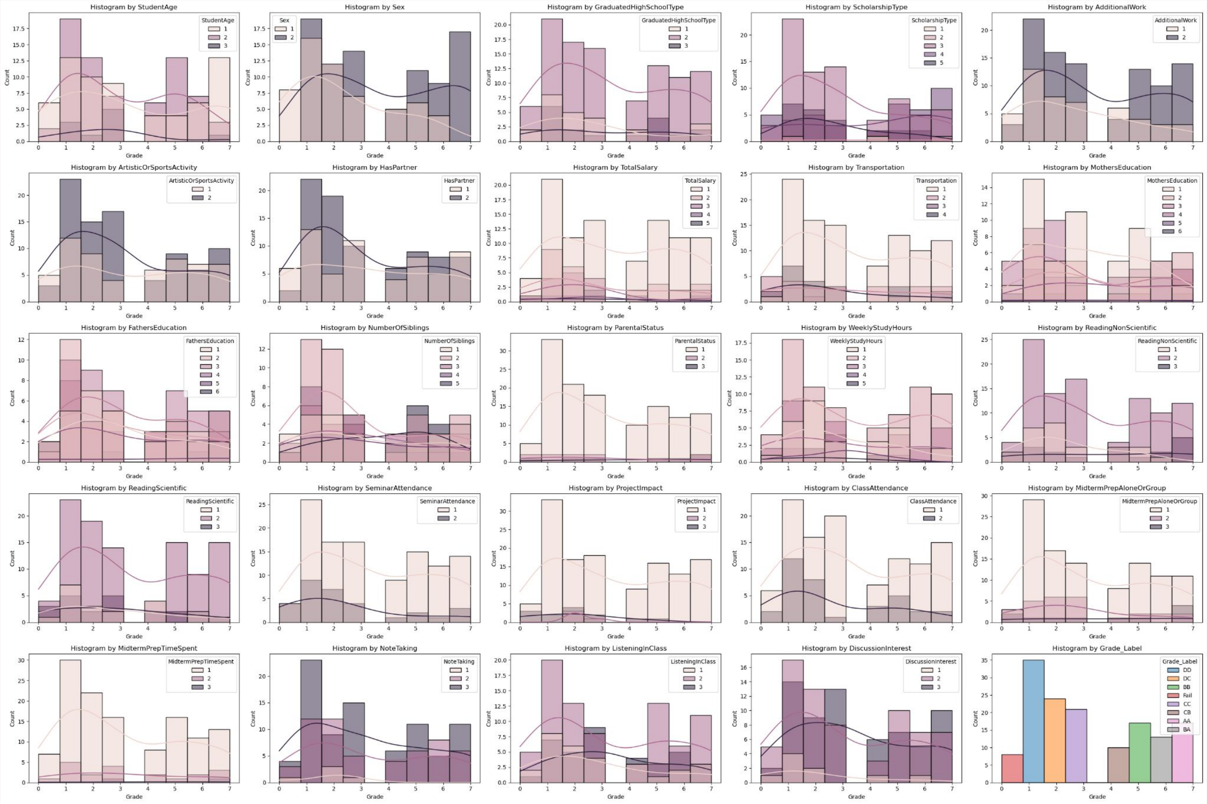
Data is fully clean!

Exploratory Analysis (1/3)

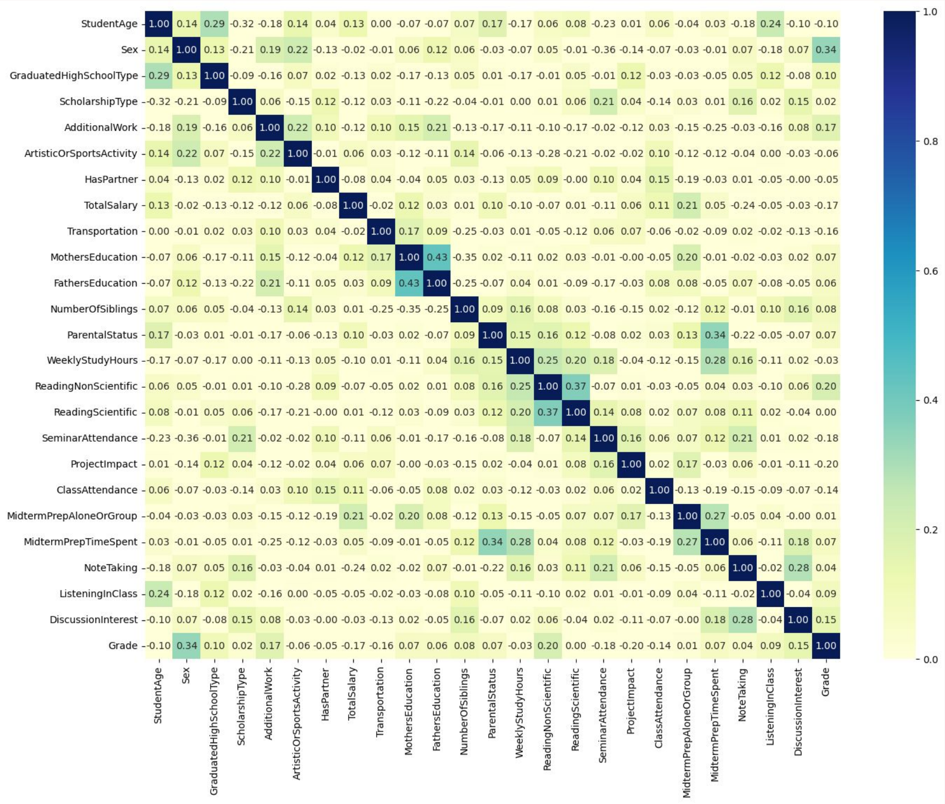
- **Target Variable: The final grade of the students**
- **Interactive Dashboard:** https://andreawcy.github.io/Project_4_Students-performance-analysis/



Exploratory Analysis (2/3)

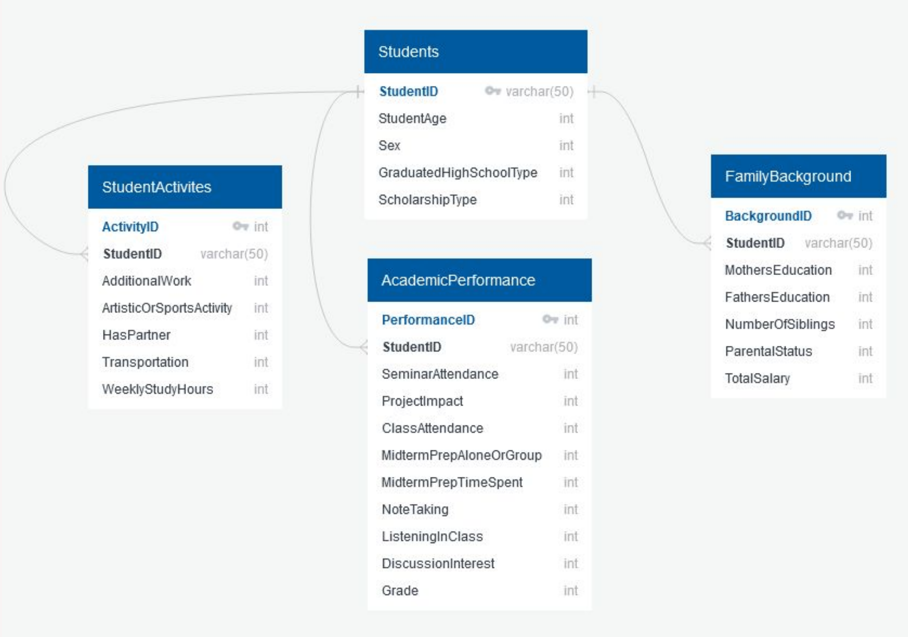


Exploratory Analysis (3/3)



	StudentAge	Sex	GraduatedHighSchoolType	ScholarshipType	AdditionalWork	ArtisticOrSportsActivity	HasPartner	TotalSalary	Transportation	MothersEducation	FathersEducation	NumberOfSiblings	ParentalStatus	WeeklyStudyHours	ReadingNonScientific	ReadingScientific	SeminarAttendance	ProjectImpact	ClassAttendance	MidtermPrepAloneOrGroup	MidtermPrepTimeSpent	NoteTaking	ListeningInClass	DiscussionInterest	Grade
StudentAge	1.00	0.14	0.29	-0.32	-0.18	0.14	0.04	0.13	0.00	-0.07	-0.07	0.07	0.17	-0.17	0.06	0.08	-0.23	0.01	0.06	-0.04	0.03	-0.18	0.24	-0.10	-0.10
Sex	0.14	1.00	0.13	-0.21	0.19	0.22	-0.13	-0.02	-0.01	0.06	0.12	0.06	-0.03	-0.07	0.05	-0.01	-0.36	-0.14	-0.07	-0.03	-0.01	0.07	-0.18	0.07	0.34
GraduatedHighSchoolType	0.29	0.13	1.00	-0.09	-0.16	0.07	0.02	-0.13	0.02	-0.17	-0.13	0.05	0.01	-0.17	-0.01	0.05	-0.01	0.12	-0.03	-0.03	-0.05	0.05	0.12	-0.08	0.10
ScholarshipType	-0.32	-0.21	-0.09	1.00	0.06	-0.15	0.12	-0.12	0.03	-0.11	-0.22	-0.04	-0.01	0.00	0.01	0.06	0.21	0.04	-0.14	0.03	0.01	0.16	0.02	0.15	0.02
AdditionalWork	-0.18	0.19	-0.16	0.06	1.00	0.22	0.10	-0.12	0.10	0.15	0.21	-0.13	-0.17	-0.11	-0.10	-0.17	-0.02	-0.12	0.03	-0.15	-0.25	-0.03	-0.16	0.08	0.17
ArtisticOrSportsActivity	0.14	0.22	0.07	-0.15	0.22	1.00	-0.01	0.06	0.03	-0.12	-0.11	0.14	-0.06	-0.13	-0.28	-0.21	-0.02	-0.02	0.10	-0.12	-0.12	-0.04	0.00	-0.03	-0.06
HasPartner	0.04	-0.13	0.02	0.12	0.10	-0.01	1.00	-0.08	0.04	-0.04	0.05	0.03	-0.13	0.05	0.09	-0.00	0.10	0.04	0.15	-0.19	-0.03	0.01	-0.05	-0.00	-0.05
TotalSalary	0.13	-0.02	-0.13	-0.12	-0.12	0.06	-0.08	1.00	-0.02	0.12	0.03	0.01	0.10	-0.10	-0.07	0.01	-0.11	0.06	0.11	0.21	0.05	-0.24	-0.05	-0.03	-0.17
Transportation	0.00	-0.01	0.02	0.03	0.10	0.03	0.04	-0.02	1.00	0.17	0.09	-0.25	-0.03	0.01	-0.05	-0.12	0.06	0.07	-0.06	-0.02	-0.09	0.02	-0.02	-0.13	-0.16
MothersEducation	-0.07	0.06	-0.17	-0.11	0.15	-0.12	-0.04	0.12	0.17	1.00	0.43	-0.35	0.02	-0.11	0.02	0.03	-0.01	-0.00	-0.05	0.20	-0.01	-0.02	-0.03	0.02	0.07
FathersEducation	-0.07	0.12	-0.13	-0.22	0.21	-0.11	0.05	0.03	0.09	0.43	1.00	-0.25	-0.07	0.04	0.01	-0.09	-0.17	-0.03	0.08	0.08	-0.05	0.07	-0.08	-0.05	0.06
NumberOfSiblings	0.07	0.06	0.05	-0.04	-0.13	0.14	0.03	0.01	-0.25	-0.35	-0.25	1.00	0.09	0.16	0.08	0.03	-0.16	-0.15	0.02	-0.12	0.12	-0.01	0.10	0.16	0.08
ParentalStatus	0.17	-0.03	0.01	-0.01	-0.17	-0.06	-0.13	0.10	-0.03	0.02	-0.07	0.09	1.00	0.15	0.16	0.12	-0.08	0.02	0.03	0.13	0.34	-0.22	-0.05	-0.07	0.07
WeeklyStudyHours	-0.17	-0.07	-0.17	0.00	-0.11	-0.13	0.05	-0.10	0.01	-0.11	0.04	0.16	0.15	1.00	0.25	0.20	0.18	-0.04	-0.12	-0.15	0.28	0.16	-0.11	0.02	-0.03
ReadingNonScientific	0.06	0.05	-0.01	0.01	-0.10	-0.28	0.09	-0.07	-0.05	0.02	0.01	0.08	0.16	0.25	1.00	0.37	-0.07	0.01	-0.03	-0.05	0.04	0.03	-0.10	0.06	0.20
ReadingScientific	0.08	-0.01	0.05	0.06	-0.17	-0.21	-0.00	0.01	-0.12	0.03	-0.09	0.03	0.12	0.20	0.37	1.00	0.14	0.08	0.02	0.07	0.08	0.11	0.02	-0.04	0.00
SeminarAttendance	-0.23	-0.36	-0.01	0.21	-0.02	-0.02	0.10	-0.11	0.06	-0.01	-0.17	-0.16	-0.08	0.18	-0.07	0.14	1.00	0.16	0.06	0.07	0.12	0.21	0.01	0.02	-0.18
ProjectImpact	0.01	-0.14	0.12	0.04	-0.12	-0.02	0.04	0.06	0.07	-0.00	-0.03	-0.15	0.02	-0.04	0.01	0.08	0.16	1.00	0.02	0.17	-0.03	0.06	-0.01	-0.11	-0.20
ClassAttendance	0.06	-0.07	-0.03	-0.14	0.03	0.10	0.15	0.11	-0.06	-0.05	0.08	0.02	0.03	-0.12	-0.03	0.02	0.06	0.02	1.00	-0.13	-0.19	-0.15	-0.09	-0.07	-0.14
MidtermPrepAloneOrGroup	-0.04	-0.03	-0.03	0.03	-0.15	-0.12	-0.19	0.21	-0.02	0.20	0.08	-0.12	0.13	-0.15	-0.05	0.07	0.07	0.17	-0.13	1.00	0.27	-0.05	0.04	-0.00	0.01
MidtermPrepTimeSpent	0.03	-0.01	-0.05	0.01	-0.25	-0.12	-0.03	0.05	-0.09	-0.01	-0.05	0.12	0.34	0.28	0.04	0.08	0.12	-0.03	-0.19	0.27	1.00	0.06	-0.11	0.18	0.07
NoteTaking	-0.18	0.07	0.05	0.16	-0.03	-0.04	0.01	-0.24	0.02	-0.02	0.07	-0.01	-0.22	0.16	0.03	0.11	0.21	0.06	-0.15	-0.05	0.06	1.00	-0.02	0.28	0.04
ListeningInClass	0.24	-0.18	0.12	0.02	-0.16	0.00	-0.05	-0.05	-0.02	-0.03	-0.08	0.10	-0.05	-0.11	-0.10	0.02	0.01	-0.01	-0.09	0.04	-0.11	-0.02	1.00	-0.04	0.09
DiscussionInterest	-0.10	0.07	-0.08	0.15	0.08	-0.03	-0.00	-0.03	-0.13	0.02	-0.05	0.16	-0.07	0.02	0.06	-0.04	0.02	-0.11	-0.07	-0.00	0.18	0.28	-0.04	1.00	0.15
Grade	-0.10	0.34	0.10	0.02	0.17	-0.06	-0.05	-0.17	-0.16	0.07	0.06	0.08	0.07	-0.03	0.20	0.00	-0.18	-0.20	-0.14	0.01	0.07	0.04	0.09	0.15	1.00

Database Design



Students	
StudentID	varchar(50)
StudentAge	int
Sex	int
GraduatedHighSchoolType	int
ScholarshipType	int

StudentActivites	
ActivityID	int
StudentID	varchar(50)
AdditionalWork	int
ArtisticOrSportsActivity	int
HasPartner	int
Transportation	int
WeeklyStudyHours	int

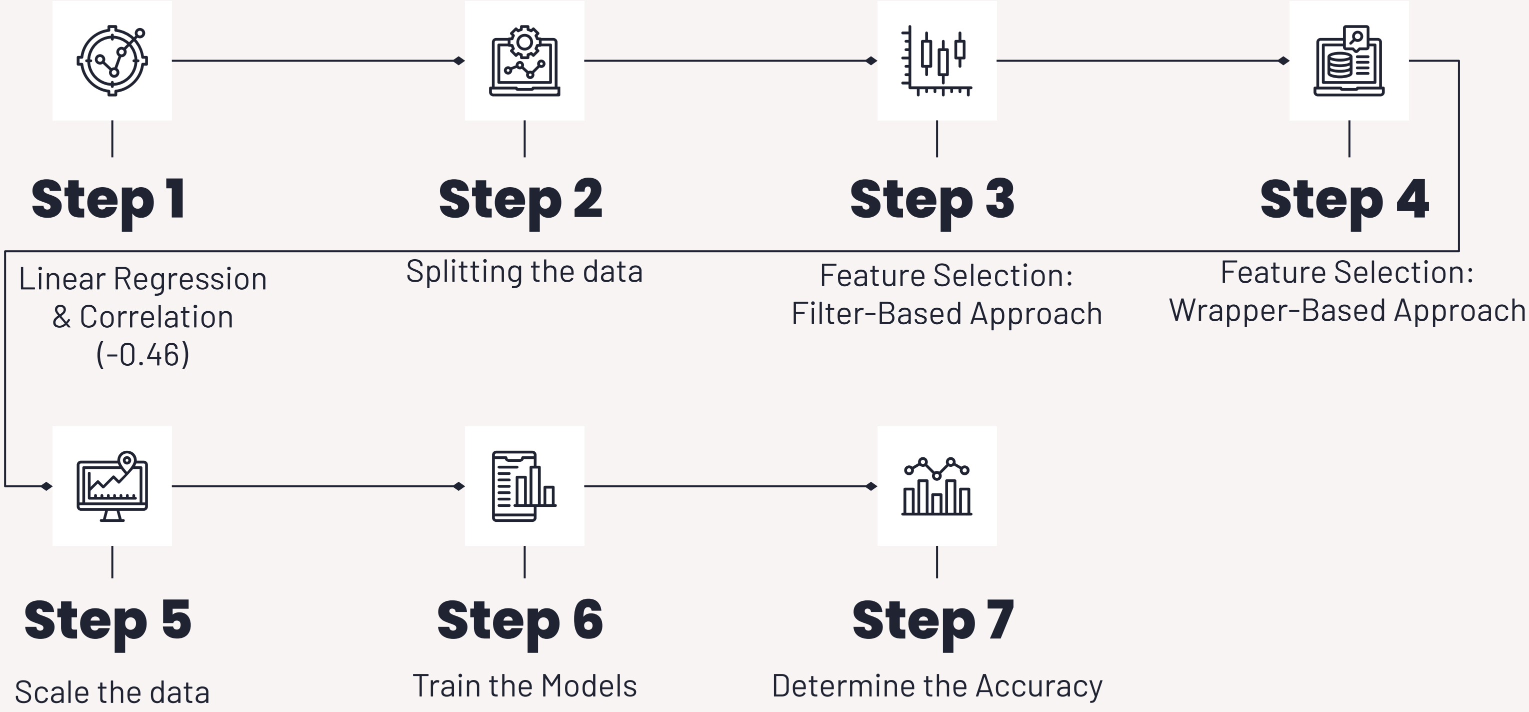
AcademicPerformance	
PerformanceID	int
StudentID	varchar(50)
SeminarAttendance	int
ProjectImpact	int
ClassAttendance	int
MidtermPrepAloneOrGroup	int
MidtermPrepTimeSpent	int
NoteTaking	int
ListeningInClass	int
DiscussionInterest	int
Grade	int

FamilyBackground	
BackgroundID	int
StudentID	varchar(50)
MothersEducation	int
FathersEducation	int
NumberOfSiblings	int
ParentalStatus	int
TotalSalary	int

The database was designed with four tables in mind, connected through primary and foreign keys. Each table stores related information, ensuring data integrity and allowing for more efficient queries.

1. **Students**
2. **Student Activities**
3. **Academic Performance**
4. **Family Background**

Modelling Experiment Design



Step 1

Linear Regression & Correlation (-0.46)

Step 2

Splitting the data

Step 3

Feature Selection: Filter-Based Approach

Step 4

Feature Selection: Wrapper-Based Approach

Step 5

Scale the data

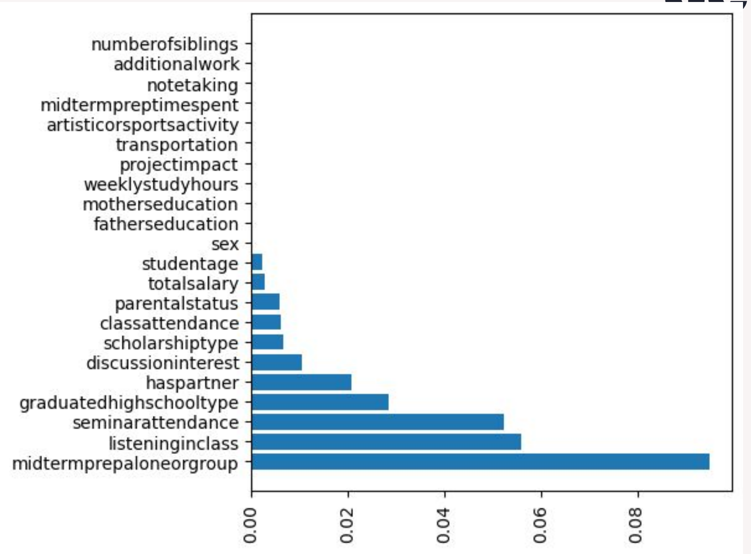
Step 6

Train the Models

Step 7

Determine the Accuracy

Filter-Based Approach of Feature Selection



Mutual Information Classification

Select subsets of features based on their relationship with the target.

Advantages:

- Faster than using a wrapper method
- More generalized and easier to interpret

Disadvantage:

- Each feature is looked at in isolation, i.e. it's prone to discarding useful features that are weak predictors on their own, but useful when combined with others

Basically, according to the mutual_info_clasif result above, in these particular train and test buckets, whether you prepped alone or in a group is the most “important” feature to consider.

Wrapper-Based Approach of Feature Selection

RandomForestClassifier

A supervised machine learning algorithm based on ensemble learning, a technique that creates multiple models using decision trees and then combines them to produce improved results.

Advantages:

- **High accuracy**— because it uses multiple decision trees, then merges them together, it lessens the variation associated with individual trees.
- **Adaptable**— can be applied to many datasets and problem areas.

Disadvantages:

- **Computationally expensive**— it uses a lot of computing power and memory.
- **High prediction time**— it can take longer to make predictions than other algorithms, making it ill-fit for real-time applications.

Wrapper-Based Approach of Feature Selection

SequentialFeatureSelector

An algorithm that iteratively adds or removes features from a dataset in order to improve the performance of a predictive model.

1) Initialize with a predictive model, number of features to select, the scoring metric, and a tolerance for improvement → 2) Fit the model on full set of features → 3) Evaluate the model on the training set using the scoring metric → 4) The feature that most improves the model's cross-validation score is added to the set, or the feature that least reduces the cross-validation score is removed from the set → 5) Repeat steps 2-4 until the desired number of features has been selected.

Advantages:

- **Simple**– it's a simple and efficient algorithm, especially for a wrapper approach.
- **Flexible**– it can be used with any type of predictive model.

Disadvantages:

- **Bias**– it can be biased towards features that are highly correlated with the target feature.
- **Computationally Expensive**– it uses a lot of computing power and memory.

Modeling the Data

1. **Choose the Feature Selection**

2. **Scale the data**
   Using StandardScaler()

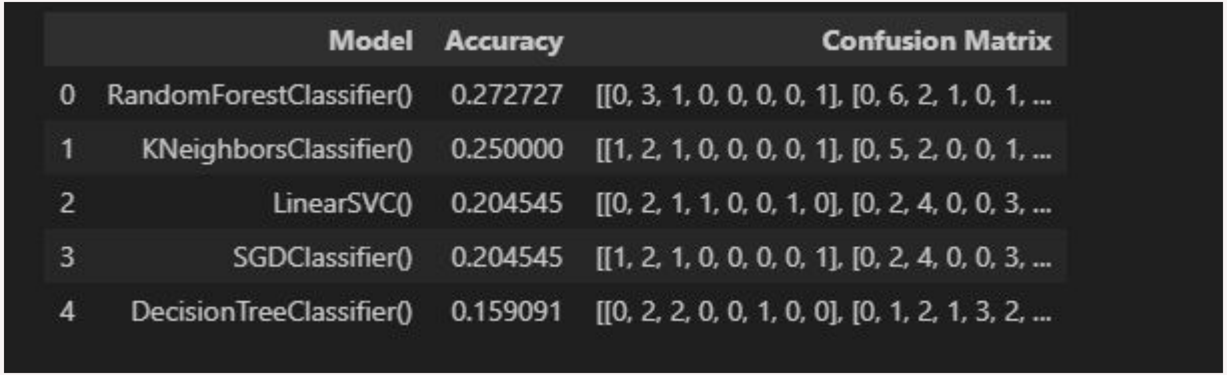
3. **Train the Models**

Models used:

- LinearSVC()
- SGDClassifier()
- DecisionTreeClassifier()
- KNeighborsClassifier()
- RandomForestClassifier()

* More models, such as a Logistic Regression model, can be found in the project notebook.

Model Evaluation



	Model	Accuracy	Confusion Matrix
0	RandomForestClassifier()	0.272727	[[0, 3, 1, 0, 0, 0, 0, 1], [0, 6, 2, 1, 0, 1, ...
1	KNeighborsClassifier()	0.250000	[[1, 2, 1, 0, 0, 0, 0, 1], [0, 5, 2, 0, 0, 1, ...
2	LinearSVC()	0.204545	[[0, 2, 1, 1, 0, 0, 1, 0], [0, 2, 4, 0, 0, 3, ...
3	SGDClassifier()	0.204545	[[1, 2, 1, 0, 0, 0, 0, 1], [0, 2, 4, 0, 0, 3, ...
4	DecisionTreeClassifier()	0.159091	[[0, 2, 2, 0, 0, 1, 0, 0], [0, 1, 2, 1, 3, 2, ...

After training each model in an iterative loop, I stored the accuracy result and the confusion matrix for each model in a Pandas dataframe. As you can see, my models weren't so accurate...

Key Findings

Based on the dataset we gathered we couldn’t predict a student’s academic performance (e.g., grade point average or success in courses) based on their personal background, study habits, and extracurricular activities because the accuracy was low in every model.

Moreover, there were no key factors that influence the success of a student.

Limitations

Data Structure

The dataset is primarily categorical, which may not provide sufficient detail for models to identify complex patterns or make accurate predictions.

Multicollinearity

In other words, it’s difficult to determine the individual effects of each variable on the student’s performance, because many of them are dependent on each other.

Limited Data

While the dataset had a wide breadth of categories, there were only a little under 150 entries in the dataset. This helps to explain the poor accuracy of our models.

Considerations for Future

Expand Data Sources:

Integrate additional data such as student surveys and extracurricular participation records.

Enhance Data Accuracy:

Implement automated validation tools and advanced imputation techniques to address data gaps and ensure accuracy.

THANKS!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**