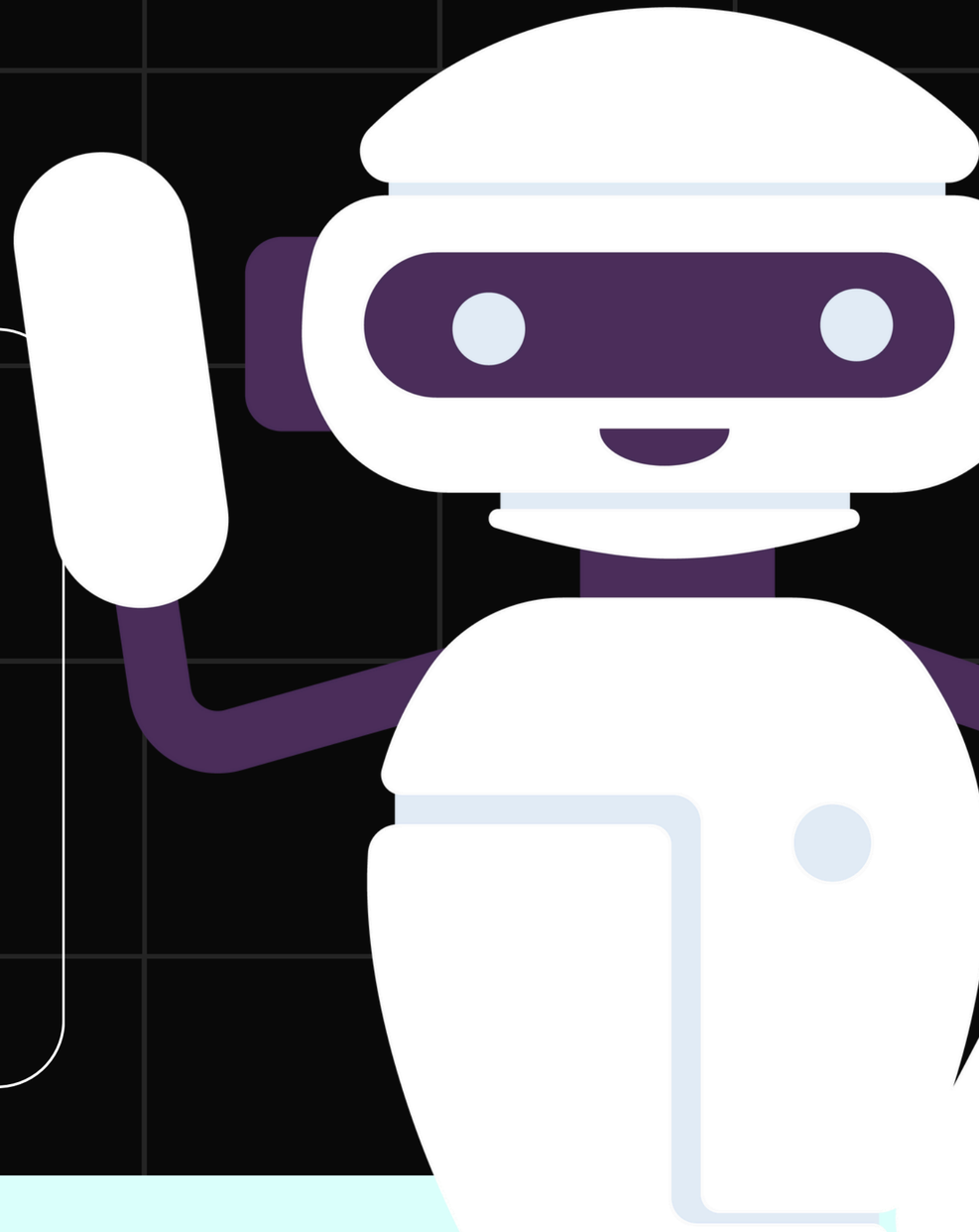
A Horror Story

Predicting the unpredictable

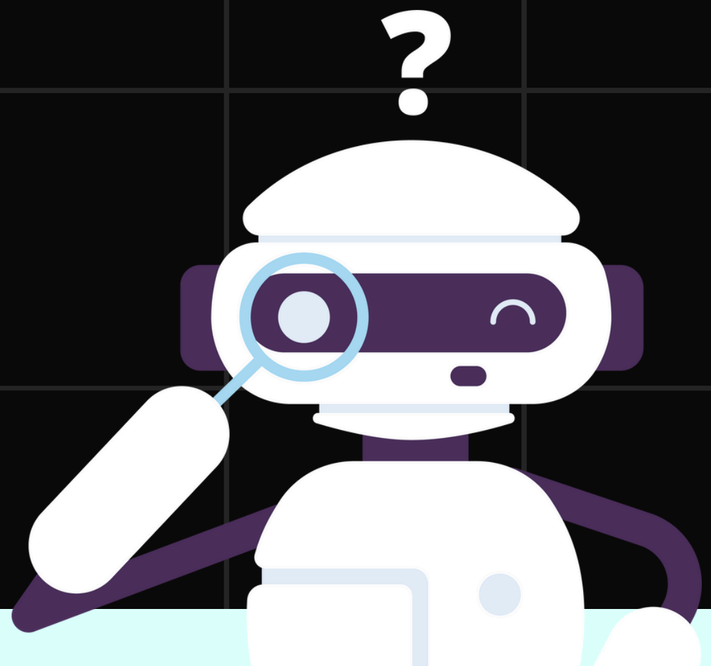
Presented by **Enia, Javier and Daniela**

Index

1. Exploring the dataset
2. Exploratory Data Analysis (EDA)
3. Transforming data & Testing data
4. Results and Conclusions



Exploring the dataset



Unknown gender

What is 0 and 1? We still don’t know it.

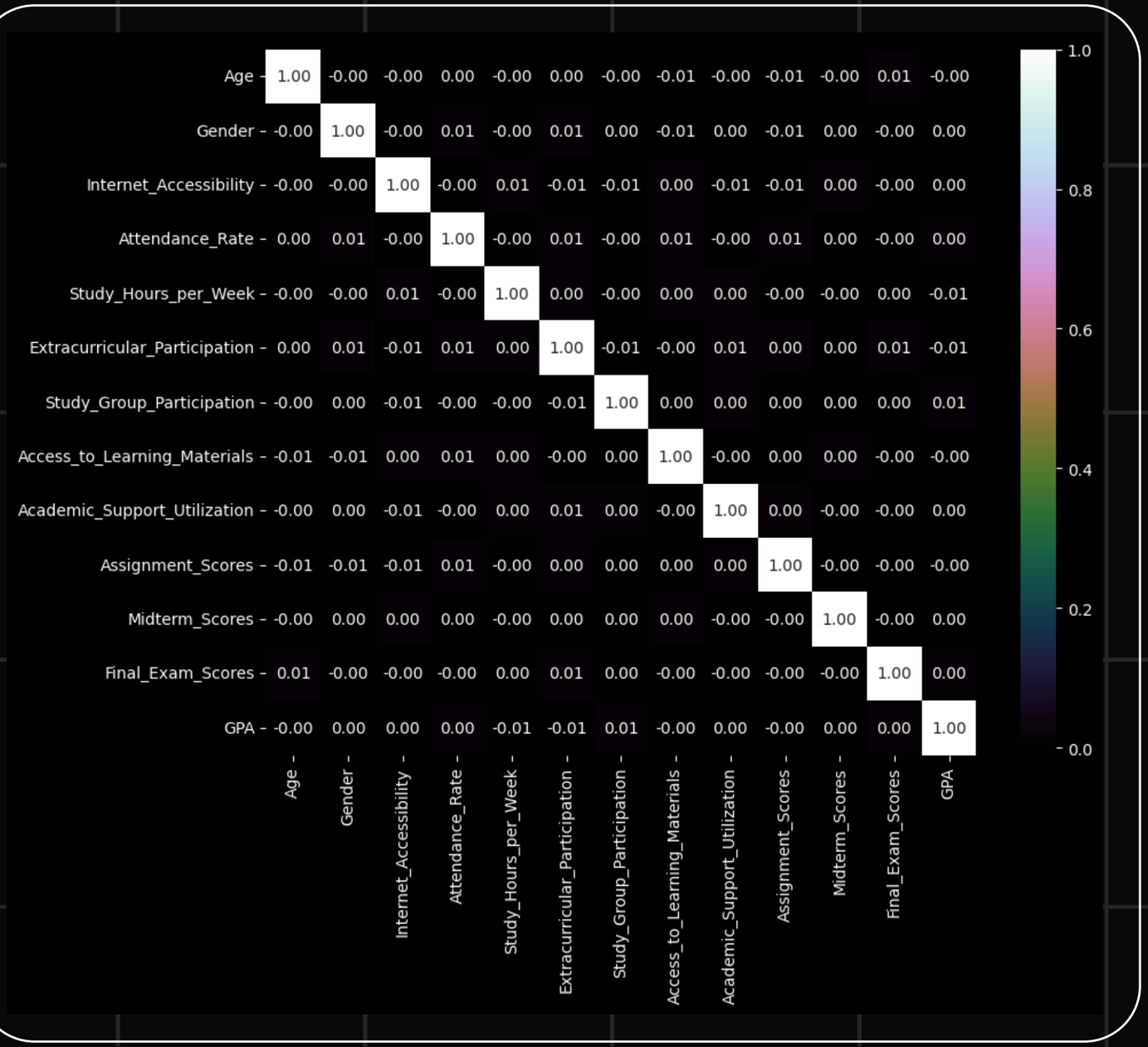
What do we want to predict?

Final Exam Score or GPA?

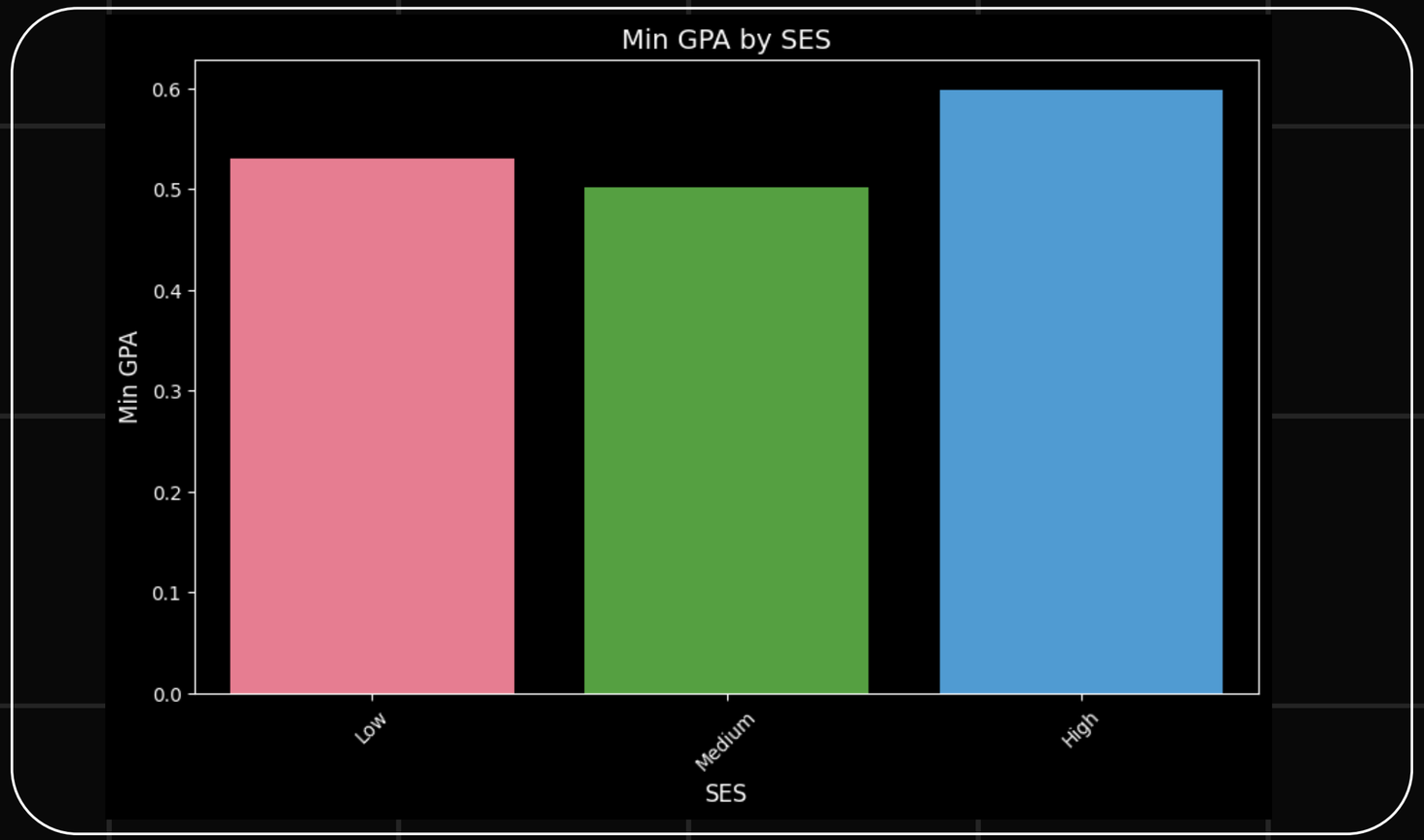
What variables affect the most the GPA?

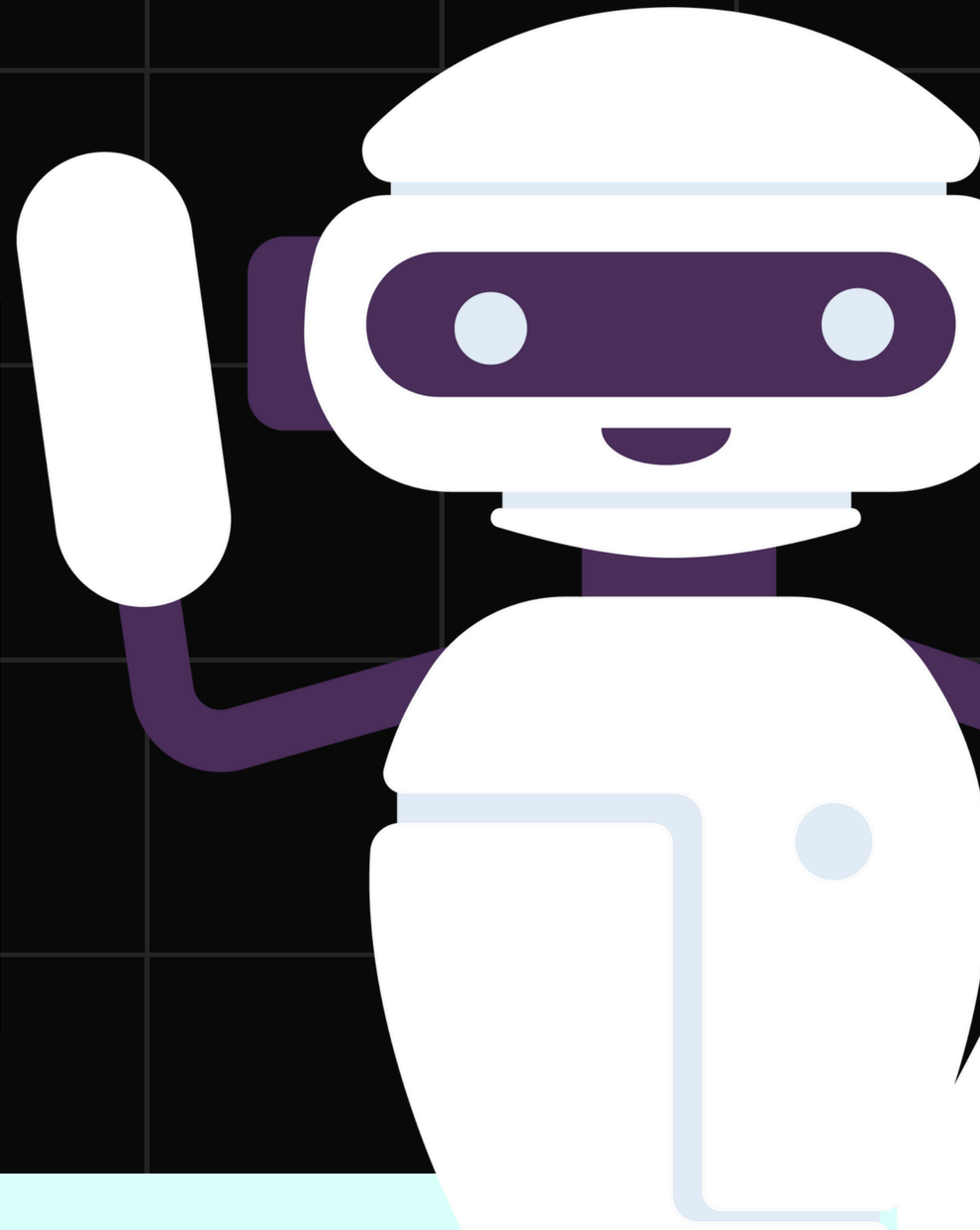
No correlation.

Heatmap

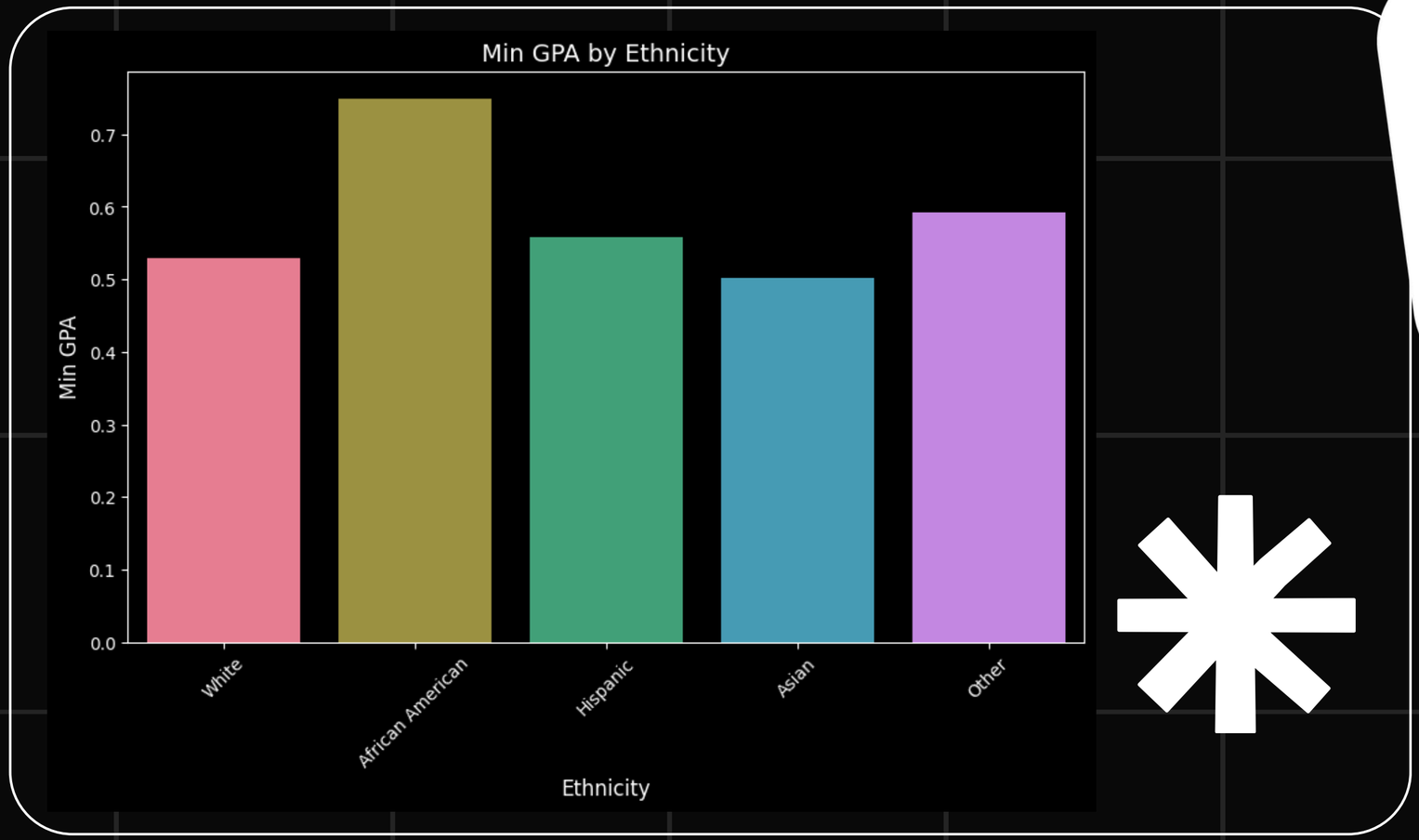

Age 1.00 -0.00 -0.00 0.00 -0.00 0.00 -0.00 -0.01 -0.00 -0.01 -0.00 0.01 -0.00
Gender -0.00 1.00 -0.00 0.01 -0.00 0.01 0.00 -0.01 0.00 -0.01 0.00 -0.00 0.00
Internet_Accessibility -0.00 -0.00 1.00 -0.00 0.01 -0.01 -0.01 0.00 -0.01 -0.01 0.00 -0.00 0.00
Attendance_Rate 0.00 0.01 -0.00 1.00 -0.00 0.01 -0.00 0.01 -0.00 0.01 0.00 -0.00 0.00
Study_Hours_per_Week -0.00 -0.00 0.01 -0.00 1.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 0.00 -0.01
Extracurricular_Participation 0.00 0.01 -0.01 0.01 0.00 1.00 -0.01 -0.00 0.01 0.00 0.00 0.01 -0.01
Study_Group_Participation -0.00 0.00 -0.01 -0.00 -0.00 -0.01 1.00 0.00 0.00 0.00 0.00 0.00 0.01
Access_to_Learning_Materials -0.01 -0.01 0.00 0.01 0.00 -0.00 0.00 1.00 -0.00 0.00 0.00 -0.00 -0.00
Academic_Support_Utilization -0.00 0.00 -0.01 -0.00 0.00 0.01 0.00 -0.00 1.00 0.00 -0.00 -0.00 0.00
Assignment_Scores -0.01 -0.01 -0.01 0.01 -0.00 0.00 0.00 0.00 0.00 1.00 -0.00 -0.00 -0.00
Midterm_Scores -0.00 0.00 0.00 0.00 -0.00 0.00 0.00 0.00 -0.00 -0.00 1.00 -0.00 0.00
Final_Exam_Scores 0.01 -0.00 -0.00 -0.00 0.00 0.01 0.00 -0.00 -0.00 -0.00 -0.00 1.00 0.00
GPA -0.00 0.00 0.00 0.00 -0.01 -0.01 0.01 -0.00 0.00 -0.00 0.00 0.00 1.00
Age
Gender
Internet_Accessibility
Attendance_Rate
Study_Hours_per_Week
Extracurricular_Participation
Study_Group_Participation
Access_to_Learning_Materials
Academic_Support_Utilization
Assignment_Scores
Midterm_Scores
Final_Exam_Scores
GPA
1.0
0.8
0.6
0.4
0.2
0.0

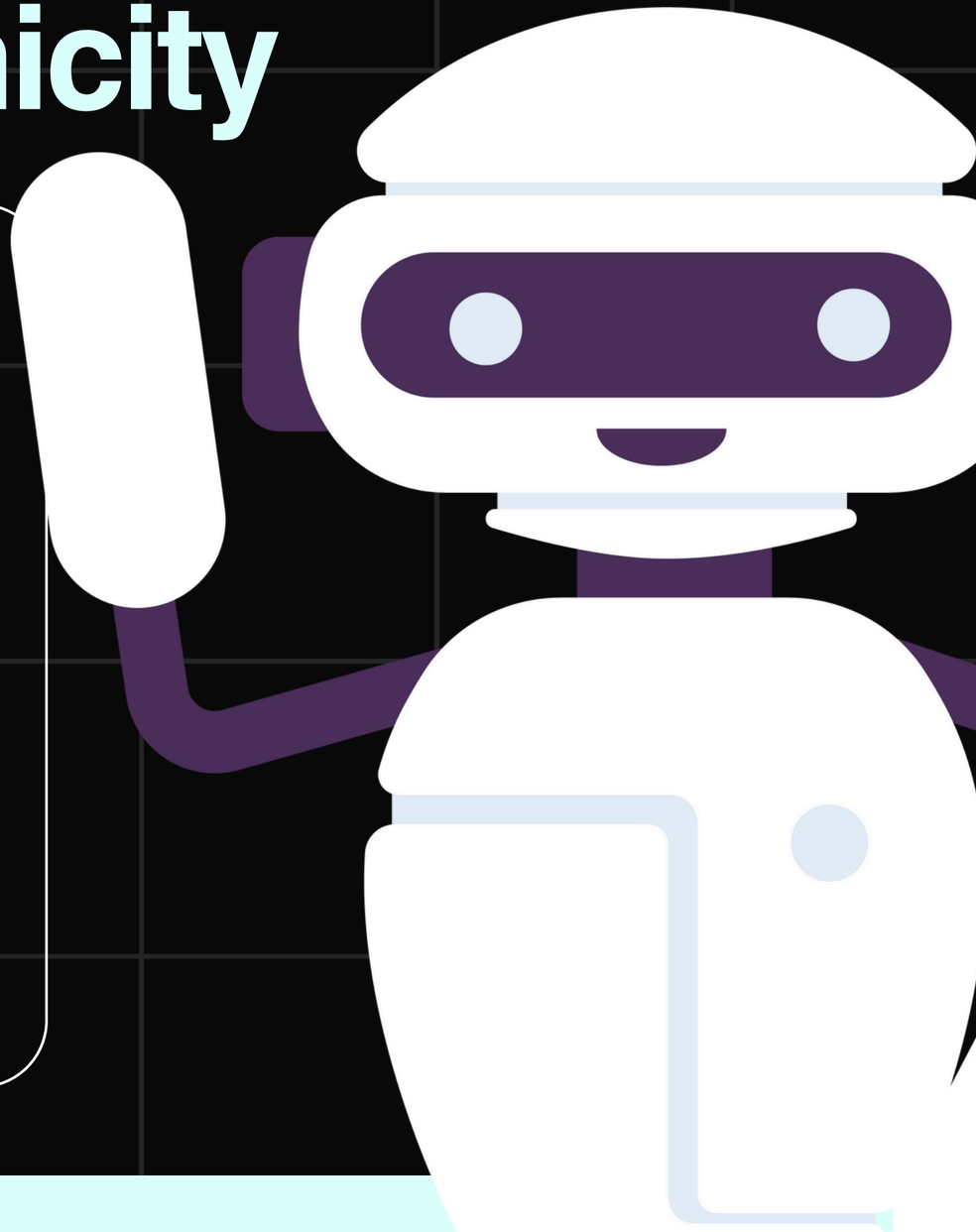
EDA. Min GPA vs SES



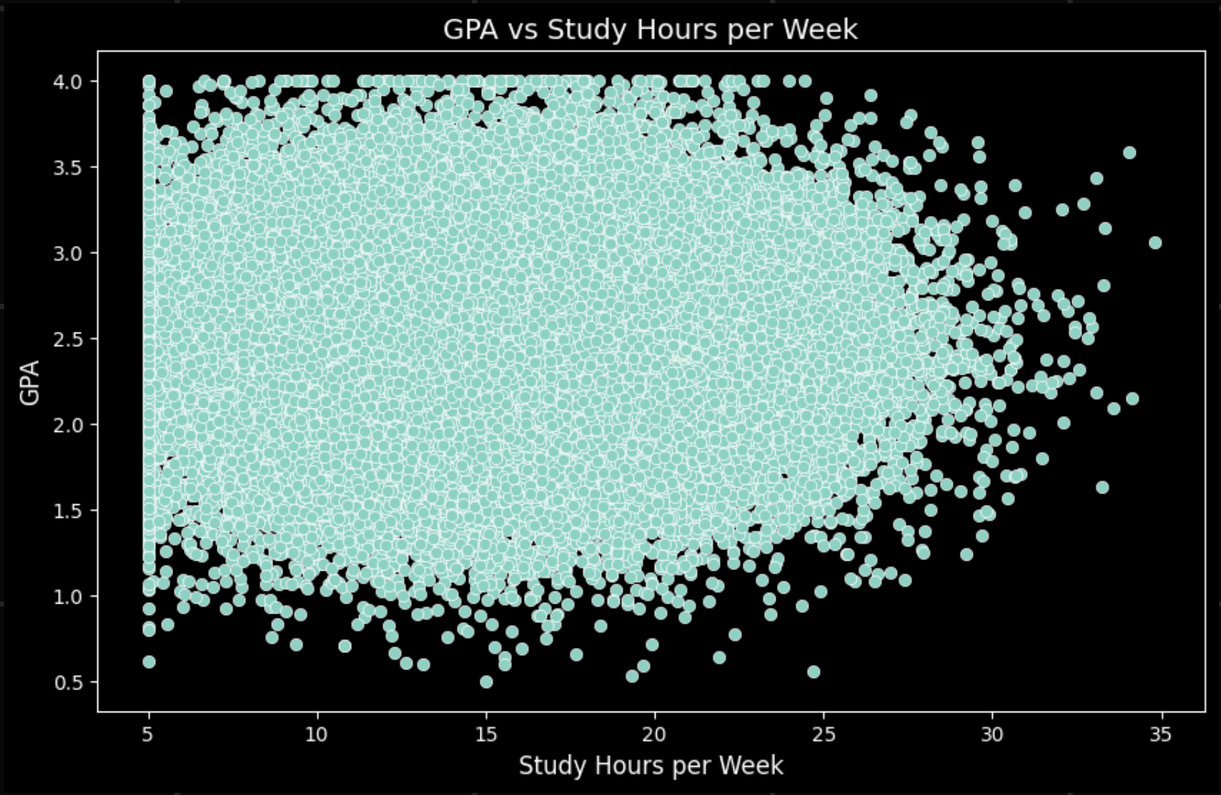


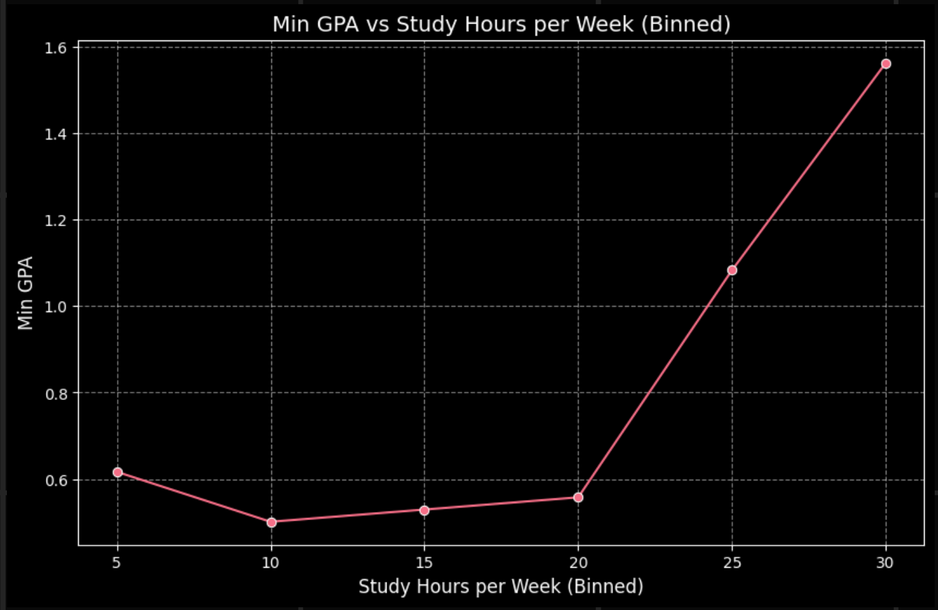
EDA. Min GPA vs Ethnicity

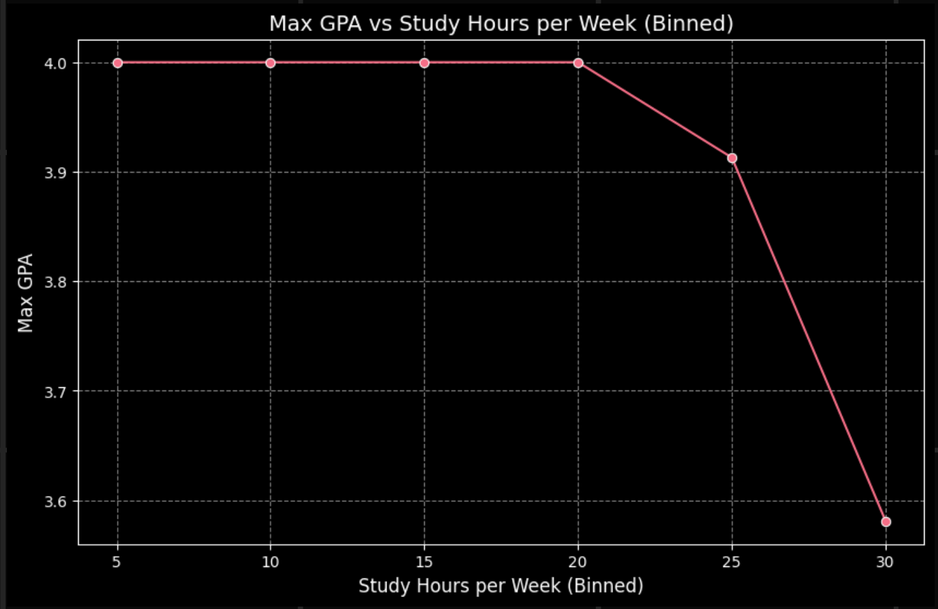




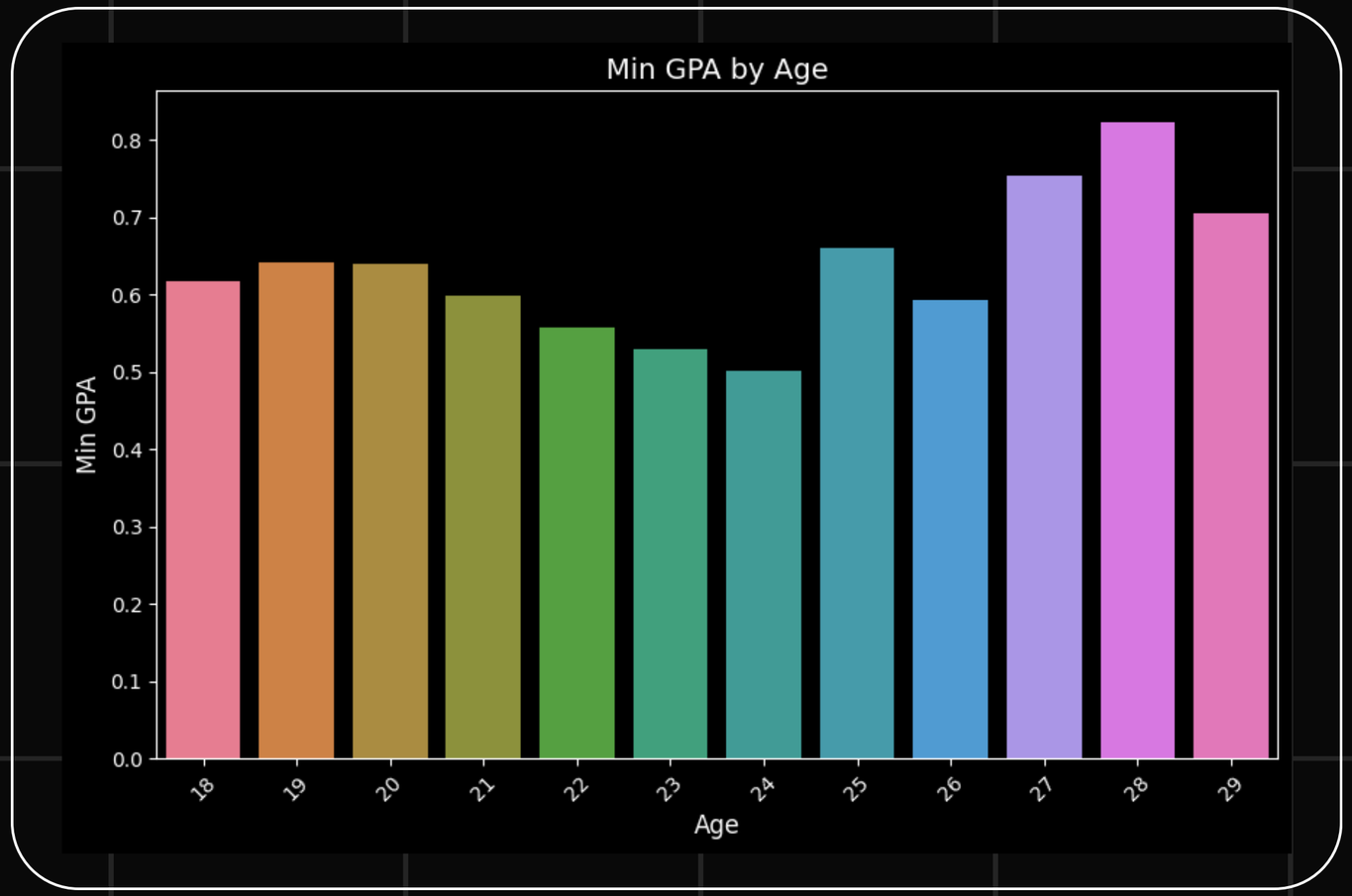
EDA. Min GPA vs Study Hours

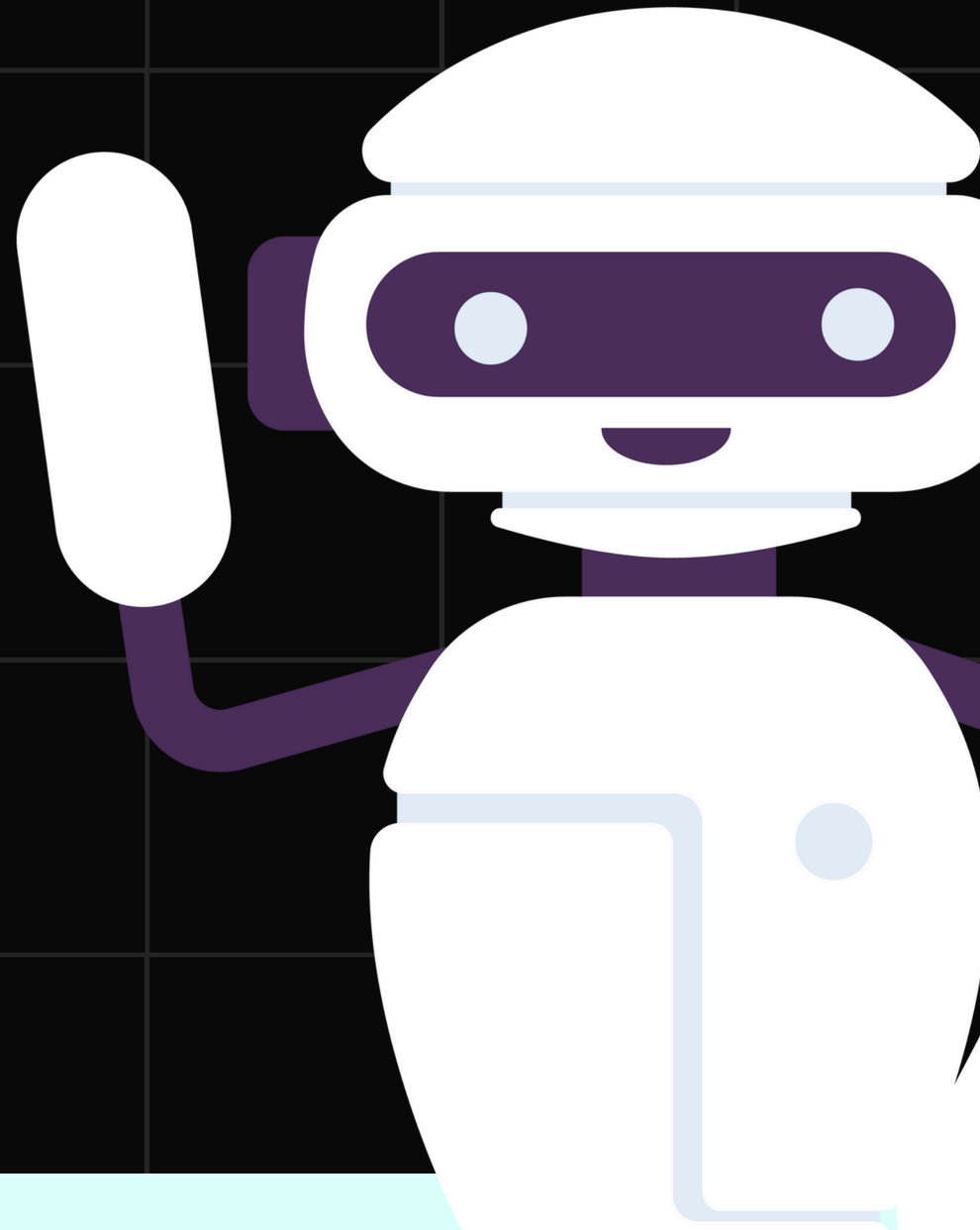






EDA. Min GPA vs Age





Transforming data

Reducing bits

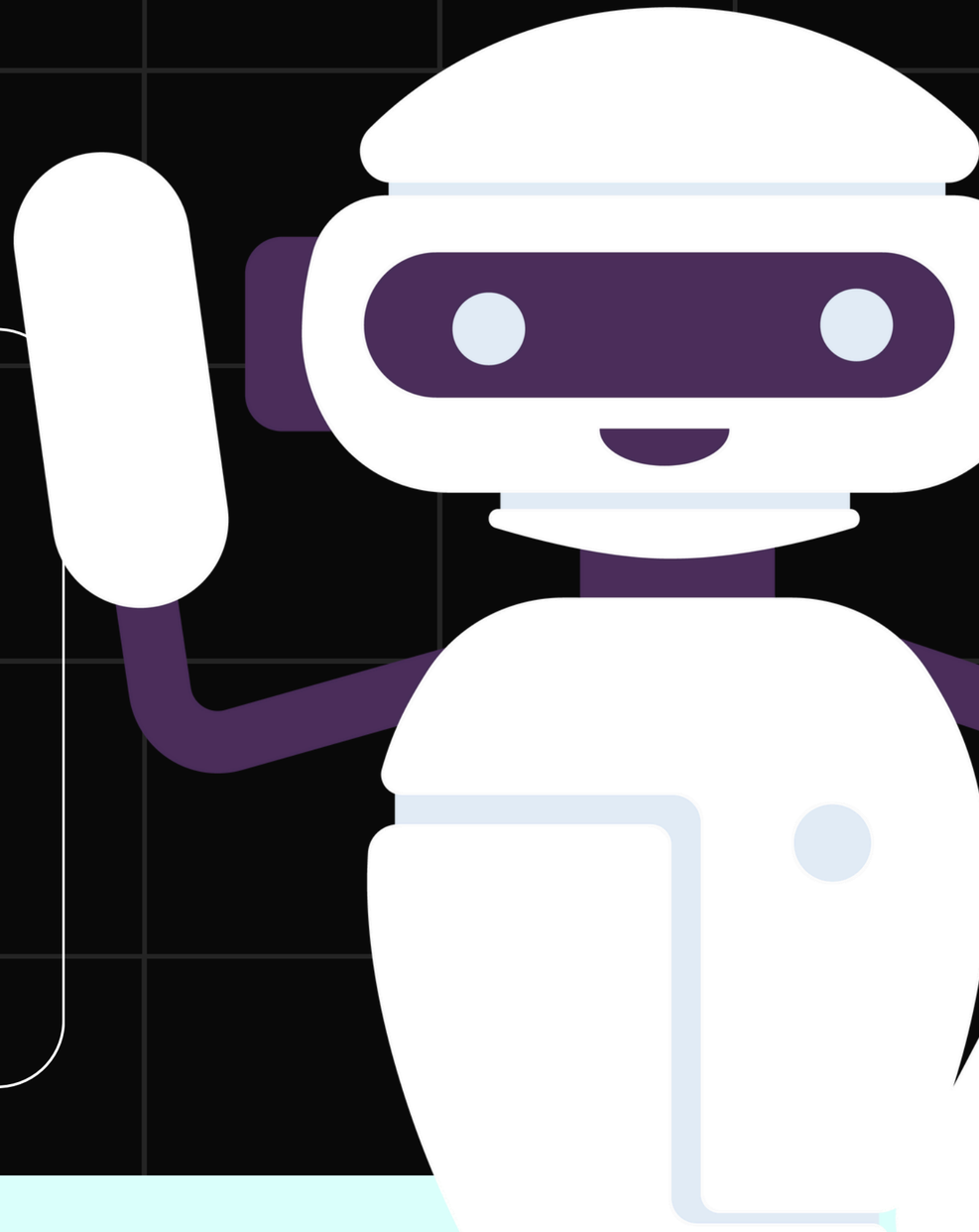
From 64 to 16 bits.

One Hot Encoding

Converting categorical variables to binary.

Balance data

More than the 80% of rows have a GPA above 2.



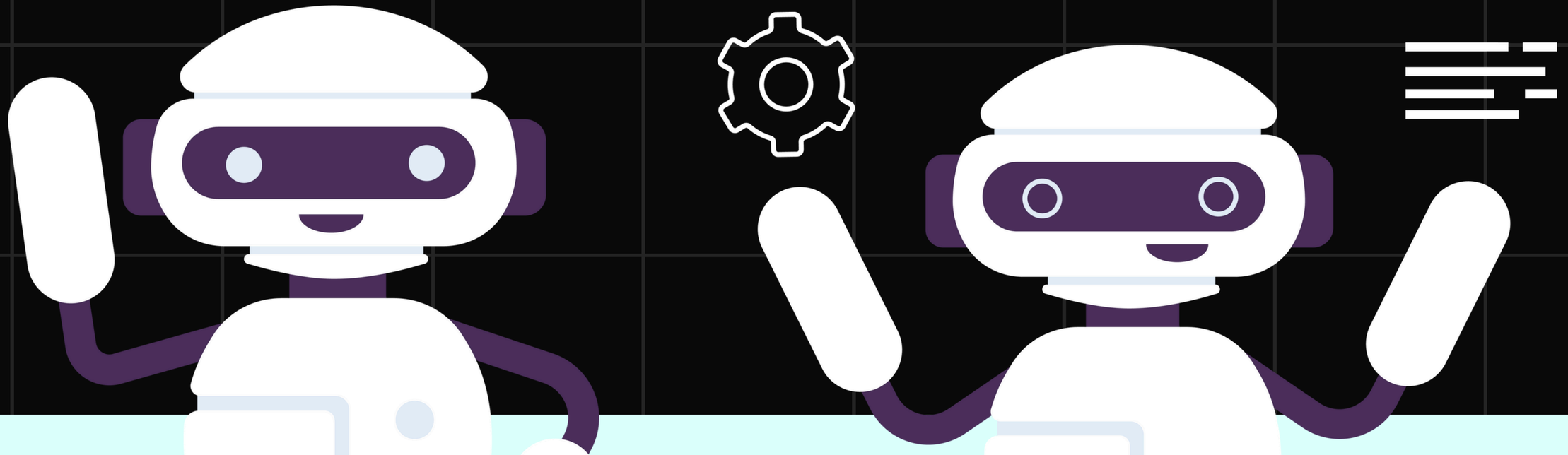
Testing the data

Tranforming and preparing data

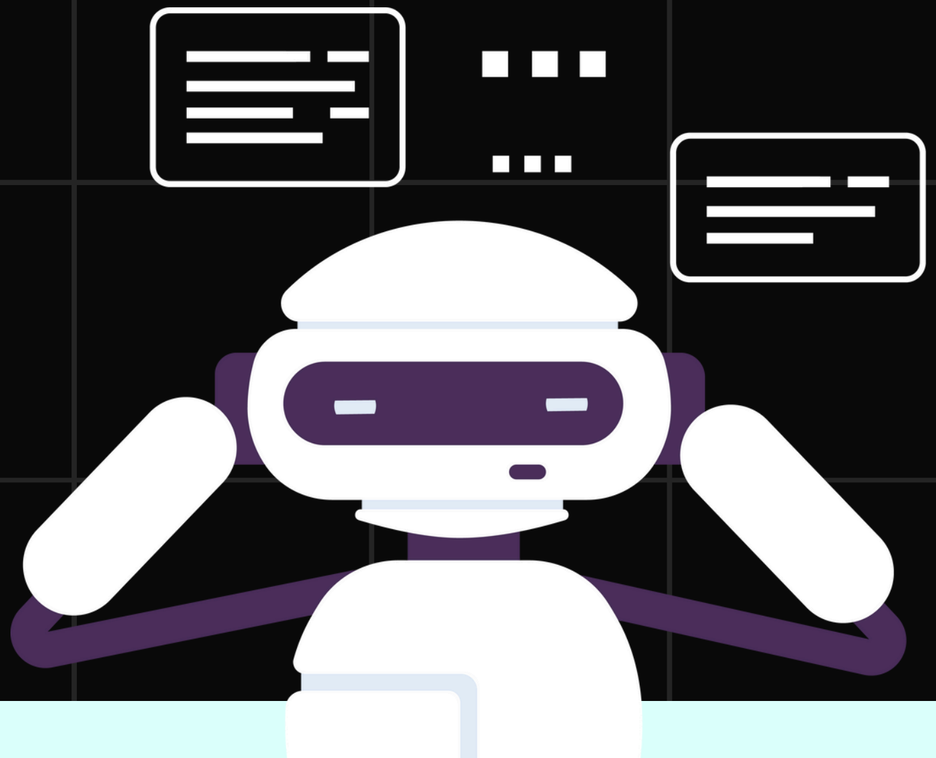
- Normalizer
- Scaler
- Encoder

Predictive Models

- Random Forest Classifier
- Gradient boosting Classifier
- Logistic Regressiom



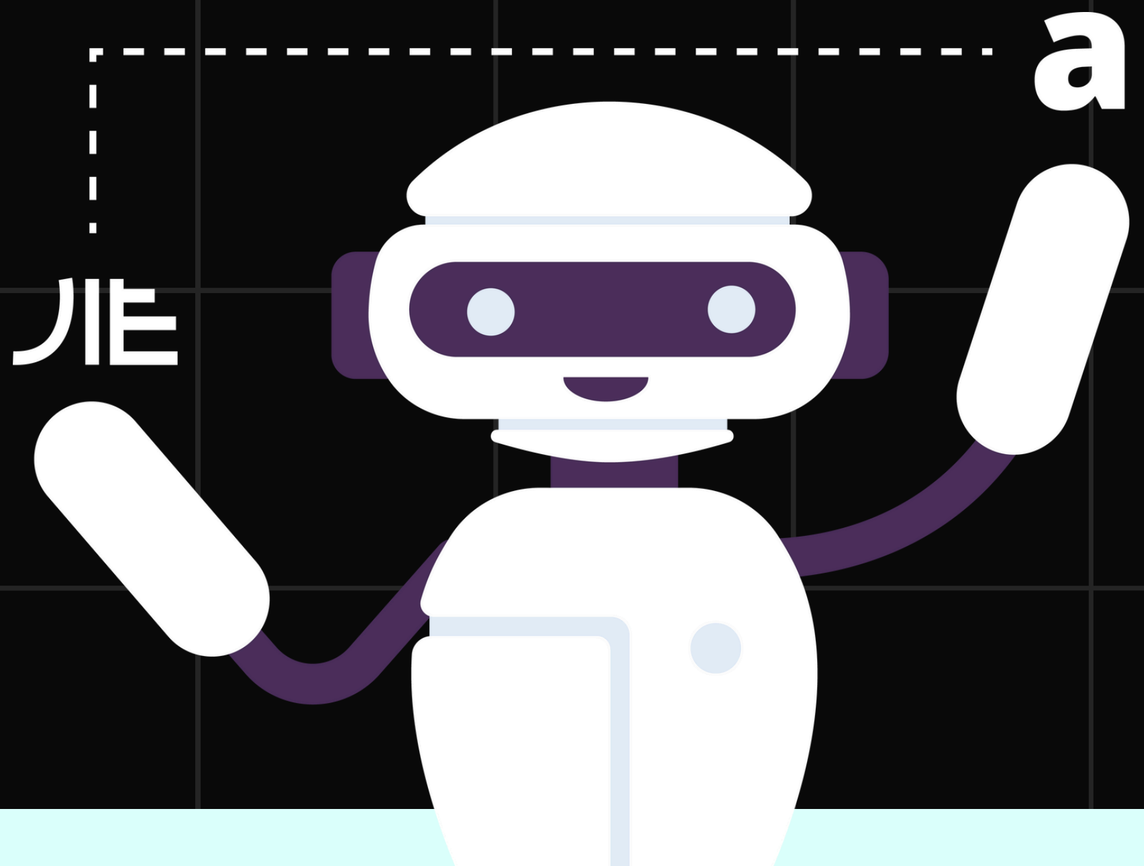
Which is the most suitable predictive model?



Predictive Model	Normalizer	Scaler	Encoder
Random Forest Classifier	0: 0.10 1: 0.84	**0: 0.99** **1: 0.98**	0: 0.10 1: 0.84
Gradient Boosting Classifier	0: 0.22 1: 0.84	0: 0.00 1: 0.84	0: 0.00 1: 0.84
Logistic Regression	0: 0.00 1: 0.84	0: 0.00 1: 0.84	0: 0.00 1: 0.84

Conclusions

- **Step 0:** Find a dataset.
- **Step 1**: Check if the data is real.
- **Step 2:** Do a heatmap.
- **Step 3:** Analyse in detail all the variables.
- **Step 4:** If dataset doesn’t meet basic requirements, go back to step 0.
- **Step 5:** Analyse different models and choose the best one.
- **Step 6:** Congrats, you have a predictive model!



Never trust Kaggle before
exploring properly the dataset.

Thankyou

@Ironhack