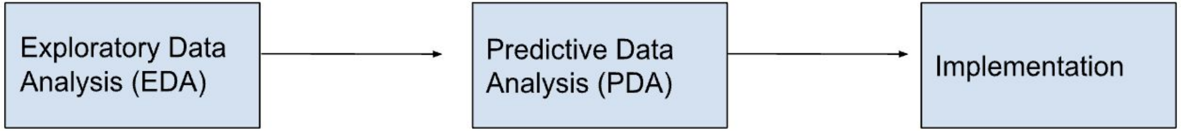
Higher Education Grade Prediction

Gaurang Bista - u3252897

Table of Content

Introduction: Problem Statement

- The aim of this project is to provide students with a prediction as to whether or not they will pass their relative course studies based on various lifestyle attributes.



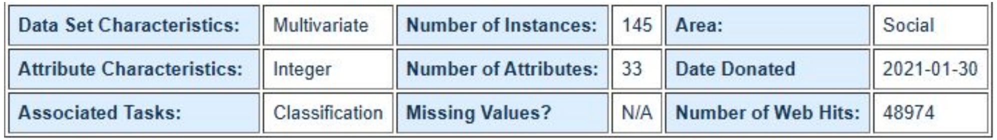
Exploratory Data Analysis (EDA)

Predictive Data Analysis (PDA)

Implementation

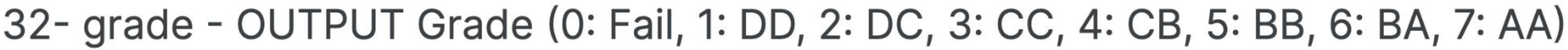
Dataset details

- Dataset name: *Higher Education Students Performance Evaluation* [3]
- Provided by **Nevriye Yilmaz** and **Boran Sekeroglu** in 2019.
- Data was collected from the Faculty of Engineering and Faculty of Educational Sciences students [4]



Data Set Characteristics:	Multivariate	**Number of Instances:**	145	**Area:**	Social
Attribute Characteristics:	Integer	**Number of Attributes:**	33	**Date Donated**	2021-01-30
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	48974

Source: UCI [4]

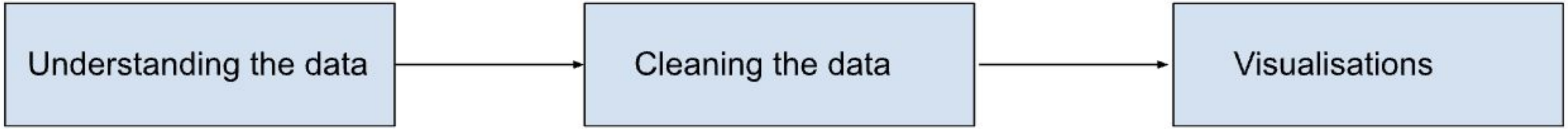


32- grade - OUTPUT Grade (0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)

Source: Kaggle [3]

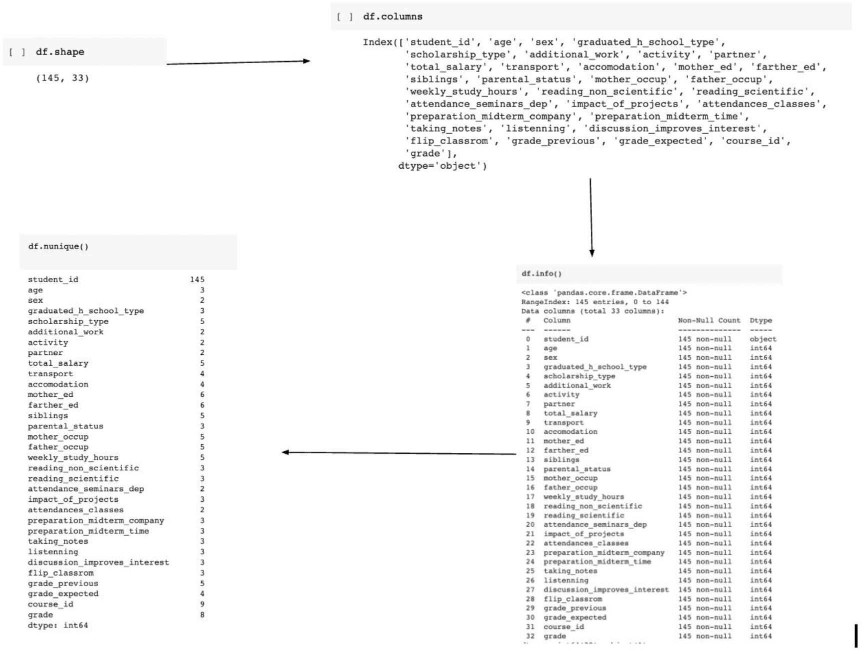
EDA Overview

- The Exploratory Data Analysis (EDA and visualisation for this project was divided into three parts
    - Understanding the data
    - Cleaning the data
    - Creating visualisations which can be used to answer 5 questions about the dataset



EDA - Outcomes Stage 1

- In this stage, the dataset was mounted to google drive and the operations below were conducted to gain a better understanding of the data.
- The dataset was checked for any null-values and the data types were also examined.



```
[ ]  df.shape

     (145, 33)
```

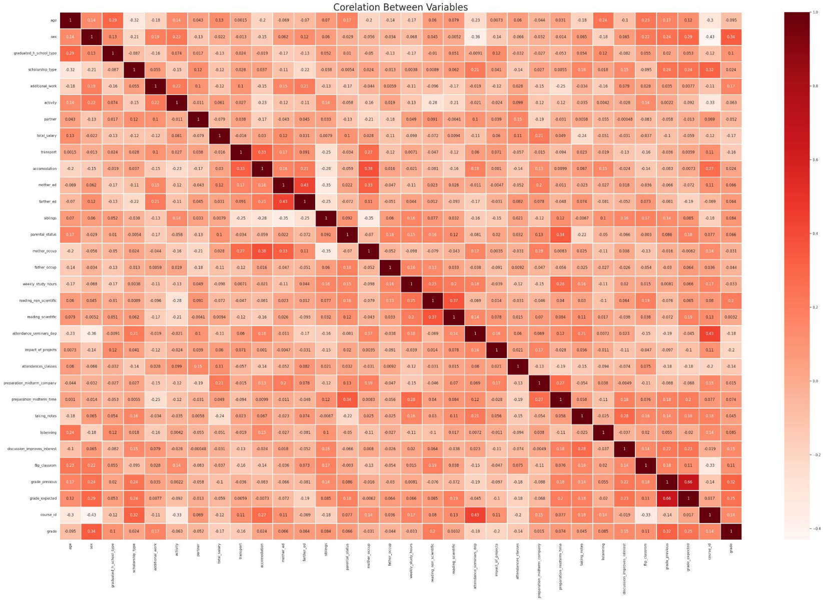
```
[ ]  df.columns

     Index(['student_id', 'age', 'sex', 'graduated_h_school_type',
            'scholarship_type', 'additional_work', 'activity', 'partner',
            'total_salary', 'transport', 'accomodation', 'mother_ed', 'farther_ed',
            'siblings', 'parental_status', 'mother_occup', 'father_occup',
            'weekly_study_hours', 'reading_non_scientific', 'reading_scientific',
            'attendance_seminars_dep', 'impact_of_projects', 'attendances_classes',
            'preparation_midterm_company', 'preparation_midterm_time',
            'taking_notes', 'listenning', 'discussion_improves_interest',
            'flip_classrom', 'grade_previous', 'grade_expected', 'course_id',
            'grade'],
           dtype='object')
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145 entries, 0 to 144
Data columns (total 33 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   student_id                    145 non-null    object
 1   age                           145 non-null    int64
 2   sex                           145 non-null    int64
 3   graduated_h_school_type       145 non-null    int64
 4   scholarship_type              145 non-null    int64
 5   additional_work               145 non-null    int64
 6   activity                      145 non-null    int64
 7   partner                       145 non-null    int64
 8   total_salary                  145 non-null    int64
 9   transport                     145 non-null    int64
 10  accomodation                  145 non-null    int64
 11  mother_ed                     145 non-null    int64
 12  farther_ed                    145 non-null    int64
 13  siblings                      145 non-null    int64
 14  parental_status               145 non-null    int64
 15  mother_occup                  145 non-null    int64
 16  father_occup                  145 non-null    int64
 17  weekly_study_hours            145 non-null    int64
 18  reading_non_scientific        145 non-null    int64
 19  reading_scientific            145 non-null    int64
 20  attendance_seminars_dep       145 non-null    int64
 21  impact_of_projects            145 non-null    int64
 22  attendances_classes           145 non-null    int64
 23  preparation_midterm_company   145 non-null    int64
 24  preparation_midterm_time      145 non-null    int64
 25  taking_notes                  145 non-null    int64
 26  listenning                    145 non-null    int64
 27  discussion_improves_interest  145 non-null    int64
 28  flip_classrom                 145 non-null    int64
 29  grade_previous                145 non-null    int64
 30  grade_expected                145 non-null    int64
 31  course_id                     145 non-null    int64
 32  grade                         145 non-null    int64
```

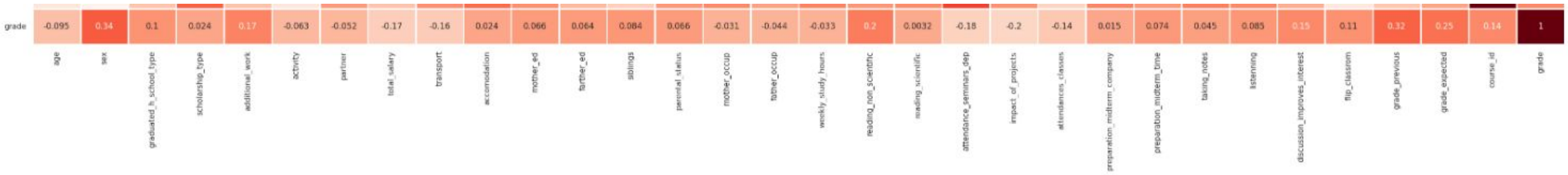
```
df.nunique()

student_id                      145
age                               3
sex                               2
graduated_h_school_type           3
scholarship_type                  5
additional_work                   2
activity                          2
partner                           2
total_salary                      5
transport                         4
accomodation                      4
mother_ed                         6
farther_ed                        6
siblings                          5
parental_status                   3
mother_occup                      5
father_occup                      5
weekly_study_hours                5
reading_non_scientific            3
reading_scientific                3
attendance_seminars_dep           2
impact_of_projects                3
attendances_classes               2
preparation_midterm_company       3
preparation_midterm_time          3
taking_notes                      3
listenning                        3
discussion_improves_interest      3
flip_classrom                     3
grade_previous                    5
grade_expected                    4
course_id                         9
grade                             8
dtype: int64
```

EDA - Outcomes Stage 2

1. Correlation chart

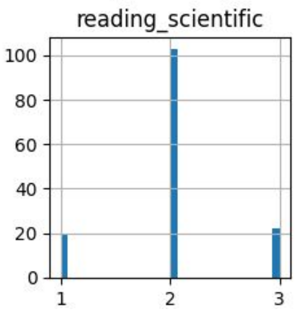


- In this stage, a sns correlation chart was created to observe the correlation rates between generic attributes and the predictive attribute.
- Most attributes that showed a weaker correlation with the student grade were dropped.

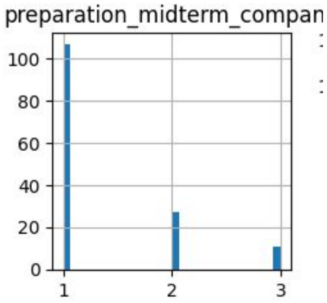


EDA - Outcome Stage 3

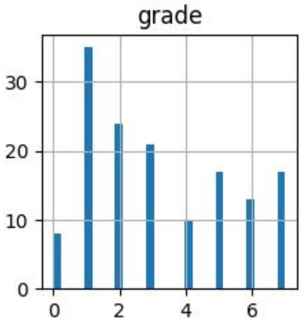
- Q1. Which grade is most commonly achieved by students in the dataset?
- Q2. What are the most common weekly study hours?
- Q3. What is the most common reading frequency for scientific text?
- Q4. How do most students prepare for midterm? Alone? With friends?
- Q5. What is the correlation coefficient between previous grade and current grade?



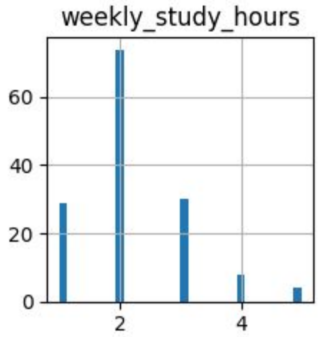
1: None, **2**: Sometimes, **3**: Often



1: alone, **2**: with friends, **3**: n/a

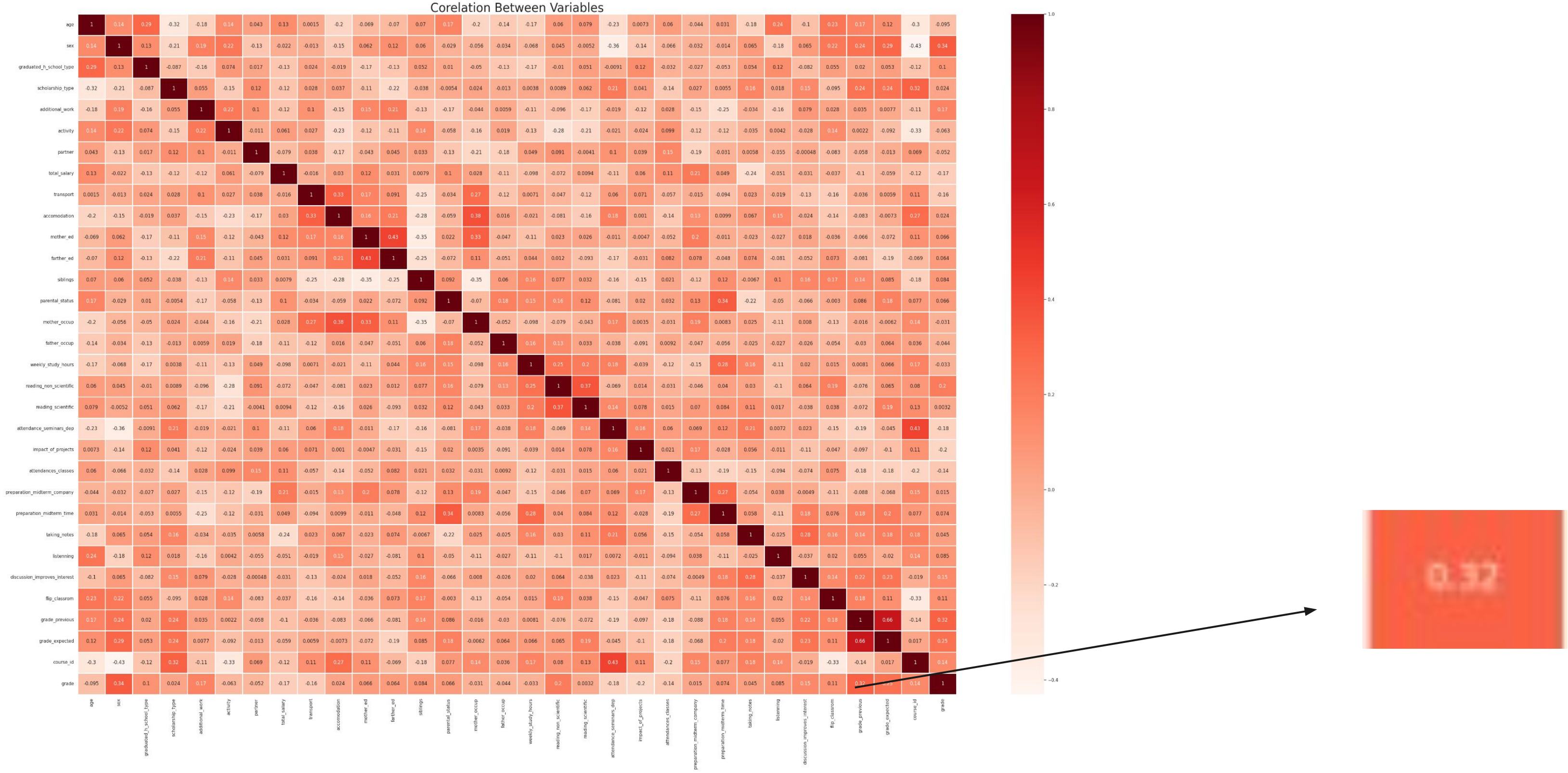


0: Fail, **1**: DD, **2**: DC, **3**: CC, **4**: CB, **5**: BB, **6**: BA, **7**: AA



1: None, **2**: <5 hours, **3**: 6-10 hours, **4**: 11-20 hours, **5**: more than 20 hours

EDA - Q5

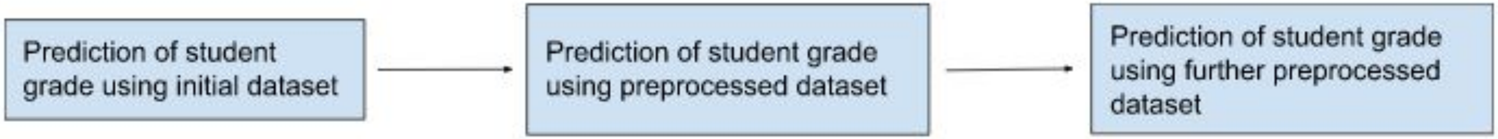


Corelation Between Variables

	age	sex	graduated_h_school_type	scholarship_type	additional_work	activity	partner	total_salary	transport	accomodation	mother_ed	farther_ed	siblings	parental_status	mother_occup	father_occup	weekly_study_hours	reading_non_scientific	reading_scientific	attendance_seminars_dep	impact_of_projects	attendances_classes	preparation_midterm_company	preparation_midterm_time	taking_notes	listenning	discussion_improves_interest	flip_classrom	grade_previous	grade_expected	course_id	grade
age	1	0.14	0.29	-0.32	-0.18	0.14	0.043	0.13	0.0015	-0.2	-0.069	-0.07	0.07	0.17	-0.2	-0.14	-0.17	0.06	0.079	-0.23	0.0073	0.06	-0.044	0.031	-0.18	0.24	-0.1	0.23	0.17	0.12	-0.3	-0.095
sex	0.14	1	0.13	-0.21	0.19	0.22	-0.13	-0.022	-0.013	-0.15	0.062	0.12	0.06	-0.029	-0.056	-0.034	-0.068	0.045	-0.0052	-0.36	-0.14	-0.066	-0.032	-0.014	0.065	-0.18	0.065	0.22	0.24	0.29	-0.43	0.34
graduated_h_school_type	0.29	0.13	1	-0.087	-0.16	0.074	0.017	-0.13	0.024	-0.019	-0.17	-0.13	0.052	0.01	-0.05	-0.13	-0.17	-0.01	0.051	-0.0091	0.12	-0.032	-0.027	-0.053	0.054	0.12	-0.082	0.055	0.02	0.053	-0.12	0.1
scholarship_type	-0.32	-0.21	-0.087	1	0.055	-0.15	0.12	-0.12	0.028	0.037	-0.11	-0.22	-0.038	-0.0054	0.024	-0.013	0.0038	0.0089	0.062	0.21	0.041	-0.14	0.027	0.0055	0.16	0.018	0.15	-0.095	0.24	0.24	0.32	0.024
additional_work	-0.18	0.19	-0.16	0.055	1	0.22	0.1	-0.12	0.1	-0.15	0.15	0.21	-0.13	-0.17	-0.044	0.0059	-0.11	-0.096	-0.17	-0.019	-0.12	0.028	-0.15	-0.25	-0.034	-0.16	0.079	0.028	0.035	0.0077	-0.11	0.17
activity	0.14	0.22	0.074	-0.15	0.22	1	-0.011	0.061	0.027	-0.23	-0.12	-0.11	0.14	-0.058	-0.16	0.019	-0.13	-0.28	-0.21	-0.021	-0.024	0.099	-0.12	-0.12	-0.035	0.0042	-0.028	0.14	0.0022	-0.092	-0.33	-0.063
partner	0.043	-0.13	0.017	0.12	0.1	-0.011	1	-0.079	0.038	-0.17	-0.043	0.045	0.033	-0.13	-0.21	-0.18	0.049	0.091	-0.0041	0.1	0.039	0.15	-0.19	-0.031	0.0058	-0.055	-0.00048	-0.083	-0.058	-0.013	0.069	-0.052
total_salary	0.13	-0.022	-0.13	-0.12	-0.12	0.061	-0.079	1	-0.016	0.03	0.12	0.031	0.0079	0.1	0.028	-0.11	-0.098	-0.072	0.0094	-0.11	0.06	0.11	0.21	0.049	-0.24	-0.051	-0.031	-0.037	-0.1	-0.059	-0.12	-0.17
transport	0.0015	-0.013	0.024	0.028	0.1	0.027	0.038	-0.016	1	0.33	0.17	0.091	-0.25	-0.034	0.27	-0.12	0.0071	-0.047	-0.12	0.06	0.071	-0.057	-0.015	-0.094	0.023	-0.019	-0.13	-0.16	-0.036	0.0059	0.11	-0.16
accomodation	-0.2	-0.15	-0.019	0.037	-0.15	-0.23	-0.17	0.03	0.33	1	0.16	0.21	-0.28	-0.059	0.38	0.016	-0.021	-0.081	-0.16	0.18	0.001	-0.14	0.13	0.0099	0.067	0.15	-0.024	-0.14	-0.083	-0.0073	0.27	0.024
mother_ed	-0.069	0.062	-0.17	-0.11	0.15	-0.12	-0.043	0.12	0.17	0.16	1	0.43	-0.35	0.022	0.33	-0.047	-0.11	0.023	0.026	-0.011	-0.0047	-0.052	0.2	-0.011	-0.023	-0.027	0.018	-0.036	-0.066	-0.072	0.11	0.066
farther_ed	-0.07	0.12	-0.13	-0.22	0.21	-0.11	0.045	0.031	0.091	0.21	0.43	1	-0.25	-0.072	0.11	-0.051	0.044	0.012	-0.093	-0.17	-0.031	0.082	0.078	-0.048	0.074	-0.081	-0.052	0.073	-0.081	-0.19	-0.069	0.064
siblings	0.07	0.06	0.052	-0.038	-0.13	0.14	0.033	0.0079	-0.25	-0.28	-0.35	-0.25	1	0.092	-0.35	0.06	0.16	0.077	0.032	-0.16	-0.15	0.021	-0.12	0.12	-0.0067	0.1	0.16	0.17	0.14	0.085	-0.18	0.084
parental_status	0.17	-0.029	0.01	-0.0054	-0.17	-0.058	-0.13	0.1	-0.034	-0.059	0.022	-0.072	0.092	1	-0.07	0.18	0.15	0.16	0.12	-0.081	0.02	0.032	0.13	0.34	-0.22	-0.05	-0.066	-0.003	0.086	0.18	0.077	0.066
mother_occup	-0.2	-0.056	-0.05	0.024	-0.044	-0.16	-0.21	0.028	0.27	0.38	0.33	0.11	-0.35	-0.07	1	-0.052	-0.098	-0.079	-0.043	0.17	0.0035	-0.031	0.19	0.0083	0.025	-0.11	0.008	-0.13	-0.016	-0.0062	0.14	-0.031
father_occup	-0.14	-0.034	-0.13	-0.013	0.0059	0.019	-0.18	-0.11	-0.12	0.016	-0.047	-0.051	0.06	0.18	-0.052	1	0.16	0.13	0.033	-0.038	-0.091	0.0092	-0.047	-0.056	-0.025	-0.027	-0.026	-0.054	-0.03	0.064	0.036	-0.044
weekly_study_hours	-0.17	-0.068	-0.17	0.0038	-0.11	-0.13	0.049	-0.098	0.0071	-0.021	-0.11	0.044	0.16	0.15	-0.098	0.16	1	0.25	0.2	0.18	-0.039	-0.12	-0.15	0.28	0.16	-0.11	0.02	0.015	0.0081	0.066	0.17	-0.033
reading_non_scientific	0.06	0.045	-0.01	0.0089	-0.096	-0.28	0.091	-0.072	-0.047	-0.081	0.023	0.012	0.077	0.16	-0.079	0.13	0.25	1	0.37	-0.069	0.014	-0.031	-0.046	0.04	0.03	-0.1	0.064	0.19	-0.076	0.065	0.08	0.2
reading_scientific	0.079	-0.0052	0.051	0.062	-0.17	-0.21	-0.0041	0.0094	-0.12	-0.16	0.026	-0.093	0.032	0.12	-0.043	0.033	0.2	0.37	1	0.14	0.078	0.015	0.07	0.084	0.11	0.017	-0.038	0.038	-0.072	0.19	0.13	0.0032
attendance_seminars_dep	-0.23	-0.36	-0.0091	0.21	-0.019	-0.021	0.1	-0.11	0.06	0.18	-0.011	-0.17	-0.16	-0.081	0.17	-0.038	0.18	-0.069	0.14	1	0.16	0.06	0.069	0.12	0.21	0.0072	0.023	-0.15	-0.19	-0.045	0.43	-0.18
impact_of_projects	0.0073	-0.14	0.12	0.041	-0.12	-0.024	0.039	0.06	0.071	0.001	-0.0047	-0.031	-0.15	0.02	0.0035	-0.091	-0.039	0.014	0.078	0.16	1	0.021	0.17	-0.028	0.056	-0.011	-0.11	-0.047	-0.097	-0.1	0.11	-0.2
attendances_classes	0.06	-0.066	-0.032	-0.14	0.028	0.099	0.15	0.11	-0.057	-0.14	-0.052	0.082	0.021	0.032	-0.031	0.0092	-0.12	-0.031	0.015	0.06	0.021	1	-0.13	-0.19	-0.15	-0.094	-0.074	0.075	-0.18	-0.18	-0.2	-0.14
preparation_midterm_company	-0.044	-0.032	-0.027	0.027	-0.15	-0.12	-0.19	0.21	-0.015	0.13	0.2	0.078	-0.12	0.13	0.19	-0.047	-0.15	-0.046	0.07	0.069	0.17	-0.13	1	0.27	-0.054	0.038	-0.0049	-0.11	-0.088	-0.068	0.15	0.015
preparation_midterm_time	0.031	-0.014	-0.053	0.0055	-0.25	-0.12	-0.031	0.049	-0.094	0.0099	-0.011	-0.048	0.12	0.34	0.0083	-0.056	0.28	0.04	0.084	0.12	-0.028	-0.19	0.27	1	0.058	-0.11	0.18	0.076	0.18	0.2	0.077	0.074
taking_notes	-0.18	0.065	0.054	0.16	-0.034	-0.035	0.0058	-0.24	0.023	0.067	-0.023	0.074	-0.0067	-0.22	0.025	-0.025	0.16	0.03	0.11	0.21	0.056	-0.15	-0.054	0.058	1	-0.025	0.28	0.16	0.14	0.18	0.18	0.045
listenning	0.24	-0.18	0.12	0.018	-0.16	0.0042	-0.055	-0.051	-0.019	0.15	-0.027	-0.081	0.1	-0.05	-0.11	-0.027	-0.11	-0.1	0.017	0.0072	-0.011	-0.094	0.038	-0.11	-0.025	1	-0.037	0.02	0.055	-0.02	0.14	0.085
discussion_improves_interest	-0.1	0.065	-0.082	0.15	0.079	-0.028	-0.00048	-0.031	-0.13	-0.024	0.018	-0.052	0.16	-0.066	0.008	-0.026	0.02	0.064	-0.038	0.023	-0.11	-0.074	-0.0049	0.18	0.28	-0.037	1	0.14	0.22	0.23	-0.019	0.15
flip_classrom	0.23	0.22	0.055	-0.095	0.028	0.14	-0.083	-0.037	-0.16	-0.14	-0.036	0.073	0.17	-0.003	-0.13	-0.054	0.015	0.19	0.038	-0.15	-0.047	0.075	-0.11	0.076	0.16	0.02	0.14	1	0.18	0.11	-0.33	0.11
grade_previous	0.17	0.24	0.02	0.24	0.035	0.0022	-0.058	-0.1	-0.036	-0.083	-0.066	-0.081	0.14	0.086	-0.016	-0.03	0.0081	-0.076	-0.072	-0.19	-0.097	-0.18	-0.088	0.18	0.14	0.055	0.22	0.18	1	0.66	-0.14	0.32
grade_expected	0.12	0.29	0.053	0.24	0.0077	-0.092	-0.013	-0.059	0.0059	-0.0073	-0.072	-0.19	0.085	0.18	-0.0062	0.064	0.066	0.065	0.19	-0.045	-0.1	-0.18	-0.068	0.2	0.18	-0.02	0.23	0.11	0.66	1	0.017	0.25
course_id	-0.3	-0.43	-0.12	0.32	-0.11	-0.33	0.069	-0.12	0.11	0.27	0.11	-0.069	-0.18	0.077	0.14	0.036	0.17	0.08	0.13	0.43	0.11	-0.2	0.15	0.077	0.18	0.14	-0.019	-0.33	-0.14	0.017	1	0.14
grade	-0.095	0.34	0.1	0.024	0.17	-0.063	-0.052	-0.17	-0.16	0.024	0.066	0.064	0.084	0.066	-0.031	-0.044	-0.033	0.2	0.0032	-0.18	-0.2	-0.14	0.015	0.074	0.045	0.085	0.15	0.11	0.32	0.25	0.14	1

PDA Outcomes

- The PDA was divided into three stages:
  - Initial dataset prediction
  - Pre -processed dataset prediction
  - Further preprocessed dataset prediction



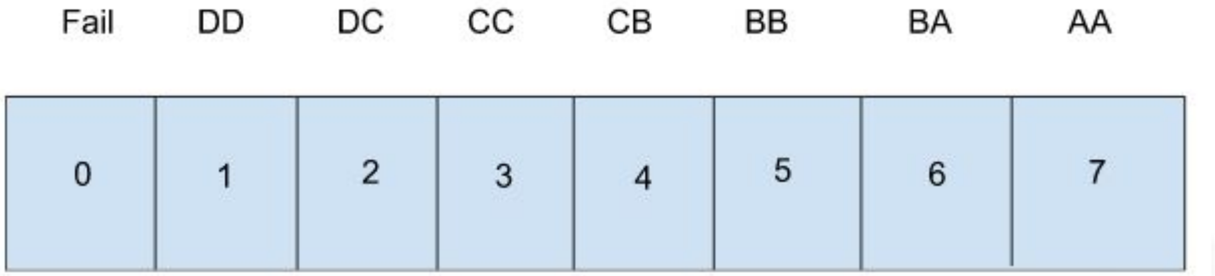
Prediction of student grade using initial dataset

Prediction of student grade using preprocessed dataset

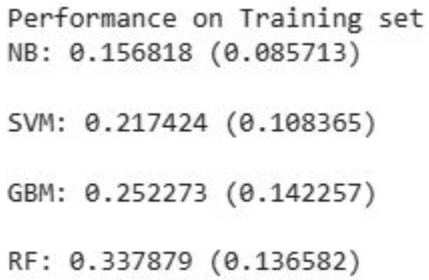
Prediction of student grade using further preprocessed dataset

PDA Outcomes - Stage 1

- This stage involved testing four predictive models Naive Bayes, Support Vector Machines, Gradient Boosting, and Random Forest on the initial dataset without preprocessing.
- Due to the dataset being small and multiple grade types, the accuracy rates of each model were low.



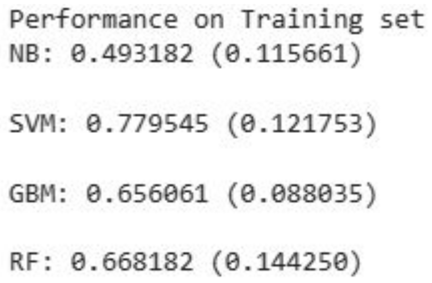
Fail	DD	DC	CC	CB	BB	BA	AA
0	1	2	3	4	5	6	7



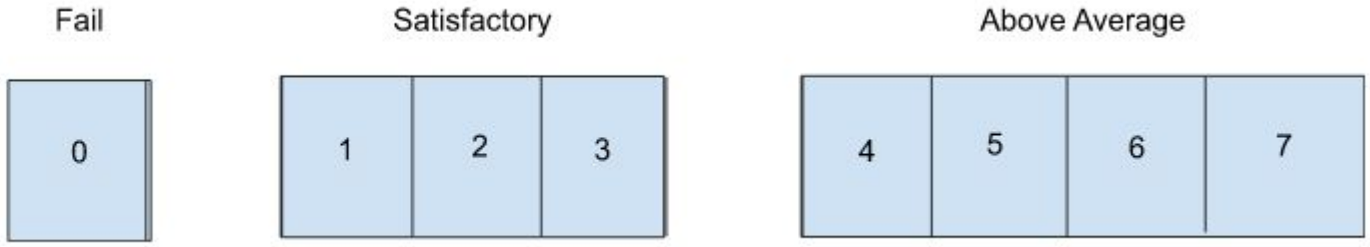
```
Performance on Training set
NB: 0.156818 (0.085713)

SVM: 0.217424 (0.108365)

GBM: 0.252273 (0.142257)

RF: 0.337879 (0.136582)
```

PDA Outcomes - Stage 2

- To improve the model prediction accuracy, the following categorisation process was conducted.



```
Performance on Training set
NB: 0.493182 (0.115661)

SVM: 0.779545 (0.121753)

GBM: 0.656061 (0.088035)

RF: 0.668182 (0.144250)
```



Fail

0

Satisfactory

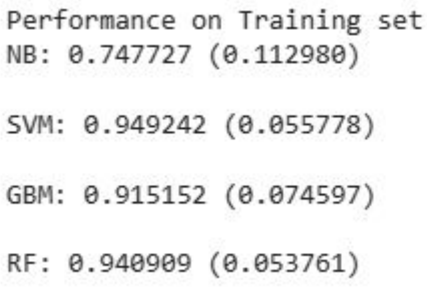
1 2 3

Above Average

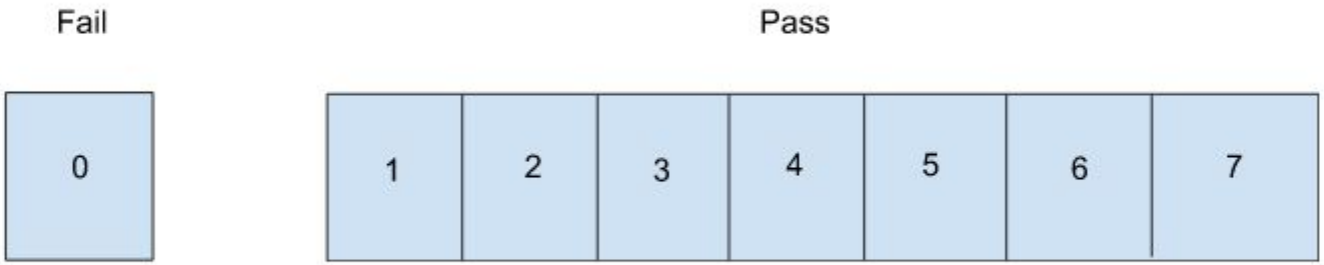
4 5 6 7

PDA Outcomes - Stage 3

- To further improve the model prediction accuracy, the following categorisation process was conducted.
- The predictive outcome grade was made to ba binary



```
Performance on Training set
NB: 0.747727 (0.112980)

SVM: 0.949242 (0.055778)

GBM: 0.915152 (0.074597)

RF: 0.940909 (0.053761)
```



Fail

Pass

0

1 2 3 4 5 6 7

Implementation/Deployment Plan and Status update

- I plan to use Tkinter for the implementation/deployment phase of the ST1 Capstone project
- The window will ask the user for their inputs for each attribute and provide a prediction as to whether or not they will pass.

Bibliography

- **[1]** J. Brownlee, “How to Choose a Feature Selection Method For Machine Learning,” Machine Learning Mastery, Nov. 26, 2019. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
- **[2]** T. Bush, “Predictive Analysis: Definition, Tools, and Examples,” pestleanalysis.com, Jun. 01, 2020. https://pestleanalysis.com/predictive-analysis/
- **[3]** “Higher Education Students Performance Evaluation,” www.kaggle.com. https://www.kaggle.com/datasets/mariazhokhova/higher-education-students-performance-evaluation (accessed May 02, 2023).
- **[4]** “UCI Machine Learning Repository: Higher Education Students Performance Evaluation Dataset Data Set,” archive.ics.uci.edu. https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset
- **[5]** C. Cote, “What Is Predictive Analytics? 5 Examples | HBS Online,” Business Insights - Blog, Oct. 26, 2021. https://online.hbs.edu/blog/post/predictive-analytics