Data Science Project

Student Academic Performance

BY: Sameer Anees Jaliawala

INTRODUCTION:

This project analysis the Student Academic Performance dataset and tries to find a way to predict the End-of-Semester grades of the students by looking at many different variables that might be able to explain how the students will perform in the future. We will also try to make interesting visualizations from the data and find conclusions to support the initial questions that we want to include in our research.

WHO:

The dataset was collected by a group of researchers from three colleges of Assam, India. These researchers are actually professors from different universities in Yemen, Morocco and India. They also have reported their findings in a research paper that they submitted to the

Indonesian Journal of Electrical Engineering and Computer Science. [2]

1. **Sadiq Hussain**
    - Examination Branch, Dibrugarh University, India
2. **Neama Abdul, Fadl Mutaher**
    - Department of Computer Science, Sana'a University, Sana'a, Yemen
3. **Najoua Ribata**
    - Lirosa Laboratory, Abdelmalek Essaâdi University, Tetuan, Morocco

NEED:

According to their research paper, titled "***Educational Data Mining and Analysis of Students' Academic Performance Using WEKA***", they have collected this data in order to analyze data mining tools and techniques for academic improvement of student performance and to prevent drop out. [2]

The three questions that could be answered by studying the data are:

1. What would be a good machine learning classification model for classifying student's end-of-semester grade, using small dataset size?
2. Can student's grades be impacted by their socio-economic backgrounds and their previous grades?
3. What are the main key indicators that could help in creating the classification model for predicting students' end-of-semester grades and how do each of these indicators correlate with their grades?

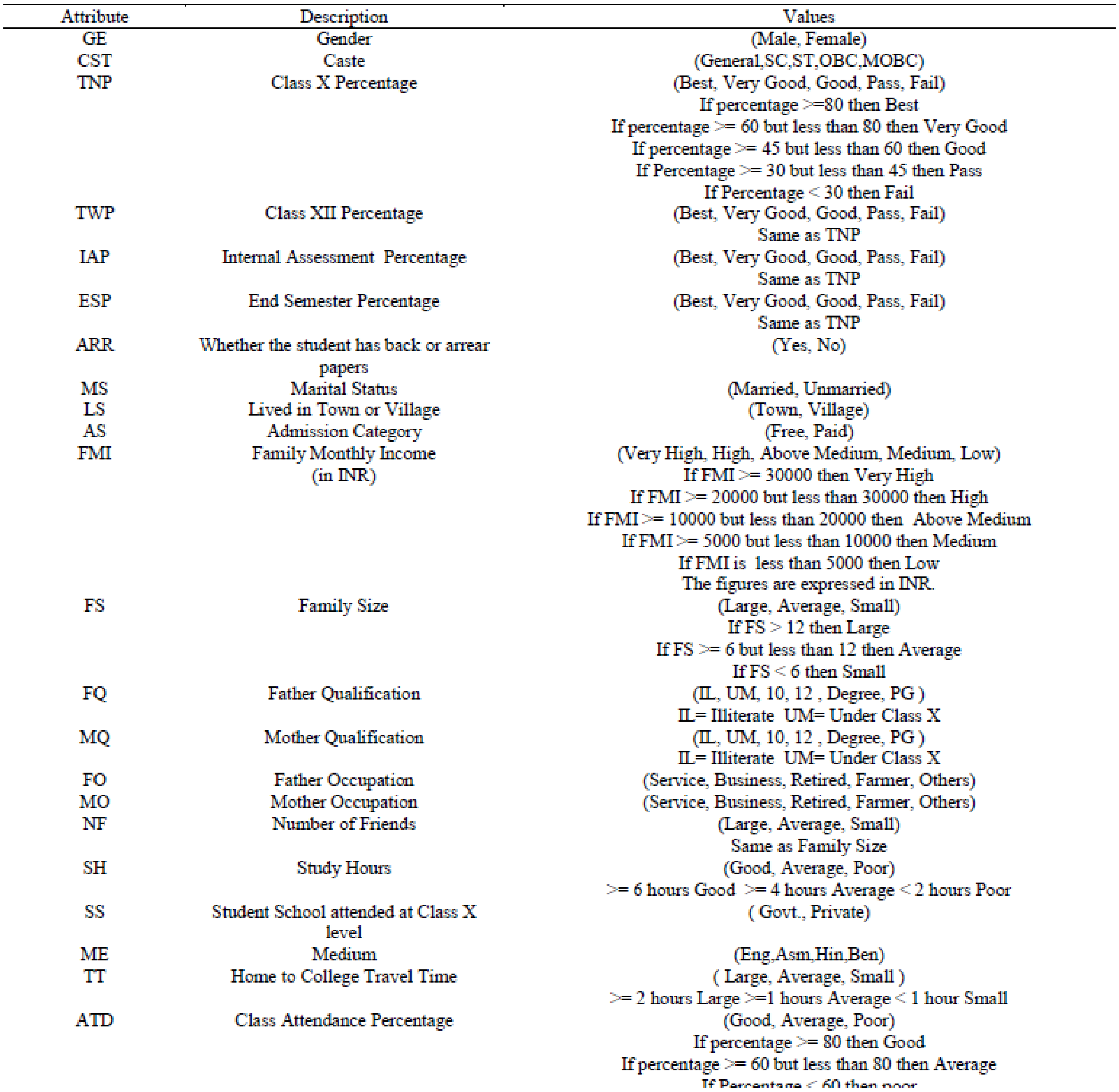
There are no privacy issues or quality issues with the data. The data is well defined and formatted; but it is not easily understandable because the column names are hard to recognize, as they used abbreviations and didn't include any legend, except for the description they mentioned in their research paper. There is also one main issue with the type of data they have provided. They have used the data mining tool known as **WEKA**, to study the data; however, the tool has changed the essence of data, as we usually deal with .CSV files.

REQUIREMENTS AND RESOURCES NEEDED:

We have used Python programming language and Jupyter notebook to analyze and study our data. We didn't use any hardware resources, as the dataset size is not that large.

DATASET DESCRIPTION:

The data consists of socio-economic, demographic as well as academic information of one hundred and thirty-one students with twenty-two attributes.



Attribute	Description	Values
GE	Gender	(Male, Female)
CST	Caste	(General,SC,ST,OBC,MOBC)
TNP	Class X Percentage	(Best, Very Good, Good, Pass, Fail) If percentage >=80 then Best If percentage >= 60 but less than 80 then Very Good If percentage >= 45 but less than 60 then Good If Percentage >= 30 but less than 45 then Pass If Percentage < 30 then Fail
TWP	Class XII Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
IAP	Internal Assessment Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ESP	End Semester Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ARR	Whether the student has back or arrear papers	(Yes, No)
MS	Marital Status	(Married, Unmarried)
LS	Lived in Town or Village	(Town, Village)
AS	Admission Category	(Free, Paid)
FMI	Family Monthly Income (in INR)	(Very High, High, Above Medium, Medium, Low) If FMI >= 30000 then Very High If FMI >= 20000 but less than 30000 then High If FMI >= 10000 but less than 20000 then Above Medium If FMI >= 5000 but less than 10000 then Medium If FMI is less than 5000 then Low The figures are expressed in INR.
FS	Family Size	(Large, Average, Small) If FS > 12 then Large If FS >= 6 but less than 12 then Average If FS < 6 then Small
FQ	Father Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
MQ	Mother Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
FO	Father Occupation	(Service, Business, Retired, Farmer, Others)
MO	Mother Occupation	(Service, Business, Retired, Farmer, Others)
NF	Number of Friends	(Large, Average, Small) Same as Family Size
SH	Study Hours	(Good, Average, Poor) >= 6 hours Good >= 4 hours Average < 2 hours Poor
SS	Student School attended at Class X level	(Govt., Private)
ME	Medium	(Eng,Asm,Hin,Ben)
TT	Home to College Travel Time	(Large, Average, Small) >= 2 hours Large >=1 hours Average < 1 hour Small
ATD	Class Attendance Percentage	(Good, Average, Poor) If percentage >= 80 then Good If percentage >= 60 but less than 80 then Average If Percentage < 60 then poor

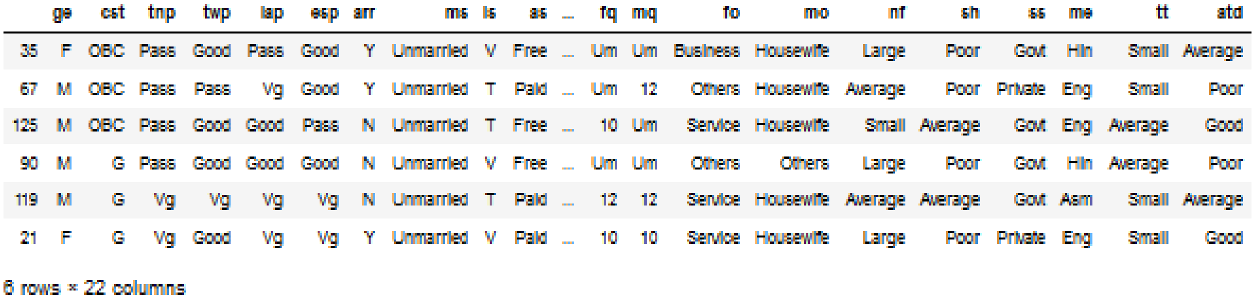
Dataset Description/Schema [2]

METADATA:

- ge's type is nominal, range is ('M', 'F')
- cst's type is nominal, range is ('G', 'ST', 'SC', 'OBC', 'MOBC')
- tnp's type is nominal, range is ('Best', 'Vg', 'Good', 'Pass', 'Fail')
- twp's type is nominal, range is ('Best', 'Vg', 'Good', 'Pass', 'Fail')
- iap's type is nominal, range is ('Best', 'Vg', 'Good', 'Pass', 'Fail')
- esp's type is nominal, range is ('Best', 'Vg', 'Good', 'Pass', 'Fail')
- arr's type is nominal, range is ('Y', 'N')
- ms's type is nominal, range is ('Married', 'Unmarried')
- ls's type is nominal, range is ('T', 'V')
- as's type is nominal, range is ('Free', 'Paid')
- fmi's type is nominal, range is ('Vh', 'High', 'Am', 'Medium', 'Low')
- fs's type is nominal, range is ('Large', 'Average', 'Small')
- fq's type is nominal, range is ('Il', 'Um', '10', '12', 'Degree','Pg')
- mq's type is nominal, range is ('Il', 'Um', '10', '12', 'Degree','Pg')
- fo's type is nominal, range is ('Service', 'Business', 'Retired', 'Farmer','Others')
- mo's type is nominal, range is ('Service', 'Business', 'Retired', 'Housewife', 'Others')
- nf's type is nominal, range is ('Large', 'Average', 'Small')
- sh's type is nominal, range is ('Good', 'Average', 'Poor')
- ss's type is nominal, range is ('Govt', 'Private')
- me's type is nominal, range is ('Eng', 'Asm', 'Hin', 'Ben')
- tt's type is nominal, range is ('Large', 'Average', 'Small')
- atd's type is nominal, range is ('Good', 'Average', 'Poor')

RESULTS/FINDINGS:

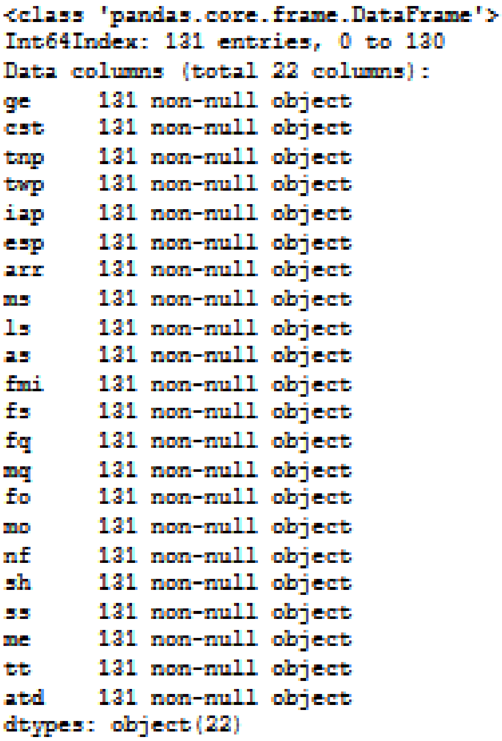
Importing Data: The data is in WEKA ".arff" file format, so we had to use a function from Scipy library to load the file in our notebook. The meta above was generated when we loaded the file. However, the data, that was loaded and then converted into pandas dataframe. was encoded in ‘utf-8’ format, so we had to decode all the columns of the dataframe. After peaking a sample of the data, the following subset of the data could be seen:



	ge	cst	tnp	twp	iap	esp	arr	ms	ls	as	...	fq	mq	fo	mo	nf	sh	ss	me	tt	atd
35	F	OBC	Pass	Good	Pass	Good	Y	Unmarried	V	Free	...	Um	Um	Business	Housewife	Large	Poor	Govt	Hin	Small	Average
67	M	OBC	Pass	Pass	Vg	Good	Y	Unmarried	T	Paid	...	Um	12	Others	Housewife	Average	Poor	Private	Eng	Small	Poor
125	M	OBC	Pass	Good	Good	Pass	N	Unmarried	T	Free	...	10	Um	Service	Housewife	Small	Average	Govt	Eng	Average	Good
90	M	G	Pass	Good	Good	Good	N	Unmarried	V	Free	...	Um	Um	Others	Others	Large	Poor	Govt	Hin	Average	Poor
119	M	G	Vg	Vg	Vg	Vg	N	Unmarried	T	Paid	...	12	12	Service	Housewife	Average	Average	Govt	Asm	Small	Average
21	F	G	Vg	Good	Vg	Vg	Y	Unmarried	V	Paid	...	10	10	Service	Housewife	Large	Poor	Private	Eng	Small	Good

6 rows × 22 columns

INSPECTING DATA: We first checked the data types of all the columns that we imported:

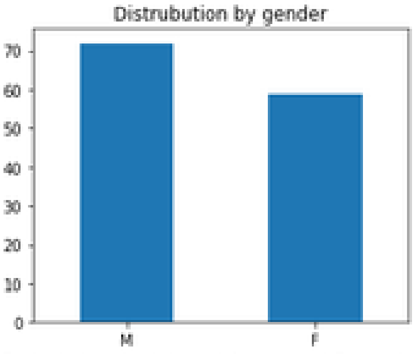


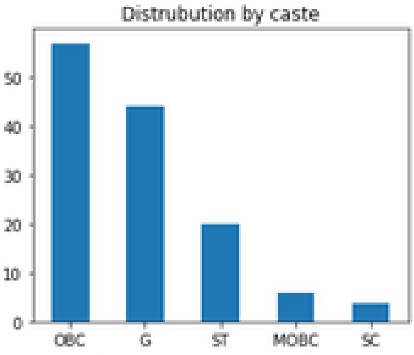
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 131 entries, 0 to 130
Data columns (total 22 columns):
ge      131 non-null object
cst     131 non-null object
tnp     131 non-null object
twp     131 non-null object
iap     131 non-null object
esp     131 non-null object
arr     131 non-null object
ms      131 non-null object
ls      131 non-null object
as      131 non-null object
fmi     131 non-null object
fs      131 non-null object
fq      131 non-null object
mq      131 non-null object
fo      131 non-null object
mo      131 non-null object
nf      131 non-null object
sh      131 non-null object
ss      131 non-null object
me      131 non-null object
tt      131 non-null object
atd     131 non-null object
dtypes: object(22)
```

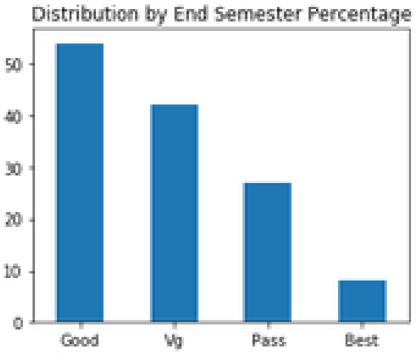
By looking at the above data, we can see that all the types of data are ‘object’. Therefore, we will need to label encode it later when we use Machine Learning algorithms on the data. We also see that the number of rows in the dataset are 131, whereas in their research paper they have claimed that the number of rows is 300. This means that the data they have uploaded for public use, has missing rows in it. Also, by further examining data in our python notebook, we realized that every column has categorical values in them, none of the values are continuously numerical.

Cleaning Data: The number of duplicated rows in the dataset are zero. There are also no missing values or null values in the dataset. Therefore, no cleaning of data was necessary as the dataset does not have any problems with it.

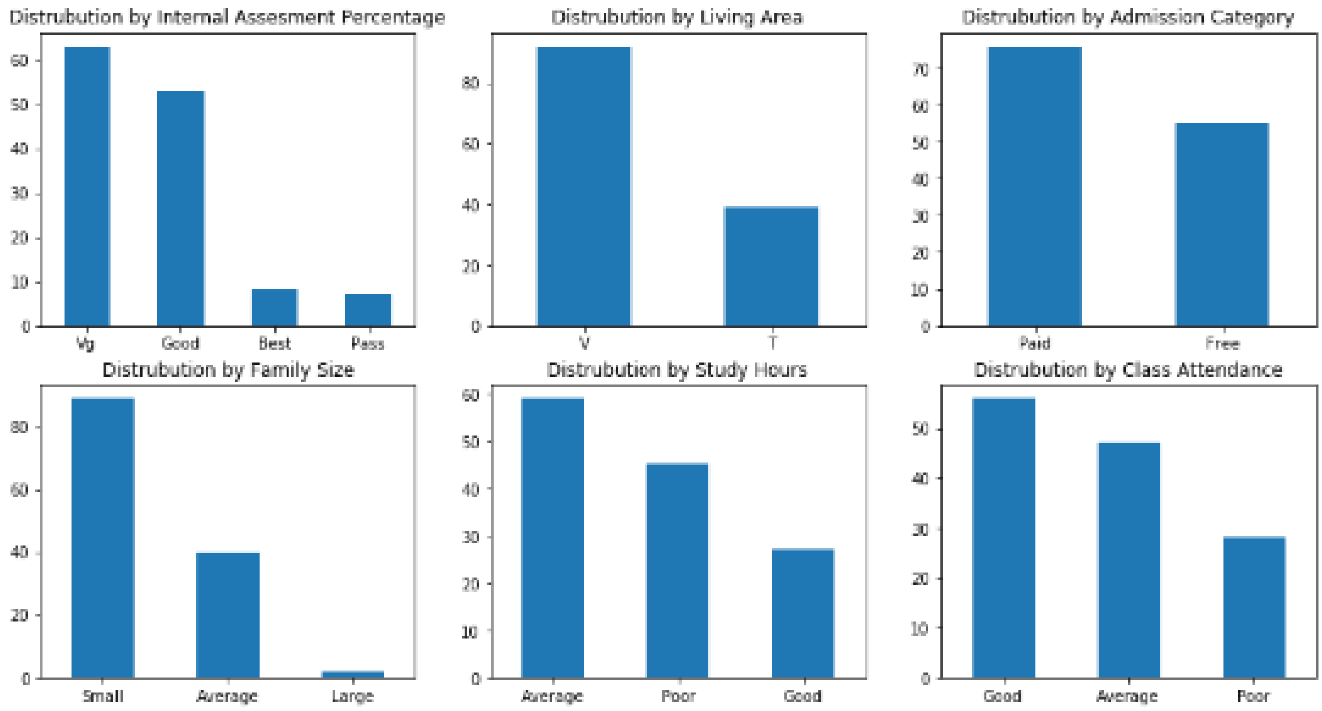
Visualizing Data: We will start by plotting the graphs of some of the columns in the dataset and try to analyze them. So, we plot bar graphs of each of the attributes against their appropriate frequencies:


Distrubution by gender


Distrubution by caste


Distribution by End Semester Percentage

By looking at the above plots, the number of males in the data set are more than females. Also, most of these students are from “Other Backward Caste” category, whereas the least of the students belong to “Schedule Caste” category. We also plotted the End-Semester Grades and found out that most of the students received Good (60 <= esp <80) grades, whereas only a few of the students received the Best grades (esp >= 80); however, none of the students failed. The rest of the bar graphs that we made are below:



Distrubution by Internal Assesment Percentage

Distrubution by Living Area

Distrubution by Admission Category

Distrubution by Family Size

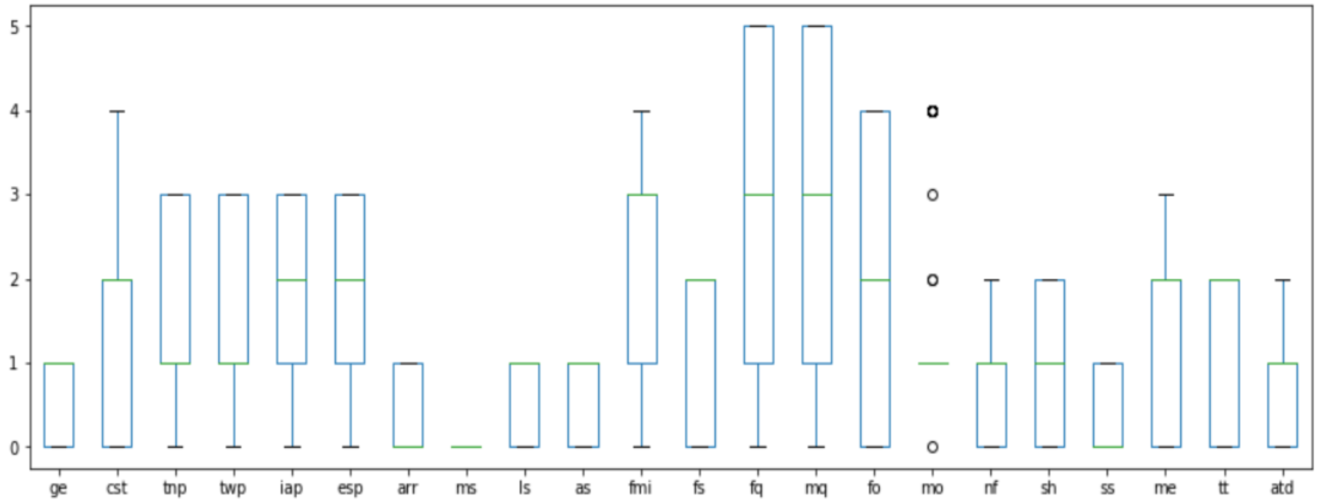
Distrubution by Study Hours

Distrubution by Class Attendance

The most interesting findings I made from these graphs are:

- Most of the students Internal Assesment percentages are very good.
- Most of the students are from Villages.
- Only a small number of students have large families or, to be specific, more than 12 students. Most of them have less than 6 people in their homes.

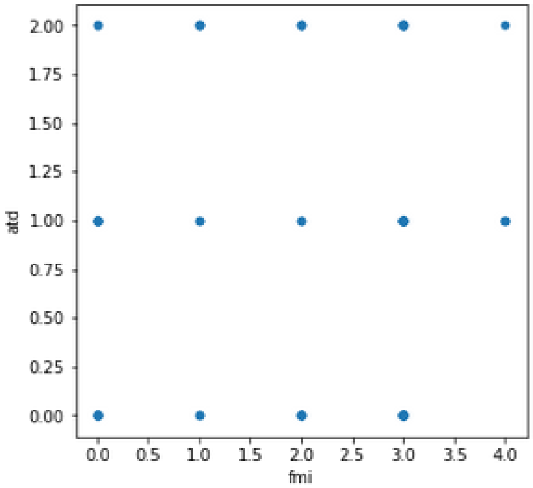
Box Plots: As the data is categorical, we were forced to first use the label encoder to change the type of data from objects into numeric. This allowed us to go ahead and plot box plots, even though it won't be the best way to visualize categorical data. Let's go ahead and look at the plot:

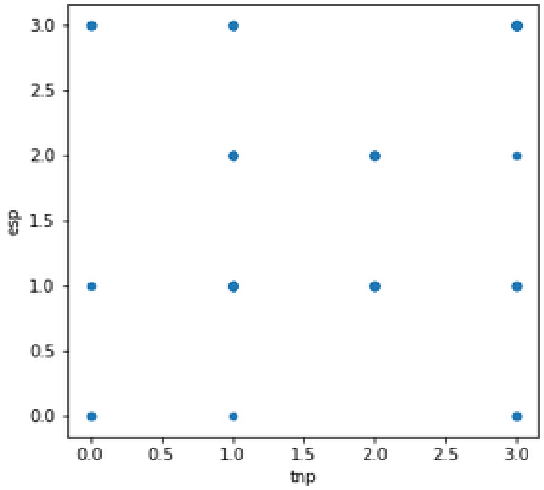


As a I briefed earlier, most of the plots right above won't make sense. We make some of the most interesting analysis from the above graph:

- The average number of students' fathers are "Illiterate". Least of the students' fathers have passed just Class Ten Examinations. The same can be seen from the box plot of Mothers Qualifications.
- Most of the students' family incomes are from Above Medium, High and Low classes.
- There is something not right about the Marital Status of the student's; seems like most of the students are unmarried/single, which makes sense as the students haven't aged much.

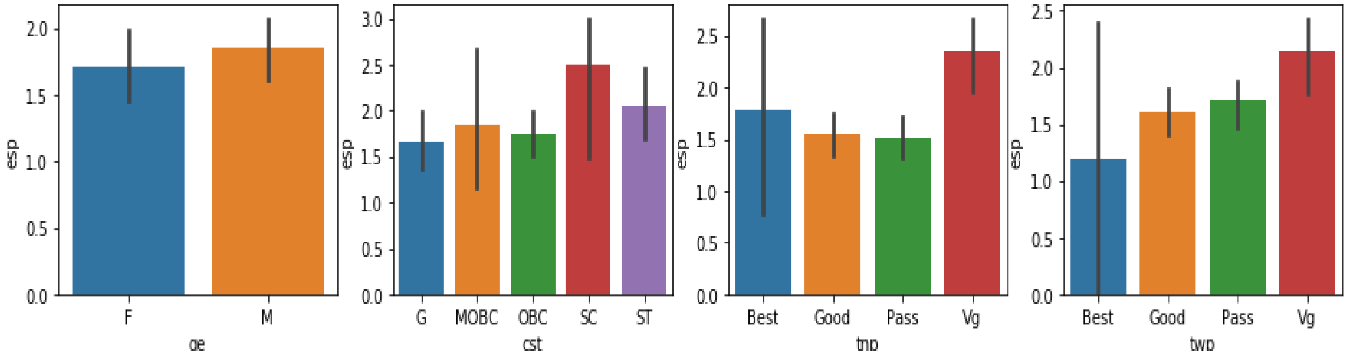
Scatter Plots: The dataset contains categorical attributes so it won't make sense plotting scatter plots of the data. Let's look at two of them:



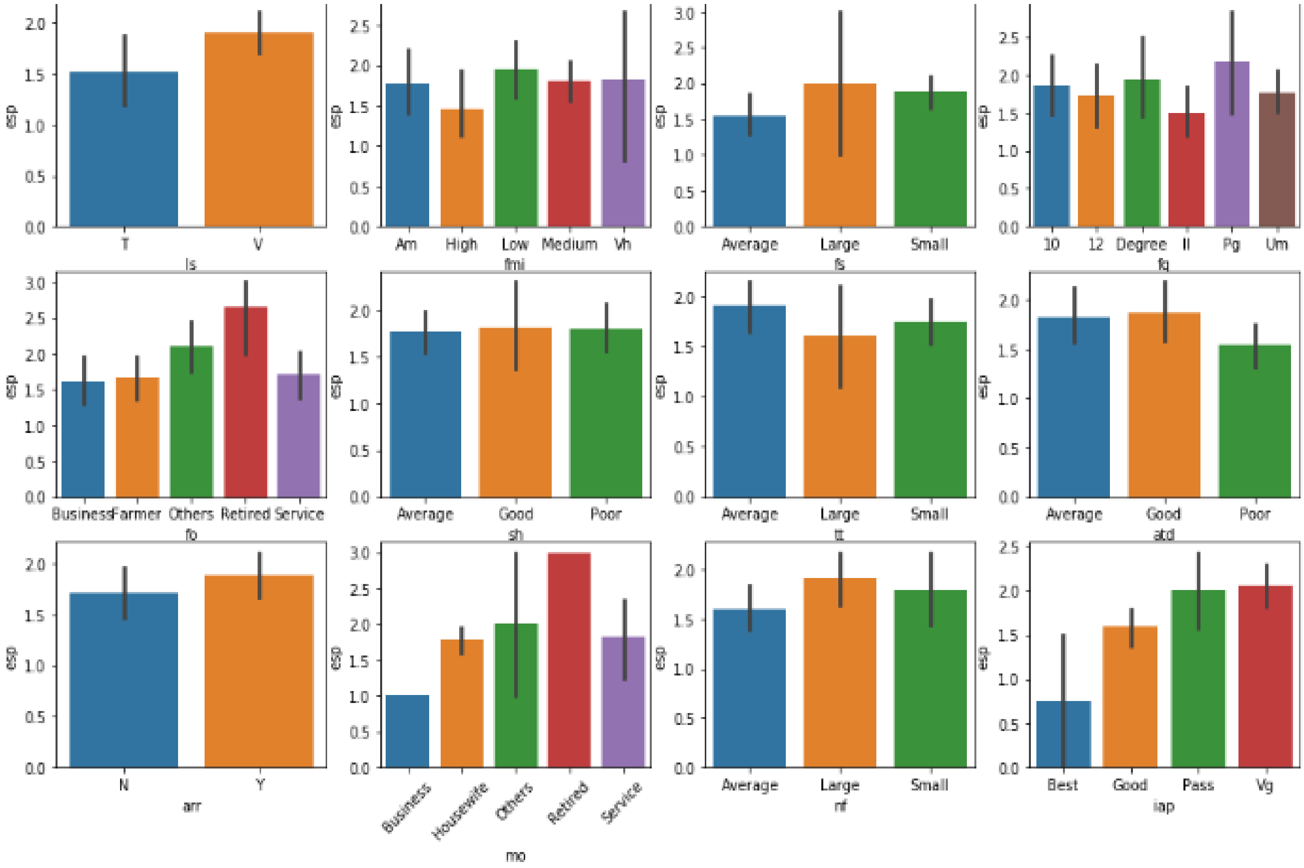


The left graph is Attendance percentage vs Family Monthly Income, whereas the right graph is End-Semester-Percentage vs Class X Percentage. The data seems evenly distributed in both graphs, that is why the scatter plots don't tell us any useful information.

Correlation Graphs: Now we will experiment with some graphs to test out which graph gives us a better visualization of response variable against other indicators. Firstly, lets try to use seaborn bar plots using some of the attributes:



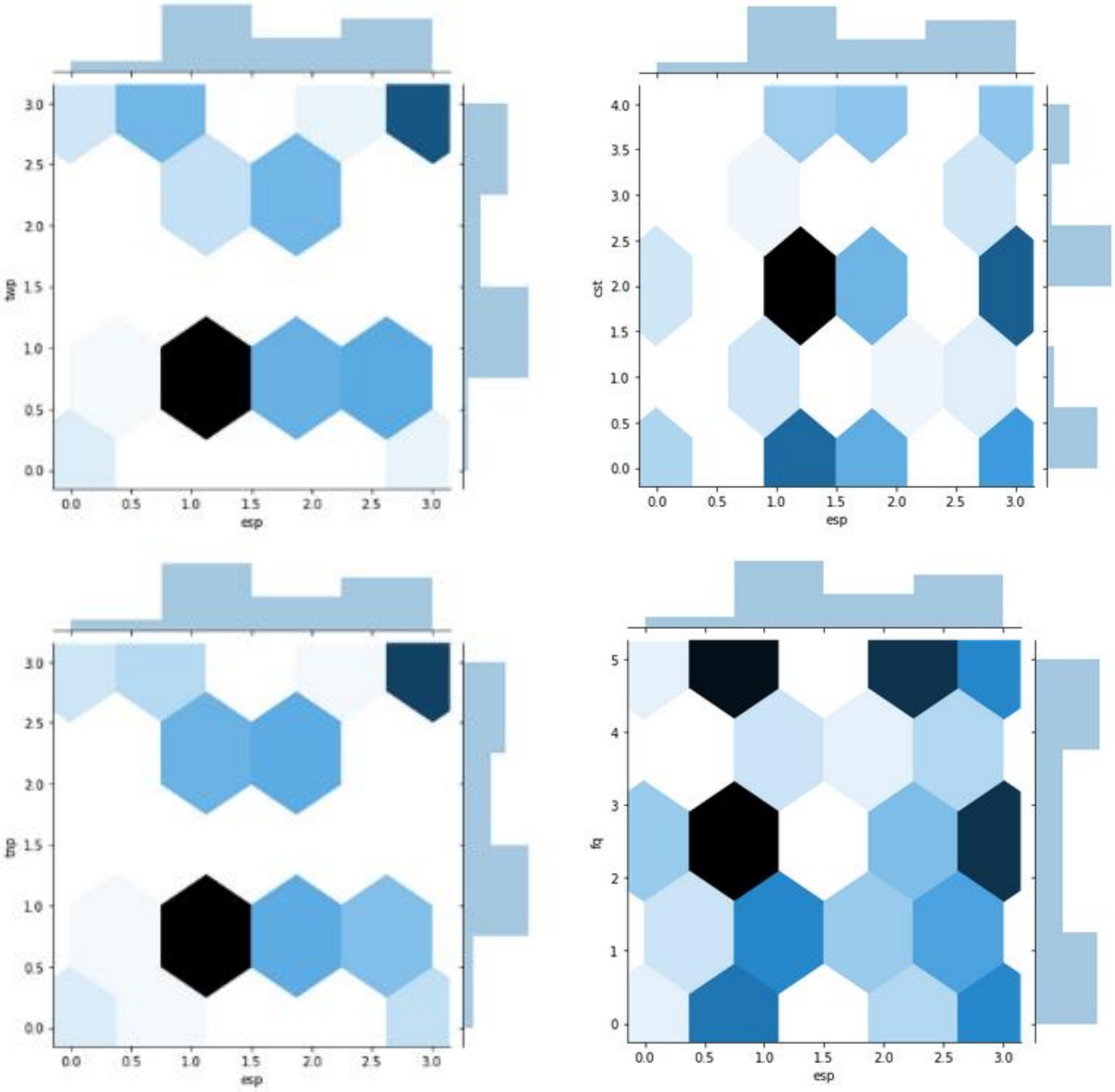
From L to R: End-Semester-Percentage vs (Gender, Caste, Class X Percentage, Class XII Percentage)



From L-R and then by Row: End-Semester vs (Living Status, Family Income, Family Size, Father Qualification, Father Occupation, Study Hours, Travel Time, Attendance Percentage, Arrear Papers, Mother Occupation, Number of Friends, Internal Assesment)

From the above graphs it seems that End Semester Percentage correlates better with columns other than Gender, Attendance, Travel Time, Study Hours and Arrears. However, this does not seem to be the optimal way to correlate data.

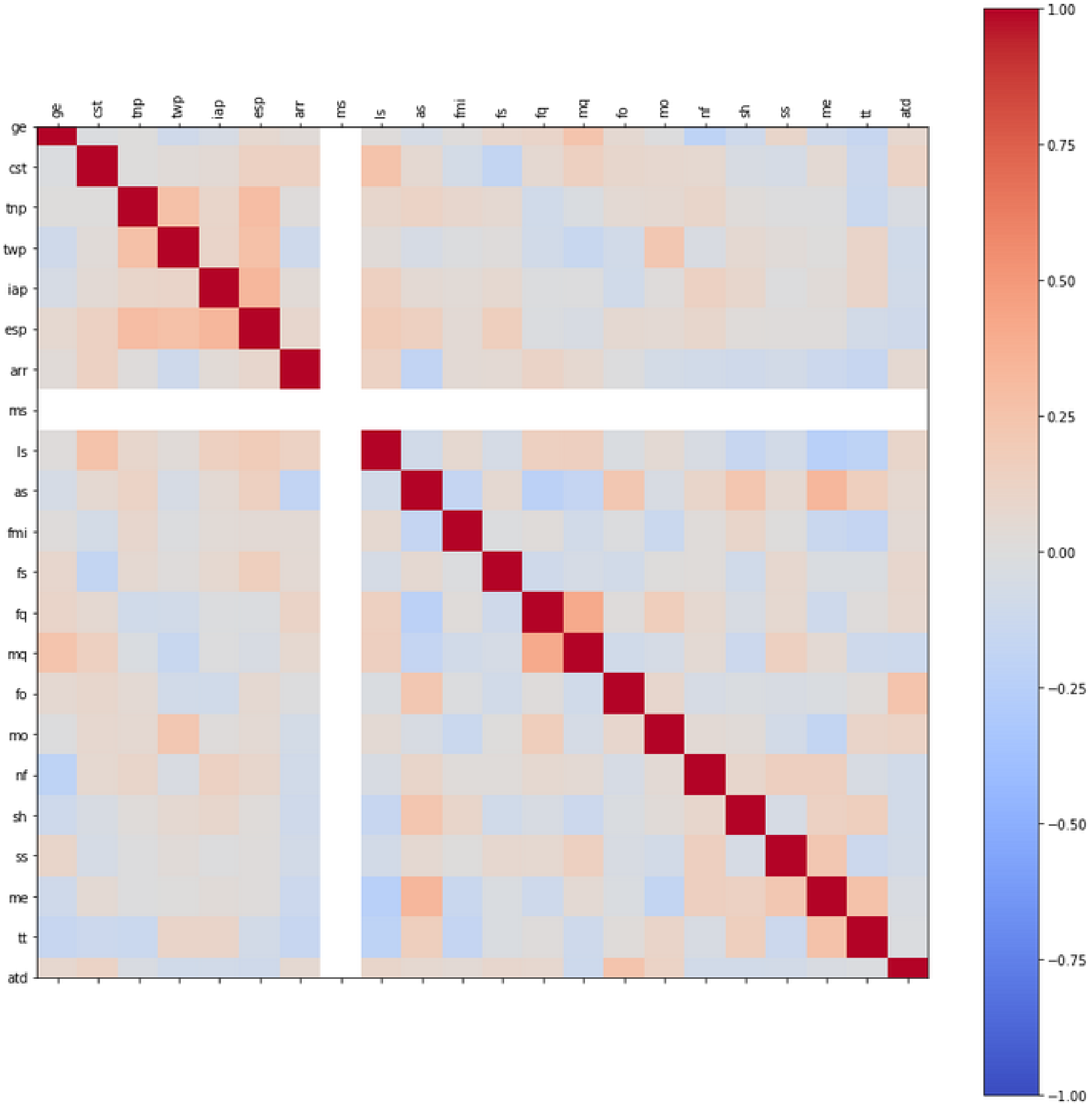
Let's now move towards jointplots, which can specifically tell us the relation between two attributes:



From L-R and then by Row: (Class XII, Caste, Class X, Father Qualification) vs End-Semester-Percentage

From the above graphs, the ones with more dark colors show us better correlation between two attributes. Therefore, Family Qualification and caste shows us better correlation than Class XII and Class X percentages. Class XII also shows better correlation than Class X percentage, which means that students’ performances are based better on recent results than on very old results. However, these graphs take too much space and time to tell us how the indicators correlate with the response variable. We need a better way to show the correlation and also let us know how each attribute's correlation differs with each other.

We will use pandas correlation (**Pearson's Correlation Coefficient**) function to give us pair wise correlation and tell us how each attribute differentiates with each other. We then use seaborn's heatmap style matrix graph to show the result:



Correlation Heat Map Matrix using Seaborn

Hypothesis: Before, showing the final correlation values for the response variable with the dependent attributes, we will first make two hypothesis tests:

Test: Pearson's Correlation Coefficient – if absolute value of coefficient is less than 0.05 then these two attributes are independent, otherwise they are dependent

Gender vs Caste: Sample of 50

- **H0**: The two attributes are independent.
- **H1**: there is a dependency between the attributes.

  Result: **Coefficient= 6.64e-17** – Therefore, Hypothesis H0 is correct.

  Why: To find out relationships between each attribute, which could directly affect the response variable we are predicting, i.e End-Semseter Grades.

Medium of study vs End-Semester-Percentage: Sample of 50

- **H0**: The two attributes are independent.
- **H1**: there is a dependency between the attributes.

  Result: **Coefficient= 0.026032** – Therefore, Hypothesis H1 is correct.

  Why: To find out if the attribute Medium is correlated with response variable, which helps in feature selection.

Feature Selection: The Final Correlations of response variable End-Semester-Percentages are:

ge	0.069780	cst	0.140074	tnp	0.293045
twp	0.265710	iap	0.333076	esp	1.000000
arr	0.085867	ls	0.184471	atd	-0.096510
as	0.148154	fmi	0.045524	fs	0.158971
fq	-0.021203	mq	-0.039767	fo	0.058970
mo	0.047507	nf	0.087892	sh	0.017952
ss	0.009439	me	0.010564	tt	-0.072989

After looking at the coeffients, we select those attributes that have an absolute value of more than 0.05. Therefore, we drop the following attributes from the dataset and then move onto splitting data into training and testing splits: `['fmi', 'fq', 'mq', 'mo', 'sh', 'ss', 'me', 'ms'].` We drop the column 'ms' (marital status), because every student was single. We will first convert data to numpy then split data, with one-thirds going into the testing dataset. No scaling would be necessary as data is categorical and not numerical.

Supervised Learning: We use the following models to predict End-Semester-Percentage, they are sorted from best to worst:

Models	Accuracy on test set
Naïve Bayes Classifier (Gaussian)	0.59
Random Forest Classifier (random state=104)	0.5681
Decision Tree Classifier (random state=33)	0.5681
Logistic Regression (with default parameters)	0.527
Neural Network (with default parameters)	0.5

Therefore, Naïve Bayes Classifier works better than other models. Neural Network was the worst model. Even though the models aren't giving a good result, we can still see that Random Forrest, Decision Trees and Naïve Bayes works better than the other two models. This is because these models work best when the data is mostly categorical. However, there is not much disparity between results to prove why certain model is performing better than other.

CONCLUSION:

To conclude, we will summarize what we have done so far and answer the questions that we asked earlier for our research. We firstly asked the question about which model will perform better, and we saw above that all models didn't perform as better as they were supposed to. The best among them was Naïve Bayes, whereas the worst one was Neural Network. The columns that we used in our feature selection and the columns that the researchers used differ, this might be since they used 300 rows, whereas we were only provided with 131 rows. **This might also be the case why we might be getting very low accuracy**. We also saw that the student's socio-economic background and previous grades can have an impact in our conclusion, but not as much as we quite expected. However, certain parameters affected the grades less than others, for example study hours. We even found out that student's most recent past grades can have a better impact on their future grades, than grades that were least recent. We showed graphs of individual attributes to analyze how each attribute signify their importance. Lastly, we went through different visualization types to show correlation of attributes with the response variable.

References

1. The data set was provided by UCI. It is publicly available from the following link: https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance#
2. Research Paper Used: *Hussain, Sadiq, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwi and Najoua Ribata. "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA." (2018). Link: https://www.semanticscholar.org/paper/Educational-Data-Mining-and-Analysis-of-Students%E2%80%99-Hussain-Dahan/46b5436be736e5a48ab127b5a856e73e46487cc4#references*