UNIVERSITY OF TEXAS AT ARLINGTON

INSY_5339 - PRINCIPLES OF BUSINESS DATA MINING

FINAL PROJECT – SPRING 2020

"STUDENT SCORES ANALYSIS AND PREDCTION"

PRADEEP MEDAGIRI
1001757740

TABLE OF CONTENT:

OBJECTIVE:

To build three machine learning models for predicting the academic performance of a student by "class" i.e., L(low), M(medium), H(high) using independent variables and to determine which model's prediction gives the best result, implementing it using SAS.

DATA SOURCE:

Data has been acquired from a well know source "Kaggle" an online community of data scientists and machine learning practitioners.
Data Source Link: https://www.kaggle.com/aljarah/xAPI-Edu-Data

Dataset consists of 480 student records and 16 features. The features are classified into three major categories:

1. Demographic features such as gender and nationality.
   - The dataset consists of 305 males and 175 females. Students from different origins such as Kuwait, Jordan, Palestine, Iraq, Lebanon, Tunis, Saudi Arabia, Egypt, Syria, USA, Iran, Libya, Morocco and Venezuela.

2. Academic background features such as educational stage, grade Level and section.
   - The dataset is collected for two semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

3. Behavioural features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.
   - Parent participation feature has two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 parents that had answered the survey and 210 had not, 292 of the parents had been satisfied from the school and 188 had not.

ATTRIBUTES:

No	Variables	Data Type	Description
1	Gender	String	Gender of the student
2	Nationality	String	Nationality of the student
3	Place of birth	String	Place of birth of the student
4	Educational Stages	String	Educational level that the student belongs
5	Grade Levels	String	The grade that the student is enrolled in
6	Section ID	String	Class the student belongs
7	Topic	String	Course
8	Semester	String	School year semester
9	Parent	String	Parent responsible for the student
10	Raised hand	Numeric	Number of times the student has raised hands
11	Visited resources	Numeric	Number of times the student visited the course content
12	Viewing announcements	Numeric	Number of times the student checked the announcements

13	Discussion groups	Numeric	Number of times the student participated in group discussions
14	Parent Answering Survey	String	Did the parent answer the survey
15	Parent School Satisfaction	String	The degree of parent satisfaction from school
16	Student Absence Days	Numeric	Number of days a student was absent
17	Class	String	Indicator of the performance of the student (L,M,H)

The performance of the student is measured by the "class" variable which is divided into three groups.
L-This stands for low level and is given to students who score from 0 to 69
M-This is allotted to students who score somewhere between 70 and 89
H-This comprises of the most successful students who score more than 89 in their tests

SUMMARY STATISTICS:

```
In [4]: data.describe()
```

Out[4]:

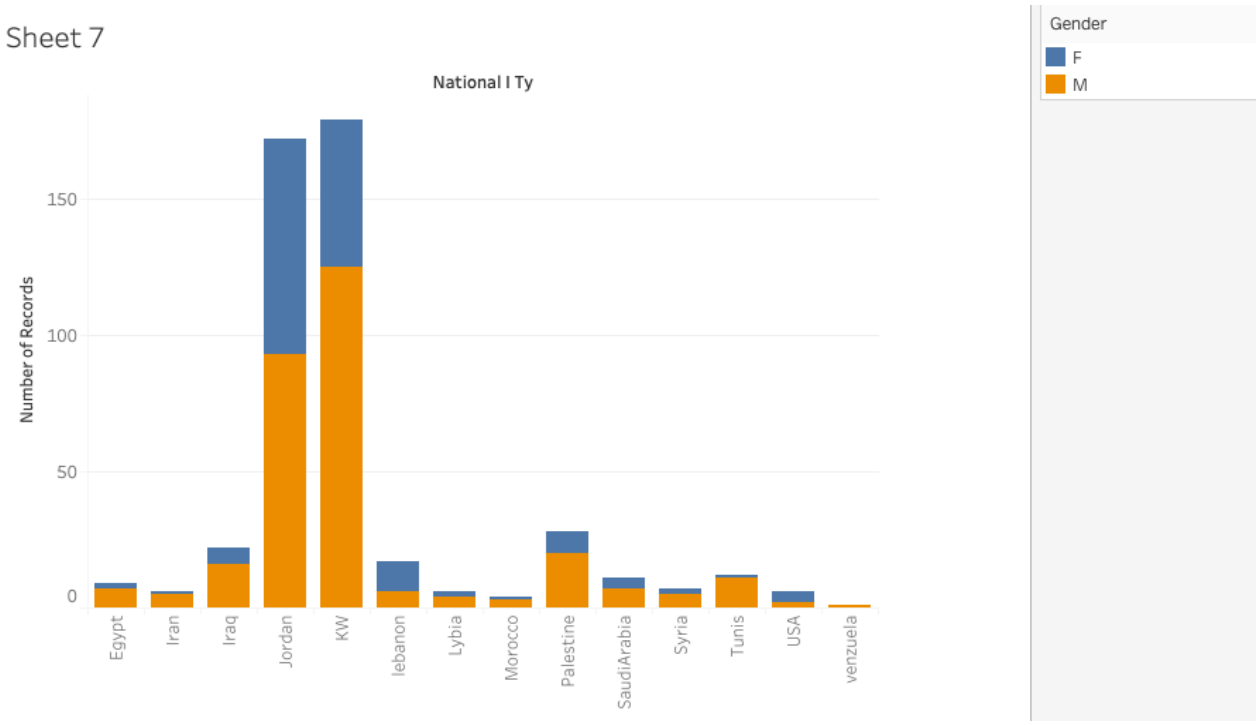
	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	480.000000	480.000000	480.000000	480.000000
mean	46.775000	54.797917	37.918750	43.283333
std	30.779223	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.750000	20.000000	14.000000	20.000000
50%	50.000000	65.000000	33.000000	39.000000
75%	75.000000	84.000000	58.000000	70.000000
max	100.000000	99.000000	98.000000	99.000000

Initial Results

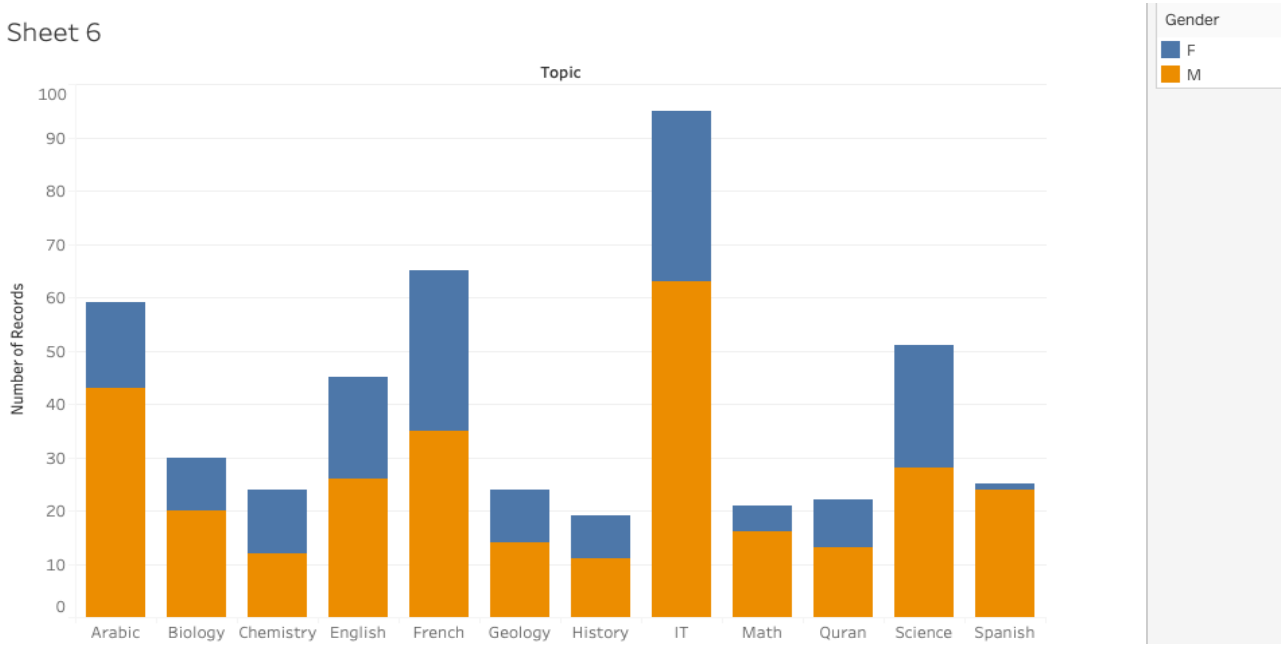
1. Data set has a total of 480 observations based on 16 features, having both integer and categorical variables.
2. Data is clean with no missing values, hence performing data pre-processing such as missing value treatments are not required.
3. In this data, by looking at one of the factors i.e., two semesters are considered, both fall and spring with 51% and 49% respectfully, so in this particular case the data is not biased and by following this approach of drilling down with more analytical steps, more insights can be drawn.

DATA VISUALIZATIONS:

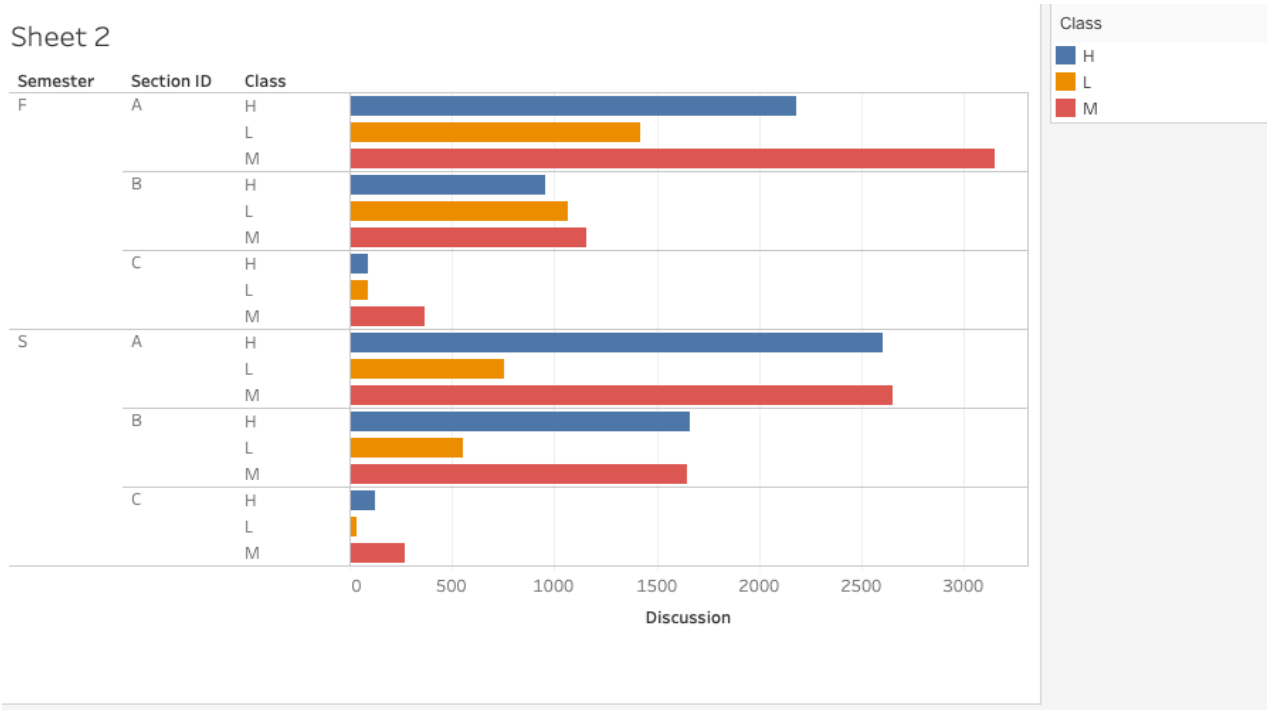
"Bar charts are used to understand the data. Visuals are produced in tableau"



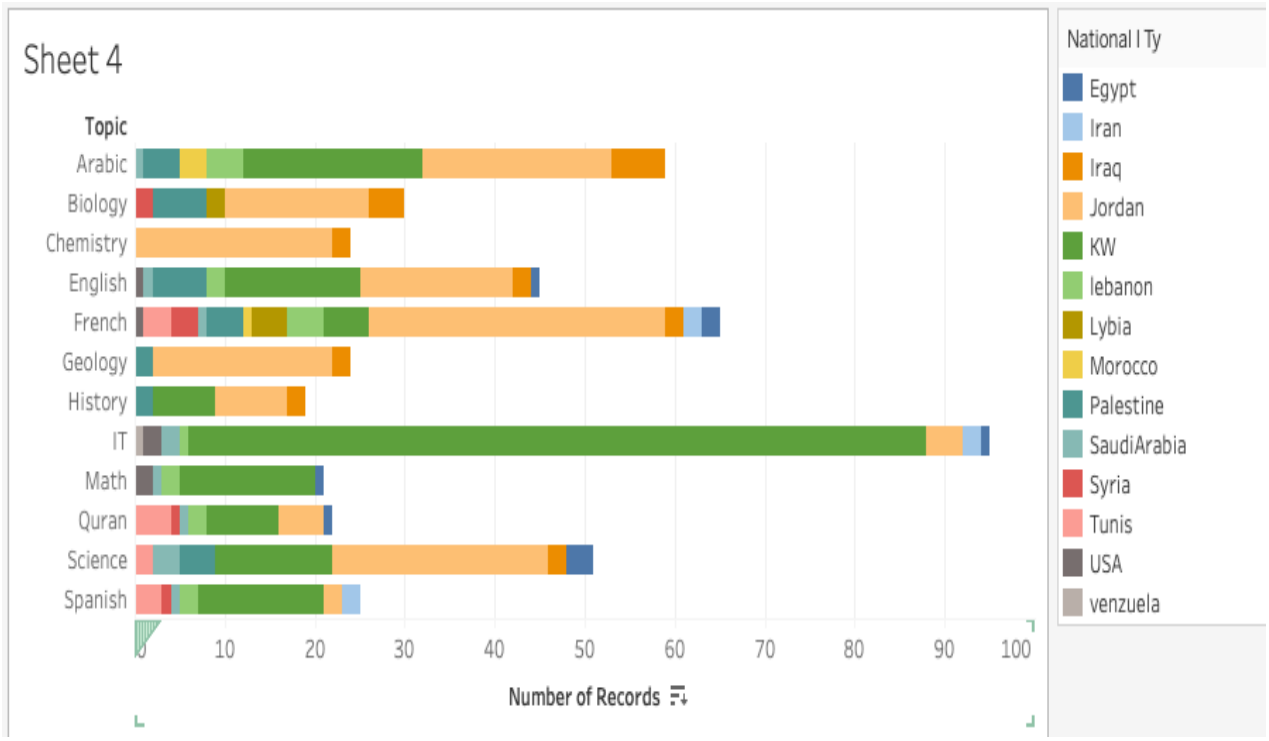
- Gender difference exists across all the countries. However, these differences did not vary much by each country.



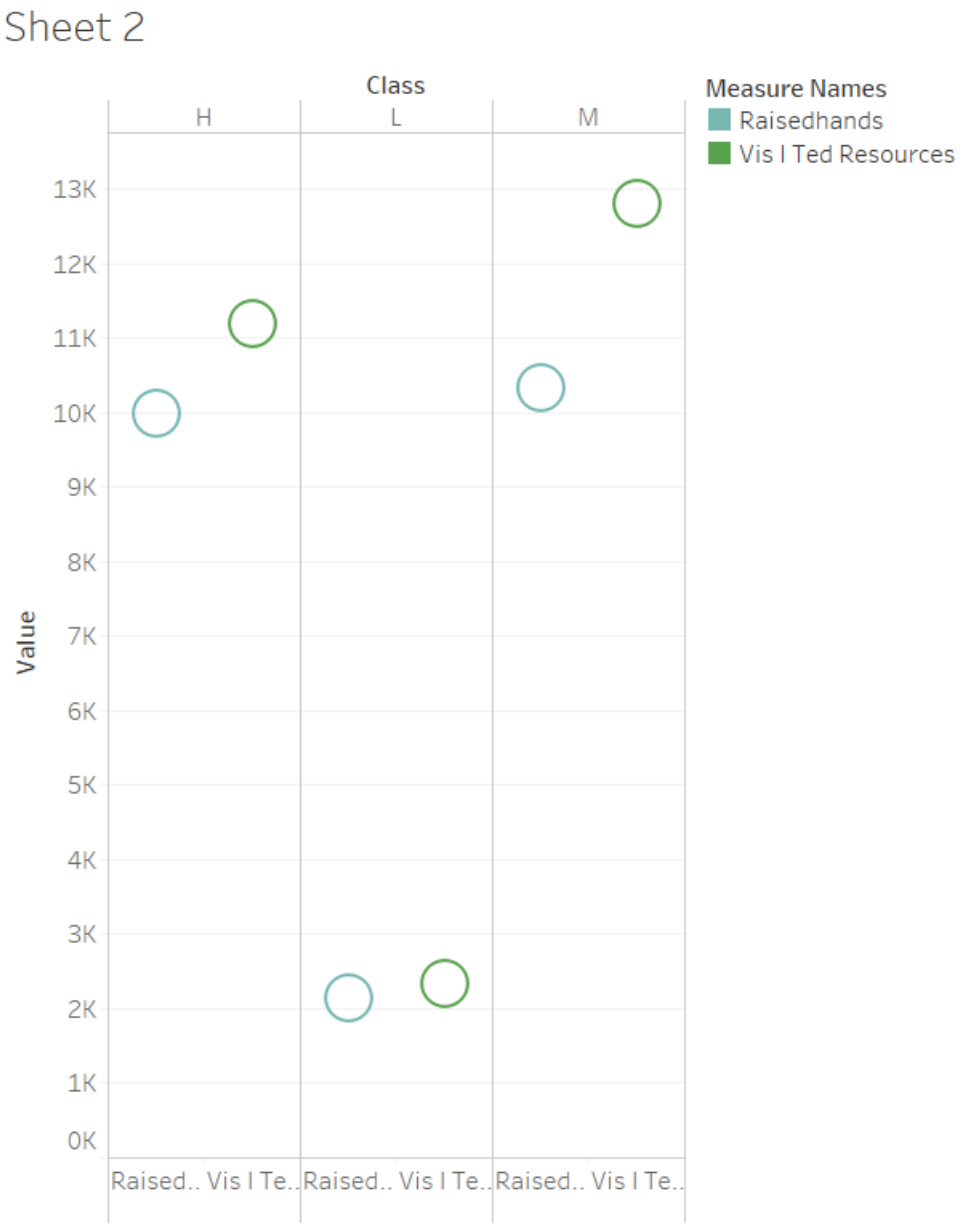
- There is almost no gender bias on choosing subjects. Only math and Spanish are dominated by one gender



- We can see the discussions in section c are very low compared to other sections in both semesters



- Some topics are dominated by a single country and some topics are evenly spread out

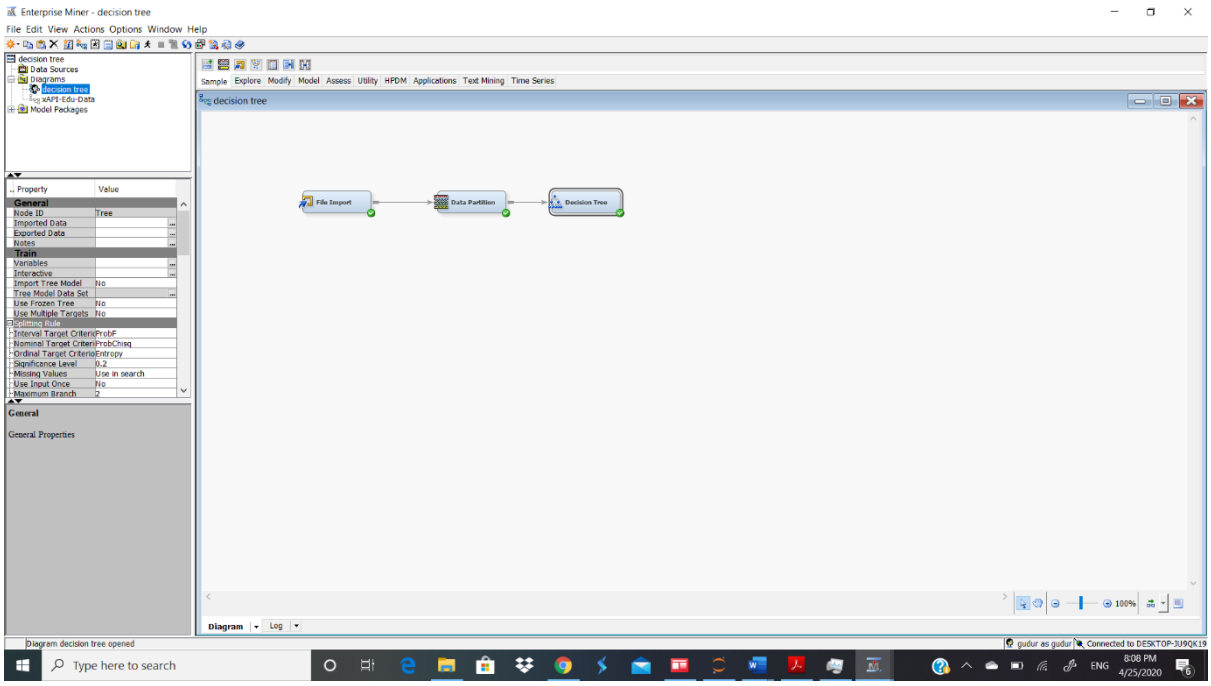


- We can see who ever exhibited low on academic performance used the resources less than the ones who exhibited high or medium level on academic performance

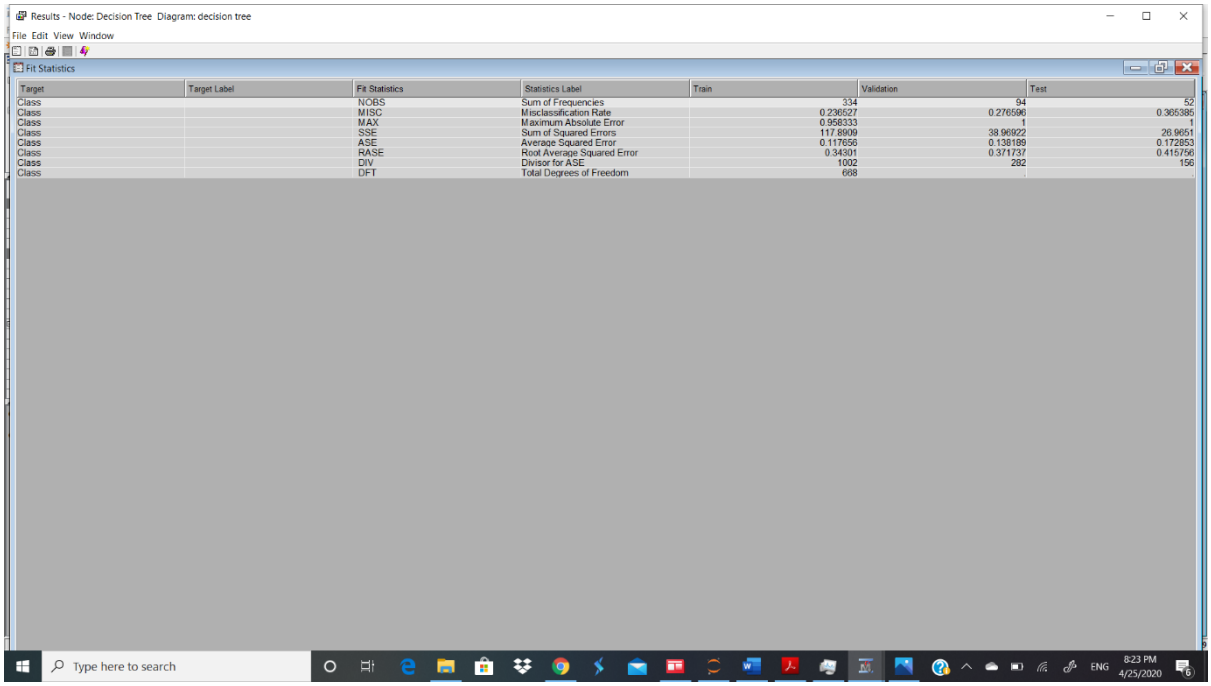
MODELS:

As target variable is categorical, so “Decision Tree”, “Random Forest” and “K Nearest Neighbors” predication techniques are used.

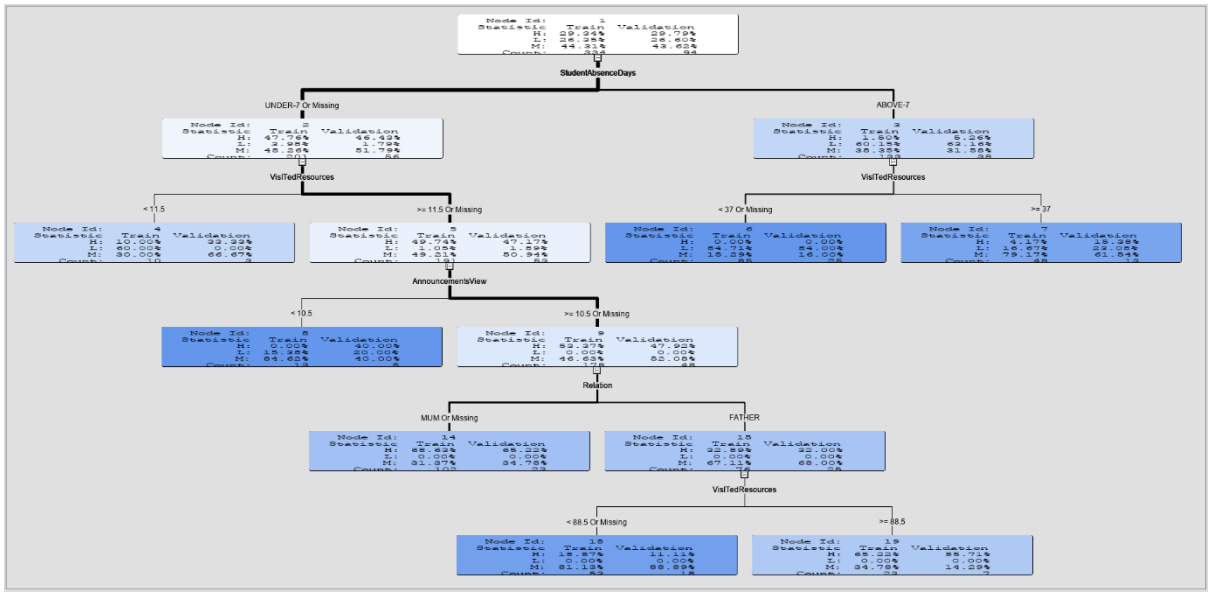
Decision Tree



- Using decision tree node to get our results

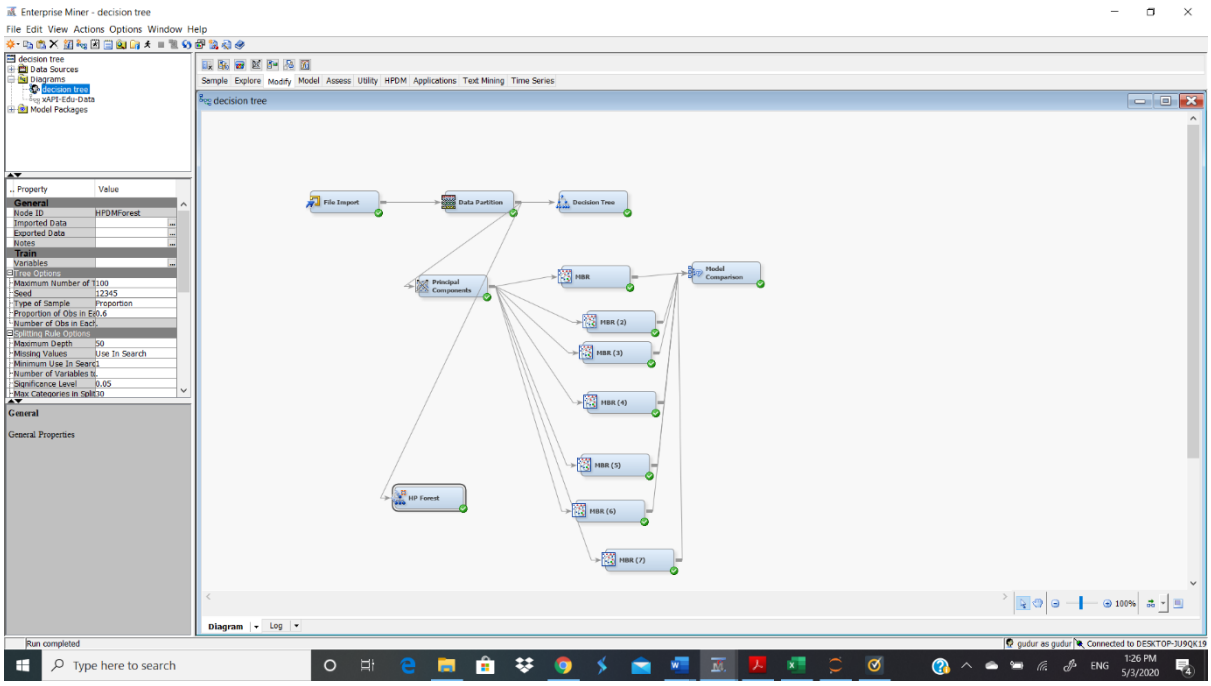


Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Class		NOBS	Sum of Frequencies	334	94	52
Class		MISC	Misclassification Rate	0.236527	0.276596	0.365385
Class		MAX	Maximum Absolute Error	0.958333	1	1
Class		SSE	Sum of Squared Errors	117.8909	38.96922	26.9651
Class		ASE	Average Squared Error	0.117656	0.136189	0.172853
Class		RASE	Root Average Squared Error	0.34301	0.371737	0.415756
Class		DIV	Divisor for ASE	1002	282	156
Class		DFT	Total Degrees of Freedom	668		



- We split the data into 70%-20%-10%(training-validation-test)
- First split occurred at Student absence days (under 7 days and above 7 days)
- We got average squared error for our target variable Class 0.11 for training data,0.13 for validation data,0.17 for test data
- Student absence days and visited resources both have high variable Importance

K Nearest Neighbor



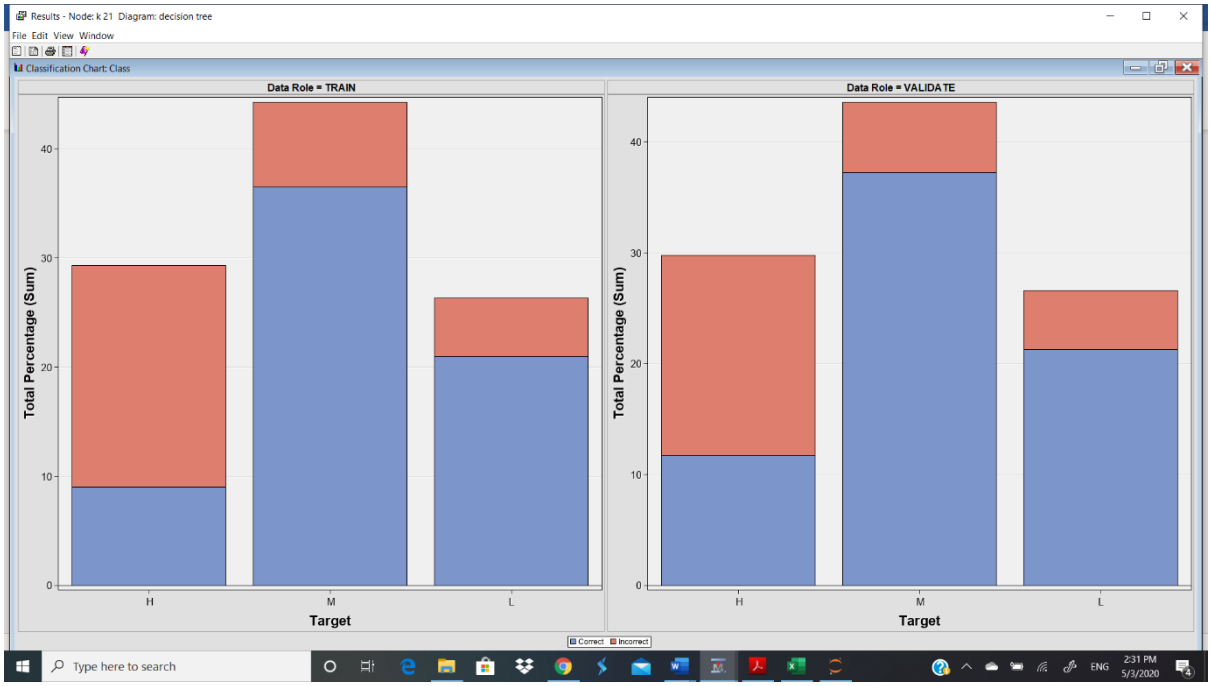
- The Memory-Based Reasoning node assumes that the variables with a model role of "input" are numeric, orthogonal to each other, and standardized. We use the Principal component node to generate numeric, orthogonal, and standardized variables that can be used as inputs for the Memory-Based Reasoning node.

- We select our k value around k=square root of the number of training data which is either 21 or 19 as we should take odd numbers to avoid confusion later. We changed k values accordingly and check them

K=7	0.39362
K=11	0.42553
K=16	0.40426
K=19	0.31915
K=21	0.29787
K=81	0.40426

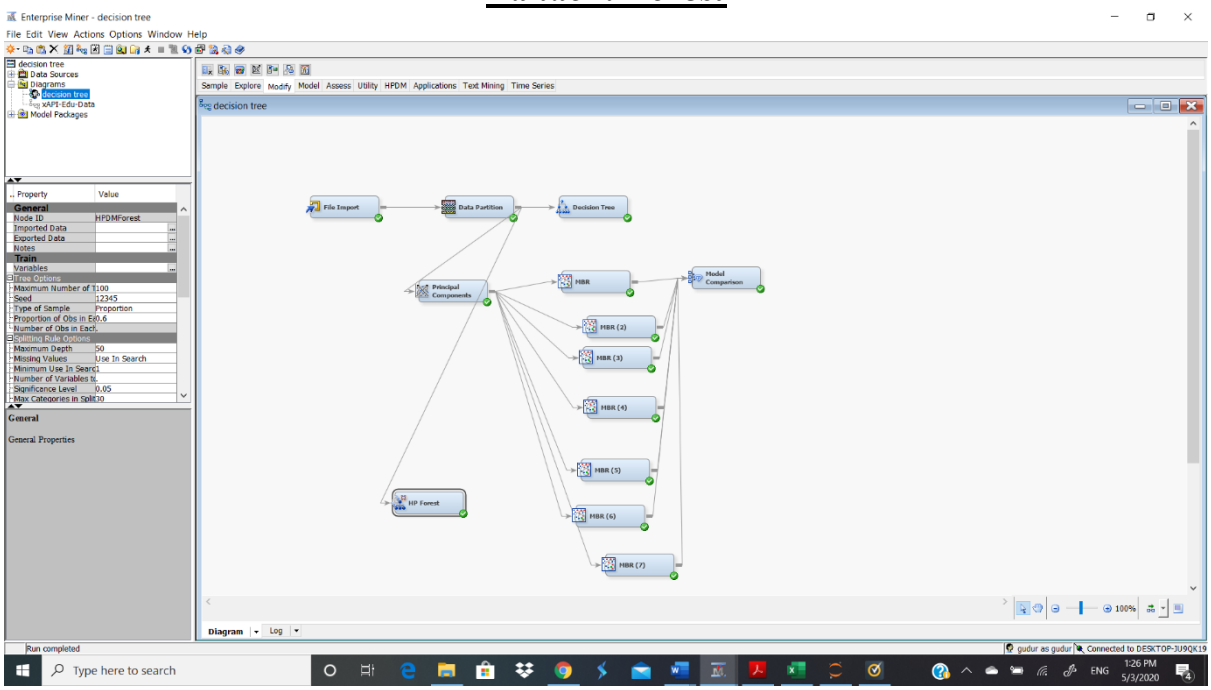
- K=21 has lowest error rate, so we pick 21 nearest neighbors
- We get the following for k-21

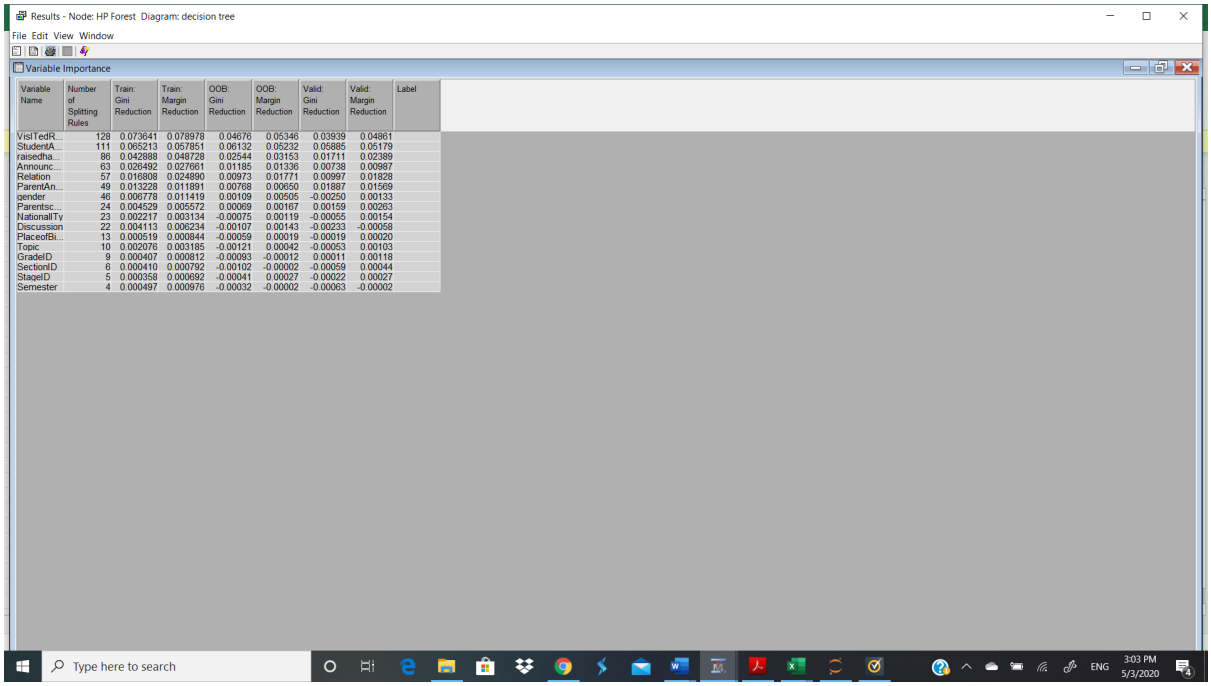
	Training	**Validation**	**Test**
Number of Wrong Classifications	112.00	28.000	21.000
Misclassification Rate	0.34	0.298	0.404



- The above picture shows how many correct or incorrect predictions have been made, class red represents incorrect and blue represents correct

Random Forest



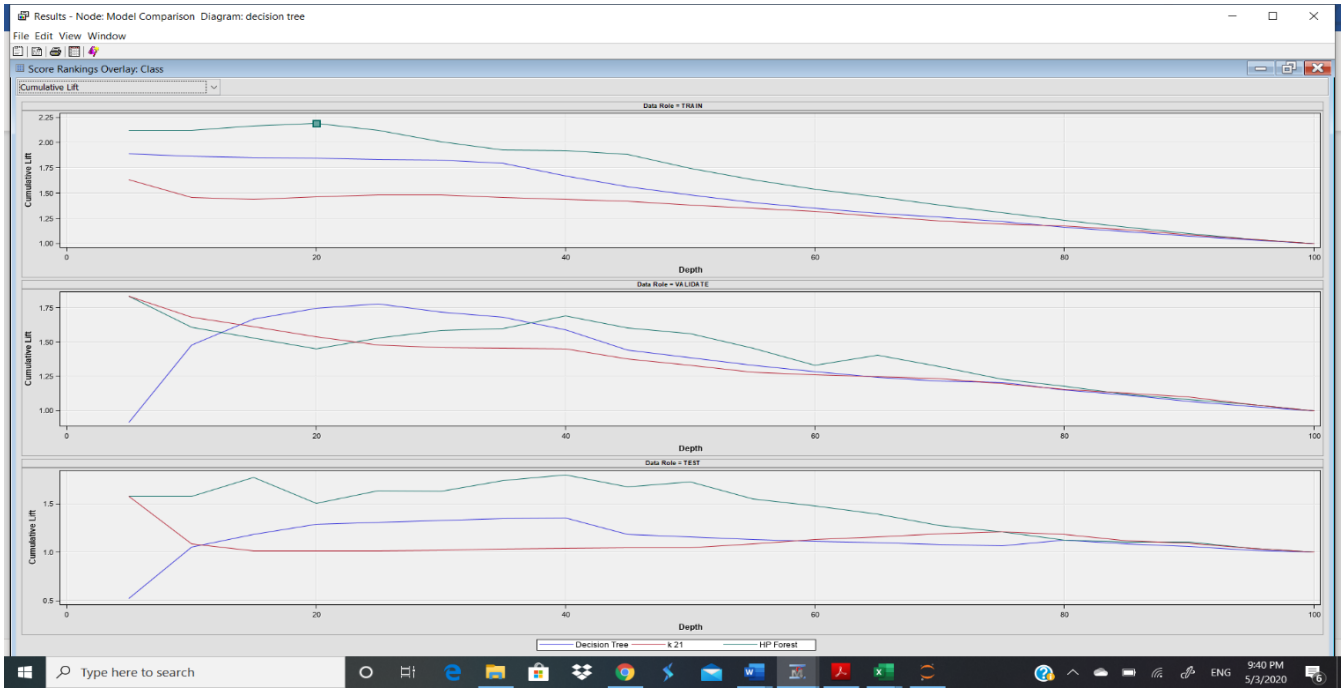


- We have observed, visited resources feature has highest variable importance followed by the student absent days feature
- For this algorithm we have

	Training	Validation	Test
Number of Wrong Classifications	49.00	26.000	13.000
Misclassification Rate	0.15	0.277	0.250

CONCLUSION-

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest	HP Forest	0.27660	0.10924	0.14671	0.13489
	Tree	Decision Tree	0.27660	0.11766	0.23653	0.13819
	MBR2	k 21	0.29787	0.15113	0.33533	0.15151



- **From the above picture we can observe cumulative lift for all the predictions**

- **Since this target variable is categorical, we should compare prediction based on misclassification rate rather than average squared error**

- **When we compare the three prediction algorithms, we can ascertain that random forest performed slightly better than both the decision tree and k nearest neighbor algorithms**

REFERENCE:

1. Students' Academic Performance Dataset https://www.kaggle.com/aljarah/xAPI-Edu-Data

2. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

3. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

4. Technologies – SAS Enterprise Miner | Tableau