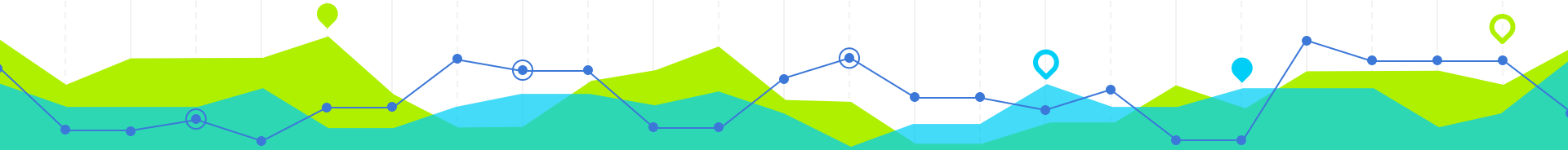


SC1015 Mini Project

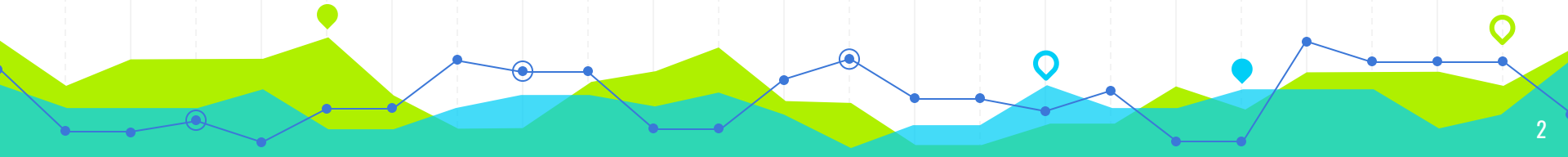


Students Performance In Exams

SC7 Team 7: Yao Xian(Robin), Akash, John

Problem Formulation

- ⦿ Does parental education really affect student results?
- ⦿ Does attending Test Prep Course affect student results?
- ⦿ Does having a proper lunch affects performance?
- ⦿ Does doing good in one subject mean they are good in another subject?



Data Introduction

```
In [3]: Students_data = pd.read_csv('StudentsPerformance.csv')
Students_data.head()
```

```
Out[3]:
```

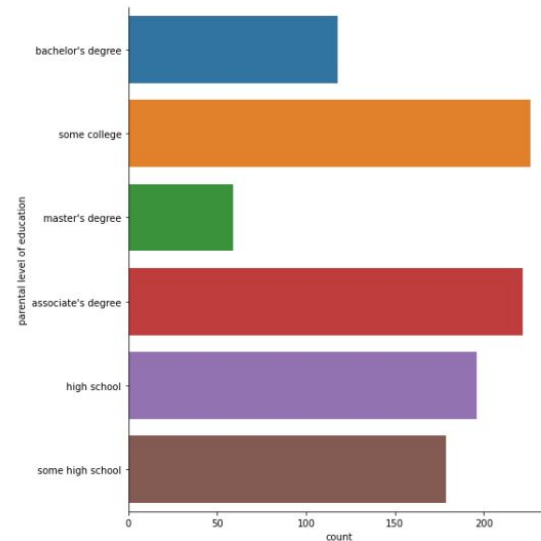
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

```
In [4]: print("Data type : ", type(Students_data))
print("Data dims : ", Students_data.shape)

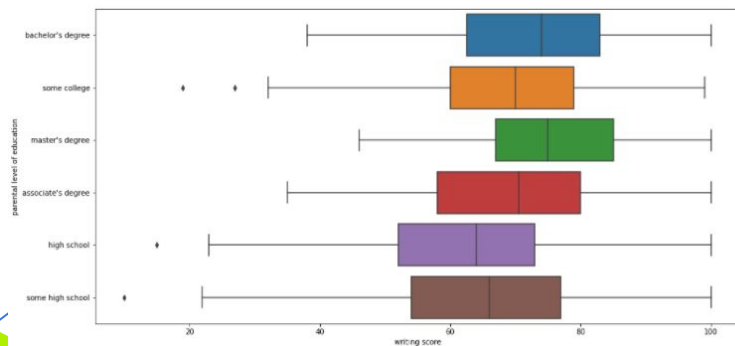
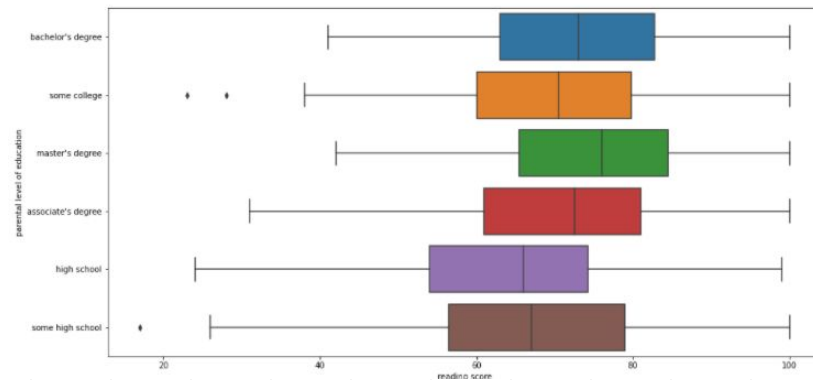
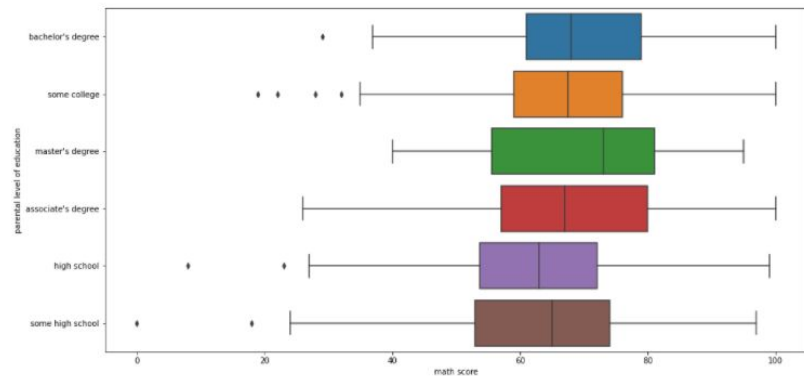
Data type : <class 'pandas.core.frame.DataFrame'>
Data dims : (1000, 8)
```

```
Number of Parental Level Types : 6
some college      226
associate's degree 222
high school      196
some high school  179
bachelor's degree 118
master's degree   59
Name: parental level of education, dtype: int64
```

```
Out[7]: <seaborn.axisgrid.FacetGrid at 0x19e3d77bf70>
```



Data Introduction





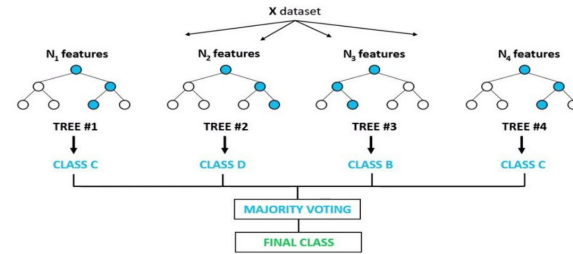
RFC VS SVM Vs Naive Bayes?

Done By: AKASH

1

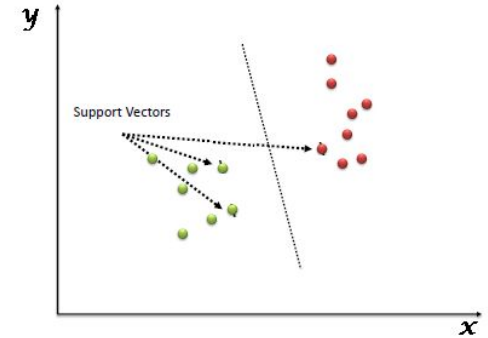
Random Forest Classifier

- Random Forest is a supervised learning algorithm
- Consists a large number of decision trees that operate as an ensemble
- Each Individual tree in the random forest spits out a class prediction
- The class with the most votes becomes our model prediction
- Fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting
- The model is generally trained using a bagging method



SUPPORT VECTOR MACHINE

- Is a supervised machine learning algorithm
- We plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.
- Then, we perform classification by finding the hyper-plane that differentiates the two classes very well .



GAUSSIAN NAIVE BAYES

- Uses various features to discriminate different objects
- Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred.
- Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other.
- Hence it is called naive.
- 3 Types of Naive Bayes Classifier : Multinomial , Bernoulli & Gaussian

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

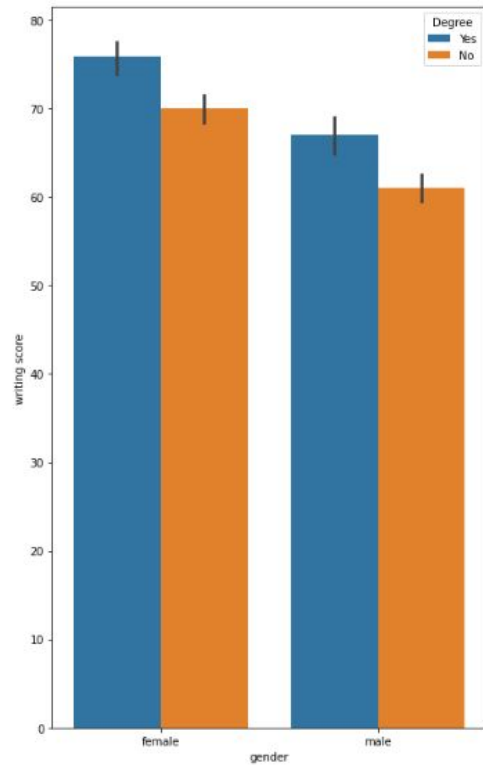
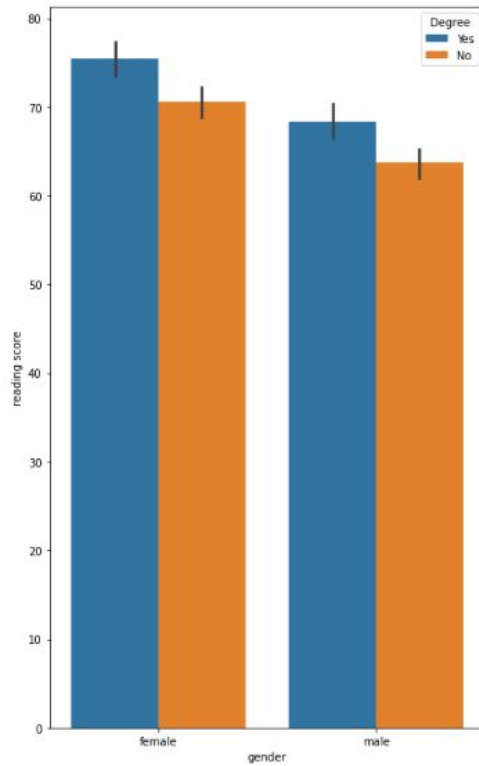
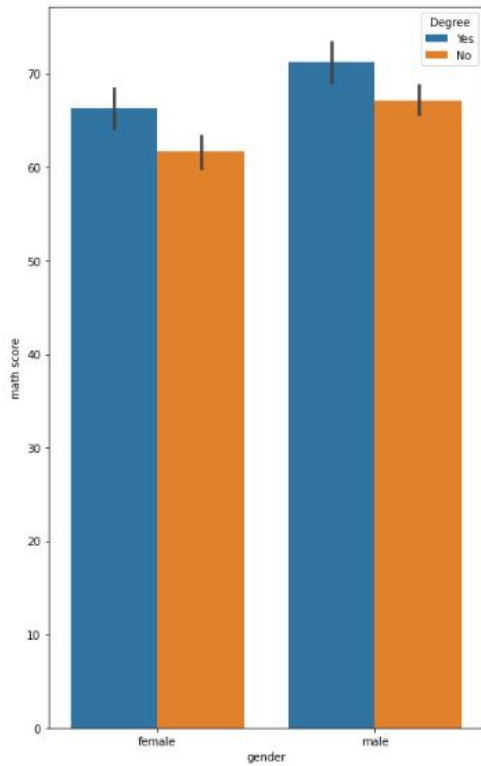


Does parental education really affect student results?

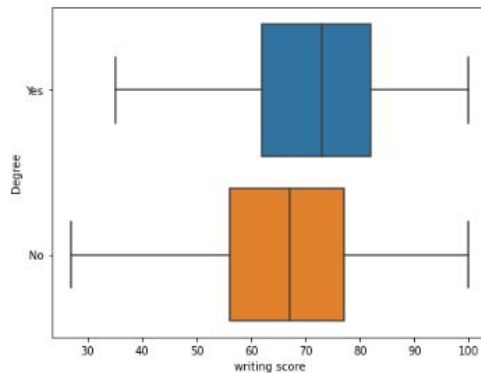
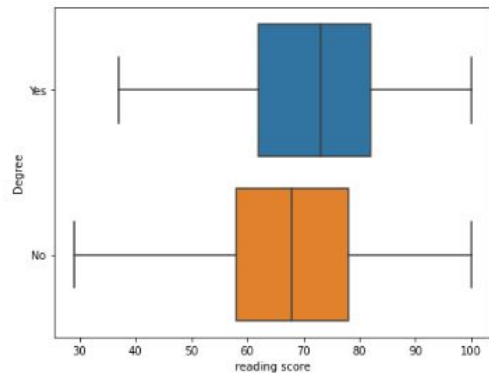
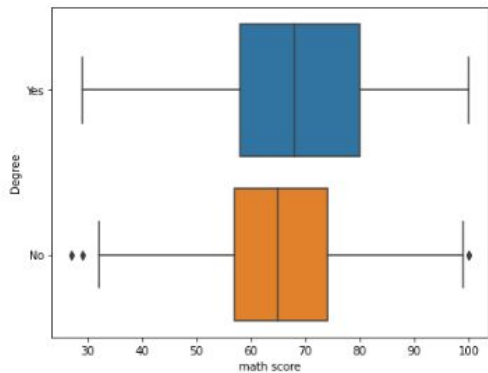
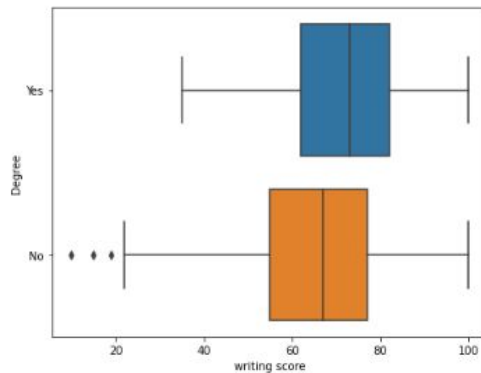
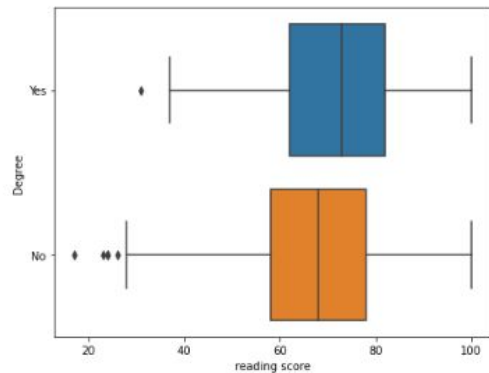
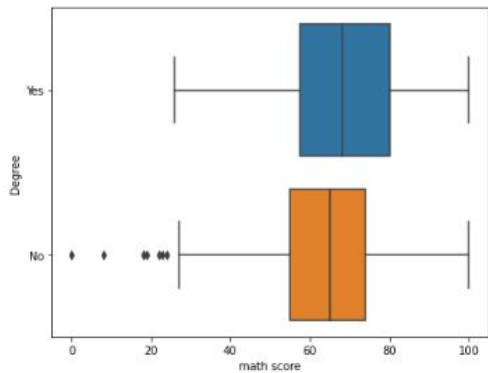
Done By: Yao Xian(Robin)

2

Data Analysis



Data Analysis



Data Analysis (Random Forest Classification)

Train Set

TPR Train : 0.987460815047022
TNR Train : 0.9957537154989384

FPR Train : 0.004246284501061571
FNR Train : 0.012539184952978056

Test Set

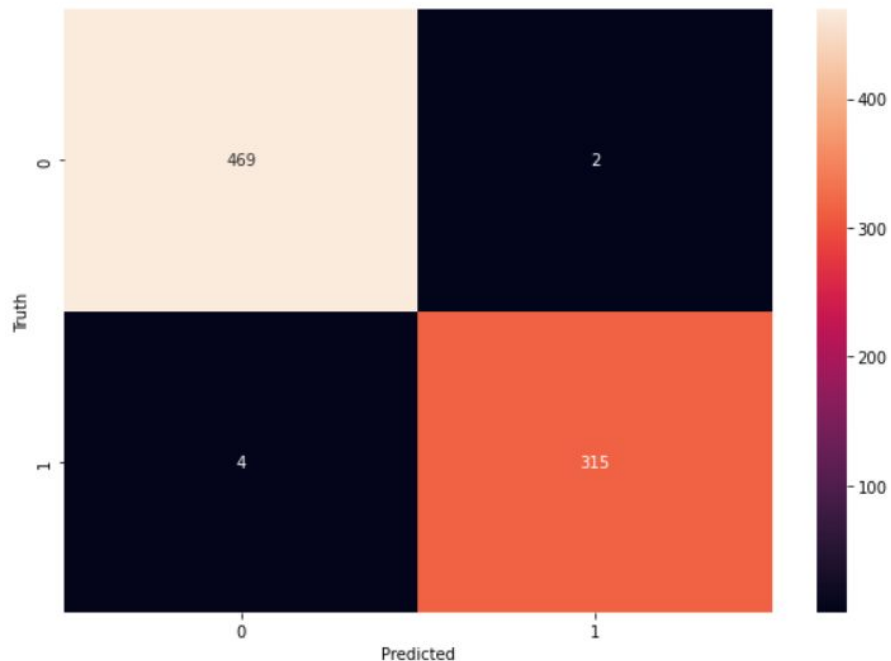
TPR Test : 0.35443037974683544
TNR Test : 0.6638655462184874

FPR Test : 0.33613445378151263
FNR Test : 0.6455696202531646

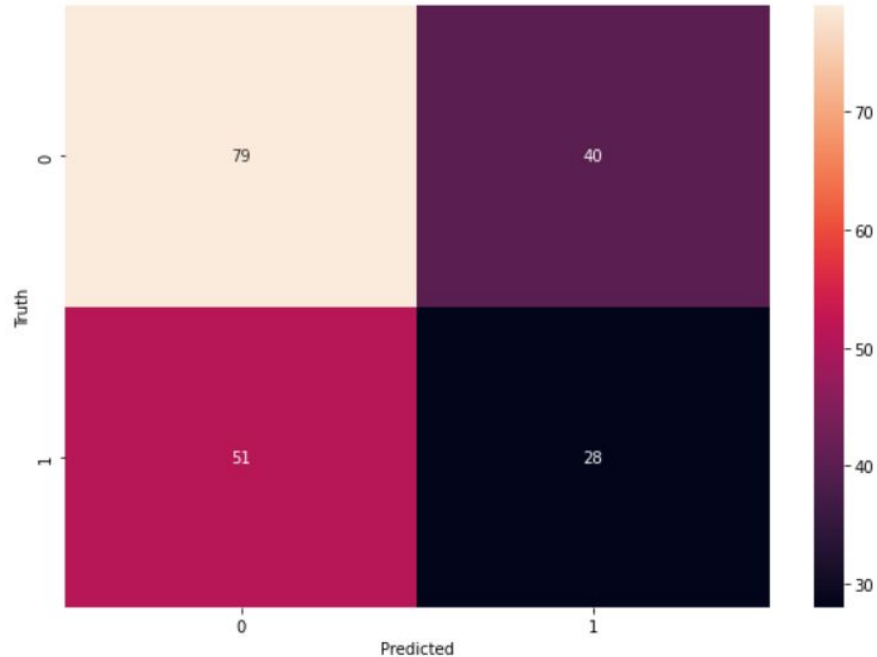


Data Analysis (Random Forest Classification)

Train Set



Test Set



Data Analysis (Naive Bayes)

Train Set

TPR Train : 0.3605015673981191
TNR Train : 0.7579617834394905

FPR Train : 0.24203821656050956
FNR Train : 0.6394984326018809

Test Set

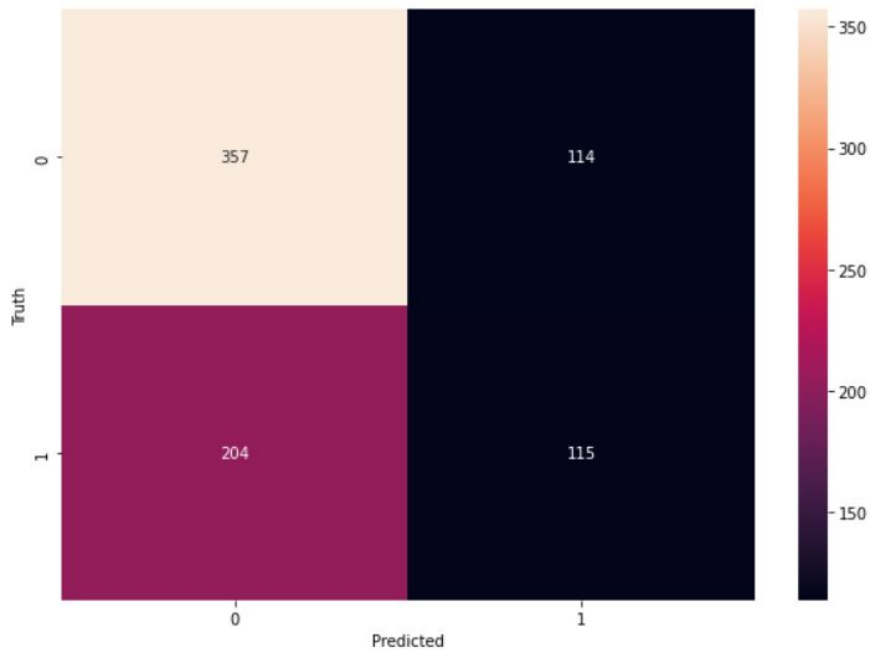
TPR Test : 0.4050632911392405
TNR Test : 0.6890756302521008

FPR Test : 0.31092436974789917
FNR Test : 0.5949367088607594

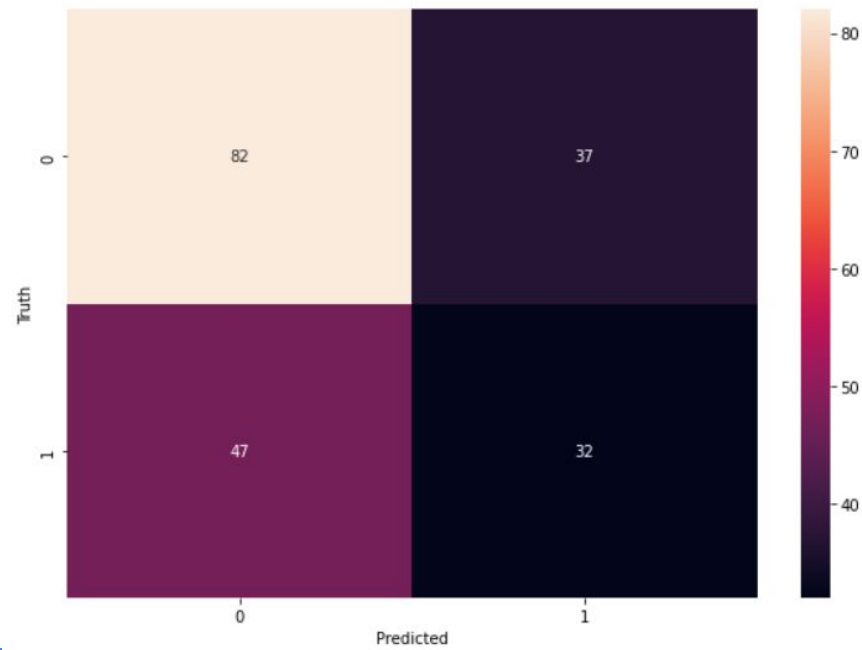


Data Analysis (Naive Bayes)

Train Set



Test Set



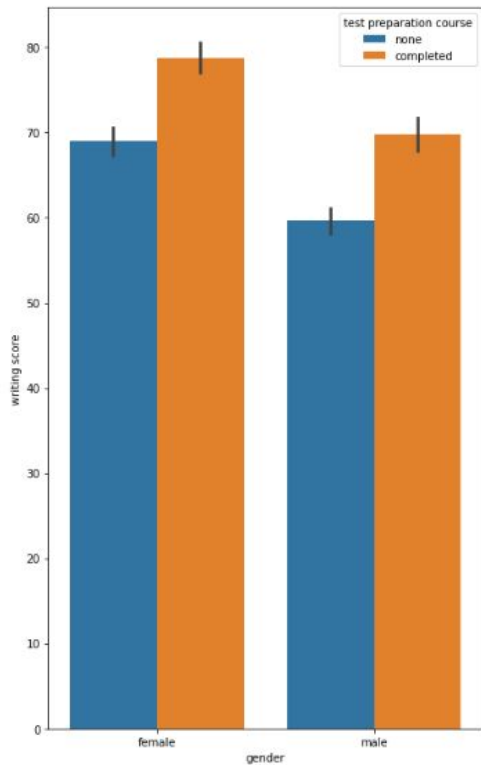
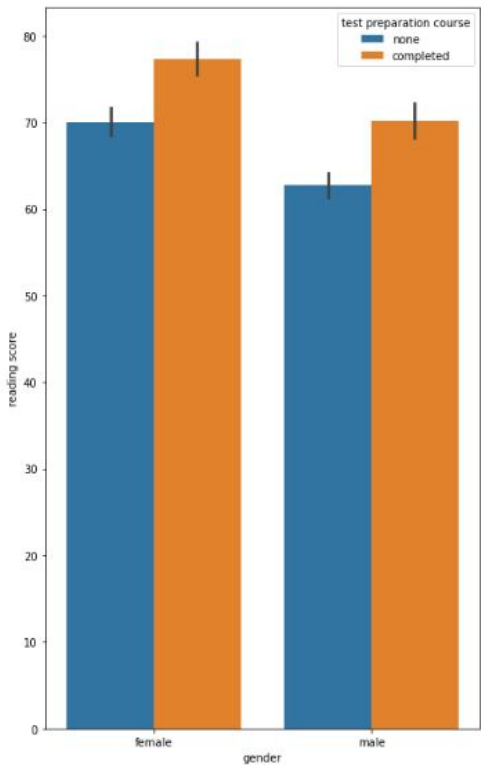
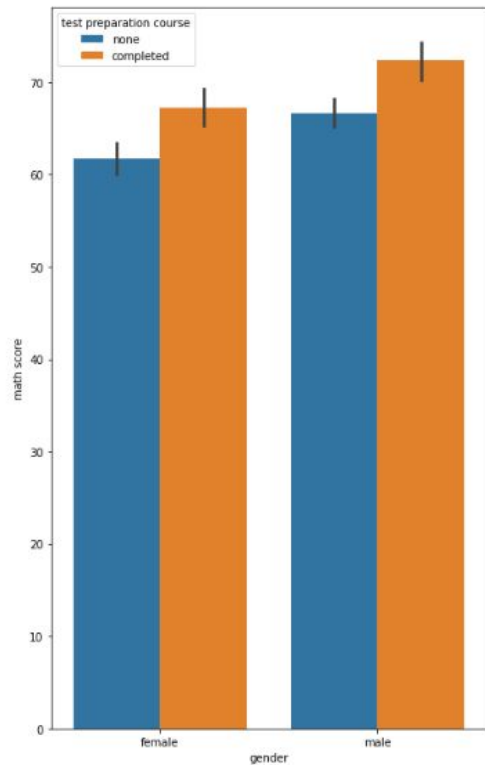


Does Test Prep Course really affect student results?

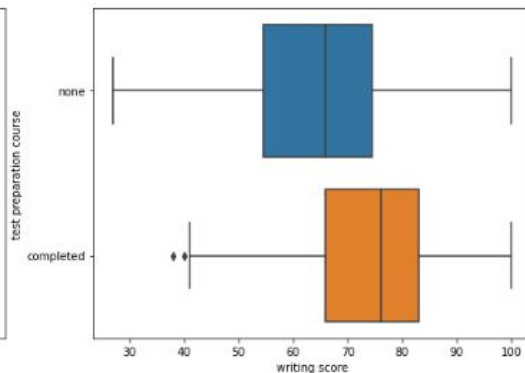
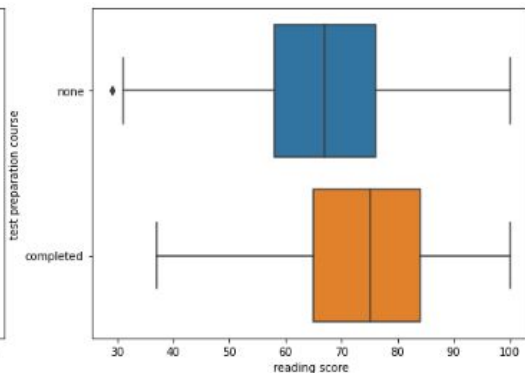
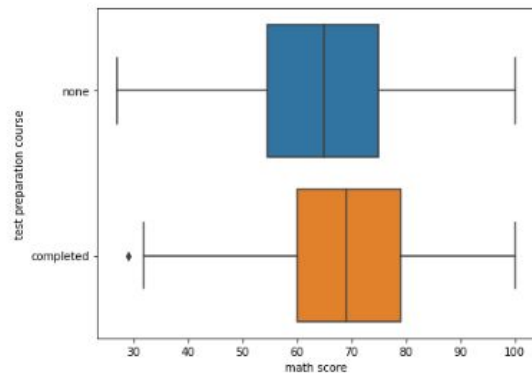
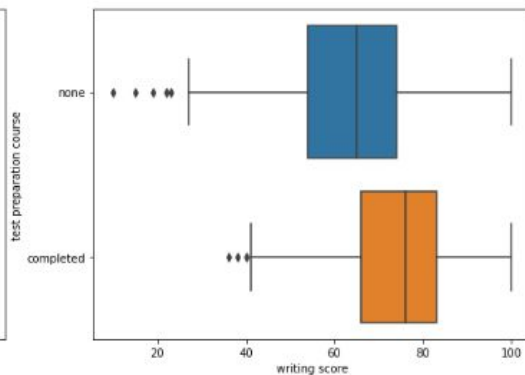
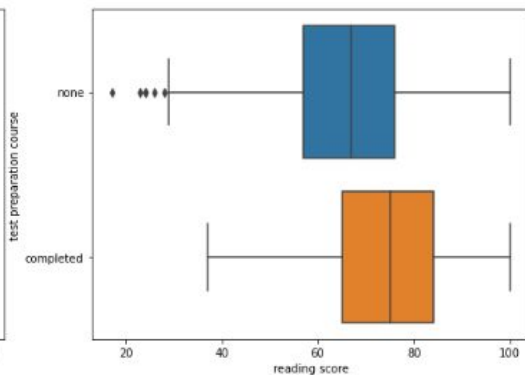
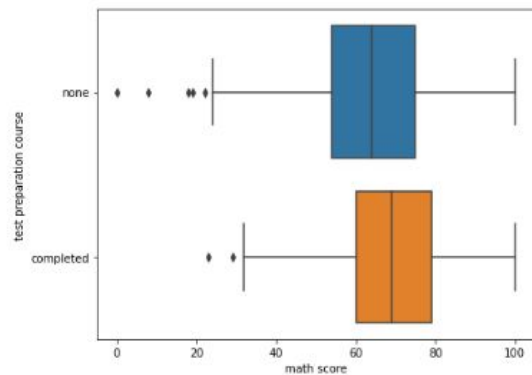
Done By: Yao Xian(Robin)

3

Data Analysis



Data Analysis



Data Analysis (Random Forest Classification)

Train Set

TPR Train : 0.996031746031746
TNR Train : 0.9790209790209791

FPR Train : 0.02097902097902098
FNR Train : 0.003968253968253968

Test Set

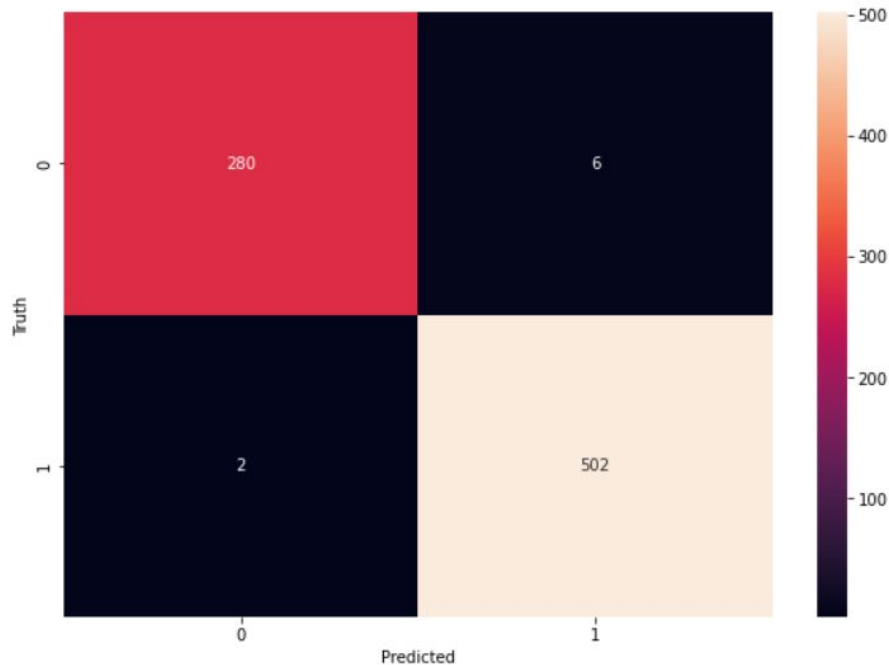
TPR Test : 0.7244094488188977
TNR Test : 0.4788732394366197

FPR Test : 0.5211267605633803
FNR Test : 0.2755905511811024

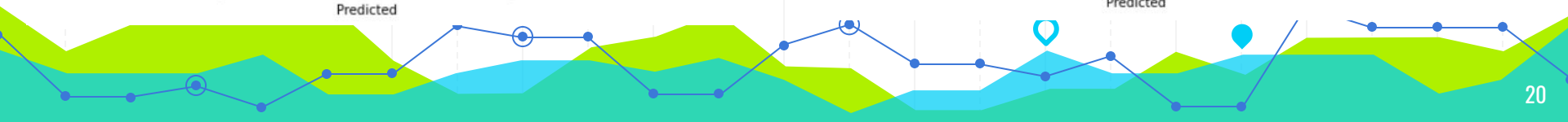
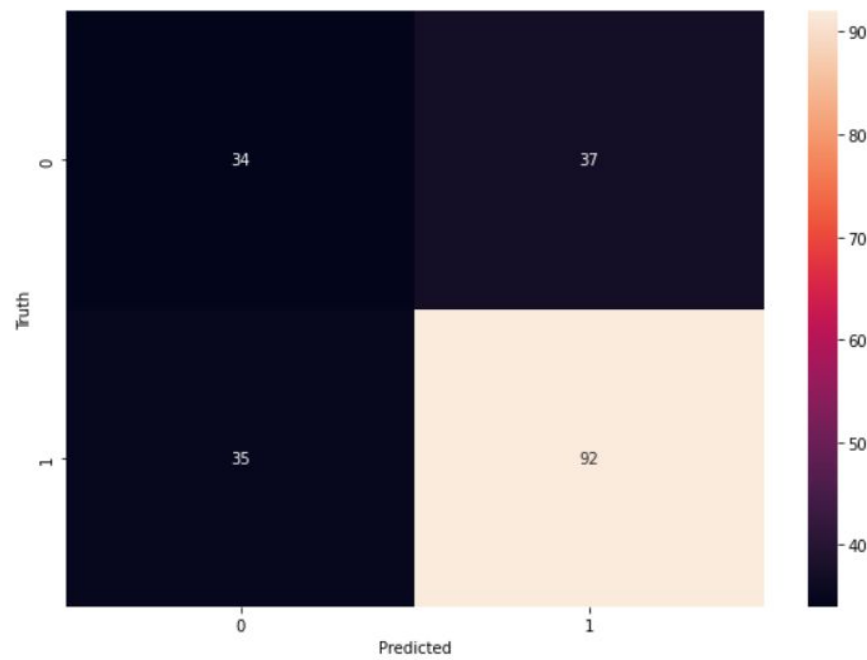


Data Analysis (Random Forest Classification)

Train Set



Test Set



Data Analysis (Support Vector Machine)

	precision	recall	f1-score	support
completed	0.60	0.47	0.52	286
none	0.73	0.82	0.77	504
accuracy			0.69	790
macro avg	0.66	0.64	0.65	790
weighted avg	0.68	0.69	0.68	790

Train Set

Test Set

	precision	recall	f1-score	support
completed	0.50	0.45	0.47	71
none	0.71	0.75	0.73	127
accuracy			0.64	198
macro avg	0.60	0.60	0.60	198
weighted avg	0.63	0.64	0.64	198



Data Analysis (Naive Bayes)

Train Set

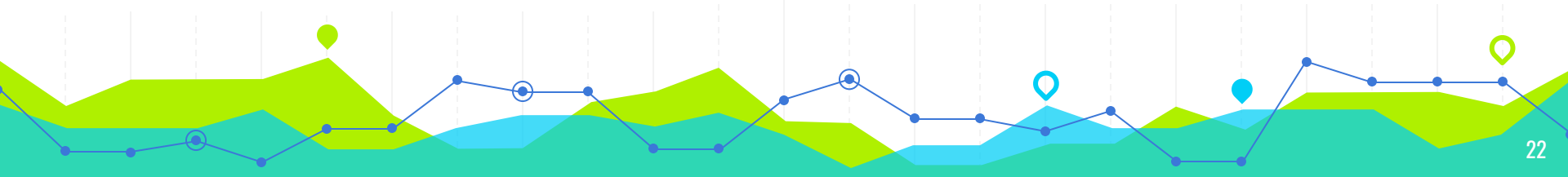
TPR Train : 0.7440476190476191
TNR Train : 0.527972027972028

FPR Train : 0.47202797202797203
FNR Train : 0.25595238095238093

Test Set

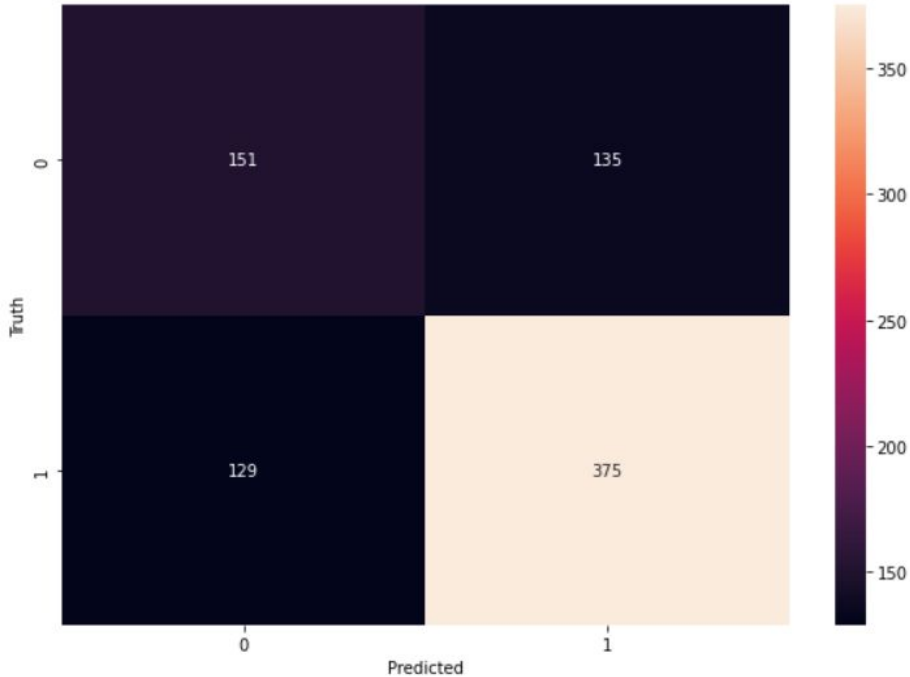
TPR Test : 0.6062992125984252
TNR Test : 0.4647887323943662

FPR Test : 0.5352112676056338
FNR Test : 0.3937007874015748

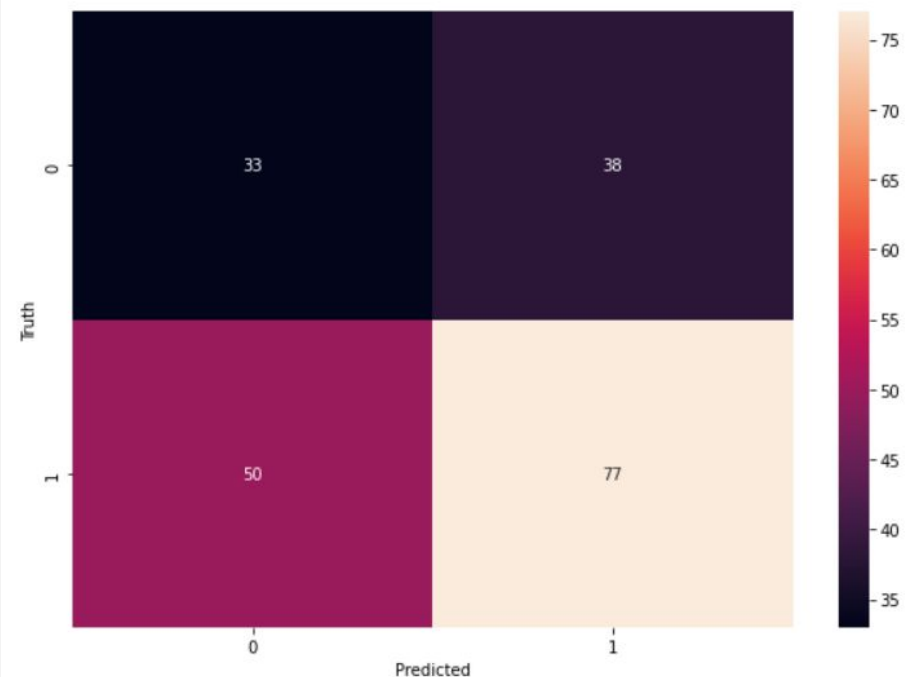


Data Analysis (Naive Bayes)

Train Set



Test Set



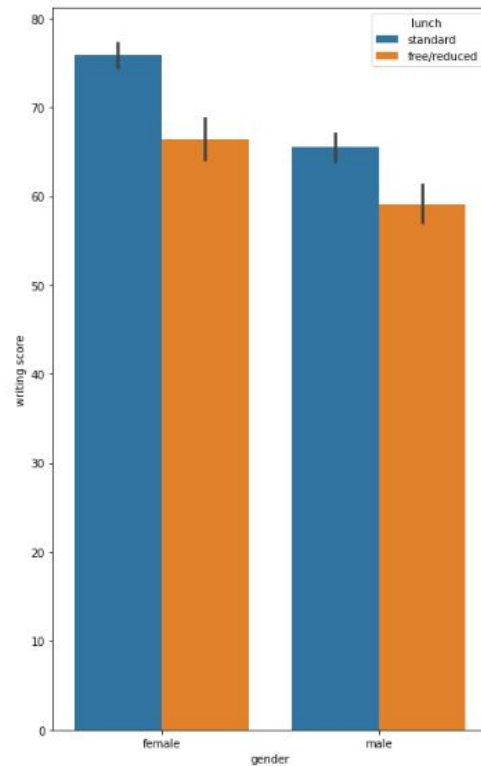
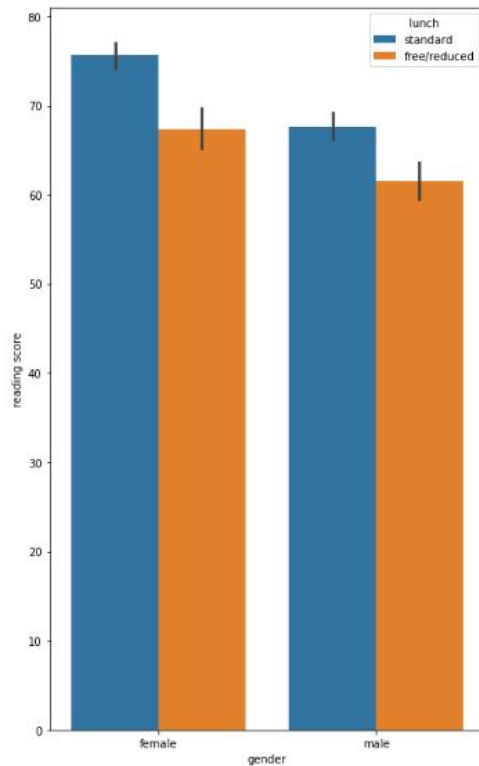
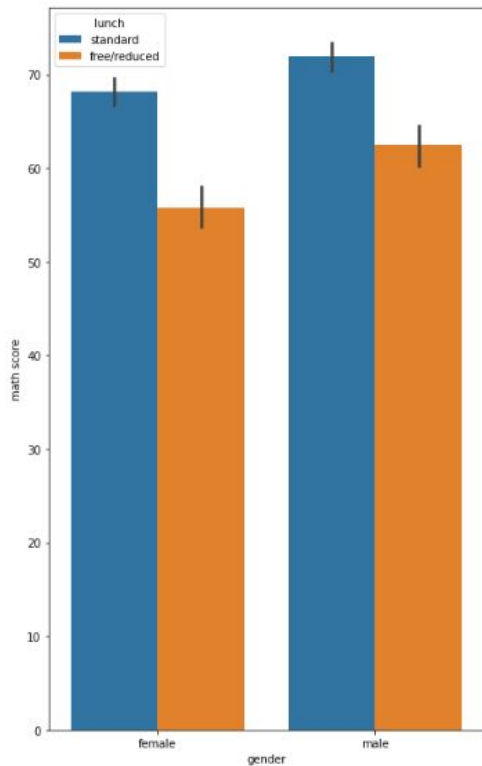


Does having a proper lunch affects performance?

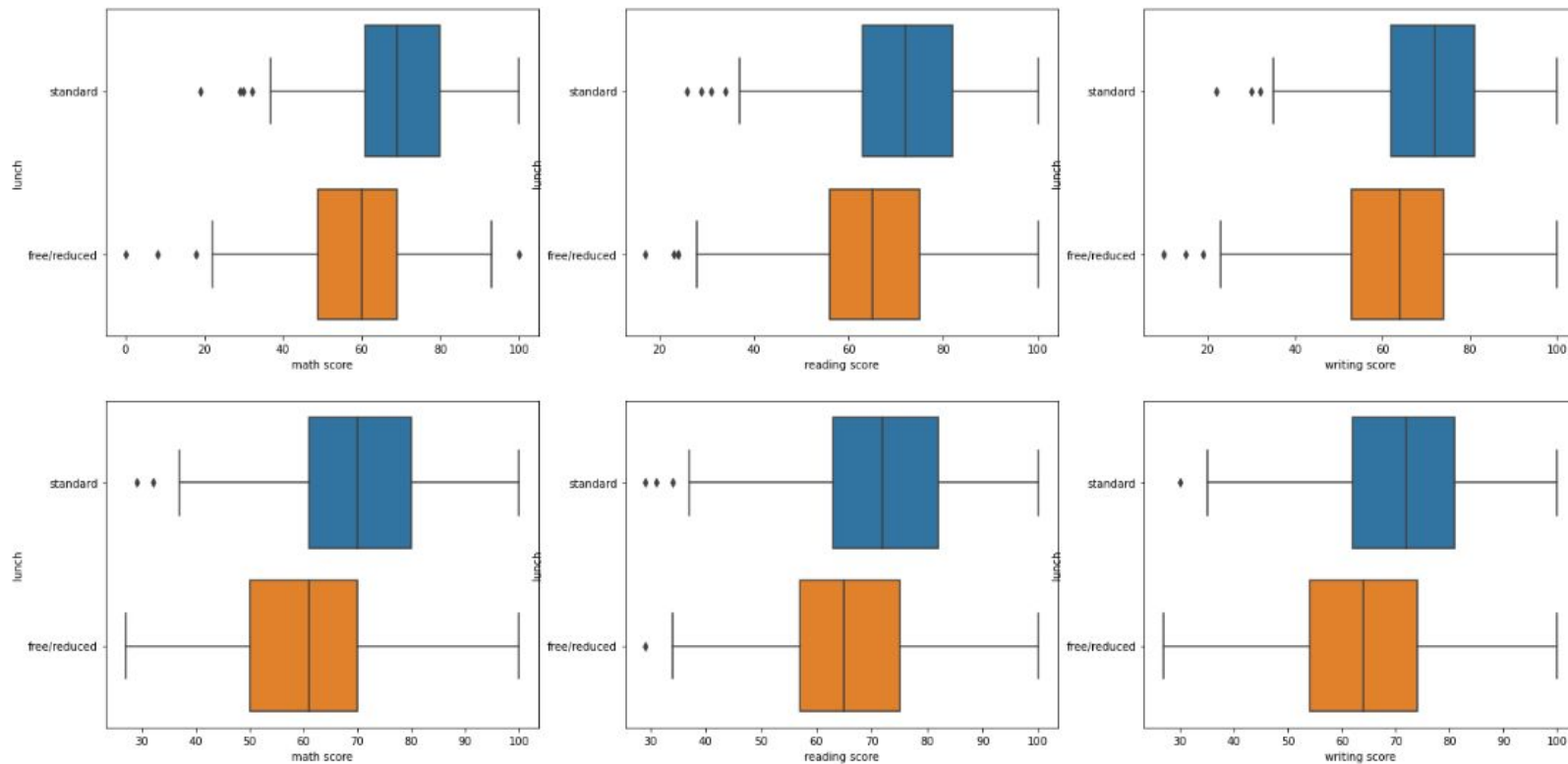
Done By: Yao Xian(Robin)

4

Data Analysis



Data Analysis



Data Analysis (Random Forest Classification)

Train Set

TPR Train : 0.9940119760479041
TNR Train : 0.9792387543252595

FPR Train : 0.020761245674740483
FNR Train : 0.005988023952095809

Test Set

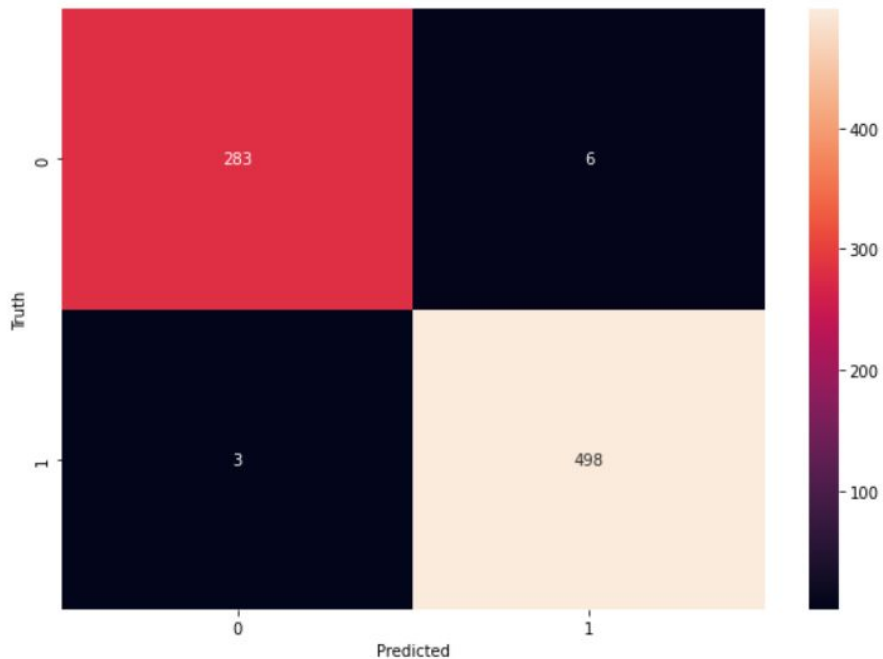
TPR Test : 0.7816901408450704
TNR Test : 0.30357142857142855

FPR Test : 0.6964285714285714
FNR Test : 0.21830985915492956

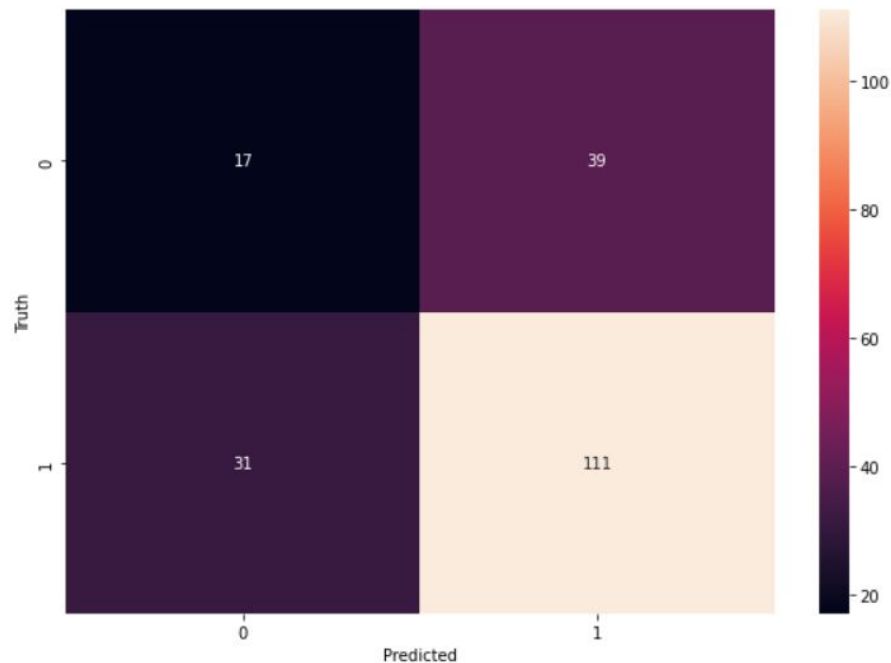


Data Analysis (Random Forest Classification)

Train Set



Test Set



Data Analysis (Support Vector Machine)

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

free/reduced	0.65	0.37	0.47	289
standard	0.71	0.88	0.79	501

accuracy			0.70	790
macro avg	0.68	0.63	0.63	790
weighted avg	0.69	0.70	0.67	790

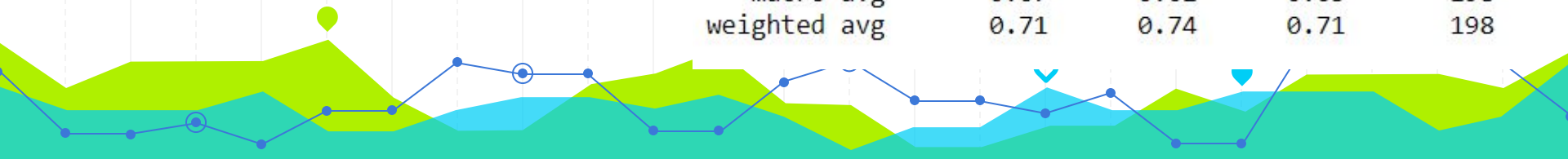
Test Set

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

free/reduced	0.56	0.34	0.42	56
standard	0.77	0.89	0.83	142

accuracy			0.74	198
macro avg	0.67	0.62	0.63	198
weighted avg	0.71	0.74	0.71	198

Train Set



Data Analysis (Naive Bayes)

Train Set

TPR Train : 0.7325349301397206
TNR Train : 0.5121107266435986

FPR Train : 0.48788927335640137
FNR Train : 0.26746506986027946

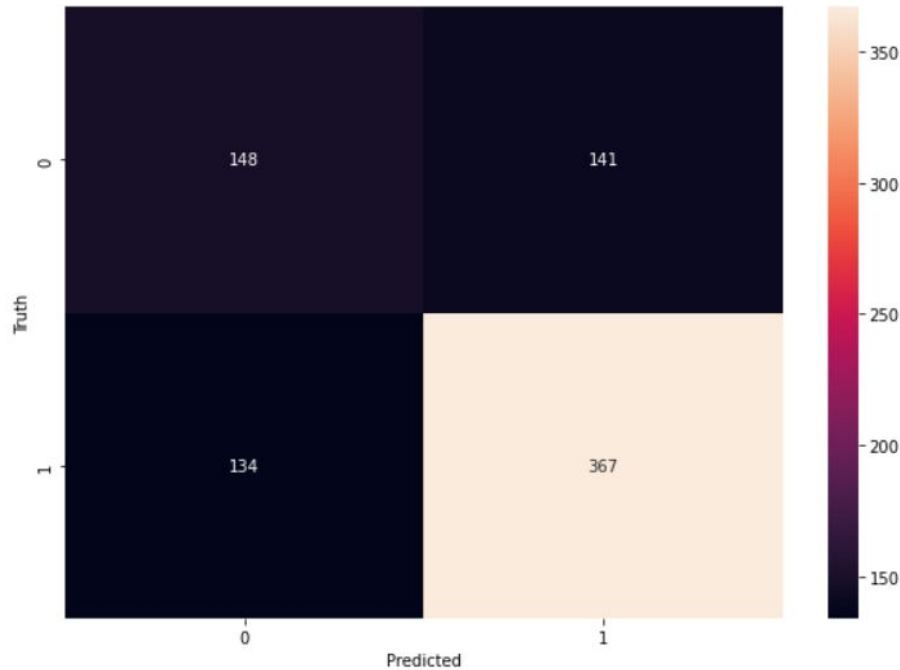
Test Set

TPR Test : 0.7887323943661971
TNR Test : 0.4642857142857143

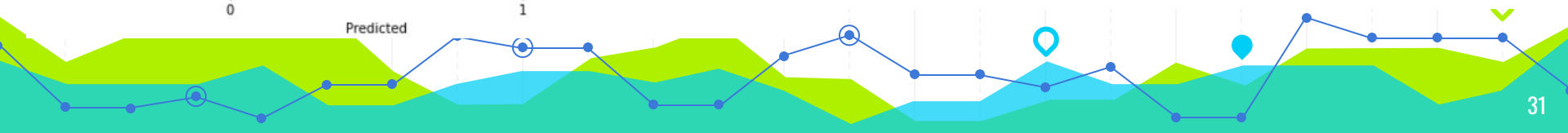
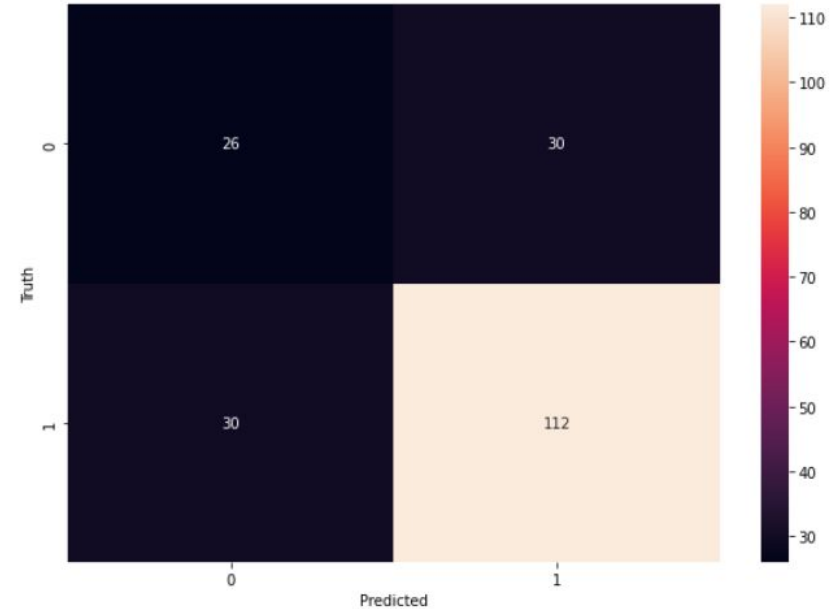
FPR Test : 0.5357142857142857
FNR Test : 0.2112676056338028

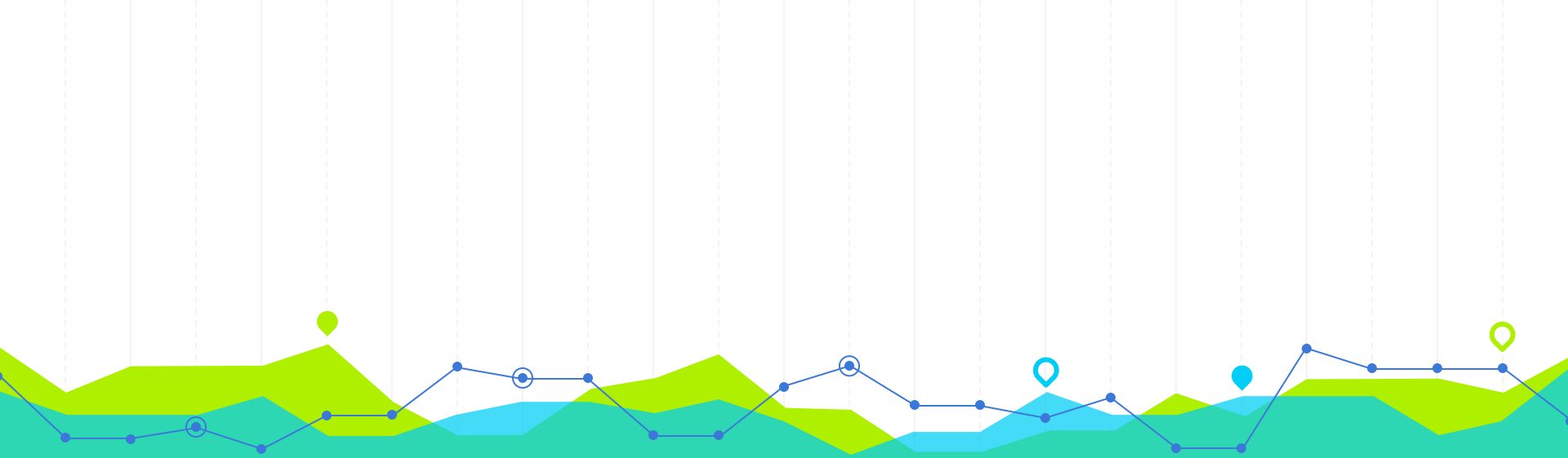
Data Analysis (Naive Bayes)

Train Set



Test Set



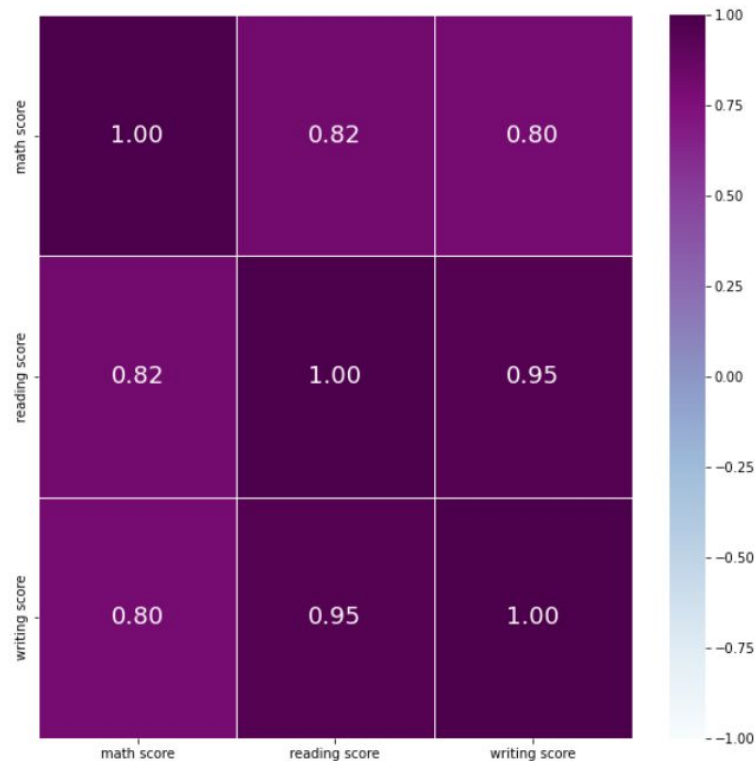


**Does doing good in one subject mean
they are good in another?**

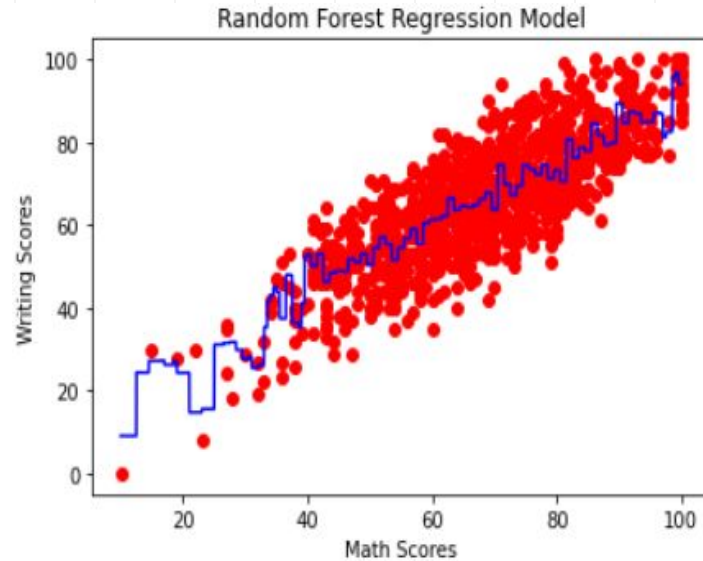
Done by: John

5

Correlation of the different scores

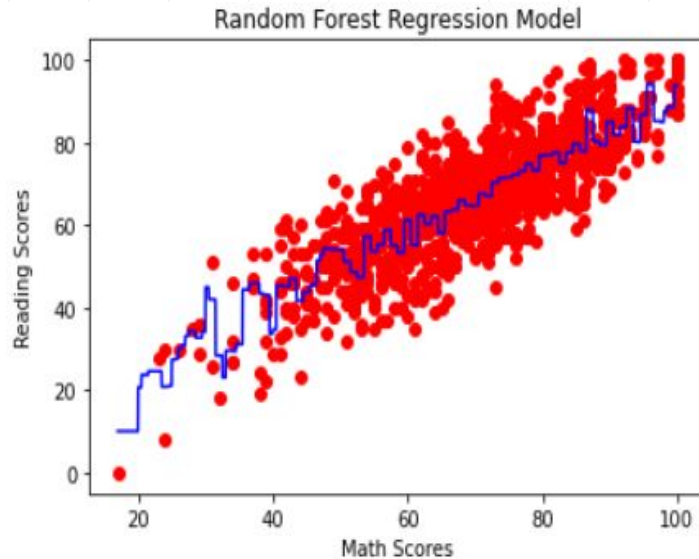


Using Writing score VS predict Math score (Random Forest Regression)



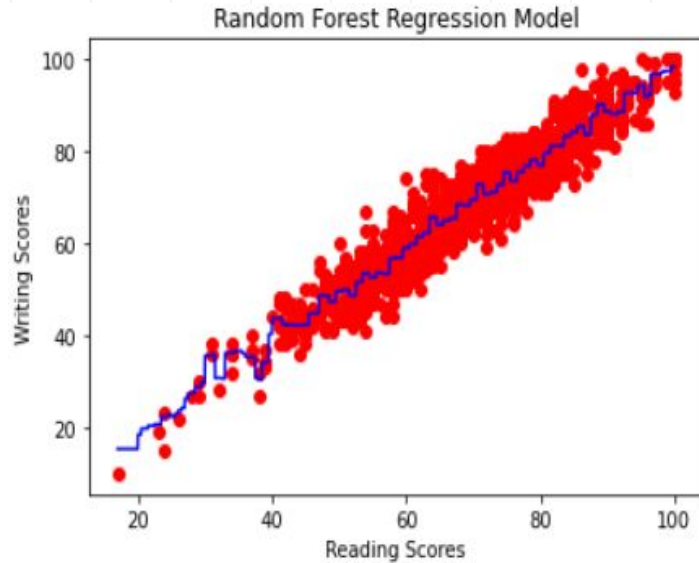
MAE	7.29
RMSE	8.92
R^2	0.69

Using Reading score VS predict Math score (Random Forest Regression)

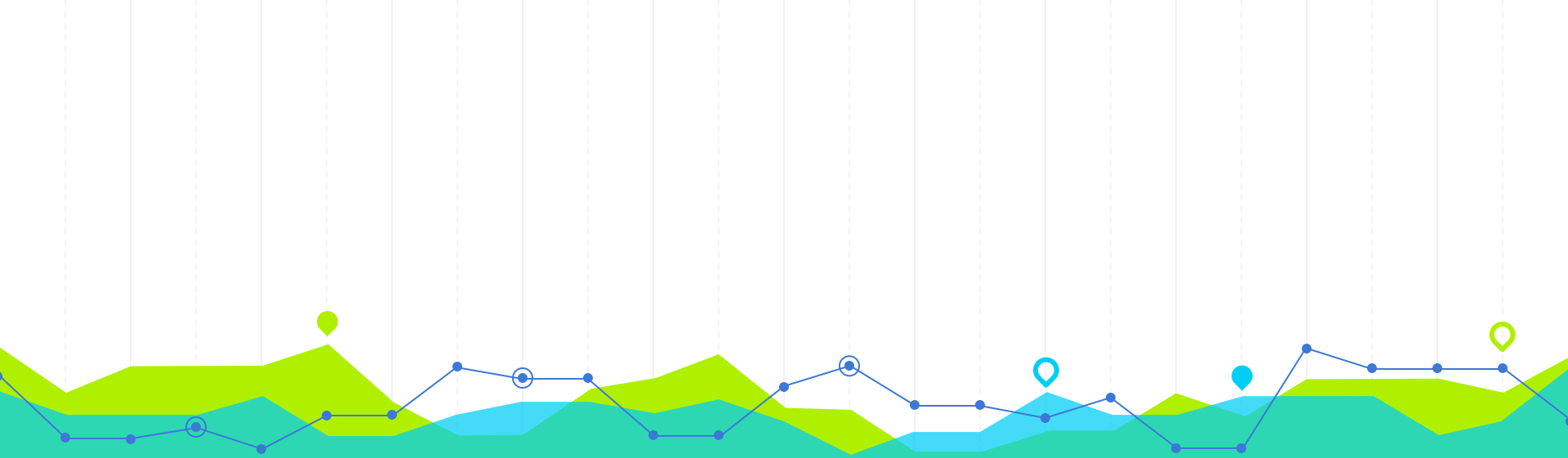


MAE	7.24
RMSE	9.09
R^2	0.65

Using Writing score VS predict Reading score (Random Forest Regression)



MAE	3.63
RMSE	4.54
R^2	0.91



Conclusion

Done by: John

6

CONCLUSION

Parents Holding Degree



Test Preparation Course



=

A+

Proper Lunch



=

A+

THANK YOU!

