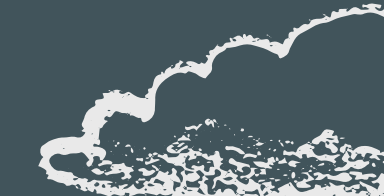
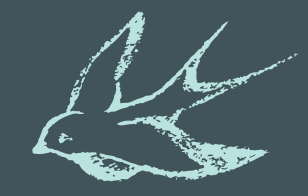








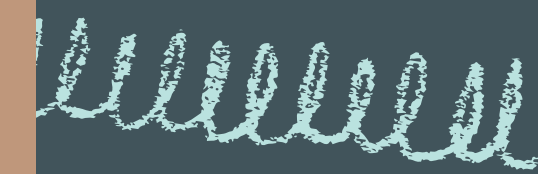

IE0005 Mini-Project

# Student Performance Prediction



Presented by:

Muhammad Ameerul Bin Azman (U2122090J),  
Heng Zi Hui (U2122657B),  
Peng Teng Kang (U2123750K),  
Guerta Uno Gabriel Yap (U2121784F)



What affects a student's grade?

Dataset taken from Kaggle

# Our Project



- Data approaches student achievement in secondary education of two Portuguese schools
- Data attributes include student grades, demographic, social and school related features
- Collected by using school reports and questionnaires

# Brief view of the Dataset

	id	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	grades
0	1	GP	F	18	U	GT3	A	4	4	at_home	...	3	4	1	1	3	6	5	6	6	D
1	2	GP	F	17	U	GT3	T	1	1	at_home	...	3	3	1	1	3	4	5	5	6	D
2	3	GP	F	15	U	LE3	T	1	1	at_home	...	3	2	2	3	3	10	7	8	10	C
3	4	GP	F	15	U	GT3	T	4	2	health	...	2	2	1	1	5	2	15	14	15	B
4	5	GP	F	16	U	GT3	T	3	3	other	...	3	2	1	2	5	4	6	10	10	C

5 rows × 35 columns

# Content



## Visualization

Standard exploration and statistical visualization of the data



## DS/ML

Usage of tools and techniques learnt from IE0005 labs



## Preparation of dataset

Cleaning, resizing/reshaping the dataset, removing outliers, balancing imbalanced classes and grouping rows/columns



## Learning something new

Using new DS/ML beyond what was covered in IE0005 labs

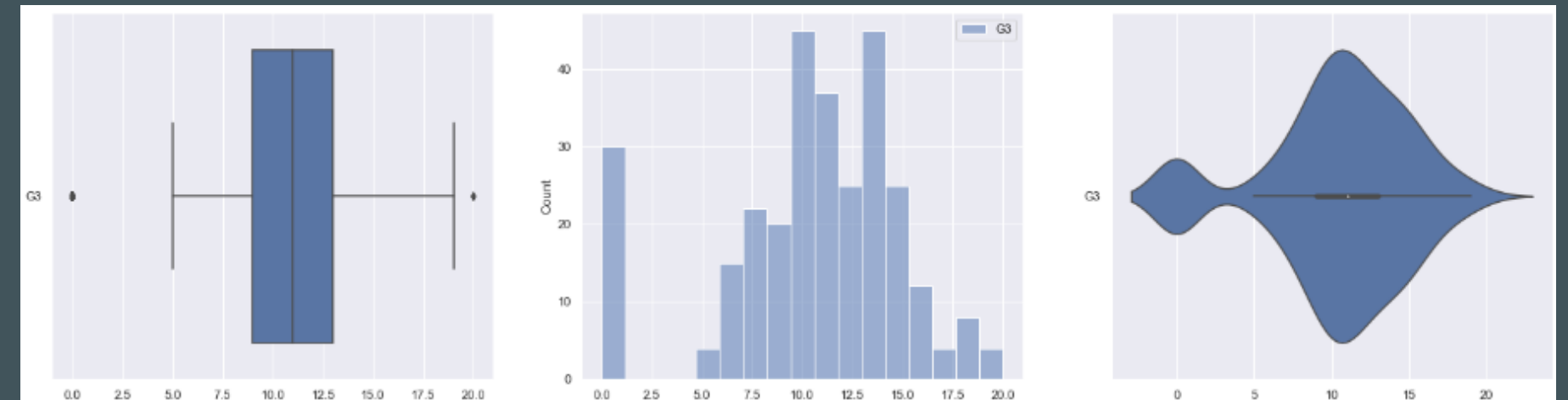


# Visualisation

## Visualisation of data

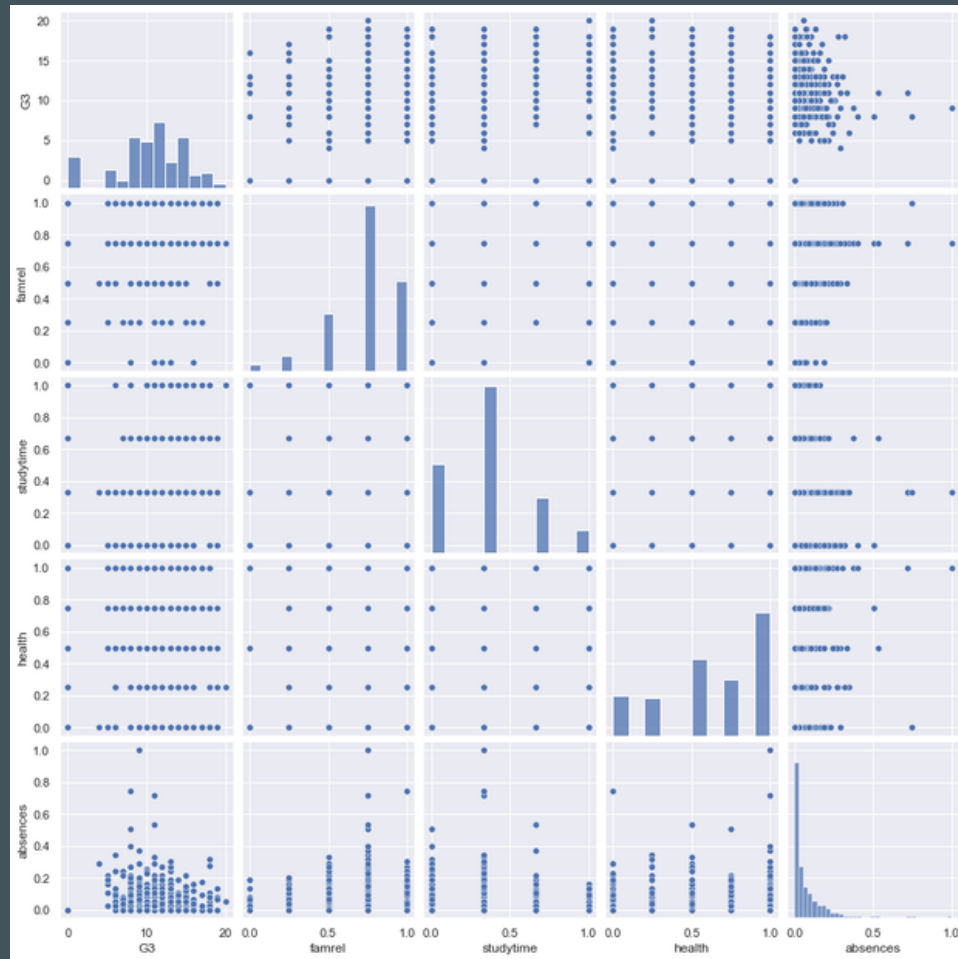
To determine the Final Grade (G3) that students will receive, we have used standard forms of visualisation of the data as taught in the course to determine the most important factors among study time, family relationship, health and absences

Some of the Visualisations used were Box Plot, Heat Map, Histogram, Violin plot and Joint Plot





# Visualisation



By using pairplots, we have compared the different variables using different plots, however what can be found from these plots is that there is a low correlation among the variables



Similarly with subplots, not much can be concluded, however it can precisely show the specifics of the variables such as how well the class did for G3 and how much the students studied for the test

# Visualisation



*Heat Maps on the other hand give a more accurate prediction due to it providing a numerical value to the correlation*



# Visualisation



*Of the chosen variables, studytime is the variable with the highest correlation based on the value of the heatmap, however this is not enough to predict the final score G3*

# Preparation of dataset

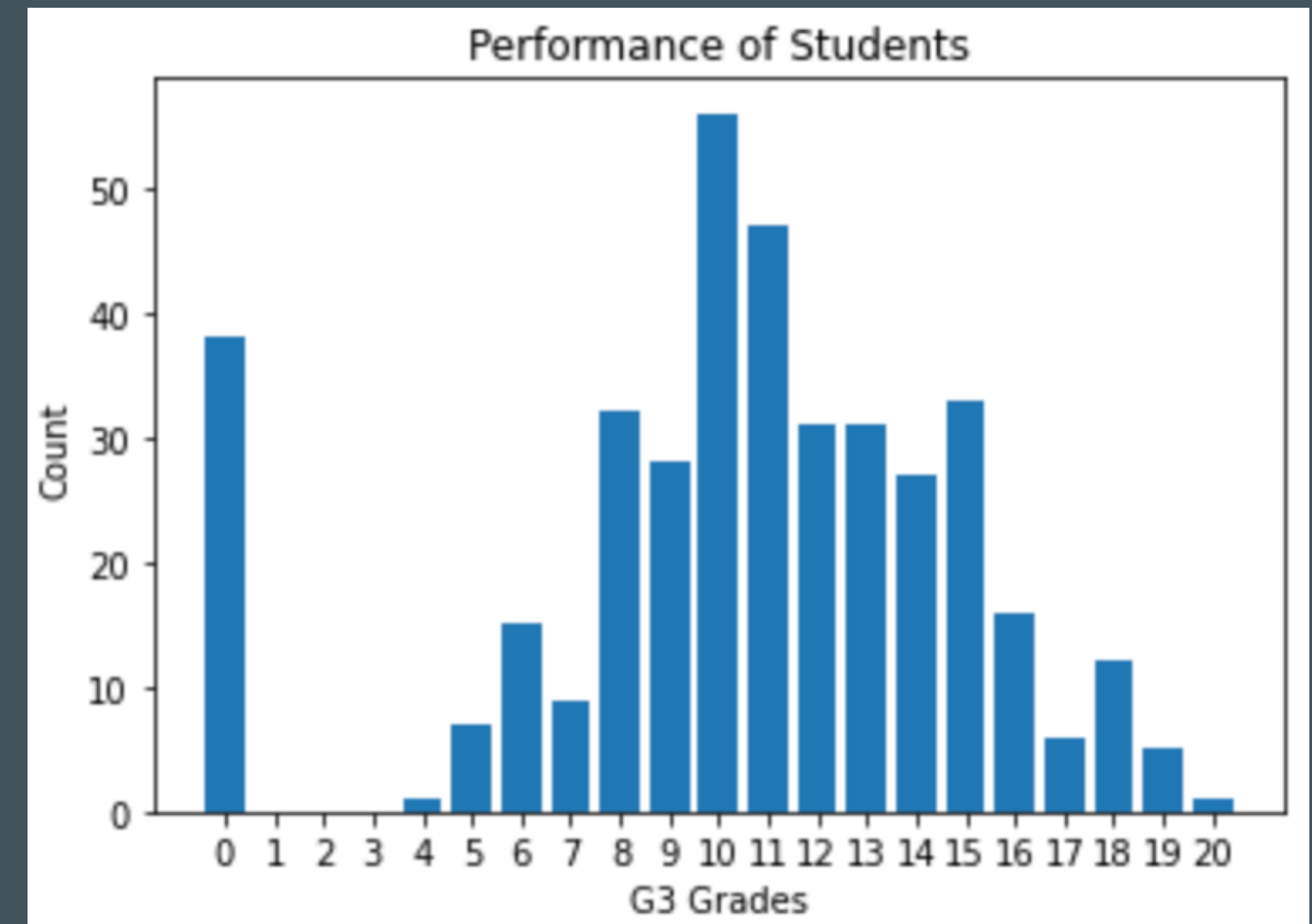
```
['GP' 'MS']  
['F' 'M']  
['U' 'R']  
['GT3' 'LE3']  
['A' 'T']  
['at_home' 'health' 'other' 'services' 'teacher']  
['teacher' 'other' 'services' 'health' 'at_home']  
['course' 'other' 'home' 'reputation']  
['mother' 'father' 'other']  
['yes' 'no']  
['no' 'yes']  
['no' 'yes']  
['no' 'yes']  
['yes' 'no']  
['yes' 'no']  
['no' 'yes']  
['no' 'yes']
```

One Hot Encoding  
(get\_dummies)

	school	sex	address	famsize	Pstatus	Mjob	Fjob	reason	guardian	schoolsup	famsup
0	GP	F	U	GT3	A	at_home	teacher	course	mother	yes	no
1	GP	F	U	GT3	T	at_home	other	course	father	no	yes
2	GP	F	U	LE3	T	at_home	other	other	mother	yes	no
3	GP	F	U	GT3	T	health	services	home	mother	no	yes
4	GP	F	U	GT3	T	other	other	home	father	no	yes

#One hot encoded data (Convert Object features into binary)

```
one_hot_encoded_data = pd.get_dummies(df_copy, columns = ['school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'F'])  
print(one_hot_encoded_data)
```



# Preparation of dataset

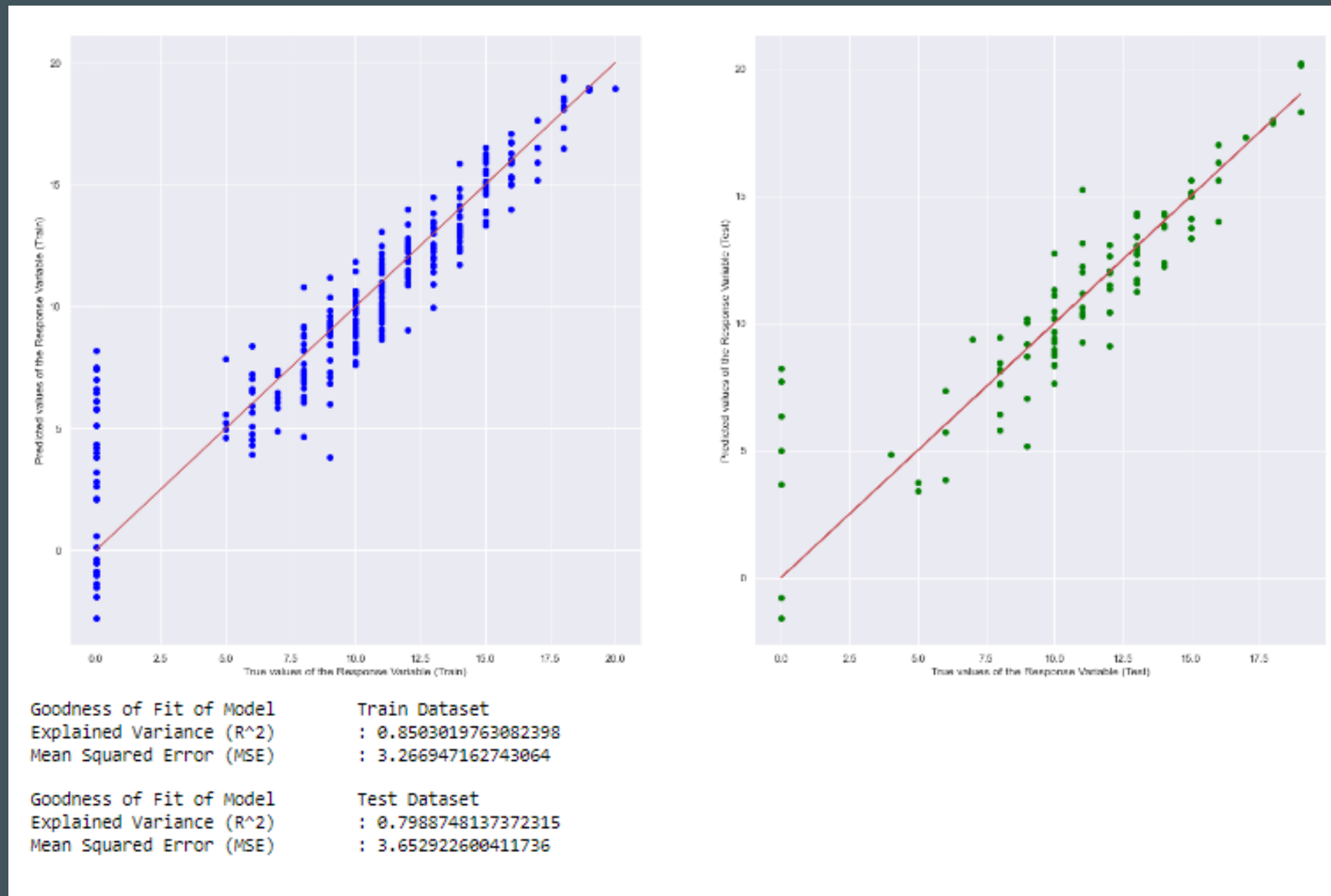
Before

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	grades
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	3	4	1	1	3	6	5	6	6	D
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	3	3	1	1	3	4	5	5	6	D
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	3	2	2	3	3	10	7	8	10	C
3	GP	F	15	U	GT3	T	4	2	health	services	...	2	2	1	1	5	2	15	14	15	B
4	GP	F	16	U	GT3	T	3	3	other	other	...	3	2	1	2	5	4	6	10	10	C

After

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	...	higher_no	higher_yes	internet_no	internet_yes	romantic_no	romantic_yes
0	18	4	4	2	2	0	4	3	4	1	...	0	1	1	0	1	0
1	17	1	1	1	2	0	5	3	3	1	...	0	1	0	1	1	0
2	15	1	1	1	2	3	4	3	2	2	...	0	1	0	1	1	0
3	15	4	2	1	3	0	3	2	2	1	...	0	1	0	1	0	0
4	16	3	3	1	2	0	4	3	2	1	...	0	1	1	0	1	0

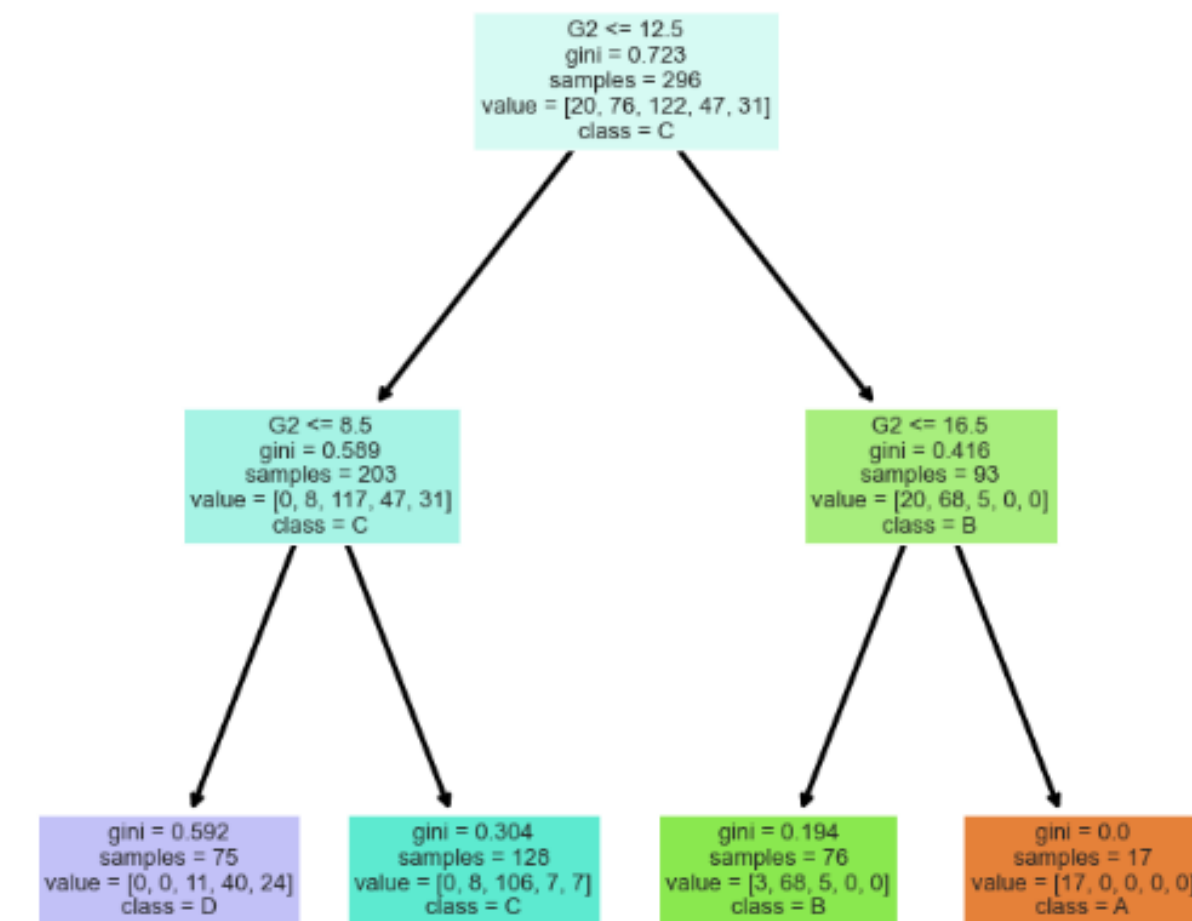
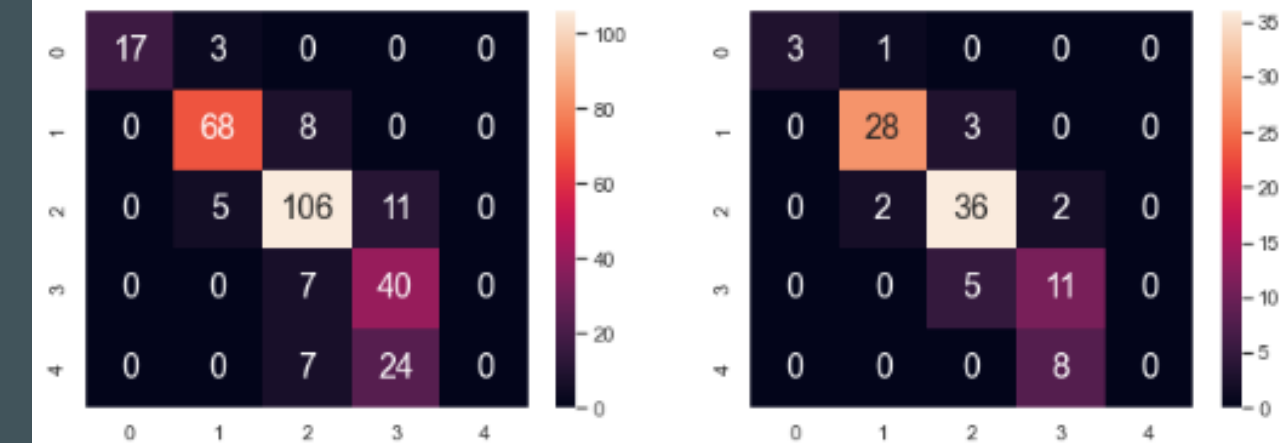
# Usage of DS/ML



Linear Regression

Goodness of Fit of Model  
Classification Accuracy : 0.7804054054054054

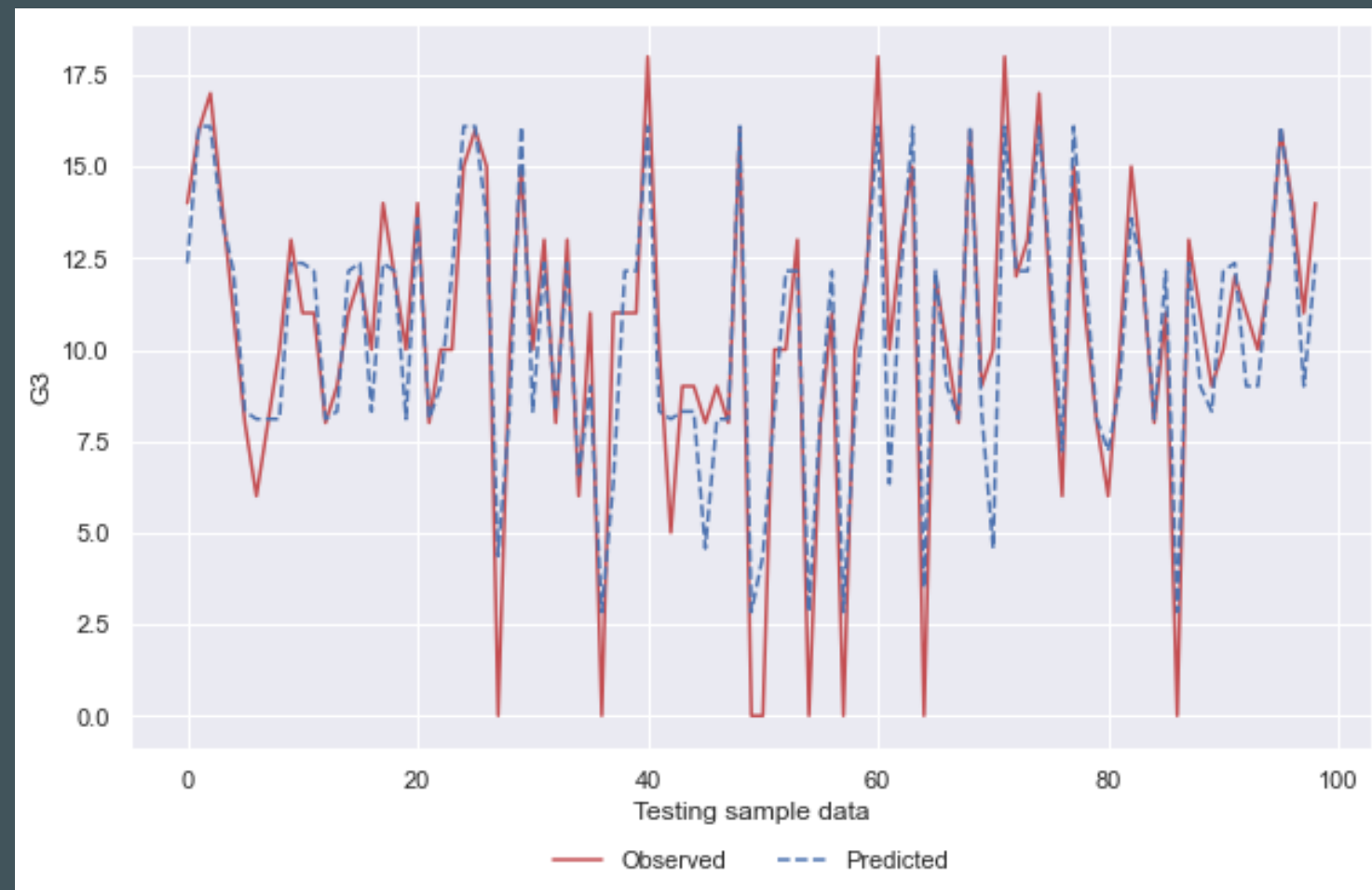
Goodness of Fit of Model  
Classification Accuracy : 0.7878787878787878



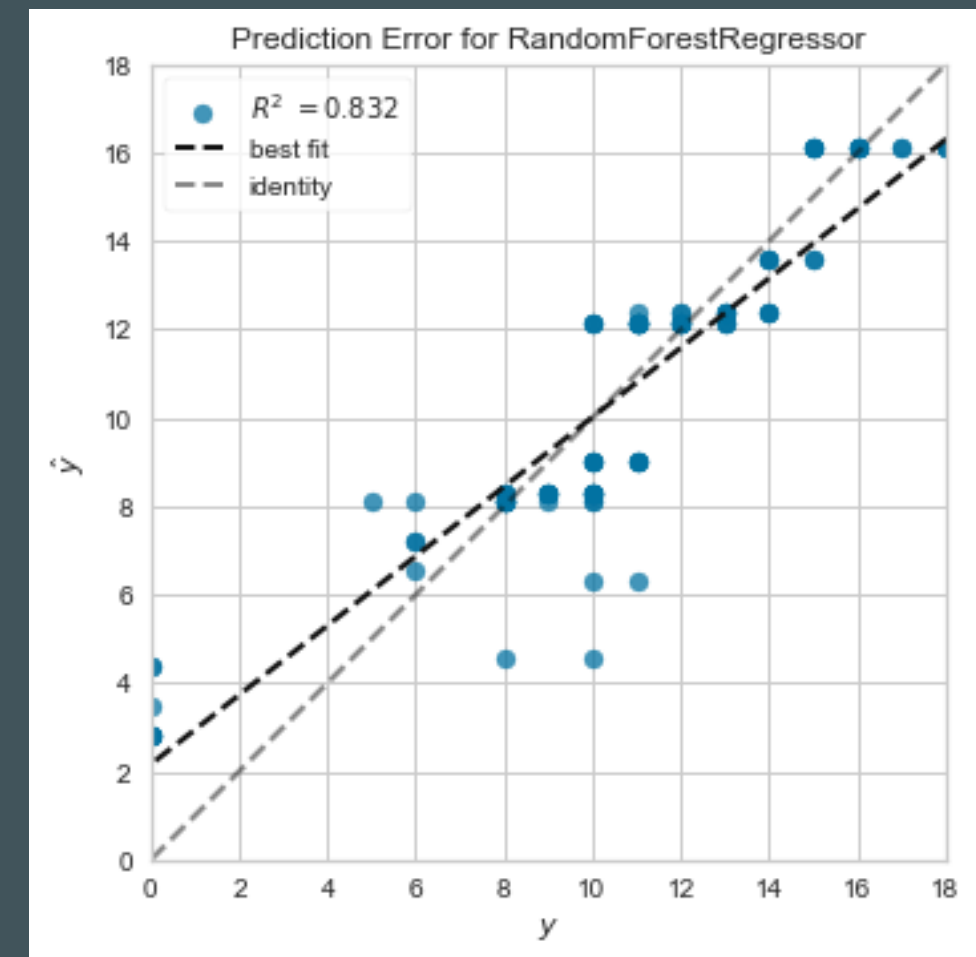
Decision Tree

# Learning something new

## 1) Random Forest regression



Testing sample data

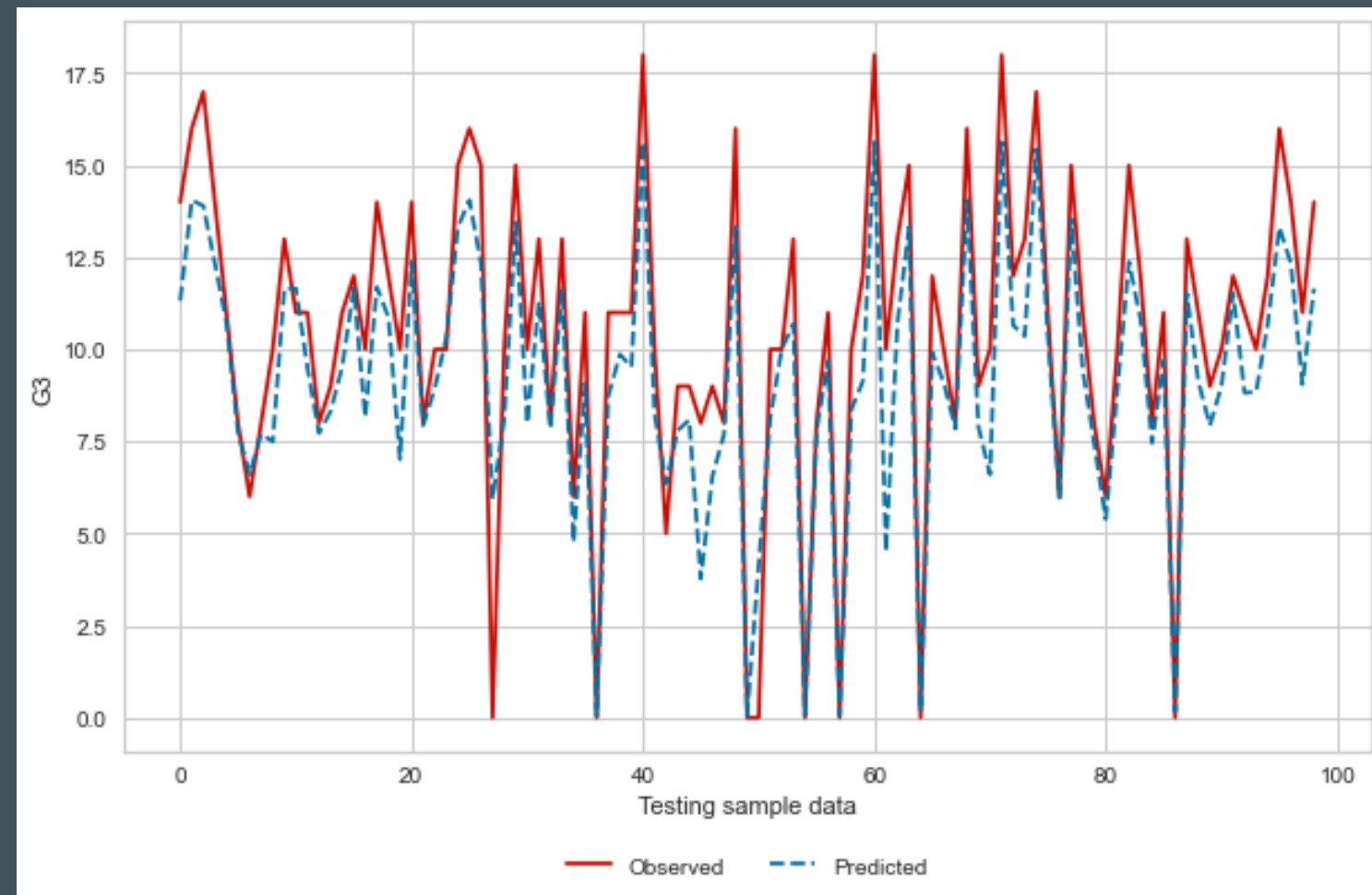


Prediction error

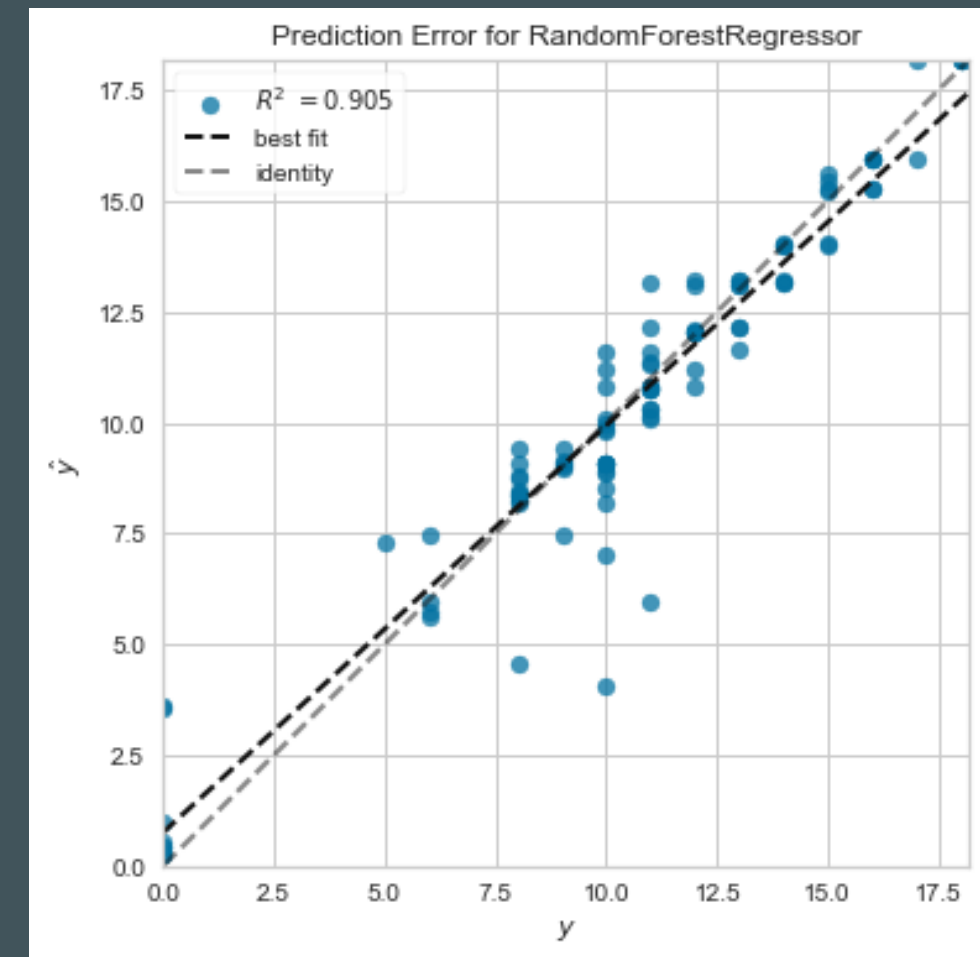




After tuning



Testing sample data



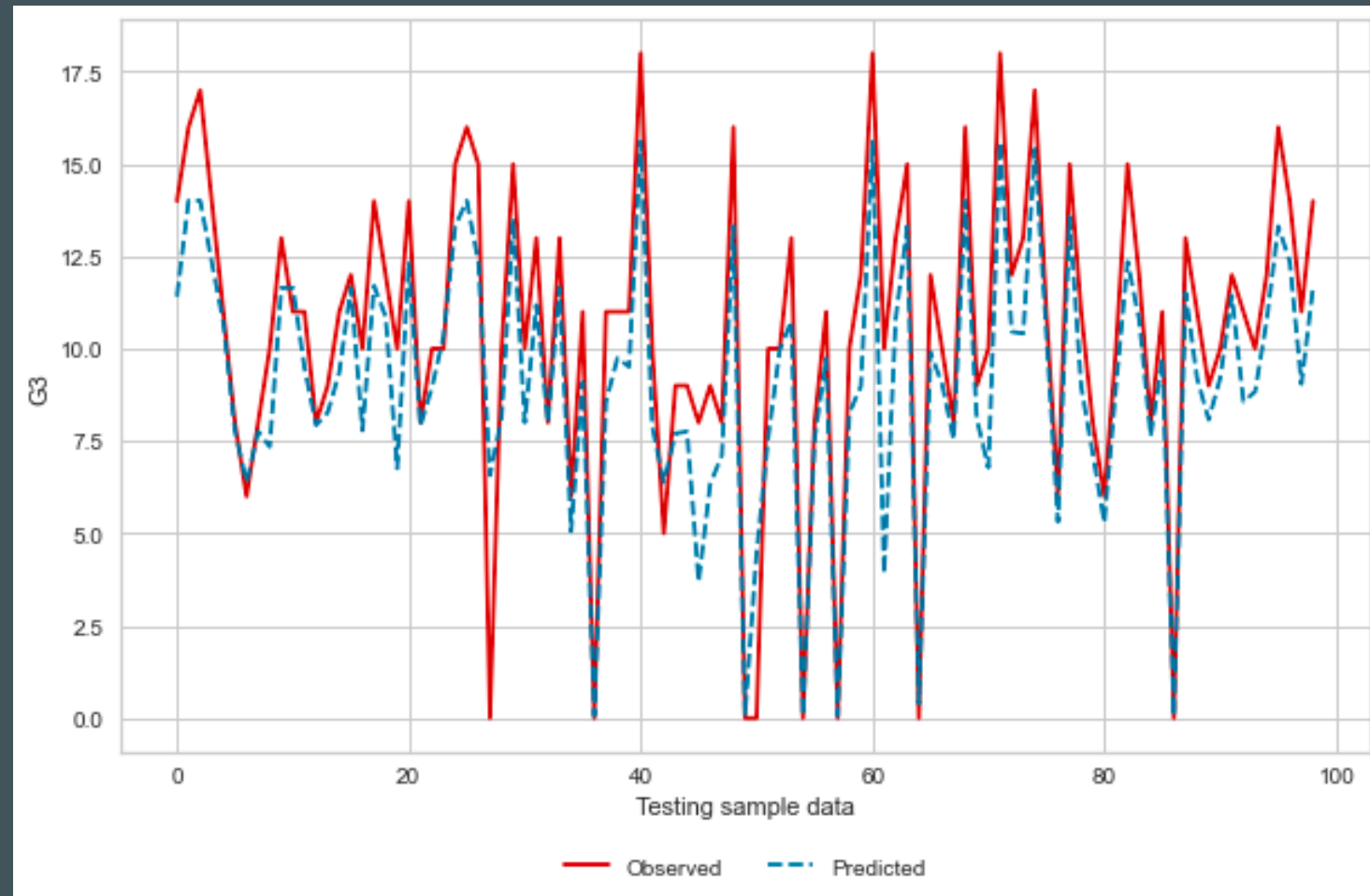
Prediction error

Conclusion: **G2** (Second Period Grade) and **Absences** are the factors that affect **G3** (Final Grade)



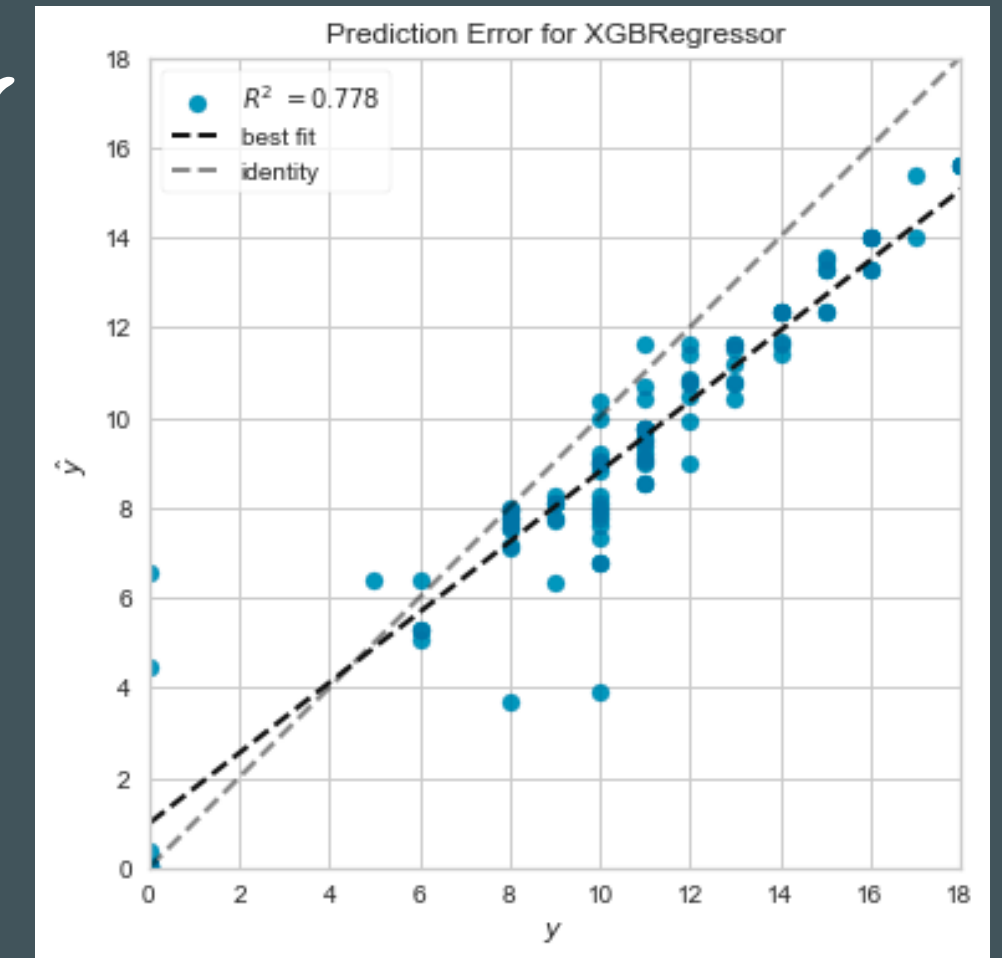
# Learning something new

## 2) Extreme Gradient Boosting Regression



Testing sample data

Prediction error



Before Tuning

\*\*\*Original Model\*\*\*

----R Square (Variance)----

Training accuracy (r\_sq) is: 0.8826155538754211

Testing accuracy (r\_sq) is: 0.7783179115786188

----Mean Squared Error----

Training mean squared error is: 2.587676181450642

Testing mean squared error is: 3.906408885956202



Learning something new

## 2) Extreme Gradient Boosting Regression

*Before Tuning*

```
***Original Model***
```

```
----R Square (Variance)----
```

```
Training accuracy (r_sq) is: 0.8826155538754211
```

```
Testing accuracy (r_sq) is: 0.7783179115786188
```

```
----Mean Squared Error----
```

```
Training mean squared error is: 2.587676181450642
```

```
Testing mean squared error is: 3.906408885956202
```

*K-Fold cross validation*

```
***Original Model***
```

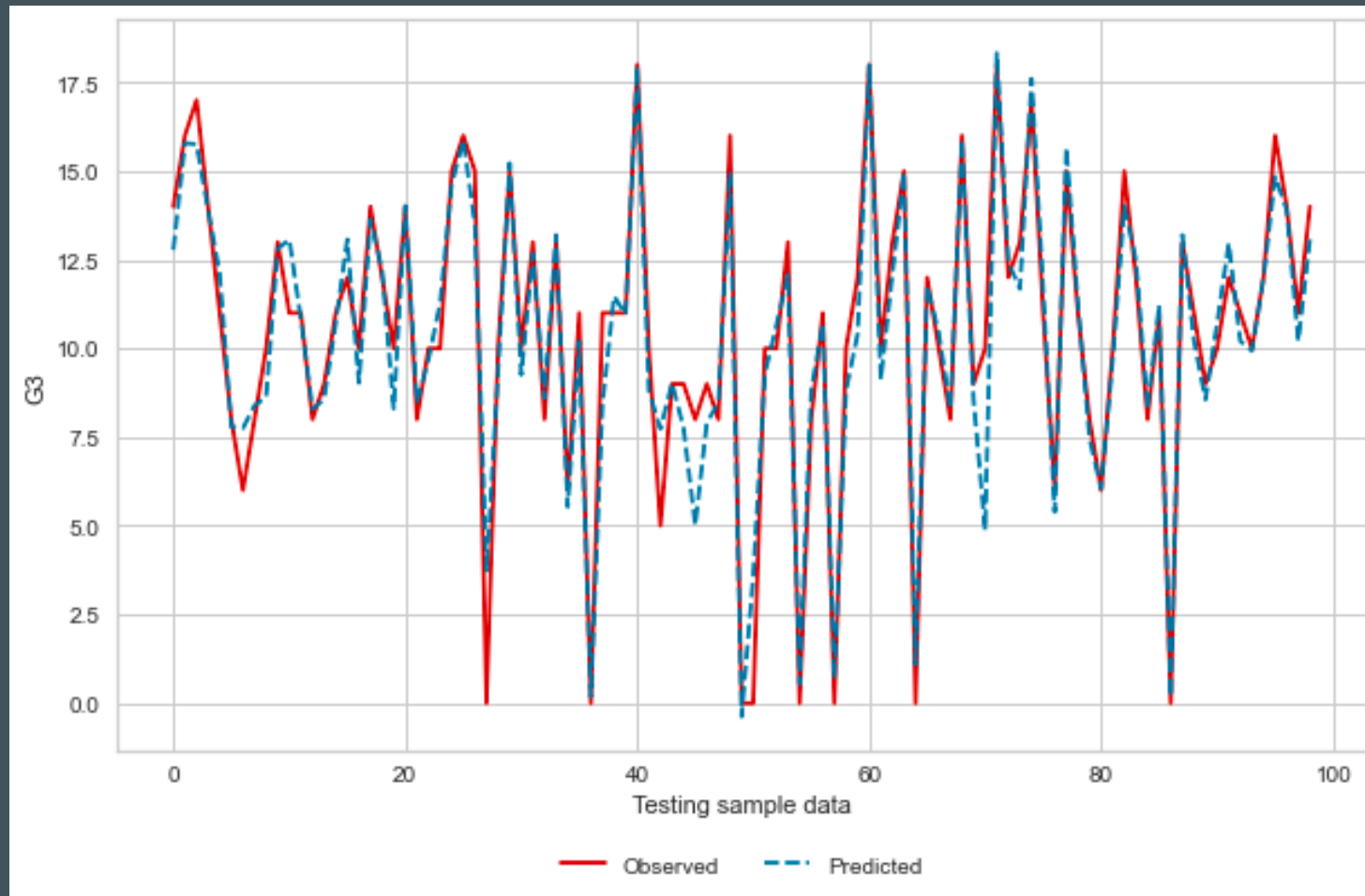
```
cross val training mean_squared_error is: 2.5614309505109425
```

```
cross val testing mean_squared_error is: 5.341055034213677
```

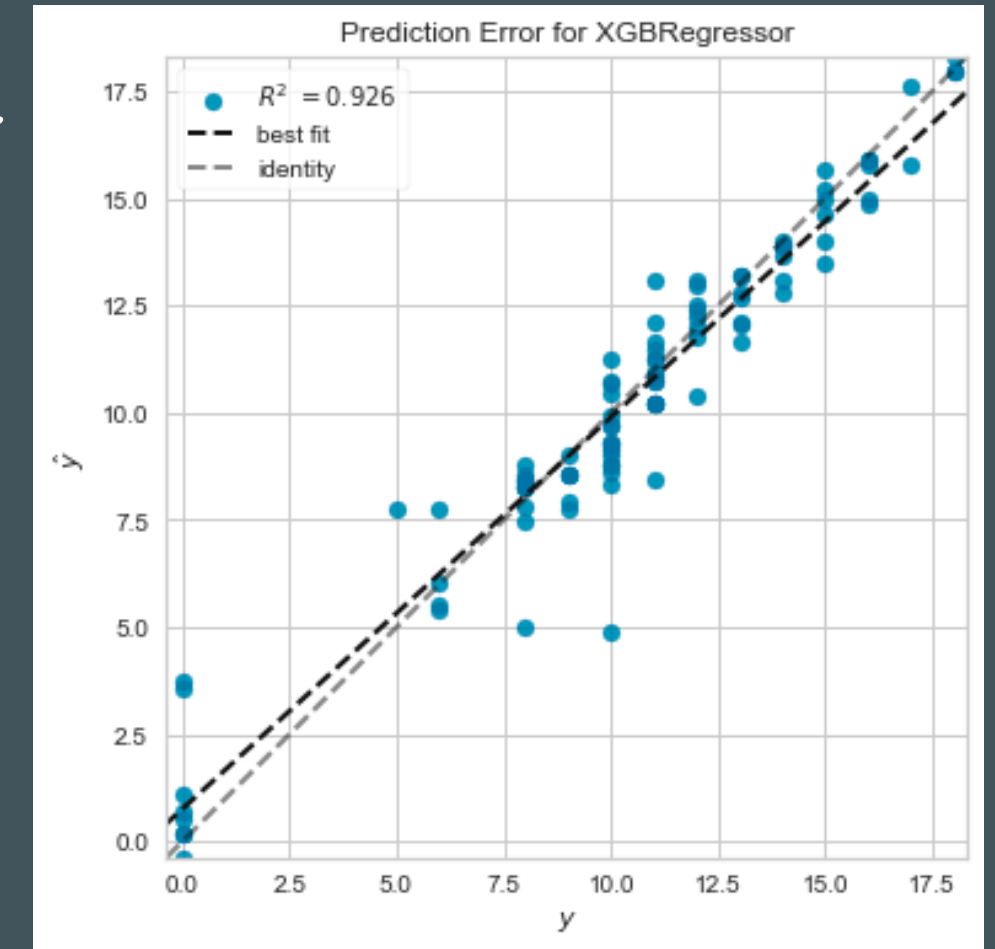
*Overfitting!*

# Learning something new

## 2) Extreme Gradient Boosting Regression



Prediction error



Testing sample data

After Tuning

Ordinary Least Squares  
Grid Search

Family Relationship  
Absences  
school\_GP  
G1  
G2

\*\*\*Final Model\*\*\*

----R Square (Variance)----

Training accuracy (r\_sq) is: 0.8901456368572296

Testing accuracy (r\_sq) is: 0.925721497930853

----Mean Squared Error----

Training MSE after tuning is: 2.4216796033716963

Testing MSE after tuning is: 1.3089113450017726

----Difference in MSE----

The testing MSE has improved by 0.16599657807894586 after tuning.

The testing MSE has improved by 2.5974975409544294 after tuning.

Learning something new

## 2) Extreme Gradient Boosting Regression

After Tuning

```
***Final Model***
----R Square (Variance)----
Training accuracy (r_sq) is:  0.8901456368572296
Testing accuracy (r_sq) is:  0.925721497930853

----Mean Squared Error----
Training MSE after tuning is:  2.4216796033716963
Testing MSE after tuning is:  1.3089113450017726

----Difference in MSE----
The testing MSE has improved by  0.16599657807894586 after tuning.
The testing MSE has improved by  2.5974975409544294 after tuning.
```

*K-Fold cross validation*

```
***Final Model***
cross val training mean_squared_error is: 2.1140157125925145
cross val testing mean_squared_error is: 2.7227312723629233
```

# Conclusion



What Affects a Student's Grades?

- Absences
- G2 [2nd Period Grade]





# Contributions



Dataset Finding and Problem Statement	All		
Data Visualisation/Analysis	Muhammad Ameerul Bin Azman	Peng Teng Kang	
Data Preparation	Heng Zi Hui		
Data Science/ Machine Learning	Guerta Uno Gabriel Yap (Decision Tree)		Peng Teng Kang (Linear Regression)
Learning something new	Guerta Uno Gabriel Yap (Random Forest Regression, yellowbrick)	Heng Zi Hui (XGBoost Regression, Tuning [OLS, Grid Search], One Hot Encoding)	Muhammad Ameerul Bin Azman (Bar Plot)





# Thank you!

*Do you have any questions before we go?*

