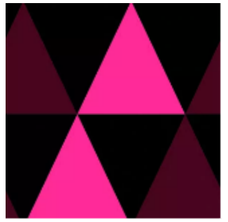
Student Grade Prediction

Predict if the student will get average score > 10 or ≤ 10



Student Grade Prediction

Predict the final grade of Portugese high school students

kaggledatasets

https://www.kaggle.com/dipam7/student-grade-prediction

Check the data

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes

	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
	3	4	1	1	3	6	5	6	6
	3	3	1	1	3	4	5	5	6
	3	2	2	3	3	10	7	8	10
	2	2	1	1	5	2	15	14	15
	3	2	1	2	5	4	6	10	10

. . .

The data has 395 rows, and 33 columns (features). Where G1, G2, and G3, are 1st period grades, 2nd period grades, and final year grade, respectively.

Ikhwanul Muslimin

github.com/waannuullll

linkedin.com/in/ikhwanulmuslimin/

Student Grade Prediction

Predict if the student will get average score > 10 or ≤ 10

Student Grade Prediction

Predict the final grade of Portugese high school students

k kaggledatasets

https://www.kaggle.com/dipam7/student-grade-prediction

The result when using all of the variables (Logistic Regression)

```
[23] # evaluate classification model - accuracy
     accuracy_test = metrics.accuracy_score(y_test,y_test_pred)
     print('Accuracy Test Data: {}'.format(accuracy_test))
```

```
Accuracy Test Data: 1.0
```

```
[24] # classification report
     print(classification_report(y_test,y_test_pred))
```

```
                precision    recall  f1-score   support

             0       1.00      1.00      1.00        39
             1       1.00      1.00      1.00        40

      accuracy                           1.00        79
     macro avg       1.00      1.00      1.00        79
  weighted avg       1.00      1.00      1.00        79
```

Wow! Our accuracy is perfect, 100%!!!

Wait... what???

Is our model correct?

Ikhwanul Muslimin

github.com/waannuullll

linkedin.com/in/ikhwanulmuslimin/

Student Grade Prediction

Predict if the student will get average score > 10 or ≤ 10

Student Grade Prediction

Predict the final grade of Portugese high school students

kaggledatasets

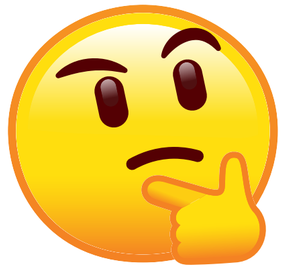
https://www.kaggle.com/dipam7/student-grade-prediction

Multicollinearity

The reason for the absurdity of our results is multicollinearity.

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	Gavg	Passed
age	1.00	-0.16	-0.16	0.07	-0.00	0.24	0.05	0.02	0.13	0.13	0.12	-0.06	0.18	-0.06	-0.14	-0.16	-0.11	-0.09
Medu	-0.16	1.00	0.62	-0.17	0.06	-0.24	-0.00	0.03	0.06	0.02	-0.05	-0.05	0.10	0.21	0.22	0.22	0.19	0.20
Fedu	-0.16	0.62	1.00	-0.16	-0.01	-0.25	-0.00	-0.01	0.04	0.00	-0.01	0.01	0.02	0.19	0.16	0.15	0.17	0.18
traveltime	0.07	-0.17	-0.16	1.00	-0.10	0.09	-0.02	-0.02	0.03	0.14	0.13	0.01	-0.01	-0.09	-0.15	-0.12	-0.10	-0.09
studytime	-0.00	0.06	-0.01	-0.10	1.00	-0.17	0.04	-0.14	-0.06	-0.20	-0.25	-0.08	-0.06	0.16	0.14	0.10	0.13	0.11
failures	0.24	-0.24	-0.25	0.09	-0.17	1.00	-0.04	0.09	0.12	0.14	0.14	0.07	0.06	-0.35	-0.36	-0.36	-0.31	-0.33
famrel	0.05	-0.00	-0.00	-0.02	0.04	-0.04	1.00	0.15	0.06	-0.08	-0.11	0.09	-0.04	0.02	-0.02	0.05	0.01	-0.02
freetime	0.02	0.03	-0.01	-0.02	-0.14	0.09	0.15	1.00	0.29	0.21	0.15	0.08	-0.06	0.01	-0.01	0.01	-0.01	-0.01
goout	0.13	0.06	0.04	0.03	-0.06	0.12	0.06	0.29	1.00	0.27	0.42	-0.01	0.04	-0.15	-0.16	-0.13	-0.17	-0.16
Dalc	0.13	0.02	0.00	0.14	-0.20	0.14	-0.08	0.21	0.27	1.00	0.65	0.08	0.11	-0.09	-0.06	-0.05	-0.14	-0.06
Walc	0.12	-0.05	-0.01	0.13	-0.25	0.14	-0.11	0.15	0.42	0.65	1.00	0.09	0.14	-0.13	-0.08	-0.05	-0.19	-0.11
health	-0.06	-0.05	0.01	0.01	-0.08	0.07	0.09	0.08	-0.01	0.08	0.09	1.00	-0.0					
absences	0.18	0.10	0.02	-0.01	-0.06	0.06	-0.04	-0.06	0.04	0.11	0.14	-0.03	1.0	-0.03	-0.03	0.03	-0.18	-0.06
G1	-0.06	0.21	0.19	-0.09	0.16	-0.35	0.02	0.01	-0.15	-0.09	-0.13	-0.07	-0.0	1.00	0.85	0.80	0.96	0.79
G2	-0.14	0.22	0.16	-0.15	0.14	-0.36	-0.02	-0.01	-0.16	-0.06	-0.08	-0.10	-0.0	0.85	1.00	0.90	0.98	0.77
G3	-0.16	0.22	0.15	-0.12	0.10	-0.36	0.05	0.01	-0.13	-0.05	-0.05	-0.06	0.0	0.80	0.90	1.00	0.98	0.74
Gavg	-0.11	0.19	0.17	-0.10	0.13	-0.31	0.01	-0.01	-0.17	-0.14	-0.19	-0.07	-0.1	0.96	0.98	0.98	1.00	0.80
Passed	-0.09	0.20	0.18	-0.09	0.11	-0.33	-0.02	-0.01	-0.16	-0.06	-0.11	-0.01	-0.0	0.79	0.77	0.74	0.80	1.00

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model (Investopedia)



Ikhwanul Muslimin

github.com/waannuullll

linkedin.com/in/ikhwanulmuslimin/

Student Grade Prediction

Predict if the student will get average score > 10 or ≤ 10

Student Grade Prediction

Predict the final grade of Portugese high school students

kaggledatasets

https://www.kaggle.com/dipam7/student-grade-prediction

	variables	VIF
0	age	1.201556
1	Medu	1.776802
2	Fedu	1.731367
3	traveltime	1.095101
4	studytime	1.148875
5	failures	1.315812
6	famrel	1.105915
7	freetime	1.196457
8	goout	1.389888
9	Dalc	1.814727
10	Walc	2.117870
11	health	1.058943
12	absences	1.116536
13	G1	4.015640
14	G2	7.924867
15	G3	6.165788

In order to eliminate the multicollinearity, we have to calculate the Variance Inflation Factor (VIF) of each variable, and drop the variable which has VIF > 5.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables (Investopedia).

So, we have to drop G2 and G3 because these variables are highly collinear with other independent variables.

Ikhwanul Muslimin

github.com/waannuullll

linkedin.com/in/ikhwanulmuslimin/

Student Grade Prediction

Predict if the student will get average score > 10 or ≤ 10

Student Grade Prediction

Predict the final grade of Portugese high school students

k kaggledatasets

https://www.kaggle.com/dipam7/student-grade-prediction

The result when considering the multicollinearity (Logistic Regression)

```
[30] # evaluate classification model - accuracy
     accuracy_test = metrics.accuracy_score(y_test,y_test_pred)
     print('Accuracy Test Data: {}'.format(accuracy_test))
```

```
Accuracy Test Data: 0.9240506329113924
```

```
[31] # classification report
     print(classification_report(y_test,y_test_pred))
```

```
                precision    recall  f1-score   support

           0       0.88      0.97      0.93        39
           1       0.97      0.88      0.92        40

    accuracy                           0.92        79
   macro avg       0.93      0.92      0.92        79
weighted avg       0.93      0.92      0.92        79
```

Our model's performance now drop from 100% to around 92.4%. It's okay, because we've already improving the overall reliability of our results!

Ikhwanul Muslimin

github.com/waannuullll

linkedin.com/in/ikhwanulmuslimin/