Influence of demographic, social and school factors on upper secondary education Mathematics test

Erick Calderon-Morales

Fall 2021

Introduction

Education is one of the main key areas of investment that dominate the political agenda in the European Union (EU). Despite the recent global economic crisis, policymakers have set ambitious goals like reducing school dropouts and increasing the percentage of the population with an educational degree. With this kind of strategy, policymakers are trying to address the increasing social inequalities that affect the citizens of the EU (Busemeyer et al, 2018).

Achieving equality in education is recognized by the politicians as key, but among the EU members, there is variation in the socioeconomic factors that influence the citizens of each country. This kind of variation among countries makes it difficult to pass general bills that reduce education inequality in the EU. Because of this, it is important to investigate each country separately in order to draw conclusions that can potentially be generalized to the whole EU. (Hippe, Araujo and da Costa, 2016).

Portugal's educational system is characterized by providing universal access to its citizens. This has led to reducing the number of schools dropouts, (having the lowest out-of-school rate at the high school level among OCDE members). Despite this, Portugal has one of the highest unemployment rates of recent graduates with upper secondary education and one of the lowest medium and high educational attainment (O.E.C.D, 2021).

In this report, I investigate the demographic, social, and school factors that influence the performance of students in the mathematics test. Specifically, I address the following questions, which factors determine the score in mathematics (A1) and which factors can be used to predict if a student fails or pass the test (A2). The main goal of this report is to develop predictive models for the final year grade (G3) using statistical learning methods.

Methods

For addressing which factors determine the score in mathematics (A1) I used multiple regression analysis, partial least squares regression, principal components regression, LASSO regression, and Ridge regression, while for addressing which factors can be used to predict if a student fails or pass the test (A2) I used decision trees, Bagging, Random forests, Random forest with gradient boosting and k-Nearest Neighbor.

The data set was split into a training set containing 75% of the observations and testing set containing the rest of the observations. For the evaluation of all the models, I used the testing set. In the case of the linear models, I used the mean squared error on the testing data set for evaluating the fit of the models, and for the classification models, I estimated the accuracy, sensitivity, and specificity of each model using the test data set.

Results

Exploratory study of the variables.

The data set used contains a total of 395 observations (with no missing values) of students from two schools Gabriela Pereira and Mousinho da Silveira. In total 34 variables were measured among students which 31 were considered as factors (Figure 1) and 3 were considered as numeric (g1, g2, g3).

In general, most of the students, come from the school called Gabriela Pereira (88%), live in most of the cases in urban areas (78%), travel approximately 1h to get to school (65%), dedicated 2h to studying (50%), failed almost no class (78%), do not receive extra class support from the school (87%), the parents support them to stay in school (61%), attended to nursery school (70%), want to take higher education (95%), have internet access (83%), have no romantic relationships (67) and have a low alcoholic consumption during the workday. In terms of family, most of the students have families with sizes greater than 3 (71%) and have good family relationships (40% and 27%) and most of the parents have at least some degree of education. Also in terms of absences, there were 6 students with more than 30 absences (Figure 1).

All students took three tests, the first-period grade (g1), the second-period grade (g2), and the final grade (g3). All of these have a score that varies from 0 to 20 being 0 the lowest score possible. Overall among students, g1 had a mean value of 10.90 with a standard deviation of 3.31, g2 had a mean value of 10.71 and a standard deviation of 3.76 and g3 had a mean of 10.41 with a standard deviation of 4.5 (Table 1). It is worth mentioning that each of these variables has a high correlation with each other which gets less strong across periods (Table 2). Finally is important to mention that the proportion of students that got a score of 0 increased across tests. In the g1 test, a total of 0 students had a score of 0 while in the g2 and g3 a total of 13 and 38 got a score of 0 respectively. I used g3 as the response variable and g1 and g2 as predictors. No outliers or skewed distributions were found in any of these variables (Figure 2).

When each categorical variable was evaluated I couldn't find any dominating trend. This makes me think that the determinants of student success in the g3 test are a combination of different socioeconomic factors (Figure 3).

The zero (zeroVar) and near-zero variance (nzv) predictors are those categorical predictors that only have a unique or a highly extreme frequency of values. These kinds of predictors are characterized by having a fraction of unique values over the sample size is low (say 10 %) and having a ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20). I evaluated this in all the categorical and integer predictors and none of them have this problem (Kuhn and Johnson, 2013) (Table 3).

Linear regression selection/dimension reduction methods (A1)

For answering which factors determine the score in mathematics (A1) I used a total of five models. The model with the worse performance was the best subset model (MLR) which showed the largest MSE in the test set while the Ridge regression and the LASSO regression showed the smallest test error (Table 4). After this process, I fitted both models to the full data set.

Figure 4 shows the coefficients from the LASSO regression and the Ridge regression. As expected (because of the high correlation between g1 and the predictors g2 and g3) g2 and g1 have a high predictive power over the scores of g3, so students that scored high in the g1 and g2 will score high in the final exam.

Classification analysis (A2 and A3)

Overall most of the models evaluated had high accuracy in classifying which students pass (g3 scores > 10) and which students fail (g3 scores < 10). The worst performing model was the kNN model while the best performing was the Random forest with gradients boosting (Table 5).

Random forest with gradients boosting model showed almost the same results like the ones obtained with the ridge regression where g1 and g2 have the largest importance on predicting which students fail and which ones pass. For these reasons, I decided to exclude g1 and g2 from the data set and re-run the models to determine which other factors predict the students that pass and fail (Figure 5).

After removing the g1 and g2 from the data set, all performance metrics decrease across all models, but still, the Random forest with gradients boosting is best performing since it has the greatest values of accuracy sensibility and specificity (Table 6).

Summary of analysis results in context

Equality in education is one of the most important challenges of the EU and for achieving this goal it is necessary to employ different statistical techniques that guide the formulation of data-driven policies. In this report, I found that Random forest and Ridge regressions are adequate techniques for understanding which factors might influence students' performance in schools of Portugal. Using Random forest I found that the main determinants of test scores are how well the student did in past tests (Figure 6). After taking out g1 and g2 from the analysis I found that the most important predictors of g3 scores are, the number of pass class failures (failures), the number of school absences (absences), and how much each student go out with friends (gouot) (Figure 6).

On the other hand, using the ridge regression I found that the reason for choosing the school, if a student wants to take higher education, and the past scores (g1, g2) had a positive effect over the final score while the number of past class failures, if the students have a romantic relationship and the parent's cohabitation status have a negative effect over the final scores (Figure 4).

Of all models considered in this report, I think this last one (Ridge regression) is the one that offers the most insights, mainly because with this model we can identify the most important predictors that have a positive and negative effect on school performance which could inform better policymakers.

References

Busemeyer, M. R., Garritzmann, J. L., Neimanns, E., & Nezi, R. (2018). Investing in education in Europe: Evidence from a new survey of public opinion. Journal of European Social Policy, 28(1), 34-54.

Hippe, R., Araujo, L., & da Costa, P. D. (2016). Equity in education in Europe. Luxembourg: Publications Office of the European Union.

Indicators, O.E.C.D. (2021). Education at a Glance 2016. Editions OECD, 90.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 43). New York: Springer.

Figures

Figure 1

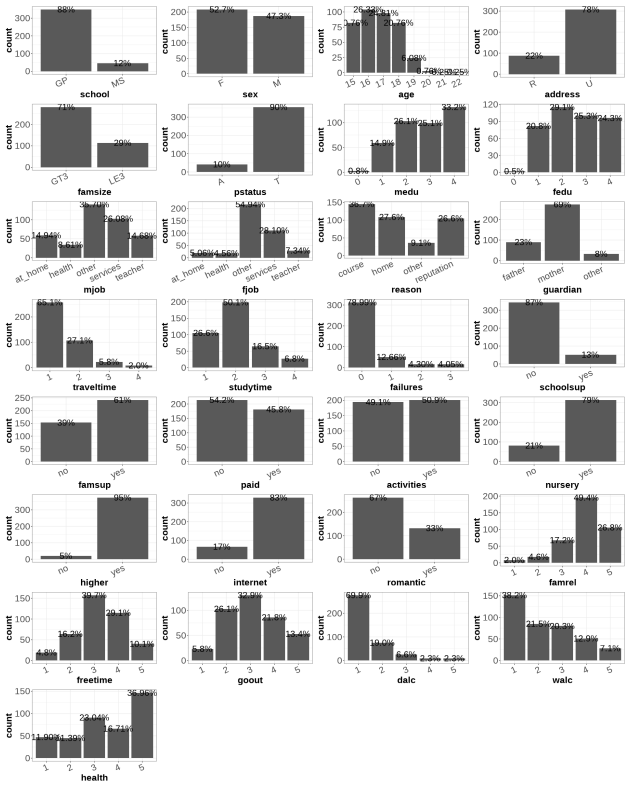


Figure 2

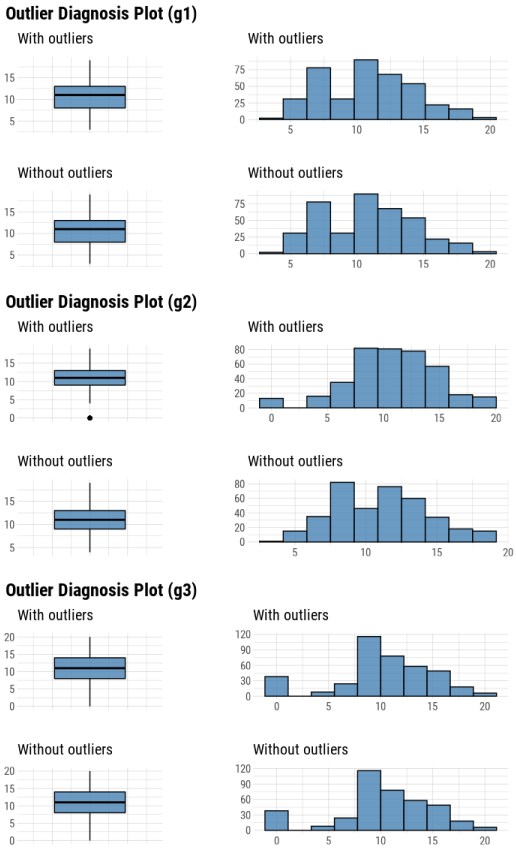


Figure 3

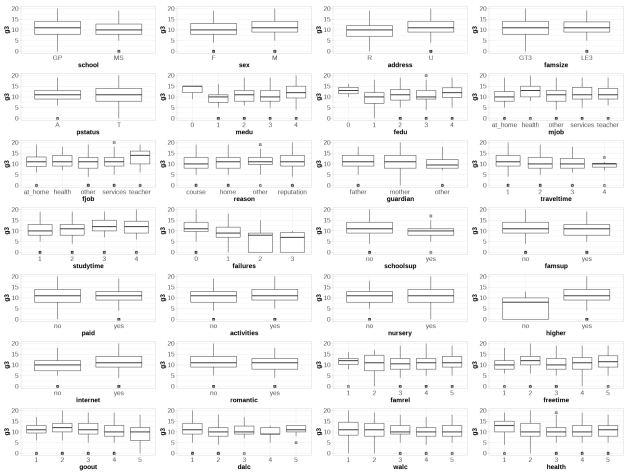


Figure 4

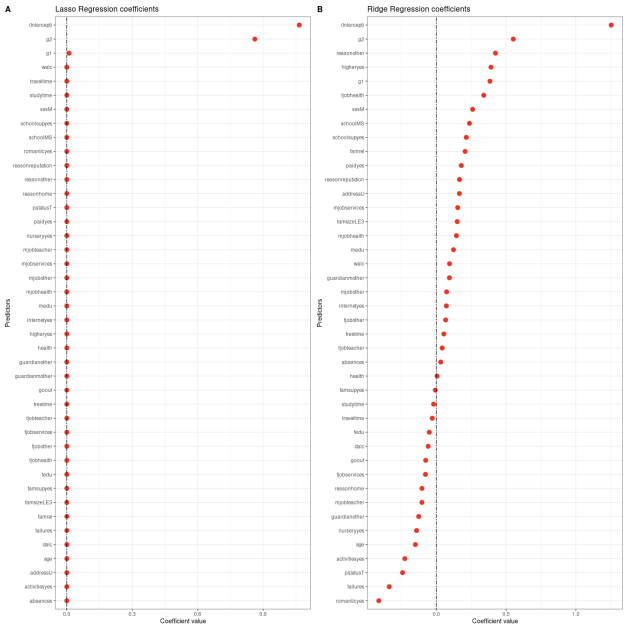

A
Lasso Regression coefficients
(Intercept)
g2
g1
walc
traveltime
studytime
sexM
schoolsupyes
schoolMS
romanticyes
reasonreputation
reasonother
reasonhome
pstatusT
paidyes
nurseryyes
mjobteacher
mjobservices
mjobother
mjobhealth
medu
internetyes
higheryes
health
guardianother
guardianmother
goout
freetime
fjobteacher
fjobservices
fjobother
fjobhealth
fedu
famsupyes
famsizeLE3
famrel
failures
dalc
age
addressU
activityesyes
absences
Predictors
0.0
0.3
0.6
0.9
Coefficient value
B
Ridge Regression coefficients
(Intercept)
g2
reasonother
higheryes
g1
fjobhealth
sexM
schoolMS
schoolsupyes
famrel
paidyes
reasonreputation
addressU
mjobservices
famsizeLE3
mjobhealth
medu
walc
guardianmother
mjobother
internetyes
fjobother
freetime
fjobteacher
absences
health
famsupyes
studytime
traveltime
fedu
dalc
goout
fjobservices
reasonhome
mjobteacher
guardianother
nurseryyes
age
activityesyes
pstatusT
failures
romanticyes
Predictors
0.0
0.5
1.0
Coefficient value

Figure 5

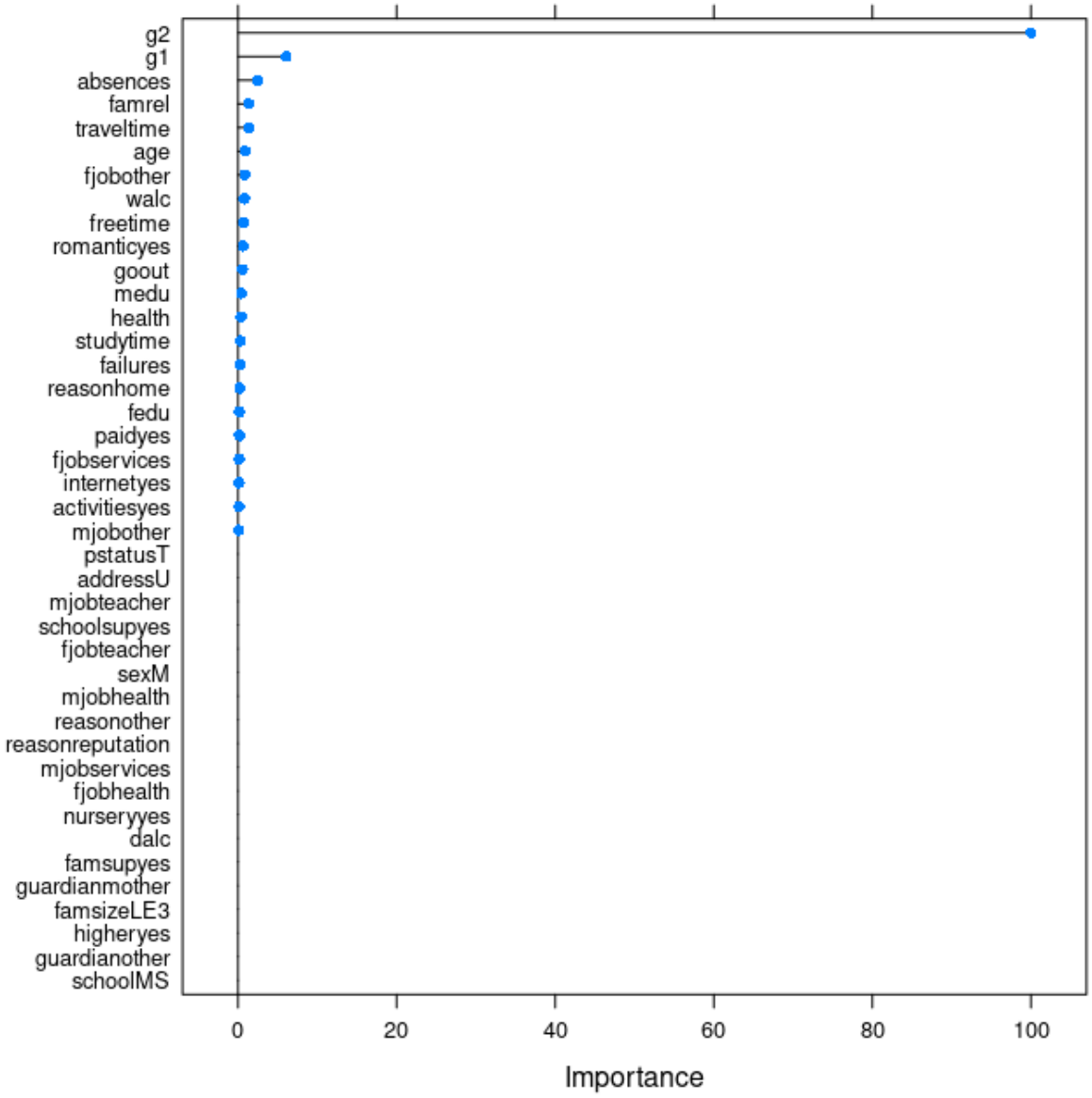
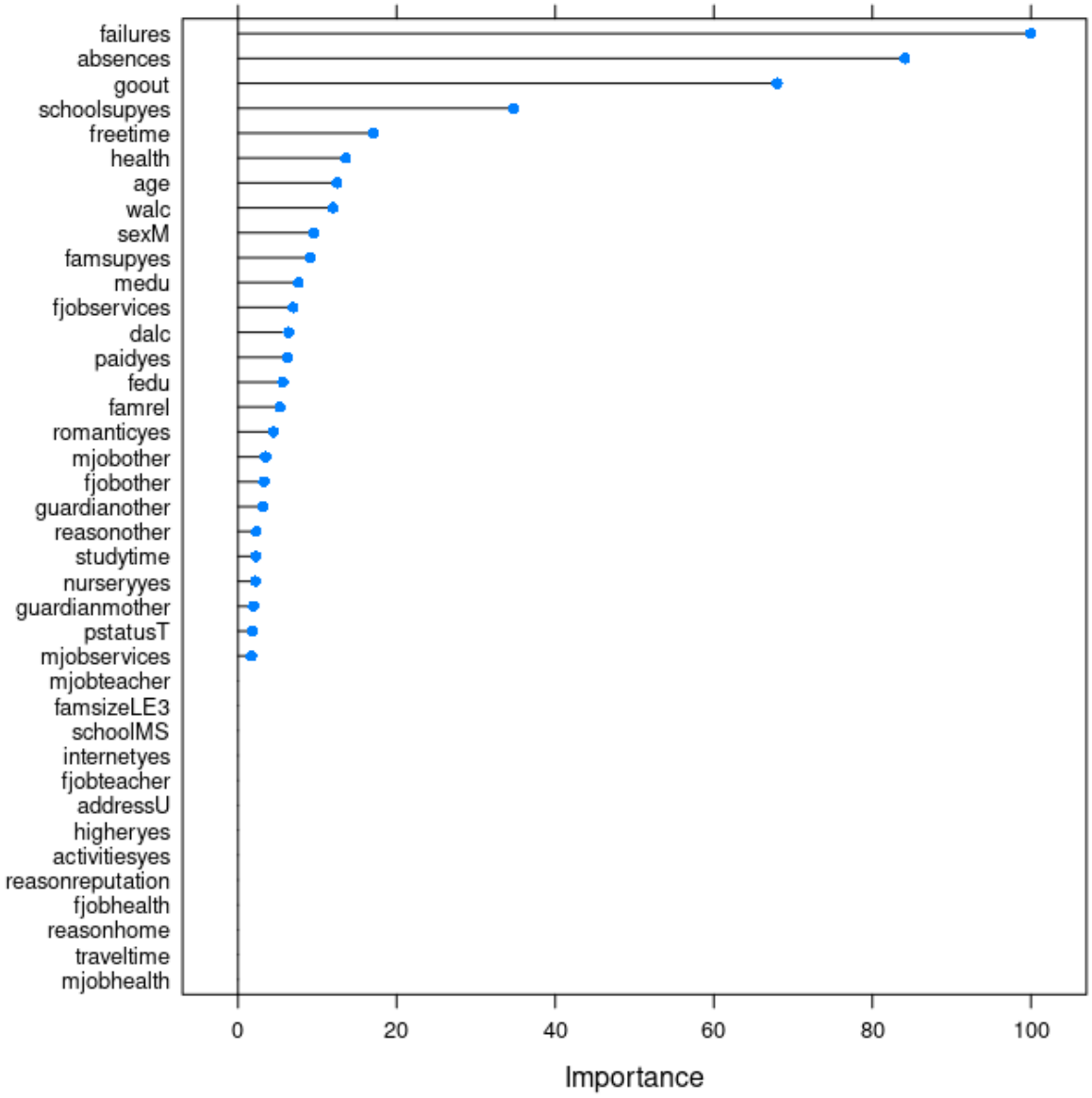

g2
g1
absences
famrel
traveltime
age
fjobother
walc
freetime
romanticyes
goout
medu
health
studytime
failures
reasonhome
fedu
paidyes
fjobservices
internetyes
activitiesyes
mjobother
pstatusT
addressU
mjobteacher
schoolsupyes
fjobteacher
sexM
mjobhealth
reasonother
reasonreputation
mjobservices
fjobhealth
nurseryyes
dalc
famsupyes
guardianmother
famsizeLE3
higheryes
guardianother
schoolMS
0
20
40
60
80
100
Importance

Figure 6



Tables

Table 1

variable	n	na	mean	sd	se_mean	IQR	skewness	kurtosis	p00	p01	p05	p10	p20	p25	p30	p40	p50	p60	p70	p75	p80	p90	p95	p99	p100
g1	395	0	10.90886	3.319195	0.1670068	5	0.2406132	-0.6938295	3	5	6	7	8	8	9	10	11	12	13	13	14	16.0	16.0	18	19
g2	395	0	10.71392	3.761505	0.1892618	4	-0.4316454	0.6277056	0	0	5	6	8	9	9	10	11	12	13	13	14	15.0	16.3	18	19
g3	395	0	10.41519	4.581443	0.2305174	6	-0.7326724	0.4034208	0	0	0	5	8	8	9	10	11	12	13	14	14	15.6	17.0	19	20

Table 2

g3	g2	g1
1.0000000	0.9048680	0.8014679
0.9048680	1.0000000	0.8521181
0.8014679	0.8521181	1.0000000

Table 3

freqRatio	percentUnique	zeroVar	nzv
7.586957	0.5063291	FALSE	FALSE
1.112299	0.5063291	FALSE	FALSE
1.061224	2.0253165	FALSE	FALSE
3.488636	0.5063291	FALSE	FALSE
2.464912	0.5063291	FALSE	FALSE
8.634146	0.5063291	FALSE	FALSE
1.271845	1.2658228	FALSE	FALSE
1.150000	1.2658228	FALSE	FALSE
1.368932	1.2658228	FALSE	FALSE
1.954955	1.2658228	FALSE	FALSE
1.330275	1.0126582	FALSE	FALSE
3.033333	0.7594937	FALSE	FALSE
2.401869	1.0126582	FALSE	FALSE
1.885714	1.0126582	FALSE	FALSE
6.240000	1.0126582	FALSE	FALSE
6.745098	0.5063291	FALSE	FALSE
1.581699	0.5063291	FALSE	FALSE
1.182320	0.5063291	FALSE	FALSE
1.036082	0.5063291	FALSE	FALSE
3.876543	0.5063291	FALSE	FALSE
18.750000	0.5063291	FALSE	FALSE
4.984848	0.5063291	FALSE	FALSE
1.992424	0.5063291	FALSE	FALSE
1.839623	1.2658228	FALSE	FALSE
1.365217	1.2658228	FALSE	FALSE
1.262136	1.2658228	FALSE	FALSE
3.680000	1.2658228	FALSE	FALSE
1.776471	1.2658228	FALSE	FALSE
1.604396	1.2658228	FALSE	FALSE
1.769231	8.6075949	FALSE	FALSE

Table 4

model	mse_train	mse_test	difference
MLR	3.666164	31.270229	27.6040650
PCR	7.982341	6.688145	1.2941958
PLS	4.353050	4.023175	0.3298759
RIDGE	4.173689	3.895818	0.2778710
LASSO	4.747010	3.541852	1.2051580

Table 5

model	Accuracy	AccuracyLower	AccuracyUpper	Sensitivity	Specificity
decision_tree	0.8989899	0.8220754	0.9504891	0.8714286	0.9655172
bagging_tree	0.9090909	0.8344310	0.9575835	0.8955224	0.9375000
random_forest	0.9090909	0.8344310	0.9575835	0.8840580	0.9666667
random_rorest_gradient_boosting	0.9191919	0.8469730	0.9644665	0.9090909	0.9393939
knn	0.7373737	0.6393278	0.8207276	0.9354839	0.4054054

Table 6

model	Accuracy	AccuracyLower	AccuracyUpper	Sensitivity	Specificity
decision_tree	0.6868687	0.5858584	0.7763516	0.7073171	0.5882353
bagging_tree	0.6565657	0.5543693	0.7491181	0.7012987	0.5000000
random_forest	0.6868687	0.5858584	0.7763516	0.7125000	0.5789474
random_rorest_gradient_boosting	0.6969697	0.5964500	0.7853320	0.7160494	0.6111111
knn	0.6464646	0.5439647	0.7399466	0.9230769	0.1176471