Predicting Student Success: A Statistical Analysis of Academic Performance Factors

Kim Lu, Sherine Tumushemeze, Gorkem Dural, and Augustine Ezirim

Abstract

This project explores whether secondary school students' academic outcomes can be effectively predicted using demographic factors, academic history, and lifestyle habits. Leveraging a real-world dataset from the UCI Machine Learning Repository, we aimed to classify whether students would pass their final grade *($G3 \geq 10$)* using statistical learning methods. Logistic regression proved to be the most accurate model, identifying the second-period grade (*G2*) as the strongest predictor of success, while lifestyle variables such as weekend alcohol consumption (*Walc*) and frequency of going out with friends *(goout*) also showed statistically significant effects. A decision tree classifier, though less accurate, provided interpretability and revealed that past academic failures were the most influential predictor, along with interactions involving social behaviour and family support. To reduce dimensionality and uncover latent patterns, we applied Principal Component Analysis (PCA), which showed that academic performance and parental education were dominant in the first principal component, while behavioural habits defined the second. Together, these methods highlight how both academic and social factors shape student outcomes and demonstrate the potential of data-driven models to inform early interventions and improve educational decision-making.

Introduction

Academic performance is a critical concern for educators, students, and policymakers alike. Identifying students who are at risk of failing can enable timely interventions that improve outcomes and reduce dropout rates. In this project, we investigate whether a student's likelihood of passing their final grade ($G3 \geq 10$) can be predicted based on demographic characteristics and study habits.

Using a real-world dataset on student performance, we apply several statistical techniques to explore this question. Logistic regression is used to model the binary outcome of pass/fail, while principal component analysis (PCA) helps identify the most influential underlying factors in the dataset. In addition, we implement a decision tree model to visualize the key predictors of academic success and gain insights into how different features interact in classifying students.

This analysis not only demonstrates practical applications of statistical learning but also highlights how data-driven approaches can support educational decision-making and early student support.

Dataset

The dataset used in this study contains comprehensive information on secondary school students, with variables that capture their academic history, demographic background, and lifestyle habits. The primary outcome of interest is whether a student passes the course, which is defined as achieving a final grade (G3) of 10 or higher. To support this analysis, we created a binary target variable (Pass) where students with $G3 \geq 10$ are classified as passing and those with $G3 < 10$ as failing.

Several predictor variables were included to assess their relationship with academic performance. Demographic features such as gender, age, residential area (urban or rural), and parental education levels offer context into the students' backgrounds. In particular, the mother's (Medu) and father's (Fedu) education levels serve as proxies for household educational environment.

The dataset also includes variables related to students' academic habits and lifestyle, such as weekly study time, number of school absences, travel time to school, and amount of free time after school. Additional features like the number of past class failures help capture prior academic challenges. These factors are expected to influence whether a student is likely to pass or fail.

Overall, the dataset is relatively clean, with no missing values and a manageable number of observations for statistical modelling. Categorical variables were recoded or transformed as needed to fit the requirements of the models used. This structure makes the dataset well-suited for logistic regression, principal component analysis, and decision tree modelling, providing a solid foundation for exploring the predictors of academic success.

Decision Tree Classification

To enhance model interpretability and explore feature interactions, we implemented a classification tree to predict whether a student would pass or fail. As shown in Figure 1, the decision tree was built using all available features, with the primary split occurring on the variable *failures*, indicating that previous academic performance was the strongest predictor of current success. Students with at least one past failure were more likely to fail again, while those with no history of failure had a higher likelihood of passing.

Additional splits in the tree revealed other contributing factors. Among students with no past failures, variables such as *goout* (frequency of going out with friends), *Walc* (weekend alcohol consumption), and *famsup* (family educational support) played a role in shaping outcomes. For example, students with low alcohol consumption and higher family support tended to perform better academically. The feature importance summary in Figure 2 reinforces the dominant influence of *failures* on prediction accuracy.

Model performance was evaluated on a test dataset. The final pruned tree achieved an overall accuracy of 70.3%, with a specificity of 91.1% and sensitivity of 28.2%. As illustrated by the ROC curve in Figure 4, the area under the curve (AUC) was 0.597, indicating modest discriminatory power. This suggests that while the model performs well in correctly identifying students who will pass, it struggles to flag those at risk of failing, a limitation reflected in Figure 3 as well.

Despite this, the decision tree provides valuable interpretability. It highlights the dominant role of prior academic failures and certain social behaviours in student performance and offers a transparent framework for decision-making in educational settings.

Logistic Regression

We used logistic regression to predict the likelihood of students passing (G3 ≥ 10), with "Pass" (1 for G3 ≥ 10, 0 for G3 < 10) as the binary outcome. Predictors included prior academic performance (failures, G1, G2), social habits (goout, Walc), and family support (famsup). The dataset was split into 70% training and 30% testing sets.

The model identified G2 as a strong predictor (coefficient: 1.88486, p = 1.54e-10), with each unit increase raising the log-odds of passing by 1.88. Walc (coefficient: 0.49050, p = 0.00877) positively influenced passing, possibly reflecting a complex interplay with student lifestyle, while goout (coefficient: -0.44619, p = 0.03979) had a negative effect, suggesting that frequent outings may detract from academic focus. Failures, famsup, and G1 were not significant ($p > 0.05$) (Appendix, Figure 5: "Logistic Regression Outcomes”). Variance Inflation Factors (VIF) were low (failures: 1.212469, goout: 1.219540, Walc: 1.381774, famsup: 1.013403, G1: 1.225734, G2: 1.261785), confirming no multicollinearity (Appendix, Figure 6: "VIF”)

Test set performance showed an accuracy of 92.41% (95% CI: 0.8934–0.9482), sensitivity of 94.72%, specificity of 87.69%, and AUC of 0.9120, as visualized in the ROC curve (Appendix, Figure 7: "ROC Curve: Simplified Model"). The model excels at identifying both passing and at-risk students, outperforming the decision tree, with the ROC curve’s steep rise indicating robust discrimination across threshold values. This high performance suggests the model could be a valuable tool for educators to prioritize support, though the unexpected positive effect of Walc warrants further investigation into how weekend alcohol consumption might correlate with other unmeasured factors, such as peer influence or stress relief.

This highlights the importance of recent grades (G2) and social habits (goout, Walc) for targeted interventions, offering a data-driven approach to enhance student outcomes. To further explore the underlying structure of the predictors and their relationships, we next apply Principal Component Analysis (PCA).

Principal Component Analysis

The principal component analysis revealed that the first few components capture a substantial portion of the total variance in the dataset. Specifically, the first principal component (PC1) alone explains approximately 21.2% of the variance (Figure 8), with the second (PC2) contributing another 13.1%, and the third (PC3) adding about 9.7%. Cumulatively, the first seven components explain over 71% of the variance, as seen in the cumulative variance plot. This suggests that a considerable amount of the original dataset's complexity can be captured using a reduced number of dimensions, which helps simplify subsequent analyses or visualizations.

Examining the loadings of PC1 (Figure 11), the variables that contribute the most include G2, G3, and G1 — the three grade variables — followed by number of past class failures, mother’s education (Medu), and father’s education (Fedu). The direction and magnitude of these loadings imply that academic performance is the most influential underlying factor in this principal component. PC2, on the other hand, shows high negative loadings for alcohol consumption (Dalc and Walc) and social activities (goout, freetime), indicating that this component may represent a behavioral or lifestyle dimension that is distinct from academic achievement.

Figure 10 demonstrates some dispersion but no strong clustering, suggesting a continuous spread of student profiles along these dimensions. Students who score high on PC1 are generally those with better academic performance and higher parental education levels. Conversely, those with low PC1 scores may be struggling academically. PC2 distinguishes students based on their social and alcohol use behaviors. Overall, PCA effectively reduces the dimensionality of the dataset while preserving key patterns and relationships, making it a powerful tool for exploratory analysis in this educational context.

Conclusion

This analysis demonstrates how statistical learning techniques can be effectively applied to predict student academic performance and uncover underlying factors contributing to success or failure. Among the three modeling approaches, logistic regression emerged as the most accurate, achieving a 92.41% accuracy rate and an AUC of 0.912, indicating excellent discriminatory ability. The variable G2 (second-period grade) was the strongest predictor, significantly increasing the odds of passing with each unit increase ($p < 0.001$). Notably, Walc (weekend alcohol consumption) also showed a positive association with passing (p = 0.00877), while goout (frequency of going out) negatively impacted performance (p = 0.03979), suggesting the influence of social behaviors.

The decision tree classifier, while offering interpretability, achieved a lower accuracy of 70.3%, with an AUC of 0.597, and was particularly limited in sensitivity (28.2%). Nevertheless, it identified past class failures as the most influential variable, with additional splits revealing the roles of social habits and family support in student outcomes. This model highlights how behavioral patterns interact with prior academic history in determining performance.

Complementing these predictive models, Principal Component Analysis (PCA) uncovered latent structures in the data. The first three principal components captured over 44% of the total variance, and the top seven explained more than 71%. Academic performance variables (G1, G2, G3, failures) and parental education levels loaded most heavily on PC1, while PC2 represented lifestyle factors such as alcohol consumption and social activity. These insights reinforce the multidimensional nature of student success, underscoring the need for holistic support strategies that address both academic and behavioral factors. Overall, this project highlights how data-driven methods can inform targeted interventions and resource allocation in educational settings.

Recommendation

Based on the analysis, we recommend that educational institutions implement early identification and support strategies to improve student outcomes. Second-period grades (G2) emerged as the most influential predictor of final performance, making them a critical checkpoint for academic monitoring. Schools should use G2 as a trigger for mid-year assessments and offer additional academic assistance to students showing signs of struggle. Moreover, students with prior academic failures represent a high-risk group and would benefit from personalized academic plans, consistent follow-ups, and access to mentorship programs. Lifestyle factors—such as frequent social outings and weekend alcohol consumption—were also found to negatively impact performance, highlighting the importance of student wellness programs that encourage balance, responsible behavior, and time management. Lastly, the PCA findings emphasize the need for holistic interventions that consider academic and socio-behavioral aspects of student life. Combining these insights can help educators design more effective and comprehensive support systems to enhance academic achievement and student well-being.

Appendix

Figure 1: Decision Tree for Student Pass Prediction

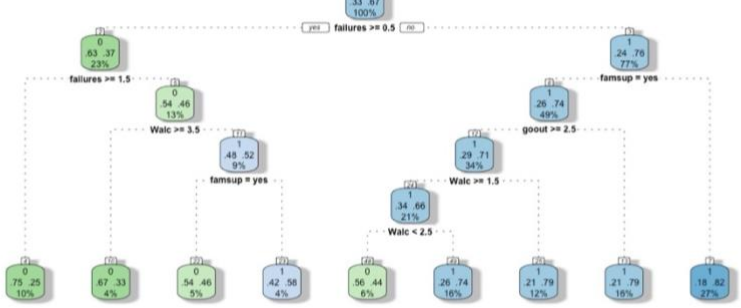


Figure 2: Feature Importance in Decision Tree Model

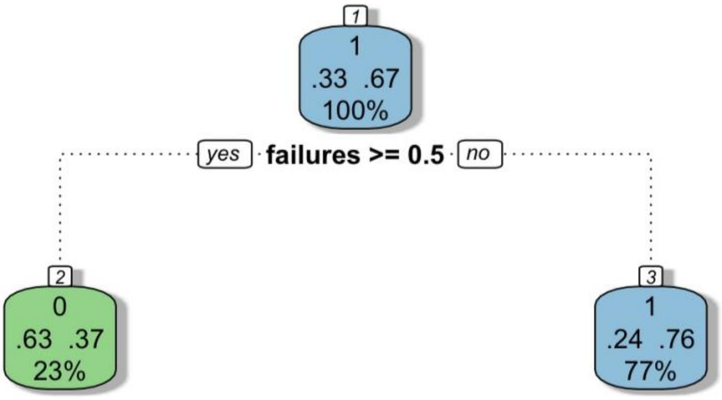


Figure 3: Confusion Matrix Heatmap for Decision Tree

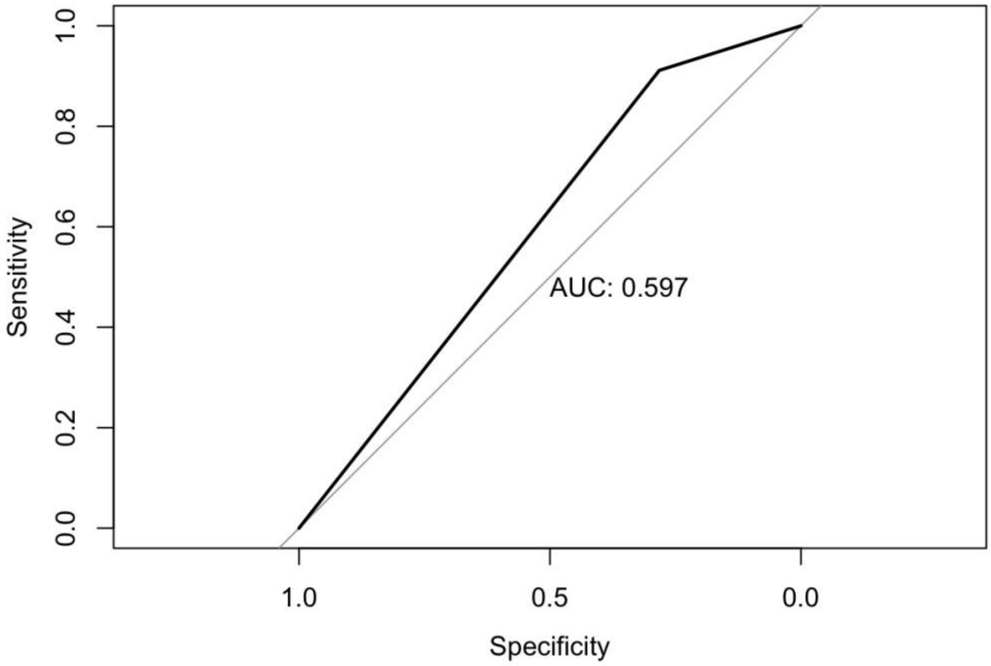


Figure 4: ROC Curve: Student Pass Prediction Tree

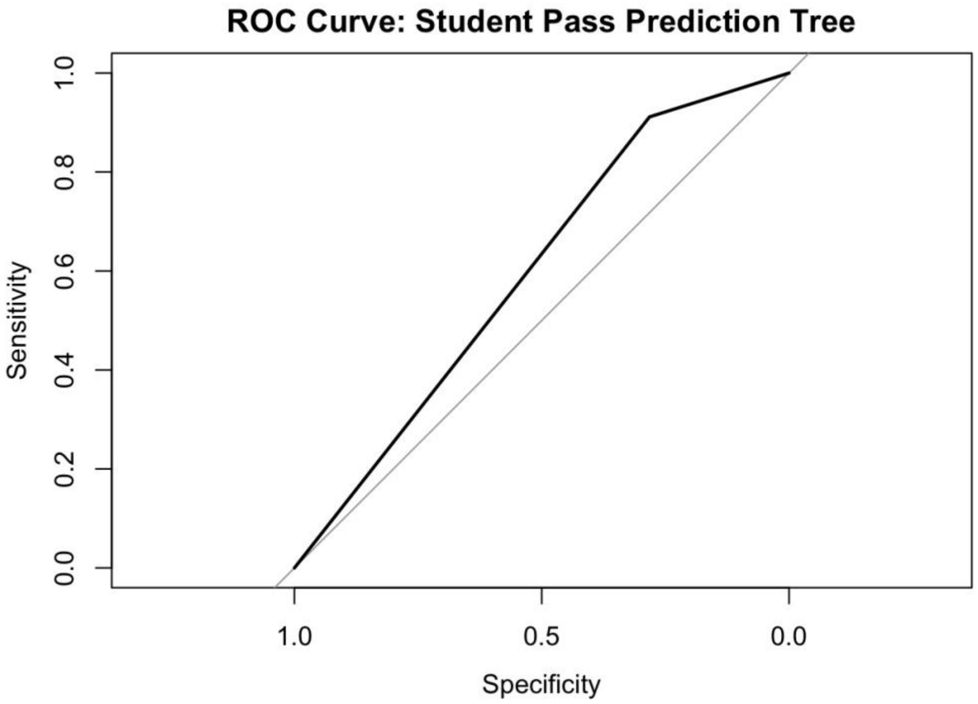
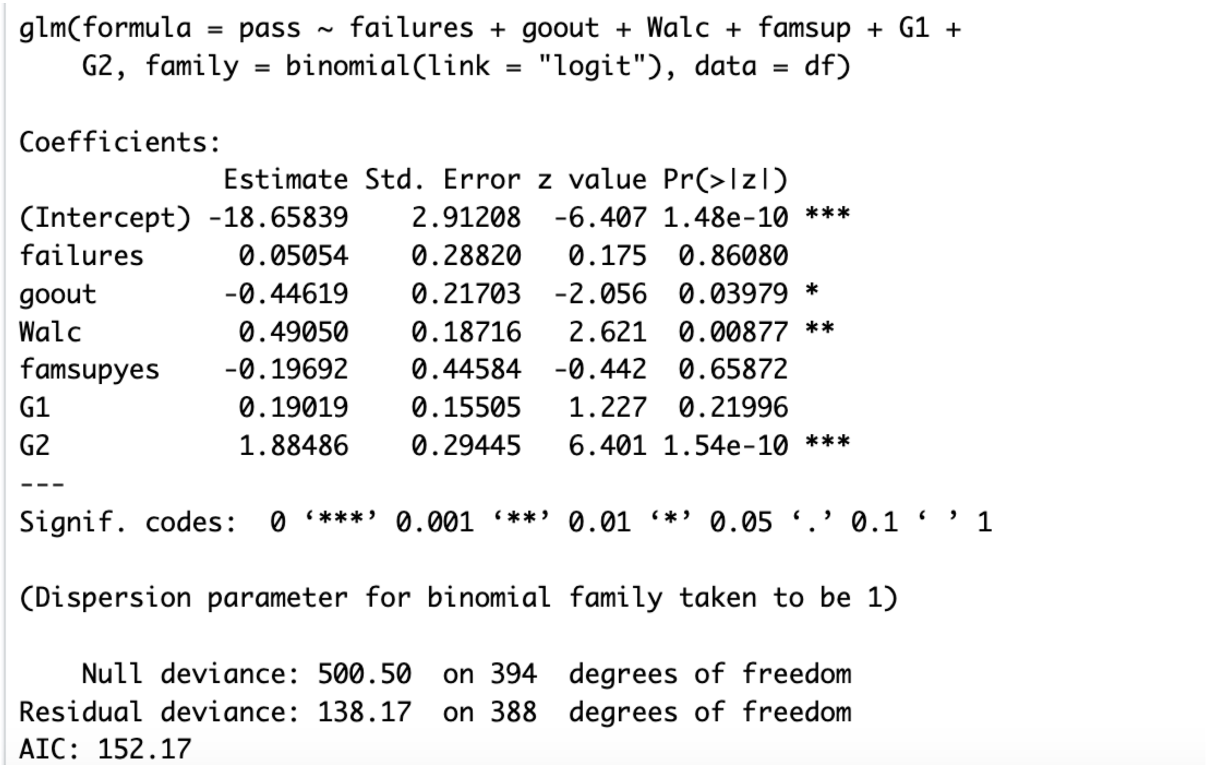
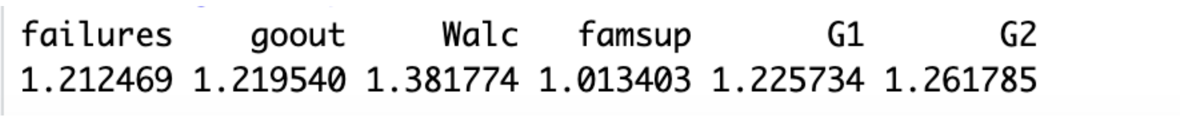


Figure 5: Logistic Regression Outcomes



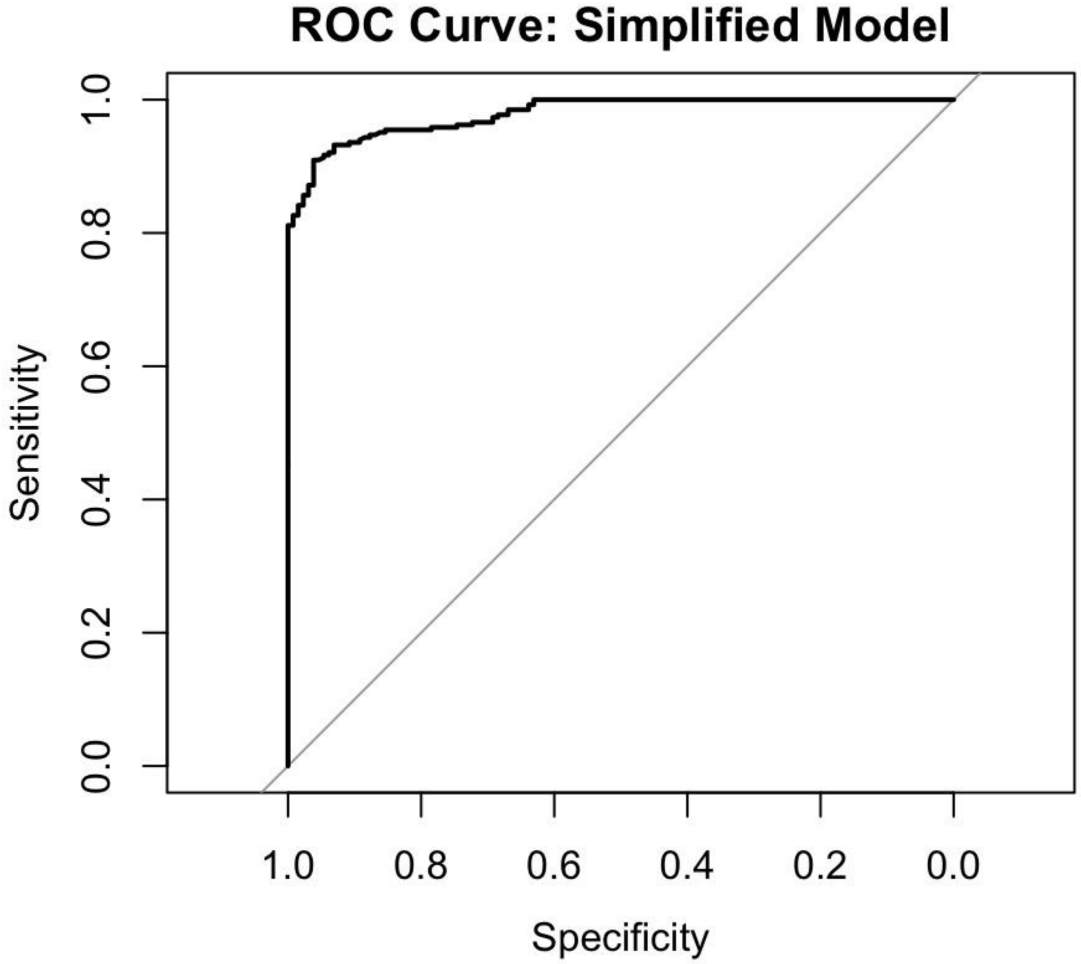
```
glm(formula = pass ~ failures + goout + Walc + famsup + G1 +
    G2, family = binomial(link = "logit"), data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.65839    2.91208  -6.407 1.48e-10 ***
failures      0.05054    0.28820   0.175  0.86080
goout        -0.44619    0.21703  -2.056  0.03979 *
Walc          0.49050    0.18716   2.621  0.00877 **
famsupyes    -0.19692    0.44584  -0.442  0.65872
G1            0.19019    0.15505   1.227  0.21996
G2            1.88486    0.29445   6.401 1.54e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 500.50  on 394  degrees of freedom
Residual deviance: 138.17  on 388  degrees of freedom
AIC: 152.17
```

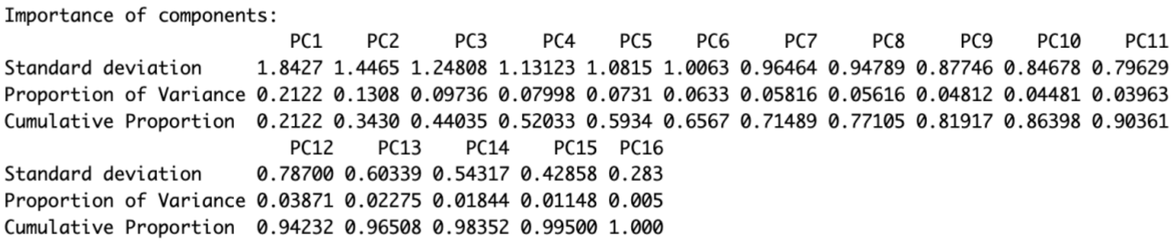
Figure 6: VIF



```
failures     goout      Walc    famsup        G1        G2
1.212469 1.219540 1.381774 1.013403 1.225734 1.261785
```

Figure 7: ROC Curve: Simplified Model





```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5    PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     1.8427 1.4465 1.24808 1.13123 1.0815 1.0063 0.96464 0.94789 0.87746 0.84678 0.79629
Proportion of Variance 0.2122 0.1308 0.09736 0.07998 0.0731 0.0633 0.05816 0.05616 0.04812 0.04481 0.03963
Cumulative Proportion  0.2122 0.3430 0.44035 0.52033 0.5934 0.6567 0.71489 0.77105 0.81917 0.86398 0.90361
                          PC12    PC13    PC14    PC15  PC16
Standard deviation     0.78700 0.60339 0.54317 0.42858 0.283
Proportion of Variance 0.03871 0.02275 0.01844 0.01148 0.005
Cumulative Proportion  0.94232 0.96508 0.98352 0.99500 1.000
```

Figure 8: Proportion of Variance

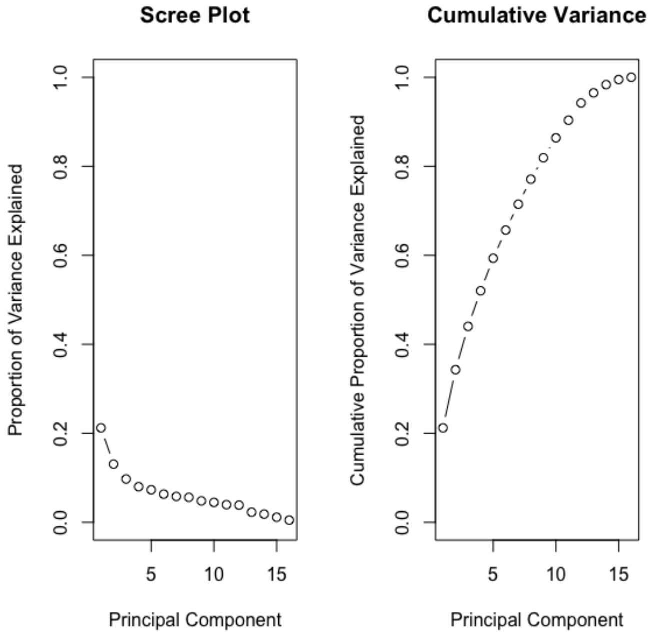


Figure 9: Scree Plot and Cumulative Variance

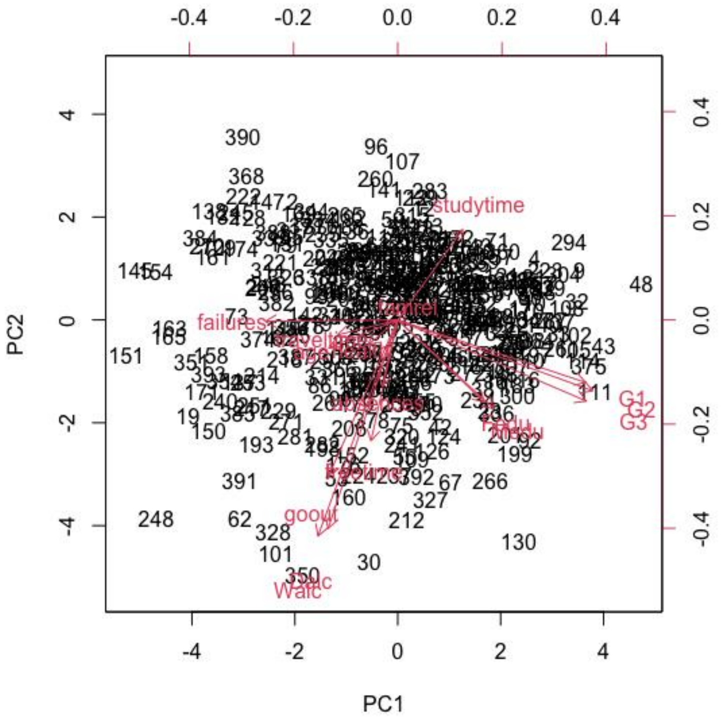
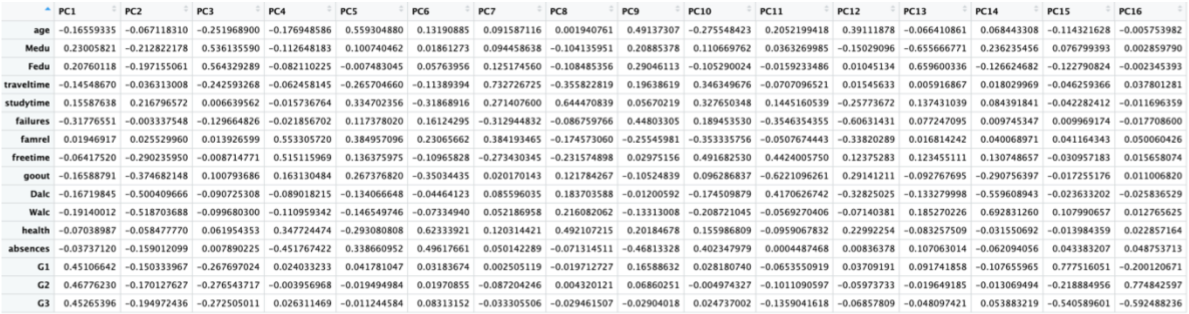


Figure 10: Bi Plot for PC1 vs PC2



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
age	-0.16559335	-0.067118310	-0.251968900	-0.176948586	0.559304880	0.13190885	0.091587116	0.001940761	0.49137307	-0.275548423	0.2052199418	0.39111878	-0.066410861	0.068443308	-0.114321628	-0.005753982
Medu	0.23005821	-0.212822178	0.536135590	-0.112648183	0.100740462	0.01861273	0.094458638	-0.104135951	0.20885378	0.110669762	0.0363269985	-0.15029096	-0.655666771	0.236235456	0.076799393	0.002859790
Fedu	0.20760118	-0.197155061	0.564329289	-0.082110225	-0.007483045	0.05763956	0.125174560	-0.108485356	0.29046113	-0.105290024	-0.0159233486	0.01045134	0.659600336	-0.126624682	-0.122790824	-0.002345393
traveltime	-0.14548670	-0.036313008	-0.242593268	-0.062458145	-0.265704660	-0.11389394	0.732726725	-0.355822819	0.19638619	0.346349676	-0.0707096521	0.01545633	0.005916867	0.018029969	-0.046259366	0.037801281
studytime	0.15587638	0.216796572	0.006639562	-0.015736764	0.334702356	-0.31868916	0.271407600	0.644470839	0.05670219	0.327650348	0.1445160539	-0.25773672	0.137431039	0.084391841	-0.042282412	-0.011696359
failures	-0.31776551	-0.003337548	-0.129664826	-0.021856702	0.117378020	0.16124295	-0.312944832	-0.086759766	0.44803305	0.189453530	-0.3546354355	-0.60631431	0.077247095	0.009745347	0.009969174	-0.017708600
famrel	0.01946917	0.025529960	0.013926599	0.553305720	0.384957096	0.23065662	0.384193465	-0.174573060	-0.25545981	-0.353335756	-0.0507674443	-0.33820289	0.016814242	0.040068971	0.041164343	0.050060426
freetime	-0.06417520	-0.290235950	-0.008714771	0.515115969	0.136375975	-0.10965828	-0.273430345	-0.231574898	0.02975156	0.491682530	0.4424005750	0.12375283	0.123455111	0.130748657	-0.030957183	0.015658074
goout	-0.16588791	-0.374682148	0.100793686	0.163130484	0.267376820	-0.35034435	0.020170143	0.121784267	-0.10524839	0.096286837	-0.6221096261	0.29141211	-0.092767695	-0.290756397	-0.017255176	0.011006820
Dalc	-0.16719845	-0.500409666	-0.090725308	-0.089018215	-0.134066648	-0.04464123	0.085596035	0.183703588	-0.01200592	-0.174509879	0.4170626742	-0.32825025	-0.133279998	-0.559608943	-0.023633202	-0.025836529
Walc	-0.19140012	-0.518703688	-0.099680300	-0.110959342	-0.146549746	-0.07334940	0.052186958	0.216082062	-0.13313008	-0.208721045	-0.0569270406	-0.07140381	0.185270226	0.692831260	0.107990657	0.012765625
health	-0.07038987	-0.058477770	0.061954353	0.347724474	-0.293080808	0.62333921	0.120314421	0.492107215	0.20184678	0.155986809	-0.0959067832	0.22992254	-0.083257509	-0.031550692	-0.013984359	0.022857164
absences	-0.03737120	-0.159012099	0.007890225	-0.451767422	0.338660952	0.49617661	0.050142289	-0.071314511	-0.46813328	0.402347979	0.0004487468	0.00836378	0.107063014	-0.062094056	0.043383207	0.048753713
G1	0.45106642	-0.150333967	-0.267697024	0.024033233	0.041781047	0.03183674	0.002505119	-0.019712727	0.16588632	0.028180740	-0.0653550919	0.03709191	0.091741858	-0.107655965	0.777516051	-0.200120671
G2	0.46776230	-0.170127627	-0.276543717	-0.003956968	-0.019494984	0.01970855	-0.087204246	0.004320121	0.06860251	-0.004974327	-0.1011090597	-0.05973733	-0.019649185	-0.013069494	-0.218884956	0.774842597
G3	0.45265396	-0.194972436	-0.272505011	0.026311469	-0.011244584	0.08313152	-0.033305506	-0.029461507	-0.02904018	0.024737002	-0.1359041618	-0.06857809	-0.048097421	0.053883219	-0.540589601	-0.592488236

Figure 11: Loadings of PCs

Reference

`Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.