Predicting Exam Performance Based On Survey Responses

David Jiang

Introduction

As part of an investigation into what factors influence performance on exams, a survey was distributed among students of the CMSC320 course at UMD. 177 survey responses with 10 different factors[1] were used to train various classifiers to determine whether a given student would pass or fail, and what letter grade they would receive. A regression analysis was also performed on the data to predict the numerical grade a student would receive given responses to the questions in the survey. Overall, none of these analyses were particularly accurate in making predictions.

Principal Component Analysis

Reducing the dimensionality of data can aid in the performance of classifiers in certain cases. Thus, PCA was performed on the survey data. Results are summarized in Figures 1 and 2:

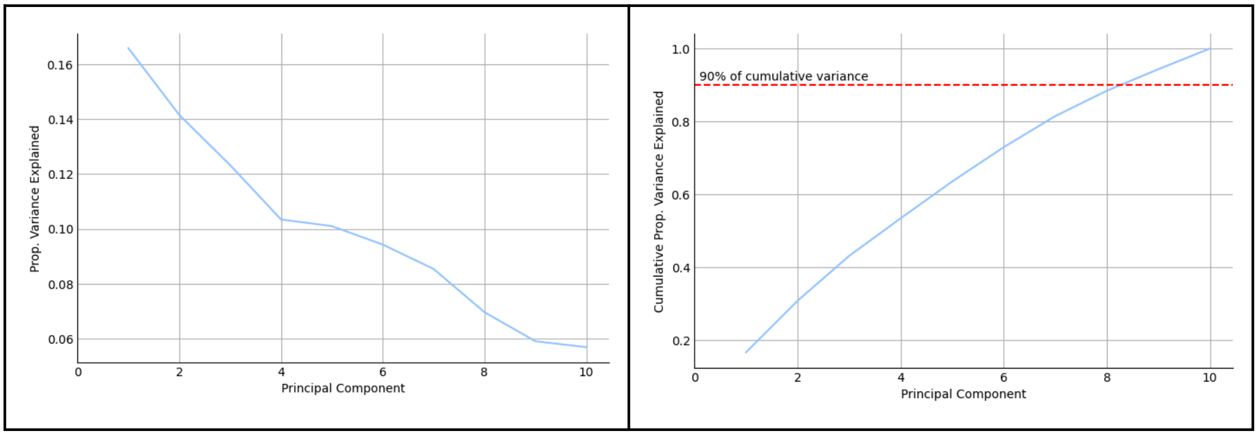

Figure 1. Variance (Proportional and Cumulative) Explained by Each Subsequent Principal Component

An elbow point in the proportional variance explained by each principal component can be seen at 4 principal components. However, principal components after the 4th still explain a substantial amount of variance in the data, as can be seen in the graph of cumulative proportional variance. Only the minimal number of principal components needed to explain 90% of the variance of the data were kept – however, this only reduced the dimensionality of the data from 10 factors to 9 principal components. As will be seen later in the report, PCA did not unilaterally improve accuracy of classification.

[1] The survey had 188 responses and 11 questions. However, one of these questions was an attention check; filtering out responses that failed the attention check left 177 responses and 10 factors.

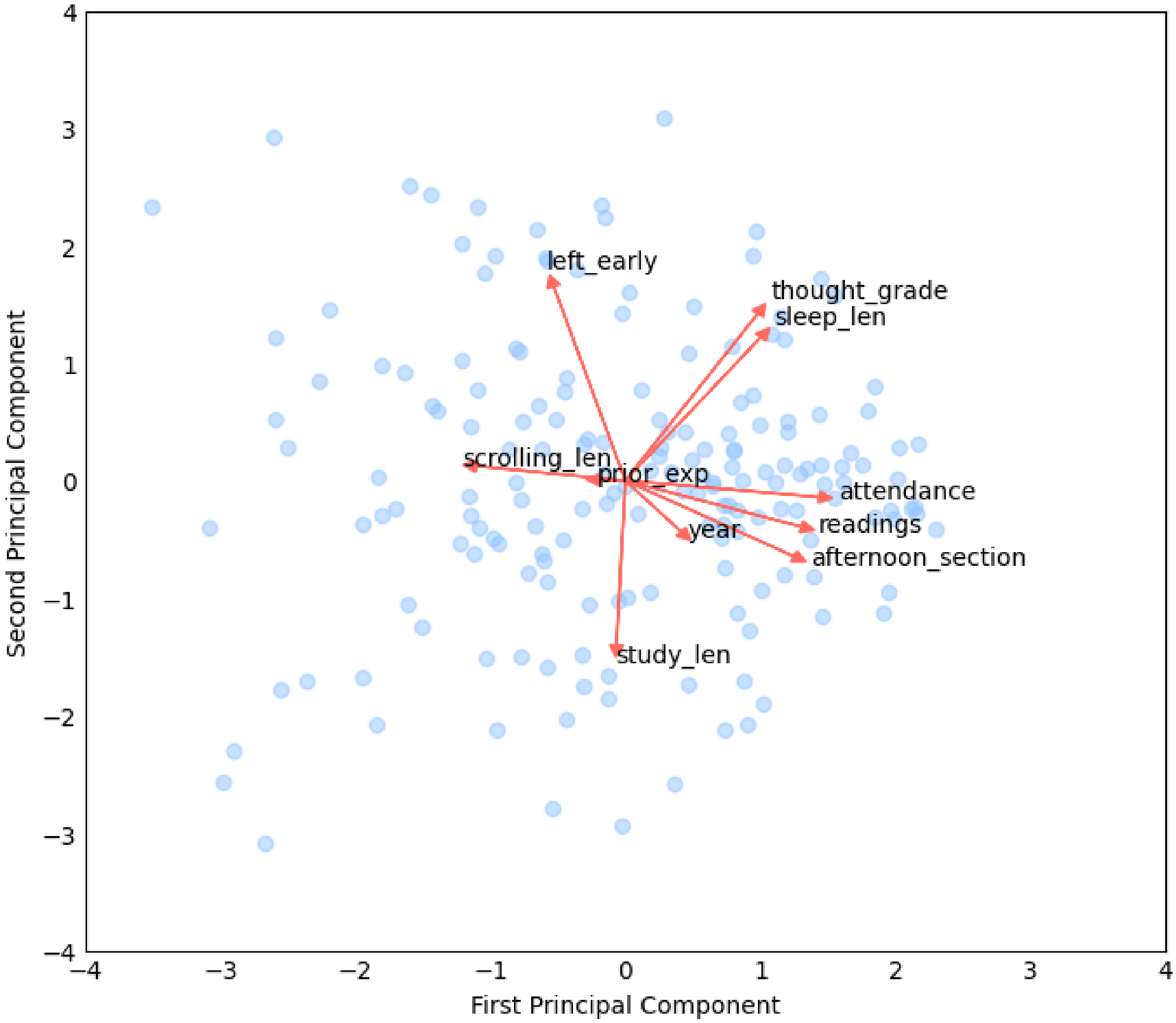

Figure 2. Survey Data Reduced to First Two Principal Components

Pass/Fail Classification

Observations were then binned into two categories: those who passed the exam (with a score a 70% or higher) and those who failed.

K-Means Analysis:

Before using supervised learning models to predict whether a student passed or failed, K-means clustering with $k = 2$ was performed on the data to investigate how separable the two classes were. Results are summarized in Figures 3 and 4 below.

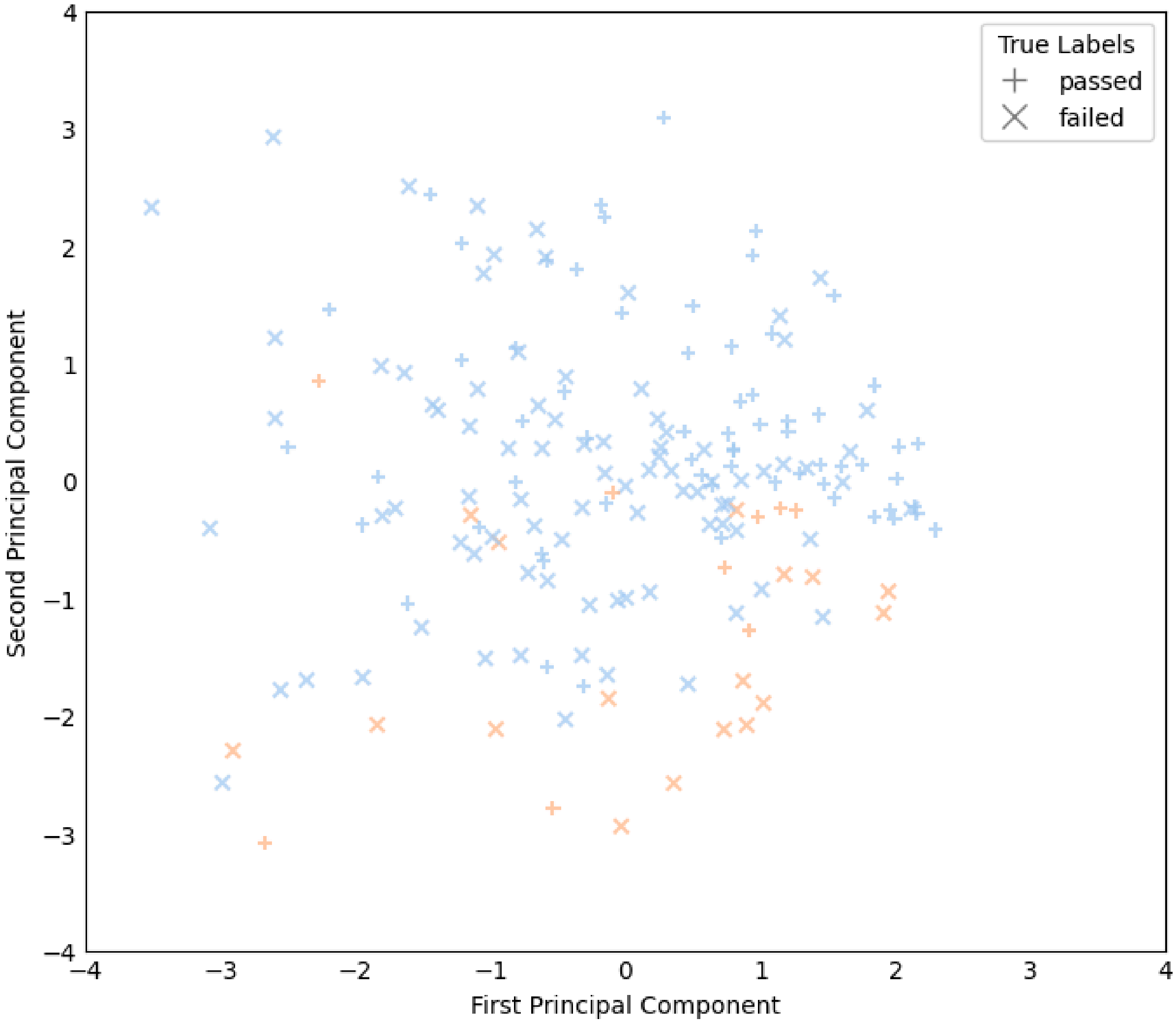

Figure 3. Data in Principal Component Space After K-Means Clustering with k = 2

In order to better communicate the purity of each cluster, clusters were optimally mapped to class labels so that k-means could be evaluated as a classifier, yielding a classification accuracy and confusion matrix. Both possible class label mappings were tested for classification accuracy, with the better of the two yielding an accuracy of 54.2%. A confusion matrix for this classifier can be found below in Figure 4.

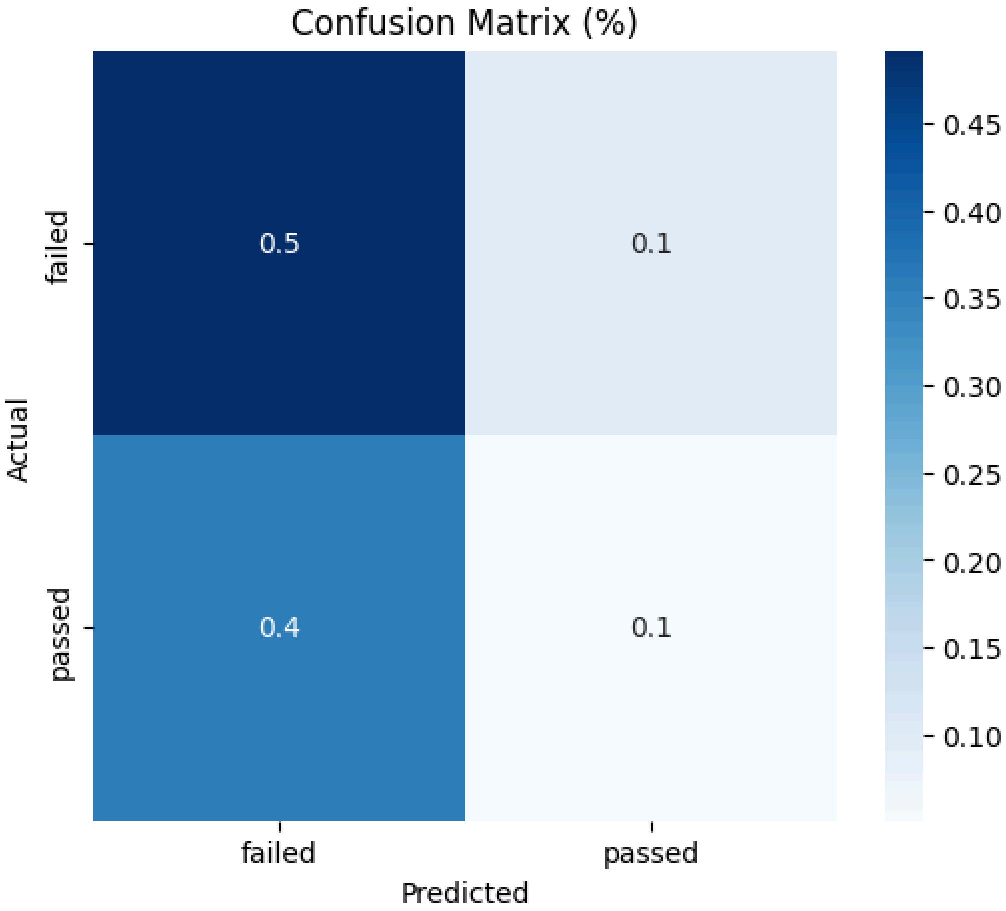

Figure 4. Confusion Matrix for Classification via K-Means Clustering with k = 2

Supervised Learning:

Next, supervised learning algorithms were used to classify the data. The classifiers used were KNN, Random Forest, and Logistic Regression. Classifiers were used with the best hyperparameters found from 10-fold cross validation for both original and PCA data. In all subsequent confusion matrices, the confusion matrix for original data is shown on the left and the confusion matrix for PCA data is shown on the right.

For KNN, the best values for *k* were found to be 13 for the original data and 17 for the PCA data. Mean accuracy across folds on the original data was 0.685 ± 0.09, while mean accuracy on the PCA data was 0.671 ± 0.09. Confusion matrices are below in Figure 5.

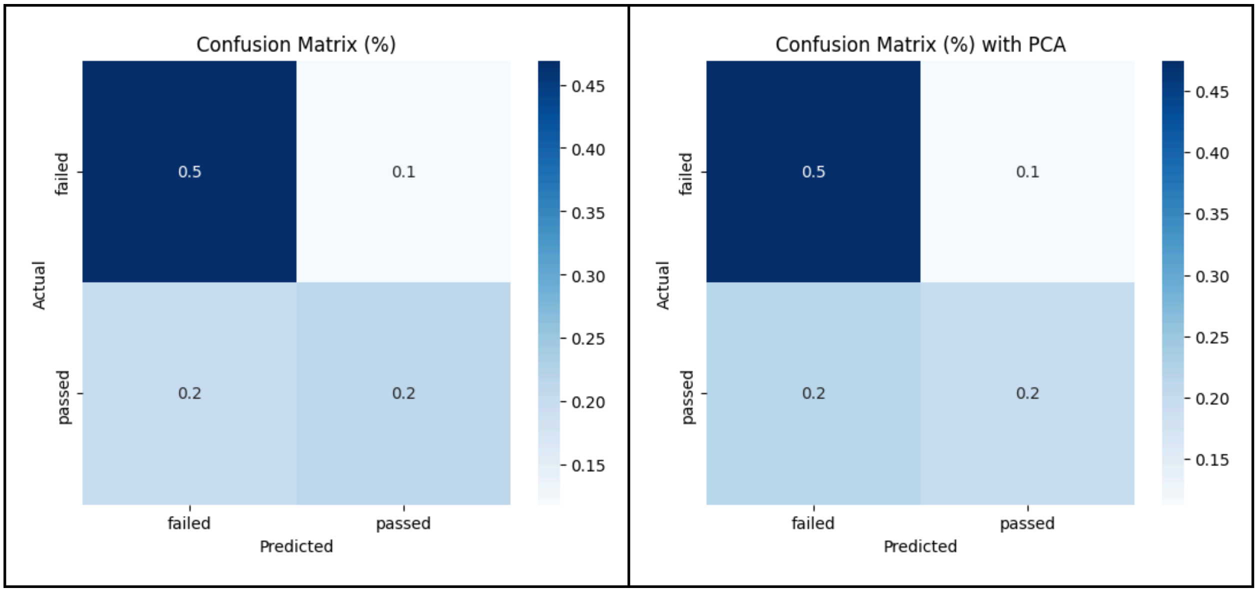

Figure 5. Confusion Matrix for Pass/Fail Classification on Original and PCA Data using KNN

For Random Forest classification, the best depth limits for its trees were found to be 3 splits for both the original and PCA data. Mean accuracy across folds on the original data was found to be 0.651 ± 0.09, while mean accuracy on the PCA data was found to be 0.610 ± 0.105. Confusion matrices are below in Figure 6.

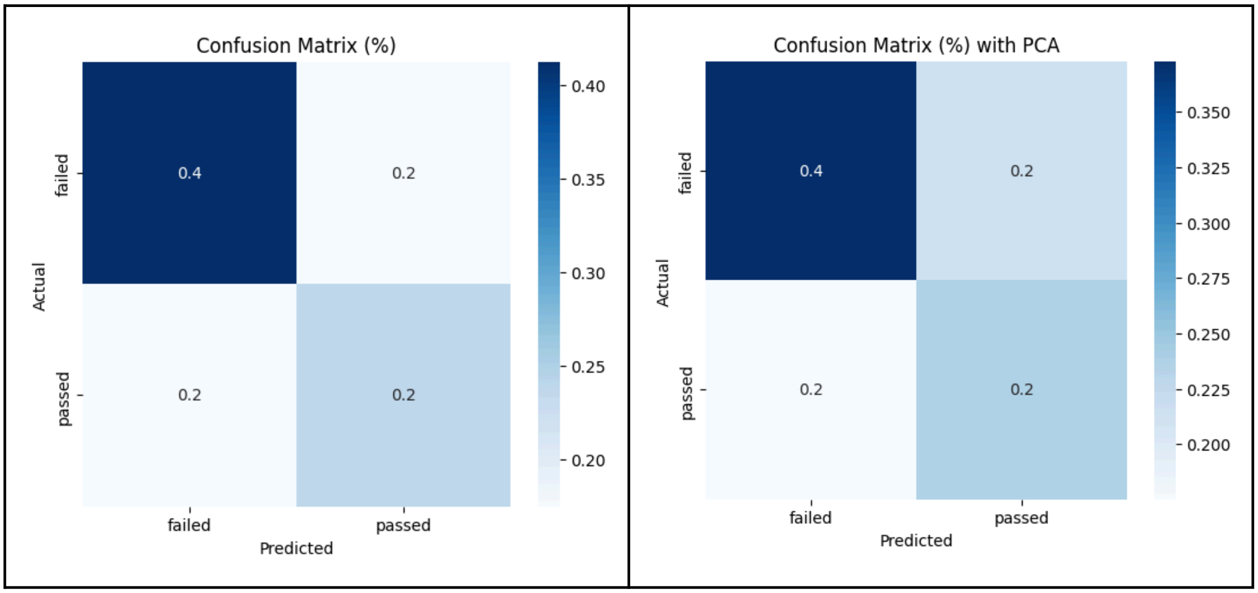

Figure 6. Confusion Matrix for Pass/Fail Classification on Original and PCA Data using Random Forest

For Logistic Regression classification, the best value for the C hyperparameter was found to be 1 for both the original and PCA data. Mean accuracy across folds on the original data was found to be 0.632 ± 0.136, while mean accuracy on the PCA data was found to be 0.633 ± 0.10. Confusion matrices are below in Figure 7.

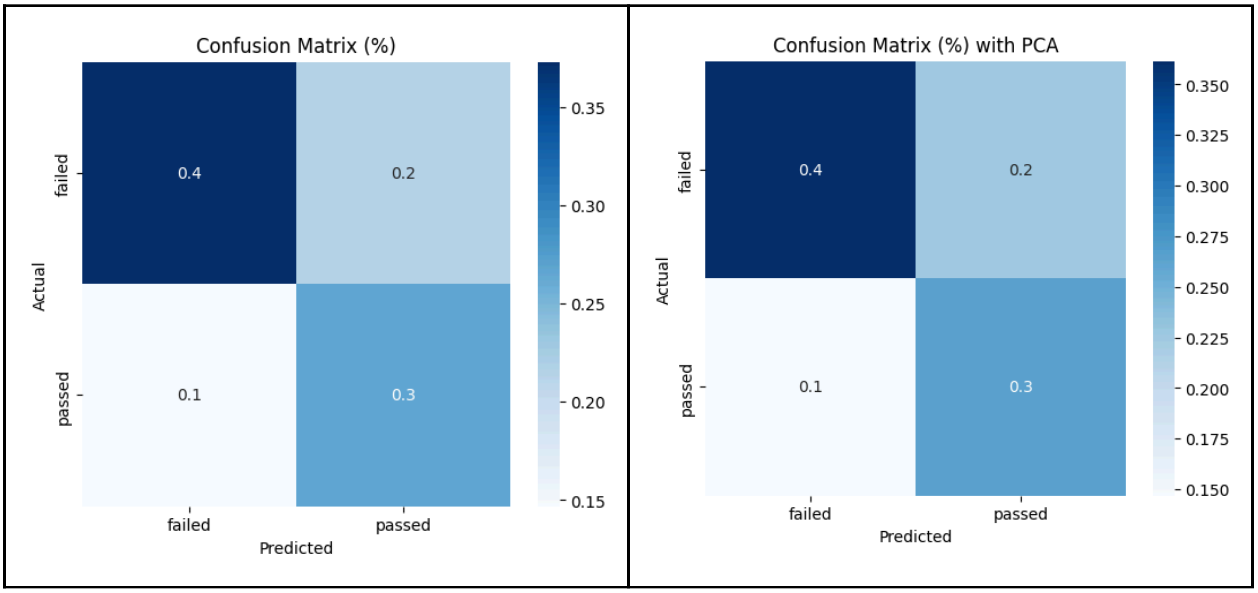

Figure 7. Confusion Matrix for Pass/Fail Classification on Original and PCA Data using Logistic Regression

Letter Grade Classification

Next, observations were binned into 5 categories, corresponding to the letter grade they received on the exam (i.e. 90-100 → A, 80-89 → B, etc.)

K-Means Analysis:

As before, in order to investigate separability of the data, k-means clustering was performed on the data, this time with *k* = 5. Results are summarized in Figures 8 and 9 below.

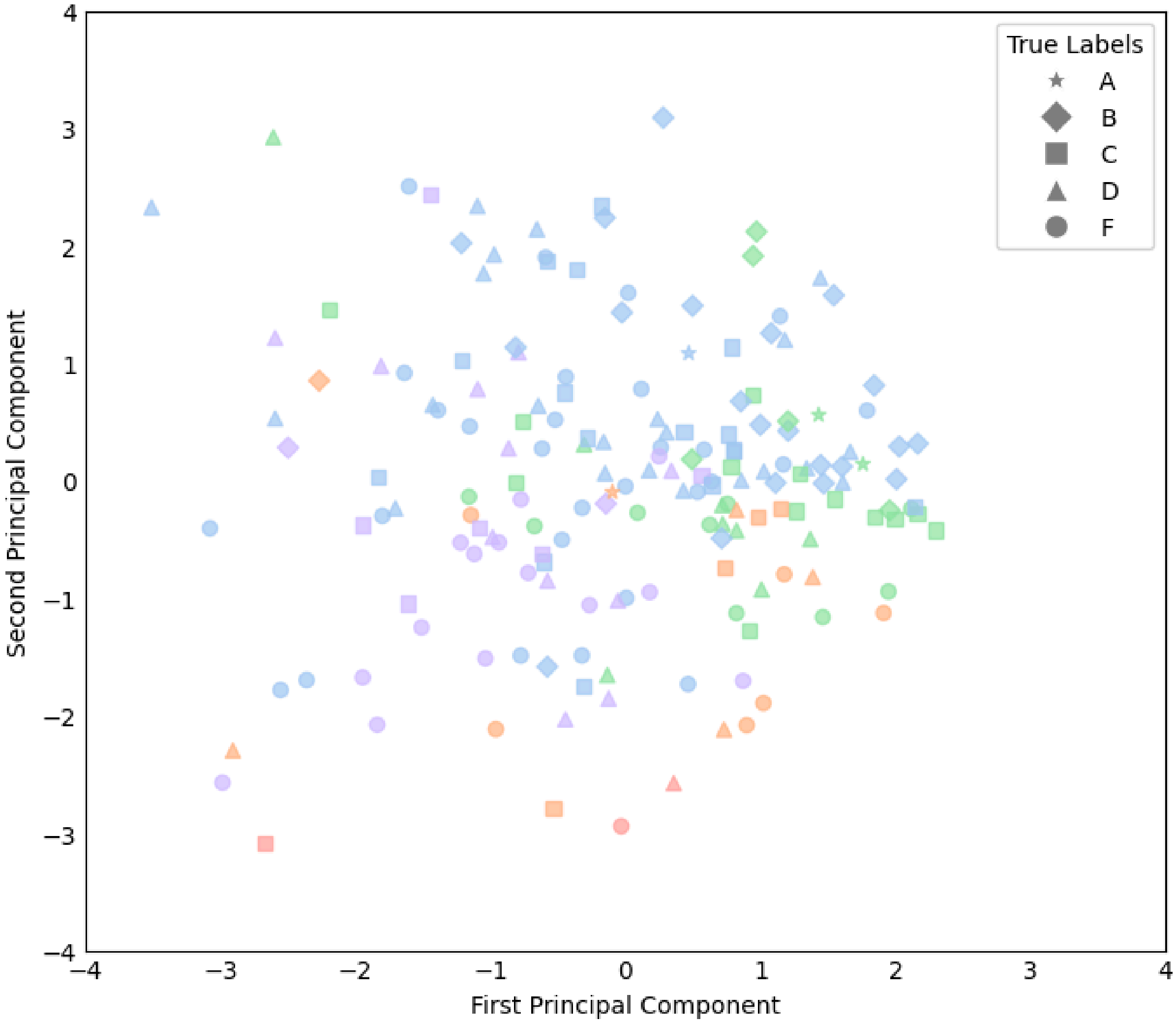

Figure 8. Data in Principal Component Space After K-Means Clustering with k = 5

Again, as before, k-means clusters assignments were optimally mapped to class labels to communicate the purity of each cluster. This was accomplished with scipy's linear_sum_assignment function, which finds the mapping of clusters to class labels that maximizes classification accuracy. However, this method only yielded an accuracy of 25.4%. A confusion matrix for this classifier can be found below in Figure 9.

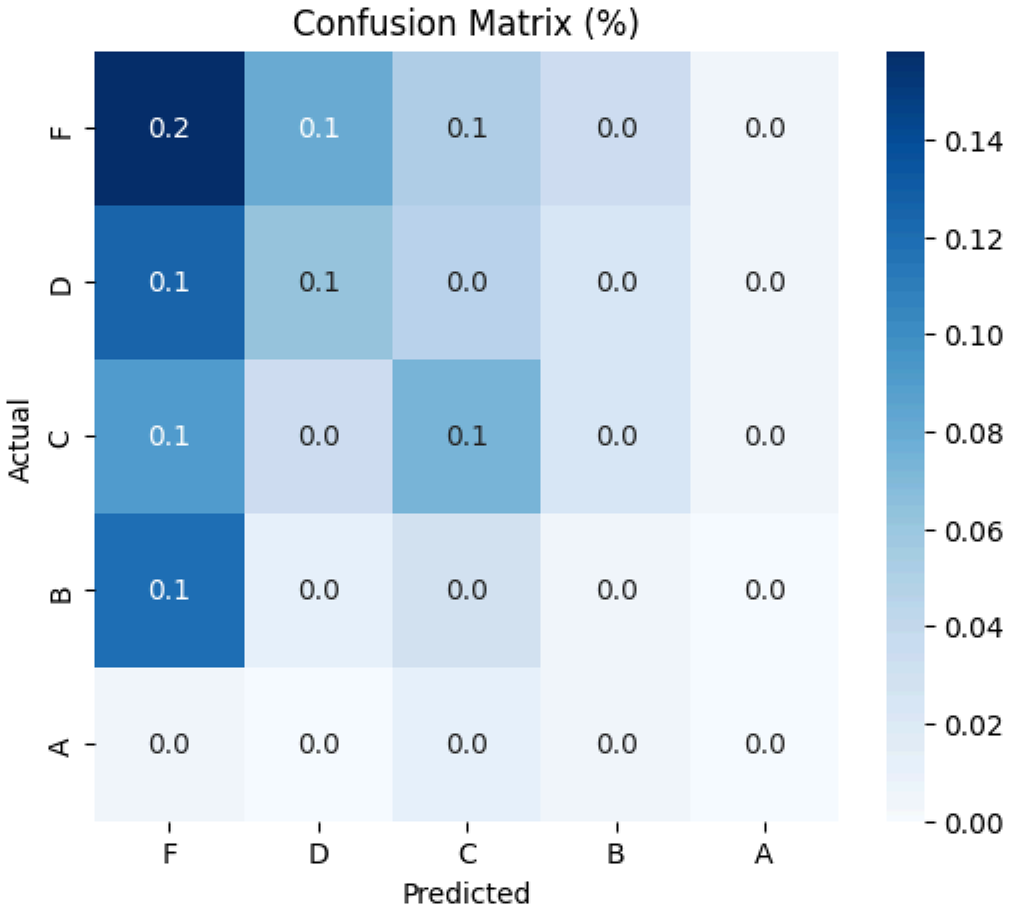

Figure 9. Confusion Matrix for Classification via K-Means Clustering with k = 5

Supervised Learning:

Next, supervised learning algorithms were used to classify the data. The classifiers used were KNN, Random Forest, and a Multilayer Perceptron (MLP). For the MLP, a fully connected network with two hidden layers of 32 and 16 neurons and a dropout rate of 0.3 was used. KNN and Random Forest were used with the best hyperparameters found from 10-fold cross validation for both original and PCA data. The MLP was re-initialized and trained for 50 epochs in each of the 10 cross-validation folds, and its performance was evaluated as the average accuracy across all folds. Again, in all subsequent confusion matrices, the confusion matrix for original data is shown on the left and the confusion matrix for PCA data is shown on the right.

For KNN, the best values for *k* were found to be 17 for the original data and 13 for the PCA data. Mean accuracy across folds on the original data was 0.333 ± 0.07, while mean accuracy on the PCA data was 0.401 ± 0.123. Confusion matrices are below in Figure 10.

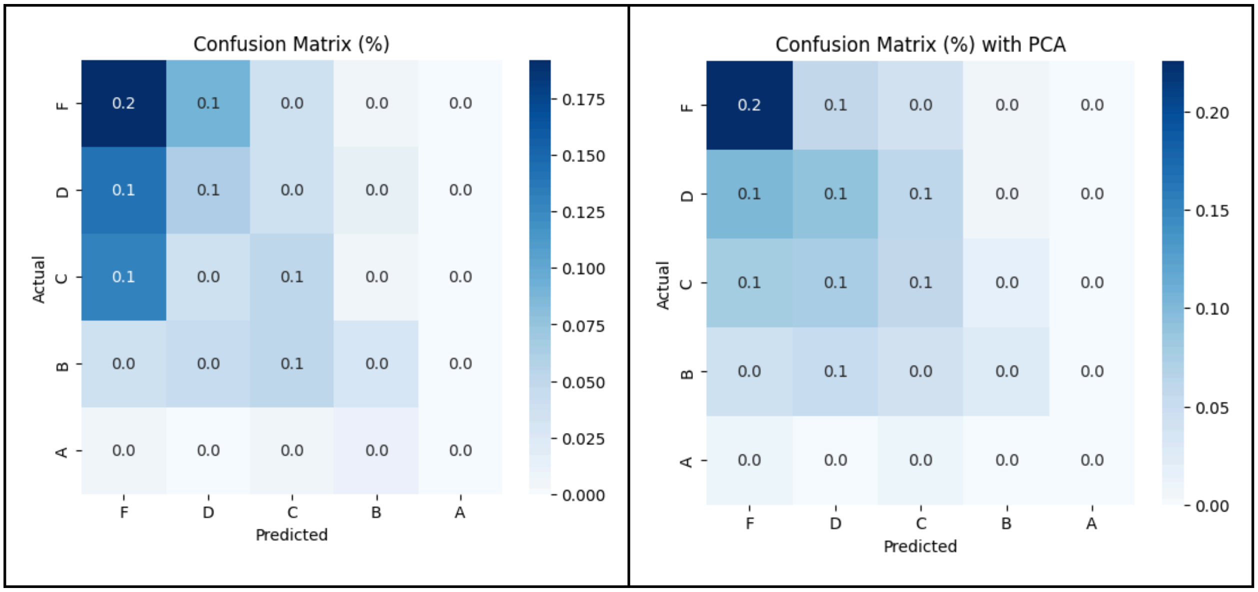

Figure 10. Confusion Matrix for Letter-Grade Classification on Original and PCA Data using KNN

For Random Forest classification, the best depth limit for its trees were found to be 3 splits for the original and 5 splits for the PCA data. Mean accuracy across folds on the original data was found to be 0.425 ± 0.07, while mean accuracy on the PCA data was found to be 0.306 ± 0.08. Confusion matrices are below in Figure 11.

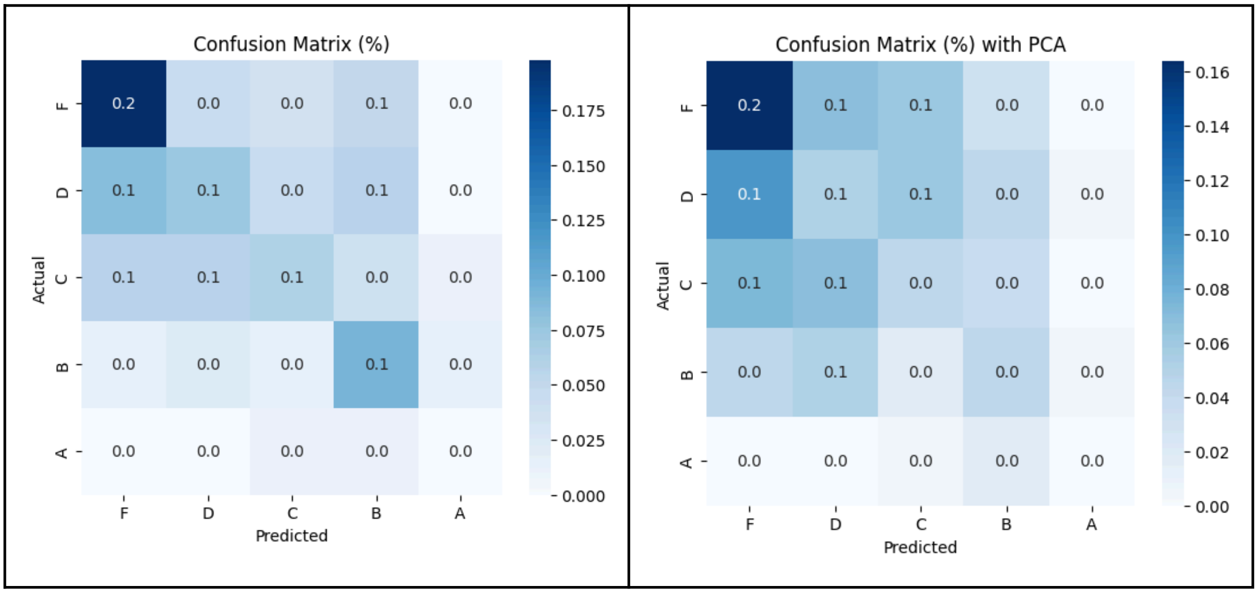

Figure 11. Confusion Matrix for Letter-Grade Classification on Original and PCA Data using Random Forest

The same MLP architecture was evaluated on both the original and PCA data. Mean accuracy across folds for the original data was 0.440 ± 0.140, while mean accuracy for the PCA data was 0.395 ± 0.130.

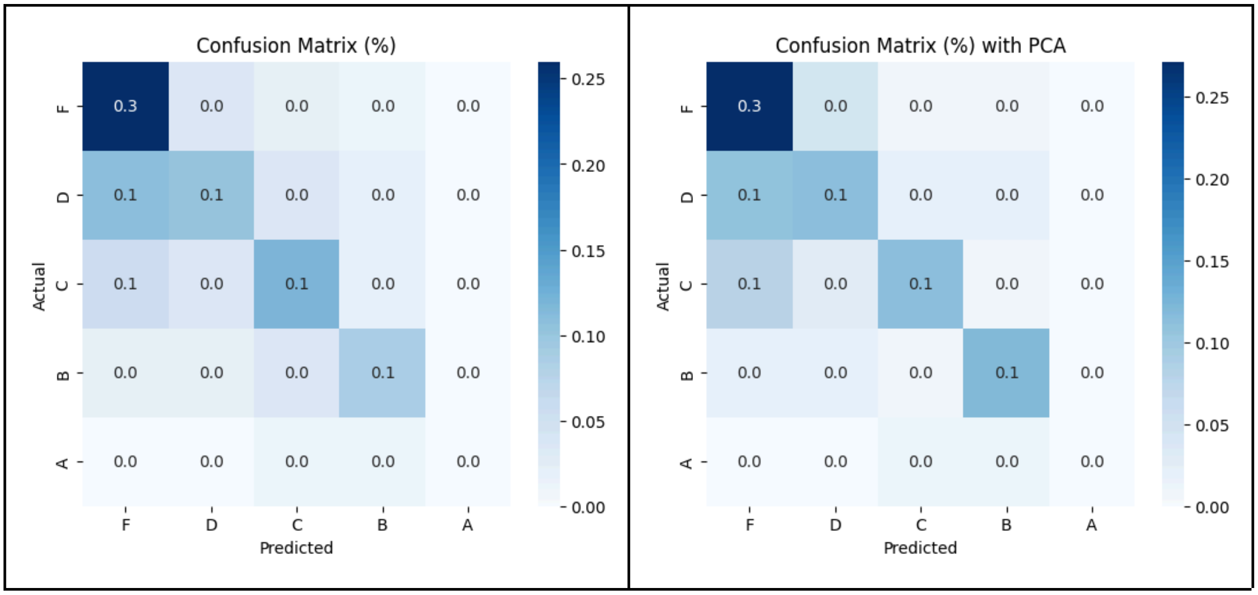

Figure 11. Confusion Matrix for Letter-Grade Classification on Original and PCA Data using MLP

Regression Analysis

Finally, regression was used to predict numerical grades. In order to reduce overfitting on noise from factors that were speculated to be irrelevant, lasso regression was used. The data was split into test and training sets, with 80% of the data being used for training. 10-fold cross validation was used on the training data to find the best value for alpha from 0.0001 to 10, which was found to be 0.471. Lasso regression with this alpha value was then fit on the entire training dataset, and its performance was evaluated on the test set. R^2 on the test set was 0.278, while MSE was 0.471. Coefficients of the model are given in the table below.

Variable	Coefficient
Intercept	42.562
attendance	0
prior_exp	0.571
afternoon_section	-1.059
study_len	0
year	0
readings	1.461
scrolling_len	0
sleep_len	2.236
left_early	0
thought_grade	3.728

Table 1. Coefficients of Lasso Regression

Discussion

Interpretation

Based on the results from the Lasso Regression, the most important predictor of the grade someone got was what grade they thought they got on the exam. Sleep duration the night before was the next most important predictor, then whether someone did the readings or not, then whether they attended the afternoon section or not, then whether they had prior experience with machine learning. All other factors were eliminated by lasso regression. Interestingly enough, the model implies that attending the afternoon section instead of the morning section had a negative effect on exam scores.

Performance

As can be seen, no model performed particularly well on the dataset. K-Means analysis revealed the data was not neatly separable, so it is not surprising that classifiers struggled to form decision boundaries between classes. This is likely due to variance in exam scores being explained by factors not captured in survey responses. Reduction to the most important principal components only improved the performance of two classifiers (predicting pass/fail with logistic regression and predicting letter grade with KNN), although given that the data was only reduced by one dimension, it is unsurprising that PCA did not improve classifier performance.

After dropping responses that failed the attention check, For pass/fail classification, 58% of responses were binned into the ‘failed’ category. With classifier performance in the 60s, this means that classifiers performed marginally better given the class imbalance. For letter grade classification, 32.8% of responses were given an ‘F’. Some models for classifying by letter grade had accuracy around or even lower than this baseline, which implies that they were not successful in making predictions in the data.

Lasso regression on the data had an R^2 value of 0.278, indicating a gap in its ability to model trends in the data. Thus, this calls into question the interpretation of the coefficients of the model given in the previous section. It is possible that the true trend in exam score is based on a non-linear function of these factors, or requires interaction terms, but such analysis is outside the scope of the author’s knowledge of exam score statistics.