Mini Project Report

ECE 20875 Fall 2024

bglavan@purdue.edu, tanouan@purdue.edu

Bogomir Glavan, Tatchi Anouan

bglavan, tanouan

Path 1

https://github.com/ECEDataScience/miniproject-f24-mglavan07.git

Student Performance Related to Video-Watching Behavior - Dataset Description

Introduction

This project aims to analyze data collected from students watching videos to conclude their performance on post-video quizzes. Specifically, information such as the fraction of the video a student watched, the number of pauses, rewinds, fast-forwards, and others were used to describe a student's behavior. The target variable in this study was the score, represented as a binary 0 or 1, which represented incorrect or correct responses at the end of the video, respectively.

The dataset contained 3,976 unique students who watched at least one of the 92 labeled videos. For each video-student pair, the dataset contained 10 different features, notated below.

Table 1: Dataset Features

Feature Name	Description
userID	A lengthy string representing the student to whom the data belongs. Students may not appear twice for the same video, but may or may not occur within other videos in the dataset.
videoID	An integer 0-92 represents the video to which data belongs. Multiple students may belong to the same video. Note that there is no video #29.
fracSpent	Fraction of time the student spent on the video in relation to the duration of the video. This includes time for pauses, rewinds, skips, etc.
fracComp	The fraction of the video the student watched relative to the length of the video, excluding rewinds.
fracPlayed	The fraction of the video the student watched relative to the length of the video, including rewinds.
fracPaused	The fraction of the video the student paused relative to the length of the video.
numPauses	The number of times the student paused the video
avgPBR	The average playback rate by the student on the video
stdPBR	The standard deviation for the playback rate by the student on the video
numRWs	The number of rewinds on the video by the student
numFFs	The number of fast-forwards on the video by the student
s	The score on the post-video quiz.

Legend: Induced in Feature Matrix, Excluded from Feature Matrix, Target Variable or Index Label

In the following report, the methodology and iterations used to build effective algorithms for clustering, regression, and classification are discussed. For each objective, the team's process will be described to understand what worked, and what didn’t, followed by a discussion of possible explanations for these results with measures of model accuracy or error.

Organizing the dataset

Before beginning clustering, the data was cleaned to effectively express the data the team intended to use. This was accomplished by sorting the data into a dictionary by the key of VideoID, with each pair being all the students, and their associated data, that watched to it. Because the project deliverables specified that students belonging to less than 5 videos should be excluded, a key search was done to remove video-student entries that had a student ID appear less than 5 times in the dataset. Finally, using the 7 features outlined above in Table 1, a feature matrix was constructed for each of the 92 videos, consisting of the valid numerical data within it. See figure 1 below.

$$X_K = \begin{bmatrix} fracSpent_1 & fracComp_1 & fracPaused_1 & numPauses_1 & avgPBR_1 & numRWs_1 & numFFs_1 \\ fracSpent_2 & fracComp_2 & fracPaused_2 & numPauses_2 & avgPBR_2 & numRWs_2 & numFFs_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ fracSpent_M & fracComp_M & fracPaused_M & numPauses_M & avgPBR_M & numRWs_M & numFFs_M \end{bmatrix}$$

Figure 1: Feature Matrix for M Students for Video K.

Clustering Students Within Each Video - Question 1

To cluster the data, the team was first faced with the decision of choosing an algorithm. As there are two main clustering algorithms, the team had to weigh the pros and cons of selecting K-means or Gaussian Mixture Models (GMMs) to cluster the data.

K-means clustering is centroid-based, which means it selects a point in Euclidean space and works by minimizing the distance between all points and their associated centers. This algorithm is effective by not requiring a prior statistical model to fit the data, however, its drawback is that it can only cluster simple data structures, where points are naturally close together.

On the other hand, GMMs are distribution-based and use probability distributions to fit data onto one of a set number of Gaussian curves. This requires having a model in mind that the data can fit around, often found by BIC scoring a number of Gaussians. However, this extra work comes with the benefit of revealing some more complex data structures that may be hidden by a simpler algorithm such as K-means.

When beginning to cluster the data, the team originally intended to use GMMs to reveal complex data structures. However, not only was the optimal value of k unknown at the beginning, but the team had no way of assuming the data could be modeled by k normal distributions. Therefore, after running algorithms to generate GMMs with a BIC score revealed through optimizing a BIC score, the team realized there was not much they could do with the model's results, as K-means would have a much more intuitive way of representing the "closeness" of the student-video data vectors.

Choosing to transition to K-means clustering, the team had to accomplish two essential steps before fitting the data: normalization and selecting a k value. For normalization, the team intended to eliminate the possible scaling bias introduced by a multi-feature dataset. As a feature such as fracSpent and fracComp are ratios concerning video length, lying typically between 0.5 and 1.1, there is little variance, and the feature is bounded. However, features such as numPauses and numFFs have extremely high variance, as some students never paused, and some students paused upwards of 20 times. Because some features are expressed as ratios whereas others are not, an unnormalized feature matrix would result in clustering algorithms that essentially only look at numFFs, numRWs, and numPauses, as from a numerical standpoint, the features expressed as ratios would all essentially have the same value. Therefore, by normalizing each feature by subtracting the feature mean and dividing it by the feature standard deviation from each data point, the data will be distributed in each feature normally, and no individual feature would have a scaling bias on the model's results. See figure 2 below.

$$X_K = \begin{bmatrix} (fracSpent_1 - \mu_1)/\sigma_1 & (fracComp_1 - \mu_2)/\sigma_2 & (fracPaused_1 - \mu_3)/\sigma_3 & (numPauses_1 - \mu_4)/\sigma_4 & (avgPBR_1 - \mu_5)/\sigma_5 & (numRWs_1 - \mu_6)/\sigma_6 & (numFFs_1 - \mu_7)/\sigma_7 \\ (fracSpent_2 - \mu_1)/\sigma_1 & (fracComp_2 - \mu_2)/\sigma_2 & (fracPaused_2 - \mu_3)/\sigma_3 & (numPauses_2 - \mu_4)/\sigma_4 & (avgPBR_2 - \mu_5)/\sigma_5 & (numRWs_2 - \mu_6)/\sigma_6 & (numFFs_2 - \mu_7)/\sigma_7 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ (fracSpent_M - \mu_1)/\sigma_1 & (fracComp_M - \mu_2)/\sigma_2 & (fracPaused_M - \mu_3)/\sigma_3 & (numPauses_M - \mu_4)/\sigma_4 & (avgPBR_M - \mu_5)/\sigma_5 & (numRWs_M - \mu_6)/\sigma_6 & (numFFs_M - \mu_7)/\sigma_7 \end{bmatrix}$$

Figure 2: Normalized Feature Matrix for M Students for Video K.

Next, for selecting a k-value, it is important to minimize the L-2 norm vector for net Euclidean distance for all clusters, however, we increasingly overfit the data as k approaches infinity. Therefore, it is advised that a k-value is selected in the "elbow" of the net distance vs k graph as shown below in Figure 3.

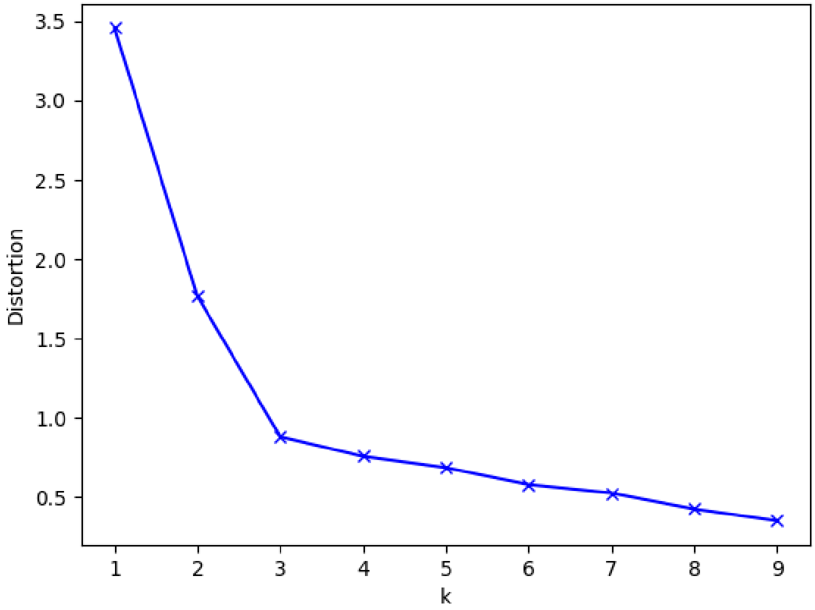

Figure 3: The Elbow Method for K-Means. Source: PythonProgrammingLanguage.com

However, when diagnosing this value of k for each team, it often corresponded with the same k value from the GMM modeling BIC score optimization. Therefore, when setting the k-value the team selected the k-value with the lowest BIC score, bounded between 1 and 8 to prevent overfitting. As any k-value over 8 is typically well beyond the “elbow” of typical k-means models, the k-values were set on bounds to prevent the BIC scores from clustering at extreme k-values all the time. When running the model, it was seen that the number of clusters often converged between 3 and 8 for most of the videos iterated through. This adaptive method of selecting a k-value was also critical to allow the analysis to be built for one video and iterated through all 92 videos.

Finally, with the data normalized and the k-value selected, the team was able to run K-means models from the Sklearn library. The results from the model in terms of raw output can be viewed in the Output section of this report.

As seen in the results, the total Euclidean distance for all points from their respective clusters between all models was 26,463 normalized units. Additionally, across all videos, there were a total of 655 clusters formed, as each video had between 3 and 8 clusters associated with it to group the 7 normalized features. This indicates that the average cluster will have a total distance of about 40.4 normalized units from itself to the sum of all its points. Given the size of the dataset and the fact that each cluster would typically have between 5 and 60 points associated, the groups can very naturally be clustered. As K-Means groups solely on an iterative algorithm of minimized Euclidean distances as its “error” a small average “error” indicates that the student-video data vectors within the dataset are naturally clumped together in Euclidean space.

While this does not necessarily reveal some complex statistical relationship as a GMM would, the performance of the team’s clustering algorithm reveals that the relationship in the data may be simplistic, as when normalized, there are groups of students who heavily interact with the video, and there are students that skip through most of it. As an individual student’s behavior is psychologically similar to its peers, it makes logical sense that a K-Means clustering algorithm would be naturally successful at grouping up students with common video-watching habits.

Predictive Regression for Total Student Performance - Question 2

First, for regression, the team decided on a model based on the inferences and known characteristics of the dataset. Given that the K-Means revealed that the data followed a simplistic interrelationship for Euclidean clustering, the team also decided it would be wise to select a relatively simple regression model, and not make any assumptions about the data trends unless the error for the first models was high.

With the goal of predicting average student performance across all quizzes, or an average student score for all the videos they watched, based on video-watching behavior, the team settled on the least squares equation form as shown below in Figure 4.

$$\hat{y} = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + b$$

Figure 4: Least Squares Equation for Predicting Average Student Performance

Note that, with a “b” in the equation, the team makes the assumption that the data will have some intercept. Therefore, the normalized feature matrix will take the form shown below in Figure 5.

$$X_K = \begin{bmatrix} (fracSpent_1 - \mu_1)/\sigma_1 & (fracComp_1 - \mu_2)/\sigma_2 & (fracPaused_1 - \mu_3)/\sigma_3 & (numPauses_1 - \mu_4)/\sigma_4 & (avgPBR_1 - \mu_5)/\sigma_5 & (numRWs_1 - \mu_6)/\sigma_6 & (numFFs_1 - \mu_7)/\sigma_7 & 1 \\ (fracSpent_2 - \mu_1)/\sigma_1 & (fracComp_2 - \mu_2)/\sigma_2 & (fracPaused_2 - \mu_3)/\sigma_3 & (numPauses_2 - \mu_4)/\sigma_4 & (avgPBR_2 - \mu_5)/\sigma_5 & (numRWs_2 - \mu_6)/\sigma_6 & (numFFs_2 - \mu_7)/\sigma_7 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (fracSpent_M - \mu_1)/\sigma_1 & (fracComp_M - \mu_2)/\sigma_2 & (fracPaused_M - \mu_3)/\sigma_3 & (numPauses_M - \mu_4)/\sigma_4 & (avgPBR_M - \mu_5)/\sigma_5 & (numRWs_M - \mu_6)/\sigma_6 & (numFFs_M - \mu_7)/\sigma_7 & 1 \end{bmatrix}$$

Figure 5: Normalized Feature Matrix for Linear Regression

Note that the above feature matrix in Figure 5 is also normalized as in the clustering model. Therefore, the team will simply be able to iterate back through the videos in the dictionary and reuse the feature matrix. Because the Linear Regression and Ridge Regression models in the Sklearn library add the column of ones to the feature matrix by default, the exact feature matrix from the clustering methodology can be passed to the Sklearn model methods, as these methods will convert the matrix to the form in figure 5 automatically.

Before passing the data through the Sklearn regression methods, the team considered and defined the measures for which they wished to determine the best models. The most common way of determining the best model is by minimizing the mean squared error (MSE) between the prediction and actual target variables for the data. However, this is not the only way to choose a model. Other measures such as R-squared quantify how much variance in the data can be explained by the model. Additionally, in a format named Ridge Regression, a form of regularization, an additional error parameter is added to minimize coefficient weights in addition to MSE, to prevent overfitting.

Wishing to prevent overfitting in the large number of models the team would generate by iterating through the videos, the team chose to select ridge regression to determine the form of the least squares equation in Figure 4. The formula for the error in a ridge regression model is displayed below in Figure 6.

$$e = \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$$

Figure 6: Formula for Error in Regularization Models such as Ridge Regression

Therefore the matrix form solution is shown below in figure 7.

$$\beta^* = (X^T X + \lambda I)^{-1} X^T y$$

Figure 7: Matrix Solution for Ridge Regression

As seen in Figures 6 and 7, ridge regression introduces a lambda parameter that requires optimization. To determine the optimal value of lambda without over- or underfitting the model, the team chose to plot MSE as a function of lambda for all videos combined as a sum and observe a value of lambda to use for the entire analysis. See the summed MSE vs lambda plot below in Figure 8.

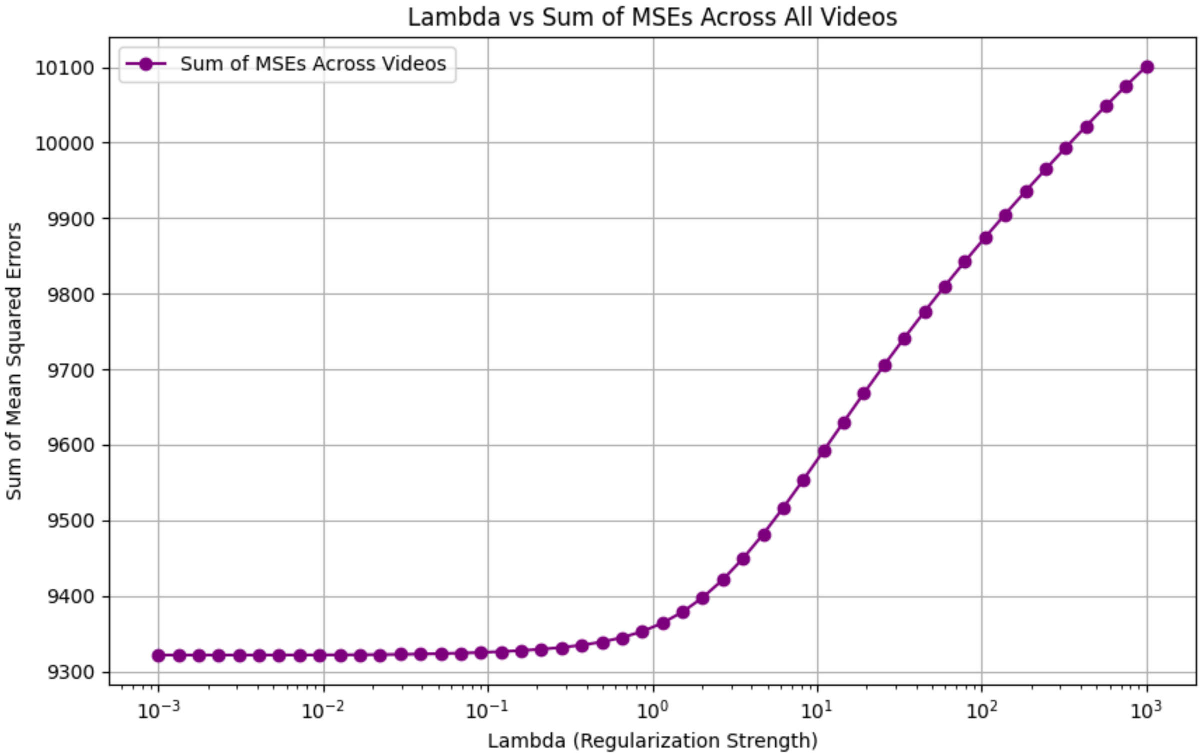

Figure 8: Summed MSE vs Lambda

As seen in the plot, the optimal value falls around 1, as any value beyond this approaching zero would favor an overfitted model, as Ridge would train with the assumption that MSE is a major driving factor in model error rather than feature weights. Therefore, for the analysis, the team chose to do ridge regression with a lambda value of 1 for the normalized data for the 7 features.

Lastly, before running the model, the team needed to compile the actual average scores for the students to make a y-vector for the “actual” values of the target variable. To do this, the team went back into the sorted dictionary where students of under five videos were watched, and for each student ID identified, iterated through all the videos to determine which videos that student watched. After locating the student ID in different videos, the score would be added to a counter, where, after the nested loop, would be divided by the total count of videos watched to get an average score, falling between 0, all incorrect, and 1, all correct.

After all of the pre-analysis steps were completed to ensure that the models would have useful and unbiased least squares equations, 92 models were trained, one for each video, to predict a student’s average score across all videos based on their behavior on one specific video. The raw output can be seen in the Output section of this document.

Summarizing the output, the MSE for the models was very low, indicating that the least squares equations were very effective at predicting the students' average performance. Additionally, looking at the normalized model coefficients, there was great variation from video to video in terms of the value of the intercept and absolute value of the weights, indicating that some videos may be more critical for student success than others, for example, if a student where to watch a certain video all the way through, they would be increasingly likely to succeed on later video quizzes, thus increasing their average score. Additionally, some coefficients were positive or negative across the vast majority of models, indicating boosters or drawbacks to student performance. For example, a6 was usually positive and a7 was usually negative, which indicates students rewinding more often get better scores more often, and students fast forwarding more often will perform worse more often.

Lastly, the team decided to implement additional features to the model for the cluster a data vector belongs to, as determined earlier in the analysis. To do this, the feature matrix was expanded, and the additional features were passed as the location of the center of the cluster for each data vector. This process was accomplished using Numpy's h-stack method to concatenate a new feature matrix before iterating through the videos again. Results for these new models can be seen in the Output section, and as the MSE shows, on average, there is a slight improvement in model accuracy when the extra features from the centroid location are added to the model.

In conclusion, the team was successful in predicting a student's average score across all quizzes based on a data vector for behavior on an individual video. With MSE values around 0.1 for the videos, this indicates that the team's model can, on average, get within 10 percentage points of the student's true quiz average by looking at one video. When adding clusters, this MSE drops noticeably, but not entirety significantly, usually by 0.01 for most videos. While the addition of cluster centroids as features improved the accuracy of the ridge regression model, this process makes the model more overfit to the training data. As clustering was performed on the same data that the model was trained on, the model is essentially able to "memorize" how students "should" perform on each video, or in other words, use the cluster location to look less at student behavior and more at trends in the training data. As there are no guarantees these same trends exist in testing data, this makes the model overfit. Further, while the moderately high regularization parameter chosen by the team works to combat this overfitting, adding clusters would be a step in the wrong direction, even though the team experienced a drop in MSE with additional features.

Classification for Individual Video Quizzes - Question 3

Lastly, the team had the goal of completing a simple classification problem in attempting to predict a student's score based on their video-watching behaviors for the corresponding video. Following the theme suggested by the success of the basic K-Means clustering model and the simple Ridge Regression model, the team chose Naive Bayes as a simple algorithm for clustering over more complex methodologies such as logistic regression or neural networks.

Additionally, since the target variable, the score was either a 0 or 1, the model could be simplified further into a binary classifier.

To begin with Naive Bayes, the team had the opportunity of passing "priors" to the model to place a bias on whether the model should predict correctness over incorrectness, or vice versa. After iterating through all the data points in the unfiltered dictionary, or, using all data vectors regardless of how many times the student ID appeared, there was a near-even split between 0 and 1 in the dataset, therefore, the team did not use any priors to improve classification accuracy.

More specifically, the team chose to generate a Gaussian Naive Bayes model with the Sklearn library using data from all video-student vectors in the 7 features used in clustering and regression. Therefore, with the multivariate Gaussian model, the team assumes that the model fits the requirements of the Gaussian Naive Bayes model. That is, features are all independent and the mean and variance for each feature are fit independently. The interpretation of this is that a student's behavior in one feature, say pausing, does not directly impact their behavior in another feature, say fast forwarding. In the situation of the behavior data, this is partially, but mostly, true.

Therefore, the Naive Bayes model will fit each data point on the Gaussian fit for the class, one class for correct, and one class for incorrect, according to the following formula, shown below in Figure 9.

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{(2\pi)^d\,|\sigma|}}e^{-\frac{1}{2}(x-\mu)^T\sigma^{-1}(x-\mu)}$$

Figure 9: Formula for Naive Bayes Gaussian Distribution, *where σ is the d x d covariance matrix with determinant $|\sigma|$*

Lastly, before running the model, the team normalized the data as done with clustering and regression, and split the data into training and testing portions. The team experimented with several train-test splits but ultimately found that using 80% of the data for training provided the best accuracy, and also avoided leaving the model with a small test dataset. Furthermore, since the dataset is so large, the team is comfortable with a high portion of training data, as there will still be a large amount of test data even at a 20% stake. Note that all random states for this project were set to 0, in the case of replication.

After running the Naive Bayes model, the team printed out the model accuracy for each video, as well as a cumulative confusion matrix for all of the models combined. Raw output is displayed in the Output section. For the Naive Bayes model, accuracies varied between 31% and 92% between videos, with a cumulative average accuracy of 53%. As this is only 3% better than theoretically guessing randomly each time, the team looked deeper into the evaluation metrics and model parameters to see if anything could change the model accuracy.

Firstly, the team noticed that some of the variances in the features were very close to 0, which, revisiting the formula from Figure 9, causes some numerical instability as we approach a divide by 0 error, and therefore, experience some odd asymptotic behaviors. To combat this, the team worked with the var_smoothing parameter to adjust the covariances by a small amount to “shift” the Gaussians for correct answers and incorrect answers to where they should be. Using values from the default 10^-9 to 10^-2, the team did not see any significant improvement in model accuracy.

Alternatively, the team decided to increase the complexity of a logistic classifier. Logistic classifiers work mathematically similar to Naive Bayes, the major difference is that it uses an exponential function with a set of feature weights, similar to what the team computed with linear regression, to place a data vector at 0 or 1. Logistic classifiers would use gradient descent to maximize this coefficient matrix, and therefore, possibly result in a classifier with different results than with Naive Bayes.

Running this logistic model with the same data yielded very similar results, however. The model accuracy was 67%. This is marginally better than Naive Bayes, however, the feature matrix shown below in Figure 10 partially explains the classifier results.

$$\begin{bmatrix}1 & 0.67\\ 0 & 0.5\end{bmatrix}$$

Figure 10: Confusion Matrix

As seen in Figure 10, the primary, and surprisingly only, error was the model predicting the student would be incorrect, when in fact the student was right. As this is a Type II error, the model is subject to not being sensitive enough when predicting a student’s performance. However, when the team attempted to introduce increased sensitivity in the Sklearn method call, there was no noticeable increase in accuracy, instead, accuracy decreased by as much as 4%.

In conclusion, with the dataset, the team deduced that it cannot classify a student’s performance on a single post-video quiz solely on their behavior for the corresponding video. While accuracy scores were achieved over 50%, they are still marginally better than a random guess and the team cannot statistically justify that the model presents any clear advantage over this 50-50 guess. Even when introducing prior biases, variance smoothing, train-test splits, and even switching to logistic regression, the team’s ultimate conclusion did not change. When considering these results in a contextual sense, sometimes a student’s performance on a post-video quiz can be the opposite of what their behaviors suggest. In other words, even if a student watches a video several times, they can still get the only question wrong, and even a student who skips through the video may get the question right. This confounding variable of student luck, whether fortunate or unfortunate, makes it very hard for the classifier to make single-quiz predictions, however, this explains why the regression model was effective over the average of all the quizzes the student attempted, as this eliminates the hidden feature of “luck.”

Output

Classification Output

```
Gaussian Naive Bayes model accuracy(in %) for video 0: 31.88405797101449
Gaussian Naive Bayes model accuracy(in %) for video 1: 69.42675159235668
Gaussian Naive Bayes model accuracy(in %) for video 2: 34.96115427302996
Gaussian Naive Bayes model accuracy(in %) for video 3: 81.37931034482759
Gaussian Naive Bayes model accuracy(in %) for video 4: 53.41981132075472
Gaussian Naive Bayes model accuracy(in %) for video 5: 62.797619047619044
Gaussian Naive Bayes model accuracy(in %) for video 6: 35.389988358556465
Gaussian Naive Bayes model accuracy(in %) for video 7: 35.60794044665012
Gaussian Naive Bayes model accuracy(in %) for video 8: 59.77011494252874
Gaussian Naive Bayes model accuracy(in %) for video 9: 58.00653594771242
Gaussian Naive Bayes model accuracy(in %) for video 10: 36.43790849673202
Gaussian Naive Bayes model accuracy(in %) for video 11: 36.12662942271881
Gaussian Naive Bayes model accuracy(in %) for video 12: 43.111111111111114
Gaussian Naive Bayes model accuracy(in %) for video 13: 47.33178654292343
Gaussian Naive Bayes model accuracy(in %) for video 14: 58.415841584158414
Gaussian Naive Bayes model accuracy(in %) for video 15: 56.127450980392155
Gaussian Naive Bayes model accuracy(in %) for video 16: 41.57608695652174
Gaussian Naive Bayes model accuracy(in %) for video 17: 41.19318181818182
Gaussian Naive Bayes model accuracy(in %) for video 18: 43.233082706766915
Gaussian Naive Bayes model accuracy(in %) for video 19: 52.53456221198156
Gaussian Naive Bayes model accuracy(in %) for video 20: 52.27272727272727
Gaussian Naive Bayes model accuracy(in %) for video 21: 37.919463087248324
Gaussian Naive Bayes model accuracy(in %) for video 22: 56.38297872340425
Gaussian Naive Bayes model accuracy(in %) for video 23: 60.416666666666664
Gaussian Naive Bayes model accuracy(in %) for video 24: 45.535714285714285
Gaussian Naive Bayes model accuracy(in %) for video 25: 46.464646464646464
Gaussian Naive Bayes model accuracy(in %) for video 26: 92.27467811158799
Gaussian Naive Bayes model accuracy(in %) for video 27: 50.51020408163265
Gaussian Naive Bayes model accuracy(in %) for video 28: 50.0
Gaussian Naive Bayes model accuracy(in %) for video 30: 60.0
Gaussian Naive Bayes model accuracy(in %) for video 31: 52.27272727272727
Gaussian Naive Bayes model accuracy(in %) for video 32: 47.80487804878049
Gaussian Naive Bayes model accuracy(in %) for video 33: 36.60130718954248
Gaussian Naive Bayes model accuracy(in %) for video 34: 50.69444444444444
Gaussian Naive Bayes model accuracy(in %) for video 35: 34.35114503816794
Gaussian Naive Bayes model accuracy(in %) for video 36: 33.939393939393945
Gaussian Naive Bayes model accuracy(in %) for video 37: 54.961832061068705
Gaussian Naive Bayes model accuracy(in %) for video 38: 47.32142857142857
Gaussian Naive Bayes model accuracy(in %) for video 39: 47.78761061946903
Gaussian Naive Bayes model accuracy(in %) for video 40: 42.0
Gaussian Naive Bayes model accuracy(in %) for video 41: 48.46153846153846
Gaussian Naive Bayes model accuracy(in %) for video 42: 53.44827586206896
Gaussian Naive Bayes model accuracy(in %) for video 43: 49.45054945054945
Gaussian Naive Bayes model accuracy(in %) for video 44: 56.470588235294116
Gaussian Naive Bayes model accuracy(in %) for video 45: 54.63917525773196
Gaussian Naive Bayes model accuracy(in %) for video 46: 44.57831325301205
Gaussian Naive Bayes model accuracy(in %) for video 47: 50.72463768115942
Gaussian Naive Bayes model accuracy(in %) for video 48: 53.62318840579711
Gaussian Naive Bayes model accuracy(in %) for video 49: 43.58974358974359
Gaussian Naive Bayes model accuracy(in %) for video 50: 39.130434782608695
Gaussian Naive Bayes model accuracy(in %) for video 51: 60.0
Gaussian Naive Bayes model accuracy(in %) for video 52: 47.5609756097561
Gaussian Naive Bayes model accuracy(in %) for video 53: 51.35135135135135
Gaussian Naive Bayes model accuracy(in %) for video 54: 64.04494382022472
Gaussian Naive Bayes model accuracy(in %) for video 55: 61.36363636363637
Gaussian Naive Bayes model accuracy(in %) for video 56: 51.28205128205128
Gaussian Naive Bayes model accuracy(in %) for video 57: 68.33333333333333
Gaussian Naive Bayes model accuracy(in %) for video 58: 55.140186915887845
Gaussian Naive Bayes model accuracy(in %) for video 59: 50.0
Gaussian Naive Bayes model accuracy(in %) for video 60: 52.80898876404494
Gaussian Naive Bayes model accuracy(in %) for video 61: 56.57894736842105
Gaussian Naive Bayes model accuracy(in %) for video 62: 50.0
Gaussian Naive Bayes model accuracy(in %) for video 63: 65.27777777777779
Gaussian Naive Bayes model accuracy(in %) for video 64: 41.17647058823529
Gaussian Naive Bayes model accuracy(in %) for video 65: 53.57142857142857
Gaussian Naive Bayes model accuracy(in %) for video 66: 40.816326530612244
Gaussian Naive Bayes model accuracy(in %) for video 67: 36.53846153846153
Gaussian Naive Bayes model accuracy(in %) for video 68: 51.13636363636363
Gaussian Naive Bayes model accuracy(in %) for video 69: 43.42105263157895
Gaussian Naive Bayes model accuracy(in %) for video 70: 62.68656716417911
Gaussian Naive Bayes model accuracy(in %) for video 71: 77.77777777777779
Gaussian Naive Bayes model accuracy(in %) for video 72: 65.97938144329896
Gaussian Naive Bayes model accuracy(in %) for video 73: 57.89473684210527
Gaussian Naive Bayes model accuracy(in %) for video 74: 48.19277108433735
Gaussian Naive Bayes model accuracy(in %) for video 75: 53.96825396825397
Gaussian Naive Bayes model accuracy(in %) for video 76: 67.1875
Gaussian Naive Bayes model accuracy(in %) for video 77: 88.0
Gaussian Naive Bayes model accuracy(in %) for video 78: 55.223880597014926
Gaussian Naive Bayes model accuracy(in %) for video 79: 63.1578947368421
Gaussian Naive Bayes model accuracy(in %) for video 80: 50.98039215686274
Gaussian Naive Bayes model accuracy(in %) for video 81: 67.74193548387096
Gaussian Naive Bayes model accuracy(in %) for video 82: 55.35714285714286
Gaussian Naive Bayes model accuracy(in %) for video 83: 88.46153846153845
Gaussian Naive Bayes model accuracy(in %) for video 84: 45.83333333333333
Gaussian Naive Bayes model accuracy(in %) for video 85: 62.5
Gaussian Naive Bayes model accuracy(in %) for video 86: 51.92307692307693
Gaussian Naive Bayes model accuracy(in %) for video 87: 54.166666666666664
Gaussian Naive Bayes model accuracy(in %) for video 88: 80.85106382978722
Gaussian Naive Bayes model accuracy(in %) for video 89: 61.224489795918366
Gaussian Naive Bayes model accuracy(in %) for video 90: 57.49999999999999
Gaussian Naive Bayes model accuracy(in %) for video 91: 42.5
Gaussian Naive Bayes model accuracy(in %) for video 92: 69.04761904761905
average accuracy is 53.270842773777%
```

```
Video ID: 0
  Normalized Coefficients without Clusters: [ 8.45872403e-06  2.15377520e-02 -2.20085823e-05
4.78421826e-03
  4.72957952e-02 -8.61358786e-04  1.60680966e-03], [0.6630155795008189]
Video ID: 1
  Normalized Coefficients without Clusters: [ 1.58660988e-05 -3.06902252e-02 -5.13300262e-06
-1.05579419e-03
  6.38296045e-02  8.75967567e-04  7.69136441e-03], [0.6593317870893912]
Video ID: 2
  Normalized Coefficients without Clusters: [-4.91543424e-05 -1.28168383e-01  1.92385295e-05
3.44454418e-03
  7.33465154e-02  3.29333980e-03 -6.21414594e-03], [0.8029628208191171]
Video ID: 3
  Normalized Coefficients without Clusters: [-9.99350848e-05  7.09576960e-03 -3.49796230e-05
1.02579228e-05
  1.64402639e-02  5.08770712e-03  2.39042806e-04], [0.7996668459904834]
Video ID: 4
  Normalized Coefficients without Clusters: [-1.82737304e-05  6.43860011e-02 -2.59234638e-05
4.09207899e-03
  5.41693020e-02 -2.69906783e-04  3.40768371e-03], [0.37743652699746927]
Video ID: 5
  Normalized Coefficients without Clusters: [-3.75950048e-04 -3.22523768e-02  3.01644895e-06
3.93215778e-03
  1.24818661e-01  1.93147630e-03  3.00402198e-03], [0.49208069016609757]
Video ID: 6
  Normalized Coefficients without Clusters: [ 3.50032783e-05  4.37836857e-02  3.60406535e-05
-1.05573901e-03
  1.32747213e-01  1.86398584e-03  1.06922798e-03], [0.6763721310026066]
Video ID: 7
  Normalized Coefficients without Clusters: [-5.00669280e-05 -4.90595552e-02 -6.13344668e-05
3.10806173e-03
  1.75751606e-01 -4.34531763e-04  5.06251296e-03], [0.4967542473959208]
Video ID: 8
  Normalized Coefficients without Clusters: [ 9.66607241e-05 -4.92900636e-02  3.40418808e-05
-1.66885233e-03
  8.93166196e-02  1.93059764e-03 -5.55838639e-04], [0.5526294173843073]
Video ID: 9
  Normalized Coefficients without Clusters: [ 8.46306979e-05  1.26085385e-01  2.27001787e-05
-3.23996800e-04
 -2.18984334e-02  4.75938926e-03 -3.74199253e-03], [0.6028692840824222]
Video ID: 10
  Normalized Coefficients without Clusters: [-5.03609832e-05 -1.44565737e-02  5.21267818e-05
1.41525577e-02
  1.61032832e-01  1.14971278e-02 -8.31965507e-03], [0.4673647428647184]
Video ID: 11
  Normalized Coefficients without Clusters: [-2.52951520e-04 -7.80302068e-03  4.46471948e-05
1.82279277e-04
```

```
 1.05645731e-01  1.55508649e-04  6.13753168e-05], [0.7239710672124213]
Video ID: 12
 Normalized Coefficients without Clusters: [-1.11600451e-04 -1.30193183e-02 -4.29741137e-06
4.24177630e-03
 9.60890110e-02 -2.33580562e-03  4.01482355e-03], [0.5350129882657404]
Video ID: 13
 Normalized Coefficients without Clusters: [ 1.57884462e-05 -1.24780386e-01  2.16984062e-05
-4.26323654e-03
 1.56064196e-01  1.06342965e-02 -1.33791993e-02], [0.5001445814290284]
Video ID: 14
 Normalized Coefficients without Clusters: [-4.27956180e-05 -1.72532635e-02  2.90115528e-05
3.40769297e-04
 9.63600373e-02  8.05984459e-05 -4.35074049e-04], [0.4949069896853139]
Video ID: 15
 Normalized Coefficients without Clusters: [-2.47443900e-05  5.39399539e-02  4.78134712e-05
-8.17255240e-03
 8.41051644e-02  1.46670950e-02 -1.13310880e-02], [0.4507754893390886]
Video ID: 16
 Normalized Coefficients without Clusters: [ 9.03943561e-05 -8.07784324e-03 -3.89328154e-05
2.78547854e-03
 4.57332844e-02  2.03016718e-03 -1.26175436e-03], [0.7912170966592325]
Video ID: 17
 Normalized Coefficients without Clusters: [-2.39500389e-05  3.73761699e-03 -1.47282111e-04
-8.61583216e-04
 1.82705883e-02  2.42442823e-03 -1.55918597e-03], [0.8731137097978046]
Video ID: 18
 Normalized Coefficients without Clusters: [-5.26035186e-05  4.63002763e-02  8.55110364e-05
8.64311445e-04
 4.15558507e-03  1.12902523e-03  2.61295624e-04], [0.3415623047881508]
Video ID: 19
 Normalized Coefficients without Clusters: [ 7.76207765e-05 -5.89110201e-02  1.34269185e-04
8.55434741e-03
 2.08566410e-01  1.03394136e-02 -2.44516866e-03], [0.2258149985532334]
Video ID: 20
 Normalized Coefficients without Clusters: [-6.64882900e-05  1.57177438e-01 -4.93821731e-05
6.24199063e-03
 7.72161524e-02  1.56872298e-02 -1.35559582e-02], [0.3767860695465331]
Video ID: 21
 Normalized Coefficients without Clusters: [-8.74656811e-06  3.99405102e-03 -7.25812826e-05
-1.62820767e-02
 1.51735450e-01  1.49415327e-03  9.35796195e-03], [0.5833248083952016]
Video ID: 22
 Normalized Coefficients without Clusters: [-7.06967528e-05 -1.67279851e-02  4.70775604e-05
3.43880012e-03
 3.94506742e-02 -3.59295014e-03 -2.84333144e-02], [0.5923947115779861]
Video ID: 23
 Normalized Coefficients without Clusters: [-0.00092847  0.08058905 -0.00010423  0.00380107
0.04783755 -0.0035674
 -0.00592786], [0.5199384022439203]
```

```
Video ID: 24
  Normalized Coefficients without Clusters: [ 1.41055956e-04  4.30474683e-02 -1.03850279e-05
4.57347226e-03
  1.65461101e-01  1.09935265e-02 -1.77219816e-02], [0.34467785251029354]
Video ID: 25
  Normalized Coefficients without Clusters: [-1.43073433e-04 -5.00836255e-02 -5.15006277e-06
2.49251328e-03
  7.59620706e-02  1.86502855e-03 -2.03333101e-03], [0.5717318267753004]
Video ID: 26
  Normalized Coefficients without Clusters: [ 8.43255216e-05 -5.15761645e-02  5.63012171e-05
2.48041341e-04
  2.13671103e-02 -5.30104227e-03  5.69594333e-03], [0.9496855835249194]
Video ID: 27
  Normalized Coefficients without Clusters: [ 1.97631315e-04  1.11211753e-01  8.18816398e-05
8.56105538e-03
 -1.55692081e-01  2.68594942e-02 -5.80892538e-03], [0.5347789430693611]
Video ID: 28
  Normalized Coefficients without Clusters: [ 1.68610752e-04 -8.94456457e-02  3.17395220e-05
3.66574305e-02
  4.53084165e-02 -6.26365377e-03 -1.81638915e-02], [0.60012984166446]
Video ID: 30
  Normalized Coefficients without Clusters: [-0.00367172  0.00467644  0.00160923  0.03851093
0.08808509  0.00908456
 -0.00706542], [0.6428616725917511]
Video ID: 31
  Normalized Coefficients without Clusters: [-3.07905106e-05 -1.67448029e-01  1.11296026e-05
2.63302777e-02
  2.06516471e-01  2.63213742e-03  6.97750046e-03], [0.43834806544392013]
Video ID: 32
  Normalized Coefficients without Clusters: [-3.63894466e-04 -9.48903147e-02  2.18496033e-05
1.13786159e-02
  9.38339207e-02  1.20949755e-02 -2.27813174e-02], [0.6135614807461376]
Video ID: 33
  Normalized Coefficients without Clusters: [ 0.00013092  0.09380051  0.00022044 -0.00011169
0.03120825  0.00698545
  0.00039932], [0.5060280164469544]
Video ID: 34
  Normalized Coefficients without Clusters: [-2.00007816e-03  6.68909196e-03  5.23348223e-05
1.13125621e-03
 -4.91271485e-02  1.17047080e-02 -1.53222319e-02], [0.5210400446679275]
Video ID: 35
  Normalized Coefficients without Clusters: [-1.01307853e-04 -7.40859520e-02 -4.58739876e-03
3.47072028e-02
  1.26331808e-01 -7.57562234e-03  3.83025103e-02], [0.1893745891593542]
Video ID: 36
  Normalized Coefficients without Clusters: [ 2.70126761e-05 -2.34088280e-01  1.12083971e-04
2.33812987e-02
  9.03625047e-02 -8.19002189e-03  2.86590913e-03], [0.3695587534817583]
Video ID: 37
```

```
  Normalized Coefficients without Clusters: [-0.00055057 -0.04534213 -0.00010203 -0.00298173
-0.02176888  0.00614588
 -0.01010609], [0.6395566788738862]
Video ID: 38
  Normalized Coefficients without Clusters: [-2.14848569e-04 -1.05490670e-01  7.32810041e-05
4.35337903e-04
  1.25392249e-01  3.41580187e-03  4.29122176e-03], [0.3342729120918173]
Video ID: 39
  Normalized Coefficients without Clusters: [-9.90237489e-05 -1.70462430e-01  1.47177346e-04
2.62951560e-02
  1.38286079e-01  1.69577485e-02 -2.00240751e-02], [0.458319193818333]
Video ID: 40
  Normalized Coefficients without Clusters: [ 0.00013975 -0.06725449 -0.00071644  0.03062092
-0.05888365  0.0038057
 -0.00419458], [0.36686376177788194]
Video ID: 41
  Normalized Coefficients without Clusters: [ 4.66410646e-05 -7.13128028e-02  2.34101786e-04
-9.03815520e-03
  1.60300475e-01 -3.89580538e-03 -1.93742329e-02], [0.7470559808916056]
Video ID: 42
  Normalized Coefficients without Clusters: [-0.00012471 -0.07288295 -0.00141184  0.02060626
0.00644794  0.04674034
 -0.02370682], [0.6752267456459513]
Video ID: 43
  Normalized Coefficients without Clusters: [-0.00014867 -0.13866374 -0.00205512  0.03670589
-0.08896596  0.00375202
 -0.00154083], [0.6952472855184636]
Video ID: 44
  Normalized Coefficients without Clusters: [ 0.00030941  0.06102343  0.00025516  0.00712439  0.1302834
0.01288669
 -0.00868279], [0.10341376356357682]
Video ID: 45
  Normalized Coefficients without Clusters: [-0.01589854 -0.17089924 -0.00024663  0.03926525
0.09549287  0.00498355
  0.00797005], [0.5118018317370702]
Video ID: 46
  Normalized Coefficients without Clusters: [ 0.00016417  0.13719121  0.00073054 -0.00793144
0.11579383 -0.00557715
  0.00641037], [0.4660864390866997]
Video ID: 47
  Normalized Coefficients without Clusters: [ 4.59953420e-05 -2.76726401e-01  3.56324256e-04
4.27536261e-02
  4.11085600e-02  4.43374375e-03  1.24291799e-01], [0.6003367532476375]
Video ID: 48
  Normalized Coefficients without Clusters: [-0.00020814  0.05059767  0.00359188  0.02798718
0.08894834  0.01668041
 -0.02349845], [0.5025514433782351]
Video ID: 49
```

```
  Normalized Coefficients without Clusters: [ 4.12858514e-05  1.41648536e-01 -9.24892478e-05
3.74250629e-02
  1.25684633e-01  2.57486010e-02  2.36196140e-02], [0.5297221648501017]
Video ID: 50
  Normalized Coefficients without Clusters: [ 0.003088   -0.03176196 -0.00026221  0.02586422
-0.07087706  0.00029731
  0.00321577], [0.7924188106628645]
Video ID: 51
  Normalized Coefficients without Clusters: [-0.00379878  0.06016988  0.00022932  0.02321414
-0.00633775  0.00220442
 -0.00924859], [0.5698820687372799]
Video ID: 52
  Normalized Coefficients without Clusters: [ 0.01591233  0.23445697  0.00112168 -0.0129896
-0.13182766  0.0239917
  0.01258137], [0.36559403195952755]
Video ID: 53
  Normalized Coefficients without Clusters: [-8.91177466e-05 -5.85178748e-02  6.13564483e-05
-1.01633836e-02
 -3.39657078e-02  1.80365952e-02  2.53226880e-02], [0.44706861964513817]
Video ID: 54
  Normalized Coefficients without Clusters: [ 0.00022863 -0.000587   -0.00029968  0.03081616  0.17121707
0.00882653
 -0.03317809], [0.5537856454325876]
Video ID: 55
  Normalized Coefficients without Clusters: [ 0.00025813 -0.01053621 -0.0003022  -0.02971443
0.23114058  0.00588654
 -0.00084816], [0.5935590451131579]
Video ID: 56
  Normalized Coefficients without Clusters: [ 0.04501575 -0.2037205  -0.00065671 -0.04460591
0.32724029  0.0410731
 -0.06056922], [0.3977106074606179]
Video ID: 57
  Normalized Coefficients without Clusters: [-0.00546341 -0.02688245  0.00293898  0.06570436
-0.27209254  0.0370405
 -0.01752052], [0.5630702399470341]
Video ID: 58
  Normalized Coefficients without Clusters: [ 2.14800291e-05 -9.07955538e-03 -3.81094337e-04
2.42757519e-02
 -7.67881038e-02 -1.25093008e-02 -2.12794729e-03], [0.9711084905215573]
Video ID: 59
  Normalized Coefficients without Clusters: [0.00103848 0.03241305 0.00012421 0.00522425 0.05385793
0.00081068
 0.0030514 ], [0.7899980745531452]
Video ID: 60
  Normalized Coefficients without Clusters: [ 0.00825237  0.36647139 -0.00042928  0.02314161
-0.29165424  0.00311067
 -0.00159814], [0.5155921878704383]
Video ID: 61
```

```
  Normalized Coefficients without Clusters: [-0.0003895   0.15645251 -0.00481264  0.01810118  0.1801806
0.0139737
  0.00123625], [0.12837126875113508]
Video ID: 62
  Normalized Coefficients without Clusters: [ 0.00023004  0.04817008 -0.00814045  0.01505035 -0.0926
0.06678939
 -0.01476113], [0.5233065451592793]
Video ID: 63
  Normalized Coefficients without Clusters: [-0.00042289  0.0213558   0.00086206  0.01421591
0.06107767  0.01749862
 -0.02779597], [0.6619534956043727]
Video ID: 64
  Normalized Coefficients without Clusters: [ 0.32875386 -0.15356542 -0.00217276 -0.00187245
-0.08034124 -0.00855934
 -0.03317756], [0.8423051290995865]
Video ID: 65
  Normalized Coefficients without Clusters: [-0.01861897 -0.20801597  0.00311212 -0.02286902
0.19020118 -0.00038207
 -0.00986745], [0.4240737930663661]
Video ID: 66
  Normalized Coefficients without Clusters: [-0.00528211  0.00964446  0.00252495  0.04317207
0.03767682  0.01580172
 -0.06180018], [0.4037990457470001]
Video ID: 67
  Normalized Coefficients without Clusters: [ 0.0001868  -0.13201494 -0.00399995  0.00409578
0.05152621  0.03069051
 -0.02361201], [0.7358733852929948]
Video ID: 68
  Normalized Coefficients without Clusters: [-0.00011686 -0.1113473  -0.00024861  0.06125785
0.03987277  0.02350824
 -0.08615675], [0.4173110687978741]
Video ID: 69
  Normalized Coefficients without Clusters: [ 3.94666386e-05  1.15895632e-01  6.13306634e-05
3.36202231e-02
  1.77509977e-02 -7.46294497e-03  2.69759747e-02], [0.5327239421425687]
Video ID: 70
  Normalized Coefficients without Clusters: [ 0.00048038 -0.02459977  0.00060333  0.04558307
-0.03578419  0.00936056
 -0.01880141], [0.7505788098413778]
Video ID: 71
  Normalized Coefficients without Clusters: [-0.00255247 -0.15132858 -0.00140485  0.06400736
0.05036908  0.00308687
 -0.0117917 ], [0.1310993990673267]
Video ID: 72
  Normalized Coefficients without Clusters: [ 3.05452525e-05  1.24403794e-01  1.09134238e-04
1.70333098e-02
 -9.45855056e-02  9.53295453e-03 -2.67137853e-02], [0.9006968030213874]
Video ID: 73
```

```
  Normalized Coefficients without Clusters: [ 0.00726977  0.01979786  0.00032778  0.01549881
-0.00296945 -0.0189677
  0.00733837], [0.8954354957187372]
Video ID: 74
  Normalized Coefficients without Clusters: [ 0.09420008 -0.13563667 -0.00127994  0.05104404
0.05861953  0.00720713
 -0.05257816], [0.5689574786893756]
Video ID: 75
  Normalized Coefficients without Clusters: [-1.66147456e-04  2.86851271e-01 -3.38039423e-03
1.79334615e-02
 -1.00733286e-01  2.64829754e-02 -1.11450979e-02], [0.46824474189427245]
Video ID: 76
  Normalized Coefficients without Clusters: [-0.19436058  0.12117481  0.00038545  0.00849862
-0.10805016  0.0519366
 -0.0636141 ], [0.8425263425140698]
Video ID: 77
  Normalized Coefficients without Clusters: [ 1.82340274e-04  4.42722095e-02  1.14235659e-04
4.21790758e-05
 -7.50244013e-02 -2.17703439e-02  3.43746559e-02], [0.9411728192159822]
Video ID: 78
  Normalized Coefficients without Clusters: [-2.28383933e-01  3.88386118e-01 -1.96798812e-04
-2.65973473e-03
 -2.19997096e-01  1.66974349e-02 -9.51058351e-03], [0.9186091035480336]
Video ID: 79
  Normalized Coefficients without Clusters: [-1.30421892e-04 -1.22488780e-01  1.23607124e-03
5.24177410e-02
  3.10001160e-01 -8.90072322e-03 -2.26834364e-02], [0.3324906829074906]
Video ID: 80
  Normalized Coefficients without Clusters: [-0.28052899  0.32121292 -0.00180811  0.01757157
-0.05049856  0.00831009
 -0.06404802], [0.5331742190357964]
Video ID: 81
  Normalized Coefficients without Clusters: [ 0.00027302  0.04544383  0.00036665  0.01580093
0.07251275  0.00806556
 -0.0181937 ], [0.6489482358662828]
Video ID: 82
  Normalized Coefficients without Clusters: [ 0.17465038 -0.1458864   0.01741271 -0.05374079
-0.18117315  0.00765984
  0.03101158], [0.7297005647501452]
Video ID: 83
  Normalized Coefficients without Clusters: [ 0.01516391  0.08245072  0.00034754  0.01795622
0.01705863 -0.00811053
 -0.00373877], [0.798274270787944]
Video ID: 84
  Normalized Coefficients without Clusters: [ 0.02246664 -0.05092139  0.00700835  0.00777304
-0.22292837 -0.00901566
 -0.06948038], [0.8938808024774023]
Video ID: 85
```

```
  Normalized Coefficients without Clusters: [ 0.00063527 -0.08341006  0.0001288   0.02611991
-0.04327076  0.00257222
 -0.00109721], [0.8482269429691031]
Video ID: 86
  Normalized Coefficients without Clusters: [-0.17617025  0.00603722  0.05596867  0.0168687
0.04975851  0.02695085
 -0.04336362], [0.7462106384445438]
Video ID: 87
  Normalized Coefficients without Clusters: [-0.07911757  0.02565485 -0.0058325   0.08097034
0.00615493 -0.00873828
 -0.01623734], [0.6815365756475058]
Video ID: 88
  Normalized Coefficients without Clusters: [-0.13382909  0.15929194 -0.07443989  0.03051365
-0.00171837 -0.02028285
  0.00714224], [0.13253171126527652]
Video ID: 89
  Normalized Coefficients without Clusters: [ 0.00217299 -0.00119211  0.00113235  0.04833554
0.12473945  0.00782174
 -0.03065158], [0.2071019798978317]
Video ID: 90
  Normalized Coefficients without Clusters: [-0.08660778  0.23230432 -0.00218931  0.04916326
-0.03212005 -0.02554492
  0.02576489], [0.15281134740226293]
Video ID: 91
  Normalized Coefficients without Clusters: [ 0.00719843  0.07727298 -0.00531801  0.08898874
-0.06142411 -0.01070746
 -0.04810492], [0.4540013225984804]
Video ID: 92
  Normalized Coefficients without Clusters: [-0.00076514  0.22231596  0.01029119  0.05338863
-0.04080865  0.01536119
 -0.01303897], [0.5695931645126946]

Summary of Ridge Regression Results:
Video ID: 0
  MSE without clusters: 0.1883
  MSE with clusters: 0.1872

Video ID: 1
  MSE without clusters: 0.2033
  MSE with clusters: 0.2024

Video ID: 2
  MSE without clusters: 0.1637
  MSE with clusters: 0.1627

Video ID: 3
  MSE without clusters: 0.1384
  MSE with clusters: 0.1368
```

```
Video ID: 4
  MSE without clusters: 0.2477
  MSE with clusters: 0.2470

Video ID: 5
  MSE without clusters: 0.2273
  MSE with clusters: 0.2259

Video ID: 6
  MSE without clusters: 0.1167
  MSE with clusters: 0.1163

Video ID: 7
  MSE without clusters: 0.2223
  MSE with clusters: 0.2210

Video ID: 8
  MSE without clusters: 0.2352
  MSE with clusters: 0.2330

Video ID: 9
  MSE without clusters: 0.2145
  MSE with clusters: 0.2134

Video ID: 10
  MSE without clusters: 0.2184
  MSE with clusters: 0.2171

Video ID: 11
  MSE without clusters: 0.1372
  MSE with clusters: 0.1368

Video ID: 12
  MSE without clusters: 0.2233
  MSE with clusters: 0.2191

Video ID: 13
  MSE without clusters: 0.2407
  MSE with clusters: 0.2394

Video ID: 14
  MSE without clusters: 0.2392
  MSE with clusters: 0.2332

Video ID: 15
  MSE without clusters: 0.2372
  MSE with clusters: 0.2359

Video ID: 16
```

```
  MSE without clusters: 0.1262
  MSE with clusters: 0.1249

Video ID: 17
  MSE without clusters: 0.0940
  MSE with clusters: 0.0935

Video ID: 18
  MSE without clusters: 0.2368
  MSE with clusters: 0.2329

Video ID: 19
  MSE without clusters: 0.2317
  MSE with clusters: 0.2278

Video ID: 20
  MSE without clusters: 0.2151
  MSE with clusters: 0.2098

Video ID: 21
  MSE without clusters: 0.1896
  MSE with clusters: 0.1879

Video ID: 22
  MSE without clusters: 0.2332
  MSE with clusters: 0.2330

Video ID: 23
  MSE without clusters: 0.2273
  MSE with clusters: 0.2246

Video ID: 24
  MSE without clusters: 0.2300
  MSE with clusters: 0.2260

Video ID: 25
  MSE without clusters: 0.2337
  MSE with clusters: 0.2307

Video ID: 26
  MSE without clusters: 0.0603
  MSE with clusters: 0.0601

Video ID: 27
  MSE without clusters: 0.2332
  MSE with clusters: 0.2247

Video ID: 28
  MSE without clusters: 0.2212
  MSE with clusters: 0.2168

Video ID: 30
  MSE without clusters: 0.1641
  MSE with clusters: 0.1630

Video ID: 31
  MSE without clusters: 0.2286
  MSE with clusters: 0.2240

Video ID: 32
  MSE without clusters: 0.2198
  MSE with clusters: 0.2151

Video ID: 33
  MSE without clusters: 0.2275
  MSE with clusters: 0.2258

Video ID: 34
  MSE without clusters: 0.2352
  MSE with clusters: 0.2307

Video ID: 35
  MSE without clusters: 0.2088
  MSE with clusters: 0.2039

Video ID: 36
  MSE without clusters: 0.2127
  MSE with clusters: 0.2034

Video ID: 37
  MSE without clusters: 0.2383
  MSE with clusters: 0.2312

Video ID: 38
  MSE without clusters: 0.2370
  MSE with clusters: 0.2239

Video ID: 39
  MSE without clusters: 0.2267
  MSE with clusters: 0.2201

Video ID: 40
  MSE without clusters: 0.1927
  MSE with clusters: 0.1896

Video ID: 41
  MSE without clusters: 0.1289
  MSE with clusters: 0.1211
```

```
Video ID: 42
  MSE without clusters: 0.2068
  MSE with clusters: 0.2047

Video ID: 43
  MSE without clusters: 0.2342
  MSE with clusters: 0.2281

Video ID: 44
  MSE without clusters: 0.2011
  MSE with clusters: 0.1975

Video ID: 45
  MSE without clusters: 0.2296
  MSE with clusters: 0.2226

Video ID: 46
  MSE without clusters: 0.2045
  MSE with clusters: 0.2022

Video ID: 47
  MSE without clusters: 0.2094
  MSE with clusters: 0.2004

Video ID: 48
  MSE without clusters: 0.1892
  MSE with clusters: 0.1875

Video ID: 49
  MSE without clusters: 0.1356
  MSE with clusters: 0.1284

Video ID: 50
  MSE without clusters: 0.1785
  MSE with clusters: 0.1627

Video ID: 51
  MSE without clusters: 0.2060
  MSE with clusters: 0.1975

Video ID: 52
  MSE without clusters: 0.2299
  MSE with clusters: 0.2168

Video ID: 53
  MSE without clusters: 0.2259
  MSE with clusters: 0.2163
```

```
Video ID: 54
  MSE without clusters: 0.1510
  MSE with clusters: 0.1443

Video ID: 55
  MSE without clusters: 0.1477
  MSE with clusters: 0.1449

Video ID: 56
  MSE without clusters: 0.2043
  MSE with clusters: 0.1923

Video ID: 57
  MSE without clusters: 0.1627
  MSE with clusters: 0.1524

Video ID: 58
  MSE without clusters: 0.0851
  MSE with clusters: 0.0793

Video ID: 59
  MSE without clusters: 0.0883
  MSE with clusters: 0.0876

Video ID: 60
  MSE without clusters: 0.2049
  MSE with clusters: 0.2008

Video ID: 61
  MSE without clusters: 0.2178
  MSE with clusters: 0.2162

Video ID: 62
  MSE without clusters: 0.2324
  MSE with clusters: 0.2207

Video ID: 63
  MSE without clusters: 0.1611
  MSE with clusters: 0.1510

Video ID: 64
  MSE without clusters: 0.1223
  MSE with clusters: 0.1142

Video ID: 65
  MSE without clusters: 0.2179
  MSE with clusters: 0.2081

Video ID: 66
```

```
  MSE without clusters: 0.2196
  MSE with clusters: 0.2121

Video ID: 67
  MSE without clusters: 0.1800
  MSE with clusters: 0.1573

Video ID: 68
  MSE without clusters: 0.2250
  MSE with clusters: 0.2174

Video ID: 69
  MSE without clusters: 0.2054
  MSE with clusters: 0.1929

Video ID: 70
  MSE without clusters: 0.1761
  MSE with clusters: 0.1727

Video ID: 71
  MSE without clusters: 0.1338
  MSE with clusters: 0.1167

Video ID: 72
  MSE without clusters: 0.0807
  MSE with clusters: 0.0790

Video ID: 73
  MSE without clusters: 0.0744
  MSE with clusters: 0.0727

Video ID: 74
  MSE without clusters: 0.2056
  MSE with clusters: 0.2024

Video ID: 75
  MSE without clusters: 0.2117
  MSE with clusters: 0.2003

Video ID: 76
  MSE without clusters: 0.1890
  MSE with clusters: 0.1749

Video ID: 77
  MSE without clusters: 0.0924
  MSE with clusters: 0.0920

Video ID: 78
  MSE without clusters: 0.1548
  MSE with clusters: 0.1371

Video ID: 79
  MSE without clusters: 0.1938
  MSE with clusters: 0.1803

Video ID: 80
  MSE without clusters: 0.2198
  MSE with clusters: 0.2114

Video ID: 81
  MSE without clusters: 0.1489
  MSE with clusters: 0.1454

Video ID: 82
  MSE without clusters: 0.2178
  MSE with clusters: 0.1956

Video ID: 83
  MSE without clusters: 0.0894
  MSE with clusters: 0.0809

Video ID: 84
  MSE without clusters: 0.2095
  MSE with clusters: 0.1847

Video ID: 85
  MSE without clusters: 0.1406
  MSE with clusters: 0.1373

Video ID: 86
  MSE without clusters: 0.1887
  MSE with clusters: 0.1788

Video ID: 87
  MSE without clusters: 0.1954
  MSE with clusters: 0.1886

Video ID: 88
  MSE without clusters: 0.1251
  MSE with clusters: 0.1174

Video ID: 89
  MSE without clusters: 0.2208
  MSE with clusters: 0.1989

Video ID: 90
  MSE without clusters: 0.1675
  MSE with clusters: 0.1490
```

```
Video ID: 91
  MSE without clusters: 0.2124
  MSE with clusters: 0.1973

Video ID: 92
  MSE without clusters: 0.1569
  MSE with clusters: 0.1449
```

Results for the Logistic Classifier

```
Cluster Centers:
 [[-2.65105884e-02  5.21519752e-01 -4.48383187e-02 -3.00371032e-03
  -2.13974361e-01 -5.49190982e-02 -1.59524787e-01]
 [ 2.42194060e+01 -1.26824692e+00 -2.57350137e-02 -5.44777475e-03
  -1.06015584e+00 -3.83272224e-02 -4.92717389e-02]
 [ 7.72304733e-03 -9.92372620e-01 -1.02480314e-02  3.70697617e-02
   6.24357909e-02  1.74884258e+00  6.40550325e+00]
 [-3.05959187e-02 -1.63030843e+00 -1.07564206e-02 -1.01161606e-02
  -6.52254126e-01  9.72040809e-02  1.61545725e-01]
 [ 1.70350206e-02 -4.54177765e-01  2.02844989e+01  1.96582211e-02
  -2.48770218e-01  5.22966931e-02  3.14031139e-01]
 [-7.22799041e-02  6.31780652e-01 -2.88273019e-02  1.56301809e+02
   1.95736598e+00  6.08402417e-01 -2.50979331e-01]
 [-3.70646182e-02  3.06260342e-01 -4.08577834e-02 -1.74908848e-02
   1.67987703e+00 -5.46669872e-02 -9.83056442e-02]]
```

```
Feature Means:
 fracSpent     24.307622
fracComp       0.764945
fracPaused    36.352843
numPauses      2.868576
avgPBR         1.112518
numRWs         2.203361
numFFs         1.544615
```

```
Feature Standard Deviations:
 fracSpent     319.334342
fracComp        0.344518
fracPaused    390.129825
numPauses      64.492769
avgPBR          0.325690
numRWs         16.102563
numFFs          6.154476
```

```
GMM Z-scores for a Sample Student:
 fracSpent    -0.074324
fracComp      0.270105
fracPaused    -0.08806
numPauses     0.048555
avgPBR        1.189725
numRWs       -0.136833
numFFs        0.398959
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.50      0.00      0.00      1605
           1       0.67      1.00      0.80      3311
```

```
    accuracy                           0.67      4916
   macro avg       0.59      0.50      0.40      4916
weighted avg       0.62      0.67      0.54      4916
```

Clustering Output

```
Video ID: 0, Number of clusters: 8
Video ID: 1, Number of clusters: 8
Video ID: 2, Number of clusters: 8
Video ID: 3, Number of clusters: 6
Video ID: 4, Number of clusters: 8
Video ID: 5, Number of clusters: 8
Video ID: 6, Number of clusters: 7
Video ID: 7, Number of clusters: 8
Video ID: 8, Number of clusters: 8
Video ID: 9, Number of clusters: 8
Video ID: 10, Number of clusters: 8
Video ID: 11, Number of clusters: 8
Video ID: 12, Number of clusters: 8
Video ID: 13, Number of clusters: 8
Video ID: 14, Number of clusters: 6
Video ID: 15, Number of clusters: 6
Video ID: 16, Number of clusters: 8
Video ID: 17, Number of clusters: 8
Video ID: 18, Number of clusters: 8
Video ID: 19, Number of clusters: 6
Video ID: 20, Number of clusters: 8
Video ID: 21, Number of clusters: 8
Video ID: 22, Number of clusters: 7
Video ID: 23, Number of clusters: 4
Video ID: 24, Number of clusters: 8
Video ID: 25, Number of clusters: 8
Video ID: 26, Number of clusters: 8
Video ID: 27, Number of clusters: 7
Video ID: 28, Number of clusters: 3
Video ID: 30, Number of clusters: 5
Video ID: 31, Number of clusters: 8
Video ID: 32, Number of clusters: 8
Video ID: 33, Number of clusters: 6
Video ID: 34, Number of clusters: 3
Video ID: 35, Number of clusters: 8
Video ID: 36, Number of clusters: 6
Video ID: 37, Number of clusters: 8
Video ID: 38, Number of clusters: 6
Video ID: 39, Number of clusters: 8
Video ID: 40, Number of clusters: 7
Video ID: 41, Number of clusters: 8
Video ID: 42, Number of clusters: 7
Video ID: 43, Number of clusters: 8
Video ID: 44, Number of clusters: 7
Video ID: 45, Number of clusters: 5
Video ID: 46, Number of clusters: 8
Video ID: 47, Number of clusters: 8
Video ID: 48, Number of clusters: 5
Video ID: 49, Number of clusters: 5
Video ID: 50, Number of clusters: 6
Video ID: 51, Number of clusters: 5
Video ID: 52, Number of clusters: 4
Video ID: 53, Number of clusters: 8
Video ID: 54, Number of clusters: 8
Video ID: 55, Number of clusters: 6
Video ID: 56, Number of clusters: 8
Video ID: 57, Number of clusters: 8
Video ID: 58, Number of clusters: 3
Video ID: 59, Number of clusters: 7
Video ID: 60, Number of clusters: 6
Video ID: 61, Number of clusters: 8
Video ID: 62, Number of clusters: 7
Video ID: 63, Number of clusters: 6
Video ID: 64, Number of clusters: 7
Video ID: 65, Number of clusters: 6
Video ID: 66, Number of clusters: 8
Video ID: 67, Number of clusters: 7
Video ID: 68, Number of clusters: 6
Video ID: 69, Number of clusters: 8
Video ID: 70, Number of clusters: 8
Video ID: 71, Number of clusters: 8
Video ID: 72, Number of clusters: 7
Video ID: 73, Number of clusters: 8
Video ID: 74, Number of clusters: 7
Video ID: 75, Number of clusters: 8
Video ID: 76, Number of clusters: 8
Video ID: 77, Number of clusters: 8
Video ID: 78, Number of clusters: 7
Video ID: 79, Number of clusters: 8
Video ID: 80, Number of clusters: 8
Video ID: 81, Number of clusters: 8
Video ID: 82, Number of clusters: 8
Video ID: 83, Number of clusters: 8
Video ID: 84, Number of clusters: 8
Video ID: 85, Number of clusters: 8
Video ID: 86, Number of clusters: 8
Video ID: 87, Number of clusters: 6
Video ID: 88, Number of clusters: 8
Video ID: 89, Number of clusters: 7
Video ID: 90, Number of clusters: 7
Video ID: 91, Number of clusters: 8
Video ID: 92, Number of clusters: 8
Total Euclidean distance from all points to their centers across all clusters: 26463.1551
Total number of clusters across all models: 655
```

References

Learn. scikit. (n.d.). https://scikit-learn.org/stable/index.html

Python. (n.d.). https://pythonprogramminglanguage.com/kmeans-elbow-method/#google_vignette