Team Member Names: Sahil Jaganmohan and Akshay Aravindan
Purdue Usernames: sjaganm and aravinda
GitHub Usernames: bullpointe and akshayaravindan
GitHub Team Name: project-s20-akshay-and-sahil
Project: Path 2: Student performance related to video-watching behavior

DATASET:

The features of this dataset include video-watching behavior, which consists of a user(userID), a video ID (VidID), the fraction of time spent watching the video (fracSpent), the fraction of the video completed (fracComp), the fraction of time the video was paused (fracPaused), the number of pauses (numPauses), the average playback rate (avgPBR), the standard deviation for the playback rate (stdPBR), the number of times the student fast-forwarded (numFFs) and re-winded (numRWs), and the student's score on the video quiz (s). The units didn't exist for most of the data points, as they were a fraction of a whole. There were 29,304 samples in total the statistics are shown in fig 1a below.

Value	Min	Max	Mean	Std Dev.
VidID	0	92	19.8304327054	22.5063577
fracSpent	0	18215.9846341	24.0765150631	308.275793868
fracComp	0	9.28	0.767870424293	0.340592996089
fracPlayed	0	439.65	0.985554633266	3.7235736127
fracPaused	0	15957.3920237	36.340134205	375.758641755
numPauses	0	10083	2.82551870052	59.1019139946
avgPBR	0	2	1.10437419398	0.315582256552
stdPBR	0	0.98	0.014450527237	0.050013334333
numRWs	0	2237	2.23802211302	15.5645896641
numFFs	0	309	1.5676016926	6.3700138764
s	0	1	0.663322413322	0.47258164434

Fig. 1a: Dataset Features and statistics

The dataset can be found at the link below:
https://ieeexplore.ieee.org/document/7218617

METHODS:

1. **How well can the students be naturally grouped or clustered by their video-watching behavior (fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs)? You should use all students that complete at least five of the videos in your analysis.**
    a. A Gaussian Mixture Model (gmm) was used to verify if students could be naturally grouped or clustered by their video-watching behavior. In order to parse the data accordingly, each feature column was normalized in order to remove large outliers affecting the model. Features that were used in the model include fracSpent, fracComp, fracPaused, numPauses, avgPBR, stdPBR, numRWs, and numFFs. Samples that were used included students that had 5 unique entries; we assumed that a data entry entails that a student completed the video. A GMM was used because each feature column was normalized and therefore a GMM model would properly create evenly distributed clusters from our samples. To verify our model, we analyzed the weights of each cluster to check that no single cluster would be over or under-represented, demonstrating groupings based on similar feature data points.

2. **Can a student's video-watching behavior be used to predict a student's performance (i.e., average score s across all quizzes)? This type of analysis could ultimately save significant time by avoiding the need for tests. You should use all students that complete at least half of the quizzes in your analysis.**
    a. We separated the data into a training set and a testing set based on user-id representing a single student's results. We assumed that the completion of a quiz corresponds to a single data entry. In order to predict the average score, we introduced a new feature for each student with the percentage of questions they got right out of the total questions taken. The training set was fed through our GMM clustering model. A Gaussian Mixture Model (gmm) was used to naturally cluster students' performance on each video quiz pair. Features that were used to build the model include fracSpent, fracComp, fracPaused, numPauses, avgPBR, stdPBR, numRWs, numFFs, and percentage. This model and approach were chosen due to the assumption that similar total video watching behavior will result in similar average performance. To verify our model, we took the testing dataset and averaged all the features used in the model for each individual student. The cluster most similar to this unique data point was used to predict the average score for that student (based on the average score for that cluster) and was compared to the actual performance.

3. **Taking this a step further, how well can you predict a student's performance on a *particular* in-video quiz question (i.e., whether they will be correct or incorrect) based on their video-watching behaviors while watching the corresponding video? You should use all student-video pairs in your analysis.**
    a. A Gaussian Mixture Model (gmm) was used to naturally cluster students performance on each video quiz pair. Features that were used in the to build the model include fracSpent, fracComp, fracPaused, numPauses, avgPBR, stdPBR, numRWs, numFFs, and score. This model and approach were chosen due to the assumption that similar video watching behavior will result in similar results on the corresponding quiz. Therefore when testing our model we could place the test sample in one of the clusters predicting the sample's score on the quiz. A test-train split was performed before creating a model to ensure verification. To verify the model test samples were placed in its "nearest" cluster to predict it's score and is compared to the actual score.

RESULTS:

1. The student video watching behavior can be naturally clustered or grouped to a certain degree. When the data was processed by our model it was able to demonstrate different grouping across different clusters that were determined by the feature data. As shown in *Fig. 2a* below we determined the 12 clusters was an optimal number of clusters for the data based on the densities.

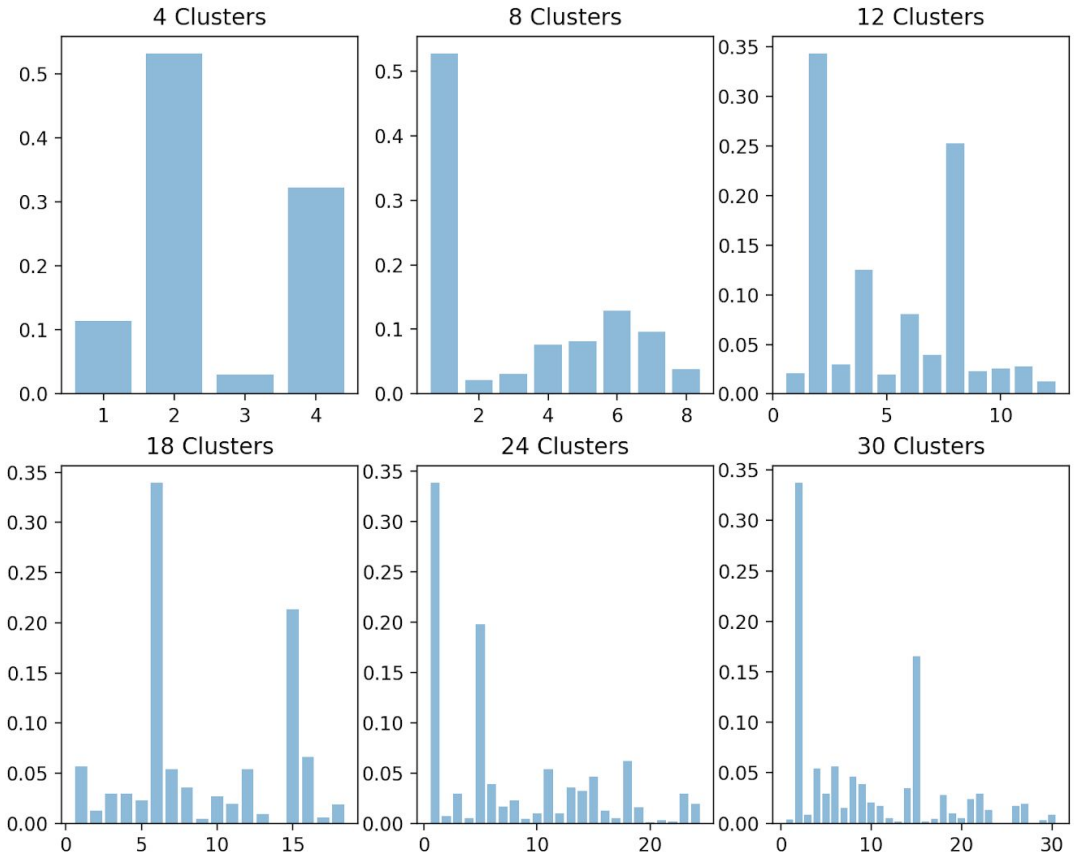

Fig. 2a: Cluster Densities for a range (4 - 30) of clusters

With 12 clusters initialized the student's video-watching behavior was functionally able to be distributed across the clusters, portrayed in *Fig. 2b* below. The bins had densities ranging from 3% to 30% portraying clusters of behaviors, however, the density across the clusters was not ideal. Therefore we can conclude that the behaviors' can be naturally clustered but not completely evenly distributed. Additionally, this can hint towards a correlation between the student's performance based on similar behaviors that will be utilized later in the report.

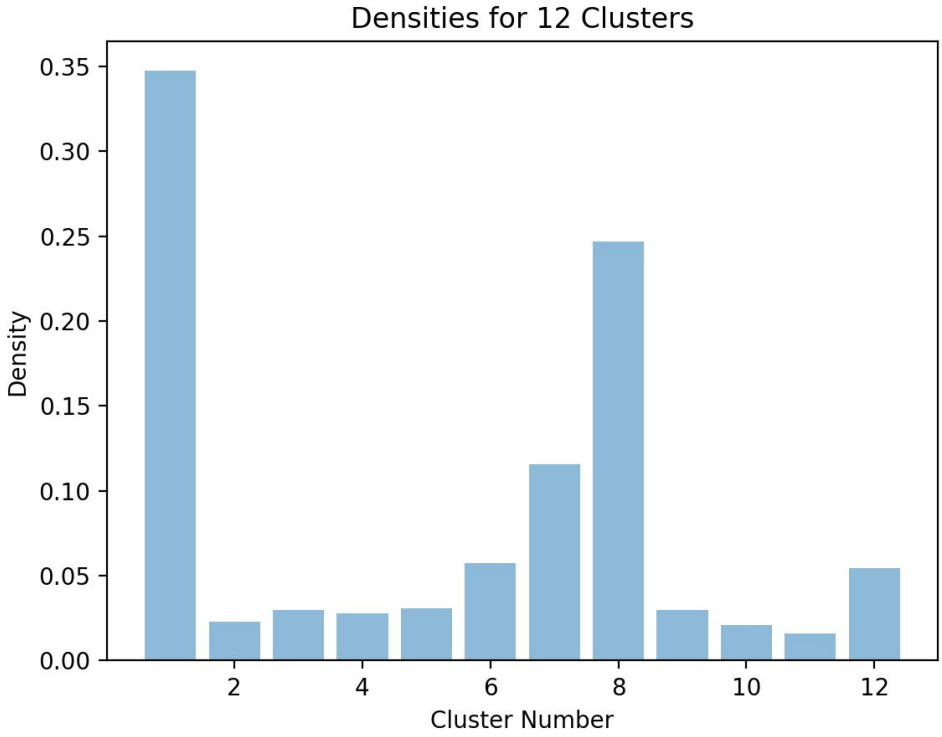

Fig. 2b: 12 Cluster Density for 12 Clusters

2. A student's video-watching behavior can predict their average performance with relatively good precision. In order to determine the student's average score, the formula in *Fig. 3a* can be used.

$$\frac{\sum(Correct\ Attempts)}{\sum(Total\ Attempts)}$$

Fig. 3a: Formula for determining average score per student

When performing the GMM clustering, it was crucial to include the average score for each person as one of the features so that each cluster would have an average score. This is portrayed in *Fig. 3b*, showing each cluster with its respective average scores. The normalized average score of the cluster "nearest" to the test data point would be the predicted score.

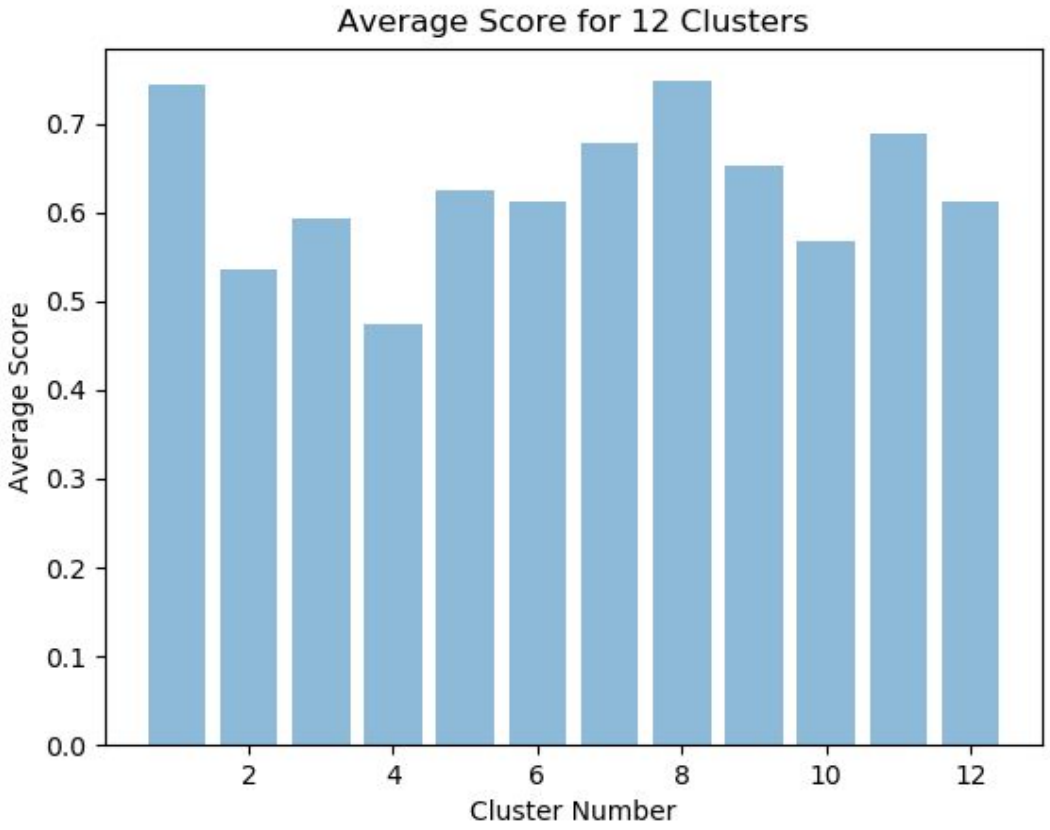

Fig. 3b: 12 Cluster Average Percentages for 12 Clusters

To determine how accurate the predicted average scores were, the Mean Squared Error (MSE) was used. The formula in *Fig. 3c* was used to determine the mean squared error.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2$$

Fig. 3c: Formula for determining MSE

The average MSE was very low in general, with the average MSE of 50 trials staying below 0.45 as shown in *Fig. 3d*. In the figure shown, there were a couple of outliers lifting the average above 0.05, however, in general, the data mostly fixed itself to achieve a flat line. These outliers could be due to the data splitting up in an undesirable way, such that a lot of data points in the testing dataset did not correlate with the data points in the training dataset.

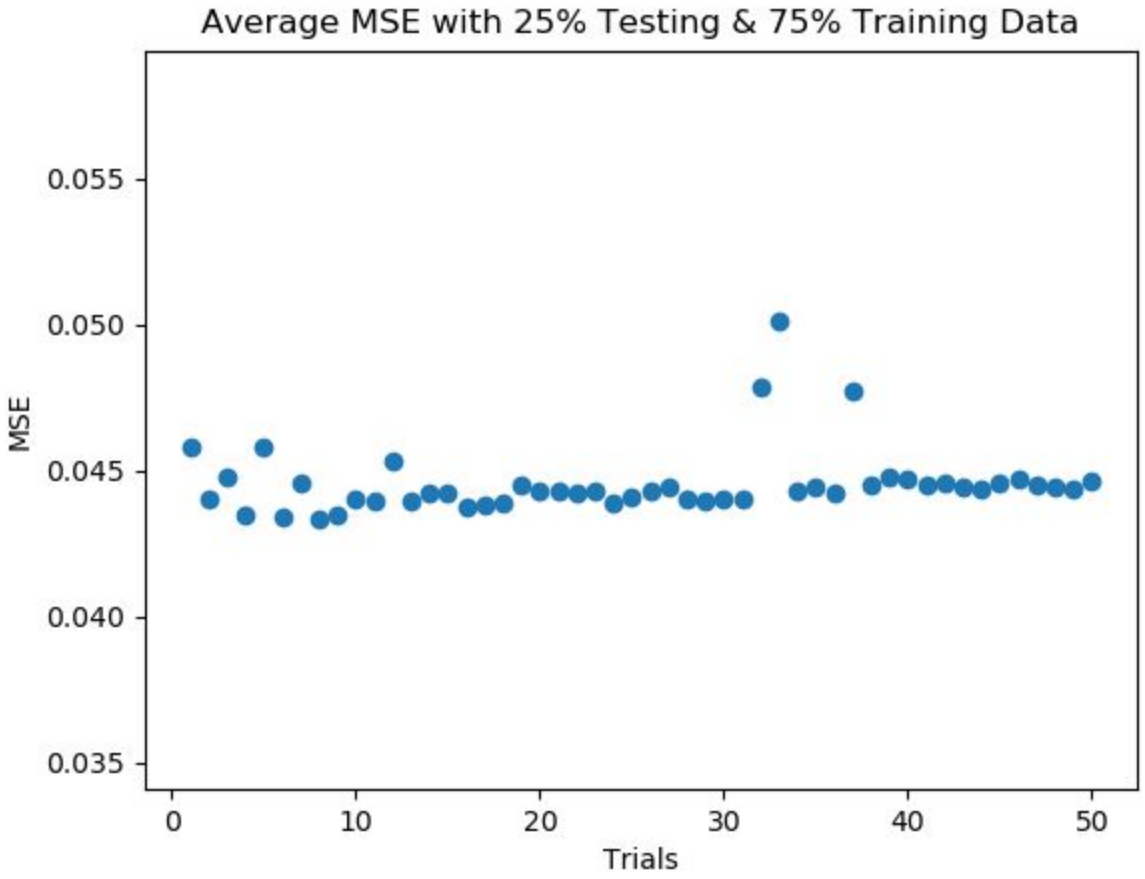

Fig 3d: Average MSE based on 50 trials

3. A student's video-watching behavior can predict their individual performance on a particular video quiz. When performing the GMM clustering, it was crucial to include the score for each person as one of the features so that each cluster would score. When normalized, the score for each cluster turned into either a 0 or a 1. This is portrayed in *Fig. 4a*, showing each cluster with its respective individual quiz score. The normalized score of the cluster "nearest" to the test data point would be the predicted score.

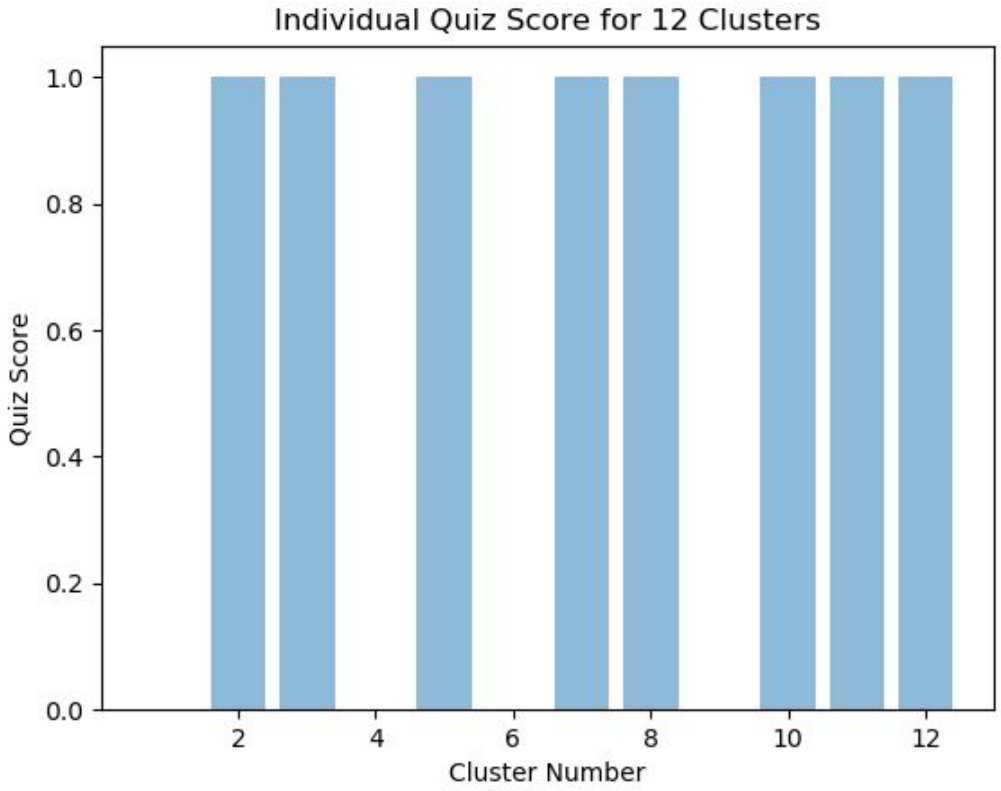

Fig 4a: 12 Cluster Individual Quiz Score for 12 Clusters

To determine how accurate the predicted quiz scores were, the Mean Squared Error (MSE) was used. The following formula was used to determine the mean squared error.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y_i})^2$$

4b: Formula for determining MSE

The average MSE was very low in general, and it was more consistent than *Fig 3d*, with fewer outliers being shown in the average MSE. The average MSE never hit above 0.04, which still means that the MSE of 50 trials is relatively accurate. A reason for the stability may be due to the fact that in this prediction, all the data is being used, however, in the previous prediction, a fewer amount of data points are being used.

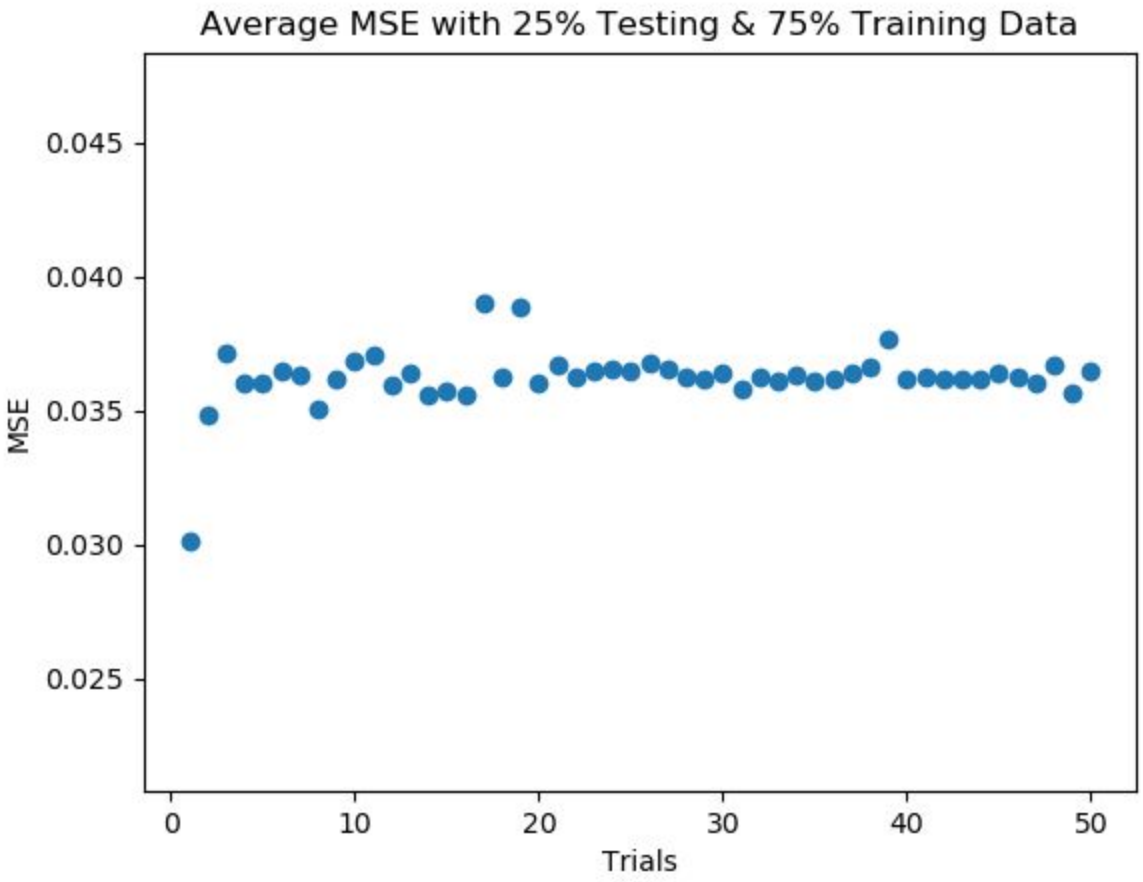

Fig 4c: Average MSE based on 50 trials