Amanda Breton and Alex Bzdel

4th May 2023

ECE 590: Brain Computer Interfaces

Instructor: Dr. Leslie Collins

Improving Feedback in Massive Open Online Course (MOOC) Learning through EEG Analysis

GitHub: https://github.com/abzdel/BCI-EEG-Classification

Introduction

While the increasing popularity and use of online learning, whether it is through resources like Khan Academy, or through online classroom settings over Zoom, has been wonderful for learning accessibility, one of the main struggles of such platforms is getting feedback from students about lesson clarity. In a normal in-person classroom setting, teachers can see their students' faces and body language such as furrowed brows, head scratching, and awkward silence when asked if anyone has any questions. However, online learning can lack real-time responses and often even the absence of facial expressions if cameras are turned off.

Graduate student researchers at Carnegie Mellon University (CMU) sought to address the confusion feedback limitation with the use of commercially available electroencephalography (EEG) headset recorder devices (Wang, Li, et. al, 2011). They designed a classifier with weak but above-chance performance results in determining if a student was confused or not. Their classifier had comparable performance to human observers watching the students' body language for signs of confusion. Our project seeks to determine if their results were reproducible and improve upon their classifier by performing a more stringent model selection, feature selection, scaling, and implementing grid search cross-validation for hyperparameter optimization.

Background

Providing more in-depth background on the CMU researchers' project, their goals were to determine if EEG can detect confusion and, if so, if it could detect confusion better than human observers. Their end goal was to provide feedback about student confusion levels during remote learning. Their original study design setup consisted of 10 students wearing a single-channel MindSet headset (figures 1 and 2) and watching Massive Open Online Course video clips. These videos were assumed to fit into either "confusing" or "not-confusing" categories with confusing videos on topics such as quantum mechanics and stem cell research, and not-confusing subjects like geometry and algebra. They had students self-report their confusion after watching each video on a scale from 1 to 7, with 1 being least confusing and 7 being most. To convert these values to a binary "confusing/not confusing", the scores less than or equal to the median were mapped to "not confusing" and scores greater than the median were mapped to "confusing". They then performed undersampling on the larger class to ensure no class imbalance.

Figure 1. Person wearing a single-channel MindSet headset.

Figure 2. Image of the single-channel MindSet headset.

For their model, the CMU researchers used a Gaussian Naive Bayes Classifier because as a model it is good for sparse & noisy training sets. They trained for two separate targets, the predefined confusion, what the researchers thought would be confusing, and the student-defined confusion, what the students scored as being confusing or not. They chose not to perform feature scaling or any form of dimensionality reduction or feature selection. Their features were the mean values of the raw signal and the frequency bands of Delta, Theta, Alpha1, Alpha2, Beta1, Beta2 , Gamma1, Gamma2, Attention, and Mediation which the MindSet also collected.

Additionally, the researchers trained student-specific and student-independent models. Student-specific models can be thought of as personalized models for each student. A single student's data is used as the dataset, where half of their videos are used for training, and half for testing. Student-independent, on the other hand, is a more generalized approach where the model is trained on all available data. This was done through LOOCV (Leave-one-out cross-validation) where the training was done on all but one student, and the testing was done on that left-out student's data.

Methods

For our expansion of this research, we used multiple methods that led to improvements in accuracy. First, we decided to test multiple models to determine which performed best. We used a handful of classifiers that are generally considered to be the industry standard for classification tasks such as this one (Gong, D., 2022). The models we used are logistic regression, KNN, SVM, random forest, decision tree, and Naive Bayes to compare against their Naive Bayes results as a baseline.

We used grid search cross-validation on each model to select the optimal hyperparameters. Grid search is a method for systematically searching through a range of hyperparameter values to find the combination that results in the best performance. Cross-validation is a technique for estimating the performance of the model on unseen data by splitting the data into training and testing sets, with different data being used for different sets on each iteration. Grid search and Cross-Validation were implemented in a two-level system, with the first level of cross-validation being used to optimize accuracy, and the second level for the grid search that also incorporated cross-validation (visualization in Figure 3).

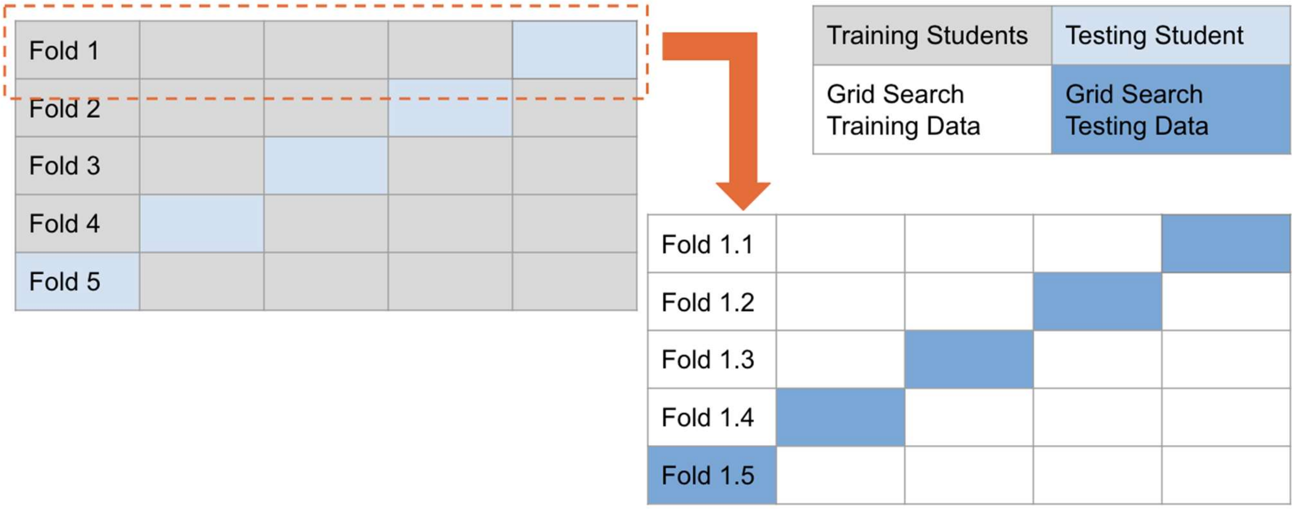


Figure 3. Visualization of the two-level cross validation and grid search system.

Additionally, we saw room for improvement in the data preprocessing steps. Specifically, we implemented a LASSO model for feature selection in our machine learning project. The LASSO (Least

Absolute Shrinkage and Selection Operator) is a regularization method used in regression analysis to prevent overfitting and improve the performance of the model (Tibshirani, 1996). By shrinking the coefficients of some of the features towards zero, LASSO can effectively reduce the complexity of the model and improve its interpretability.

To use LASSO for feature selection, we first fit a LASSO regression model to the training data using a range of regularization parameters. We then evaluated the performance of the resulting model using cross-validation and used the LASSO coefficients to rank the features according to their importance. The features with non-zero coefficients were selected as the most relevant ones. This approach allowed us to identify the most important features for our classifier, and improve its accuracy and interpretability.

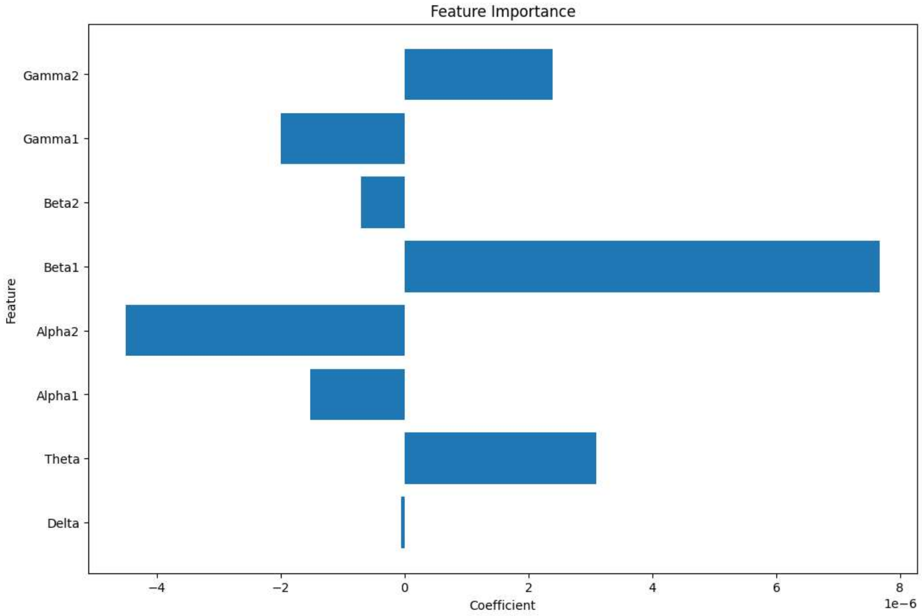


Figure 4. Feature Importance ascertained by fitting a LASSO model to the data.

We analyzed a correlation heatmap to inform our feature selection as well. Specifically, we looked for features highly correlated with the targets, but not highly correlated with one another as to avoid multicollinearity and the issues that come with it (Yoo, et. al, 2014).

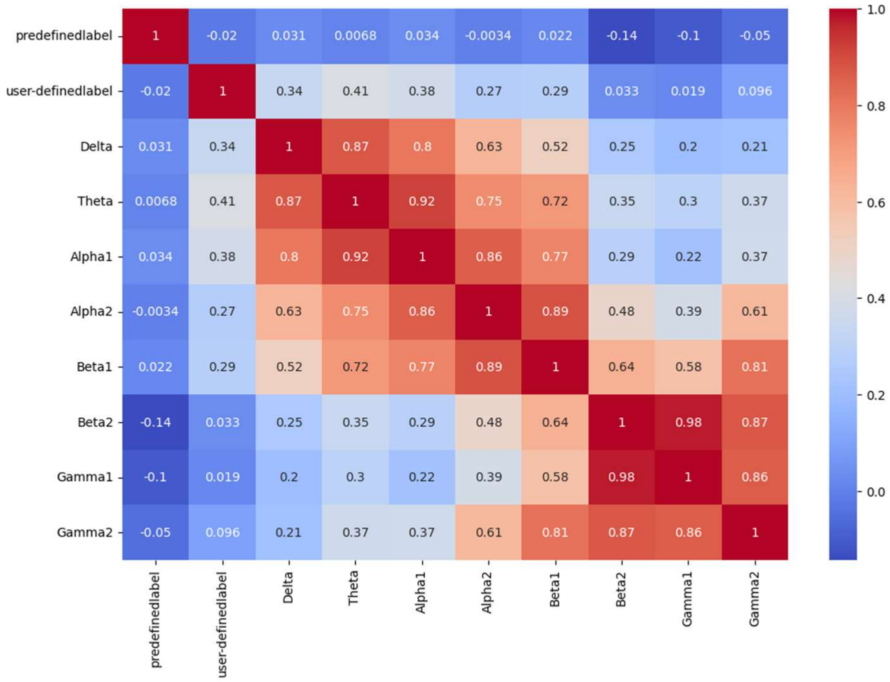


Figure 5. Correlation heatmap. This only represents a subset of the features - those with extremely low correlations were removed from the plot for readability.

Through our feature selection process, we decided to use a feature set of Beta1, Alpha1, Gamma2, and Theta. Unfortunately, most of the good candidate variables were highly correlated with one another. We believe the subset we chose balances the tradeoff between dealing with a moderate level of multicollinearity and using the best individual features for the target variables.

We also performed a separate feature investigation that we did not end up using in the final pipeline and results. The CMU researchers hypothesized that the theta band signal played an important role in being a good feature to distinguish between the two classes because in neuroscience the theta wave correlated with memory, learning, and spatial navigation. We were curious to see if we could generate statistical features from the theta band that would make good features. We generated the following statistical features from the theta band: variance of the theta band and the mean, minimum, maximum, and variance. of the theta power. We then performed LASSO Regression fit to determine if these new features were important in classification. The results can be seen in Figure 6 below and show that the newly generated theta features were not significant in classification, so they were not used.

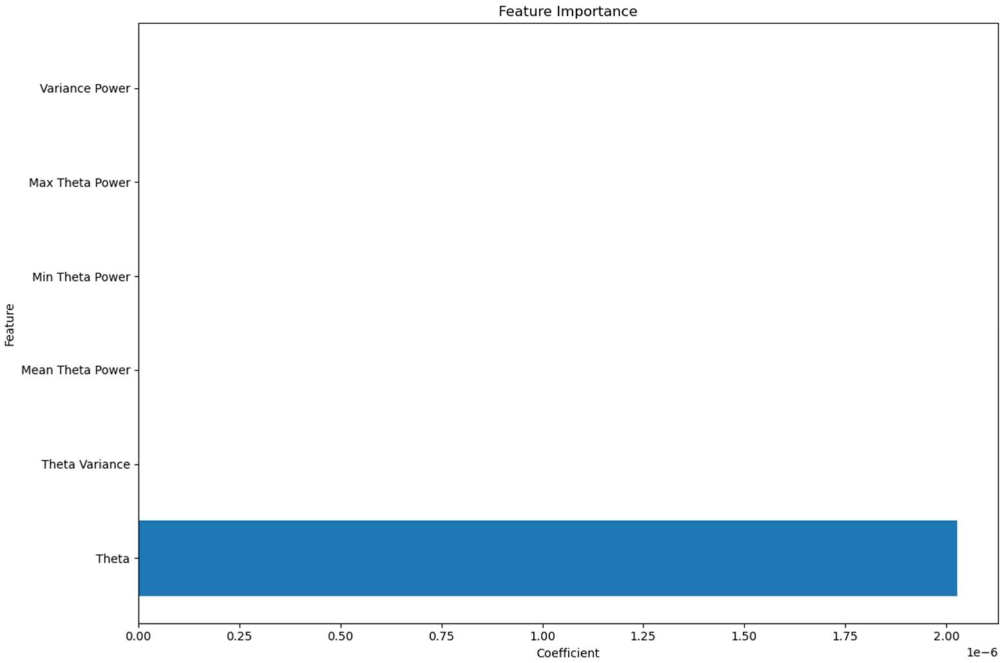


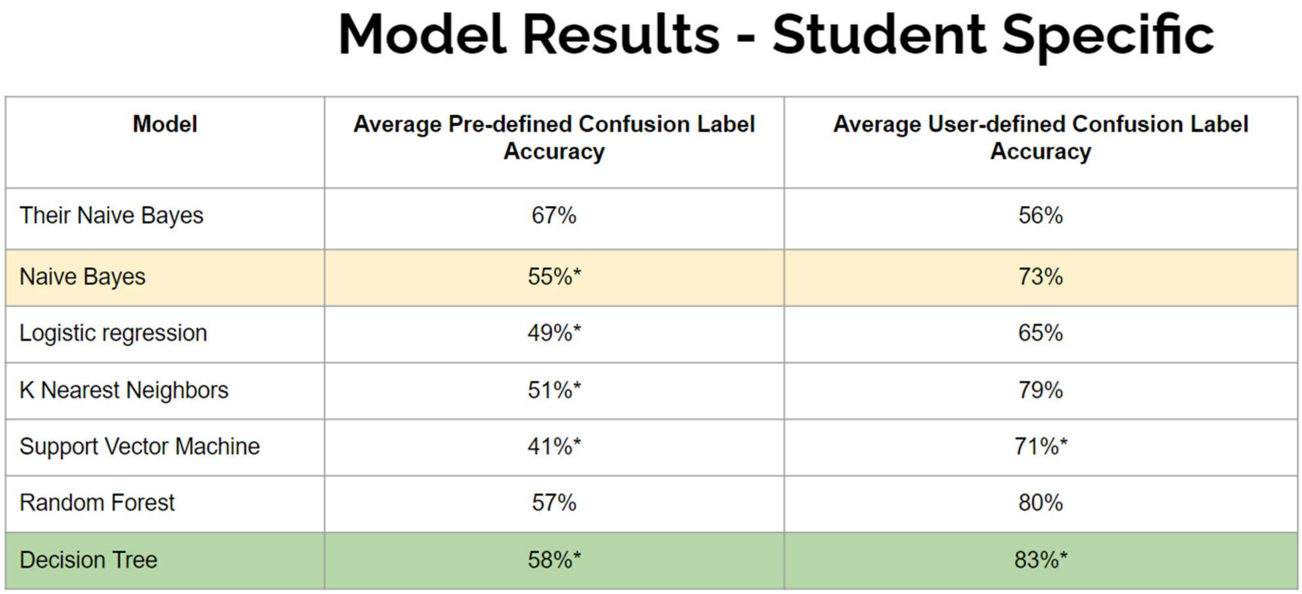
Figure 6. Feature Importance ascertained by fitting a LASSO model to the generated statistical theta band features.

We also scaled our data before training to ensure that all features were on a similar scale. Scaling of features is a common pre-processing step in machine learning that can help improve the performance, generalizability, and optimization speeds of models. We decided to scale via standardization, or mapping each variable to have zero mean and unit variance (Ahsan et. al, 2021).

To review our process, we decided to train six models. Each model was trained using the raw data, scaled data, feature-selected data, and scaled feature-selected data. Each model also had its hyperparameters optimized through grid search and its generalizability tested through cross-validation.

Results

We were able to see improvements in three out of the four cases. The one case where we could not improve the results of the paper was for the models predicting the predefined confusion label for student-specific training data. The original paper had a Gaussian Naive Bayes model that achieved 67% accuracy on its training set, whereas our best model (decision tree) achieved 58% (9% decrease). However, for student-specific models aimed at predicting student-defined confusion, we improved the model score from 56% to 83% (27% increase).



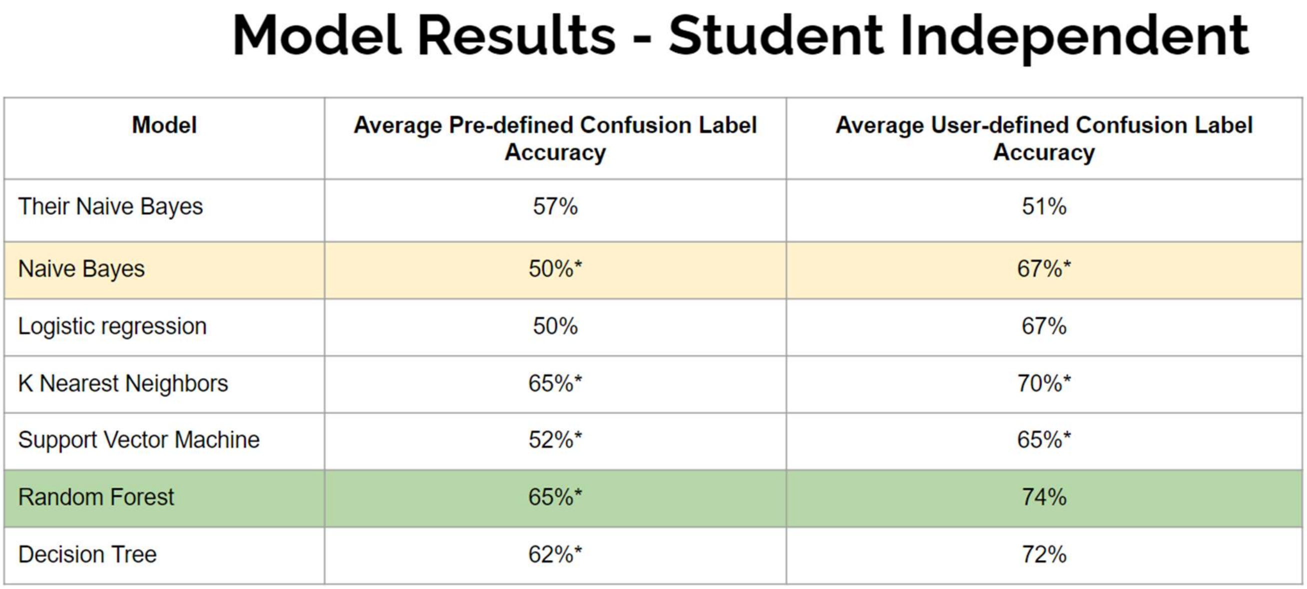
Model Results - Student Specific

Model	Average Pre-defined Confusion Label Accuracy	Average User-defined Confusion Label Accuracy
Their Naive Bayes	67%	56%
Naive Bayes	55%*	73%
Logistic regression	49%*	65%
K Nearest Neighbors	51%*	79%
Support Vector Machine	41%*	71%*
Random Forest	57%	80%
Decision Tree	58%*	83%*

Figure 7. Results for Student Specific models. Asterisks indicate that the model was improved through standardization, feature selection, or both.

Our decision tree was the best model for both predefined and student-defined confusion. The models that achieved the accuracies in Figure 7 were achieved by standardizing the data and using the entire feature set rather than selecting the subset of features we mentioned in Methods.

For the student-independent set of experiments, we were able to improve the accuracies of both predefined and student-defined confusion. For predefined, we increased the accuracy from 57% to 65% (8% increase). For student-defined, we increased the accuracy from 51% to 74% (23% increase). Both results were achieved with random forest models. The best random forest model for predefined confusion was trained on the standardized and feature-selected dataset. The best model for student-defined confusion, on the other hand, performed best when we used the entire feature set with no standardization.



Model Results - Student Independent

Model	Average Pre-defined Confusion Label Accuracy	Average User-defined Confusion Label Accuracy
Their Naive Bayes	57%	51%
Naive Bayes	50%*	67%*
Logistic regression	50%	67%
K Nearest Neighbors	65%*	70%*
Support Vector Machine	52%*	65%*
Random Forest	65%*	74%
Decision Tree	62%*	72%

Figure 8. Results for Student Independent models. Asterisks indicate that the model was improved through standardization, feature selection, or both.

While the random forest models performed the best, it takes significantly more computing power and time to run these models under the criterion that we chose. Figure 8 shows that we see moderately worse results from our decision tree and KNN models. This could be a valuable trade-off for future research, as both of these models take much less time to train than random forests.

We speculate that random forests and decision trees were the best performers because of their hierarchical approach to decision-making and how each branch works as an if-else statement. The way that decision trees and random forests branch out is similar to how humans make decisions which can make them more interpretable. Additionally, these models do not require a lot of data preprocessing, so it is likely that these models were not hindered by the single-channel data being very noisy as often seen in EEG data (The 365 Team, 2021).

We also wanted to see if our models were overfitting and evaluate their generalizability. We plotted the training and test accuracy for each cross-validation fold to see how well the accuracies lined up.

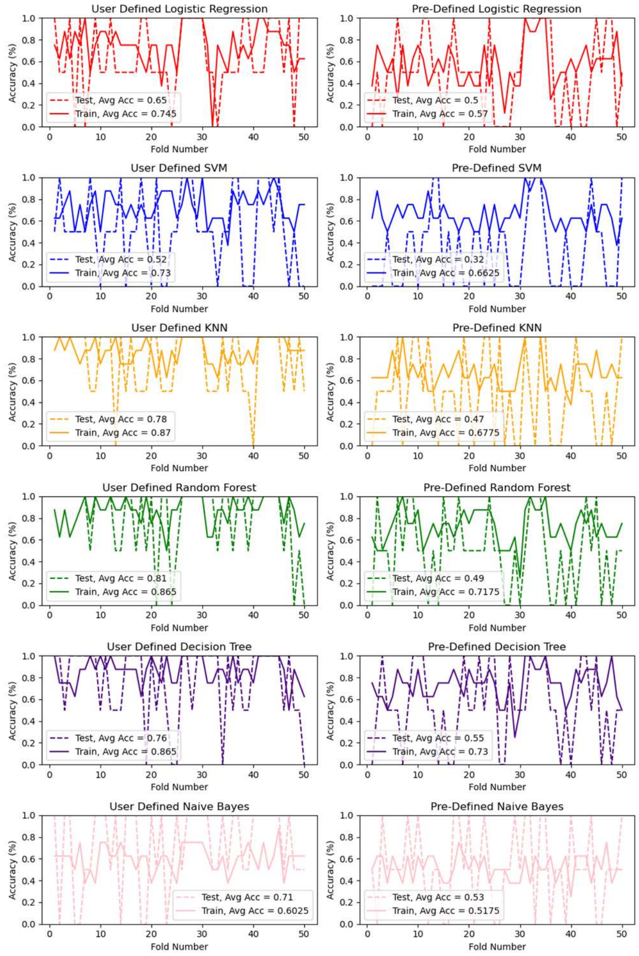


Figure 9. Student specific test versus training accuracy with non-regularized data.

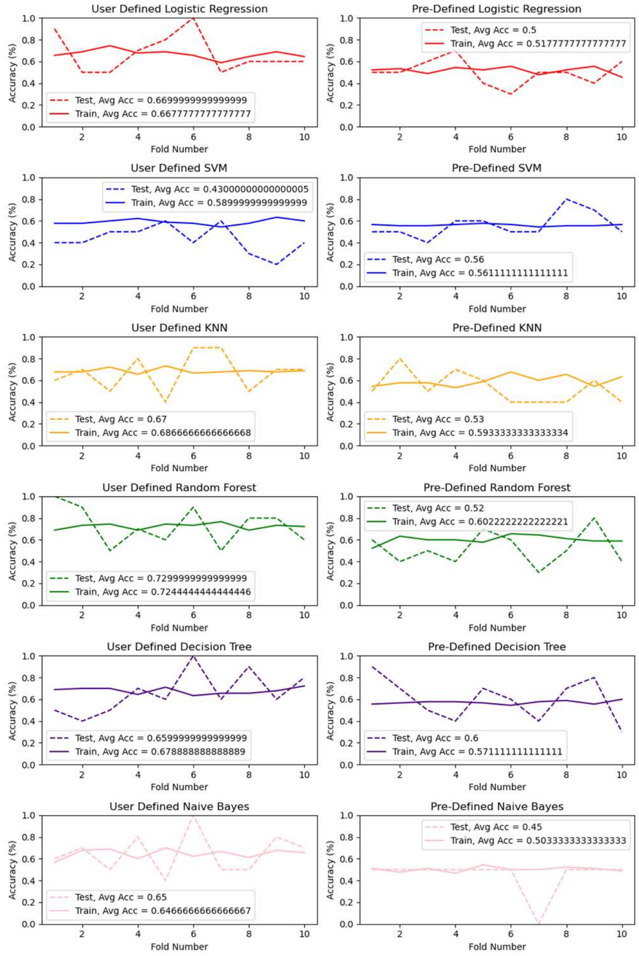
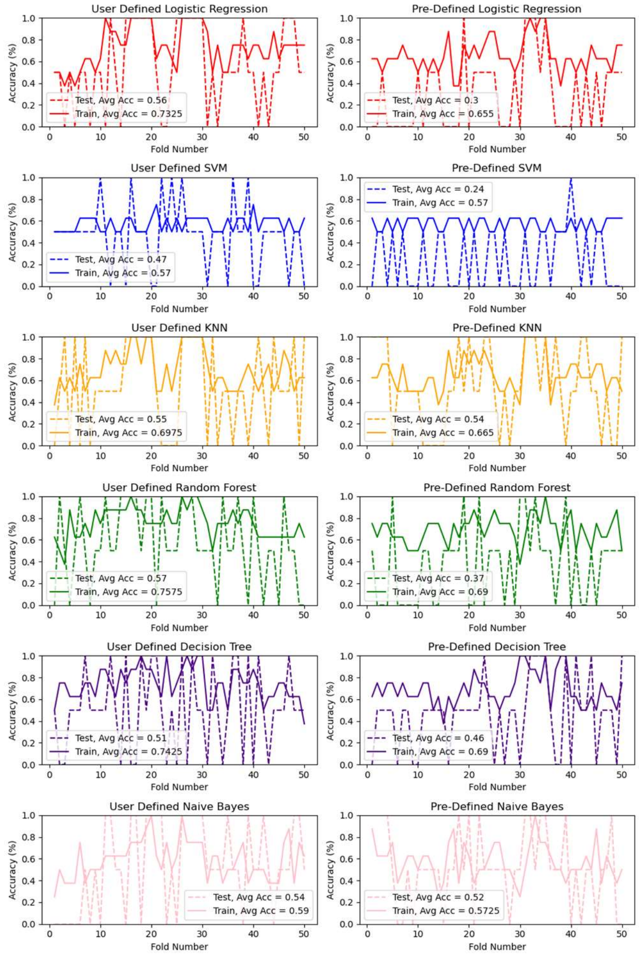


Figure 10. Student independent test versus training accuracy with non-regularized data.

The student-specific models were overfitting and not very generalizable. This was likely due to the small dataset for this model only working with a few videos for training and testing. The student-independent models were still susceptible to some overfitting, but were a lot better in their generalizability compared to the student-specific model. Overall, the random forest and decision tree models had the best accuracy/generalizability tradeoff.

Additionally, we wanted to see if regularizing the data helped with overfitting and generalizability. We reran the plots from above with the regularized data results to get the plots seen in Figures 11 and 12.



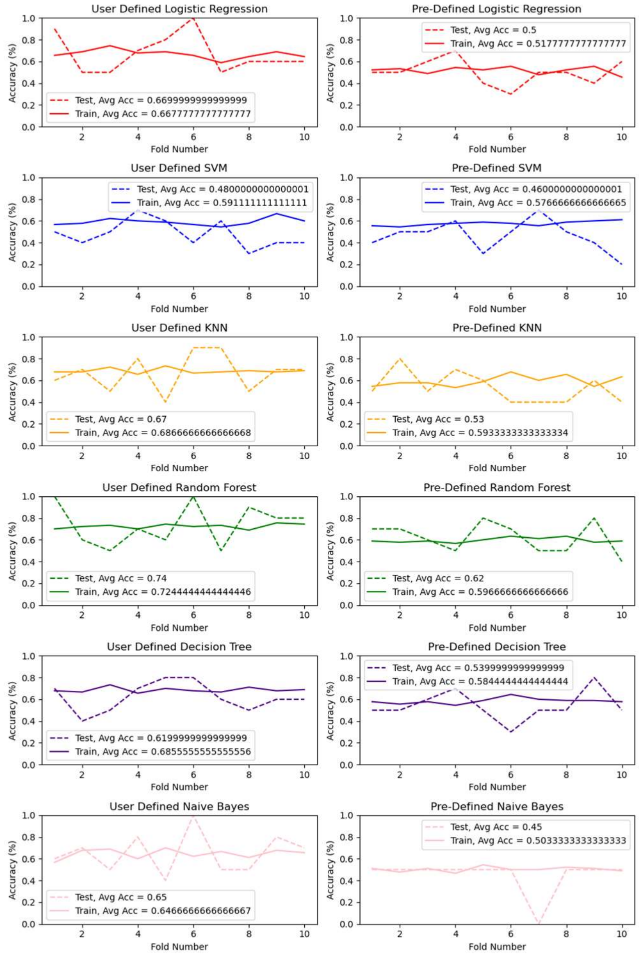


Figure 11. Student specific test versus training accuracy for each model with regularized data.

Figure 12. Student Independent test versus training accuracy with regularized data.

Overall, we saw a slight improvement in the generalizability of the user-defined models, enough to consider that standardizing the data is a necessary preprocessing step, especially when considering that average model accuracy improved as well with standardization.

Discussion

There are some limitations to be taken into consideration when discussing this kind of research. First, the data collection methods were imperfect. The original researchers only had 10 subjects, and one of the subjects ended up with corrupted, unusable data. The method of calculating the predefined confusion metric is very subjective and imprecise. A physics student, for example, may have a relatively easy time with the quantum mechanics videos meant to be confusing for everyone.

The lack of data also presents an issue when testing. We, like the original authors, ended up with a very small testing set to work with. We considered synthesizing data, but any reasonable amount of synthesizing would lead to too much of our data being made up. For example, each student watched 10 videos. If we synthesized even 5 extra samples for a student, ⅓ of that student's data would be statistically generated and, therefore, would lead to less confidence in our results.

Should future researchers look to improve upon our findings, it may be wise to re-collect the data with a larger population. We also believe the student-defined confusion to be a much better predictor of whether a student is confused, so the emphasis can solely be put on this target. If the appropriate resources are available, the models could also be strengthened significantly with a multi-channel headset to derive a wider set of features.

Citations

Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3), 52. MDPI AG. Retrieved from http://dx.doi.org/10.3390/technologies9030052

Gong, D. (2022, July 12). *Top 6 Machine Learning Algorithms for Classification*. Medium. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

The 365 Team. Introduction to Decision Trees: Why Should You Use Them? (2021, November 17). 365 Data Science. https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/

Wang, H., Li, Y., Hu, X.S., Yang, Y., Meng, Z., & Chang, K.K. (2013). Using EEG to Improve Massive Open Online Courses Feedback Interaction. *International Conference on Artificial Intelligence in Education*.

Yoo, W., Mayberry, R., Bae, S., Singh, K., Peter He, Q., & Lillard, J. W., Jr (2014). A Study of Effects of MultiCollinearity in the Multivariable Analysis. *International journal of applied science and technology*, *4*(5), 9–19.