# Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models

Debopam Sanyal
University of Illinois at Urbana–Champaign
Champaign, IL, USA
dsanyal2@illinois.edu

Nigel Bosch
University of Illinois at Urbana–Champaign
Champaign, IL, USA
pnb@illinois.edu

Luc Paquette
University of Illinois at Urbana–Champaign
Champaign, IL, USA
lpaq@illinois.edu

## ABSTRACT

Supervised machine learning has become one of the most important methods for developing educational and intelligent tutoring software; it is the backbone of many educational data mining methods for estimating knowledge, emotion, and other aspects of learning. Hence, in order to ensure optimal utilization of computing resources and effective analysis of models, it is essential that researchers know which evaluation metrics are best suited to educational data. In this article, we focus on the problem of wrapper feature selection, where predictors are added to models based on how much they improve model accuracy in terms of a given metric. We compared commonly-used machine learning algorithms including naive Bayes, support vector machines, logistic regression, and random forests on 11 diverse learning-related datasets. We optimized feature selection based on nine different metrics, then evaluated each to address research questions about how effective each metric was in terms of the others (e.g., does optimizing for precision also result in good F1?) as well as calibration (i.e., are predictions produced by models accurate probabilities of correctness?). We provide empirical evidence that the Matthews correlation coefficient (MCC) produced the overall best results across the other metrics, but that root mean squared error (RMSE) selected the best-calibrated models. Finally, we also discuss issues related to the number of features selected when optimizing for each metric, as well as the types of datasets for which certain metrics were more effective.

## Keywords

Feature selection, Metrics, Machine learning, Student models

## 1. INTRODUCTION

Machine learning is a popular method for building predictive models that automatically estimate various aspects of learning. These models, in turn, can be applied to study the processes of learning or teaching, or to automatically guide students as they learn. Training models is a complex process, however. The space of possible machine learning models is far too large to fully explore, and thus the search space is typically narrowed by focusing on candidate models that appear promising via some measure of correctness (agreement with ground truth labels, for supervised classification), such as Cohen's kappa or $F_1$ [16, 40]. One common methodological step that involves model selection (narrowing the search space) is *wrapper forward feature selection* [29], a process wherein features are added one at a time to a model based on which feature produces the largest gain in model correctness. Changing the correctness metric by which features are evaluated can have a significant impact on the final selected model (which we demonstrate in this paper); however, little is known about exactly what these impacts are for different correctness metrics. In this paper, we address this problem by performing feature selection based on different metrics and comparing the resulting models.

Previous work in the area of examining correctness metrics for educational data mining has largely focused on what those metrics reveal about models [40, 10]. Related work has shown, for example, that area under the receiver operating characteristic curve (AUC or AUROC) ignores the scale of model predictions [40], and that $F_1$ can be increased by over-predicting the positive class [10]. From such findings we can generate hypotheses about the properties of models that result from relying on those metrics during feature selection. For example, we might expect recall- and $F_1$-based feature selection to favor models that over-predict the positive class. However, there is little empirical evidence to support such hypotheses, which we aim to provide in this paper.

We explore a wide variety of correctness metrics for feature selection, evaluating them on 11 education-related datasets, to empirically measure relationships between feature selection metrics and resulting models. We include well-known and extensively-used metrics like AUC, Cohen's kappa, and others, as well as metrics that are less-commonly used but perhaps equally valuable, like the Matthews correlation coefficient and the minimum proper AUC. We experiment with metrics and datasets across four commonly-used machine learning classifiers, including support vector machine, naive Bayes, logistic regression and random forest. These algorithms have been frequently applied with great success in educational data mining and related research [24, 21, 43, 9], including in situations where high-dimensional data require feature selection [27, 49, 34].

To the best of our knowledge, ours is the first work to explicitly test differences between correctness metrics in the context of feature selection. Our results are valuable for future educational data mining research and practice by providing guidance to machine learning experts who wish to make evidence-based decisions about their model building methods. In particular, we characterize metrics in terms of the models that result from performing feature selection based on those metrics, which will help researchers decide on appropriate metrics based on the desired properties of their resulting models.

## 2. RELATED WORK

While previous research and other projects in this area is limited, there have been a few relevant research projects with findings that significantly informed our current work. In this section, we describe metrics evaluated in this study along with examples where they were used in previous work, then discuss directly-related work on evaluating metrics in educational data mining.

### 2.1 Metrics and their Usage

**Accuracy.** In this paper, accuracy refers to the proportion of correctly classified instances, though in other contexts it may refer more generally to any measure of how well a model's predictions align with ground truth values. Accuracy is one of the most straightforward metrics to calculate and understand, and thus has been reported frequently in machine learning studies [35, 12]. However, previous research has noted flaws with accuracy. In situations where labels are imbalanced, accuracy is often attenuated [25] or inflated [10] depending on the rate at which the model predicts the majority class. Despite possible flaws, it is commonly examined and is often the default correctness measure in machine learning software [39], including in wrapper feature selection software [41], so we include it in this paper.

**AUC.** AUC measures model correctness in terms of true positive rate across every possible false positive rate (i.e., across all possible decision thresholds). Chance level AUC is 0.5, while a perfect model has AUC = 1 and a completely incorrect model has AUC = 0. AUC is a valuable metric for its clear interpretability and effectiveness in the face of class imbalance [25], and has often been reported as an evaluation metric on educational datasets (e.g., [26, 23, 40, 37]). However, it only measures correctness in terms of the order of predicted values, not their scale [40], so it is unclear whether selecting features based on AUC will result in models that may have poorly-scaled predictions (an issue we explore in this paper). A related metric is the area under the precision–recall curve (AUPRC) [44], which also considers all possible decision thresholds. We have not yet included AUPRC in analyses, but expect that its behavior with respect to scale of predictions may be similar to AUC.

**MPAUC.** In situations where models provide only binary predictions, an approximation of AUC can be calculated by measuring the minimum proper AUC (MPAUC) of the quadrilateral formed by the single available decision threshold [38], as shown in Figure 1. We refer to this metric as MPAUC for the sake of brevity when reporting results, though it is not typically abbreviated in previous literature. It differs from AUC in that it measures the area for a "curve" defined by a single point instead of many points as in AUC. Its advantage is that it is applicable even when continuous decision thresholds are not available. MPAUC has been utilized as a metric for feature selection in prior educational data mining research [9], but it is unclear how it compares to alternatives we explore in this paper.

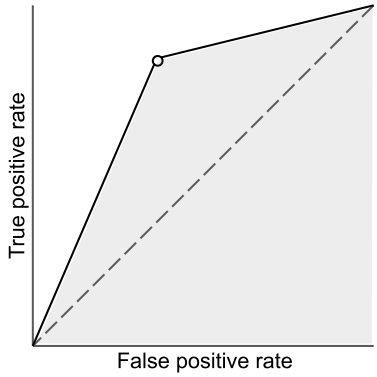


Figure 1: Example MPAUC (shaded area).

**MCC.** The Matthews Correlation Coefficient (MCC) measures the correlation between two binary variables (predicted labels and actual labels) [30], and is equivalent to Pearson's $r$ for two binary variables (i.e., ). MCC ranges from -1 to 1, where 0 indicates chance level and 1 indicates perfect classification. MCC is especially useful in binary classification models where there is class imbalance, since its chance level is not affected by imbalance. MCC is simply a correlation coefficient between the true and predicted class. It is only defined for binary variables. While it is not common in educational data mining research, it has been occasionally reported [8, 1] and is valued in other machine learning fields [15].

**Recall.** Recall is the proportion of a certain label class (typically the positive class) that was correctly identified as being in that class [46, 4]. Recall is an informative measure for understanding model correctness, especially in situations where it is important to focus on one class (e.g., in situations where false negatives are costly). However, it can be inflated by over-predicting the positive class [10] and is thus not often reported as the sole measure of model correctness, so it is unclear whether it is appropriate as a metric for feature selection.

**Precision.** Precision is similar to recall; it is the proportion of instances predicted as being in the positive class that were correct predictions. Like recall, it is typically only reported in conjunction with other correctness metrics, but unlike recall it cannot be inflated by over-predicting the positive class [10]. However, in some cases it can be maximized by predicting the positive class for only a few of the highest-confidence instances.

**$F_1$.** $F_1$ is defined as the harmonic mean of precision and recall, and thus avoids some of issues of recall (favoring over-prediction of the positive class) and precision (under-predicting the positive class). However, it can be inflated by over-predicting the positive class [10], so it is unclear

whether selecting features based on $F_1$ will favor models that over-predict the positive class or not.

**RMSE.** RMSE (root-mean-square error) measures the Euclidean distance between predictions and ground truth labels. Since RMSE is an error metric, lower values are better, with 0 indicating no error. It is commonly associated with regression problems, since it can be easily calculated for continuous labels, but is also effective for binary classification with models that produce continuous-valued probability predictions [40, 13]. Previous research has noted that RMSE is especially effective for optimizing probabilistic predictions [40]; thus, we expect that selecting features based on RMSE might also produce models with well-calibrated probabilities (where model confidence matches the probability that the model is correct). Like AUC, RMSE does not require setting a decision threshold, unlike the other metrics we consider in this study. We refrained from using close variants like Mean Absolute Deviation (MAD) or Error (MAE), since previous work has noted issues with these metrics for model selection [40].

**Kappa.** Cohen's kappa ( ) was developed as a measure of agreement between human annotators [16], but has often been utilized as a machine learning correctness metric by measuring the agreement between ground truth labels and predicted labels [10]. Like correlation measures, kappa ranges from -1 to 1 where 0 is random chance and 1 indicates perfect classification.

## 2.2 Research on Metrics in Educational Data Mining

Previous research has focused on metrics primarily in terms of the perspective that metrics have on a model of students, or on the properties of the model that are highlighted (or hidden) by particular metrics.

In one previous project, researchers focused on evaluating the properties of metrics that require continuous (probability-like) predictions [40]. In particular, they focused on AUC, RMSE, mean absolute error (MAE), and log likelihood (LL). They noted that for some applications (e.g., prediction of probability that a student has mastered a specific skill) metrics such as AUC do not favor well-calibrated models. They also compared metrics in terms of how often they agreed on picking the best model out of a pool of 20 simulated datasets, finding that RMSE and LL frequently agreed (17 out of 20) but others agreed much less often; the second-highest agreement was between RMSE and AUC, on 7 out of 20 datasets. This is especially relevant to the work in this paper, where we compare properties of metrics applied across 11 real-world datasets.

In similar previous work, researchers compared the properties of metrics that require binary or categorical predictions, rather than continuous predictions [10]. They noted that F1 is influenced by the base rate of the positive class in data, in line with other research on Cohen's kappa, AUC, and other metrics [25]. However, they also noted that F1 (and recall) are influenced by the predicted rate of classifiers. This finding is especially relevant to the current research because it is possible that feature selection will favor models and features that tend to predict more of the positive class when selecting based on these metrics.

# 3. FRAMING THE PROBLEM

The goal of this paper is, broadly speaking, to provide empirical results that illustrate the relationships 1) among different metrics, and 2) between metrics and models, when metrics are employed for forward feature selection.

Sequential feature selection is a type of wrapper (model-based) feature selection in which a feature is added to or removed from a model, the model is re-trained, and the quality of the feature in question is assessed based on improvement in model correctness (as measured by some metric). In this study, we specifically performed forward feature selection by adding one feature at a time, stopping when all features were added or when the model had not improved for three consecutive features, then returned the set of features with maximum correctness among all the combinations explored. Our work focuses primarily on the effects of utilizing different metrics for the step in which model correctness is assessed, which drives the entire feature selection process. We define four research questions (RQs) to explore this problem:

**RQ1: When selecting features based on a speci c metric, how do the results vary in terms of the other metrics?** Addressing this question will inform decisions about which metric to apply during feature selection by showing the relationships between metrics. For example, some low-cost applications may benefit from high recall (e.g., automatically selecting the most relevant material for students to review) while other higher-cost applications may require high precision (e.g., automatically predicting when a teacher should intervene to redirect learning behaviors). In these examples, we may wish to optimize feature selection for different metrics, but it is crucial to understand how that might influence other metrics; e.g., does optimizing feature selection for AUC tend to produce models that are also good in terms of Cohen's kappa, recall, and the other metrics?

To address RQ1 we define the *ranking* of a metric with respect to all the other metrics. Specifically, given a set of metrics $M$, a selection metric $X \ 2 \ M$ has rank 0 with respect to another metric $Y \ 2 \ M$ if selecting features based on $X$ results in the best[1] value of $Y$ compared to selecting features based on all other metrics in $M$. Likewise, a metric $Z \ 2 \ M$ has rank 1 with respect to $Y$ if selecting features based on $Z$ produces the second-best value of $Y$ compared to all other metrics in $M$, and so on. Generally, we expect that selecting features for some metric $X \ 2 \ M$ will have rank 0 with respect to itself ($X$), though this is not necessarily always true. Furthermore, some metrics may be generally better than others in terms of rank, if they tend to favor models with well-rounded properties that satisfy each metric. We thus calculate the *mean ranking* of each metric as the mean of all rankings for a metric with respect to itself and all other metrics (nine in total, in this paper), as a way to discover which feature selection metrics tend to yield models that satisfy the wide range of criteria imposed by different metrics.

[1]"Best" meaning highest for most metrics, but lowest for RMSE since it is an error metric.

RQ2: How do different feature selection metrics impact model calibration? As previous work noted, some metrics do not penalize models for being poorly calibrated [40]. However, it remains unclear how large of an effect using different metrics during feature selection may have on the calibration of the resulting model. We address this research question by calculating CAL scores (described in Sec. 4.4) for models selected based on each metric [12].

RQ3: How do different feature selection metrics impact the predicted rates of models? Certain correctness metrics favor over- or under-prediction of the positive class more than others. For example, accuracy for a problem with imbalanced classes can be increased simply by biasing predictions of the positive class in the same direction as the imbalance in the data [10]. We might expect that relying on accuracy for feature selection could thus result in models that over or under-predict the positive class, but it is unclear how problematic these effects may be, which we measure in addressing this research question.

RQ4: Do some feature selection metrics tend to result in more parsimonious models (fewer features) than others? In addressing this research question, we further characterize the models that result from applying different metrics during feature selection, and highlight cases where feature selection may fail (by selecting too few features) or unnecessarily increase model complexity (by selecting an unusually large number of features).

# 4. EXPERIMENTS

We performed a variety of experiments to address our research questions, consisting of training and testing machine learning classifiers with forward feature selection. Experiments required approximately 11 months of continuous run time[2], given that we performed extensive hyperparameter selection with 4 classifiers, 11 datasets, and 9 feature selection metrics, as detailed in this section.

## 4.1 Classifiers

As mentioned in the Introduction, we trained models including random forest, support vector machines, naive Bayes, and logistic regression. These machine learning algorithms represent a variety of methods with differing assumptions and levels of flexibility, and which are frequently employed in educational data mining research [18, 5, 21, 43, 20, 7, 11, 45]. Moreover, with the possible exception of random forest, these models quite often benefit from feature selection to avoid problems of over-fitting (e.g., when a logistic regression has nearly as many parameters as instances) [33] and collinearity (e.g., when two very similar features incorrectly double the impact of a relationship in a naive Bayes model).

## 4.2 Cross-validation

We utilized student-level four-fold cross-validation, training each model on data from 75% of students and testing it on the remaining 25% of students, then repeating a total of four times until each student was in the testing data exactly once. This procedure ensured that data from the same student was never present in training and testing at the same time, which was crucial given that some of our datasets had multiple instances per student.

We performed nested (within training data) student-level four-fold cross-validation for evaluating hyperparameters and selecting features. Specifically, for every possible combination of hyperparameters, we performed forward feature selection, then stored the best result from the feature selection process (according to the current selection metric). Finally, we retrained the model using the best set of hyperparameters, including the best features, on all training data, and applied it to the testing data. Hyperparameter selection and feature selection did not involve the testing set in any way.

There are two common strategies for evaluating the results of cross-validation. The first, *macro-level averaging*, consists of calculating the desired correctness metric for each fold and averaging across folds (four folds, in our case). The second strategy, *micro-level averaging*, involves storing the predictions of each fold and calculating the correctness metric once at the end based on all predictions. We evaluated both strategies to assess possible differences on the feature selection process.

## 4.3 Hyperparameters

We extensively tested common hyperparameters for each classification algorithm to ensure models had a chance to fit to the very different properties of our datasets (e.g., type of data, number of features, size of dataset).

For random forest we set the number of trees at 50 (significantly increasing this proved infeasible for an already-long run time). We varied the minimum number of samples required to create a branch in each tree, trying 5 different values (2, 4, 8, 16, or 32). This hyperparameter controls model complexity by restricting how fine-grained the decisions in each tree can be. We also varied the number of features randomly chosen for building each tree, testing 4 options including proportions of .25, .50, .75, and the square root of the number of features (the default setting). This hyperparameter controls how different trees are from each other in terms of the features from which they are trained. In total, there were 5  4 = 20 combinations of hyperparameters for random forest.

We trained SVMs with the radial basis function (RBF) kernel, which has a hyperparameter  that controls the size (radius of influence) of each RBF kernel. We tried values for  of 0.001, 0.01, 0.1, 1, and 10. Similarly, we tuned $C$, the SVM complexity hyperparameter, over the same set of 5 possible values. There were thus 5  5 = 25 hyperparameter combinations for SVM.

Naive Bayes has little in the way of hyperparameters to tune, apart from the distribution assumption to use. We assumed a Gaussian distribution for all models, and thus did not perform grid search across hyperparameters.

We trained logistic regression models with $L_2$ regularization, and tuned the strength of regularization as a hyperparameter over the space of 5 possible values: 0.001, 0.01, 0.1, 1, and 10.

[2]Experiments were run on an Intel Core i7 4.2 GHz processor (using a single core) with 32 GB memory and 256 GB storage.

Finally, we experimented briefly with hyperparameters related to class imbalance in the datasets, after noting that models frequently learned to only predict the majority class. We initially experimented with re-weighting instances of the minority class with higher weight set as a hyperparameter, but ultimately found that generating synthetic minority-class data via SMOTE (Synthetic Minority Over-sampling TEchnique [14]) was more consistently effective across our datasets without requiring hyperparameter tuning.

## 4.4 Measuring Model Calibration

Calibration refers to how well a model's predicted probabilities match the probability that those predictions are correct. For example, given a set of 100 instances where model predictions are all 0.7, we would expect 70 of the instances to be the positive class, and 30 to be in the negative class. If more than 70 are true positives, the model is under-confident for those 100 instances, while if fewer are true positives, the model is overconfident. Good model calibration is desirable so that predictions are interpretable as probabilities, allowing decision thresholds to be set in meaningful ways (e.g., triggering an intervention only if the model is at least 90% confident, knowing that it will thus result in a 10% false positive rate).

We measured calibration by calculating CAL scores [12]. The CAL score for a model is calculated by sorting all $N$ instances according to predicted probability, then dividing into $N - 99$ sliding windows of 100 instances (sliding by 1 instance). For each window, we calculated the absolute difference between the base rate of the positive class for those 100 instances and the mean predicted probability for the same instances. The CAL score consists of the mean of those absolute differences across all windows, and can be interpreted as the mean absolute error in model confidence.

## 4.5 Datasets

### *4.5.1 Video-based Engagement Detection Datasets*

We obtained six datasets from a study that measured students' self-reported engagement during an essay writing task [31], during which students' faces were recorded by a video camera. Students made verbal judgments of their engagement in the moment (concurrently) in response to auditory probes. One week later, they made retrospective judgments of their engagement by viewing video clips of themselves that were recorded during the essay writing task. There were 23 students who made a total of 530 judgments of engagement during the writing task and 1,325 retrospective judgments. Researchers extracted three sets of features from videos: 1) heart rate, estimated via photoplethysmography [32]; 2) animation units (ANUs), a set of facial feature descriptors provided by the Microsoft Kinect SDK, which are analogous to facial action units (AUs) [19]; and 3) local binary patterns in three orthogonal planes (LBP-TOP) [50], which capture facial textures and how those textures change over time.

There were thus two sets of labels and three sets of features, for a total of six video-related datasets. We refer to the two heart rate datasets as `video-hr-c` (concurrent labels) and `video-hr-r` (retrospective labels). Similarly, we refer to the two animation unit datasets as `video-anu-c` and `video-anu-r`, and the two LBP-TOP datasets as `video-lbp-c` and `video-lbp-r`.

### *4.5.2 Cognitive Tutor Algebra Datasets*

We obtained two datasets from a study [36] in which 59 students interacted with a computerized learning environment called Cognitive Tutor Algebra [3]. Students used Cognitive Tutor Algebra for an entire year as part of their regular mathematics curriculum. Researchers labeled 10,397 sequences of student actions in the learning environment for the presence of "gaming the system" behavior, where students attempt to progress through material by exploiting features of the learning environment (e.g., requesting hints repeatedly, guessing many answers) [6].

Researchers extracted two sets of features. Pattern features captured the presence or absence of 60 different sequences of actions that were designed to be similar to patterns identified by domain experts. We refer to the dataset with pattern features as `cta-pf` in this paper. The second set of features consisted of 25 count features. Count features captured the number of times 6 different actions occurred as well as the number of times 19 different events occurred. Events were identified by domain experts, and included things like pausing between attempts to answer a problem or trying to reuse an answer in multiple steps of a problem. We refer to the dataset with 25 count features as `cta-c` in this paper.

### *4.5.3 Student Survey Datasets*

Two additional datasets came from surveys obtained from 788 students at two different secondary schools during the 2005–2006 school year [17]. The survey consisted of 30 questions, including demographics, which school they attended (of two possibilities), and other variables. We one-hot encoded variables with categorical answers. Labels in both datasets consisted of course grades recorded on a 0–20 scale. We converted these to binary labels by splitting on the median into high and low grades, so that all datasets would be comparable binary classification problems.

One of the datasets came from students in a mathematics course (`math`, with 395 students) and the other from a Portuguese language class (`portuguese`, with 649 students). Some students were in both classes; thus, the total number of students was less than the sum of the classes.

### *4.5.4 Educational Process Mining Dataset*

We also extracted features from an educational process mining (`epm`) dataset. Students worked on electronics exercises in a software environment called DEEDS (Digital Electronics Education and Design Suite). Students' actions in the learning environment were timestamped and logged, and included mouse movements, keystrokes, and information about the exercises being solved. Grade data were provided for five learning sessions, from which we extracted features including time spent on activities, number of actions, mean, standard deviation, and other summary features from problem-level data. In total, 115 students participated, but grades and action log data were not available for all students in every session. Grades were recorded on a numeric scale, though we again converted these to classification problems via median split to maintain consistency with other datasets.

# 5. RESULTS AND DISCUSSION

We focus results on the four research questions outlined in Section 3; we also provide model correctness results in the

Appendix, but do not focus on these results here since the goal of this work is to compare metrics rather than focus on improving over previously-published models. Our experiments to address the research questions included 4 different machine learning algorithms, 2 methods of calculating results during cross-validation, and 11 datasets. The different machine learning algorithms yielded similar patterns for our primary research question (RQ1), with only a few exceptions (Figure 2). Similarly, results differed little across macro- and micro-averaging methods (Figure 3). Thus, we aggregated across classification algorithms and averaging methods to address our research questions without unnecessarily dividing results into 8 (2 averaging levels  4 classifiers) subsets.

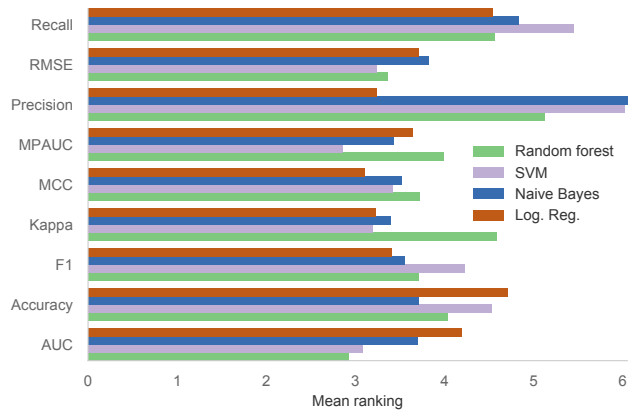


Figure 2: Mean ranking for each machine learning algorithm and feature selection metric. \Log. Reg." refers to logistic regression.

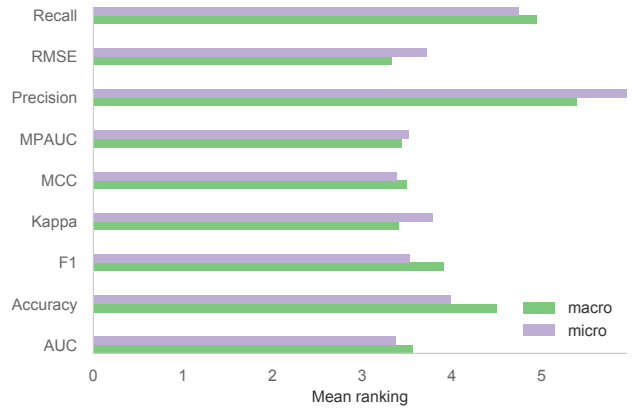


Figure 3: Mean ranking for feature selection metrics when calculating results via macro-level versus micro-level averaging.

## 5.1 Mean Rankings

RQ1 asks *When selecting features based on a specific metric, how do the results vary in terms of the other metrics?* Results in Table 1 show that MCC was, on average, the best (lowest) across models and datasets. Mean ranking for MCC averaged 3.441 across all datasets, while AUC and MPAUC were similar with mean rankings of of 3.468 and 3.476 respectively. Low rank for MCC indicates that, across 11 datasets, selecting features based on improvement in MCC yielded better results (in terms of itself and the other 8 metrics) than selecting features based on any of the other metrics. Specifically there were 3.441 correctness metrics on average for which selecting features based on some metric other than MCC yielded better results than MCC.

Conversely, precision was the worst-performing metric in terms of producing good results for other correctness metrics, with a mean ranking of 5.672. Recall and accuracy both had mean rankings above 4, while all other selection metrics had rankings  3.5.

There was also some notable variation across datasets. Possible causes of variations include the differing types of features in the datasets (binary, continuous, counts, etc.), class imbalance, and problem difficulty (e.g., signal to noise ratio). A handful of datasets had significantly lower mean rank values for a specific metric when compared other metrics and the average value across all datasets for the metric itself. For example, in the `portuguese` dataset, AUC was a particularly effective metric. AUC's mean ranking was 1.764, indicating that selecting features based on AUC in that dataset was almost always better (in terms of itself and the other metrics) than optimizing for those metrics was. In other datasets like `video-lbp-c`, the best metric had a much higher mean ranking. Similarly, metrics like $F_1$ and Accuracy had unusually low mean rank values for the `math` and `video-hr-r` datasets, respectively. In such cases, one metric did not frequently outperform the others.

We also explored RQ1 visually by counting the number of datasets for which each metric had at least a certain ranking or better (Figure 4), much like constructing a receiver operating characteristic curve requires finding predictions above every possible threshold. In Figure 4, higher curves are better, indicating that there were more datasets where the metric had a desirable ranking. The curve for precision was clearly lowest, followed by recall and then accuracy. The rest of the metrics were similar to one another, though the consistency of MCC is apparent from the fact that it was the first metric to achieve a certain ranking across all datasets.

## 5.2 Probability Calibration

RQ2 asks *How do different feature selection metrics impact model calibration?* The features that are selected can influence how well it is theoretically possible to calibrate a model. For example, a model with two binary features can only output four possible values, and thus it is quite likely the model will be unable to output predicted probabilities that closely align with the true probability that the model's prediction is correct or not.

Results show that RMSE easily produced the best results (Table 2), with a mean calibration score (CAL) of 0.166 and the best CAL score in 8 of the 11 datasets. Recall had the worst calibration score averaged across models and datasets, followed by precision, accuracy and $F_1$.

## 5.3 Positive Class Predicted Rate

RQ3 asks *How do different feature selection metrics impact the predicted rates of models?* The predicted rate of models is in some respects related to model calibration, since

Table 1: Mean ranking for each metric and dataset. Lower is better, indicating that a metric, on average, yielded better results in terms of itself and the other metrics. Values range from 0 (selecting features for that metric always produced the best score in terms of itself and the other metrics) to 9 (the number of metrics). The best metric for each dataset is highlighted in green, while the worst is in red.

| Dataset | Accuracy | AUC | F1 | Kappa | MCC | MPAUC | Precision | RMSE | Recall |
|---|---|---|---|---|---|---|---|---|---|
| cta-c | 6.319 | 2.653 | 2.153 | 2.681 | 3.167 | 4.181 | 4.778 | 3.528 | 6.542 |
| cta-pf | 5.403 | 2.222 | 4.903 | 4.625 | 3.986 | 2.208 | 6.069 | 3.736 | 2.847 |
| video-anu-c | 2.958 | 4.069 | 3.583 | 3.403 | 4.194 | 3.806 | 6.556 | 5.250 | 2.181 |
| video-hr-c | 3.847 | 5.111 | 4.514 | 3.764 | 3.556 | 4.181 | 5.153 | 2.333 | 3.542 |
| video-lbp-c | 3.806 | 3.389 | 3.986 | 3.528 | 3.319 | 3.986 | 5.875 | 4.097 | 4.014 |
| video-anu-r | 4.931 | 3.306 | 3.431 | 5.139 | 2.931 | 3.264 | 4.764 | 3.542 | 4.694 |
| video-hr-r | 2.000 | 4.389 | 4.111 | 4.333 | 3.583 | 4.722 | 5.319 | 3.306 | 4.236 |
| video-lbp-r | 3.833 | 2.361 | 6.458 | 2.528 | 4.056 | 3.069 | 4.597 | 3.306 | 5.792 |
| epm | 3.222 | 5.319 | 3.208 | 2.458 | 3.125 | 2.694 | 6.056 | 3.556 | 6.361 |
| math | 5.556 | 3.569 | 1.583 | 3.333 | 2.222 | 2.653 | 6.472 | 4.056 | 6.556 |
| portuguese | 4.819 | 1.764 | 3.028 | 3.792 | 3.708 | 3.472 | 6.750 | 2.125 | 6.542 |
| Mean | 4.245 | 3.468 | 3.723 | 3.598 | 3.441 | 3.476 | 5.672 | 3.530 | 4.846 |
| Std. dev. | 1.282 | 1.172 | 1.327 | 0.862 | 0.573 | 0.776 | 0.781 | 0.843 | 1.605 |

Table 2: Mean calibration score of each metric for each dataset. Lower is better, where 0 indicates that predicted probabilities exactly matched the probability that that model's predictions were correct. The best metric for each dataset is highlighted in green, while the worst is in red.

| Dataset | Accuracy | AUC | F1 | Kappa | MCC | MPAUC | Precision | RMSE | Recall |
|---|---|---|---|---|---|---|---|---|---|
| cta-c | 0.387 | 0.149 | 0.164 | 0.155 | 0.204 | 0.225 | 0.228 | 0.106 | 0.408 |
| cta-pf | 0.337 | 0.282 | 0.269 | 0.268 | 0.269 | 0.260 | 0.403 | 0.252 | 0.261 |
| video-anu-c | 0.271 | 0.281 | 0.281 | 0.263 | 0.278 | 0.261 | 0.320 | 0.257 | 0.271 |
| video-hr-c | 0.199 | 0.223 | 0.215 | 0.210 | 0.207 | 0.217 | 0.237 | 0.171 | 0.213 |
| video-lbp-c | 0.284 | 0.257 | 0.272 | 0.241 | 0.248 | 0.255 | 0.308 | 0.232 | 0.280 |
| video-anu-r | 0.235 | 0.217 | 0.233 | 0.228 | 0.220 | 0.223 | 0.235 | 0.214 | 0.257 |
| video-hr-r | 0.199 | 0.194 | 0.209 | 0.199 | 0.198 | 0.205 | 0.217 | 0.173 | 0.202 |
| video-lbp-r | 0.211 | 0.193 | 0.239 | 0.201 | 0.219 | 0.214 | 0.249 | 0.199 | 0.249 |
| epm | 0.067 | 0.147 | 0.069 | 0.060 | 0.066 | 0.071 | 0.099 | 0.063 | 0.184 |
| math | 0.086 | 0.132 | 0.138 | 0.141 | 0.137 | 0.130 | 0.083 | 0.091 | 0.137 |
| portuguese | 0.072 | 0.114 | 0.131 | 0.113 | 0.140 | 0.117 | 0.133 | 0.066 | 0.207 |
| Mean | 0.213 | 0.199 | 0.202 | 0.189 | 0.199 | 0.198 | 0.228 | 0.166 | 0.243 |
| Std. dev. | 0.106 | 0.059 | 0.068 | 0.065 | 0.063 | 0.063 | 0.096 | 0.073 | 0.070 |

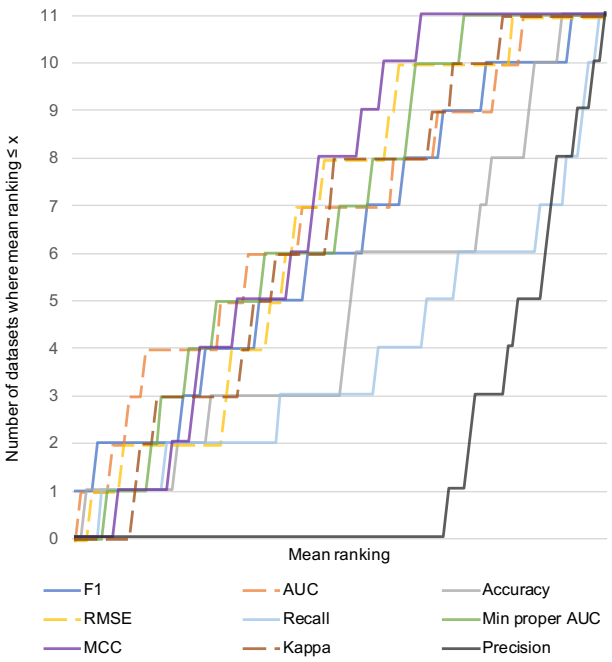


Figure 4: Step graph for mean rankings of metrics used for wrapper feature selection across all datasets. The left edge of the $x$ axis indicates the best (lowest) ranking, while the right indicates the worst (highest). The $y$ axis indicates the number of datasets that have mean rank $x$.

a model that severely over- or under-predicts the positive class is unlikely to be well-calibrated (e.g., a model that always predicts 100% confidence for the positive class will have very poor calibration for any negative-class instances). Results reflect this calibration–predicted-rate relationship (Table 3), showing that selecting features based on recall resulted in the largest mean absolute difference between actual base rate and predicted rate (0.233), while RMSE was close to best (0.080). Selecting features based on accuracy (proportion correct) did not produce inaccurate predicted rates (mean absolute difference = 0.079), however, despite relatively poor model calibration.

For imbalanced datasets where classification is imperfect, accuracy can be inflated by over-predicting the majority class [10, 25]. However, Table 3 shows that selecting features based on accuracy did not have this effect, perhaps because we applied SMOTE to reduce the impact of class imbalance during training. Conversely, selecting features based on recall increased the positive class predicted rate for most datasets, since doing so can inflate recall regardless of the presence of class imbalance [10]. Similarly, selecting features based on precision often resulted in under-prediction of the positive class (10 out of 11 datasets).

## 5.4 Number of Features Selected

Selecting features based on precision yielded the fewest numbers on average (4.173), while selecting based on RMSE yielded the most (10.523). Selecting features based on AUC also yielded more features (10.006, on average) than other metrics except RMSE.

These patterns are likely due to the fact that adding relatively unimportant features to a model will offer only marginal improvement, and may not be enough to shift predictions above or below the decision threshold. All of the metrics that require a decision threshold (accuracy, $F_1$, kappa, MCC, MPAUC, precision, and recall) resulted in fewer features than the threshold-free metrics of AUC and RMSE. For example, adding a feature that applies to only a few instances may help push the probability decision for those few instances in the right direction, but may not change the binary decision for those instances and thus may not be selected when evaluating based on threshold-based metrics.

# 6. LIMITATIONS AND FUTURE WORK

There are a few limitations to the experiments in this paper. First, the datasets that we analyzed represent only a handful from among thousands of educational datasets that researchers and others have collected over the years. Our datasets are also quite diverse, measuring very different student characteristics. Thus, we have only a sparse sampling of the space of educational datasets, and datasets that vary notably from those reported on here could exhibit different trends. Future work is especially needed in this area to discover specific properties of datasets (e.g., number of features, type of features) that inform which metrics are likely to be successful for wrapper feature selection. Such analysis is only possible with a large enough number of datasets to enable statistical comparisons at the dataset level.

Second, the metrics we examined also only represent a subset of many possible. Many other metrics are closely related to those we studied (e.g., informedness, markedness, balanced accuracy), but may not exhibit exactly the same patterns. We selected a diverse mix of commonly reported metrics and some less-common metrics, all of which have been shown to be useful in previous research.

Third, we explored only four of the most prominent machine learning classifiers from among many possible options. We chose these classifiers because they are represented in many education-related research endeavors, but results for other classifiers may differ. Perhaps most importantly, deep neural networks are increasingly popular for educational data mining research [2, 28, 48, 47, 42], but were not considered here. Wrapper feature selection is perhaps less common for deep neural networks, given the high computational cost of model training, but correctness metrics often play a similar role in the model selection process for neural networks – for example, when deciding when to stop training a model. In future work we will explore issues of model selection for neural networks as well.

Fourth, averaging across the four classifiers is a limitation as well. While classifiers performed somewhat similarly, Figure 3 shows some exceptional cases. For example, kappa performed poorly with random forest, and precision performed well with logistic regression. As part of future work, we will explore classifier-based analysis of metrics in more depth, including statistical analyses (e.g., Friedman test) where we consider a large number of classifiers as judges that are ranking metrics.

Table 3: Mean predicted rate of the positive class for models with features selected based on each metric, for each dataset. Base rate indicates the actual proportion of the positive class in the dataset. The last row refers to the mean absolute di erence between predicted rate and base rate across datasets. Green highlighting indicates the closest match to the true base rate, while red indicates the predicted rate furthest away in each row.

| Dataset | Base rate | Accuracy | AUC | F1 | Kappa | MCC | MPAUC | Precision | RMSE | Recall |
|---|---|---|---|---|---|---|---|---|---|---|
| cta-c | 0.068 | 0.060 | 0.179 | 0.170 | 0.169 | 0.215 | 0.250 | 0.149 | 0.111 | 0.695 |
| cta-pf | 0.068 | 0.029 | 0.084 | 0.084 | 0.084 | 0.084 | 0.087 | 0.003 | 0.064 | 0.085 |
| video-anu-c | 0.776 | 0.612 | 0.502 | 0.591 | 0.562 | 0.534 | 0.554 | 0.353 | 0.533 | 0.557 |
| video-hr-c | 0.776 | 0.669 | 0.688 | 0.669 | 0.724 | 0.705 | 0.703 | 0.681 | 0.753 | 0.663 |
| video-lbp-c | 0.776 | 0.631 | 0.526 | 0.586 | 0.617 | 0.563 | 0.548 | 0.385 | 0.588 | 0.607 |
| video-anu-r | 0.733 | 0.610 | 0.637 | 0.567 | 0.594 | 0.616 | 0.611 | 0.581 | 0.657 | 0.568 |
| video-hr-r | 0.733 | 0.718 | 0.658 | 0.703 | 0.690 | 0.705 | 0.683 | 0.627 | 0.732 | 0.702 |
| video-lbp-r | 0.733 | 0.590 | 0.614 | 0.529 | 0.610 | 0.572 | 0.560 | 0.389 | 0.629 | 0.590 |
| epm | 0.237 | 0.312 | 0.405 | 0.321 | 0.309 | 0.319 | 0.331 | 0.214 | 0.315 | 0.585 |
| math | 0.410 | 0.373 | 0.505 | 0.627 | 0.491 | 0.537 | 0.552 | 0.120 | 0.484 | 0.730 |
| portuguese | 0.425 | 0.437 | 0.524 | 0.609 | 0.509 | 0.573 | 0.532 | 0.085 | 0.473 | 0.840 |
| Mean $j\Delta j$ | | 0.079 | 0.126 | 0.135 | 0.098 | 0.123 | 0.128 | 0.210 | 0.080 | 0.233 |

Table 4: Number of features in each dataset (N) and mean number of features selected by each metric. The highest number of selected features for each dataset is highlighted in light blue, while the lowest is highlighted in gray.

| Dataset | N | Accuracy | AUC | F1 | Kappa | MCC | MPAUC | Precision | RMSE | Recall |
|---|---|---|---|---|---|---|---|---|---|---|
| cta-c | 25 | 2.531 | 10.313 | 8.781 | 8.750 | 5.750 | 4.094 | 7.188 | 12.969 | 2.875 |
| cta-pf | 60 | 10.469 | 35.219 | 25.125 | 24.625 | 26.125 | 28.156 | 1.000 | 28.094 | 27.969 |
| video-anu-c | 42 | 3.656 | 5.031 | 3.875 | 4.500 | 4.531 | 4.625 | 3.219 | 5.438 | 4.031 |
| video-hr-c | 7 | 3.313 | 3.188 | 2.594 | 3.563 | 3.875 | 3.531 | 2.750 | 4.094 | 2.969 |
| video-lbp-c | 2304 | 3.563 | 6.344 | 4.031 | 5.750 | 5.781 | 4.844 | 2.000 | 7.656 | 3.625 |
| video-anu-r | 42 | 4.281 | 6.063 | 5.063 | 5.594 | 4.906 | 4.813 | 3.875 | 7.688 | 4.281 |
| video-hr-r | 7 | 3.531 | 3.563 | 3.031 | 3.938 | 3.781 | 3.563 | 3.938 | 4.906 | 2.719 |
| video-lbp-r | 2304 | 8.344 | 12.656 | 6.500 | 9.125 | 9.500 | 9.469 | 6.500 | 16.781 | 4.750 |
| epm | 38 | 6.375 | 6.344 | 6.219 | 7.125 | 6.688 | 5.813 | 7.156 | 7.406 | 1.000 |
| math | 43 | 7.000 | 8.719 | 6.438 | 8.656 | 7.781 | 7.781 | 4.313 | 8.250 | 1.375 |
| portuguese | 43 | 10.844 | 12.625 | 7.719 | 11.344 | 9.531 | 9.719 | 3.969 | 12.469 | 1.313 |
| Mean | 446.818 | 5.810 | 10.006 | 7.216 | 8.452 | 8.023 | 7.855 | 4.173 | 10.523 | 5.173 |

# 7. CONCLUSION

As the field of educational data mining develops, and machine learning becomes increasingly popular for modeling student outcomes, it is imperative to deeply understand each step of the process and the influence researchers' choices have on models. Our experiments offer insight into the large differences that can arise from machine learning design decisions, specifically for feature selection. We showed that selecting features based on some metrics is rarely advisable (especially precision), and that the choice of metric has impacts not only on correctness measures but on other important properties of the resulting models, including calibration and size (number of features).

We found that MCC produced the overall best results across the other metrics in terms of mean ranking as a measure of well-rounded correctness across metrics. MCC was not the best selection metric for all the datasets; in fact, it was the most effective only for 3 of the 11 datasets we analyzed in this study. However, it was more consistently well-ranked than the other metrics. On the other hand, RMSE produced the best-calibrated models, which can also be an important consideration for applying student models that might benefit from easily-adjustable decision thresholds.

Student models are the driving forces in adaptive learning software. Thus, enhancing them will lead to better software for students and teachers. The results of this project will enable researchers to more accurately build models which predict student outcomes by informing the correctness metrics relied upon for feature selection. In particular, we suggest utilizing metrics like MCC and RMSE (if calibration is desirable) to yield models with well-rounded accuracy across metrics. We suggest avoiding recall, precision, and accuracy, even though accuracy is the default setting in some machine learning software.

# 8. REFERENCES


[1] R. Ade. Students performance prediction using hybrid classifier technique in incremental learning. *International Journal of Business Intelligence and Data Mining*, 15(2):173–189, Jan. 2019.

[2] F. Ai, Y. Chen, Y. Guo, Y. Zhao, Z. Wang, and G. Fu. Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 240–245, 2019.

[3] V. Aleven, B. McLaren, I. Roll, and K. Koedinger. Toward tutoring help seeking. In *International Conference on Intelligent Tutoring Systems*, pages 227–239. Springer, 2004.

[4] H. Almayan and W. Al Mayyan. Improving accuracy of students' final grade prediction model using pso. In *2016 6th International Conference on Information Communication and Management (ICICM)*, pages 35–39. IEEE, 2016.

[5] E. A. Amrieh, T. Hamtini, and I. Aljarah. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8):119–136, 2016.

[6] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 383–390, New York, NY, USA, 2004. ACM.

[7] R. S. Baker and P. S. Inventado. Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer, 2014.

[8] N. Bosch, R. W. Crues, G. M. Henricks, M. Perry, L. Angrave, N. Shaik, S. Bhat, and C. J. Anderson. Modeling key differences in underrepresented students' interactions with an online STEM course. In *Proceedings of TechMindSociety '18*, pages 6:1–6:6, New York, NY, 2018. ACM.

[9] N. Bosch and S. K. D'Mello. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, in press.

[10] N. Bosch and L. Paquette. Metrics for discrete student models: Chance levels, comparisons, and use cases. *Journal of Learning Analytics*, 5(2):86–104, 2018.

[11] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, 66:541–556, 2018.

[12] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 69–78, New York, NY, 2004. ACM.

[13] P. Chaudhury, S. Mishra, H. K. Tripathy, and B. Kishore. Enhancing the capabilities of student result prediction system. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pages 1–6, 2016.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2011.

[15] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, Jan. 2020.

[16] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[17] P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*. EUROSIS-ETI, 2008.

[18] T. Devasia, T. Vinushree, and V. Hegde. Prediction of students performance using educational data mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 91–95. IEEE, 2016.

[19] P. Ekman and W. V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists

Press, 1978.

[20] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang. Predicting students performance in educational data mining. In *2015 International Symposium on Educational Technology (ISET)*, pages 125–128. IEEE, 2015.

[21] W. Hämäläinen and M. Vinni. Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, pages 57–71, 2011.

[22] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali. Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52(1):381–407, 2019.

[23] M. Hussain, W. Zhu, W. Zhang, J. Ni, Z. U. Khan, and S. Hussain. Identifying beneficial sessions in an e-learning system using machine learning techniques. In *2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 123–128. IEEE, 2018.

[24] D. Ifenthaler and C. Widanapathirana. Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning*, 19(1-2):221–240, 2014.

[25] L. A. Jeni, J. F. Cohn, and F. De la Torre. Facing imbalanced data–Recommendations for the use of performance metrics. In *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction*, pages 245–251, Sept. 2013.

[26] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, 5(3):597–610, 2012.

[27] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[28] B.-H. Kim, E. Vizitei, and V. Ganapathi. GritNet: Student performance prediction with deep learning. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*, pages 625–629, 2018.

[29] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[30] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, Oct. 1975.

[31] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28, 2016.

[32] H. Monkaresi, R. A. Calvo, and H. Yan. A machine learning approach to improve contactless heart rate monitoring using a webcam. *IEEE Journal of Biomedical and Health Informatics*, 18(4):1153–1160, July 2014.

[33] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004)*, ICML '04, pages 78–85, New York, NY, 2004. ACM.

[34] R. Nilsson, J. M. Pena, J. Björkegren, and J. Tegnér. Evaluating feature selection for svms in high dimensions. In *European Conference on Machine Learning*, pages 719–726. Springer, 2006.

[35] Z. Papamitsiou and A. A. Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[36] L. Paquette and R. S. Baker. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments*, 27(5-6):585–597, 2019.

[37] L. Paquette, N. Bosch, E. Mercier, J. Jung, S. Shehab, and Y. Tong. Matching data-driven models of group interactions to video analysis of collaborative problem solving on tablet computers. In J. Kay and R. Luckin, editors, *Proceedings of the 13th International Conference of the Learning Sciences (ICLS) 2018, Volume 1*, pages 312–319, London, UK, 2018. International Society of the Learning Sciences.

[38] S. Parodi, V. Pistoia, and M. Muselli. Not proper roc curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC bioinformatics*, 9(1):410, 2008.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Nov. 2011.

[40] R. Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 7(2):1–19, 2015.

[41] S. Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), Apr. 2018.

[42] J. M. Reilly and C. Dede. Exploring stealth assessment via deep learning in an open-ended virtual environment. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, pages 643–646, 2019.

[43] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura. Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472, 2013.

[44] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3), 2015.

[45] F. Siraj and M. A. Abdoulha. Uncovering hidden information within university's student enrollment data using data mining. In *2009 Third Asia International Conference on Modelling & Simulation*, pages 413–418. IEEE, 2009.

[46] N. Tasnim, M. K. Paul, and A. S. Sattar. Performance