An Analysis of Educational Diagnostic Questions

Bhavana Jonnalagadda Erik Whitfield

2023-12-13

Table of contents

Abstract

We investigate the validity of diagnostic test items used in the NeurIPS 2020 Education Challenge, which aimed to improve technologies for developing individualized learning resources. The focus is on understanding students' response patterns and assessing the quality of diagnostic questions. We address research questions related to the evidence supporting subject domains in the test items and the ability of machine learning models to predict student performance for informed educational decision-making. The analysis involves clustering methods, such as k-prototypes and Gower Distance, to assess the similarity of factors in the dataset. Additionally, several supervised learning models are applied to predict student performance, including logistic regression, support vector machines, and gradient boosting. The findings suggest that machine learning models could provide insights into students' learning and performance, contributing to the potential for precision educational interventions. However, limitations, such as limited demographic information, lack of student response times, and unknown student characteristics, highlight areas for improvement and future studies in the field.

1 Introduction

1.1 Background

In education, diagnostic testing plays a significant role in helping educators understand the needs and capabilities of students. It allows educators to tailor their teaching strategies to fit the unique circumstances of different learners, and it assists them in identifying areas in which students may need special interventions (Cronbach, 1942). Failure to draw attention to students' weaknesses as early and accurately as possible can have consequences that reach far beyond the context in which a diagnostic test is administered. Consequently, the questions of how to devise diagnostic testing items that are responsive to students' abilities and how to best interpret diagnostic data that has already been collected are not trivial.

In the context of educational measurement and assessment, the validity of a test is dependent on the degree to which it actually measures what it claims to measure (Borsboom et al., 2004). A test is only valid if it is sensitive to varying levels of some latent trait that causally produces the outcomes observed in test scores. If one can show that performance on a test is likely driven by variance in some attribute that is irrelevant to the construct that the test purports to measure, then the validity of the test is threatened. The importance of test validity in education reaches far beyond any particular testing instrument. In a broader sense, validity can be understood as referring to the extent to which evidence supports the interpretations and suggested uses of test scores (Cronbach & Meehl, 1966).

A necessary condition for the reasonable use of any test is a convincing body of evidence for the validity of a particular interpretation of test scores. In this regard, there are several ways that validity evidence can be weak or otherwise insufficient (Borsboom et al., 2004). For instance, a test might fail to capture some relevant aspects of the construct it was designed to measure. A test might also be exceedingly susceptible to the influence of processes that are entirely unrelated to its intended purpose. For example, a reading comprehension test would be subject to accusations of construct-irrelevance if the content of the test elicited an emotional reaction that interfered with the test takers performance. There are numerous other threats to test validity, but what they have in common is that they call into question the claims and decisions that are made based on test results. The question of validity is especially relevant in diagnostic testing, the purpose of which is to inform decisions about the path that a student's instructional plan should follow. This is the topic with which the NeurIPS 2020 Education Challenge was concerned (Wang, Lamb, Saveliev, Cameron, Zaykov, Hernández-Lobato, et al., 2021). In an effort to improve the technologies that are used to develop individualized learning resources for students, the competition tasked participants with devising new methods to understand students' response patterns and assess the quality of diagnostic questions. The challenge utilized data from an online education service provider called Eedi consisting of tens of thousands of multiple-choice questions that were administered over two years to a range of students from elementary to high-school grades. The test items were all targeted to different skills in mathematics, including Algebra, Geometry, and Statistics. The matter of validity is immediately relevant to the challenge. In order for the instructional decision making that results from the administration of diagnostic tests to be sound, the tests themselves must be valid. In order to be valid, the diagnostic test questions need to capture variation in respondents' mathematical abilities. The diagnostic items can be considered valid if there is evidence that they do in fact measure what the test administrators claim that they measure.

1.2 Research Questions

This analysis intends to test the validity of the test items used in the challenge and look for construct-irrelevant factors that may have influenced the responses to test items. We take up the following research questions:

- Do the item response data provide evidence to support the subject domains present in the test?
- Can relatively accessible machine learning models predict student performance with enough precision to inform educational decision making?

2 Methods

2.1 Data

The data used is from the NeurIPS 2020 Education Challenge (Wang et al., 2020), which is in the format of question-answer pairs of mathematical questions posed to students and their answers (and demographic). There are more than a 200 million data points in the full dataset, so we use a subset of about only a million data points. We join across the multiple tables that the data is present in, and combine them in order to have full information for each data point.

The data used format can be seen in Table 2.1. We dropped unused columns (`CorrectAnswer, AnswerValue, SchemeOfWorkId`), and the transformations for feature engineered columns are listed in the descriptions. The description of the columns used are as follows:

- **QuestionId**: ID of the question answered. Numeric.
- **UserId**: ID of the student who answered the question. Numeric.
- **AnswerId**: Unique identifier for the (QuestionId, UserId) pair, used to join with associated answer metadata (see below). Numeric.
- **IsCorrect**: Binary indicator for whether the student's answer was correct (1 is correct, 0 is incorrect). Categorical.
- **SubjectId**: Each subject covers an area of mathematics, at varying degrees of gran- ularity. We provide IDs for each topic associated with a question in a list. Example topics could include "Algebra", "Data and Statistics", and "Geometry and Measure". These subjects are arranged in a tree structure, so that for instance "Factorising" is the parent subject of "Factorising into a Single Bracket". We provide details of this tree in an additional file subject metadata.csv which contains the subject name and tree level associated with each SubjectId, in addition to the SubjectId of its parent subject. Categorical.
- **Category1**: Feature engineered. The first-level category of the question (given that there is hierarchical categories). Categorical.
- **Gender**: The student's gender, when available. 0 is unspecified, 1 is female, 2 is male and 3 is other. Categorical.
- **Age**: Feature engineered. The student's age, as calculated from `DateAnswered - DateOfBirth`. Numeric.
- **PremiumPupil**: Whether the student is eligible for free school meals or pupil premium due to being financially disadvantaged. Categorical.
- **DateAnswered**: Time and date that the question was answered, to the nearest minute. Time sequence/numeric.
- **Confidence**: Percentage confidence score given for the answer. 0 means a random guess, 100 means total confidence. Numeric.
- **GroupId**: The class (group of students) in which the student was assigned the question. Categorical.
- **QuizId**: The assigned quiz which contains the question the student answered. Categorical.

Table 2.1: The source data used in this paper, tranformed by unions across several csvs, some columns dropped, and some created columns.

	QuizId	Gender	PremiumPupil	SubjectId	Age	Category1
0	86	2	False	[3, 49, 62, 70]	12	Algebra
1	39	0	False	[3, 32, 144, 204]	nan	Number
2	39	0	False	[3, 32, 37, 220]	nan	Number

Table 2.1: The source data used in this paper, tranformed by unions across several csvs, some columns dropped, and some created columns.

	QuizId	Gender	PremiumPupil	SubjectId	Age	Category1
3	115	0	False	[3, 49, 81, 406]	nan	Algebra
4	78	2	False	[3, 71, 74, 180]	11	Geometry and Measure

2.1.1 Exploratory Analysis

We perform initial EDA, which can be seen in Section 3 (Results). The EDA performed, in order to gain insight into the data, is as follows:

- Summary statistics of each column used.
- A sunburst plot of all the subject categories found in the question dataset.
- A histogram of the proportion of questions answered correctly by each student.

2.2 Statistical Methods Analysis

2.2.1 Clustering

We perform clustering in order to answer the question: are the distances/similarity coefficients between the factors of the dataset, indicative of the subject category ID given to each data point? More specifically, when the columns given above are clustered, do we achieve a clustering similar to the labels of `Category1` given to each data point? In order to achieve this, we must take into account the categorical factors in the dataset (such as `QuizId` and `Gender`) and the fact that we cannot simply compute minimal Euclidian distances between the values, even when ordinal. We utilize 2 different methods for handling categorical factors when clustering.

2.2.1.1 Using the kmodes library

We use the `kprototypes` algorithm from the `kmodes` library. k-modes is used for clustering categorical variables. It defines clusters based on the number of matching categories between data points. (This is in contrast to the more well-known k-means algorithm, which clusters numerical data based on Euclidean distance.) The k-prototypes algorithm combines k-modes and k-means and is able to cluster mixed numerical / categorical data (Vos, 2015--2021).

2.2.1.2 Using the Gower Distance

Gower's Distance can be used to measure how different two records are (Gower, 1971). The records may contain combination of logical, categorical, numerical or text data. The distance is always a number between 0 (identical) and 1 (maximally dissimilar). The metrics used for each data type are described below:

- quantitative (interval): range-normalized Manhattan distance
- ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
- nominal: variables of k categories are first converted into k binary columns and then the Dice coefficient is used

This distance metric can be used to calculate a distance matric between all points in the dataset, which can then be used by standard hierarchical clustering. We use the `scikit-learn` package with its Agglomerative clustering algorithm, and cluster across multiple linkage types (as different types of linkage can produce vastly different clusters) (Pedregosa et al., 2011).

2.2.1.3 Performance metrics

In order to measure how well the clustering results approximate the question category labels given, we use the Rand index for similarity. It is a measure of similarity between two different clusterings of the same set of data; the measure essentially considers how each pair of data points is assigned in each clustering (Rand, 1971). A value of 0 indicates no similarity (clusterings do not agree on any pair of points), and 1 indicates perfect matching in clustering labels. A form of the Rand index, called the adjusted Rand index, is adjusted for the chance grouping of elements.

2.2.2 Supervised Learning Models

We also run various supervised learning models on the data, in order to answer the following question: can non-deep learning (aka not neural network) models learn, based on the given factors in the data, whether a student will answer a question correctly? Specicifally, we run models on a transformed version of the dataset in order to predict the label column of `IsCorrect`. The transformations performed on the columns of the dataset are:

- Numerical values were min-max normalized to 0-1.
- The timeseries column (`DateAnswered`) was transformed into an integer.
- The categorical columns were one-hot encoded, where each category in each factor recieves its own column of 0-1 values (with 1 indicating that value is present), essentially creating a sparse matrix subset.

We run 7 different models on the dataset, with all implemented in `scikit-learn` (Pedregosa et al., 2011). The models used are as follows:

- **Logistic Regression**: A simple logistic regression classifier, where parameters (for each factor plus bias/intercept) are fitted to a linear model.
- **Logistic Regression with Stochastic Gradient Descent (SGD)**: SGD improves on gradient descent by replacing the gradient with an esimation of it, reducing computational complexity.
- **Perceptron**: A linear predictor which uses a set of weights with the feature vector to output a binary clasisfier.
- **Linear Support Vector Machine**: Maps training examples to points in space so as to maximise the width of the gap between the two categories. Used to perform linear classification.
- **Decision Tree**: A model where decisions are represented by a tree/flowchart structure, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules (Wikipedia contributors, 2023).
- **Random Forest Classifier**: A meta-model where a number of decision trees are fitted on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (Pedregosa et al., 2011).
- **Histogram-based Gradient Boosting Classification Tree**: A meta-model that fits multiple gradient-boosted decision tree classifiers, that has support for NaN values, categorical values, and is computionally quick dues to binning inputs into histograms instead of naive evaluation.

2.2.2.1 Performance metrics

To evaluate the models, we use standard performance metrics that are used for supervised learning on binary classification. The metrics we display are:

- **Accuracy**: The ratio of correct predictions to all predictions. In other words, the total of the green squares in a confusion matrix divided by the entire matrix. This is arguably the most common concept of measuring performance. It ranges from 0-1 with 1 being the best performance.
- **Precision**: The ratio of true positives to the total number of positives (true positive + true negative).
- **Recall**: The ratio of true positives to the number of total correct predictions (true positive + false negative).

- **F1 Score**: Known as the harmonic mean between precision and recall. Precision and Recall are useful in their own rights, but the F1-Score is useful in the fact it's a balanced combination of both precision and recall. It ranges from 0-1 with 1 being the best performance.
- **Support**: The number of true instances for each label.

In addition, we use several visualizations to display perfomance of the models:

- **ROC Curve**: A plot of the true positive rate vs the false positive rate, as a curve. We examine the AUC (area under the curve) to determine how well that a randomly chosen positive example is indeed labeled positive. If it follows the straight diagonal line, the AUC is low and therefore the classifier is no better than chance. If there's a high AUC, then the classifier is performing well. The baseline AUC is 0.5, a a perfect classififer has 1.0
- **Confusion Matrix**: A matrix showing the amount of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- **Precision-Recall Curve**: A model can improve in precision or recall, but not both. A PR curve shows that tradeoff, and how well it performs in both. The curve is constructed by calculating and plotting the precision against the recall for a single classifier at a variety of thresholds. A perfect classifier would have a line that starts high and straight, and curves down only near the end of the recall axis. The summary value for the curve is the AP, or average precision; higher values towards 1 are better.

3 Results

3.1 Exploratory Data Analysis

For all used columns in the dataset, we show their statistical summaries in Table 3.1.

Table 3.1: Summary values of all the used columns in the dataset.

Table 3.2

	QuestionId	UserId	AnswerId	IsCorrect	DateAnswered	Confidence	GroupId
count	1382727.0	1382727.0	1382727.0	1382727.0	1382727	346428.0	1382727.0
unique					177615		
top					2020-03-03 11:12:00		
freq					165		
mean	468.2	3036.3	754427.9	0.5		73.9	196.2
std	273.6	1770.6	435619.0	0.5		31.2	114.6
min	0.0	1.0	0.0	0.0		0.0	0.0
25%	233.0	1515.0	377293.5	0.0		50.0	95.0
50%	468.0	3009.0	754453.0	1.0		75.0	198.0
75%	703.0	4565.0	1131772.5	1.0		100.0	300.0
max	947.0	6147.0	1508916.0	1.0		100.0	389.0

Table 3.3

	QuizId	Gender	PremiumPupil	SubjectId	Age	Category1
count	1382727.0	1382727.0	1382727	1382727	727345.0	1382727
unique			2	62		3
top			False	[3, 32, 42, 211]		Number
freq			1185347	147497		685691
mean	61.4	1.0			11.4	
std	31.3	0.8			0.6	
min	0.0	0.0			0.0	
25%	37.0	0.0			11.0	
50%	63.0	1.0			11.0	
75%	86.0	2.0			12.0	
max	119.0	3.0			38.0	

Some features of note are:

- On average, all question-answser pairs are answered correct about 50% of the time.
- Students are on average 75% confident in their answers.
- The average student age is 11 years old, with a couple outliers of student being 38.

We show a plot of all the subject categories found in the question dataset Figure 3.1, proportioned by their count appearance in the dataset. Note that for the later clustering/analysis, we only use the 1st level hierarcical categories, aka `Algebra, Geometry and Measure, Number`.

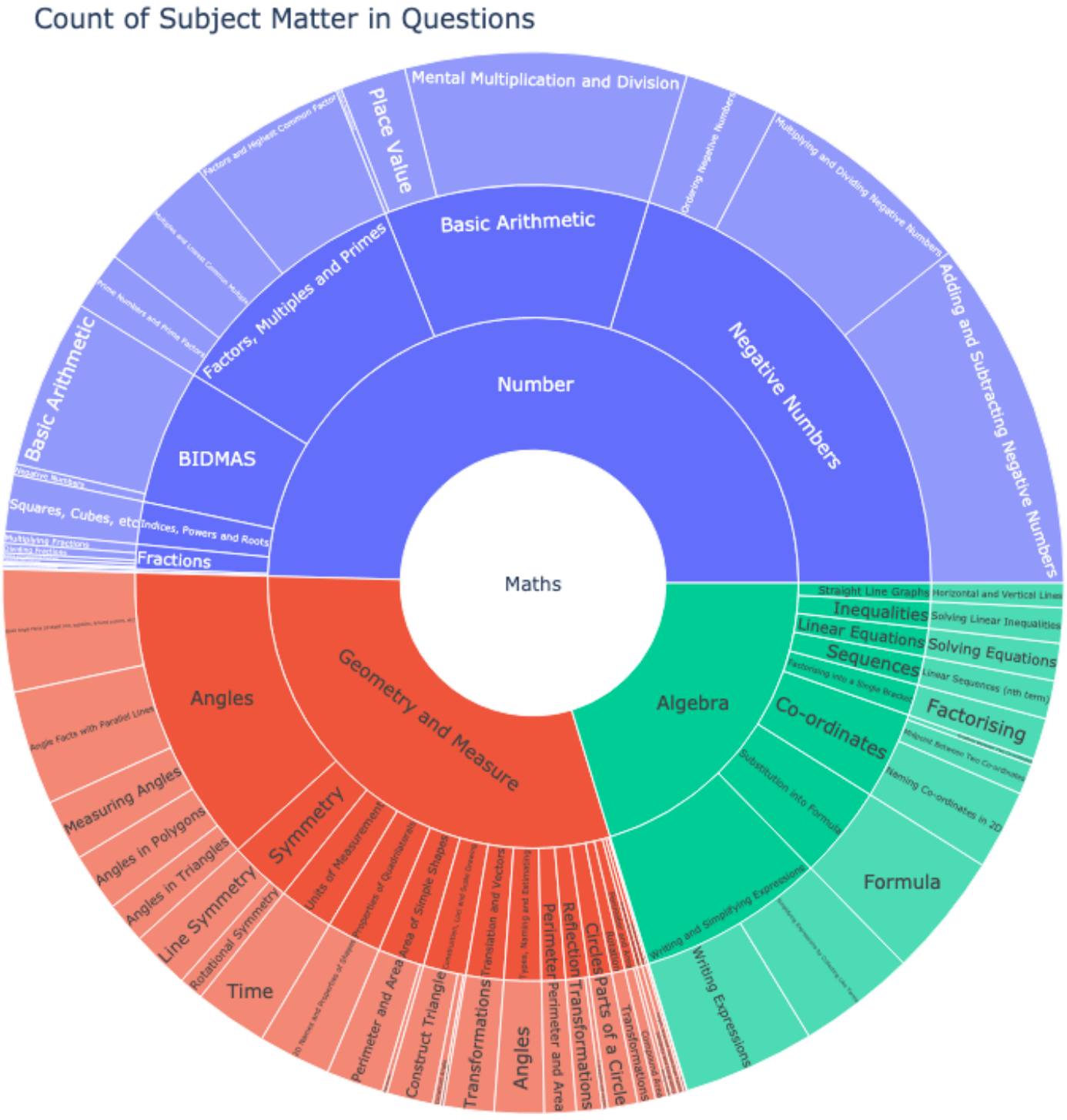


Figure 3.1: A sunburst plot of all the hierarchical subject categories found in the question dataset, with ratios correct to their count/appearance.

We observe that the category `Number` has the most questions, with `Geometry and Measure` having the most 2nd level categories.

In Figure 3.2, we show the proportion of questions answered correctly by each student as a histogram.

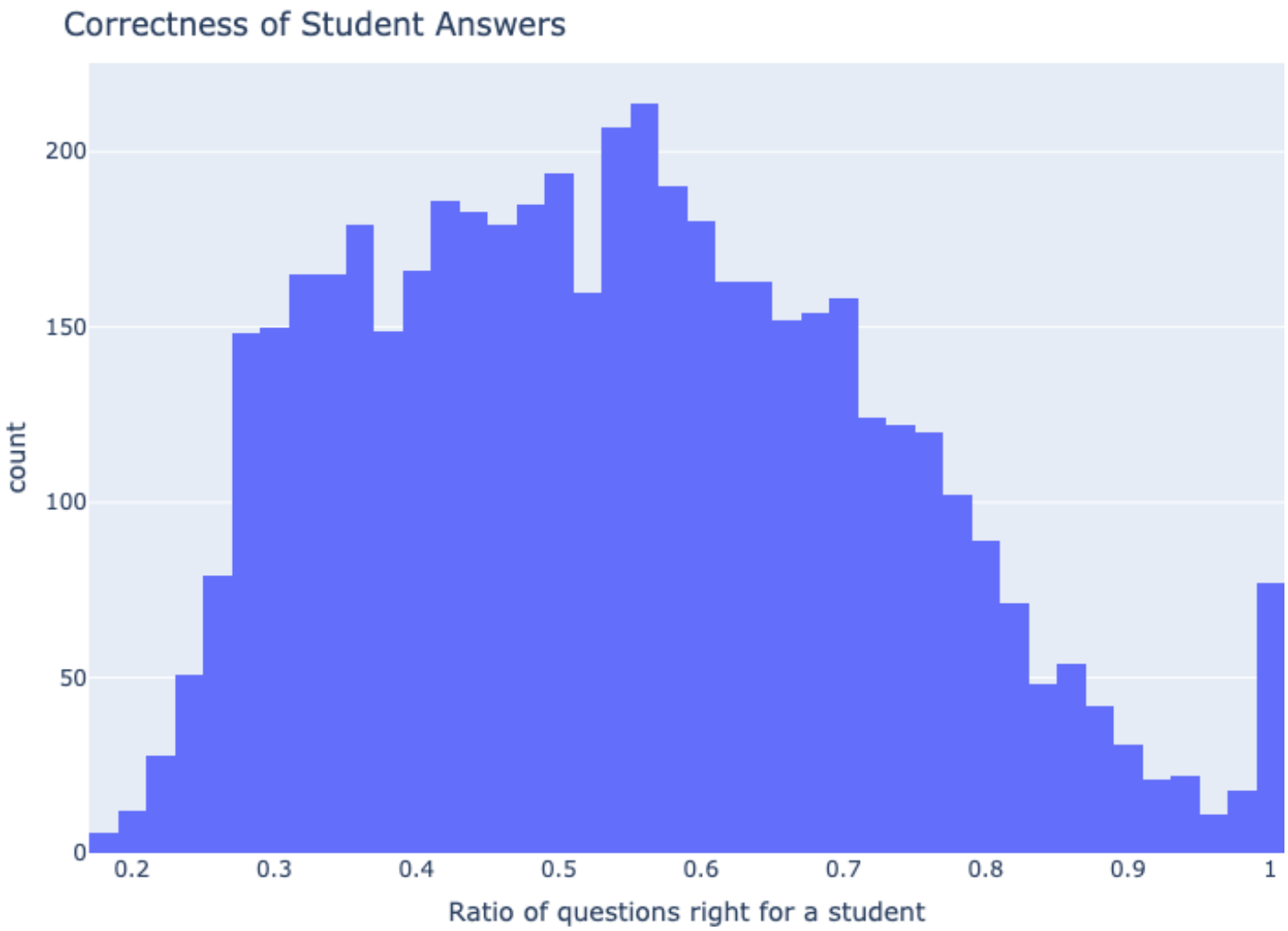


Figure 3.2: A histogram of the proportion of questions answered correctly by a student.

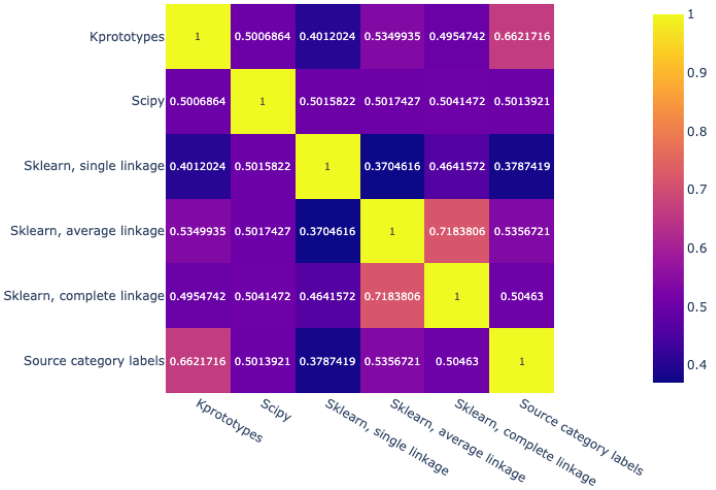
The proportions generally seem to follow a normal distribution, which is expected and confirms the proper spread of data.

3.2 Clustering

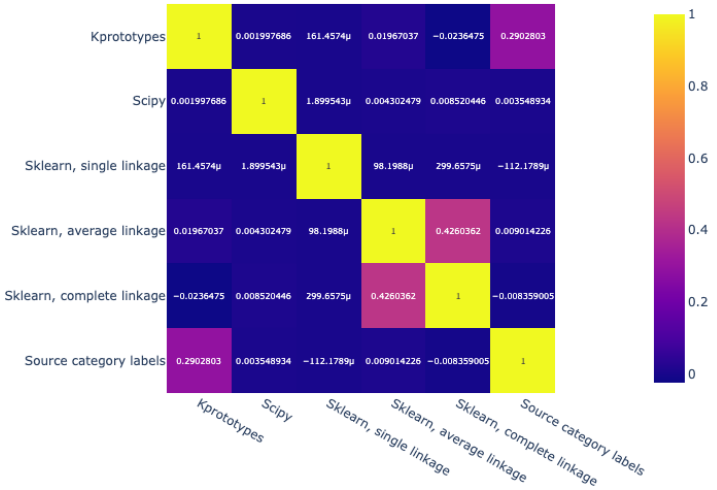
Clustering was performed as described in Section 2 (Methods); the `kmodes` library was used on the base dataset, the Gower distances matrix was computed in order to handle categorical variables, and the distance matrix was used with `scikit-learn` library for Agglomerative clustering. In addition, for the Agglomerative clustering, 3 different linkages between distances were used: single, average, and complete.

3.2.1 Perfomance

In order to judge the clustering output, the Rand index was calculated as described in the previous section. We compute the Rand index, and the adjusted Rand index, not only between the source labels and the computed clusters but also between each clustering method. The rand index results are displayed in Figure 3.3, as a heat map matrix between all clustering types and the source labels. The highest Rand index with the source category labels was the `Kprototypes` clustering algorithm, with a value of 0.66.



(a) Rand Index Matrix



(b) Adjusted Rand Index

Figure 3.3: Matrices of the rand index as compared across all clustering methods, and compared to the original source category labels. The scipy and sklearn clustering methods were done using a precomputed Gower matrix.

3.3 Supervised Learning Models Performance Metrics

The supervised learning models were run as described in the previous section, with their default values used for most parameters (such as loss function). For most models, the one modified parameter was the number of iterations run in order to train the model, by increasing it to 1000.

For the best performing model (`HistGradientBoostingClassifier`), we made the following modifications in an attempt to improve its performance even further:

- The data was not transformed from the original source columns, leaving the categories nad numbers unchanged (with the exception of `DateAnswered`, which was still converted to an int).
- The maximum number of leaf nodes was increased to 80.
- The category factors were given as categorical input.

This resulted in the new best performing model, `HistGradientBoostingClassifier2`. The summary and performance statistics for each model are summarized in Table 3.4, with the metrics as described in the previous section.

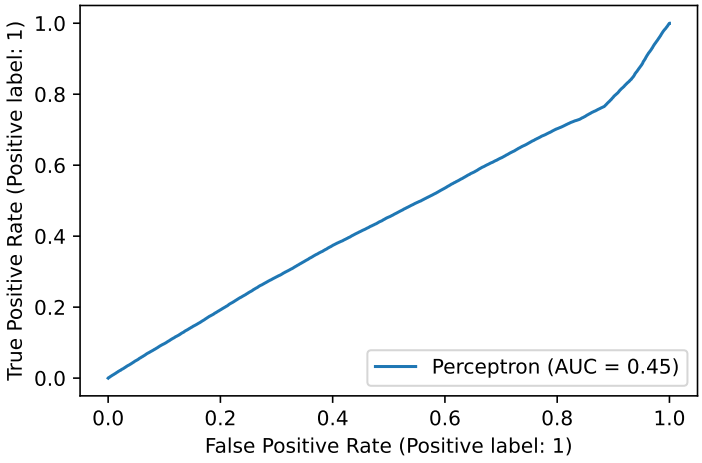
Table 3.4: The performance values and some settings for the supervized learning models ran, sorted by accuracy.

Table 3.4

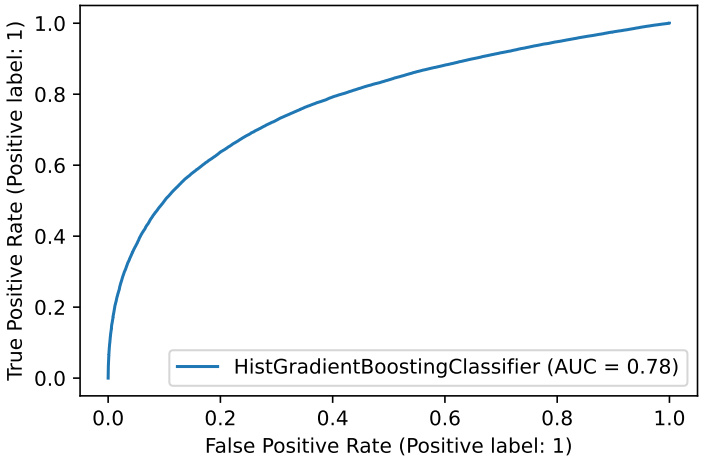
Model	Accuracy	precision	recall	f1-score	support	TrainIters	LossFcn
HistGradientBoostingClassifier2	0.71441	0.71472	0.71441	0.71454	138273.00000		log_loss
HistGradientBoostingClassifier	0.66020	0.65997	0.66020	0.66007	138273.00000		log_loss
RandomForestClassifier	0.65193	0.65193	0.65193	0.65193	138273.00000		
DecisionTreeClassifier	0.61077	0.61074	0.61077	0.61075	138273.00000		
LinearSVC	0.57034	0.56629	0.57034	0.55989	138273.00000	12	squared_hinge
LogisticRegression	0.57028	0.56623	0.57028	0.56008	138273.00000	[0]	
SGDClassifier	0.56954	0.56555	0.56954	0.55500	138273.00000	41	log_loss
Perceptron	0.47191	0.45654	0.47191	0.45497	138273.00000	11	perceptron

3.3.1 Performance Plots

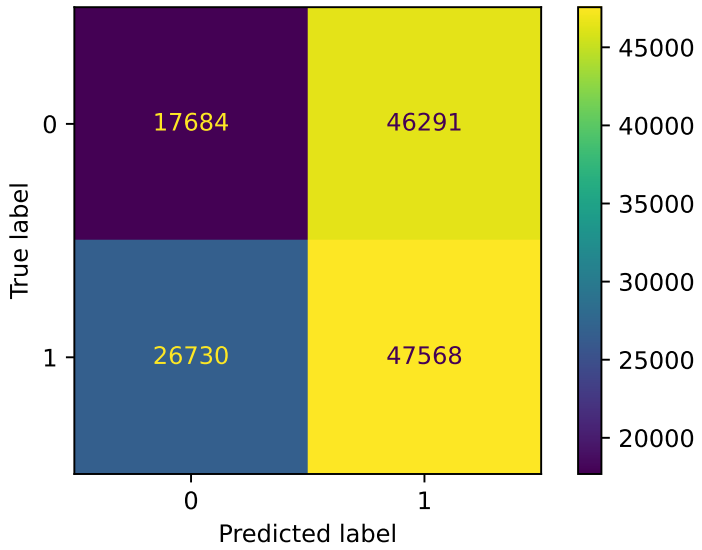
We display the ROC curve, confusion matrix, and precision-recall curves for the worst (`Perceptron`) and best (`HistGradientBoostingClassifier2`) performing models in Figure 3.4. We observe that we only see "typical" ROC and PR curves for the best performing model.



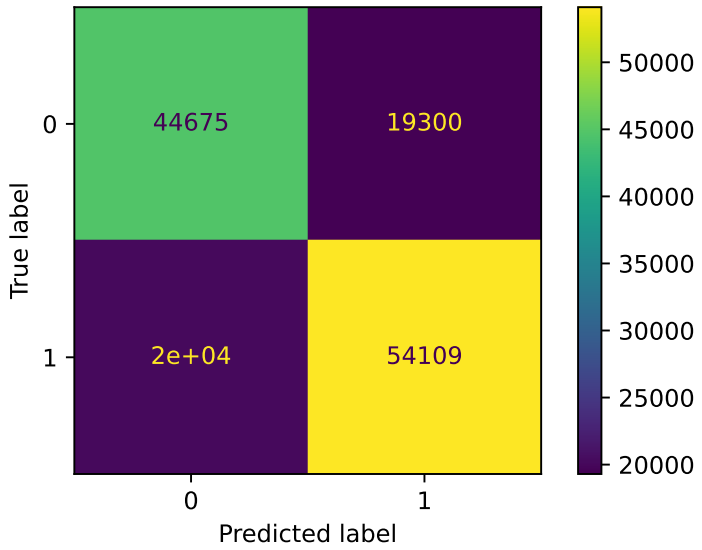
(a) Perceptron: ROC Curve



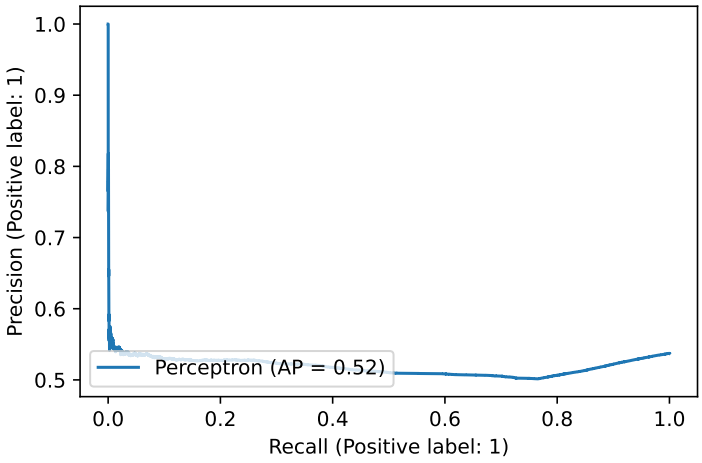
(b) HistGradientBoostingClassifier: ROC Curve



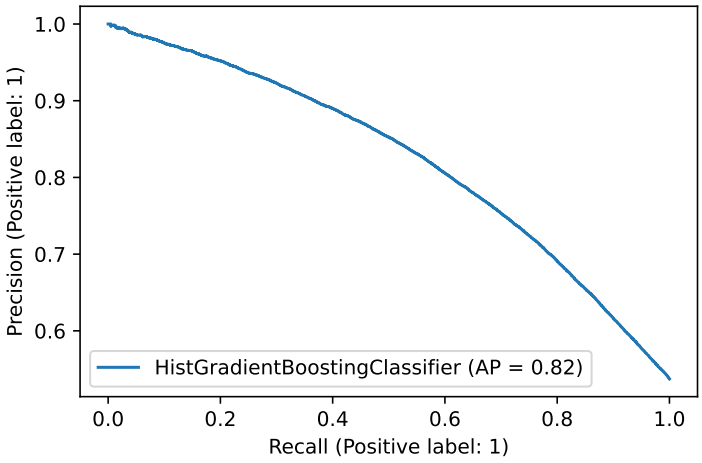
(c) Perceptron: confusion matrix



(d) HistGradientBoostingClassifier: confusion matrix



(e) Perceptron: precison-recall curve



(f) HistGradientBoostingClassifier: precison-recall curve

Figure 3.4: The ROC, confusion matrix, and precision-recall curves for the best and worst performing model.

4 Conclusions

High-quality machine learning models have been of keen interest in the field of education since at least as early as the 1980s (Murphy, 2019). They have routinely demonstrated their potential for enhancing the quality of diagnostic testing. Given that the fundamental idea of machine learning in education is to use large amounts of student response data to monitor the current state of a students' learning and estimate the future performance of students on subsequent material, presumably after receiving a relevant course of instruction. One of the primary advantages of employing models in educational diagnostic testing is their capacity to analyze complex data sets. These models can use data from various sources, including student performance records, behavioral patterns, learning styles, and assessment results. By scrutinizing this multidimensional information, these models can discern correlations that would otherwise go unnoticed and identify subtle trends that might evade human perception (Yang et al., 2021). Consequently, they facilitate the early detection of learning challenges, allowing educators to intervene proactively and tailor educational strategies to suit individual student needs.

4.1 Findings

From examining the results in Figure 3.3, we can see that the clustering methods performed only adequately; if we choose to consider the adjusted Ran Index instead of the non-adjusted (as a reminder, the adjusted index accounts for the chance or random grouping of elements), we only get a maximum score of 0.29 similarity, indicating a low match in groupings with the source category labels.

From Table 3.4, we can see that the best performing model achieved and accuracy and F1 score of 0.71. This is an adequate model, that gives statistically better results than random chance; however, compared to deep learning models implemented on this data which achieved performance of 0.94 and greater (haradai1262, 2020; shshen-closer, 2021), our models do not measure up. Nevertheless, a simplistic supervised learning model achieving such performance is a good indicator of the predictive power of the features in the input dataset of diagnostic questions. Additionally, the ROC curves and precision-recall curve of the best performing model (Figure 3.4) show typical and moderately well-perfoming predictive performance.

While the analysis conducted here did not produce any evidence to threaten the validity of the diagnostic test items in question and did not provide any reason to conclude that construct-irrelevant factors were driving student responses, we feel that there is reason to be optimistic about the relative success of the machine learning models. To the extent that the models were successful in generating information that could be used to guide educational decision making, they provide evidence in favor of using such models in other tasks such as machine grading, precision educational interventions, and automated tutoring. Machine grading is one area in which accurate and efficient models could ease the burden on educators by limiting the time they need to spend trying to provide meaningful and actionable feedback to students (Spector et al., 2016).

With regard to precision educational interventions, the applications of effective machine learning models span a broad spectrum of contexts. For example, existing research points to potential uses for predicting dropouts and disciplinary problems, as well as other matters related to attrition and retention (Luan & Tsai, 2021). So, even though this analysis focused on an application for predicting response patterns to mathematics questions, there is no reason that the same models could not be used for prediction and diagnosis in other areas of education. This is especially relevant in the aftermath of the massive interruptions to students' education during and after the pandemic when educators across the country are racing to find ways to remedy the severe learning loss that took place.

4.2 Limitations

One limitation of this study is that there was relatively little student demographic information available. If a richer set of variables had been available in the data set, then it is at least conceivable that construct-irrelevance or differential item functioning might have emerged from the data. There was also no information about student response times, which is another indicator that is commonly used to investigate whether certain families of items operate in different ways among different groups of respondents. A third limitation is that very little is known about the students who are present in the sample. Fairly simple information about how much exposure students have had to certain kinds of questions and what kind of instruction they receive was unavailable. Any of these limitations on their own or in tandem could make for very different results with regard to test validity.

4.3 Future Studies

Future iterations of this kind of work might attempt to focus the predictive abilities of the machine learning models to behavioral and social-emotional characteristics of students based on school-level data that is typically recorded. For example, attendance, chronic absences, discipline referrals, and expulsions are all kinds of information that are routinely collected by schools. This kind of information combined with standard metrics of academic performance might lead to incredibly useful insights that allow teachers and administrators to identify and design interventions for the most at-risk students.

5 Link to Video Presentation

The Youtube video of the presentation can be found at TODO

References

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9365–9374.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295x.111.4.1061

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research*, *36*(3), 206–217. http://www.jstor.org/stable/27528353

Cronbach, L. J., & Meehl, P. E. (1966). Construct validity in psychological tests. In *Readings in clinical psychology* (pp. 29–52). Elsevier. https://doi.org/10.1016/b978-1-4832-0087-3.50007-3

Ghosh, A., & Lan, A. (2021). BOBCAT: Bilevel optimization-based computerized adaptive testing. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 2410–2417). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/332

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. http://www.jstor.org/stable/2528823

haradai1262. (2020). Solution of NeurIPS education challenge 2020. In *GitHub repository*. GitHub. https://github.com/haradai1262/NeurIPS-Education-Challenge-2020

Luan, H., & Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, *24*(1), 250–266. https://www.jstor.org/stable/26977871

Murphy, R. F. (2019). *Artificial intelligence applications to support k–12 teachers and teaching: A review of promising applications, opportunities, and challenges*. RAND Corporation. http://www.jstor.org/stable/resrep19907

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.1080/01621459.1971.10482356

shshen-closer. (2021). TOP1-for-task-2-in-the-NeurIPS-2020-education-challenge. In *GitHub repository*. GitHub. https://github.com/shshen-closer/TOP1-for-task-2-in-the-NeurIPS-2020-Education-Challenge

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L. (Joy), Mukama, E., Warusavitarana, A., Dona, K. L., Eichhorn, K., Fluck, A., Huang, R., Bridges, S., Lu, J., Ren, Y., Gui, X., Deneen, C. C., Diego, J. S., & Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of Educational Technology & Society*, *19*(3), 58–71. http://www.jstor.org/stable/jeductechsoci.19.3.58

Vos, N. J. de. (2015--2021). *Kmodes categorical clustering library*. https://github.com/nicodv/kmodes.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernandez-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2021). *Results and insights from diagnostic questions: The NeurIPS 2020 education challenge*. https://arxiv.org/abs/2104.04034

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2020). Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv Preprint arXiv:2007.12061*.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., & Zhang, C. (2021). *Instructions and guide for diagnostic questions: The NeurIPS 2020 education challenge*. https://arxiv.org/abs/2007.12061

Wikipedia contributors. (2023). *Decision tree — Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=Decision_tree&oldid=1178109845

Yang, C. C. Y., Chen, I. Y. L., & Ogata, H. (2021). Toward precision education: Educational data mining and learning analytics for identifying students' learning patterns with ebook systems. *Educational Technology &*

Society, *24*(1), 152–163. https://www.jstor.org/stable/26977864