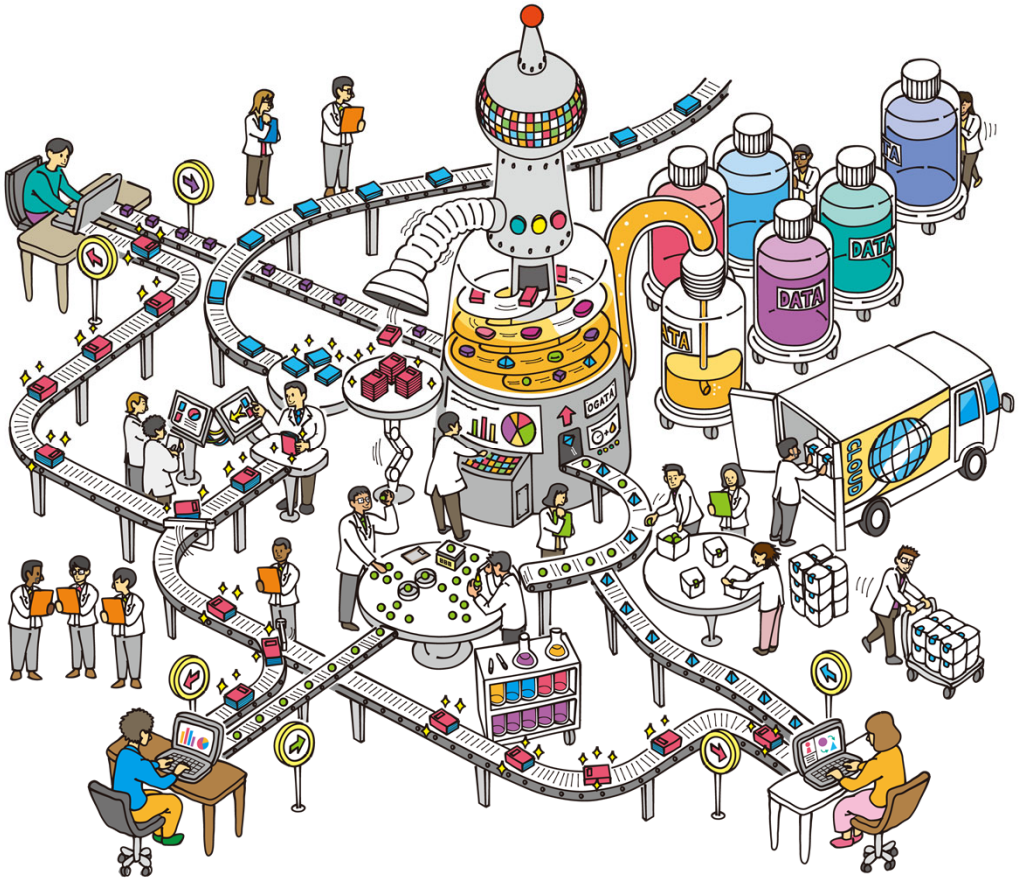


京都大学
KYOTO UNIVERSITY

教育データ解析コンテストについて

緒方広明
京都大学学術情報メディアセンター
Learning and Educational Technologies Research Unit

本研究の一部は、内閣府総合科学技術・イノベーション会議の「SIP第2期/ビッグデータ・AIを活用したサイバー空間基盤技術」（管理法人：国立研究開発法人新エネルギー・産業技術総合開発機構）によって実施されました。

目的

- 内閣府・戦略的イノベーション創造プログラム（SIP）第2期ビッグデータ・AIを活用したサイバー空間基盤技術（管理法人：NEDO）に採択された「エビデンスに基づくテーラーメイド教育の研究開発」の研究では、教育・学習支援システムを利用して、教育データの収集と分析を実施している。

- 本企画は、イベントを通して、教育データの解析技術を、社会全体で向上していくことを目的。

教育データ解析コンテスト実行委員会

- 京都大学
- 九州大学
- NTTラーニングシステムズ
- 東京大学
- 慶応義塾大学
- 明治大学
- NTT
- NTTコミュニケーションズ
- アセンブローグ

https://sites.google.com/view/sipdatachallenge/

結果発表・表彰式

コンテストの結果発表及び表彰式は、3月9日（火）13時より、オンライン形式で行われます。

当日のプログラムの予定は以下の通りです。

結果発表・表彰式プログラム

13:00-13:10　データチャレンジンコンテストの概要、緒方広明（京都大学学術情報メディアセンター教授）

13:10-13:15　デジタル教材のログ分析のためのライブラリOpenLA、島田敬士（九州大学 大学院システム情報科学研究院・教授）

13:15-13:30　コンテストの結果発表と受賞者からの発表

13:30 まとめ

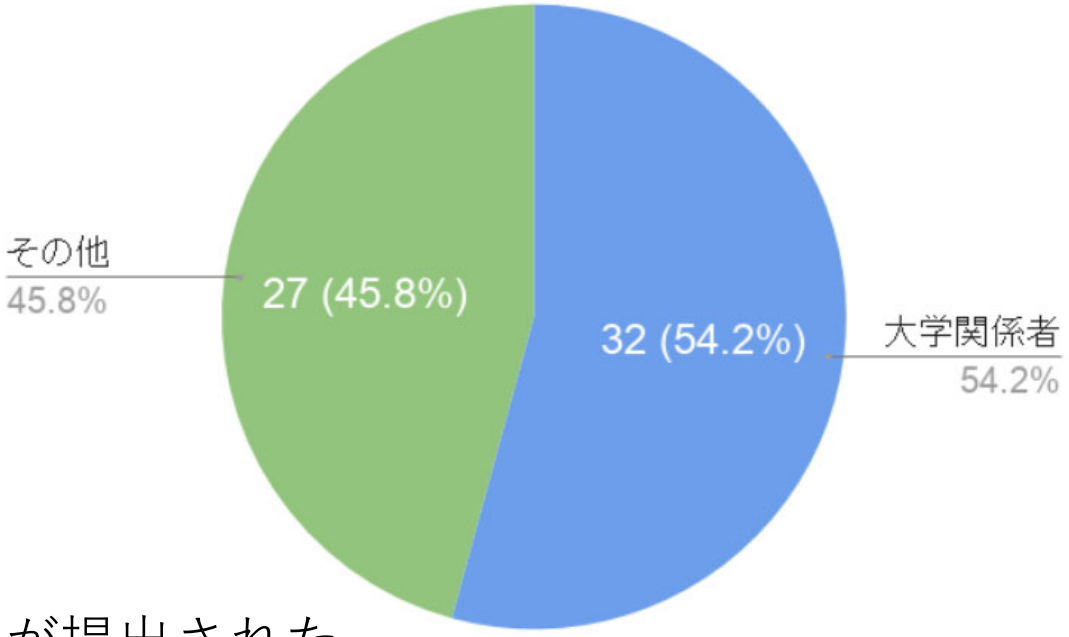
表彰式は専用URLにてライブ配信いたします。
ご視聴を希望されるの方は、下記メールアドレスまでご連絡ください。

日程

- 募集開始：2020年12月1日
- データダウンロード開始：2020年12月7日
- 参加申し込み締切：2021年1月31日
- 作成プログラムの提出開始（順位表の公開開始）：2021年1月15日
- 作成プログラムの提出締切：2021年2月15日(23:59 JST)
- 結果発表・表彰式：2021年3月9日 13:00～（オンライン）

参加者

- **合計：59チーム**
- 大学関係者：32チーム
- 企業などその他：27チーム



最終的に**19チーム**からプログラムが提出された

参加の条件

SIPデータチャレンジのデータセットを使用する条件として、次の事項に必ず同意して下さい。

1.データセットは、SIPデータチャレンジのためのみに使用すること。

2.SIPデータチャレンジが終了した後、ダウンロードしたデータセットはすべて削除すること。

3.本データをWeb pageやSNS等に掲載したり、第三者に提供しないこと。

4.本データを商品化しないこと、また、悪意のある方法、意図しない方法で掲載や使用しないこと。

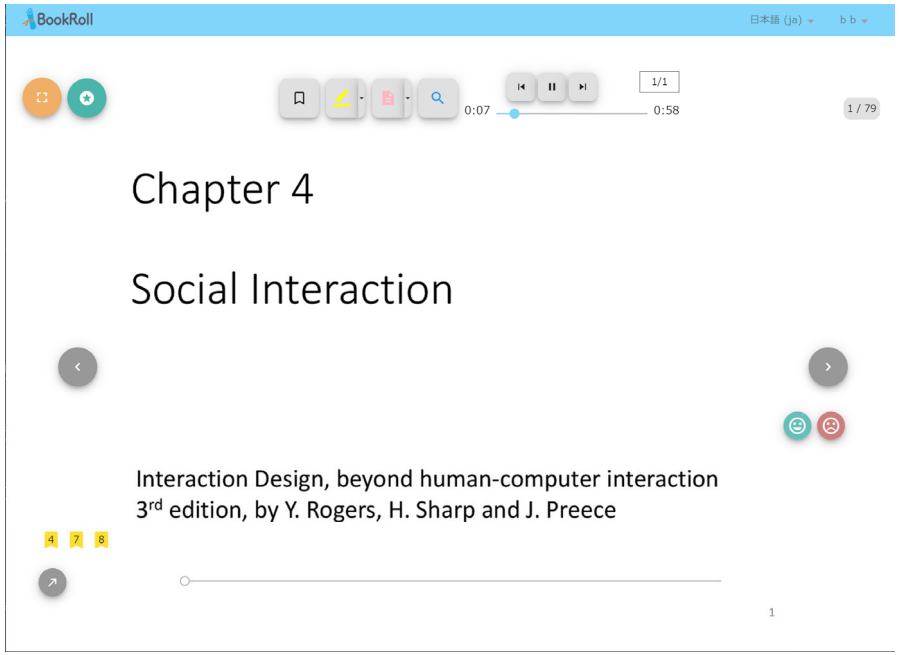
5.本データを用いて作成したプログラムは、指定した方法で主催者側に提供すること。

本データセットを論文などで説明する際は、以下の論文を引用すること。

- Brendan Flanagan, Hiroaki Ogata, Learning Analytics Platform in Higher Education in Japan, Knowledge Management & E-Learning (KM&EL), Vol.10, No.4, pp.469-484, 2018.
- Hiroaki Ogata, Misato Oi, Kousuke Mohri, Fumiya Okubo, Atsushi Shimada, Masanori Yamada, Jingyun Wang, and Sachio Hirokawa, Learning Analytics for E-Book-Based Educational Big Data in Higher Education, In Smart Sensors at the IoT Frontier, pp.327-350, 2017.

ツールの紹介:BookRoll

- デジタル教材配信システム**BookRoll**を用いて教材の閲覧ログデータを収集
  https://www.let.media.kyoto-u.ac.jp/project/digital-teaching-material-delivery-system-bookroll/
- NIIシンポジウムで紹介第7回(5/8)、第8回(5/15)で紹介



教員がデジタル教材（教科書、補助資料等）をPDF形式で登録すれば、学生は授業中・予習/復習時に、それをウェブブラウザで閲覧できる。

音声も

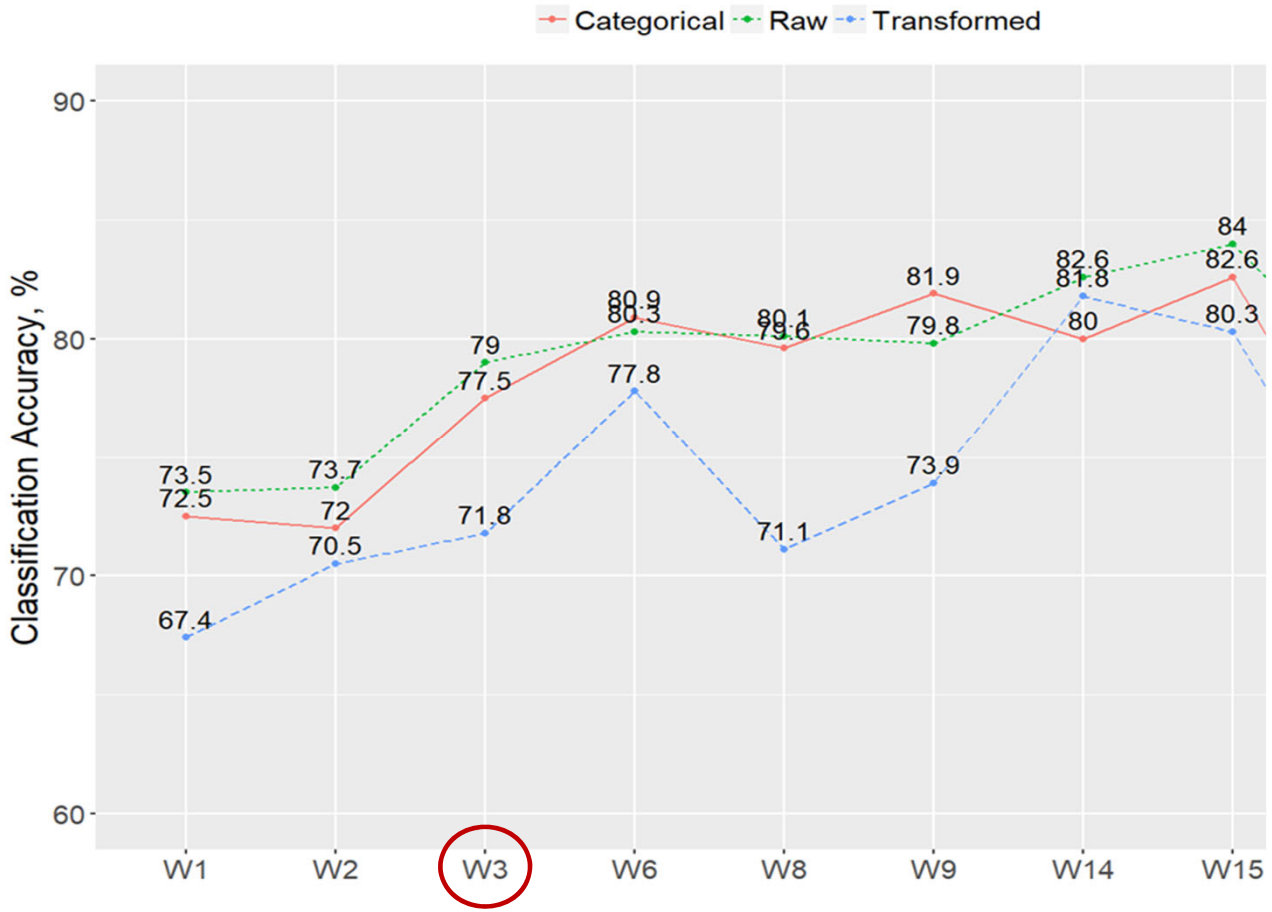
学生に元のPDFをダウンロードされないので、内容が拡散しない。

BookRoll上での学生の行動は学習ログとして記録される。

Open, Close, Next, Prev,Add_Marker, Delete_Marker, Add_memo, Delete_memo, など15種類のAction

これまでの研究：学生の成績予測

[Akcapinar et. al., SLE, 2019]



- 教員が早期に単位を落としそうな学生を発見
- 学生は成績予測を確認して、学びに向かう態度を改善

Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an Early-Warning System for Spotting At-Risk Students by using eBook Interaction Logs. *Smart Learning Environments, 6*(4), 1-15. doi:doi.org/10.1186/s40561-019-0083-4

予測結果

[Akcapinar et. al., SLE, 2019]

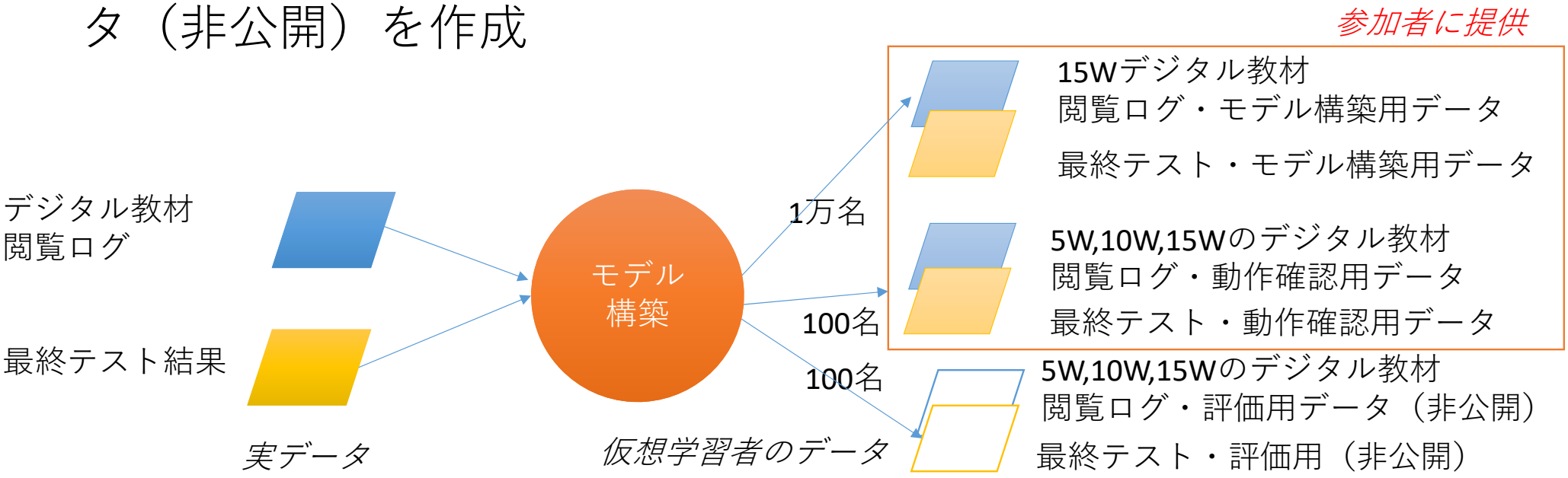
Table 4 Average scores of the models in terms of Accuracy and Kappa

Algorithm	Raw Data		Transformed Data		Categorical Data	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
Adaboost	0.790	0.580	0.783	0.567	0.748	0.494
bartMachine	0.811	0.620	**0.813**	**0.625**	**0.792**	**0.583**
gbm	0.795	0.589	0.793	0.586	0.782	0.565
glm	0.753	0.504	0.728	0.454	0.680	0.359
J48	0.813	0.625	**0.833**	**0.665**	0.766	0.530
JRip	0.795	0.587	0.782	0.564	0.755	0.509
knn	**0.813**	**0.627**	0.798	0.595	**0.805**	**0.611**
naive_bayes	**0.823**	**0.646**	0.801	0.601	**0.811**	**0.621**
nnet	0.710	0.420	0.780	0.558	0.754	0.505
rf	**0.823**	**0.647**	**0.824**	**0.644**	0.782	0.563
rpart	0.752	0.501	0.759	0.516	0.727	0.454
svmLinear	0.776	0.550	0.749	0.498	0.684	0.369
xgbLinear	0.798	0.596	0.780	0.560	0.726	0.452

Note: Bold values show the top three best-performed algorithms' results for each data form

データセットの作成について

- デジタル教材の閲覧ログと最終テストの成績
- SIPの実証校で蓄積したデータからモデルを構築して、1万人の**仮想学習者**のデータを作成
- 同じモデルを用いて、100名の動作確認用データと評価用データ（非公開）を作成

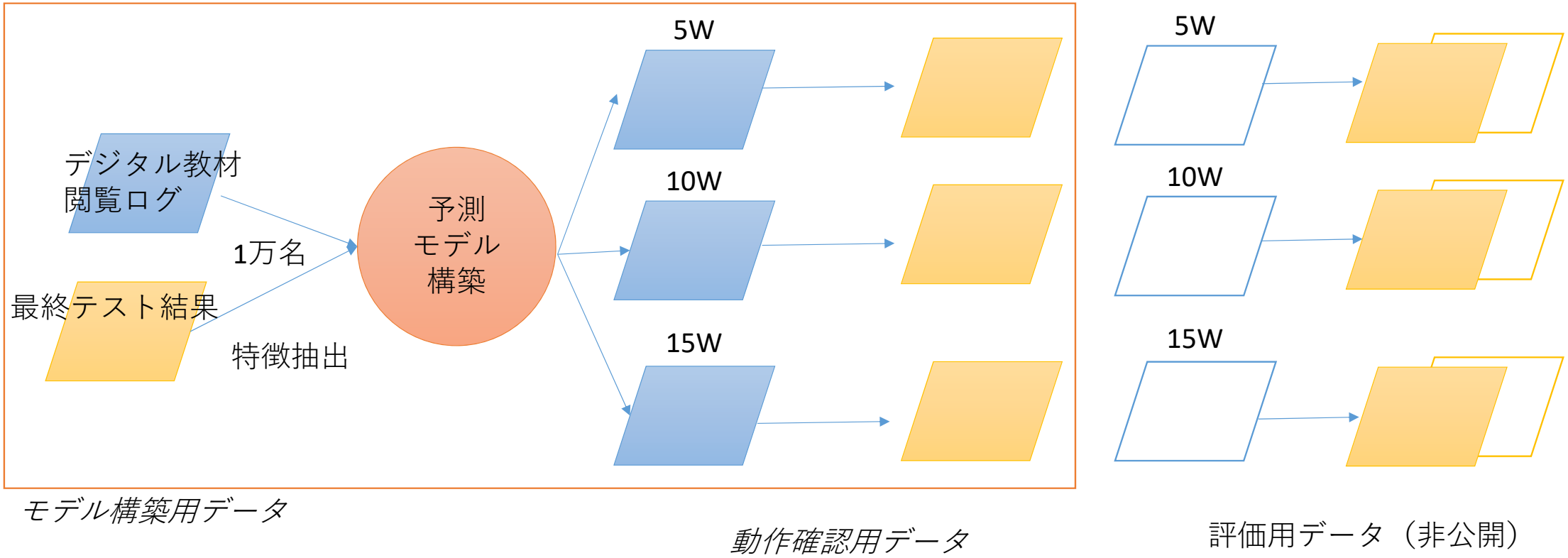


成績の予測について

100名
デジタル教材閲覧ログ

最終成績
(0点～100点の100段階)

評価用データを用いた
予測結果と比較



モデル構築用データ

動作確認用データ

評価用データ（非公開）

評価方法について

- 予測結果は、指標としてRMSE（平均平方二乗誤差）を使用
- コース開始から5週経過時点、10週経過時点、15週経過時点のデータから予測した最終成績のRMSEの平均値で評価。
- 再現性を確保するため、参加者はプログラムをインストールしたDockerイメージを提出する。
- 非公開の評価データを用いて京大と九大の研究室で別々に評価して、評価結果を照合。

結　果

順位発表(1/18-2/10)

順位	チーム名（ニックネーム）	スコア
1	**のざモト**	**16.50**
2	NTTCS	16.59
3	chscc	17.47
4	polly	18.39
5	khiroyuki1993	21.91
6	上智大学 データサイエンス特論チームA	22.98
7	熊本大学 喜多研究室	25.65
8	mh	25.86
9	koudou-ai	30.12
10	mi-fujita	31.75
11	Nakamoto	4.556.74

20210210 | 20210209 | 20210201 | 20

全体の最終順位

順位	チーム名	スコア
1	まるちゃん	10.89
2	NTTCS	14.71
3	mi-fujita	15.38
4	chscc	15.91
5	のざモト	16.29
6	mh	17.84
7	khiroyuki1993	21.91
8	上智大学 データサイエンス特論チームA	23.08
9	熊本大学 喜多研究室	25.25
10	mprg	25.89
11	koudou-ai	30.12

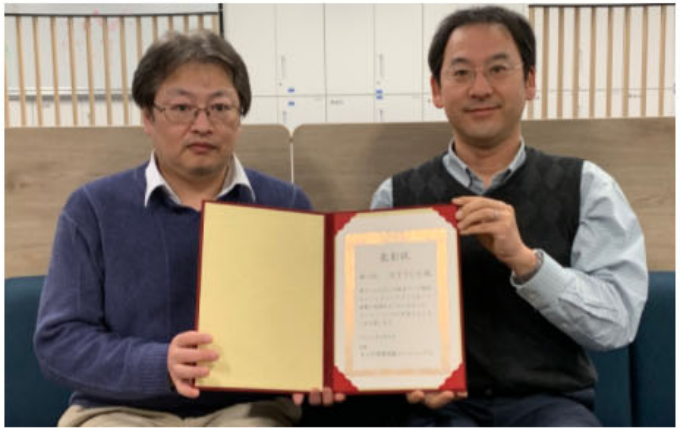
入賞者発表

第1位　まるちゃん様

スコア：10.89

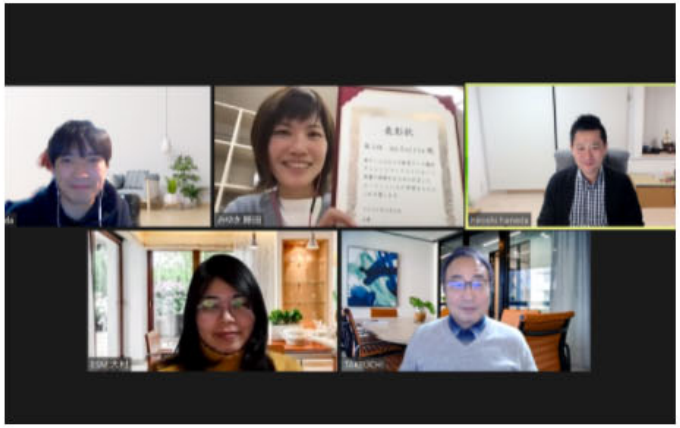
ソースコード：
https://github.com/mtmaru/sip-data-challenge

第2位　NTTCS様



スコア：14.71

第3位　mi-fujita様



スコア：15.38

副賞(図書券5万円)はNTTラーニングシステムズ様からご提供頂きました。

「まるちゃん」様の方法（代理で発表）

データ拡張

- 途中経過における予測に対応させるため、学習データ (15週経過時点のデータ) から疑似的に5週経過時点のデータと10週経過時点のデータを生成し、学習データに加えた。

特徴抽出

- 主な仮説：
- 仮説1：成績上位者に特有の操作パターンがある。
- 例）下位者は一方向にページを送り続けるが、上位者は頻繁にページをジャンプする、など。
- 仮説2：好成績を収めるうえで重要な教材やページがある。
- 例）易しい教材がある、テストに出る内容が書かれているページがある、など。

仮説に対応する特徴量

仮説1：成績上位者に特有の操作パターンがある (※1、※2)。

- 操作の種類のn-gram (1-gram、2-gram、3-gram) 別の操作数
- マーカーの種類別の操作数
- メモの長さ
- 操作時間
- 操作別の操作時間
- デバイスの種類別の操作数
- 時間別の操作数
- 曜日別の操作数
- 週別の操作数

仮説2：好成績を収めるうえで重要な教材やページがある。

- 各教材の閲覧フラグ
- 教材別の操作数
- 教材別・ページ番号別の操作数
- 教材別・ページ番号別の滞在時間

※1 受講した教材の数の影響を除くため、まず教材別に操作数をカウントし、その値を教材で集約した。
※2 集約は、`min` `median` `max` `mean` `std` `sum` を用いた。

手法

- 特徴選択：Null importanceを用いて、約17,000個の特徴量から約650個の重要な特徴量を抽出
- モデル：LightGBM
- 後処理：予測結果が0を下回る場合は0に、100を超える場合は100に置き換え
- 評価：学習データとテストデータの分け方を変えて、ホールドアウト検証を10回繰り返し、その平均値で評価した。
- 1. 5週経過時点のデータ、10週経過時点のデータ、15週経過時点のデータを生成
- 2. 各データから100人ずつ、計300人を抽出し、テストデータにする。
  残りを学習データにする。
- 3. 学習データでモデルを学習させる。
- 4. テストデータの総合成績を予測し、RMSEを求める。
- 5. 1-4を10回繰り返し、RMSEの平均を求める。
- 参考にした解法[2019 Data Science Bow 1st Place Solution - Kaggle](https://www.kaggle.com/c/data-science-bowl-2019/discussion/127469)

この他のこれまでの開催実績

2018

- ICCE - 5th ICCE workshop on Learning Analytics (LA) & Joint Activity on predicting student performance

2019

- ICCE - 6th ICCE Workshop on Learning Analytics (LA) - Scaling Up Evidence-based Institutional LA Practices
- ACM LAK - International Workshop on Technology-Enhanced and Evidence-Based Education and Learning
- IEEE T4E - Technology-enhanced and evidence based education Experience sharing and the road forward.

2020

- ICCE - 7th Workshop on Learning Analytics (LA) Technologies & Practices for Evidence-based Education.
- ACM LAK - Data Challenge: Predicting Performance Based on the Analysis of Reading Behavior

2021

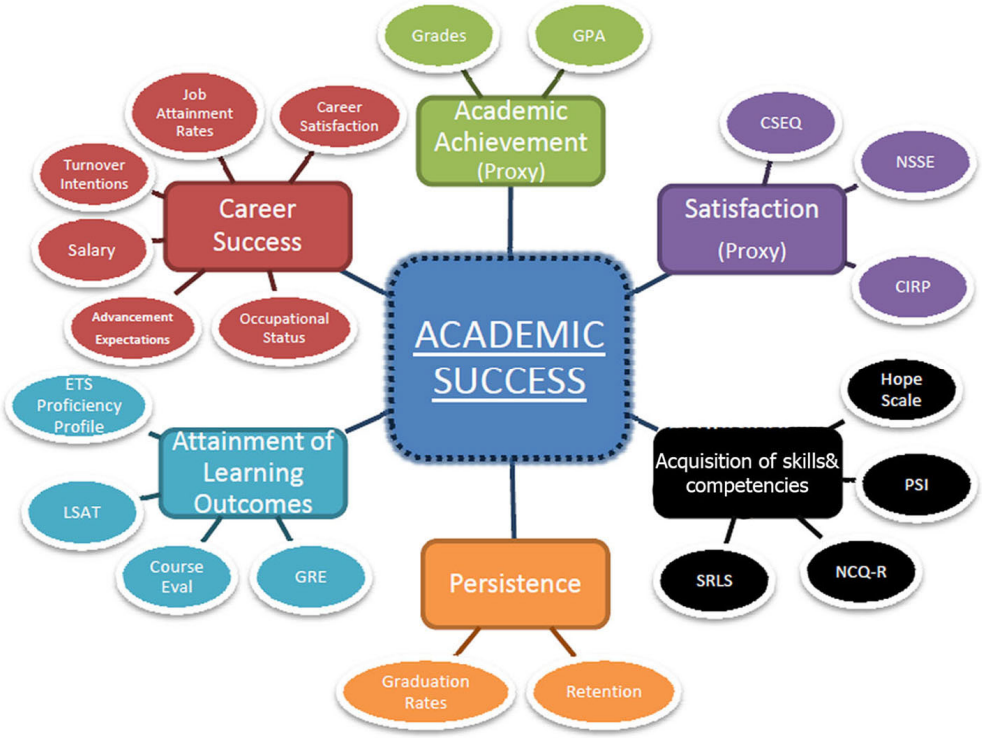
- ACM LAK - Data Challenge: Predicting Performance Based on the Analysis of Reading Behavior

【この他、国際会議EDMなどでも、Data challenge WSを開催している】

関連研究：教育データを用いた予測

予測に用いるデータ

予測する数値



Alyahyan and Düştegö, Predicting academic success in higher education: literature review and best Practices, International Journal of Educational Technology in Higher Education (2020) 17:3, https://doi.org/10.1186/s41239-020-0177-7

おわりに

- 教育改善のためには教育・学習活動のプロセスのデータを蓄積して、利活用することが重要となります。
- 今後も、社会全体での教育データ解析技術の向上と、教育データの利活用の推進のために、教育データ解析コンテストを継続して開催したいと思います。引き続き、どうぞよろしくお願い申し上げます。
- **非営利型一般社団法人　エビデンス駆動型教育 研究協議会**
- **BookRollなどを用いたデータ分析技術の研究交流など**
- https://sites.google.com/view/ederc/

是非、ご入会ください