

複数オンライン学習コースへのアクセスログを用いた 不登校傾向学生の予測

渡邊健太¹

概要： 本研究は大学で不登校傾向の学生についての調査とその学生を支援するための早期予測を目的としている。今回はコロナ禍での学習オンライン環境 Moodle の複数科目への学習ログを用いて予測を行った。データにはオンデマンド形式の授業である情報基礎を受講している学生の全授業のオンライン上アクセスログと情報基礎での出欠表を使用した。本研究では不登校傾向の学生を情報基礎の出席状況と他の科目へのアクセスから定義した。また、複数科目のアクセスログから1週間ごとに複数の変数を作成し、週ごとに予測を行った。機械学習には不均衡データに対する手法、リサンプリング学習をした後にアンサンブル学習を行った。予測精度は主に変数の作成により上げていき、不登校傾向の学生の特徴を調べていった。

キーワード： 不登校, 早期予測, オンライン, 学習ログ, 複数科目, 不均衡データ

Prediction of students who tend to be absent from school using access logs to multiple online learning courses

KENTA WATANABE^{†1}

Abstract: The purpose of this study is to investigate students who tend to be absent from school at university and to make early predictions to support these students. In this study, we used the learning logs of multiple subjects in Moodle, the online learning environment of COVID-19 pandemic. In this study, we used the online access logs of all the classes and the attendance sheets of the students who were taking the on-demand class, Basic Information. In this study, students with a tendency to be absent from school were defined based on their attendance status in Basic Information and access to other courses. Multiple variables were created for each week from the access logs of multiple subjects, and predictions were made for each week. For machine learning, we used a method for imbalanced data, resampling learning, and then ensemble learning. The prediction accuracy was mainly improved by creating variables to examine the characteristics of students who tend to be absent from school.

Keywords: absent, early predictions, online, learning logs, multiple courses, imbalanced data

1. はじめに

大学では高校までと比べて自由な選択が可能になってくが、その反面学生自身が履修登録や単位など多くの決定をする必要がある。そのため、心理的な負担がかかり不登校を起こしやすい時期である[1]。

そこで本研究の目的はLMSの複数の学習ログを用いて不登校傾向学生を早期に予測を行い、不登校に至る前に支援することである。さらに、不登校傾向学生の予測への有効な特徴量の調査である。

本研究では2021年度熊本大学の全学の1年生を対象に開講される情報基礎科目を受講した学生を対象に全科目でのオンラインでのアクセスログと情報基礎科目の出席表を使用し、機械学習を用いた不登校傾向の学生の予測を行った。今回、情報基礎科目を5回連続欠席した学生を対象に情報基礎科目以外の全科目でのアクセスログを見ていき、アクセスログが少ない週が連続3回以上の学生を不登校傾向の学生として定義して予測を行った。

予測する変数はアクセスログから週ごとに複数作成した。機械学習には不均衡データに対する手法[2]であるリサンプリング学習した後にアンサンブル学習を行い、一週間ごとに予測を行った。そして、変数の改良するによって予測の精度向上を目指していき、不登校傾向のある学生の特徴を見つける。

変数では説明変数を前の週との倍率で作成していく。そして、学習習慣を考慮した変数で予測を行った。機械学習ではアンダーサンプリングした後ランダムフォレストで予測を行った。

評価はPrecision（適合率）、Recall（再現率）、F値、AUCで行った。学習習慣を考慮した変数ではPrecisionの値は良くても10%程度であるが、Recallの値は平均で95%を超えるものもあった。よって学習習慣を考慮した変数ではほとんどの陽性を正しく予測することが可能だ。しかし、Precisionの値が低いことから陰性を間違って予測してしまうことが多い。

したがって、本研究では不登校傾向のある学生を陽性と予測することができた。しかし、不登校傾向のある学生で

¹ 熊本大学
Kumamoto University

はない学生も陽性と予測してしまい、不登校傾向ではない学生をうまく予測できなかった。

不登校傾向のある学生は学習習慣を考慮した変数が分類に有効である。不登校になる前に週の時間帯と曜日のアクセス回数に変化がある可能性が高い。特に試験前のアクセスの変化は分類に有効である。しかし、週の時間帯と曜日のアクセス回数に変化があった場合でも不登校にならない学生も多くいた。週の時間帯と曜日のアクセス回数に変化があった学生の調査が必要である。

2. 序論

2.1 研究背景

近年では大学への進学率は上昇しており、多くの人が大学に行くようになった。しかし、その進学者の中でも学習意欲が高くない学生も存在しており、動機がはっきりしない状態で大学へ入学してくる者も少なからずいる。また、大学ではアイデンティティの確立、環境の変化、卒業後の職業選択など多くの事柄に直面する。そのため、不適応を起こしやすい時期であり、学習意欲がない学生はより不登校になってしまう可能性がある。

不登校になってしまう学生は学業への不安、学習意欲の低下などの悩みを抱えている。よって、不登校になってしまう学生を学習への復帰が手遅れになる前に発見し、話し合いを行い、悩みを解決する必要がある。

そこで本研究では不登校になる可能性のある学生を早期に予測を行う。

2.2 研究目的

本研究では学習オンライン環境 Moodle での複数の学習ログを用いて、不登校傾向のある学生を早期に予測する。また、複教科目の学習ログから作成する変数を改良して、対象データに対しての有効な特徴量、また不登校傾向の学生の学習傾向を検討する。

2.3 先行研究

学生の予測の先行研究として近藤らの研究がある[3]。この研究では4年分の新入生の学生のデータからドロップアウトする学生の予測を行っている。実際には3年次の在籍状況を学士課程の初期段階で予測している。使用データは性別、学部、入試種別などの10項目の特徴量である。機械学習に複数のモデルを使用して、データセットは学習データで単年度と複数の年度、テストデータには学習データとは違う年度のものを使用して比較を行っている。結果としては1年次の第5週の終了時点ではおよそ40%、春学期終了時点では50~60%の予測が可能であった。

本研究ではオンラインでのアクセスログのデータを使用して変数の作成を行い、不登校傾向がある学生を予測する。アクセスログのみを使用することでデータ収集の手間が省

け、より詳しく学習履歴についての特徴を調査が可能である。またアクセスログから不登校傾向のある学生の学習習慣の特徴を反映していることが予想される。不登校傾向のある学生は全体に対して非常に少ないと考えられるので不均衡なデータとなることが予測される。そのため、機械学習はデータをアンダーサンプリングしてランダムフォレストを使用し予測を行う。

3. 先行方法

本研究では不登校に関する調査、予測を行う。しかし、不登校の学生は、大学では履修登録が本人に任せられているため早期発見が難しく、授業の出席日数や単位取得状況から不登校と判断されても再び学習に取り組むのは手遅れであることが多い[4]。よって、不登校に至る前に早期発見が必要である。本研究では不登校傾向がある学生について予測を行っていく。不登校傾向のある学生とは、特定の教科を連続で欠席しており、かつ他の科目のアクセスログも少ない学生と定義する。

3.1 研究環境

本研究は2021年度熊本大学の全学の1年生を対象に開講される情報基礎科目を受講した学生1799人を対象としている。この科目は教員が週ごとにオンライン学習環境 Moodle 上へ教材をアップして、学生が週に一回以上アクセスし学習する非同期型の授業である。授業は全15回あり、対象週の課題を週の金曜日までに提出することで出席となる。

3.2 使用データ

本研究では2021年度熊本大学の情報基礎科目を受講している学生1799人のアクセスログデータを使用する。期間は2021年の4月19日から8月6日までの情報基礎科目が開講されている間のデータを使用した。また、アクセスログはオンライン学習環境 Moodle へのアクセスのデータであり、対象の学生1799人の期間内に受講している全科目の Moodle へのアクセスのデータを使用しています。アクセスログの中身については図1に示す。

例

```
235157376,"\\mod_book\\event\\chapter_viewed","mod_book","viewed","chapter","book_chapters",23049,"r",2,1250164,70,594941,54627,93458,"N",0,"N";1619103395,"web",1619103395,"IN"
```

- コースURL : <https://md.kumamoto-u.ac.jp/course/view.php?id=93458>
- ログの「説明」 : The user with id '54627' viewed the chapter with id '23049' for the book with course module id '594941'.
- ログの「時間」 : 2021年04月22日23:56(unixtime=1619103395)

図1 アクセスログの中身

Figure 1 Contents of the access log.

アクセスログのデータの構成は図 1 のようになっており、21 個の要素が入っている。それぞれの要素はカンマで区切られている。左から順に id, eventname, component, action, target, objecttable, objectid, crud, edulevel, contextid, contextlevel, contextinstanceid, userid, courceid, relateduserid, anonymous, other, timecreated, origin, ip, realuserid の情報を持つデータが入っている。このアクセスログデータを使用して変数を作成して予測を行っていく。

次に情報基礎科目の連続 5 回以上欠席した学生の出欠表とユーザー ID を使用した。これは今回、不登校傾向のある学生を予測するため特定の科目を連続で欠席している学生のデータが必要であるため使用している。

3.3 データの前処理

情報基礎科目を受講している学生 1799 人の中に全教科に一度もアクセスがない学生を削除していく処理を行う。これは一度もアクセスがなかった学生はすでに不登校の学生であり、予測には必要ないからである。実際にアクセスログがなかった学生は 8 人存在したので削除を行い、1791 人の学生に絞った。

3.4 変数の作成

本研究は機械学習で予測を行うため変数を作成していく。

3.4.1 目的変数

本研究で予測を行う不登校傾向のある学生とそうでない学生に目的変数を作成していく。今回の目的変数は不登校傾向のある学生が陽性”1”，そうでない学生が陰性”0”として二値分類を行う。

本研究では不登校傾向のある学生を特定の教科を連続で欠席しており、かつ他の科目のアクセスログも少ない学生と定義している。したがって、情報基礎科目を連続 5 回以上欠席した学生で、かつ他の科目でのアクセスログが少ない場合に不登校傾向のある学生とする。これは情報基礎科目でのみ欠席傾向で他の科目には十分なアクセスがある場合は不登校傾向があるとは言えないからである。よって、情報基礎科目以外の全科目でのアクセスログを見ていく。情報基礎科目を欠席し始めた、またしているタイミングに他の科目でもアクセスが減少している学生を陽性として目的変数を作成していく。

目的変数を作成手順はまず他の科目でのアクセスの変化を見ていくため情報基礎科目以外でのアクセスでヒストグラムを作成する。ヒストグラムから閾値を決定し、情報基礎科目以外での活動の多さを決める。そして活動が少ないことが週ごとにわかる表を作成して、情報基礎科目の出欠表と情報基礎科目以外での表で目的変数の作成を行っていく。

まず、情報基礎科目を連続 5 回以上欠席した学生の出欠表とユーザー ID のデータを使用していく。図 2 に情報基礎科目を連続 5 回以上欠席した学生の出欠表を示す。

userid	第1回	第2回	第3回	第4回	第5回	第6回	第7回	第8回	第9回	第10回	第11回	第12回	第13回	第14回	第15回
0 41107	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-
1 42107	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-
2 46727	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3 46960	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-
4 46017	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5 45844	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6 50422	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7 50192	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8 50423	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9 50194	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10 50426	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11 49267	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12 50435	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13 49551	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14 49277	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15 51635	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-

図 2 情報基礎科目の出欠表

Figure 2 Attendance sheets for basic information subjects.

図 2 では情報基礎科目を連続 5 回以上欠席した学生 53 人のユーザー ID と一週ごとの出欠のデータが入っている。

次に情報基礎科目以外の科目への週ごとの 1 科目あたりのアクセス数のヒストグラムを作成していく。情報基礎科目を受講し、かつアクセスのある学生 1791 人を対象に作成していく。また、期間は情報基礎科目が開講されている 2021 年の 4 月 19 日から 8 月 6 日までのアクセスログを使用していく。図 3 に第 1 週の 1 科目あたりのアクセス数のヒストグラムを表示していく。

第1週にアクセスをしていない学生:2人
第1週
期間:2021年 04月19日~04月25日
総アクセス回数: 737925回
最大回数: 86回
最小回数: 0回
平均: 25.46475199086653
標準偏差: 9.757892838720343
 σ の下側信頼限界の値: 15.706859152146189
下側信頼限界外:217人/1791人

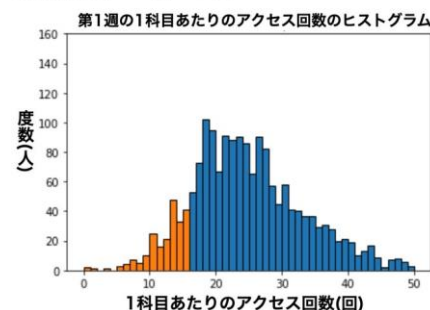


図 3 第 1 週の 1 科目あたりのアクセス数のヒストグラム

Figure 3 Histogram of the number of accesses per subject in the first week.

今回、学生の活動の多さの境界を決定する閾値を σ とする。 σ の信頼区間の下限値以下の場合に活動が少ないとして扱う。 σ の下側信頼限界は全体の約 16%なのでこれを下回ると活動が少ないと言えるため σ を閾値とする。ヒスト

グラムでは σ の下側信頼限界外の学生の色を変えている。週ごとに約200人前後の欠席傾向のある学生がいる。

次に週ごとのヒストグラムから情報基礎科目以外の表を作成していく。対象は情報基礎科目を連続5回以上欠席していた学生53人である。 σ の下側信頼限界外だった週は活動が少ないとして作成していく。作成した情報基礎科目以外の表を図4に示す。

userid	第1回	第2回	第3回	第4回	第5回	第6回	第7回	第8回	第9回	第10回	第11回	第12回	第13回	第14回	第15回
0 41107	-	-	-	-	-	-	-	1	1	1	1	-	1	-	1
1 42107	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2 46727	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3 46960	1	-	-	-	-	-	-	-	1	-	-	-	-	-	1
4 46017	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5 45844	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-
6 50422	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
7 50192	-	-	-	-	-	-	-	1	1	1	-	-	-	-	-
8 50423	-	1	1	1	1	-	1	1	1	1	-	1	1	1	1
9 50194	-	-	-	-	-	-	-	1	1	-	-	-	-	1	1
10 50426	1	-	-	-	-	-	-	1	1	-	-	-	-	-	-
11 49267	-	-	-	-	-	1	-	-	1	-	-	-	-	-	1
12 50435	-	-	-	1	-	1	-	1	-	-	-	-	1	-	-
13 49551	-	-	-	-	-	-	-	-	1	1	-	-	-	-	1
14 49277	1	-	1	-	-	1	-	1	-	-	-	-	1	-	1
15 51635	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

図4 情報基礎科目以外の科目の表

Figure 4 Table of subjects other than basic information subjects

図4は1では活動がある週であり、1以外では活動が少ない週を表している。情報基礎科目以外の表は図2の情報基礎科目と比べて活動が多い週が多い。これは、情報基礎科目は欠席しているが他の科目ではアクセスがあり不登校傾向とは言えない学生がいる。

次に二つの表から教師データを作成していく。二つの出席表でどちらも1ではない週を活動が少ない週とする。どちらかの表が1、もしくはどちらも1の場合は1として作成していく。また、最後の週である第15週のみ情報基礎科目の出欠表をそのまま反映させる。これは他の科目はすでに終了している可能性があるためである。作成した教師データを図5に示す。

1: ['41107', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
2: ['42107', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
3: ['46727', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
4: ['46960', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
5: ['46017', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
6: ['45844', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
7: ['50422', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
8: ['50192', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
9: ['50423', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
10: ['50194', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
11: ['50426', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
12: ['49267', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
13: ['50435', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
14: ['49551', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
15: ['49277', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
16: ['51635', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
17: ['51412', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
18: ['50934', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
19: ['51417', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
20: ['49324', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
21: ['51187', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
22: ['50894', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
23: ['50899', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
24: ['50331', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
25: ['50560', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
26: ['49410', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
27: ['49849', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
28: ['55585', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
29: ['54199', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']
30: ['54464', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1', '1']

図5 二つの出欠表から作成した教師データ

Figure 5 Teacher data generated from two attendance sheets.

図5の教師データで3回連続以上活動が少ない学生を赤色、すべての週で活動が少ない学生を青色で表示している。本研究では教師データで3回連続以上活動が少ない学生を不登校傾向のある学生とする。これは連続3回欠席すると半月以上欠席していることとなり、その後の学習に支障が出るためである。また、本研究では不登校傾向が出始めた学生を早期に予測するため、すべて欠席の学生6人は今回対象から削除していく。したがって、今回のデータでは学生が全体で陽性35人、陰性1750人であった。

次に教師データから目的変数を作成していく。本研究では不登校傾向がある学生を早期に予測するため、不登校傾向の特徴が出始めたタイミングで予測を行う。よって、目的変数は週ごとによって変化するように作成する。今回は3回連続で活動が少なくなり始める最初の週からその学生を陽性"1"とする。例えば7週から9週に連続で活動が少ない学生でも第3週時点では陰性"0"とする。そして第7週から陽性"1"となるように目的変数を作成する。

3.4.2 説明変数

本研究では学習オンライン環境 Moodle アクセスログから変数を作成して予測を行う。アクセスによって変数を作成していくが個人差が大きくなるので学生ごとに比較する。週ごとに変数を作成していき、その週の変数を前の週の変数で割り倍率をとって説明変数とする。

今回は学習習慣を考慮した変数を作成する。生活リズムが登校回避感情を生じる原因となる可能性があること[5]から週の時間帯と曜日のアクセス回数の変化を示す変数を作成する。まず曜日ごとに作成をする。一週間を月曜日から日曜日までに分けて、日ごとのアクセス回数の計7個を作成する。

次に学習習慣を考慮した変数として時間帯と曜日ごとに作成した変数を作成する。一週間を月曜日から日曜日までに分けて、一日を6時間ごとの時間帯に早朝0:00~6:00、朝6:00~12:00、昼12:00~18:00、夜18:00~24:00に分ける。そしてそれぞれのアクセス回数の計28個を作成する。以上の変数を週ごとに作成して予測を行っていく。

3.5 機械学習

2005年の国立大学のアンケート調査で学生の約2.5%が休学に至ったことがあり、約1.5%が退学、約6%が留年をしている[6]。また、大学の新生42名の中で不登校に至った学生は1名という結果[7]から、全体から不登校の学生の割合は非常に少ないことがわかる。本研究の不登校傾向のある学生については最後の週までも1785人中35人と全体の約2%以下という結果である。よって、今回扱うデータは陽性が陰性に比べて極端に少ない不均衡データである。

不均衡データの予測は陽性の分類精度が低くなってし

まう。よって、不均衡データに対する2クラス分類アルゴリズムに関する主流なアプローチで予測を行う。今回はリサンプリング学習とアンサンブル学習を組み合わせたハイブリッドモデルを使用していく。

リサンプリング学習とは、学習データの陽性と陰性を同数になるようにサンプリングして学習させる方法である。その中でも多数の陰性をサンプリングして陽性と同数程度とするアンダーサンプリングを使用していく。アンダーサンプリングを使用することでデータの不均衡性を取り除き、陰性データを削除するので計算コストが向上する。

アンサンブル学習とは、複数の弱学習器を作成してその結果を統合する方法である。その中でもブーストラッピング法によって選ばれたデータで学習を行い、構築された複数の弱学習器を統合するバギングを使用していく。バギングを使用することでバリエーションが低下しやすく、学習を並列で行うことが可能である。今回、バギングは決定木による複数の弱学習器を統合させるランダムフォレストで予測を行っていく。ランダムフォレストを選択した理由は、ブーティングや他のバギングの手法と比較したときに同程度のサンプル数の実行で安定した結果が得られたからである。ランダムフォレストの決定木の数デフォルトの100個として使用していく。

今回は5分割のシャッフル分割交差検証を行っている。またサンプルの0.4をテストデータ、0.6を訓練データとしている。シャッフル分割交差検証は分割するときに毎回訓練データ分のサンプルを選び出し、テストデータは訓練データと重複しないサンプルを選び出す手法である。今回アンダーサンプリングを使用するので陰性データは陽性の分しか学習データに使用できない。よって、5分割ほとんど変わらない陰性データを使用すると過学習となり精度が悪くなる可能性があるため、毎回陰性データをランダムに選び出すシャッフル分割交差検証を使用する。

3.6 評価指標

今回扱うデータ不登校傾向のある学生は全体の学生に比べて非常に少ない。したがって陽性が陰性に比べて極端に少ない不均衡データである。不均衡データに対する分類精度の評価は、陰性と陽性の精度のバランスを考慮した評価が必要である。accuracyのみで評価を行う場合、すべて陰性と予測してもデータの陽性の数が非常に少ないので評価がよくなってしまふ。したがって、accuracyでは正確な評価が困難であるため、accuracy以外の評価指標が必要である。

したがって、本研究ではPrecision（適合率）、Recall（再現率）、F値、AUCを使用していく。

4. 実験

本研究の実験では不登校傾向のある学生を次のような方法で行う。まず予測を行う週は第1~8週までの期間とする。よって、変数は第1~2週、第2~3週、第3~4週、第4~5週、第5~6週、第6~7週、第7~8週の7回予測を行う。2週目はゴールデンウィークにより情報基礎科目の授業が2週にわたっているが、予測も同じように2週分を1週として変数を作成する。また1~8週の期間は今回不登校傾向のある学生を早期に予測を行うことを目的としており、第8週から次の科目が始まるため第8週までとしている。目的変数は不登校傾向のある学生を陽性、そうではない学生を陰性として2クラス分類を行う。また目的変数はそれぞれの週で変化していき、対象の週から3回連続活動の少ない学生とそれ以前にすでに3回連続活動が少ない学生を陽性とする。

機械学習にはアンダーサンプリングとランダムフォレストを使用したハイブリッドモデルを使用しており、5分割のシャッフル分割交差検証を行う。結果は予測を5回行ったものの平均をスコアとして算出している。また、スコアはPrecision（適合率）、Recall（再現率）、F値、AUCの4つである。表1に1~8週目での目的変数を示す。

表1 週ごとの目的変数

Table 1 Objective variables by week.

	1~2 週	2~3 週	3~4 週	4~5 週	5~6 週	6~7 週	7~8 週
陽性	15	17	18	22	24	24	25
陰性	1770	1768	1767	1763	1761	1761	1760

曜日ごとの学習習慣を考慮した変数7個と時間帯と曜日ごとの学習習慣を考慮した変数28個を使用する曜日ごとの学習習慣を考慮した変数は月曜日から日曜日の曜日ごとのアクセス回数の計7個の変数を作成する。時間帯と曜日ごとの学習習慣を考慮した変数は月曜日から日曜日のそれぞれの曜日を4つの時間帯に分けてそれぞれの時間帯にあったアクセス回数の計28個を使用して変数を作成していく。そしてそれぞれ対象の週で出た数値を前の週との倍率をとった変数として使用していく。

5. 実験結果

学習習慣を考慮した変数を使用した2つを比較している。表2にそれぞれのPrecision（適合率）、Recall（再現率）、F値、AUCの結果を示す。

表 2 実験結果

Table 2 Results of experiments

週	説明変数	Precision	Recall	F 値	AUC
1~2	曜日	0.0262	0.966	0.0508	0.914
	時間帯と曜日	0.0226	0.967	0.0441	0.911
2~3	曜日	0.0346	0.743	0.0660	0.830
	時間帯と曜日	0.0315	0.829	0.0605	0.862
3~4	曜日	0.0354	0.704	0.0660	0.800
	時間帯と曜日	0.0380	0.889	0.0725	0.892
4~5	曜日	0.0387	0.911	0.0742	0.895
	時間帯と曜日	0.0479	0.933	0.0906	0.914
5~6	曜日	0.0596	0.940	0.112	0.924
	時間帯と曜日	0.0715	0.980	0.133	0.949
6~7	曜日	0.102	0.880	0.181	0.916
	時間帯と曜日	0.0946	0.980	0.174	0.961
7~8	曜日	0.0814	0.840	0.144	0.883
	時間帯と曜日	0.667	0.940	0.123	0.927

図 6, 図 7 に週の変化による予測の結果を示している。

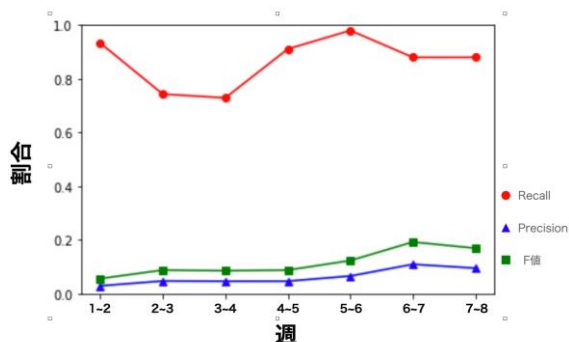


図 6 曜日ごとの変数の予測結果

Figure 6 Prediction results of variables for each day of the week.

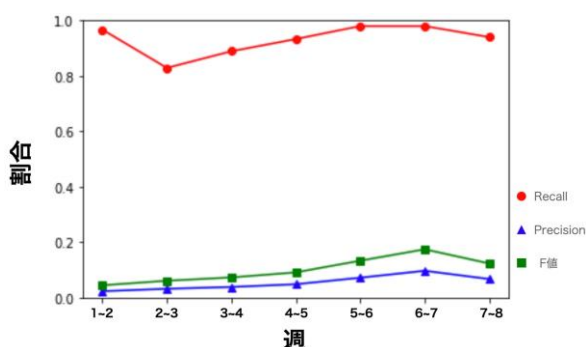


図 7 時間帯と曜日ごとの変数の予測結果

Figure 7 Prediction results for variables by time of day and day of week.

図 6, 図 7 は縦軸が割合, 横軸が週である。

次に図 8 に曜日ごとの学習習慣を考慮した変数, 時間帯と曜日ごとの学習習慣を考慮した変数の AUC の変化を示している。

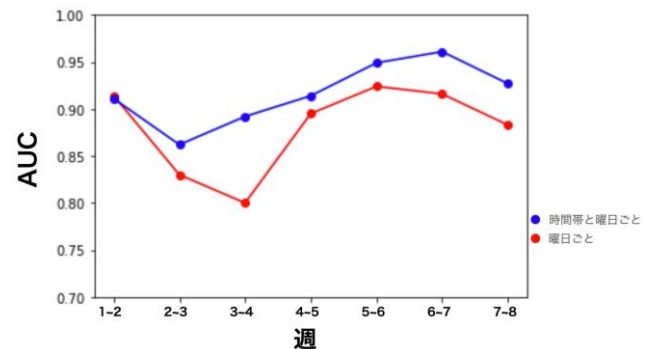


図 8 それぞれの変数の AUC の比較

Figure 8 Comparison of AUC for each variable

図は縦軸が AUC, 横軸が週である。また, 横軸は 0.70~1.00 の範囲を表示している。時間帯と曜日ごとの学習習慣を考慮した変数の方が精度は良い。

6. 考察

全体的に Precision はほとんどが 1 割以下で非常に低く, Recall はほとんどが 8 割を超えており高い。図 6, 図 7 からどちらの変数でも 2~4 週で精度が悪くなっているが後半は精度が向上している。図 8 からは曜日だけでなく時間帯ごとに作成した方が AUC 値は高い。

学習習慣を考慮した変数では後半に精度が向上しているのは, 7~8 週の期間に多くので試験が行われるためであると考えられる。5~7 週に試験の勉強のためにいつもとは違う時間のアクセスが増え, 不登校傾向のある学生との差が顕著となり予測精度が向上した。また 2~4 週で精度が悪くなっているのは 2 週目にゴールデンウィークで期間が長くなっているためであると考えられる。比較したことより, 学習習慣を考慮した変数は曜日ごとだけでなく, 時間帯で分けた方が不登校傾向のある学生の予測に有効である。よって不登校傾向のある学生を予測するのに学生の学習習慣の変化を見ることが有効であると考えられる。

7. 結論

本研究では複数の科目のアクセスログから不登校傾向のある学生の予測を行った。実験の結果より、学生の学習習慣の変化を示す特徴量が予測には有効であるという結果となった。そして不登校傾向のある学生の予測は試験前のアクセス回数の変化に注目することが有効である。今回の実験から不登校傾向のある学生の早期予測は Recall が高くなったが Precision は低い結果となった。不登校傾向のある学生を分類はできているが不登校傾向のない学生を誤って分類している。しかし、不登校傾向のある学生を正しく分類できているので早期に予測して支援することが可能である。

今後の課題として Precision の向上のために不登校傾向のある学生と誤って不登校傾向のある学生と判断した学生の分類ができるような特徴量の調査が必要である。また、今回は対象の週から欠席が始まっている学生を予測したが、1,2 週前から欠席が始まる学生の予測を行っていく。そして今回は半年分のデータを使用しているが年度の違うデータや対象の学生を増やすことが必要である。

謝辞 本研究を進めるにあたり、多くのご指導を賜りました。総合情報統括センター研究室の杉谷賢一教授、中野裕司教授、久保田真一郎准教授には心から感謝いたします。また同研究室の戸田真志教授、右田雅裕助教授、武蔵泰雄教授にも多くのご助言をいただきました。厚く御礼申し上げます。そして、同研究室の皆様にも大変お世話になりました。ここに深く感謝の意をお伝えいたします。

参考文献

- [1] 三宅典恵, 岡本百合, 大学生のメンタルヘルス. 心身医学. 2015, vol. 55, no. 12, p. 1360-1366.
- [2] 藤原和樹, 繁野麻衣子, 住田潮, 不均衡データに対する多段階学習を用いたアンサンブルモデルによる 2 クラス分類アルゴリズムの提案. 情報処理学会論文誌数理モデル化と応用. 2019, vol. 12, no. 3, p 10-17.
- [3] 近藤伸彦, 畠中利治, 学士課程における大規模データに基づく学修状態のモデル化. 教育システム情報学会誌. 2016, vol. 33, no. 2, p94-103.
- [4] 坂原泰子, 石川健介, 大学生の連続欠席者における不登校傾向得点の検討. 日本教育心理学会第 59 回総合発表論文集. 2017, p411.
- [5] 成田正樹, 宮戸美樹, 大学生の登校回避感情が生じる要因. 日本心理学会第 82 回大会. 2018, p242.
- [6] 内田千代子, 大学生における休・退学、留年学生に関する調査 第 28 報. 「休・退学、留年学生調査」事務局（茨城大学保健管理センター内）.2008.
- [7] 栗原久, 萩野基行, 大学入学時の自記式健康度調査(THI)による長期授業欠席リスクの高い学生の予測. 東京福祉大学・大学院紀要. 2012, vol. 2 no. 2, p115-121.