ラーニング・アナリティクス、性格特性、教学 IR データを活用した GPA と学校生活満足度の予測モデルの開発

片瀬拓弥

Development of prediction model of GPA and school life satisfaction by Learning Analytics, personality traits and data of Institutional Research for education

Takuya Katase

要旨

本研究では、ラーニング・アナリティクス、性格特性、教学 IR データを活用し、１年春学期末の GPA と学校生活満足度を予測するモデルを開発した。モデル開発手法として、k-means 法によるクラスター分析、線形重回帰分析及びニューラルネットワークを採用した。予測精度を計算した結果、線形重回帰モデルの最大決定係数は、（GPA，学校生活満足度）＝（0.279，0.305）であり、ニューラルネットワークモデルの最大決定係数は、（GPA，学校生活満足度）＝（0.831，0.479）となった。

キーワード：ラーニング・アナリティクス、教学 IR、GPA、学校生活満足度、予測モデル

１．はじめに

昨今、教育工学という学問分野において、ラーニング・アナリティクス（以下、LA）と教学 IR との関連性について報告がされている（松田・渡辺 2017）。LA とは、学習管理システム（Learning Management System：以下、LMS）などの「学習状況を把握し最適化させるため、学習者とそれを取りまく文脈に関わるデータを測定、収集、分析、報告する方法（LAK'11 2011）」であり、教学 IR とは「大学の意志決定を支援するための情報収集やデータ分析活動の中でも教学部門に特化されたもの

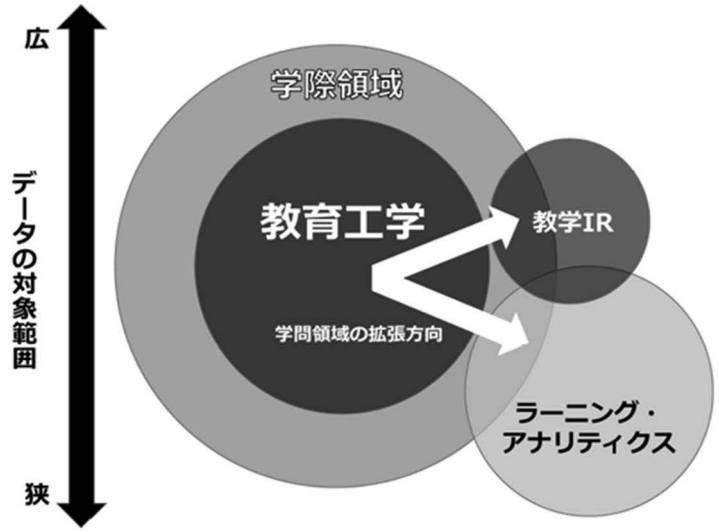

図１　教育工学の広がりとラーニング・アナリティクス、教学 IR（松田・渡辺 2017 より転載）

（高橋ら 2014)」とされている。**図１**は、松田・渡辺（2017）が提示した「教育工学の広がりとラーニング・アナリティクス、教学 IR」の位置づけ図である。さらに松田・渡辺（2017）は「LA と教学 IR は教育工学の隣接分野であると考えられるので，新たな学際性の形成領域として、両者との連携を積極的に図ることが、複雑な営みである教育をひもとく手がかりの拡大につながるであろう」と述べている。また、船守（2014）も、教学 IR で扱う学務情報（入試、履修、成績、学生生活、就職等）と LA で扱う LMS の学習ログ等の「両者を組み合わせた方が精度のよい解析と、それに伴う的確な学習支援となることは十分推測される」とし、両者を組み合わせた「精緻な教学 IR」の必要性を指摘している。つまり、**図１**の「LA と教学 IR の領域が重なる分野」の研究が求められている。

さて、このような研究領域において、片瀬 (2017) は、LA データ、性格特性、教学 IR データ (hyper-QU 大学版）を組み合わせた学生支援モデルを試作した。hyper-QU 大学版とは、河村（2010）によって開発・標準化された学生生活に関する記名式アンケート調査であり、学校生活満足度などの教学 IR データを含んでいる。片瀬（2017）は、学生支援ニーズを予測するためには、分析データとして「LA データ、教学 IR データ」とともに「性格特性」も考慮すべきであるとしている。

そこで、本研究では、LA データ、性格特性、教学 IR データ（入試種別、出身高校偏差値、日本語プレースメントテスト；以下、日本語 PL テスト）を活用し、１年春学期末の GPA と学校生活満足度の予測モデルを開発することを目的とする。ただし、モデル開発手法として、k-means 法によるクラスター分析（教師なし機械学習)、線形重回帰モデル及びニューラルネットワークモデル（教師あり機械学習）を採用する。

２．予測モデルの開発方法

２．１　予測モデル開発工程の概略

図２は、本研究における予測モデルの開発工程図である。**図２**の A～F 部は、それぞれ、A 部が「LA データとしての活用するリメディアル教材の LMS 学習ログ」、B 部が「性格特性を決定するために活用する性格検査（調査データ)」、C 部が「k-means 法によるクラスター分析（教師なし機械学習)」、D 部が「教務学生部等から提供を受ける教学 IR データ」、E 部が目的変数である「１年春学期末の GPA と学校生活満足度（hyper-QU)」を示している。そして、F 部が「線形重回帰モデル及びニューラルネットワークモデル（教師あり機械学習）による予測モデルの開発」を示す。

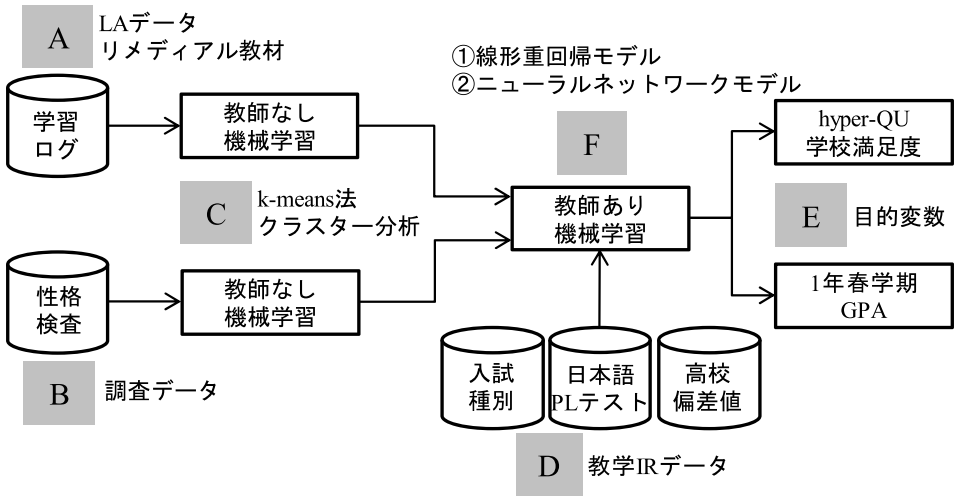

図２　予測モデルの開発工程図

次節から、これらの段階ごと具体的方法を述べていくことにする。

２．２　ＬＡデータの活用

図２のA部のLAデータの活用について述べる。本研究の説明変数として活用するLAデータは、本校の入学前オンライン学習の LMS として活用しているラインズドリルベーシックコース（2018）の学習ログとする。このドリルの目的は、高校までの５教科（国語、数学、英語、理科、社会）のリメディアル教材（演習ドリルと実力診断テスト）を行い、入学生の基礎学力の確認と向上を目指すものである。ただし、本校では「理科」を除外して実施している。また，入学者全員に対して入学年度5月上旬までに課題範囲の全クリアを課しているため、比較的強い「外発的動機づけ」がされている。このラインズドリルは、パソコンだけでなくスマホやタブレットからも学習可能なeラーニングシステムである。学習ログは、ラインズ社によって一次加工（学生ごとの項目別集計）が施されたCSV形式として提供されている。本研究で活用する学習ログは、一次加工が施された以下の変数とする。

①ログイン回数（入学年度5月上旬までのログイン回数）
②総学習時間（入学年度5月上旬までの総学習時間）
③ログイン1回ごとの平均学習時間（②／①）
④クリア率（ベーシックコースのクリア率；0～100%）

分析は、①～③について、それぞれ正規化を行い「偏差値（Z得点）」に変換後、**図２**のC部「クラスター分析」に投入する。一方、④については、正規化を行わず「生データ」のまま、**図２**のF部の説明変数として投入する。

２．３　性格検査（調査データ）の活用

図２のB部の性格検査（調査データ）の活用について述べる。本研究の性格検査として活用するのは、NEO-FFI（NEO Five Factor Inventory）である。NEO-FFIは、NEO-PI-R（Revised NEO Personality Inventory）の短縮版であり、5つの性格因子各12項目、計60項目で構成される（下仲ら 2011）。5因子性格検査（BIG5 Personality Inventory）の一種であり、因子として「外向性、調和性、誠実性、開放性、神経症傾向」からなる。対象学生らには、入学年度5月までに研究同意を得た上でデータを取得している。分析は、各性格因子について正規化を行い「偏差値（Z得点）」に変換後、**図２**のC部「クラスター分析」に投入する。

２．４　k-means法によるクラスター分析

図２のC部の「k-means法によるクラスター分析（教師なし機械学習）」について述べる。k-means 法は、機械学習において「教師なし学習」のクラスタリング問題を解くためのアルゴリズムである。片瀬（2017）の先行研究では、LAデータの学習ログから得られる「学習スタイル」のクラスター数を3つとしており、本研究も同様に「3 クラスター（L1、L2、L3）」とする。また、性格検査から得られる「性格特性」のクラスター数を4つとしており、本研究でも同様に「4クラスター（P1、P2、P3、P4）」とする。このクラスター分析により、次段階の「教師あり機械学習」へ投入する情報量が削減されるとともに、投入前の説明変数について、パターン情報という有意味な解釈を与えることができる。学習スタイルと性格特性のパターン情報は、**図２**のF部の「教師あり機械学習」に投入される前に、それぞれをダミー変数（0 or 1）に変換して分析する（ケース１）。さらに、k-means法によるクラスター分析は、原理的に各クラスター中心からの距離が最小となるクラスターを所属クラスターとし

ている。つまり、データごとに持っている距離情報は、所属クラスターを決定する要因であると同時に各クラスター中心からの距離という 2 つの情報を含んでいる。よって、この距離情報は、ダミー変数とほぼ同じ意味を持ちながら情報量が多く、予測モデルの精度向上に役立つ可能性がある。したがって、この距離情報についても、**図２**の F 部の「教師あり機械学習」に投入して予測モデルの開発に用いる（ケース２）。3 つの学習スタイルのダミー変数（L1、L2、L3）と各クラスター中心からの距離変数をそれぞれ（L1d、L2d、L3d）とし、4 つの性格特性のダミー変数（P1、P2、P3、P4）と各クラスター中心からの距離変数をそれぞれ（P1d、P2d、P3d、P4d）とする。ただし、結果の解釈的に同じ情報の混在を避けるため､各ダミー変数とそれに対応する距離変数の同時投入はしないものとする。

２．５　教学ＩＲデータの活用

図２の D 部の教学 IR データの活用について述べる。本研究で活用する教学 IR データは、本校の教務学生部から提供を受ける入試種別データ、出身高校偏差値、入学前に実施する日本語 PL テスト結果である。入試種別は、AO・自己推薦入試、特別推薦入試、指定校・公募推薦入試、センター・一般入試の 4 パターンである。それぞれをダミー変数（0 or 1）に変換し、略記号として、AO・自己推薦入試（Adm)、特別推薦入試（Pdm)、指定校・公募推薦入試（Sdm)、センター・一般入試（Cdm）とする。これらのダミー変数を**図２**の F 部の「教師あり機械学習」に投入する。また、出身高校偏差値は、民間会社が提供している情報を取得した（みんなの高校情報 HP 2018)。日本語 PL テストは、市販されている基礎学力判定テスト（日本語）を利用する（旺文社教学支援サービス 2018)。出身高校偏差値と日本語 PL テストは、「生データ」のまま、**図２**の F 部の「教師あり機械学習」に投入する。

２．６　ＧＰＡと学校生活満足度（目的変数）

図２の E 部にある 2 つの目的変数について述べる。

①GPA

1 つ目の目的変数は、教学 IR 分析においてよく活用されている一般的な成績指標である GPA を採用する。本校の GPA は、原則、以下の計算式で定義されている。

GPA＝{ 秀(S)評価の単位数×4＋優(A)評価の単位数×3＋良(B)評価の単位数×2
＋可(C)評価の単位数×1 } ／ 履修済単位数の合計

②学校生活満足度

2 つ目の目的変数である学校生活満足度は、市販されている hyper-QU 大学版を活用する。hyper-QU 大学版は、河村（2010）によって開発・標準化された学生生活に関するアンケート調査のことである。河村（2010）が開発・標準化した学校生活満足度尺度は、2 つの下位尺度（X 軸：侵害点、Y 軸：承認点）から構成され、XY 座標上の位置から 4 群（満足群、非承認群、侵害行為認知群、不満足群）に分けることで学校生活満足度を表現している。本研究では、これら 2 つの下位尺度点を加工し、一次元の「学校生活満足度の偏差値」を新たに定義することにした。計算方法は、まず承認点と侵害点をそれぞれ正規化し、承認 Z 偏差と侵害 Z 偏差に変換する。次に侵害 Z 偏差が、学校生活満足度に対して反転項目であることを考慮し、R 侵害 Z 偏差=（100－侵害 Z 偏差）を定義し、値を反転する。その後、原点（各 Z 偏差の平均 50）を中心とした反時計回り 45 度の座標回転を行い、以下の計算式をもって、「学校生活満足度の偏差値」と定義する。そして、この数値をモデル開発の目的変数とする。

学校生活満足度の偏差値＝1／$\sqrt{2}$ ×{ (承認 Z 偏差－50)　＋ (R 侵害 Z 偏差－50) } ＋ 50

２．７　線形重回帰モデル及びニューラルネットワークモデル

図２のF部にある「教師あり機械学習」によるモデル開発方法について述べる。まず、これらのモデル開発の元となる説明変数の候補と目的変数は以下である。説明変数は、4 つの入試種別のダミー変数、3 つの学習スタイル（LI、L2、L3）のダミー変数又はクラスター中心からの距離変数（L1d、L2d、L3d）、4 つの性格特性（P1、P2、P3、P4）のダミー変数又はクラスター中心からの距離変数（P1d、P2d、P3d、P4d）、LA データの LMS 課題のクリア率、出身高校偏差値、日本語 PL テストの計 14 個である。目的変数は、１年春学期末の GPA と学校生活満足度の偏差値の 2 つである。

①線形重回帰モデル

線形重回帰モデルでは、ダミー変数の性質上、14 個の全変数をモデルに投入すると多重共線性（線形結合）がいくつか発生してしまう。したがって、それぞれのダミー変数の中から 1 つずつ除外してモデル開発を行う。つまり、ダミー変数を用いた線形重回帰モデル（ケース１）の説明変数は 11 個となる。除外するダミー変数は、各目的変数との単相関分析から判断する。各目的変数に対する線形重回帰分析の決定係数（R^2）を予測精度の目安とする。一方、学習スタイルと性格特性のクラスター中心からの距離変数をモデル開発に投入する場合、多重共線性の有無を確認し、問題無ければ全変数を投入する。つまり、距離変数を用いた線形重回帰モデル（ケース２）の説明変数は最大 13 個となり、各目的変数に対する線形重回帰分析の決定係数（R^2）を予測精度の目安とする。

②ニューラルネットワークモデル

このモデルは、非線形モデルであるため、変数間に線形結合が発生していても問題がない。よって、14 個の全変数をモデルに投入する。解析ツールは、MATLAB（2018）を使用した。ネットワークの学習にはベイズ正則化法を用い、全データの 50%をモデル開発用、25%をモデルの過適合防止用、残り 25%を完全独立したモデル検証用として使用する。また、隠れニューロン数は 10 個とし、数十回の学習試行の中でモデル検証用の決定係数（R^2）が最大となる予測モデルを選択する。

３．予測モデルの開発と検証

３．１　対象者

対象者は、本校の X 年度～（X+3）年度に入学した 1 年生、計 331 名を対象とした。教務学生部か

表１　データの記述統計量と単相関係数（n=331）

変数名	M	SD	単相関係数											
			目的変数		LAデータ				調査データ（性格検査）					教学
			GPA	SAT	LogCn	SumLn	AveLn	Crp	E	A	C	O	N	PLT
GPA	2.58 (0.39)		—											
SAT	50.0 (12.1)		.13 *	—										
LogCn	31.4 (21.4)		.07	.04	—									
SumLn	10h11m 5h22m		-.04	-.04	.43 **	—								
AveLn	0h22m 0h10m		-.12 *	-.04	-.47 **	.38 **	—							
Crp	0.95 (0.16)		.35 **	.13 *	.21 **	.30 **	.08	—						
E	26.4 (6.91)		.01	.46 **	.11 *	.06	-.03	.16 **	—					
A	31.8 (5.43)		.16 **	.34 **	.12 *	.09	.03	.12 *	.38 **	—				
C	27.0 (6.13)		.17 **	.33 **	.14 **	.11 *	-.01	.07	.28 **	.28 **	—			
O	27.8 (5.76)		.10	.01	-.07	-.12 *	-.02	-.02	.05	.04	-.05	—		
N	27.9 (6.13)		.05	-.28 **	.02	.07	.01	.03	-.19 **	-.22 **	-.37 **	.23 **	—	
PLT	549.7 (86.7)		.28 **	-.16 **	-.20 **	-.24 **	.04	.03	-.27 **	-.07	-.24 **	.26 **	.06	—
HiZ	48.1 (5.63)		.28 **	.09	-.06	-.22 **	-.10	.11 *	.03	.06	-.04	.13 *	.04	.28 **

** p<.01 ，* p<.05

ら提供された各種個人情報データは、モデル開発に必要なデータを統合後、個人情報が特定できないように匿名化処理を行った。また、欠損データについては、ペアワイズ方式による分析とした。

３．２　データの記述統計

モデル開発に必要なデータの記述統計量を示すため、各変数の略記号を以下のカッコ内の英文字として定義する。以後の結果は、この略記号を用いる。**表１**に各種データ（入試種別のダミー変数を除く）の記述統計量と単相関係数を示す。

目的変数：１年春学期末の GPA（GPA）、学校満足度の偏差値（SAT）

LA データ：ログイン回数（LogCn）、総学習時間（SumLn）、平均学習時間（AveLn）、クリア率（Crp）

調査データ：外向性（E）、調和性（A）、誠実性（C）、開放性（O）、神経症傾向（N）

教学 IR データ：日本語 PL テスト（PLT）、出身高校偏差値（HiZ）、

表１によれば、目的変数である GPA と SAT に対し、LA データの一部や性格検査の一部が有意な相関を持っていることが分かる。

３．３　k-means 法によるクラスター分析結果

クラスター分析は、①学習スタイルのクラスター分析、②性格特性のクラスター分析に分けて行う。また、クラスターの解釈を容易にするため、投入前に各変数を偏差値（Z 得点）に変換してから分析を行った。

①学習スタイルのクラスター分析

図３に学習スタイルのクラスター分析結果を示す。各クラスターの数値が平均的範囲内かどうかの解釈は「±0.5σ（偏差値 45～偏差値 55）」を判断目安とする。L1 は、全変数において平均を下回っているものの、各偏差値は 45 付近となっており、179 名（54%）の学生が「L1 クラスター」と分類された。所属する人数割合を考慮すると比較的「平均的な学習スタイル」といえる。L2 は、総学習時間と平均学習時間が比較的長いスタイルである．つまり、１回の平均学習時間が「長時間型の学習スタイル」といえる。100 名（30%）の学生が「L2 クラスター」と分類された。L3 は、ログイン回数が非常に多く、平均学習時間が短い学習スタイルである．つまり、１回の平均学習時間が「短時間型の学習スタイル」といえる。52 名（16%）の学生が「L3 クラスター」と分類された。

②性格特性のクラスター分析

図４に性格特性のクラスター分析結果を示す。各クラスターの数値が平均的範囲内かどうかの解釈

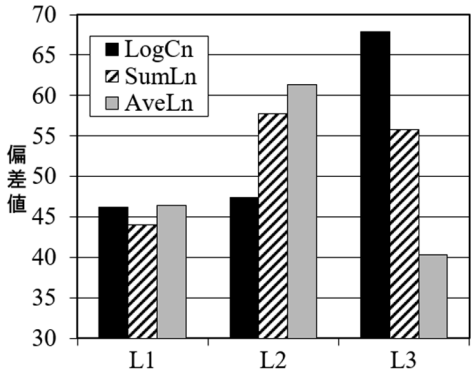

図３　学習スタイルのクラスター分析

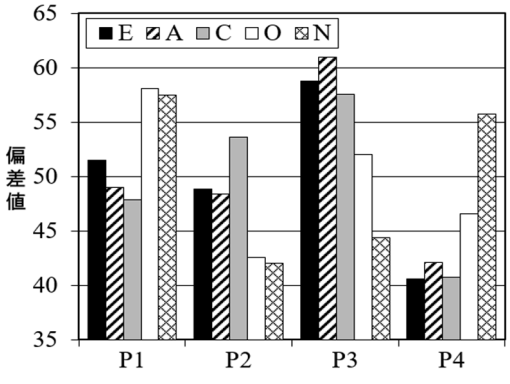

図４　性格特性のクラスター分析

は①と同様とする。P1 は、外向性、調和性、誠実性が平均的、開放性と神経症傾向が上位の性格特性である。つまり、開放性の解釈の一種である知的好奇心は高いが、情緒が比較的安定していない性格特性である。93 名（28%）の学生が「P1 クラスター」と分類された。P2 は、外向性、調和性、誠実性が平均的、開放性と神経症傾向が下位の性格特性である。つまり、知的好奇心は低いが、情緒は比較的安定している性格特性である。P1 とは相対的に逆の性格特性ともいえる。87 名（26%）の学生が「P2 クラスター」と分類された。P3 は、外向性、調和性、誠実性が高く、神経症傾向が低い性格特性である。つまり、一般的に社会適応力が高いと考えられている性格特性である。76 名（23%）の学生が「P3 クラスター」と分類された。P4 は、外向性、調和性、誠実性が低く、神経症傾向が高い性格特性である。つまり、社会適応力があまり高くない可能性のある性格特性である。P3 とは相対的に逆の性格特性ともいえる。75 名（23%）の学生が「P4 クラスター」と分類された。

３．４　線形重回帰モデルによる予測モデルの開発

予測モデルの開発に先立って、入試種別、性格特性、学習スタイルの各ダミー変数と GPA と SAT との単相関分析を行った。その結果、GPA に対する単相関係数が最も低い説明変数をそれぞれ除外することにした。なぜなら、各ダミー変数の内、それぞれ 1 つについては、原理的に多重共線性（線形結合）が発生するためである。単相関分析結果に従い、GPA 予測モデルでは、Sdm、P2、L1 を除外し、SAT 予測モデルは、Adm、P2、L3 を除外した。一方、学習スタイルと性格特性の各クラスター中心からの距離である（L1d、L2d、L3d）と（P1d、P2d、P3d、P4d）についても単相関分析を行い、多重共線性（線形結合）の有無を確認した。その結果、全てにおいて多重共線性が発生していないことを確認したため、これらの全変数をモデル開発に活用することにした。

さて、クラスター分析により分類された「学習スタイルと性格特性」について、以下のようなケースに分けて分析結果を説明する。

表２　線形重回帰分析結果

ケース１

説明変数			目的変数 GPA β	目的変数 SAT β
ダミー変数	入試種別	Adm	-.080	none
		Pdm	.088 †	
		Sdm	none	
		Cdm		
ダミー変数	性格特性	P1		-.135 *
		P2	none	none
		P3		.224 ***
		P4	-.150 **	-.332 ***
	学習スタイル	L1	none	.082
		L2	-.107 *	
		L3		none
		PLT	.223 ***	-.102 *
		HiZ	.174 ***	.083
		Crp	.301 ***	.093 †
決定係数（R^2）			.277	.268
n			323	322

ケース２

説明変数			目的変数 GPA β	目的変数 SAT β
ダミー変数	入試種別	Adm	-.081	none
		Pdm	.089 †	.078
		Sdm	none	
		Cdm		
距離変数	性格特性	P1d		
		P2d		
		P3d	-.087 †	-.356 ***
		P4d	.094 †	.282 ***
	学習スタイル	L1d		-.075
		L2d		
		L3d	-.119 *	
		PLT	.240 ***	-.086 †
		HiZ	.198 ***	.104 *
		Crp	.258 ***	
決定係数（R^2）			.279	.305
n			323	322

注）$^{***}p<.001$, $^{**}p<.01$, $^{*}p<.05$, $^{\dagger}p<.05$　none:投入なし

ケース１：クラスター情報がダミー変数の場合
ケース２：クラスター情報が各クラスター中心からの距離の場合

表２にケース１とケース２の線形重回帰分析（ステップワイズ変数増減法、変数投入基準:p 値<0.2）の結果を示す。**表２**を見ると、GPA 予測に 5%以下の有意確率で寄与している標準偏回帰係数 (β) は、ケース１では P4、L2、PLT、HiZ、Crp であり、ケース２では L3d、PLT、HiZ、Crp である。両者に共通する説明変数は、PLT、HiZ、Crp である。次に SAT 予測に 5%以下の有意確率で寄与している標準偏回帰係数 (β) は、ケース１では P1、P3、P4、PLT であり、ケース２では P3d、P4d、HiZ となった。両者に共通する説明変数は、P3（P3d）と P4（P4d）である。P3 と P3d、P4 と P4d の符号が反転しているのは、ケース１では「あるクラスターに属する場合に 1」としている一方、ケース２では「あるクラスターに属する場合は、その距離の数値が最も小さい」からであって、符号反転に矛盾はない。これらを総合的に見ると、GPA 予測では、学習スタイルと性格特性の双方からの影響はあると考えられるが高いとはいえない。一方、SAT 予測では、性格特性の影響が明らかに存在するが、学習スタイルからの影響はほぼないと考えられる。また、GPA と SAT の全ケースにおいて、PLT の寄与が存在する可能性が高い。PLT は GPA 予測ではプラスの影響であるが、SAT 予測ではマイナスの影響である。このことは、PLT 得点の高い学生達は「本校に不本意入学」している可能性があり、学校不適応感が

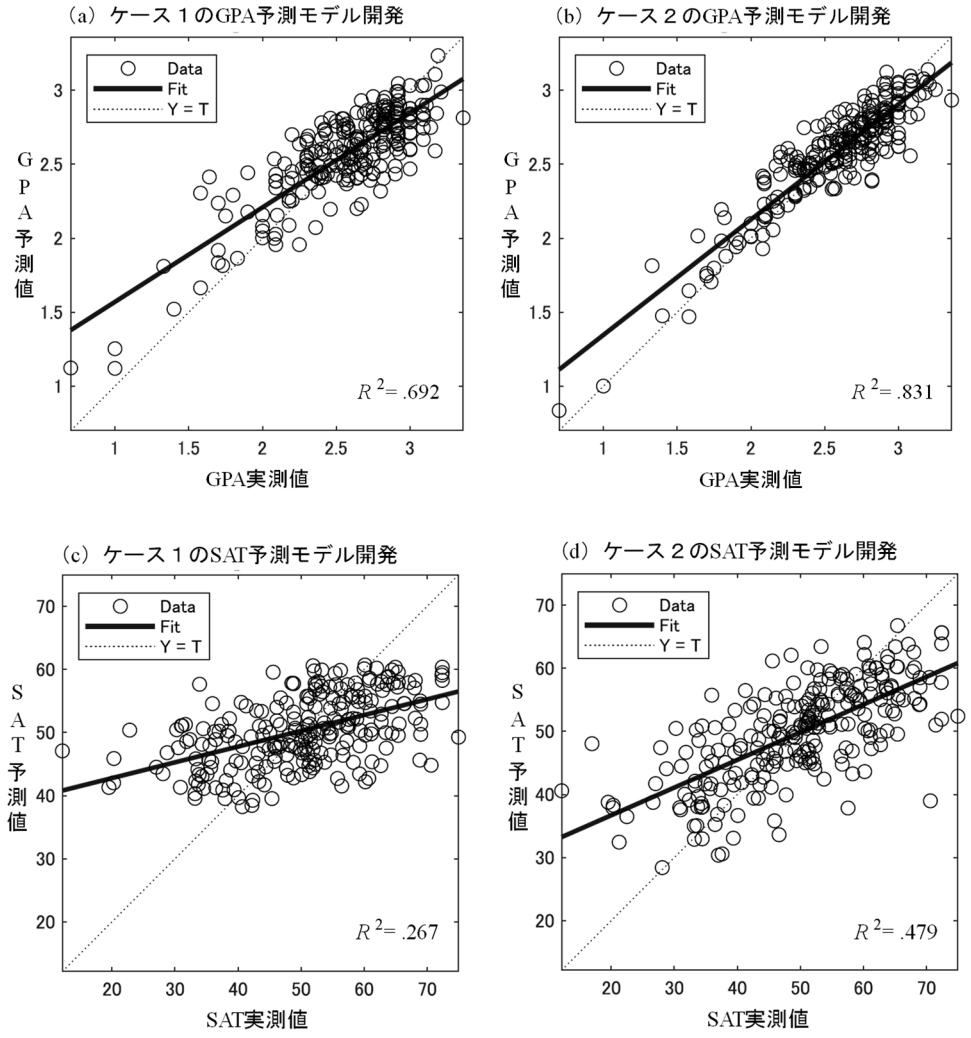

図５　ニューラルネットワークモデルによる GPA 及び SAT 予測モデルの開発

影響しているのかもしれない。最後に、各ケースにおける GPA 予測と SAT 予測の決定係数（R^2）を比較するといずれもケース２の方が若干ながら高くなっている。よって、線形重回帰モデルの最大決定係数（予測率）は、ケース２の（GPA，SAT）＝（0.279，0.305）とする。

３．５　ニューラルネットワークモデルによる予測モデルの開発

先にも述べたが、ニューラルネットワークモデルは非線形モデルのため、変数間に線形結合が発生していても問題はない。よって、14 個の全変数をモデル開発に投入する。

図５(a)～(d)にニューラルネットワークモデルによる予測モデル開発の結果を示す。

(a)ケース１の GPA 予測モデル開発　(b)ケース２の GPA 予測モデル開発

(c)ケース１の SAT 予測モデル開発　(d)ケース２の SAT 予測モデル開発

図５(a)(b)によれば、GPA 予測モデルにおいて、決定係数は（ケース１，ケース２）＝（0.692，0.831）となり、ケース２の方が高くなっている。この結果は、ケース２のクラスター情報の方が多いことを考えると予想通りといえる。一方、線形重回帰モデル（ケース２）の決定係数が 0.279 であることと比較すると約 55%の予測精度の向上が見られ、ニューラルネットワークモデルの予測率がかなり高いことを示している。しかし、その要因については、さらなる検討が必要である。

次に**図５**(c)(d)によれば、SAT 予測モデルにおいて、決定係数は（ケース１，ケース２）＝（0.267，0.479）となり、やはりケース２の方が高くなっている。一方、線形重回帰モデル（ケース２）の決定係数が 0.305 であることと比較すると約 17%の予測精度の向上に留まった。GPA 予測モデルと同様な変数を投入しながら、ニューラルネットワークモデルでも予測精度があまり向上しなかった要因について、今後の検討が必要である。

４．まとめと今後の課題

本研究では、LA データ、性格特性、教学 IR データを活用し、１年春学期末の GPA と学校生活満足度を予測するモデルを開発した。モデル開発手法として、k-means 法によるクラスター分析、線形重回帰分析及びニューラルネットワークを採用した。予測精度を計算した結果、線形重回帰モデルの最大決定係数は、（GPA，学校生活満足度）＝（0.279，0.305）であり、ニューラルネットワークモデルの最大決定係数は、（GPA，学校生活満足度）＝（0.831，0.479）となった。

本研究のようにクラスター分析を前段処理として活用することで、直接投入する説明変数の数を削減することができる。変数削減の代表的な手段としては「主成分分析」があるが、本研究のように「k-means 法のクラスター分析による各クラスター中心からの距離」という方法もあることを示した。この方法の利点は、投入変数がいくら増えたとしても、分類されたクラスターは「パターン情報」という一次元に集約されることである。さらにニューラルネットワークモデルでは、一般的に投入する説明変数が多くなるほど予測精度は高くなる。しかし、その一方で計算過程はブラックボックスになりがちである。この点、本研究では、説明変数の意味解釈が容易なクラスター分析（教師なし機械学習）を前段処理に活用し、その後、線形重回帰モデル（教師あり機械学習）をもって予測モデルの概略をつかみ、その後、予測精度を向上させる目的でニューラルネットワークモデル（教師あり機械学習）を活用するモデルを開発したことになる。

今後、特に SAT 予測モデルの精度を向上させるために新たな説明変数を探索する必要がある。そのため、SAT の線形重回帰モデルにおいて、予測が大きく外れている学生たちを対象に半構造化面接やアンケート調査等を行い、予測精度の向上とその予測による早期の学生対応方法を模索していきたい。

謝辞

本研究は JSPS 科研費 JP18K02882 の助成を受けた。また、データ収集・提供に際しては、本校の学生や教務学生部職員の協力を受けた。

参考文献

船守美穂（2014）『デジタル技術は高等教育のマス化問題を救えるか？－MOOCs，教育のビッグデータ，教学 IR の模索』，情報知識学会誌，24(4)，pp.424-436.

片瀬拓弥（2017）『教学 IR のための学生支援モデルの試作－学習スタイルと性格特性のクラスタリングを活用して－』, 日本教育工学会研究会報告集 リフレクション活動の支援/インストラクショナルデザイン/一般（JSET17-4），pp.159-166.

河村茂雄（2010）『hyper-QU（大学版）』，図書文化社，東京

ラインズドリルベーシックコース（2018）『清泉女学院 SJC ラーニング ベーシックコース』，ラインズ株式会社，https://lines-drill.education.ne.jp/seisen-jc/basic/PC/ (Accessed 2019.01.04)

LAK'11（2011）『About the 1st International Conference on Learning Analytics and Knowledge 2011』，https://tekri.athabascau.ca/analytics/ (Accessed 2019.01.04)

MATLAB（2018）『Neural Network Toolbox 入門ガイド(R2018a)』，Math Works，Inc.，https://jp.mathworks.com/help/pdf_doc/deeplearning/index.html (Accessed 2019.01.04)

松田岳士，渡辺雄貴（2017）『教学 IR，ラーニング・アナリティクス，教育工学』，日本教育工学会論文誌，41(3)，pp.199-208.

みんなの高校情報 HP（2018）『長野県高校偏差値一覧 2018 年度版』，株式会社イトクロ，https://www.minkou.jp/hischool/exam/nagano/deviation/ (Accessed 2019.01.04)

旺文社教学支援サービス（2018）『基礎学力判定テスト（プレースメントテスト）』，株式会社旺文社，https://www.obunsha.co.jp/06/cramschool/AchievementSystem/test.html (Accessed 2019.01.04)

下仲順子，中里克治，権藤恭之，高山緑（2011）『日本語版 NEO-PI-R, NEO-FFI 使用マニュアル改訂増補版』，東京心理，東京

高橋哲也，星野聡孝，溝上慎一（2014）『学生調査と e ポートフォリオならびに成績情報の分析について －大阪府立大学の教学 IR 実践から』，京都大学高等教育研究，20，pp.1-15.

SUMMARY

In this research, we developed a model that predicts school life satisfaction and GPA in the spring semester of 1st year by using Learning Analytics, personality traits, data of Institutional Research for education. However, as a model development method, cluster analysis by k-means method, linear multiple regression model and neural network model was adopted. As the result of calculating the prediction accuracy of the model, the maximum decision coefficient of the linear multiple regression model is (GPA, school life satisfaction) = (0.279, 0.305), and the maximum decision coefficient of neural network model is (GPA, school life satisfaction) = (0.831, 0.479).