高校生向けデータサイエンス教材の開発

村上綾菜[1,a)]　伊藤 貴之[1,b)]

概要：政府が AI 戦略を掲げ，高校生の段階から AI やデータサイエンスの基礎を学ぶよう，提言されている．しかし，既存の学習教材は，文章の説明のみのものや，プログラミングの知識を必要とするものが多く，初学者にとって難解である．本報告では，内容は高校生レベルのデータサイエンスを扱いながらも，インタラクティブで楽しい Web アプリケーション教材を提案する．学習内容は判別分析を取り上げ，生徒は判別分析のためのデータクレンジング作業を通じて判別分析の仕組みを理解する．本報告では，本教材を学部 2 年生および学部 3 年生の授業において使用した際の使用ログを収集し，解析した．さらに，本教材使用後にアンケートを実施し，本教材を使用した学生の理解度をはかる．

1. はじめに

技術のデジタル化が進み，デジタル機器が身近になるにつれて，AI やデータサイエンスをはじめとする情報科学技術の重要性が急激に注目されている．この傾向は教育分野においても同様で，政府 [1] は全ての高等学校卒業生が「数理・データサイエンス・AI」に関する基礎的なリテラシーを習得することを今後の教育の目標の一つとして掲げている．

高等学校においてデータサイエンスおよび AI に関する教育を推進するにあたり，高校生に適した学習内容の教材が必要である．そして学習の手段についても注意しなければならない．森山ら [2] の研究によると，ICT に対する苦手意識が情報活用の実践力を身に付けたいという気持ちを減衰させてしまい，情報科の学習に対して自己効力感が十分に高まらないことが，高校生を対象としたアンケートにより確認されている．特に，コンピュータを使用することに不慣れな生徒が一定数いることを踏まえると，データサイエンスの教育においても高度な ICT スキルを前提としない教材の開発が求められる．

本報告では，高校生を対象としたインタラクティブなデータサイエンス教材の一事例を提案する．本学習ツールでは，学習内容の一例として判別分析を採用しており，生徒は訓練データのクレンジング作業を通じて判別分析の仕組みを理解することを期待している．データクレンジングとは削除や修正などの操作によってデータの品質を高める工程であり，本教材では例外的なデータ要素の削除作業を通してデータの品質向上を体験してもらう．

我々は，本教材の学習目標を以下の 3 点とした．ここで定めた学習目標の実現が本教材の大きな特徴ともいえる．

目標 1: 生徒が高校生の学力の範囲内で判別分析の概念を理解することができる．

目標 2: プログラミングを知らなくても，また複雑な操作を覚えなくても，生徒がオンライン教材を使いこなせる．

目標 3: 生徒が楽しくデータサイエンスを学習することができる．

我々は，この学習目標の達成を目標として本教材を開発する．同時に，高校生が使用することを鑑み，本ツールには主に 4 点の工夫を施すこととした．

1 点目として，**目標 1** のために，データの視覚表現を 2 次元の散布図に限定する．これには，高校生が図を理解しやすいようにする目的がある．学習指導要領 [3] によると，高校生は数学 I データの分析の章で「散布図や相関係数を用いて 2 つのデータの相関を把握し説明すること」を学ぶ，とある．つまり，高校生は 2 次元の散布図を用いてデータを分析することには慣れていると考えられる．一方で高校生は，変数が 3 つ以上のグラフは見慣れていないので，理解するのに時間がかかり，負担になりやすい．少ない変数で分析することで，判別分析の精度に限界が生じるデメリットと，データが読み取りやすく傾向もつかみやすいメリットを比較し，今回は後者を重要視し使用するデータの変数を 2 つに制限した．

2 点目として，**目標 1** のために，生徒の操作に対して即時のフィードバックを与える．具体的には，散布図上の点を削除するたびに，散布図上で判別分析の境界線が動く．

[1] お茶の水女子大学
[a)] g1720541@is.ocha.ac.jp
[b)] itot@is.ocha.ac.jp

この動きにより，作業のたびに自分の動作が正しいかどうかを確認できるので，試行錯誤して課題を遂行することが期待できる．

3 点目として，**目標 2** のために，全ての操作を GUI ベースにするとともに，GUI 部品を必要最低限にし，クリック操作のみでほぼ全ての作業を行えるようにする．GUI の工夫により，生徒にとって負担になる可能性があるキーボード入力を排除する．これには，学習内容以外の負担を減らすことで学習内容へ集中を促す目的がある．

4 点目として，**目標 3** のために，高校生にとって身近なデータを題材とする．今回は J-POP 等の楽曲を題材とした．データの中に自分の知っている曲を見つけることで，データが身近であることに気がつき，同時にデータサイエンスも身近であることを理解することを期待する．

我々は本教材を，お茶の水女子大学附属高等学校の 1 年生の情報科の授業にて実習題材として使用した．また，お茶の水女子大学理学部情報科学科の学部 2.3 年生にも，本教材の被験者としてユーザ実験に参加してもらった． 本報告では，ユーザ実験に参加した学部生の使用ログを解析することで，本教材の操作の傾向についても議論する．

2. 関連研究

データサイエンス手法の学習を目的とした既存の学習教材は大きく分けて 2 種類に分類できる．1 種類目は，Web サイトやアプリケーションを利用したインタラクティブな学習ツールである．具体的には，科学の道具箱 [4] や Bowland Japan が公開している教材 [5] が挙げられる．これらの教材の利点は，パソコンの操作を苦手とする生徒でも学習に取り組むことができ，ゲーム感覚で楽しく学ぶことができる点である．一方で，現在このようなインタラクティブな教材で高校生を対象とした事例はまだ少なく，充実しているとはいいがたい．

2 種類目として，分析ソフトやプログラミング言語を用いた本格的なデータ分析環境があげられる．文部科学省が提示する高等学校情報科「情報 I」教員研修用教材 [6] においては Excel を活用した単回帰分析が紹介されており，高等学校情報科「情報 II」教員研修用教材 [7] においては Python や R を用いた分析方法の例がソースコードとともに紹介されている．神部ら [8] の研究では，R を用いた統計教育への問題解決型教材を採用しているが，機器の操作方法の習得へ意識が奪われている学生がかなり見受けられたと報告されている．これらの教材の課題は，プログラミング初学者やプログラミングを苦手とする生徒には苦手意識が先行してしまい，データサイエンス自体も難解に感じる可能性があることである．その一方で，実践レベルの専門的な分析手法が学べるという利点があり，一定以上の学力と PC スキルを有する高校生の学習レベルには適している面がある．

3. データサイエンス教材の開発

本章では我々が開発したデータサイエンス教材の概要と詳細について説明する．我々は R の Web アプリケーション作成パッケージである shiny を用いて本教材を実装した．本教材を Web アプリケーションにて実装した理由は，本教材使用者が事前のインストール作業等の煩雑な環境構築をすることなく使用できるようにするためである． 我々が開発した Web アプリケーション教材のメイン画面のスナップショットを図 1 に示す．この教材において，画面左側には，判別分析に使用するデータを散布図で表示する．画面右側には，ユーザが使用する各種機能に関するボタン，および判別分析の精度表示のための GUI 部品が搭載されている．

図 1 教材メイン画面のスナップショット．

3.1 使用したデータ

本教材の判別分析に使用したデータは，600 曲以上の J-POP の楽曲の音響特徴量である．具体的には，Librosa[9] を用いて楽曲の音響信号から音響特徴量を算出し，これを説明変数とした．我々は，高校生がデータを理解しやすいように，音響特徴量に対して次元削減を適用して 2 次元に圧縮し，それを座標値にして散布図として表示した．次元削減には UMAP を使用した．判別のラベルは楽曲の発売年を適用し，手作業で付与した．今回は楽曲の発売年を「1999 年以前」と「2000 年以降」で 2 種類に分類してラベルとした．散布図上では，青い点が 1999 年以前に発売された楽曲，赤い点が 2000 年以降に発売された楽曲を示す．

3.2 学習教材

本教材においてユーザに課する作業は「散布図上の点をクリックすることによる例外的なデータの削除」であり，その結果としてユーザに求める目標を「分類精度の向上」であるとした．この作業をユーザが円滑に遂行できるように，本教材では以下の機能を実装する．

- 分類精度の確認

- 判別の境界線の確認
- 散布図上のデータの詳細情報の確認
- 削除したデータの一覧表示の確認
- テストデータプロット位置の確認

ユーザは上記の機能を適切に使用しながら，判別分析の精度を向上するためにどの点を削除すべきかを考えるものとする．本教材は以下の 3 つの画面で構成されている．

3.2.1 メイン画面

図 1 に示したメイン画面は，ユーザがデータクレンジング作業を遂行するための画面である．この画面の散布図上で，ユーザは例外と思われる点をクリック操作によって削除する．散布図には判別の境界線が直線で表示されている．点を削除するたびに本教材では判別分析を再度実行し，その結果として散布図上の境界線が変動する．マウスのカーソルを散布図上の点に重ねると，その点に対応する楽曲の詳細情報が散布図の下に表示される．ユーザが例外的な点を削除する際には，散布図上のプロット位置からだけでなく，楽曲の詳細情報も同時に読んで総合的に判断することもできる．

判別精度を評価するために，我々は，楽曲データ全体をあらかじめ訓練データとテストデータに分割した．本教材では，判別分析による境界線決定のための学習に訓練データを使用し，判別精度の検証のためにテストデータを使用する．以下，テストデータに付与されたラベルを正解ラベル，判別分析によってされたラベルを予測ラベルと称する．テストデータは，予測の際に正解ラベルを隠しておき，判別後に正解ラベルと予測ラベルを照合することで，判別精度を算出する．言い換えれば，正解ラベルと予測ラベルが一致する楽曲は，正しい判別がなされた楽曲とみなすことができる．本教材において，判別精度は画面右側に表形式で表示される．そのスナップショットを図 2 に示す．

図 2 教材メイン画面における精度表示のスナップショット．

表で最も右の列の「件数」は，表の各項目に該当するテストデータの楽曲数を表示する．表形式で精度を表示した理由は主に 2 つある．

1 つ目の理由は，精度の算出について多様な方法を認めることで教員の自由な授業展開を実現するためである．「精度」の算出方法は様々で，実際には分析目的に合わせて適切な方法を選択する．

単純に正答率を精度とするのであれば，正解ラベルと予測ラベルが一致した件数を全てのテストデータの件数で割ることで算出できる．また，表から再現率・適合率を算出する形で精度を求めても良い．これら 3 種類の求め方以外にも，表から該当件数を数えることで，簡易的に分類精度の向上を確認することもできるだろう．このように，表形式で表示することで多様な精度算出方法の学習につながり，生徒のレベルに合わせて使い分けることができる点で，ただ単に計算後の精度を数値のみで表示するよりも効果的だと考えられる．

2 つ目の理由は，精度の計算方法を生徒が学習する機会を与えるためである．例えば数値のみで「精度 86%」と表示してしまうと，データや散布図と精度の関係を生徒が考える機会を奪う可能性がある．件数を表形式で示すことで，パーセント表示の精度を求めるには自分で計算しなければならないので，データ件数と精度の関係を学ぶ機会を与えることができる．

3.2.2 テストデータ確認画面

図 3 に示すテストデータ確認画面で，ユーザはテストデータのプロット位置と境界線の位置関係を確認することができる．その画面のスナップショットを図 3 に示す．これにより，ユーザは図 2 の表に表示されているデータ件数が示す該当テストデータが，散布図上のどの点に対応するのかを確認できる．また，メイン画面と同様に，散布図上の点にマウスオーバすることで，その点に対応する楽曲の詳細情報を読むことができる．

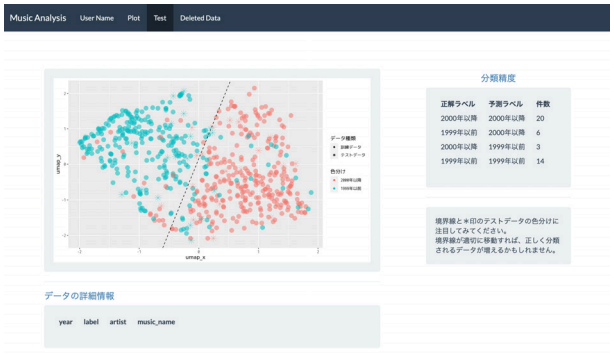

図 3 教材テストデータ表示画面のスナップショット．

テストデータ確認画面における散布図の一例を図 4 に示す．ここでテストデータの配色は，テストデータを構成する各点の正解ラベルにもとづく．散布図中央に表示されている点線は，判別の境界線を表す．

ここで図 4 の中心部を参照されたい．前述の通り，散布図上にて青色で描画された点（以下「青点」と表記）は，

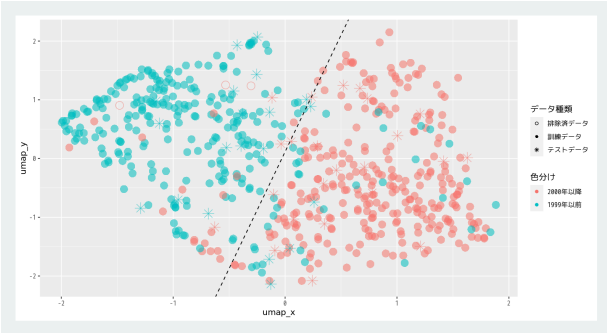

図 4　教材テストデータ表示画面の散布図を拡大表示した図.

1999 年以前に発売された楽曲に対応する．一方で，散布図上にて赤色で描画された点（以下「赤点」と表記）は，2000 年以降に発売された楽曲に対応する．図 4 には境界線の左側に青点が多く位置するが，赤点もいくつか存在することがわかる．これらの点は，本来なら赤点であるにもかかわらず境界線の左側に位置することから，青点であると予測されるであろう．この点が図 2 における分類精度の表の中で，正解ラベルが 2000 年以降（赤），予測ラベルが 1999 年以前（青）に該当する 3 件のうちの 1 つあることがわかる．

この画面においてテストデータのプロット位置を確認することが，分類精度の向上のためにどの点を削除するかを考察するためのヒントになる．

3.2.3　削除データの一覧表示画面

削除した点の一覧表示画面のスナップショットを図 5 に示す．

year	label	artist	sex	music_name	umap_x	umap_y
1988	before	美空ひばり	f	みだれ髪	-1.88	0.30
2003	after	氷川きよし	m	白雲の城	-1.09	-0.90
2018	after	BTS (防弾少年団)	m	FAKE LOVE	-0.53	1.25
2015	after	ジャスティン・ビーバー	m	What Do You Mean?	-1.48	0.91
2014	after	One Direction	m	Story Of My Life	-1.28	0.36
2009	after	U2	m	Get On Your Boots	-1.71	0.63
2017	after	アリアナ・グランデ、ジョン・レジェンド	n	美女と野獣	-0.92	-0.53
2011	after	レディー・ガガ	f	The Edge Of Glory	-1.93	0.19

図 5　教材削除データ表示画面のスナップショット.

この画面では，ユーザが削除した点の詳細情報が一覧表示される．これを適切に使用することで，削除したデータの傾向を理解し報告する，という発展課題を設定することも考えられる．

4.　操作ログの解析

本報告では，お茶の水女子大学理学部情報科学科の学部 2,3 年生合計 26 名を被験者として本ツールを用いて学習してもらい，その操作ログを記録した．操作ログでは，被験者が所定の操作を行うごとに，その時刻と操作内容を記録する．記録内容は以下の項目である．

- 散布図再描画のタイミング
- テストデータの閲覧タイミング
- 削除データ一覧表示のタイミング
- マウスホバーで参照したデータ
- 削除したデータ及び削除後の判別精度

図 6 に示す可視化結果では，学生被験者群が削除した楽曲を表示しており，学生が楽曲を削除した順に濃い色が割り当てられている．具体的には，各学生の作業イベントごとに，本教材使用開始時刻と削除時刻との時間差を計算し，その差分が大きい点ほど濃い色になるよう割り当てる．削除した点の座標値に差分時間を考慮した濃さの点をプロットする処理を，全学生に対して適用した．散布図のプロット位置は，本ツール上の散布図と同じである．配色は本教材と同様に，1999 年以前に発売された楽曲に青，2000 年以降に発売された楽曲に赤を割り当てている．

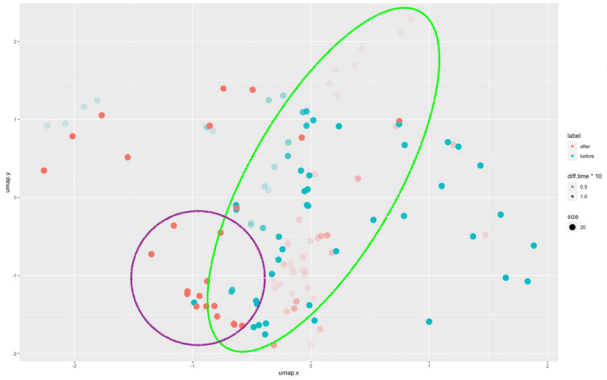

図 6　可視化結果 (1)．プロット位置は本教材上の散布図と一致する．色の濃さは削除されたタイミングを示しており，色の濃い点ほど早期に削除された点であることを示す.

ここで，散布図中央の緑の楕円形の枠に注目すると，色の濃い点と薄い点が混在している．この緑の枠は，学生が観察する散布図における判別の境界線付近に対応する．学生は，いくつかの点を早い段階で削除したものの，境界線付近の多くの点の削除を後回しにしており，削除操作に対して慎重になっていたことがわかる．一方で，緑の枠の外側では，色の薄い点はあまり見られない．緑の枠の外側は，境界線から離れた位置である．例えば，図 1 内の散布図と照合すると，緑の枠の左側は青い点が大半を占めるエリアである．学生は，青い点が集中するエリアに赤い点が混在し，かつそれが境界線から離れた位置であれば，早い段階でその赤い点を削除していたことがわかる．このことから，学生が外れ値の概念を理解して本教材に取り組んでいることがわかる．

続いて，図 6 の左下の紫色の円形の枠を参照されたい．ここは，境界線付近であるのにもかかわらず，早い段階で点が削除されていることがわかる．紫の枠内の点には後述

する独特な傾向があることがわかっている．ここから学生は，この傾向を比較的早く発見し，このエリアの点を早く削除した学生が一定数いたことが予想される．

図 7 に示す散布図では，本教材で使用したデータのうち氷川きよしの楽曲のみを黄色でハイライトする．プロット位置は本教材上の散布図と一致する．図 6 と図 7 を比較されたい．図 6 の紫の枠で囲まれたエリアと，図 7 が示す氷川きよしの楽曲のプロット位置がほぼ一致することが読み取れる．

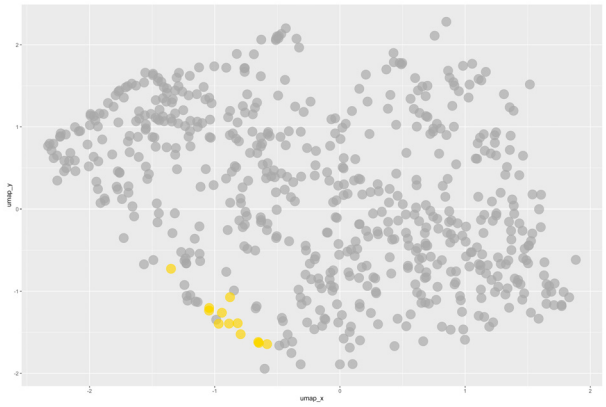

図 7 可視化結果 (2)．使用データのうち氷川きよしの楽曲を黄色でハイライトした．

ここから，図 6 に示す散布図にて，紫色の枠周辺には氷川きよしの楽曲が集中していたことがわかる．このことが，図 7 の紫の枠内では境界線が近いにも関わらず早い段階で削除された点が多い，という前述した現象の一理由にもなるだろう．また，この考察から学生被験者の中には，単に散布図の点の分布を観察するだけでなく，楽曲情報を読んで歌手ごとの傾向を推測しながら効率的にデータクレンジング作業を進めた学生が一定数存在した，ということが示唆される．

5. ユーザテストにおける評価

我々は，お茶の水女子大学学部生 51 名を対象にユーザテストを実施した．対象学生は学部 2 年生および学部 3 年生である．当大学にはデータサイエンスの選択科目が新規開講されたばかりで，判別分析の実習に関しては大半の学生が初学者である．

このユーザテストでは，本教材の使用前に，概要説明と操作説明をそれぞれ 10 分ほど実施した上で，学生参加者に本教材を使用してもらった．そして使用後にアンケートを実施した．

アンケートの結果を図 8 に示す．図 8 帯グラフの上段は，「本教材のやるべきことと目標を理解して学習に取り組めたか」という設問に対する回答の集計結果を示す．下段は，「本ツールで使用する音楽データを理解できたか」というという設問に対する回答の集計結果を示す．

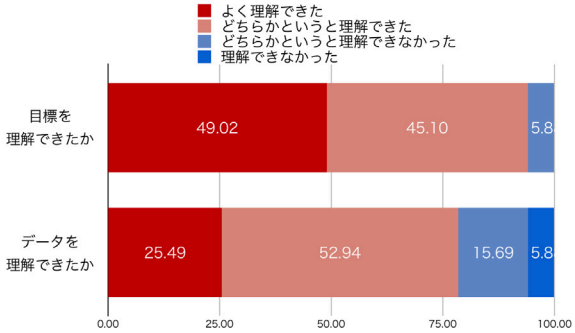

図 8 アンケート結果．学生の理解度を図る設問の回答結果を示す．

5.1 本教材に対する理解度

図 8 の帯グラフの上段に注目されたい．9 割以上の学生が，本教材の目標を理解して作業に取り組めたことが読み取れる．この理由として，学生が本ツールを使用する前に実施した合計 20 分ほどの丁寧な説明時間が効果的だったことが考えられる．ここから，学習効果を高めるには，教員がある程度の時間をかけて説明をした上で，演習として本教材を使用する，という流れが適すると考えられる．もう一つの理由として，GUI が簡潔であることに加えて．本教材の内部にも操作説明ウィジェットの表示を開始するボタンがあり，すぐに操作を確認できる仕組みがあったことがあげられる．アンケートの自由記述欄に「操作方法の確認機能がとてもわかりやすかった」との記述があった．ここから，操作説明ウィジェットが教材の内部にあることが，事前説明だけでは理解しきれなかった学生にとって有用であったと考えられる．

5.2 使用データに対する理解度

図 8 の帯グラフの下段に注目されたい．学生の 7 割以上が本ツールで使用する楽曲データの内容を大方理解していることがわかる．一方で，図 8 全体を見ると，使用データに対する理解度が全体的に低いことがわかる．特に，上段と異なり，下段には「理解できなかった」を選択した学生が 6%いる．

今回のユーザテストでの事前説明では，判別分析の説明に重点を置いたおり，楽曲データの生成方法についてはほとんど触れなかった．そのため，どのように楽曲から使用データが作られたのかに疑問を抱く学生が何人かいたことが予測される．

5.3 全体を通じて

自由記述欄に「どのように x 軸 y 軸を決めたのか，とても気になった」との回答があった．こうした興味は教育現場において，前後の授業との適切かつ自然なつながりを生み出すことが期待される．例えば，次元削減の学習をした後に，今回の判別分析の学習へ入る場合を考える．この時に，楽曲から抽出した音響特徴量を 2 次元に削減すること

を説明すると，学生もどんなところで次元削減が使われているかを知ることができる．もう一つの例として，判別分析の学習の後に次元削減の学習をする場合を考える．この場合，生徒は楽曲データの生成方法に興味があるので次元削減の学習も意欲的に取り組むことができると考えらえる．このように，どのような場面および目的で次元削減を使用するかを理解した上で次元削減の学習できるため，学習効果が高まることが期待される．

6. まとめ・今後の課題

本報告では，高校生を主対象として，判別分析を例題としたオンライン型のデータサイエンス教材を提案した．我々は，データサイエンス初学者に本教材を使用してもらい，その使用ログを解析した．この結果，本教材の使用者は，外れ値の概念を理解し，判別分析の境界線から離れた点を優先的に削除することを発見した．また，一部の学生参加者は，削除した点に対応する楽曲の傾向を読み取り，それを判別分析の作業に反映できていることを発見した．

今後の課題として，ユーザ全体を分類精度の向上を効率的に達成したグループとそうでないグループに分類し，その違いに着目したい．Boroujeni・Dillenbourg[10] の研究では，ページ遷移をマルコフモデルで表現した後，類似した手順でクラスタリングし各クラスターの学習特徴を分析する．本教材においても，3 つのページが存在するので，そのページ推移との関係にも着目し，さらなる解析を進めたい．

もう一つの課題として，判別分析以外のデータサイエンスの題材もサポートした Web アプリケーションとしての開発を進めたい．高等学校情報科の指導要領 [11] に記載されている学習内容のうち，共通科目である情報 I の「情報通信ネットワークとデータの活用」においては，データの収集整理に適切な方法を選択し，実行・評価・改善する能力を身につけることが期待されている．ここでは，アンケート等の統計的分析・可視化に止まらず，テキストマイニングや単回帰分析等についても記載されている．さらに，選択科目である情報 II においては，重回帰分析・分類・機械学習といった専門的な分析手法についても触れられている．このような時代の潮流の中で，本報告のようなインタラクティブなデータサイエンス教材の需要は高いと考えられる．本報告で取り上げた判別分析にとどまらず，単回帰分析などの他の理論を題材とした教材の開発，およびその精度をあげるためのデータクレンジングなどの各作業のサポートと解析を進めたい．

謝辞 音楽データをご提供くださった株式会社レコチョク社の関係者の皆様に感謝の意を表します．
本教材を授業で使用してくださったお茶の水女子大学附属高等学校の先生方に感謝の意を表します．
本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．

参考文献

[1] 首相官邸ホームページ,“AI 戦略 2019 人・産業・地域・政府全てに AI ”, https://www.kantei.go.jp/jp/singi/ai_senryaku/pdf/aistratagy2019.pdf，2019.
[2] 森山, 原田, 福井, 中尾, 小倉, 近澤, 山下,” 高校生の ICT に対する苦手意識と情報活用実践力および自己効力感との関連性”，兵庫教育大学 研究紀要, 第 57 巻, pp. 65-75, 2020.
[3] 文部科学省，“高等学校学習指導要領解説 数学編”，pp. 24-26，2009
[4] 科学の道具箱，https://riaska-net.com/contents/cp0530/contents/07.html#.
[5] Bowland Japan,“ボーランド・ジャパンの教材紹介”, https://bowlandjapan.org/materials_jp.
[6] 文部科学省，“高等学校情報科「情報 I」教員研修用教材（本編）”， 第 4 章 情報通信ネットワークとデータの活用，pp. 184-191，2019.
[7] 文部科学省，“高等学校情報科「情報 II」教員研修用教材（本編）”，第 3 章 情報とデータサイエンス（前半），pp. 124-143，2019.
[8] 神部，玉田,” プログラミング活用による統計教育への問題解決型教材の開発に向けて”，Informatio：江戸川大学の情報教育と環境, pp. 29-32，2020.
[9] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, “librosa: Audio and Music Signal Analysis in Python”, Proc. of the 14th Python in Science Conf.(SCIPY), pp. 18-24, 2015.
[10] Mina Shirvani Boroujeni, Pierre Dillenbourg, “Discovery and Temporal Analysis of Latent Study Patterns in MOOC Interaction Sequences”, LAK ’18: International Conference on Learning Analytics and Knowledge, pp. 206-215, 2018.
[11] 文部科学省，“高等学校学習指導要領解説 情報編”，pp. 35-40， 2018.