データインフォームド型データサイエンス教材開発の試み

Development of Data Science Education Materials for Data-Informed Learning

章 志華* 山本 敏幸*
Zhihua ZHANG Toshiyuki YAMAMOTO

抄 録

データサイエンス教育において，データ分析の過程で直面する課題や結果の解釈・活用について十分に理解し学べる教材の開発が求められている。本研究では，オーセンティックな学びを促進するデータインフォームド型データサイエンス教育教材を提案する。教材はデータに基づく問題解決のプロセスを重視し，受講生がデータと共に思考する能力を養うことを目的とする。特に，プログラミング苦手意識が高い文系学生に対して，ノンプログラミング環境でのデータアナリティクスやインタラクティブなビジュアライゼーションコンテンツで学習することが必須である。

1. はじめに

近年，データサイエンスはビジネス，政策，社会サービスに至るまで多くの分野で重要な役割を果たしている。これに伴い関連する人材の需要が高まっており，データサイエンス教育も急速に拡大している。STEM（科学，技術，工学，数学）分野だけでなく，文系の学生に対してもその重要性が認識され始めている。文部科学省は，この状況に対応するために「数理・データサイエンス・AI 教育プログラム認定制度」を全国向け推進しており，これを受け，各大学は独自の取り組みを通じて，データサイエンス分野の学部・学科の強化を進めている。本学のような私立文系の大学においても，2022 年度より「データサイエンス副専攻」を設置し，文部科学省のこの認定制度に適合したカリキュラムを構築した上で，全学向けのデータサイエンス教育プログラムを実施している。このプログラムは 2023 年 8 月に文部科学省の「数理・データサイエンス・AI 教育プログラム認定制度（リテラシーレベル)」に認定された。また，筆者らが所属する社会学部では，2021 年度よりデータサイエンス専攻を設置し，「高度データサイエンス人材育成プログラム」を開講した。本プログラムでは，社会や企業におけるデータに基づく意思決定や問題解決に必要な実践的な能力と，データサイエンスの応用知識および関連技術を体系的に習得できる人材育成を目指している。この「高度データサイエンス人材育成プログラム」も 2023 年 8 月に文部科学省の「数理・データサイエンス・AI 教育プログラム認定制度（応用基礎レベル)」に認定された。

データサイエンスは，統計学や機械学習などの複数の分野を統合した学問分野であり，データを中

＊ 関西国際大学社会学部 教育総合研究所学内研究員

心に置いて価値を創造する方法論である。しかし，その学際性やデータ中心主義の特性は，教育の現場では抽象的で理解しにくいという課題がある。データサイエンスを教育する課程を設計する際には，各大学が統計学手法や機械学習手法の基礎に，プログラミング技法や数学が注目されているが，学生の多様性，特に文系の学生にはハードルが高い。多くの文系学生は，数学やプログラミングといったデータサイエンスの基礎技術に対して苦手意識を持っている。この点において，筆者らは「データサイエンス実践演習」や「人工知能の基礎」などの科目の教育実践で実感している。例えば，Python プログラミングを用いたデータ分析では，しばしば思う通りに授業展開ができない。コーディング作業においてプログラミング技法やアルゴリズムの基礎への理解，そして問題解決背後に関わる統計学や数学知識の説明などに多くの時間が取られ，授業趣旨通りの展開がスムーズにできない。

また，従来のデータサイエンス教育では，統計学や機械学習などの理論やアルゴリズムを重点的に学習する方法で行うことが多く，データ分析の過程で直面する様々な課題や，データ分析結果の解釈や活用方法などについて十分に学ぶことができない。しかし，文系学生の今後のキャリアを考えれば，むしろ後者の方が大切である。特に，データに基づいて意思決定や問題解決の場合，ビジネス領域に関する基本的な理解とその領域の知識のほかに，データを扱うための基本的なプログラミング技術やデータ分析ソフトウェアツールの利活用などに関わるコンピュータ知識や IT スキルも必要である。また，データを収集・分析・解釈することで，現状を理解したり，未来を予測したりする統計学の知識も欠かせない。このように，データサイエンスの学習では，学際的な知識や技能を統合的に運用し，データ分析スキルとデータ活用能力を実際のデータセットを用いて課題解決に取り組む能力が必要である。そのため，学生にデータサイエンスの問題解決のプロセスを理解してもらうことが極めて重要である。このような考え方を持つ教育コンテンツを本研究では，データインフォームド型（Data-Informed）データサイエンス教材と呼ぶ。

一方で，ノンプログラミングで学生がデータサイエンスの概念をよりよく理解し，実際のデータを使用してデータサイエンスのスキルを学習するのに役に立つ環境はないのか，あるいは，プログラミングスキルがそれほどなくても，グラフィカルユーザーインターフェース（GUI）を通じてデータの管理・加工，探索的データ分析，結果の共有と伝達を行うことができるスキルを身につければ，今後のキャリアではきっと役に立つ教材の開発はできないのか。こうした発想から，本研究は Tableau や Exploratory のような BI 系よりも優れたデータ分析ツールに注目した。Tableau や Exploratory は，直感的に操作でき，データの前処理，データアナリシス，データの可視化，そしてダッシュボードなどを通して分析結果を共有し，活用方法をインタラクティブ的に学ぶことができる。これらのツールを用いて，実データセットで学べば，受講生が容易にデータインフォームドの考え方や知識を習得できる。

以上の議論を踏まえると，本研究では，データインフォームド型データサイエンス教育教材の開発方法を提案する。この方法で開発した教材は，データに基づく問題解決のプロセスを重視し，学生がデータと共に思考する能力を養うことを目的とする。特に，プログラミング苦手意識が高い文系学生に対して，ノンプログラミング環境でのデータアナリティクスやインタラクティブなビジュアライゼ

ーションコンテンツで学習することが可能である。しかし，このようなデータインフォームド型データサイエンス教育アプローチを用いた文系私大生向けの教材が見当たらない。そこで本研究は，文系大学生を対象に, ノンプログラミング環境を用いたデータサイエンス教育を行う教材の開発を試みる。

2. データサイエンスのプロセスと教材の設計方針

データサイエンスはビッグデータ社会において, 解決課題の定義, データの取得, データの前処理, データの蓄積，データマネジメント，データ処理，探索的データ分析，データの推論，データの可視化および結果の共有などに用いられる新しいアプローチとなっている。社会のさまざまなデータから何らかの価値を生み出す際に，特にデータインフォームド型問題解決のアプローチはデータに基づく意思決定のプロセスにおける科学的方法論の一つである。

これまでデータインフォームド型問題解決のプロセスとして，いくつかの展開するフレームが提案されている。例えば，問題解決に向けたデータに基づく知識発展のサイクルとしてのデータサイエンスのサイクルや, データマイニングのためのフレームワークとして CRISP-DM（Cross-Industry Standard Process for Data Mining）などがある。データサイエンス教育では，学生達にこうしたデータサイエンスのプロセスを理解させ，さらに実例的なデータセットを用いて分析のスキルなどを訓練することが重要である。またシンプルな UI（User Interface）を持ち，直観的な操作でデータを扱うことが可能なデータ分析ツール Exploratory は，簡潔なデータサイエンスのワークフローを実現している。今回の教材開発において，この三つのプロセスモデルを参考に，研究を進めている。以下の節でそれぞれについて簡単に説明する。

2.1 データサイエンスのサイクル

Richard D. De Veaux, et al. は「Curriculum Guidelines for Undergraduate Programs in Data Science」（2017）[1)] で指摘されているように，データサイエンスにおいては，図 1 に示す「課題抽出と定式化」，「データの取得・管理・加工」，「探索的データ解析」，「データ解析と推論」，「結果の共有・伝達，課題解決に向けた提案」というプロセスがあり，問題解決とともにデータからの知識発見や新たな課題発見を可能にするために，知識発見のサイクルが生じる。また，データサイエンス教育先進国であるアメリカ合衆国の The National Academies は，データサイエンスの学部教育に関する報告書において，「データ洞察力」を構成する 10 分野の知識領域を指摘している。その 10 分野とは，「データの記述・可視化」，「データの取得・管理・加工」，「統計基礎」，「数学基礎」，「計算基礎」，「モデリングと評価」，「ドメイン知識の考慮」，「コミュニケーションとチームワーク」，「倫理に配慮した課題解決」，「ワークフローと再現性」である[2)]。数理・データサイエンス教育強化拠点コンソーシアム・カリキュラム分科会（2019）の「データサイエンス教育に関するスキルセットおよび学修目標」-第 1 次報告（リテラシーレベル）[3)] では，図 1 に示すサイクルの中で各要素が位置付けられている。ただし，「倫理に配慮した課題解決」はすべてのプロセスにおいて重要であり，「ワークフローと再現性」はサイクルを持続させるための作業手順の保持と再現性であり，各ステップに割り当

てられるわけではないと指摘されている。

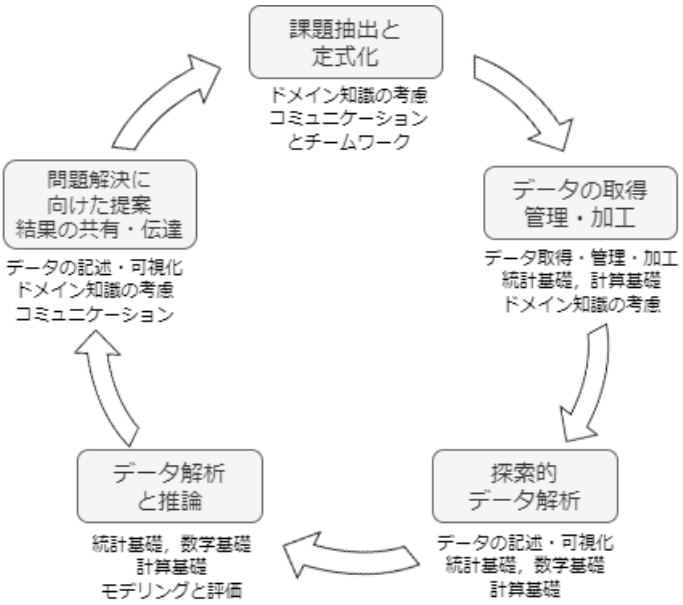


図 1　データサイエンスのサイクル

2.2 データサイエンスのプロセス

データマイニングプロジェクトを成功に導くためのフレームワークとして，CRISP-DM が提案された。CRISP-DM は，業界や使用技術に依存しない汎用的なプロセスモデルであり，データマイニングプロジェクトのコスト削減，信頼性の向上，再現性の確保，管理の容易化，迅速化を目的としている。さらに，計画，プロジェクトチーム内外のコミュニケーション，文書化などにも役立つことが実証されている[4)]。データインフォームド型問題解決はデータをもとにした意思決定や施策作りを助ける方法であり，データを理解したり，見せたり，説明したりすることや，データを解析して予測したり，発見したり，分類したりすることにおいて，過程の多くはこの CRISP-DM のフェーズと重なる。よって，CRISP-DM は現在データサイエンスのプロセスとして考えられている。

図２は CRISP-DM の６フェーズを示している。以下，各フェーズの概要を簡潔に説明する。

（1）　ビジネスの理解

業務課題や分析の目的と要件を理解し，明確化する。このフェーズは，業務の現状やプロジェクトに求められている役割から，分析で突き詰めるべき課題を明確にする重要なフェーズとなる。

（2）　データの理解

ビジネスの理解で策定した業務課題を解決するために，現状どのようなデータがあるかを確認し，実施したい分析をするために十分なデータが揃っているか，不十分である場合どのようなデータが更に必要かを確認する。外れ値，欠損値の確認，データの分布，データ同士の相関関係もここで把握する。

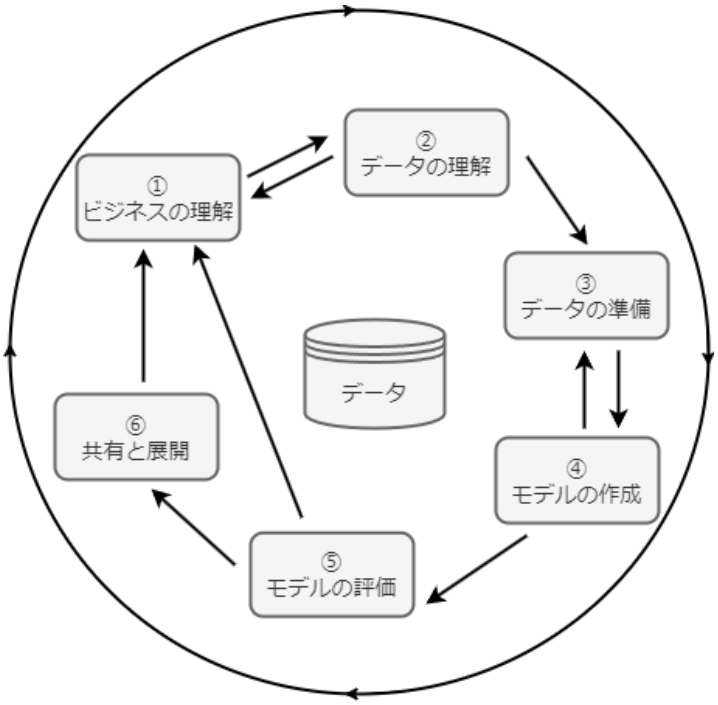


図 2　データサイエンスのプロセス（CRISP-DM）

（3）　データの準備

データを，分析に適した形に整形する。多くの場合，入手できる生データは雑然データとなっているため，データの統合や変換，表記ゆれの対処，外れ値の除去，NA 値の扱いなど，様々な処理を行うことで，整然データになるように整える。

（4）　モデリングの作成

課題を解決するために，統計学的なアプローチや機械学習の手法を使ってモデルを構築する。データアナリティクスとも呼ばれる。一般に，目的ごとに複数のモデルを作成し，あらかじめ定めた指標をもとに，評価指標をもとにどのモデルを利用するかを評価結果から決定する。

（5）　モデルの評価

前のフェーズで作成したモデルが最初のビジネス理解で定義した目標を達成するために適切か，または十分な精度になったかをビジネス観点から評価する。

（6）　共有と展開

分析結果を組織内で共有し，ビジネスアクションとして適用するための具体的な計画を立案・実施する。得られた予測値の精度がどんなに高かったとしても，それを共有し，ユーザーが意思決定できるような形で提供できなければ，価値を最大化することができない。そのために，予測結果をどのように表現するか（プレゼンテーションの方法）も重要になる。例えば，Tableau はダッシュボードによる予測結果の可視化や，Tableau Server のオンラインを使った組織内でのインサイトの共有とコラボレーションができる。

2.3 Exploratory のデータサイエンスのワークフロー

Exploratory は, データサイエンスの民主化を目指したデータ分析ツールである。シンプルな UI で, データラングリング，データ可視化，統計や機械学習を用いたデータアナリティクス，そしてダッシュボードやレポートの作成といったデータサイエンスの基本には欠かせないタスクを簡単に行うことができる。R を基盤にして便利な分析やアナリティクス方法を備えたため，何万とあるオープンソースのパッケージを使ってさらに機能を拡張することができる。

図 3 は Exploratory のデータ処理ワークフローを示している。以下その概要を簡潔に説明する。

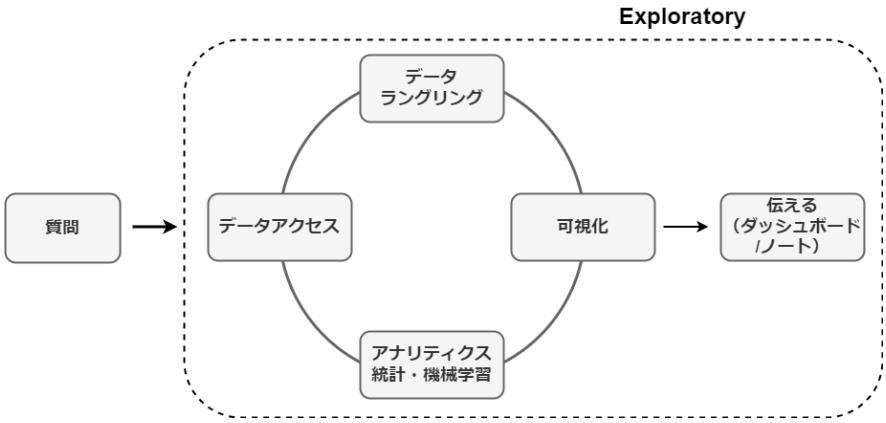


図 3　Exploratory のデータサイエンス・ワークフロー

データ処理は事前に持つ質問（問題）に関わるデータへアクセスし，データの前処理やクリーニングを含むデータラングリング（Data wrangling）を経て，統計学・機械学習モデルの手法を用いてアナリティクスを行う。処理結果を様々なグラフで可視化し，ダッシュボードやノートといった機能を用いて，可視化結果をインサイトの共有や伝えることができるといった特徴がある。

ここでいうデータラングリングとは，一般にデータ分析や機械学習などの目的でデータを使用するために，データを収集，クリーニング，変換，統合するプロセスのことである。データの質を向上させ，分析や機械学習のためにデータをより扱いやすく有用にすることを目的とし，データを生の雑然状態から望ましい整然データへと変換することや構造化させることで，分析や機械学習の精度と効果を向上させることにつながる。

2.4 データインフォームド型教材の設計方針

データインフォームドと言えば, データ駆動（Data-Driven）について言及する必要がある。一般に, どちらもデータに基づいた意思決定や問題解決の思考方法であるが, データ駆動型のアプローチでは, 収集されたデータを分析し，パターンや傾向を特定して意思決定を行う。データの分析によって得られる洞察が，意思決定の主要根拠となる。また，データ駆動型のアプローチは客観的で事実に基づいており，感情や主観的な要素が少ない傾向がある。一方，データインフォームド型の意思決定もデー

タに基づいているが，データだけでなく，領域の経験やノウハウ，洞察なども考慮に入れる。つまり，データインフォームド型のアプローチでは，データを重要な情報源として活用しながらも，意思決定においてデータサイエンティストの判断や直感も重要視された意思決定を行う。要するに，データ駆動型の意思決定は主にデータに依存し，客観的なアプローチを強調する。データインフォームド型の意思決定は，データを尊重しつつも，専門家の洞察や判断も考慮に入れるアプローチである。どちらのアプローチも異なる状況やコンテキストにおいて有用であり，適切な方法を選択する際にはその特性を考慮する必要がある。本研究では，敢えて両者を厳密に区別しないが，学生がデータと共に思考し，データ分析のプロセスへ取り組む際に独自の考えも取入れながら問題解決の手法を学ぶという点において，データインフォームドを使用した点がポイントとなる。

本研究では，上記述べたデータサイエンスのサイクルやプロセスの原則を考慮し，データに基づく問題解決のプロセスを重視したデータインフォームド型データサイエンス教育教材を開発している。

教材の設計方針は次に示すとおりである。

（1）教材の設計原則：データサイエンスのサイクルに基づく。

（2）データサイエンスのプロセスの理解：実データでサイクルを回して実践には，CRISP-DM に基づく実データセットを用いた課題解決モデルに従う。

（3）データインフォームド型の局面コンテンツの理解：データの理解，データ準備，データ可視化などのデータ分析の局面的コンテンツは，Exploratory や Tableau を用いて展開する。

（4）データモデリングの理解：統計学的・機械学習的アナリティクスのモデル理解は Anaconda 環境や Google Colab 環境の Jupyter Notebook を用いる。典型的アナリティクスモデルの場合は Exploratory を使用する。

3. 教材開発のアプローチ・使用ツールおよび素材

前項で述べた教材の設計方針に従い，データインフォームド型教材の開発を行っている。ここでは，教材開発のアプローチ，使用するツールおよびデータ素材等について説明する。

3.1 教材開発のアプローチ

データインフォームド型データサイエンス教育教材は，実際のデータセットを教材の中核に据え，学生がデータ分析を通して知識を体験的に習得できるようにしなければ意味がない。従来の理論中心の教材とは異なり，データ分析の実践的なスキルを育成することに重点を置いている。一般的なデータインフォームド型教材開発のアプローチ例として，次のようなものが挙げられる。

- データ分析プロジェクト教材（PBL 型）：企業や自治体と連携し，実社会の課題解決に取り組む。
- シミュレーション教材: データ分析の各ステップを体験できるシミュレーション教材を用いる。
- オープンデータ教材: 政府機関などが公開しているオープンデータを用いて，データ分析を行う。
- データ分析コンテスト教材: 実際のデータセットを用いて，課題解決に取り組む。
- ケーススタディ教材: 過去のデータ分析事例を題材に，分析方法や結果の解釈を学ぶ。

データインフォームド型のデータサイエンス教育教材とは，データ，コード，インタラクティブなビジュアライゼーションなどを一体化した教材となる。このタイプの教材は，学生がデータと共に思考し，データサイエンスの基礎概念や理論をよりよく理解するとともに，実際のデータを使用してデータ分析の過程で直面する様々な課題の対処，分析結果の解釈や活用する際の知識やスキルを訓練するのに役立つものである。

データインフォームド型データサイエンス教育教材は，このような考えを持つ新しいスタイルの教材となり，特に，実際のデータセットを用いて，データ分析の全過程を体験できるようなコンテンツで構成される。具体的には，以下のような特徴がある。

・実際のデータセットを用いる:
教材は，実社会の様々な応用領域の問題に関連する実際のデータセットを用いており，学生はデータ分析を通してこれらの問題について理解を深めることができる。

・データ分析の全過程を体験できる:
教材は，データの収集・前処理，統計分析，機械学習，結果の解釈・活用まで，データ分析の全過程を体験できるような内容になっている。

・データ分析の課題について学べる:
教材は，データ分析の過程で直面する様々な課題についても解説しており，学生はこれらの課題を克服するためのスキルを身につけることができる。

・データ分析結果の解釈・活用方法について学べる:
教材は，データ分析結果の解釈や活用方法についても解説しており，学生はデータ分析結果を意思決定に役立てるためのスキルを身につけることができる。

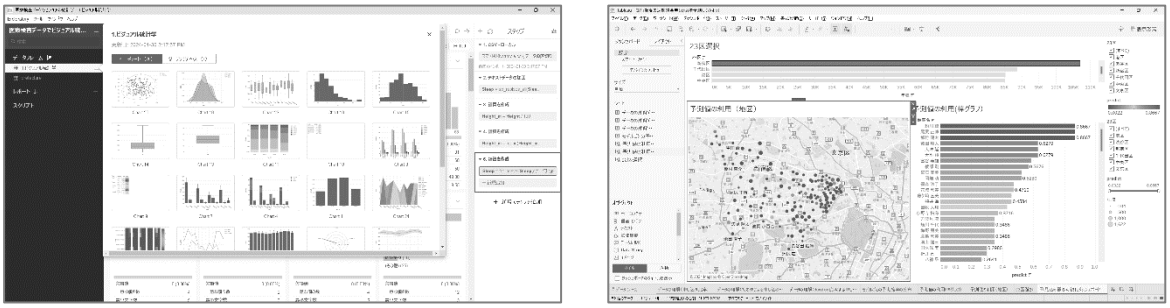
3.2 教材開発および教育実践のための環境

本研究で開発する教材のアプローチは，学生がデータと対話しながら，ノンプログラミング環境でデータ分析の基礎を学べるようにすることである。そのために，教材コンテンツの開発と教育実践において，Exploratory と Tableau という二つのデータ分析ツールを採用した。これらのツールは，データの可視化や操作を容易に行えるだけでなく，今後の教育実践にも適応できるツールとなる。また，データアナリティクスにおいて重要なデータモデリングのスキルを身につけるために，Jupyter Notebook や Visual Studio Code というプログラミング環境を使用した。プログラミング環境では，データアナリティクスに必要な機械学習パッケージである scikit-learn や科学技術計算用のライブラリ Scipy などを利用する。プログラミング環境の構築には，受講生が自分のデバイスを持ち込む BYOD 方式を採用する。この方法では，開発環境の構築が比較的に容易である。

Tableau は，データの可視化に特化しており，様々な種類のグラフやダッシュボードを簡単に作成できる。また，Tableau は，テキストファイルやエクセルファイルだけでなく，データベースやクラウドサービスなどにも直接接続ができ，データの結合や変換，フィルタリングなどをグラフィカルに操作できる。さらに，Tableau Server や Tableau Online などを使って，作成した分析結果を組織内や外部に

共有できる。一方，Exploratory は，データを探索的に分析することができるツールである。R 言語を基盤としており，様々な統計や機械学習のアルゴリズムを使ってデータを分析することができる。特にコードを書かなくても，直感的な UI でデータの加工，可視化，ダッシュボード作成などができる。

図３に Exploratory および Tableau のそれぞれの利用者環境を示す。図中（１）は Exploratory の分析チャート，（２）は Tableau のダッシュボードのシステム画面のスクリーンショットである。



（１）Exploratory の分析チャート　　（２）Tableau のダッシュボード

図 4　教材開発および教育実践のための利用環境

また，データアナリティクスにおいて重要なデータモデリングのスキルを学習に必要な場合，予備科目のプログラミング演習などで Jupyter Notebook や Visual Studio Code の準備ができている。

3.3 教材開発の使用素材について

データサイエンスの教育において実データセットを用いることは，学生に実践的な経験を与え，彼らのスキルを実世界で適用できるようにするために重要である。その理由は以下の項目が挙げられる。

１）データ分析の結果を実社会や実問題に応用する能力を養うことができまる。データサイエンスは，実社会の課題を解決するために用いられる実用的な技術である。実データセットは，実際のビジネスや社会分野の問題に関わるデータであるため，そのデータ分析の学習は，今後実社会での実問題に対する技能や解決策として活用することができる。実データセットを用いることで，データ分析の意義や価値を認識し，データから新しい価値を生み出すことへの理解が高まる。さらに，学生は実際のデータに基づいて問題を発見し，解決策を検討する経験を積むこともできる。

２）データの特徴や問題点を理解し，適切な分析手法や可視化方法を選択する能力を養うことができる。データサイエンスでは，データの収集，加工，分析，解釈などのスキルが求められる。実データセットを用いることで，受講生はこれらのスキルをより実課題に近き形で体験できまる。実データセットは，必ずしも整った形で提供されているわけではない。理想的な整然データとは異なり，欠損値や外れ値，ノイズなどを含んでいることが多く，それらをどのように扱うかが分析の質に影響する。また，分析の目的や背景に応じて，異なる視点や指標で分析することができる。実データセットを用いることで，データに対する感覚や洞察力を磨くことができ，受講生はデータの欠損や異常値などの問題を解決しながら，データ分析を行う必要があり，データ分析のスキルを身につけられる。

3）データ分析のコミュニケーションや協働の能力を養うことができ，データサイエンスの総合的応用力を養うことができる。実データセットは，多くの場合，複数の人や組織が関わって収集や提供されているデータであるため，その分析には，データの出典や品質，利用条件などを確認することが必要である。また，その分析結果は，分析者だけでなく，他の人や組織にも共有やフィードバックすることが必要である。実データセットを用いることで，データ分析の釈明と共有などに必要なコミュニケーションや協働の重要性やその方法を学ぶことができる。

本研究では，データサイエンスのサイクルで示す課題の定式化やビジネスの理解において，学生にとって現場の実務経験などから難しい点があることから，教材開発では，このフェーズを省略した。また，実データセットについても，現場に近い雑然データの入手は困難であるため，ある程度整ったデータセットを使用することとした。表 1 はその概要を示している。

表１　教材開発に使用している素材

	内容	説明
C-1	・SSDSE（教育用標準データセット） ・SSDSE-C（家計調査の「品目分類」によるデータ） （そのほか SSDSE-A～F も使用する）	独立行政法人統計センターが作成・公開している，データ分析のための汎用素材として利用できる統計データで, 主要な公的統計を地域別に一覧できる表形式のデータセットである。
C-2	・Bank＋Marketing（銀行顧客の定期預金申込データ） ・その他の機械学習モデルを適用できるデータセット	機械学習でよく利用されている公開データで, 電話によるテレマーケティングデータ。2008 年から 2011 年間のボルトガルの銀行顧客の行動履歴，経済指標, 顧客が実際に預金を申し込んだかどうかの記録情報などを含む。
C-3	・Airbnb_東京宿泊データ （Exploratory 社のサイトから参照） ・医療検査データセット	データ分析の局面的ビューの可視化やデータチャートを利用する。 データ可視化の多様な表現の作成参考にする。

4. 教材の事例紹介

本研究プロジェクトで提案するデータインフォームド型学習コンテンツの一部を紹介する。主な関連項目は以下の通りである。

1）データの理解とデータの準備（前処理）。

2）データの局面的アナリティクス（統計学的分析ピューに関わるチャート）。

3）データモデルの作成と評価（機械学習モデルの実践）。

4）データ共有のためのダッシュボード作成。

4.1 データの理解とデータの準備について

データの理解は，分析対象となるデータセットの全貌を理解し，データの意味や背景，傾向や特徴，関係や因果性などを把握することである。その目的は，データに基づいて正確かつ効果的なビジネスの意思決定を行うためと言える。こうした基本素養は教材のデータセットの理解を通して徐々に養っていく。Exploratory のデータサマリビューは，分析データの各列の統計情報とその分布を素早く確認できる機能で，データの型に応じて，バーチャートや欠損値の割合などを表示することができる。

図５は，Exploratory のデータサマリビュー機能で確認した SSDSE-C 家計調査の「品目分類」によるデータの全貌（図は部分）を示す。列項目（特徴量）ごとの基本統計情報，欠損状況などデータの傾向や特徴を直感的に把握するのに役に立つのである。

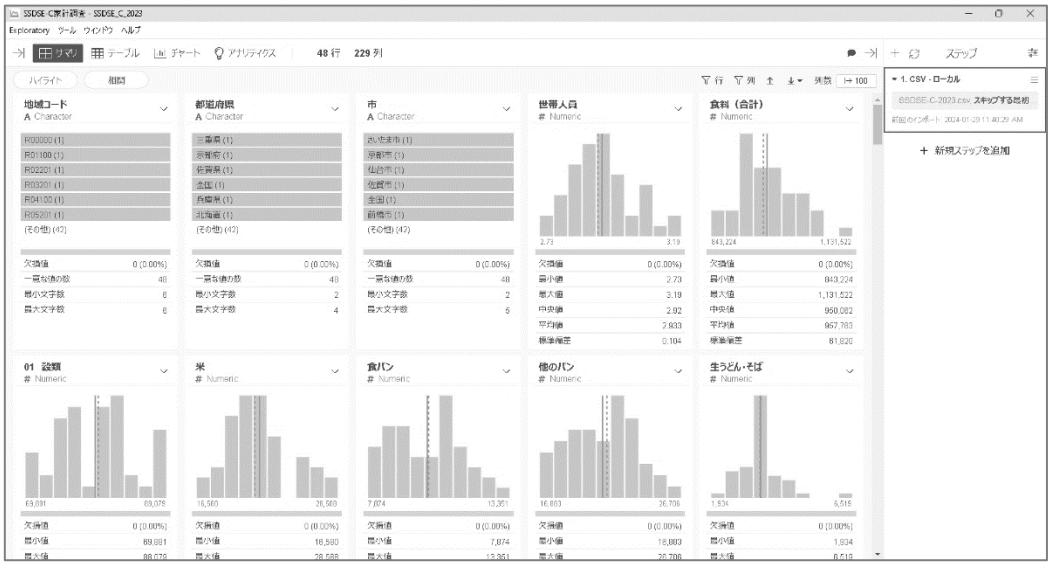


図 5　Exploratory の SSDSE-C 家計調査の「品目分類」によるデータの全貌

こうしたデータの基本を学生達に理解するとともに，後のさらなる分析のためにデータの整然作業をも行う。例えば，データ標記ゆれの解消，商品名の名寄せ，属性の抽出，数値データの抽象化，データの型の変換，欠損値の対応なども行う。こうしてデータの理解をさらに深めることにより，今後データの設計や偏りを評価し，適切な分析方法や可視化手法を選択し，仮説を立てて検証する能力を養うことにも役に立つのである。

4.2 データの局面的アナリティクス

データの理解を深めるためには，特定の項目間における相関性や因果関係を見る必要がある。例えば，Bank＋Marketing（銀行顧客の定期預金申込データ）教材では，Campaign（現在キャンペーンにおける顧客とのコンタクト回数）と Duration（最終コンタクト時間秒数）の関係を焦点に当て，サブピュー的に理解したい場合，可視化して確認すればよい。図６は Exploratory のチャートビューで作成した結果を示す。

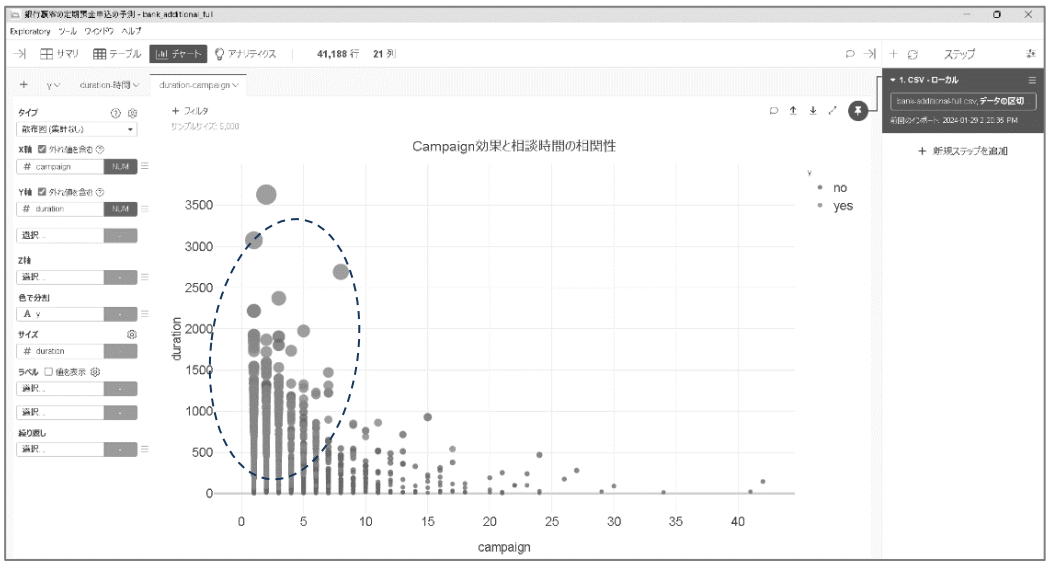


図 6　Exploratory でデータ理解（Campaign と Duration の相関性チャート）

可視化結果の散布図の点では，図 6 中の楕円点線で囲まれた部分が成約（「Yes」＝実際に申込のあった顧客）を表わしている。コンタクトの回数が少なく，かつ，コンタクトの時間が長いところに集まっていることが分かる。逆に，コンタクトの回数が多くても，話す時間が短いところには成約がない（図中右裾の部分）。これは，相手に「親切」だと感じさせるより，「しつこい」と思われている印象が強いということになる。こうした事例分析を通して，学生に営業の難しさを理解しつつ，顧客に対して最適な連絡回数や時間の長さを判断する基準を得ることができる。

Exploratory のサマリビューから特定の特徴に対して，チャートビューやアナリティクスビューに移動することで，データの理解を深め，データをさらに深く探索することができる。また，Tableau も特定項目の可視化や分析チャートの作成が得意としている。関連教材の学習を通して，学生はこの種のツールを習熟し，今後のキャリアで素早く活用できることが期待される。

4.3　データモデルの作成と評価（機械学習モデルの実践）

データサイエンスは，大量のデータを体系的・法則的に整理し，課題解決や新たな知見の発見につなげる分析手法である。これには，統計学や機械学習モデルなどがよく用いられ，様々なビジネスシーンに活用されている。本研究で開発している教材は，データ分析で広く用いられる典型的統計学および機械学習モデルの学習に必要な例題を備えている。例えば，線形回帰，ロジスティック回帰，K-Means 法, 決定木, ランダムフォレスト, 主成分分析 PCA などがある。こうしたモデルは, Jupyter Notebook や Visual Studio Code を用いてプログラミングで学ぶことも可能であるが，Exploratory のアナリティクスビューでノンプログラミング的に活用できるようになっているため，教材素材のデータセットを基に学習事例を設けている。または，Tableau と Jupyter Notebook を連携して，データサイエンスのプロセスにおけるフェーズごとに学習することができる。

図 7　Exploratory でデータアナリティクス（K-means 法の例）

図 7 は，Exploratory のアナリティクスの K-means 法による「家計調査」の「牛肉」「豚肉」「鶏肉」の消費支出から見た県民特性のクラスタリング分析の結果を示している。家計調査の消費支出は 10 大分類に分けられ，「食料」はその一つで，212 品目が含まれている。その中の中分類「肉類」には小分類「生鮮類」があり，「牛肉」「豚肉」「鶏肉」の項目が含まれている。この結果を通じて，受講生は生鮮肉の「嗜好」に関する地方の傾向を見ることができる。

こうした典型的モデルを活用して，ノンプログラミングでデータ分析を行うほかに，必要な場合，モデルの作成及び評価を含めたデータサイエンスのプロセスに対する理解力の育成も欠かせない。

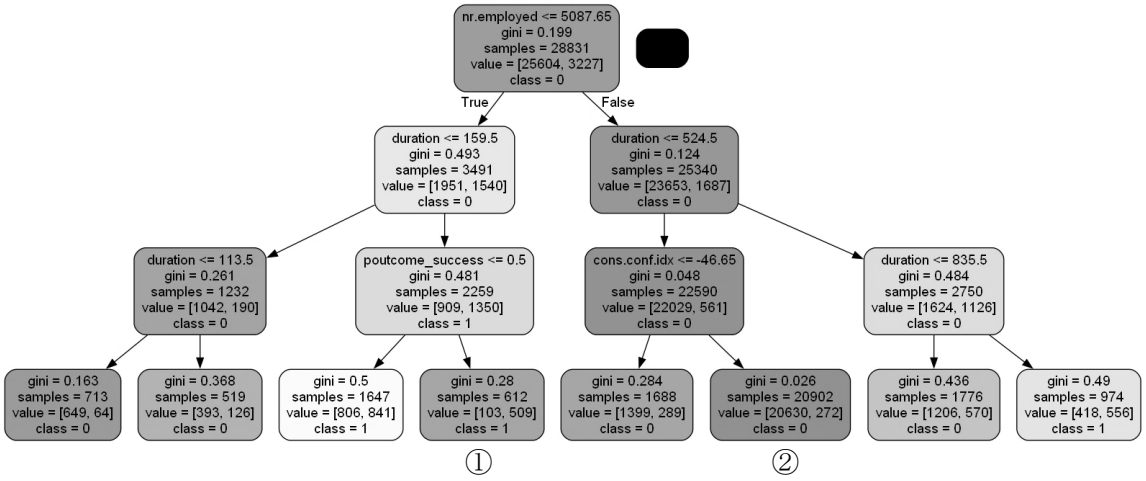


図 8　Jupyter Notebook で決定木モデルに関する学習ユニット教材の一例

この場合は Jupyter Notebook を活用したモデル作成の学習ユニットも用意している。図８は Tableau と Jupyter Notebook を連係プレイした Bank＋Marketing（銀行顧客の定期預金申込データ）学習例の一部を示している。Python の scikit-learn に含まれる決定木分類モデルを利用して，定期預金を申し込む確率を推論する学習内容である。濃いグレー部分（図中①）は目標変数で，「定期預金の申し込みが有（yes）」が多い分類，逆に深いダーク部分（図中②）は定期預金の申し込みが少ない分類になる。

この学習ユニットでは，データセットを７対３の割合で訓練データと検証データに分割し，訓練データを学習してモデルを作成し，検証データを用いてモデルを検証する。図は，決定木の深さ max_depth=3；ノートに含まれる最小サンプル数 min_samples_leaf=500 に設定した後，モデルに fit した結果を graphviz ツールで可視化したものである。その後，モデルのチューニングを実施し，モデルの精度変化を確認する。こうした過程でモデルの精度を向上した後，すべてのデータに対してモデルを利用した「定期預金を申し込む確率」を算出し，Tableau を使って可視化をしながらモデルの評価とデータの予測値を生成する。さらに，作成したモデルで新しいデータにたして推論し，その結果を利用することになる。詳細についてここでは割愛するが，文系の学生にはハードルのある学習内容となる。

4.4 分析結果の共有（情報のダッシュボード化）

データ共有のためのダッシュボードは，ビジネスの意思決定をサポートする重要なツールとなる。これは複数のチャートやアナリティクスの結果を組み合わせて，洞察を提供する役割を果たす。特に学生には，データ分析結果を分析者だけでなく他の人と共有し，フィードバックを受けることが重要である。そのためには，わかりやすいダッシュボードを作成し，データ分析の釈明と共有に必要なコミュニケーションスキルや協働の方法を学ぶことが必要である。このような教育的アプローチにより，学生は将来のビジネス環境で効果的にデータを活用し，意思決定に寄与できるようになる。

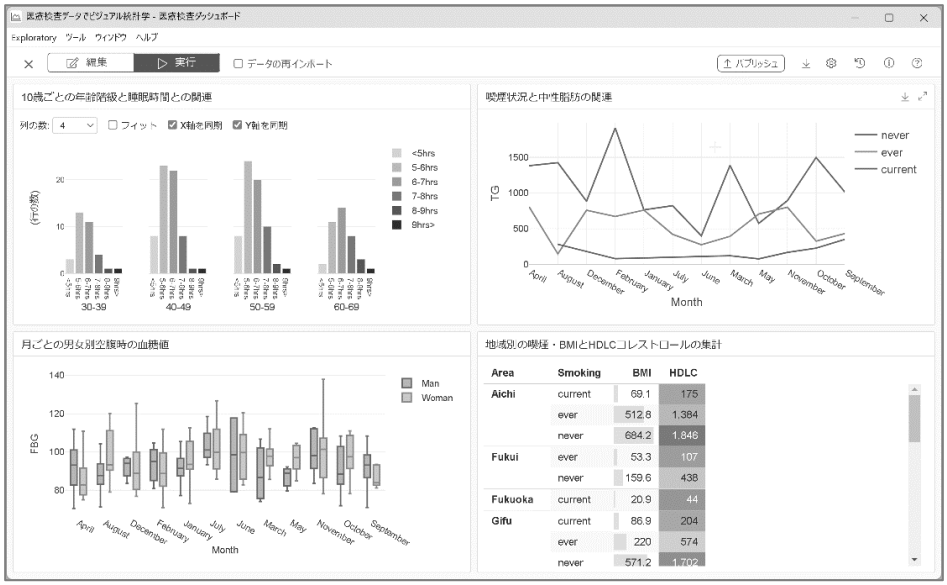


図 9　情報共有のダッシュボード

Tableau は，連携性が高く，ビジュアライゼーションが秀逸で，直感的な操作感の UI を持ち，リアルタイム更新が可能で，ユーザーニーズに応えた分析力を備えている。一方，Exploratory も，データを公開・共有でき，分析結果や作成したチャートを入れたり，テキスト説明文を入れたり，また画像を貼り付けたり，などなど自由度の高いレポートを作成できる。さらには，それをサーバー上にパブリッシュすることで他の人と共有することができる。

図９は，Exploratory のレポート機能を活用した学習ユニット教材の一例である。このダッシュボードは，医療検査のデータセットを分析した結果から得られた血糖値，血圧，コレステロール，睡眠時間などの関連チャートを基に作成されている。

データインフォームド型教材では，このようなダッシュボードやノートの作成技法も取り入れてある。わかりやすいダッシュボードを作成し，データ分析の釈明と共有に必要なコミュニケーションスキルや協働の方法を学ぶことで，将来のビジネス環境で効果的にデータを活用し，意思決定に寄与するためのスキルを養う一助となり，今後のキャリアにおいて競争力を高めるのに役に立つのである。

5. 結び

データサイエンス教育の文脈において，データ分析プロセス中に生じる課題への対処および分析結果の解釈及び応用能力の向上に対する教材の開発が重要視されている。現在，データに基づく問題解決プロセスに焦点を当て，学生がデータを理解し，それを用いて論理的に思考する能力を育成する目的の教材は不足している。特に，プログラミングに苦手意識を持つ文系学生にとって，プログラミング不要のデータアナリティクス手法やインタラクティブなビジュアライゼーションを取り入れた教材の必要性が高まっている。この種の教材は，学生がデータ分析の基礎技術を習得するだけでなく，データに基づいた意思決定能力の育成にも寄与することが期待される。

本研究では，データサイエンス教育における教材開発に注目し，その開発アプローチおよび成果について詳細に分析を行った。データインフォームド型の手法を採用することで，学生がデータサイエンスの核心概念および技術を理解し，実際のデータセットを活用して問題解決スキルを向上させるための教材開発の手法を提案した。主要な内容は，以下の通りである。

本研究における教材設計の基本方針は，データサイエンスのサイクルを根幹に据えつつ，実際のデータセットを用いた問題解決モデルに基づくことにある。教育用の開発環境としては，Exploratory，Tableau，Jupyter Notebook といったツールを活用しており，これらを通じてデータの可視化，分析，そしてモデリングの各技術を学生が容易に理解し学べるように各学習ユニットを細かく設計している。さらに，実際のデータセットの使用を通じて，学生は理論と実践が統合された学習体験を得ることができる。

本研究における教材開発アプローチは，データサイエンスのプロセス全体を体験することに焦点を当てている。具体的には，データの理解から始め，データ準備，可視化，モデリング，分析結果の共有に至るまでを包含する。この方法論により，学生はデータサイエンスの理論と技術を包括的に習得し，実際のデータを用いた問題解決能力を育成することができる。

本研究プロジェクトは，本稿執筆時点で進行形であり，既にデータサイエンス専攻内の「専門演習」や「データサイエンス実践演習」などの科目で教材の一部が実装されている。しかし，学習者の学習成果や教材効果の評価については，将来の研究で詳細に検討される予定である。今後の教材開発における課題には，以下の点が挙げられる。

・実践的スキルの強化：

理論的な知識に加え，実際のビジネスや社会問題に対応する応用スキルの教育を強化することが求められる。これには，実世界のケーススタディの導入や，実データを用いたプロジェクトベースの学習が有効である。特にデータの収集や現場の生データの前処理に関わる部分は欠けていたが，今後 PBL 手法を通じて教材を改善する必要がある。

・異なる分野への適応：

教材をさまざまな業界や専門分野に適応させることで，学生の関心とニーズに対応する多様な教育コンテンツの提供が必要である。また，データの倫理的使用とセキュリティに関する教育を強化することも，データサイエンスの教育において極めて重要である。今後，データプライバシー，データの正確性，透明性の確保，偏見の回避などの倫理的に重要なテーマを組み込むことを視野に入れた教材の開発が必要である。

・カリキュラムの継続的更新：

データサイエンス分野は急速に進化しているため，教材内容の定期的な更新が不可欠である。これには，最新のデータサイエンス技術，ツール，アプローチを取り入れることが含まれる。

本研究の重要点は，理論と実践のバランスを保ちつつ，データサイエンスの教育におけるデータインフォームドアプローチの効果的な実装にある。教材の設計方針，使用ツール，教材開発のアプローチは，学生がデータサイエンスの基本概念と実践的なスキルを習得するための効果的な基盤を提供している。今後の課題としては，教材の内容を現実世界の応用に密接に結びつけ，学生が実際のビジネスや社会問題への対応能力を高めることが求められる。これらの取り組みにより，データサイエンス教育はさらにその効果と範囲を拡大し，高度データサイエンス人材を育成するための重要な役割を果たすことになる。最後に，本研究で提案したアプローチは，文系私立大学に限らず，様々な場面におけるデータサイエンス教育カリキュラムの実装に参考となることを期待している。

謝　辞

本研究は，「2023 年度 関西国際大学教育総合研究所」研究プロジェクト助成を受けたものである。

参考文献

1) Richard D. De Veaux，et al.，Curriculum Guidelines for Undergraduate Programs in Data Science, http://dstf.acm.org/ ，Annual Review of Statistics and Its Application，Vol. 4:15-30 (Volume publication date March 2017)

2) ACM Data Science Task Force，Task Force Final Report，「Computing Competencies for Undergraduate Data

Science Curricula」，http://dstf.acm.org/ ，(January 2021)
3) 数理・DS 教育強化拠点コンソーシアム カリキュラム分科会，「データサイエンス教育に関するスキルセットおよび学修目標」-第 1 次報告（リテラシーレベル）」，（2019 年 11 月）
4) Rüdiger Wirth and Jochen Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, Proceedings of the 4th International Conference on The Practical Applications of Knowledge Discovery and Data Mining (4), pp. 29–39, (2000).
5) https://exploratory.io/（アクセス：2024 年 1 月）
6) 章 志華，山本敏幸，「Python プログラミングで学ぶデータサイエンスのための高校数学基礎に関するオンデマンド教材の構築」，『教育総合研究叢書』第 16 号，pp. 109-123，（2023 年 3 月）
7) Zhihua ZHANG, T. Yamamoto，The Course Design Of "Basic Data Science" Taking into Account Both Face-To-Face and On-Demand Teaching and Effect Analysis, IEEE eXplore Conference Proceedings，（2023 年 11 月）
8) Zhihua ZHANG, T. Yamamoto and K. Nakajima, Development of Education Curriculum in the Data Science Area for a Liberal Arts University，Proceedings of IFIP WCCE 2022，（August 2022）
9) National Academies of Sciences，Engineering，and Medicine 2018，Data Science for Undergraduates: Opportunities and Options，Washington, DC: The National Academies Press，(2018)
10) 藤井亮輔・鈴木康司, R でできるビジュアル統計学 学会・論文発表に役立つデータ可視化マニュアル, 株式会社　金芳堂，（2021 年）
11) 岩橋智宏・今西航平・増田啓志，Tableau で始めるデータサイエンス，株式会社 秀和システム，(2019 年)
12) 山本章博，「データサイエンス・プロセスから見るデータサイエンス・カリキュラム」，情報知識学会誌，Vol.31，No.4，pp.452-461（2021 年）
13) 大橋真也，「高等学校共通教科「情報」におけるデータサイエンス」－新学習指導要領解説およびその他の資料から見えること－，コンピュータ＆エデュケーション，Vol.52，pp.18-25（2022 年）

Abstract

In data science education, there is a need to develop teaching materials that provide students with a thorough understanding of the challenges that arise in the data analysis process, as well as the interpretation and utilization of results. In this study, we propose a data-informed data science education teaching material. The teaching material emphasizes the process of problem-solving based on data, with the aim of cultivating students' ability to think with data. In particular, it is possible for students with a high sense of programming difficulty, such as students in the humanities, to learn data analytics in a non-programming environment and interactive visualization content.