データサイエンティスト検定™
リテラシーレベル

書籍（第3版）掲載の模擬試験（50問）の解説

※『最短突破　データサイエンティスト検定（リテラシーレベル）公式リファレンスブック　第3版』に掲載した問題の解説です。「初版」、「第2版」に掲載した模擬試験の解説ではありませんので、ご注意ください。

問題番号	解答	解説	該当ページ
1	c	固有ベクトルは行列の線形変換に対して不変の方向を示し、固有値はそのベクトルがどのように伸縮するかを示します。よって正解は**c**です。	34
2	d	勾配ベクトルとは、偏微分係数を並べたものをいいます。xで偏微分すると$2x+2y$、yで偏微分すると$2x-2y$となるので、$(x, y)=(-1, 2)$の勾配ベクトルは$(2, -6)$となり、**d**が正解となります。	38
3	a	問題のベン図では、集合Aと集合Bのどちらか片方にのみ含まれる要素の集合が表されています。これは対称差集合と呼ばれ、「$A \triangle B$」と表現されることもあります。よって、正解は**a**です。	43
4	b	標準偏差は分散の平方根を取った値であるため、まずは分散を求めます。データの平均は$(6+1+0-2+7)/5=2.4$と求められます。ここから、分散は$\{(6-2.4)^2+(1-2.4)^2+(0-2.4)^2+(-2-2.4)^2+(7-2.4)\}/5=12.24$となります。この平方根を取ると、標準偏差は3.50とわかります。よって**b**が正解です。	51
5	a	今回の問題では、数学のテストのスコアと読書量には、一方が変化すれば他方が変化するという関係性（相関関係）があるということを主張しており、原因と結果の関係（因果関係）があるとは主張できないことに注意が必要です。そのため、相関関係を主張している**a**が適切だとわかります。	56
6	b	指数関数のグラフが直線になるのは、対数スケールにした場合です。特に、$y=a^x$の形の関数は、y軸を対数スケールに設定したときx軸に対して直線のグラフになります。よって**b**が正解です。	63

問題番号	解答	解説	該当ページ
7	a	仮説検定では、否定したい仮説を帰無仮説、主張したい仮説を対立仮説に設定します。この問題では、運動プログラムが参加者の体力向上につながったかどうかを検証しています。よって、帰無仮説は、一般的に「体力が向上しない」または「効果がない」となります。対立仮説は、証明しようとしている仮説であり、この場合は「運動プログラム後の参加者の体力が向上する」となります。よって、適切な組み合わせは**a**です。	73
8	a	因果推論において、特定の処置の効果を測定するためには、処置を行う「実験群（処置群）」と、処置を行わない「対照群」に分けて比較・分析する必要があります。よって、正解は**a**となります。	80
9	d	現在の需要データと予測に乖離がある場合、その乖離の原因を捉えるか、いずれかのデータに誤りがあると考えて対策を講じる必要があります。分析結果を鵜呑みにして、現実と乖離したまま進めても、生じている誤差を解決する行動にはつながりません。よって、**d**が正解となります。	88
10	b	データ解析結果を他部門に伝える際には、**a**のようにデータだけでは伝わりません。また**d**のように集計結果を網羅的に大量に出しても見てもらえないことが多いです。**c**の示唆を提示することは大事ですが、示唆のみだとなぜそうなったのかがわからず、現場が考えることができなくなります。よって、正解は、**b**の分析結果と関連性を提示しつつ、会議で議論することになります。	109

問題番号	解答	解説	該当ページ
11	d	**a**.原点を0以外に設定すると、見た目と実数にずれが生じ、誤った解釈を導く可能性があるため不適切です。 **b**.ヒストグラムは1つの定量属性のみを受け入れるため、区間ごとの度数データを違う色で表現してもかえって見にくく不適切です。 **c**.3D円グラフで立体的に示すと、実際の割合と視覚的な感覚がずれ、誤った解釈を導く可能性があるため不適切です。 **d**.増加していることを強調するために矢印を重ねることは、適切な表現です。よって**d**が正解です。	112
12	c	箱ひげ図では、第1四分位数と第3四分位数に基づいて「箱」が描かれます。箱の外に延びる「ひげ」は、通常、第1四分位数と第3四分位数から算出される四分位偏差を用いて箱の外に描かれます。外れ値は、そのひげの先を超えた値として表示されます。これはデータの分布において通常期待される範囲を超える値として識別されるためです。よって**c**が正解です。	117
13	c	Accuracy（正解率）は、正解のレコード数をすべてのレコード数で割ることで求められます。ここでは正解のレコード数は$110+120=230$であり、総レコード数は$110+120+20+30=280$であるため、Accuracyは$230/280=0.82$と求められ、正解は**c**となります。	125
14	b	過学習（overfitting）とは、機械学習モデルが訓練データに対して過度に最適化されてしまい、訓練データに含まれない新しいデータに対する汎化性能が低下する現象を指します。過学習の状態では、モデルは訓練データに対しては非常に高い精度を示しますが、未知のデータ（テストデータ）に対しては精度が著しく低下します。よって正解は**b**です。なお、**c**や**d**は二項分類で50%の精度であるため、未学習の状態にあると言えます。	132

問題番号	解答	解説	該当ページ
15	d	5-foldsクロスバリデーションでは、データセットを5等分し、そのうちの1つを検証データとして使用し、残りの4つを訓練データとして使用します。これを全ての分割で繰り返します。つまり、それぞれの分割では、20,000件のデータのうち、4,000件(全体の1/5)が検証データとして使用され、残りの16,000件がトレーニングデータとして使用されますが、全ての分割の中では、データが少なくとも一度は検証データとして使用されます。したがって、クロスバリデーションのプロセスが終了した後で「使用されないデータ」は存在しません。よって**d**が正解です。	137
16	c	連合学習では、データを共有せず、各端末で学習されたモデルのパラメータを中央サーバーで統合して最終的なモデルを構築します。この方法により、プライバシーが保護されつつ、複数の端末で学習された知識を活用して高度なモデルを構築することが可能となります。よって、正解は**c**です。	143
17	b	建設機械の故障検知において、モデルの複雑化による精度改善よりもデータの質と量を向上させることが効果的であることがあります。ノイズのフィルタリングやデータの補完、最新データの利用などを行うことで、検知モデルがより正確に故障を検知することができます。よって**b**が正解です。なお、複雑化は精度向上を期待できる反面、過学習になる可能性もあり注意が必要です。	144
18	a	深層学習は、大量のデータから複雑な関係を抽出し、その学習を通じて予測や分類などのタスクで高い性能を発揮することができる技術です。よって正解は**a**です。**b**は深層学習の範囲をカバーした表現でなく、**c**は様々な問題への適用は追加の調整が必要であり、**d**は計算資源として高度なハードウェアが必要という観点でそれぞれ正しい表現とは言えません。	145

問題番号	解答	解説	該当ページ
19	a	時系列データには長期的な傾向(トレンド)、季節的な変動、周期的な変動、そして不規則な変動が含まれているため、それらを全て考慮したうえで、目的に合わせて分析することが重要です。よって、最も適切でないアプローチは、長期トレンドを無視し、季節成分やノイズにのみ焦点を当てる**a**といえます。	147
20	b	デンドログラムを見ていくつかのグループに分ける際には、指定したグループの数だけクラスターが存在している箇所に横線を引くことで判断できます。クラスターが4つ存在するのは、4つに分ける場合、(A,B)(C,D)(E)(F,G)となるときです。よって、**b**が正解です。	153
21	a	有向グラフと無向グラフの主な違いは、エッジの向きです。有向グラフではエッジが向きを持ち、その向きが意味を持ちます。一方、無向グラフではエッジに向きがなく、エッジは単純に接続関係を表します。よって、**a**が正解です。	154
22	a	コンテンツベースフィルタリングは、ユーザーの好みに基づいて類似したコンテンツを推薦します。一方、協調フィルタリングは、ユーザー同士の行動や評価の類似性を基に推薦します。よって、**a**が正解です。	155
23	a	畳み込み処理において、入力画像の周囲を画素値で埋めることで出力のサイズを調整する操作をパディングと呼びます。よって、**a**が正解です。 なお、ストライドは畳み込み処理においてフィルタが移動する幅、プーリングは特徴マップのサイズを小さくする操作、カーネルは畳み込み処理で使用されるフィルタを指す用語です。	164

問題番号	解答	解説	該当ページ
24	c	大規模言語モデルの場合、モデルの構造は通常非常に複雑であり、多数のパラメータを含んでいます。ハルシネーションは、データの不足や誤り、学習プロセスの特性、または入力に対するモデルの解釈の仕方によって生じることが多く、単純な構造が原因とは一般的には考えられません。よって、**c**が正解です。	171
25	d	ER図はエンティティ間の関係性、特に構造的な関連を視覚化するために用いられるツールで、データの論理的な構造やエンティティ間の関係性に焦点を当てています。一方で、エンティティ間で交換されるデータの量やそのバランスを直接示すものではありません。よって、**d**が説明できないため、**d**が正解です。	188
26	b	第一正規化は、テーブル内の各カラムが原子的な値を持ち、重複するグループのデータが存在しないようにすることを目的としています。一方で、第二正規化は、第一正規化を満たした上で、各テーブルがその主キーに完全に依存する非キー属性のみを含むようにすることを目的としています。これは、主キーの一部にのみ依存する非キー属性を排除することによって達成されます。よって、**b**が正解です。	189
27	b	Apache Sparkは、大規模データ処理のための統合分析エンジンです。Apache Sparkはイミュータブル(不変)のデータ構造を使用しており、特定のレコードを指定しての変更操作はSparkのモデルにはそぐわないため、推奨されていません。データ変更が必要な場合は、変更を適用した新しいデータセットを生成する方法が一般的です。よって、**b**が正解です。なお、**a**は「構造化データの処理に適しており、非構造化データの処理はサポート外」という表現、**c**は「クラスター内のノード数に依存しない」という表現、**d**は「処理するデータの量に応じて実行するサーバの数を指定する必要」(動的な割当が可能)という表現がそれぞれ誤りです。	193

問題番号	解答	解説	該当ページ
28	a	入社年が2015年以前:これはHireYear <= 2015で表現できます。月収が300,000円以上であるが400,000円未満:これはSalary >= 300000 AND Salary < 400000で表現できます。そして、この2つの条件のいずれかを満たす従業員を抽出したいので、これらの条件はORで結ぶ必要があります。よって、**a**が正解です。	196
29	a	「https?」は、"http"または"https"で始まることを意味し、ウェブページのURLにおける一般的なプロトコルを指定しています。「s?」は"s"が0回または1回出現することを許容し、"http"と"https"の両方にマッチし、**a**と**d**がこの条件を満たします。 次に、最後の「(/.*)?」は、スラッシュで始まるURLのパス部分を示し、任意の文字列にマッチします。この部分はオプショナルであり、URLにパスが含まれない場合も考慮しています。URLはパスを含むことが多いため、この表現を入れておく必要があります。よって、**a**が正解となります。	197
30	d	文字列を日付型に変換する関数としては、TO_DATE関数が挙げられます。これは第一引数に変換したい文字列、第二引数に変換後のフォーマットを指定することで日付型に変換することができます。今回は年(Year)・月(Month)・日(Day)をYYYYMMDDで指定しています。よって、**d**が正解です。	204
31	c	BIツールを用いてダッシュボードやグラフを作成する際、データの選択、可視化したいカテゴリの明確化、データのコンテキストを明確化するタイトルや説明の入力などは重要な操作であり、これらのステップは、データを正確に理解し、適切に伝達するために必要不可欠です。一方で、**c**.グラフに表示するデータの色分けの基準を設定する 作業は、データの理解や解釈において重要な役割を果たしますが、データを選択し、その意味を解釈する基本的なプロセスに比べると、必要性が低いと考えられます。よって**c**が正解です。	212

問題番号	解答	解説	該当ページ
32	b	選択肢は、「正常」または「エラー」のテストの結果と、「ホワイトボックス」または「ブラックボックス」のテストの種類の組み合わせです。このシナリオでは、テストが具体的なエラーを示しているわけではなく、正しい動作（特定の日付に対する曜日の判定）を確認しているため、「正常」が適切です。 また、テストについては、内部構造に触れることなく、ある特定の入力（日付）に対する期待される出力（曜日）のみを確認していますので、「ブラックボックス」が適切です。よって、**b**が正解です。	219
33	d	画像系の学習済予測モデル、例えば画像内の文字認識（OCR: Optical Character Recognition）は、機械学習やディープラーニングモデルを用いた一般的な処理の一つです。これらのモデルは、画像データから情報を検出し、それをデータに変換することが可能です。よって、正解は**d**です。	223
34	c	問題文のSQLクエリは、国別に2010年から2020年までの温室ガス排出量の合計を集計し、合計排出量の降順（多い順）で国名とともに表示する操作を示しています。よって、**c**が正解です。	229
35	d	公開鍵暗号化方式では、公開鍵と秘密鍵の2つの鍵を使用します。公開鍵は誰にでも公開することができ、これを使ってデータを暗号化します。一方、その暗号化されたデータは対応する秘密鍵を持つ人だけが復号できます。この方法により、データの安全な送信が可能になります。よって、**d**が正解です。	235

問題番号	解答	解説	該当ページ
36	a	OAuthを使用する場合、まず認証サーバーに対してアクセストークンを要求する必要があります。これには、クライアントIDとクライアントシークレットなどの認証情報が使用されます。アクセストークンを受け取った後、そのトークンをHTTPリクエストのAuthorizationヘッダに付与してAPIを呼び出します。これにより、APIサーバーはリクエストを認証し、適切なデータへのアクセスを許可します。よって、**a**が正解です。 なお、**b**はセキュリティリスクが高く、一般的に推奨されません。**c**は、OAuth認証とは直接関係ありません。**d**のOAuthのエンドポイントは、認証情報の交換に使用され、データ提供サービスのAPI自体ではありません。	238
37	a	Few-shot Promptingは、生成AIに少量のサンプルを与え、それを元に複数の文やアイデアを生成させる技法の一つです。これを利用することで、特定のテーマやトピックに関連する内容を生成させることができます。よって、**a**が正解です。	244
38	c	すでに開発に着手している製品の市場導入準備を進める際には、これまでに収集した既存のデータを効果的に活用し、改善点が見つかれば部分的に戦略を修正する手法が効果的です。逆にすべてのデータを基に一から立て直すことは、時間とリソースの無駄になりかねません。よって、**c**が正解です。	254
39	d	アンケート結果の中で、オンライン広告の効果については言及していません。他の内容はすべて言及があり、その内容を踏まえた仮説として適切です。よって、**d**が正解です。	255

問題番号	解答	解説	該当ページ
40	a	データを取り扱う場合は、不正行為をしないことが極めて重要です。問題のように、健康データの扱いにおいて特に慎重であるべきヘルスケア領域においては、データの偽造や改ざんは不適切な行為であり、データが欠落している事実を透明にすること、正確なデータの取得を試みること、分析における限界や仮定を明確に示すことが重要です。よって、**a**が正解です。	259
41	c	アンケート設計では特に設問数が増えないよう冗長な設問設計は避ける必要があります。**a**は年齢、**b**はジャンル、**d**は時間で重複する設問が存在しています。よって、正解は**c**です。	264
42	c	生成AIは定型的なテキスト生成やデータ分析、書類作成などのタスクを効率化できますが、従業員のパフォーマンス評価のような主観的判断や詳細な個人の評価を含む業務は自動化することが難しいです。パフォーマンス評価は個々の実績、能力、貢献度、そして行動の質的な側面を総合的に評価する必要があり、現在の生成AIではこの複雑さを完全に理解し再現することが困難なため、自動化には適していません。よって、**c**が正解です。	275
43	b	分析プロジェクトにおいてデータを入手する際は、その前に分析に必要になるデータを整理することが重要です。現状の把握や仮説出しを怠り、先にデータを集めようとすると、現在使用することのできるデータありきの仮説になってしまい、適切な課題解決につながらないことがあります。よって、仮説無しに入手可能なデータで分析しようとしている**b**が正解です。	281

問題番号	解答	解説	該当ページ
44	b	大規模言語モデルの使用時には、モデルが生成する内容が常に正確であるとは限らないため、出力結果を鵜呑みにせず、事実確認や二次的な検証を行うことが重要です。ハルシネーションは、現在の大規模言語モデルの限界の一つであり、この問題を根本から解決することは現状では難しい状況です。よって、正解は**b**です。	284
45	a	データの可視化や分析の結果が得られたら、仮説通りの結果になっているか、またそうでない場合はどのような原因が考えられるかを解釈する必要があります。**a**の選択肢にある異常値や外れ値が見つかった時、それらのデータが集計ミスによって発生したものなのか、または適切なものなのかをデータ取得者に確認する必要があり、自身の判断で削除してしまうと本来得られるはずであった貴重な情報を見逃してしまう可能性があるため、不適切であるとわかります。よって、**a**が正解です。	290
46	c	クラウドコンピューティングはプライベートネットワーク内でのみ利用されるという部分が誤りで、実際には、クラウドコンピューティングはプライベートだけでなく、パブリックやハイブリッドなどの形態でも広く利用されています。よって、**c**が正解です。	302

問題番号	解答	解説	該当ページ
47	b	予測的データ分析では、過去の情報を利用して定期的に起きる事象の予測や検知を自動で行うことが可能です。**a**の小売業における四半期の予測の場合は季節変動を捉えることが不可欠であり、むしろ同じ季節同士を比較するほうがより良い結果が得られるでしょう。**c**の公共交通機関の場合は、イベントや天候なども大きく影響する要因として取り込む必要があり、日次での最適化も行われていません。**d**の金融業界における投資ポートフォリオの予測は突発的な市場の変動や経済危機などの予測できない要素が存在し困難なケースが多いです。一方、**b**は、個人の健康改善の可能性に加え、医療費の削減や公衆衛生の向上にも寄与する可能性があります。よって、**b**が正解です。	302
48	c	与えられたデータからヒストグラムの適切な階級数を求める公式として、スタージェスの公式が挙げられます。階級数をk、サンプルサイズをNとすると、スタージェスの公式は$k=\log_2(N)+1$と表せます。これに今回のサンプルサイズである$N=40$を代入すると、$k=6.3$となり、およそ6個の階級を作るとよいことがわかります。実際、データの最小値は51、最大値は600であるため、階級幅の目安はおよそ$(600-51)/6=91.5$となります。よって、最も近いものを選ぶと、**c**が正解となります。 なお、スタージェスの公式は階級数の目安としてよく用いられますが、公式に入れたものが必ず正しいとは限らないため、必ず元データを見て自身で検討することが必要です。	308

問題番号	解答	解説	該当ページ
49	a	相関関係とは、2つの物事の間で一方が変化すれば他方も変化するような関係をいいます。また、因果関係とは2つ以上の物事が原因と結果の関係にあることをいいます。ここでは、運動習慣と2型糖尿病の発症率の関係性に注目しており、運動習慣が2型糖尿病の発症率リスクを低減させる可能性が高く、原因と結果の関係性があると仮説を立てていると推察されるため、相関関係も因果関係もあると考えることができます。よって**a**が正解です。	308
50	b	ここでは、各学部や研究科を層とみなし、それぞれから無作為に学生を選び出して調査する手法を採用しています。これは、「層化無作為抽出法」と呼ばれる手法で、母集団内の異なる層（この場合は学部や研究科）を代表する標本を得るために用いられます。よって、**b**が正解です。	308