CompTIA DataX 認定資格試験 出題範囲

試験番号：DY0-001

試験について

CompTIA DataX認定資格試験は、以下の必要な知識とスキルを証明します。

- データサイエンスのオペレーションとプロセスを理解し、実装する。
- 数学的・統計的手法を適切に適用し、データ処理とクリーニング、統計モデリング、線形代数、微積分の概念の重要性を理解する。
- 機械学習モデルを適用し、ディープラーニングの概念を理解する。
- 適切な分析およびモデリング手法を活用し、正当なモデルの推奨を行う。
- 業界のトレンドやデータサイエンスの専門的な応用についての理解を実証する。

試験開発

CompTIA**の認定資格試験は、ITプロフェッショナルに必要とされるスキルと知識に関して、専門分野のエキスパートによるワークショップ、および業界全体へのアンケートの調査結果に基づいて策定されています。**

CompTIA認定教材の使用に関するポリシー

CompTIA Certifications, LLCは、無許可の第三者トレーニングサイト（通称「ブレインダンプ」）とは提携関係になく、これらが提供するいかなるコンテンツも公認・推薦・容認しません。CompTIAの認定資格試験の受験準備にこのような教材を使用した個人は、CompTIA受験者同意書の規定に基づいて資格認定を取り消され、その後の受験資格を停止されます。CompTIAでは、無許可教材の使用に関する試験実施ポリシーをよりよく理解していただくために、認定資格試験を受験される全員の方にCompTIA認定資格試験実施ポリシーをご一読いただくようご案内しております。CompTIAの認定資格試験を受験するための学習を始める前には、必ずCompTIAが定めるすべてのポリシーをご確認ください。受験者はCompTIA受験者合意書を遵守することが求められます。個々の教材が無許可扱いになるかどうかを確認するには、CompTIA (examsecurity@comptia.org)までメールにてご確認ください。

注意事項

箇条書きで挙げられた項目は、すべての試験内容を網羅するものではありません。この出題範囲に掲載されていない場合でも、各分野に関連する技術、プロセス、あるいはタスクを含む問題が出題される可能性があります。CompTIAでは、提供している認定資格試験の内容に現在必要とされているスキルを反映するため、また試験問題の信頼性維持のため、継続的な試験内容の検討と問題の改訂を行っています。必要な場合、現在の出題範囲を基に試験を改訂する場合があります。この場合、現在の試験に関連する資料・教材等は、継続的にご利用いただくことが可能です。

試験情報

試験番号	DY0-001
問題数	最大90問
出題形式	単一/複数選択、パフォーマンスベーステスト
試験時間	165分
推奨経験	データサイエンティストとして最低5年間の実務経験
合格スコア	合格/不合格の記載のみ、得点表記はなし

試験の出題範囲（試験分野）

下表は、この試験における試験分野（ドメイン）と出題比率の一覧です。

試験分野		出題比率
1.0	数学と統計	17%
2.0	モデリング、分析、結果	24%
3.0	機械学習	24%
4.0	オペレーションとプロセス	22%
5.0	データサイエンスの専門的応用	13%
合計		**100%**

1.0 数学と統計

1.1 与えられたシナリオに基づいて、適切な統計手法や概念を適用することができる。

- **t検定**
- **カイ二乗検定**
- **分散分析(ANOVA)**
- **仮説検証**
- **信頼区間**
- **回帰パフォーマンス指標**
  - R^2
  - 調整R^2
  - 二乗平均平方根誤差(RMSE)
  - F検定
- **ジニ係数**
- **エントロピー**
- **情報利得**
- ***p*値**
- **第一種過誤と第二種過誤**
- **受信者操作特性/曲線下面積(ROC/AUC)**
- **AIC/BIC（赤池情報量規準/ベイズ情報量規準）**
- **相関係数**
  - ピアソン積率相関係数
  - スピアマン順位相関係数
- **混同行列**
  - 分類器のパフォーマンス指標
    - 正確度(Accuracy)
    - 再現率(Recall)
    - 精度(Precision)
    - F1スコア
    - マシューズ相関係数(MCC)
- 中心極限定理
- 大数の法則

1.2 確率と合成モデリングの概念とその使用法を説明することができる。

- **分布**
  - 正規分布
  - 一様分布
  - ポアソン分布
  - *t*分布
  - 二項分布
  - ベキ分布
- **歪度**
- **尖度**
- **分散不均一性と分散均一性の比較**
- **確率密度関数(PDF)**
- **確率質量関数(PMF)**
- **累積分布関数(CDF)**
- **確率**
  - モンテカルロシミュレーション
  - ブートストラップ法
  - ベイズの定理
  - 期待値
- **欠測の種類**
  - ランダムな欠測
  - 完全にランダムな欠測
  - ランダムではない欠測
- **オーバーサンプリング**
- **層別**

1.3 線形代数と微積分の基本概念の重要性について説明することができる。

- **線型代数**
  - 順位
  - スパン
  - トレース
  - 固有値/固有ベクトル
  - 基底ベクトル
  - 単位行列
  - 行列とベクトルの演算
    - 行列の乗算
    - 行列の転置
    - 逆行列
    - 行列の分解
  - 距離メトリック
    - ユークリッド
    - 放射状/ラディアル
    - マンハッタン
    - コサイン
- **微積分**
  - 偏導関数
  - 連鎖律
  - 指数
  - 対数

1.4 さまざまなタイプの時間モデルを比較対照することができる。

- **時系列**
  - 自己回帰(AR)
  - 移動平均(MA)
  - 自己回帰積分移動平均(ARIMA)
- **縦断分析**
- **生存分析**
  - パラメトリック
  - ノンパラメトリック
- **因果推論**
  - DAG（有向非巡回グラフ）
  - Difference-in-differences（差分の差分法）
  - A/Bテスト
  - ランダム化比較試験

2.0 モデリング、分析、結果

2.1 与えられたシナリオに沿って、探索的データ分析(EDA)の方法またはプロセスを利用することができる。

- **一変量解析**
- **多変量解析**
- **オブジェクトの動作と属性の識別**
- **チャートとグラフ**
  - 棒グラフ
  - 散布図
  - 箱ひげ図
  - 折れ線グラフ
  - バイオリン図
  - ヒートマップ
  - 相関プロット
  - ヒストグラム
  - サンキーダイアグラム
  - Q-Qプロット (quantile-quantile plot)
  - 密度プロット
  - 散布図マトリックス
- **フィーチャタイプの識別**
  - カテゴリ変数
  - 離散変数
  - 連続型変数
  - 順序変数
  - 名義変数
  - 二項変数

2.2 与えられたシナリオに基づいて、データの一般的な問題を分析できる。

- **一般的な問題**
  - 疎（スパース）データ
    - 疎行列
    - 疎ベクトル
  - 非線形性
  - 非定常性
  - 遅延の観測(Lagged observations)
  - 差分の観測 (Difference observations)
  - 多重共線性
  - 季節性評価
  - 粒度のずれ
  - 不十分な特徴
  - 多変量の外れ値

2.3 与えられたシナリオに基づいて、データの強化と拡張の技術を適用することができる。

- **フィーチャーエンジニアリング**
- **データ変換**
  - One-Hotエンコーディング
  - ラベルエンコーディング
  - クロスターム
  - 線形化
    - 対数
    - 指数
  - Box-Cox変換
  - 正規化
  - ビニング
  - 比率
  - ピボット
- **ジオコーディング**
- **スケーリング**
- **標準化**
- **追加データソース**
  - データ拡張
  - **データセット**
  - 合成データ

2.4 与えられた設定に基づき、モデル設計の反復プロセスを実施することができる。

- **設計と仕様**
  - 制約
    - 時間
    - リソース
    - 物理的ハードウェア
    - 費用
- **パフォーマンス評価**
  - 統計的指標
  - トレーニング時間とコスト
  - 時間経過に伴う推論パフォーマンス
  - モデル診断プロット
    - 残差vs適合値
- **モデルの選択**
  - 文献レビュー/リテラチャーレビュー
  - ハイパーパラメータのチューニング
  - 実験管理
  - モデル構築の反復
- **要件の妥当性検証**

2.5 与えられたシナリオに基づいて、実験とテストの調査結果を分析し、最終的なモデルの推奨と選択を正当化することができる。

- **ベースラインに対するベンチマーク**
- **従来のプロセスに対するベンチマーク**
- **仕様テストの結果**
- **最終パフォーマンス指標**
- **ビジネス要件を満たす**
  - ビジネスのニーズとウォンツと現実を区別する

2.6 与えられたシナリオに基づいて、結果を解釈し、適切な方法と媒体を通じて伝達することができる。

- **ビジュアライゼーションとレポートの種類**
- **レポートのためのデータ選択**
- **同僚や利害関係者への効果的なコミュニケーションとレポートの考慮事項**
  - ビジネスエグゼクティブの利害関係者のタイプ
  - ビジネスドメインの利害関係者のタイプ
  - 同僚/専門家の利害関係者のタイプ
- **適切なビジュアライゼーション/レポート作成のためのデータタイプ、寸法、および集計レベルを考慮する**
- **意図せず誤解を招くような図表や報告を避ける**
- **チャートのアクセシビリティ**
  - フォントの選択とサイズ
  - 色の選択
  - コンテンツのタグ付け
  - アクセシビリティの有効性
  - 政府規制の影響
- **データとモデルの文書化**
  - コードの文書化
  - データ辞書（データディクショナリー）
  - メタデータ
  - 変更記録

3.0 機械学習

3.1 与えられたシナリオに基づいて、機械学習の基礎概念を適用できる。

- **損失関数**
  - 分散最小化
- **偏り(Bias)と分散(Variance)のトレードオフ**
  - 過剰適合
  - 過少適合
- **変数/特徴の選択**
  - 特徴の重要性
  - 多重共線性
  - 相関行列
  - 分散拡大係数(VIF)
- **クラスの不均衡と緩和**
  - 少数派クラスのオーバーサンプリング
  - 多数派クラスのアンダーサンプリング
  - 合成少数派オーバーサンプリング手法(SMOTE)
- **正則化**
- **交差検証（クロスバリデーション）**
  - *k*-fold交差検証
- **次元の呪い**
- **オッカムの剃刀/パーシモンの法則**
- **サンプル内vsサンプル外**
- **内挿vs外挿**
- **アンサンブルモデル**
- **ハイパーパラメータのチューニング**
  - グリッド探索
  - ランダム探索
- **分類**
  - 二項分類
  - 多クラス（多項式）分類
- **レコメンダシステム**
  - 協調フィルタリング
  - 交互最小二乗法(ALS)
  - 類似度ベース
- **リグレッサー**
- **埋め込み**
- **事後分析モデルの説明可能性**
  - 大域的説明
  - 局所的説明
- **解釈可能なモデル**
- **モデルドリフトの原因**
  - データドリフト
  - 概念/コンセプトドリフト
- **データ漏洩**
  - 転移学習
  - コールドスタート問題

3.2 与えられたシナリオに基づいて、適切な統計的教師あり機械学習の概念を適用できる。

- **線形回帰モデル**
  - 通常の最小二乗法(OLS)
    - 前提条件
  - 重み付き最小二乗法
  - Ridge
  - LASSO
  - 弾性ネット
- **ロジスティック回帰モデル**
  - プロビット
  - ロジット
- **線形判別分析**
- **二次判別分析(QDA)**
- **アソシエーションルール**
  - 確信度/信頼度
  - リフト値
  - 増強
  - 支持度
- ナイーブベイズ

3.3 与えられたシナリオに基づいて、適切なツリーベース教師あり機械学習の概念を適用できる。

- **決定木**
- **ランダムフォレスト**
- **ブースティング**
  - 勾配ブースティング
  - XGBoost
- **ブートストラップ集約（バギング）**

3.4 ディープラーニングに関連する概念を説明できる。

- **人工ニューラルネットワークアーキテクチャ**
  - パーセプトロン
  - 人工神経
  - 多層パーセプトロン
  - 活性化関数
    - ReLU（整流化線形ユニット）
    - シグモイド
    - Tanh関数
    - ソフトマックス
  - レイヤーの種類
    - 入力
    - 非表示
    - プーリング
    - 出力
- **ドロップアウト**
- **バッチ正規化**
- **早期停止**
- **スケジューラ**
- **バックプロパゲーション**
- **ワンショット学習**
- **ゼロショット学習**
- **フューショット学習**
- **ディープラーニングフレームワーク**
  - PyTorch
  - TensorFlow/Keras
  - AutoML（自動機械学習）
- **オプティマイザ**
  - Adamオプティマイザ
  - モーメンタム
  - 平方根平均二乗伝播(RMSprop)
  - 確率的勾配降下法
  - ミニバッチ
- **モデルの種類**
  - 畳み込みニューラルネットワーク(CNN)
  - リカレントニューラルネットワーク(RNN)
  - 長・短期記憶(LSTM)
  - 敵対的生成ネットワーク(GAN)
  - オートエンコーダー
  - トランスフォーマー

3.5 教師なし機械学習に関連する概念を説明できる。

- **クラスタリング**
  - *k*平均法
    - シルエットスコア/エルボー法
  - 階層型
  - ノイズの密度ベース空間クラスタリング分析(DBSCAN)
- **次元削減**
  - 主成分分析(PCA)
  - *t*分布型確率的近傍埋め込み法(*t*-SNE)
  - 一様多様体近似と射影(UMAP)
- ***k*近傍法(KNN)**
- **特異値分解(SVD)**

4.0 オペレーションとプロセス

4.1 さまざまなビジネス機能におけるデータサイエンスの役割を説明できる。

- **コンプライアンス、セキュリティ、プライバシー**
  - 個人を特定できる情報(PII)
  - 専有
  - 機密データの匿名化
  - データの難読化
  - データ使用規制
- **対策、測定基準、重要業績評価指標(KPI)**
- **要件の収集**
  - 費用便益分析に基づく提案を行う
  - ビジネスニーズを最適なソリューションに変換する
  - 関連する適用範囲

4.2 さまざまなタイプのデータを入手するプロセスと目的を説明できる。

- **生成データ**
  - 調査
  - 管理
  - センサー
  - トランザクション
  - 実験
  - データ生成プロセス
- **合成データ**
  - 費用便益
  - 作成プロセス
  - 制限
  - サンプリング
  - 根拠
- **民間/公共データ**
  - 費用便益
  - 可用性
  - ライセンス
  - 規制

4.3 データインジェストとストレージの概念を説明できる。

- **インフラストラクチャ要件**
  - リソースサイジング
  - グラフィック処理装置(GPU)/テンソル・プロセッシング・ユニット(TPU)
- **データの形式**
  - 一般的な形式
    - カンマ区切り値(CSV)
    - JavaScript Object Notation (JSON)
    - Parquet
  - 圧縮フォーマット
  - 構造化ストレージ
  - 半構造化ストレージ
  - 非構造化ストレージ
- **ストリーミング**
- **バッチ処理**
- **パイプラインの実装**
- **オーケストレーション/自動化**
- **永続性**
- **リフレッシュサイクル**
- **アーカイブ**
- **データリネージ**

4.4 与えられたシナリオに基づいて、一般的なデータラングリングの技術を実装することができる。

- **マージ/結合**
  - キーの定義
  - データマッチング
    - 一致率
    - ファジー結合
  - 観測追跡
  - ユニオン
  - 交差
  - 結合の種類
- **クリーニング**
  - 日付/時刻の標準化
  - 正規表現
  - 重複排除
  - 単位変換/標準化
  - コードの欠落
- **データエラー**
  - 特異的(Idiosyncratic)
  - 体系的(Systematic)
- **外れ値**
  - 識別
  - ウインザー化/カットポイント
  - 誤差vs有効データ点
- **データのフラット化**
  - Extensible Markup Language (XML)
  - JSON
- **インピュテーションタイプ**
- **Ground truthラベル**

4.5 与えられたシナリオに基づいて、データサイエンスのライフサイクルを通じてベストプラクティスを実装できる。

- **データサイエンスのワークフローモデル**
  - データマイニングのための業界横断標準プロトコル(CRISP-DM)
  - データマネジメント協会(DAMA)
- **バージョン管理**
  - コード
  - データ
  - ハイパーパラメータ
  - モデル
- **統合開発環境(IDE)**
- **依存関係ライセンス**
- **アプリケーションプログラミングインターフェース(API)によるアクセス**
  - データのアクセスと取得
  - モデルエンドポイント/モデルサービス
- **プロセスの文書化**
  - マークダウン
  - ドックストリング
  - 適切なコードコメント
  - 参照データとドキュメンテーション
- **クリーンなコード手法**
- **単体テストの記述**

4.6 データサイエンスにおけるDevOpsとMLOpsの原則の重要性を説明できる。

- **データのレプリケーション**
- Continuous Integration/Continuous Deployment（CI/CD：**継続的インテグレーション/継続的デプロイメント**）
- **モデルのデプロイメント**
- **コンテナのオーケストレーション**
- **仮想化**
- **コードの分離**
- **モデルのパフォーマンスモニタリング**
- **モデルの検証**
  - オンライン
  - オフライン
  - モデルのA/Bテスト

4.7 さまざまなデプロイ環境を比較対照することができる。

- **コンテナー化**
- **クラウドデプロイ**
- **クラスターデプロイ**
- **ハイブリッドデプロイ**
- **エッジデプロイ**
- **オンプレミスデプロイ**

5.0 データサイエンスの専門的応用

5.1 最適化の概念を比較対照することができる。

- **制約付き最適化**
  - ネットワークトポロジー
    - 巡回セールスマン
  - スケジュール
  - 線形ソルバー
    - シンプレックス法
  - 非線形ソルバー
  - 価格設定
  - リソースの割り当て
  - バンドル
  - バウンダリーケース
- **制約なし最適化**
  - 多腕バンディット
  - 多腕バンディット
  - 局所極大・極小の発見

5.2 自然言語処理(NLP)の概念について、用途と重要性を説明することができる。

- **トークン化/bag of words**
- **単語の埋め込み**
  - *n*-gram
- **TF-IDF（単語頻度ー逆文書頻度）**
- **文書用語行列**
- **編集距離**
- **大規模言語モデル**
  - Word2Vec
  - GloVe
- **テキスト準備**
  - レンマタイゼーション
  - ストップワード
  - 拡張
  - 文字列インデックス
  - ステミング
  - 品詞タグ付け（POSタグ付け）
- **トピックモデリング**
  - 潜在ディリクレ割り当て
- **曖昧性解消**
- **自然言語処理アプリケーション**
  - センチメント分析/感情分析
  - 質疑応答/対話
  - 固有表現認識(NER)
    - 自動タグ付け
  - テキスト生成
  - 一致モデル
  - 音声認識と生成
  - テキスト要約
  - 自然言語理解(NLU)
  - 自然言語生成(NLG)

5.3 コンピュータービジョンの概念について、用途と重要性を説明することができる。

- **光学文字認識**
- **オブジェクト/セマンティックセグメンテーション**
- **オブジェクト検出**
- **追跡**
- **センサーフュージョン**
- **データ拡張**
  - フィルターアプリケーション
  - ローテーション
  - オクルージョン
  - スプリアスノイズ
  - 反転
  - スケーリング
  - 穴
  - マスキング
  - トリミング

5.4 データサイエンスのその他の専門的応用の使用目的を説明することができる。

- グラフ解析/グラフ理論
- ヒューリスティクス
- 貪欲アルゴリズム
- 強化学習
- イベント検出
- 不正検知
- 異常検知
- マルチモーダル機械学習
- エッジコンピューティングの最適化
- 信号処理

CompTIA DataX DY0-001略語リスト

下記はCompTIA DataX DY0-001認定資格試験で使用される略語の一覧です。受験者には、試験準備の一環として、これら用語を復習し、理解することをお勧めします。

略語	詳細説明
AIC-BIC	Akaike Information Criterion - Bayesian Information Criterion
ALS	Alternating Least Squares
ANOVA	Analysis of Variance
API	Application Programming Interface
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the Curve
CDF	Cumulative Distribution Function
CI/CD	Continuous Integration/Continuous Deployment
CNN	Convolutional Neural Network
CRISP-DM	Cross-industry Standard Process for Data Mining
CSV	Comma-separated Values
DAG	Directed Acyclic Graph
DAMA	Data Management Association
DBSCAN	Density-based Spatial Clustering Analysis with Noise
EDA	Exploratory Data Analysis
FFNN	Feed Forward Neural Network
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
GUID	Globally Unique Identifier
HDBSCAN	Hierarchical Density-based Spatial Clustering Analysis with Noise
HPC	High-performance Computing
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
IP	Internet Protocol
JSON	JavaScript Object Notation
KNN	*k*-Nearest Neighbors
KPI	Key Performance Indicator
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-term Memory
MA	Moving Average
MAC	Media Access Control
MCC	Matthews Correlation Coefficient
ML	Machine Learning
NER	Named-entity Recongition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OLS	Ordinary Least Squares
OS	Operating System
PCA	Principal Component Analysis
PDF	Probability Density Function
PII	Personally Identifiable Information
PIP	Preferred Installer Program
POS	Part of Speech
QDA	Quadratic Discriminant Analysis
Q-Q	Quantile-Quantile
RegEX	Regular Expression
ReLU	Rectified Linear Unit
REST	Representational State Transfer
RPC	Remote Procedure Call
RMS	Root Mean Square
RMSE	Root Mean Square Error
RMSprop	Root Mean Square Propagation
RNN	Recurrent Neural Network
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
RPC	Remote Procedure Call

略語	詳細説明
RSS	Residual Sum of Squares
SARIMA	Seasonal Auto-regressive Integrated Moving Average
SMOTE	Synthetic Minority Oversampling Technique
SOAP	Simple Object Access Protocol
SVD	Singular Value Decomposition
SVM	Support Vector Machines
SVN	Subversion
TF-IDF	Term Frequency Inverse Document Frequency
TPU	Tensor Processing Unit
t-SNE	*t*-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VIF	Variance Inflation Factor
WSL	Windows Subsystem for Linux
XML	Extensible Markup Language

CompTIA DataX DY0-001ハードウェアとソフトウェア一覧

本リストはDataX DY0-001認定試験の受験準備としてお役立ていただくためのハードウェアとソフトウェアのリストです。トレーニングを実施している企業でも、トレーニングの提供に必要な実習室コンポーネントを作成したい場合に役立ちます。各トピックに箇条書きで挙げられた項目は例であり、すべてを網羅するものではありません。

機材
- CUDA互換GPU搭載ワークステーション
- クラウドプロバイダー上のGPU

ソフトウェア
- Linuxカーネルベースのオペレーティングシステム（推奨）
- Windowsオペレーティングシステム
  - リージョナルパック
  - ユニコード
  - Linux用Windowsサブシステム(WSL)
  - Dockerデスクトップ
- CoderPad
- PythonまたはR
  - 関連パッケージ（ビジュアライゼーション、モデリング、クリーニング、機械学習）
- ノートブック環境/ツールセット
- Visual Studio Code
- Git

その他
- 大規模データセット
- 小規模データセット
- 各種データセット