Date Issued: 12.21.2023
Date Modified: 3.21.2024
Version: 1.3
Approved by: SME COMMITTEE

CertNexus Certified Data Science Practitioner™ (CDSP) Exam DSP-210

Exam Information

Exam Objective Statement

The exam will certify that the successful candidate has the knowledge, skills, and abilities required to answer questions by collecting, wrangling, and exploring datasets, applying statistical models and artificial-intelligence algorithms, to extract and communicate knowledge and insights.

Candidate Eligibility

The Certified Data Science Practitioner™ (CDSP) exam requires no application fee, supporting documentation, or other eligibility verification measures for you to be eligible to take the exam. An exam voucher will come bundled with your training program or can be purchased separately here. Once purchased, you will receive more information about how to register for and schedule your exam through Pearson Vue. You can also purchase a voucher directly through Pearson Vue. Once you have obtained your voucher information, you can register for an exam time here. By registering, you agree to our Candidate Agreement included here.

Exam Prerequisites

There are no formal prerequisites to register for and schedule an exam. Successful candidates will possess the knowledge, skills, and abilities as identified in the domain objectives in this blueprint. It is also strongly recommended that candidates possess the following knowledge, skills, and abilities:

- A working level knowledge of programming languages such as Python® and R
- Proficiency with a querying language
- Strong communication skills
- Proficiency with statistics and linear algebra
- Demonstrate responsibility based upon ethical implications when sharing data sources
- Familiarity with data visualization

You can obtain this level of skill and knowledge by taking the following course offerings, which is available through training providers located around the world, or by attending an equivalent third-party training program:

- Introduction to Programming with Python®
- Advanced Programming Techniques with Python®
- Using Data Science Tools in Python®
- R Programming for Data Science
- DSBIZ™ (Exam DSZ-210)
- DEBIZ™ (Exam DEB-110): Data Ethics for Business Professionals *or* Certified Ethical Emerging Technologist™ (CEET)

Exam Specifications

Number of Items: 90, of which 75 count toward your score

Passing Score: TBA

Duration: 120 minutes (**Note**: Published exam times include the 10 minutes you are allotted for reading and signing the Candidate Agreement and reviewing exam instructions.)

Exam Options: Online through Pearson OnVUE or in person at Pearson VUE test centers.

Item Formats: Multiple Choice/Single Response

Exam Description

Target Candidate:

The Certified Data Science Practitioner™ (CDSP) exam is designed for professionals across different industries seeking to demonstrate the ability to gain insights and build predictive models from data.

The Certified Data Science Practitioner™ (CDSP) exam will test them on the following domains with the following distribution:

Domain	# of Items
1.0 Defining the need to be addressed through the application of data science	**6**
2.0 Extracting, Transforming, and Loading Data	**16**
3.0 Performing exploratory data analysis	**23**
4.0 Building models	**17**
5.0 Testing models	**4**
6.0 Operationalizing the pipeline	**5**
7.0 Communicating findings	**4**
Total	**75**

The information that follows is meant to help you prepare for your certification exam. This information does not represent an exhaustive list of all the concepts and skills that you may be tested on during your exam. The exam domains, identified previously and included in the objectives listing, represent the large content areas covered in the exam. The objectives within those domains represent the specific tasks associated with the job role(s) being tested. The information beyond the domains and objectives is meant to provide examples of the types of concepts, tools, skills, and abilities that relate to the corresponding domains and objectives. All of this information represents the industry-expert analysis of the job role(s) related to the certification and does not necessarily correlate one-to-one with the content covered in your training program or on your exam. We strongly recommend that you independently study to familiarize yourself with any concept identified here that was not explicitly covered in your training program or products.

Objectives

Domain 1.0 Defining the need to be addressed through the application of data science

Objective 1.1 Identify the project scope

- Identify project specifications, including objectives (metrics/KPIs) and stakeholder requirements
- Identify mandatory deliverables, optional deliverables
- Determine project timeline
- Identify project limitations (time, technical, resource, data, risks)

Objective 1.2 Understand challenges

- Understand terminology
  - Milestone
  - POC (Proof of concept)
  - MVP (Minimal Viable Product)
- Become aware of data privacy, security, and governance policies
  - GDPR
  - HIPAA
  - California Privacy Act
- Obtain permission/access to stakeholder data
- Ensure appropriate voluntary disclosure and informed consent controls in place

Objective 1.3 Classify a question into a known data science problem

- Identify references relevant to the data science problem
  - Optimization problem
  - Forecasting problem
  - Regression problem

    - Classification problem
    - Segmentation/Clustering problem
- Identify data sources and type
    - Structured/unstructured
    - Image
    - Text
    - Numerical
    - Categorical
- Select modeling type
    - Regression
    - Classification
    - Forecasting
    - Clustering
    - Optimization
    - Recommender systems

Domain 2.0 Extracting, Transforming, and Loading Data

Objective 2.1 Gather data sets

- Read data
    - Write a query for a SQL database
    - Write a query for a NoSQL database
    - Read data from/write data to cloud storage solutions
        - AWS S3
        - Google Storage Buckets
        - Azure Data Lake
- Become aware of first-, second-, and third-party data sources
    - Understand data collection methods
    - Understand data sharing agreements, where applicable
- Explore third-party data availability
    - Demographic data
    - Bloomberg
- Collect open-source data
    - Use APIs to collect data
    - Scrape the web
- Generate data assets
    - Dummy or test data
    - Randomized data
    - Anonymized data
    - AI-generated synthetic data

Objective 2.2 Clean data sets

- Identify and eliminate irregularities in data (e.g., edge cases, outliers)

        - Nulls
        - Duplicates
        - Corrupt values
    - Parse the data
    - Check for corrupted data
    - Correct the data format
    - Deduplicate data
    - Apply risk and bias mitigation techniques
        - Understand common forms of ML bias
            - Sampling bias
            - Measurement bias
            - Exclusion bias
            - Observer bias
            - Prejudicial bias
            - Confirmation bias
            - Bandwagoning
        - Identify the sources of bias
            - Sources of bias include data collection, data labeling, data transformation, data imputation, data selection, and data training methods
            - Use exploratory data analysis to visualize and summarize the data, and detect outliers and anomalies
            - Assess data quality by measuring and evaluating the completeness, correctness, consistency, and currency of data
            - Use data auditing techniques to track and document the provenance, ownership, and usage of data, and applied data cleaning steps
        - Mitigate the impact of bias
            - Apply mitigation strategies such as data augmentation, sampling, normalization, encoding, validation
        - Evaluate the outcomes of bias
            - Use methods such as confusion matrix, ROC curve, AUC score, and fairness metrics
        - Monitor and improve the data cleaning process
            - Establish or adhere to data governance rules, standards, and policies for data and the data cleaning process

Objective 2.3 Merge and load data sets

- Join data from different sources
    - Make sure a common key exists in all datasets
    - Unique identifiers

- Load data
    - Load into DB
    - Load into dataframe
    - Export the cleaned dataset
    - Load into visualization tool
- Make an endpoint or API

Objective 2.4 Apply problem-specific transformations to data sets

- Apply word vectorization or word tokenization
    - Word2vec
    - TF-IDF
    - Glove
- Generate latent representations for image data

Domain 3.0 Performing exploratory data analysis

Objective 3.1 Examine data

- Generate summary statistics
- Examine feature types
- Visualize distributions
- Identify outliers
- Find correlations
- Identify target feature(s)

Objective 3.2 Preprocess data

- Identify missing values
- Make decisions about missing values (e.g., imputing method, record removal)
- Normalize, standardize, or scale data

Objective 3.3 Carry out feature engineering

- Apply encoding to categorical data
    - One-hot encoding
    - Target encoding
    - Label encoding or Ordinal encoding
    - Dummy encoding
    - Effect encoding
    - Binary encoding
    - Base-*N* encoding
    - Hash encoding
- Split features
    - Text manipulation
        - Split
        - Trim
        - Reverse
    - Manipulate data
    - Split names

        - Extract year from title
    - Convert dates to useful features
    - Apply feature reduction methods
        - PCA
        - *t*-SNE
        - Random forest
        - Backward feature elimination
        - Forward feature selection
        - Factor analysis
        - Missing value ratio
        - Low-variance filter
        - High-correlation filter
        - SVD
        - False discovery rate
        - Feature importance methods

Domain 4.0 Building models

Objective 4.1 Prepare data sets for modeling

- Decide proportion of data set to use for training, testing, and (if applicable) validation
- Split data to train, test, and (if applicable) validation sets, mitigating data leakage risk

Objective 4.2 Train models

- Define models to try
    - Regression
        - Linear regression
        - Random forest
        - XGBoost
    - Classification
        - Logistic regression
        - Random forest classification
        - XGBoost classifier
        - naïve Bayes
    - Forecasting
        - ARIMA
    - Clustering
        - *k*-means
        - Density-based methods
        - Hierarchical clustering
- Train model or pre-train or adapt transformers
- Tune hyperparameters, if applicable

    - Cross-validation
    - Grid search
    - Gradient decent
    - Bayesian optimization

Objective 4.3 Evaluate models

- Define evaluation metric
- Compare model outputs
    - Confusion matrix
    - Learning curve
- Select best-performing model
- Store model for operational use
    - MLflow
    - Kubeflow

Domain 5.0 Testing models

Objective 5.1 Test hypotheses

- Design A/B tests
    - Experimental design
        - Design use cases
        - Test creation
    - Statistics
- Define success criteria for test
- Evaluate test results

Domain 6.0 Operationalizing the pipeline

Objective 6.1 Deploy pipelines

- Build streamlined pipeline (using dbt, Fivetran, or similar tools)
- Implement confidentiality, integrity, and access control measures
- Put model into production
    - AWS SageMaker
    - Azure ML
    - Docker
    - Kubernetes
- Ensure model works operationally
- Monitor pipeline for performance of model over time
    - MLflow
    - Kubeflow
    - Datadog
- Consider enterprise data strategy and data management architecture to facilitate the end-to-end integration of data pipelines and environments
    - Data warehouse and ETL processes
    - Data lake and ETL processes

        - Data mesh, microservices, and APIs
        - Data fabric, data virtualization, and low-code automation platforms

Domain 7.0 Communicating findings

Objective 7.1 Report findings

- Implement model in a basic web application for demonstration (POC implementation)
    - Web frameworks (Flask, Django)
    - Basic HTML
    - CSS
- Derive insights from findings
- Identify features that drive outcomes (e.g., explainability, interpretability, variable importance plot)
- Show model results
- Generate lift or gain chart
- Ensure transparency and explainability of model
    - Use explainable methods (e.g., intrinsic and post-hoc)
        - Visualization
        - Feature importance analysis
        - Attention mechanisms
        - Avoiding black-box techniques in model design
        - Explainable AI (XAI) frameworks and tools
            - SHAP
            - LIME
            - ELI5
            - What-If Tool
            - AIX360
            - Skater
            - Et al
    - Document the model lifecycle
        - ML design and workflow
        - Code comments
        - Data dictionary
        - Model cards
        - Impact assessments
    - Engage with diverse perspectives
        - Stakeholder analysis
        - User testing
        - Feedback loops
- Participatory design

Objective 7.2 Democratize data

- Make data more accessible to a wider range of stakeholders

- Make data more understandable and actionable for nontechnical individuals
    - Implement self-service data/analytics platforms
- Create a culture of data literacy
    - Educate employees on how to use data effectively
    - Offer support and guidance on data-related issues
    - Promote transparency and collaboration around data

Recertification Requirements

The *Certified Data Science Practitioner™ (CDSP)* certification is valid for 3 years from the date that it is initially granted. In order to maintain a continuously valid certification, candidates can recertify via one of the following options:

1. Retake the most recent version of the exam before their certification expires.
2. Earn and submit enough continuing education credits (CECs) to recertify without retaking the exam.

Certified Data Science Professional (CDSP) Acronyms

Acronym	Expanded Form
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
ETL	Extract, Transform, and Load
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
KNN	K-Nearest Neighbors
KPI	Key Performance Indicator
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
PCA	Principal Component Analysis
RBF	Radial Basis Function Kernel
REST API	Representational State Transfer Application Programming Interface
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
SVM	Support Vector Machines
SQL	Structured Query Language
TF-IDF	Term Frequency–Inverse Document Frequency
CSV	Comma-separated Values