PUC IT

CITIZEN DATA SCIENTISTS USING DATAIKU

Subject Code: STCDS403

Total Hours: 40 Credits: 4

Course Learning Objectives (CLO)

This 5-level certification is designed to provide IT professionals with a guide about Dataiku from the initial steps through progressive mastery of the platform - including an introduction to the Dataiku platform, understanding the visual tools in Dataiku for building machine learning models, building data pipelines, code integrations, and learning proper MLOps practices and how to implement them.

UNIT 1: Core Designer **[7 hours]**

- Understand the core concepts of Dataiku (project, Flow, datasets ..)
- Build simple workflows within Dataiku with the visual tools
- Share your results via charts and dashboards
- Learn best practices through the design of a Dataiku project - without the ML part

Basics 101: Create the Project, Create the Dataset, Explore Your Data, and Course Checkpoint.

Basics 102: Prepare Your Data, Interactive Visual Statistics, Group the Data, Explore the Flow, and Course Checkpoint.

Basics 103: Data Enrichment with the Join Recipe, Dataiku Lab, Reporting Tools, and Course Checkpoint.

Visual Recipes 101: Distinct Recipe, Group By Recipe, Join Recipe, Pivot Recipe, Prepare Recipe, Sample/Filter Recipe, Sort Recipe, Split Recipe, Stack Recipe, Top N Recipe, Window Recipe, Fuzzy Join Recipe, Hands-On: Airport Traffic, and Course Checkpoint.

Integration With SQL Databases: Introduction to Integration with SQL Databases, Technical Prerequisites, Connecting to PostgreSQL, Quiz.

Dataiku DSS & SQL: Introduction to DSS & SQL, Move data to DB, SQL Code Recipes, In-Database Charts, SQL Notebooks, and Course Checkpoint.

Core Designer Certificate

UNIT 2: ML Practitioner **[8 hours]**

- Create, evaluate, and tune ML Models with Dataiku's Visual ML tool
- Deploy your model from Lab to Flow and apply it
- Understand and explain your model with Responsible AI tools
- Analyze your data with the Interactive Stats tool

Machine Learning Basics: Create the Model, Evaluate the Model, Tune the Model, Explainable AI, and Course Checkpoint.

Scoring Basics: Deploy the Model, Scoring Data, Model Lifecycle Management, and Course Checkpoint.

Interactive Visual Statistics: The Interactive Statistics Interface, Univariate, and Bivariate Analysis Fit Curves and Distributions, Correlation Matrix, Principal Component Analysis, Statistical Testing, and Course Checkpoint.

Intro to Machine Learning: Introduction to Machine Learning, Predictive Modeling, Prediction: Regression, Prediction: Classification, Clustering, Course Checkpoint.

Partitioned Models: Partitioning, Partitioned Models, training partitioned Models, course checkpoint.

NLP - The Visual Way: introduction to NLP, Preparing Text Data, Handling Text Features for ML, Course Checkpoint.

Time Series Basics: Introduction to Time Series basics, Data Types, and Formats, Components, Objectives of Time series analysis, and Course checkpoint.

Time Series Preparation: Time Series Preparation, Resampling, Interval Extraction, Windowing, Extrema Extraction, Course checkpoint.

ML Practitioner certification

UNIT 3: Advanced Designer **[10 hours]**

- Create advanced data pipeline

- Optimize your flow with variables, partitions, and reusable items
- Automate your data pipelines with the scenarios and implement rules with the metrics and checks
- Learn about Dataiku's extensibility with the plugins

Variables 101: Introduction to Variables, Coding Variables, and Course Checkpoint.

Visual Recipies 102: Common steps in Recipes, Window Recipe (Advanced), Prepare Recipe - Advanced Formula & Regex, Pivot Recipe (Advanced), Top N Reicpe (Advanced), and Course Checkpoint.

Plugin Store: Plugins in Dataiku, Plugin store usage, Hands-On Plugin Store, and Course checkpoint.

Flow Views & Actions: Flow Views, Tags, & More Views, Schema Propagation & consistency checks, Connection changes & Flow Item Reuse, Dataset Building Strategies, Flow Actions, and Course Checkpoint.

Automation: Metrics & Checks, Scenarios, Custom Metrics, Checks & Scenarios, Course Checkpoint.

Dataiku Application Tutorials: Introduction to Dataiku APplications, Creating a visual application, creating an application as a recipe.

Advanced Partitioning: Partitioning, Running Jobs with partitioned Datasets, Redispatching, and collection partitions, advanced partitioning: FIle based using partition Redispatch, Advanced partitioning: Column-Based (SQL Based), Partitioning Concepts, Partitioning in a Scenario, and Course checkpoint.

Advanced Designer Certificate

UNIT 4: Developer **[8 hours]**

- Integrate code into your Dataiku projects.
- Effectively manage code across multiple projects and instances.
- Connect to Dataiku from your preferred IDEs and drive Dataiku through APIs.
- Create plugins that your non-coding colleagues can use to enhance their work in Dataiku.

Code in Dataiku: Learn how to write, explore, run, and debug code in Dataiku using the languages and tools of your choice.

Shared Code: Learn about the most common ways you can share code in Dataiku including project libraries, notebooks, and code samples.

Custom ML Models: Learn to create and use custom ML models in Dataiku's visual ML interface.

Variables for Coders: Master the concept of project variables.

Visualization: Learn how to create custom code visualizations with code such as webapps and static insights, and share them with other users in Dataiku.

Managed Folders: Get started using managed folders in Dataiku.

Dataiku APIs: Get started interacting with Dataiku objects and instances through code.

Plugin Development: Learn to develop plugins, distribute them, and collaborate on plugin development.

Developer Certification

UNIT 5: MLOps Practitioner **[7 hours]**

- Identify project refactoring and documentation techniques to perform before deploying to production
- Perform batch deployment, monitor, and update projects on the Automation node
- Design, deploy, and monitor API services

Production Concepts: What is involved in MLOps, Six components of model development that impact MLOps, ML Model Packaging, Gain Control of MLOps Processes, Govern for MLOps, Monitoring, and Feedback Loop.

Preparing for Production: Automation Best Practices, Pipeline Optimization Best Practices, Workflow Documentation Best Practices, and Model Monitoring Basics.

Projects in Production: Preparation of Automation Node, Project Deployer, Batch Deployer, Automatically Updating Deployments.

Real-Time APIs: Real-Time API Deployment, Create API Service, Deploy and Monitor API Services.

MLOps Practitioner certification

Course Outcomes: On completion of this course, students will have:

- Proficiency in using the Dataiku DSS platform to build end-to-end data pipelines
- Data cleaning and transformation skills to prepare data for analysis
- Collaboration skills to work effectively with others on the platform
- Communication skills to effectively present and share your findings with stakeholders
- Machine learning skills to build predictive models using various techniques
- Ability to build and deploy machine learning models in a production environment using the Dataiku MLOps framework
- Experience working with cloud platforms such as AWS and Azure
- Understanding of core concepts of Dataiku (project, Flow, datasets, etc.)
- Creation of simple workflows and sharing results via charts and dashboards
- Analyzing data with the Interactive Stats tool and creating advanced data pipelines
- Automating data pipelines with scenarios and implementing rules with metrics and checks
- Management of code across projects and instances, and connection to Dataiku through APIs
- Identification of project refactoring and documentation techniques to perform before deploying to production
- Performance of batch deployment, monitoring, and updating of projects on the Automation node
- Design, deployment, and monitoring of API services.

SKILL-BASED EXERCISE (SBE):

Note: - These Projects/activities are only indicative; the faculty member can innovate

Assignments/ Mini Projects on: -

- Process, Study, and Analyse Credit Card data to map regions with high credit card fraud.
- Building a Credit Card fraud detection system using Dataiku's auto ML
- Enrich the Credit card data and perform by using Dataiku visual recipes and automate the process using scenarios.
- Build a Dataiku visual application that can be used by a user to pass credit card data and get results if the transaction is fraudulent or not. Send the results to the user by email.
- Deploy ML Model into Production using Dataiku deployer node.
- Build a Dataiku API which can be used to access a Dataiku Project flow from external programs.
- Build a Dataiku Project to classify x-ray images using Dataiku image processing capabilities.
- Build a Dataiku Project to forecast Bitcoin Prices every 15 mins.

References:

1. Dataiku: https://www.dataiku.com/
2. Dataiku Learning Paths: https://academy.dataiku.com/page/learning-paths
3. Core Designer: Basics 101, Basics 102, Basics 103, Visual Recipes 101, Integration With SQL Databases, Dataiku DSS & SQL, Core Designer Certificate
4. ML Practitioner: Machine Learning Basics, Scoring Basics, Interactive Visual Statistics, Intro to Machine Learning, Partitioned Models, NLP - The Visual Way, Time Series Basics, Time Series Preparation, ML Practitioner certification
5. Advanced Designer: Variables 101, Visual Recipes 102, Plugin Store, Flow Views & Actions, Automation, Dataiku Application Tutorials, Advanced Partitioning, Advanced Designer Certificate.
6. Developer: Code in Dataiku, Shared Code, Custom ML Models, Variables for Coders, Visualization, Managed Folders, Dataiku APIs, Plugin Development, and Developer Certification.
7. MLOps Practitioner: Production Concepts, Preparing for Production, Projects in Production, Real-Time APIs, MLOps Practitioner certification