AI CERTs™

AI+ Data ™ (5 Days)

Program Detailed Curriculum

Executive Summary

The AI+ Data certification equips professionals with vital skills for data science. It covers key concepts like Data Science Foundations, Statistics, Programming, and Data Wrangling. Participants delve into advanced topics such as Generative AI and Machine Learning, preparing them for complex data challenges. The program includes a hands-on capstone project focusing on Employee Attrition Prediction. Emphasis is placed on Data-Driven Decision-Making and Data Storytelling for actionable insights. Personalized mentorship, immersive projects, and cutting-edge resources ensure a transformative learning journey, preparing individuals for success in AI and data science.

Course Prerequisites

- Basic knowledge of computer science and statistics (beneficial but not mandatory)
- Keen interest in data analysis
- Willingness to learn programming languages such as Python and R

Module 1

Foundations of Data Science

1.1 Introduction to Data Science

- **What is Data Science?:** Explores the fundamentals of data science, including its methodologies, tools, and applications, providing a foundational understanding of this interdisciplinary field.
- **Importance of Data Science:** Examines the significance of data science in modern society, highlighting its role in driving innovation, decision-making, and competitive advantage across industries.

1.2 Data Science Life Cycle

- **Understanding the Business Problem:** Focuses on identifying and defining business challenges, ensuring data science projects address organizational objectives effectively for impactful solutions.
- **Data Preparation:** Equips with skills to collect, clean, and preprocess data, laying the foundation for accurate analysis and effective model building.
- **Exploratory Data Analysis (EDA):** Explores data visually and statistically to understand patterns, relationships, and anomalies, guiding further analysis and model development.
- **Modeling the Data:** Introduces various machine learning algorithms and techniques to build predictive models, optimizing performance and accuracy for solving business problems.
- **Evaluating the Model:** Covers methods to assess model performance and reliability, ensuring that developed models meet business requirements and provide actionable insights.
- **Deploying the Model:** Guides through the process of implementing models into production environments, enabling organizations to derive value from data-driven insights effectively.

1.3 Applications of Data Science

- **Examples of Data Science Applications Across Various Industries:** Explores how data science drives innovation and optimization in diverse sectors, showcasing real-world applications and their impact on industry.

Module 2

Foundations of Statistics

2.1 Basic Concepts of Statistics

- **Descriptive Statistics:** Covers methods to summarize and visualize data, providing insights into its central tendencies, variability, and distribution for initial analysis.
- **Inferential Statistics:** Explores techniques to draw conclusions and make predictions about populations based on sample data, enabling informed decision-making and hypothesis testing.

2.2 Probability Theory

- **Probability Distributions:** Explores various probability distributions, including normal, binomial, and Poisson, to model uncertainty and analyze random phenomena in diverse contexts.
- **Central Limit Theorem:** Examines the central limit theorem, demonstrating how it facilitates the use of sample statistics to approximate population parameters reliably.

2.3 Statistical Inference

- **Hypothesis Testing:** Covers procedures to assess the validity of hypotheses using sample data, enabling informed decisions in research and decision-making processes.
- **Confidence Intervals:** Explores methods to estimate population parameters with uncertainty, providing intervals that likely contain the true parameter value with specified confidence levels.

Module 3

Data Sources and Types

3.1 Types of Data

- **Structured, Semi-structured, and Unstructured Data:** Explores various types of data formats and their characteristics, enabling effective management and analysis in diverse data environments.

3.2 Data Sources

- **Databases, APIs, Web Scraping:** Explores methods to access and retrieve data from databases, application programming interfaces (APIs), and websites, facilitating data acquisition for analysis.

3.3 Data Storage Technologies

- **Relational Databases (SQL), NoSQL Databases (MongoDB):** Contrasts relational and NoSQL databases, covering SQL querying and MongoDB operations, enabling proficiency in managing diverse data storage systems.
- **Hands-on Exercise:** Querying Structured Data and Handling Semi-structured Data: Provides practical experience in querying relational databases with SQL and manipulating semi-structured data using appropriate tools and techniques.

Module 4

Programming Skills for Data Science

4.1 Introduction to Python for Data Science

- **Basics of Python Programming:** Covers Python syntax, data types, and variables, laying the foundation for programming proficiency in data science tasks.
- **Python Libraries for Data Science:** Explores essential Python libraries such as NumPy, Pandas, and Matplotlib for data manipulation and visualization tasks in data science projects.

4.2 Introduction to R for Data Science

- **Basics of R Programming:** Introduces R programming fundamentals, including vectors, data frames, and functions, providing a basis for statistical analysis and visualization in R.
- **R Libraries for Data Science:** Examines key R libraries like dplyr and ggplot2, enhancing data manipulation and visualization capabilities for effective data analysis and presentation.
- **Hands-on Exercise:** Data Manipulation and Visualization with Python and R Libraries: Provides practical experience in manipulating and visualizing data using Python and R libraries, reinforcing programming skills in data science contexts.

Module 5

Data Wrangling and Preprocessing

5.1 Data Imputation Techniques

- **Overview of Data Imputation:** Explore the significance of data imputation in filling missing values, ensuring data completeness and integrity for robust analysis.
- **Overview of Missing Data:** Provides an understanding of missing data, its implications, and methods to address it, crucial for accurate and reliable data analysis.
- **Overview of Imputation:** Explore techniques for filling missing data in datasets, including mean imputation, interpolation, and advanced methods.
- **Common Imputation Techniques:** Covers various data imputation methods including mean, median, mode, KNN, and regression imputation, equipping with tools to handle missing data effectively.

5.2 Handling Outliers and Data Transformation

- **Identifying and Handling Outliers:** Explores methods to detect and manage outliers, ensuring data integrity and reliability in statistical analysis and modeling.
- **Data Transformation Techniques:** Normalization, Standardization: Covers normalization and standardization methods to scale and transform data, facilitating accurate comparisons and analysis in data science tasks.
- **Hands-on Exercise:** Data Cleaning, Imputation, and Preprocessing with Python and R Libraries: Provides practical experience in cleaning, imputing, and preprocessing data using Python and R libraries.

Module 6

Exploratory Data Analysis (EDA)

6.1 Introduction to EDA

- **Purpose and Goals of Exploratory Data Analysis:** Explores the objectives and methods of EDA, enabling effective data exploration to uncover patterns, anomalies, and insights.
- **Common Techniques: Summary Statistics, Data Visualization:** Covers essential techniques such as summary statistics and data visualization to gain insights and make informed decisions from data exploration.

6.2 Data Visualization

- **Types of Visualizations: Histograms, Scatter Plots, Box Plots:** Explores various visualization types, including histograms, scatter plots, and box plots, to represent and analyze different data distributions effectively.
- **Choosing the Right Visualization for Different Types of Data:** Guides in selecting appropriate visualizations based on data types, ensuring clear communication and interpretation of insights.
- **Hands-on Exercise:** Creating Visualizations using Python's Matplotlib and Seaborn, and R's ggplot2

Module 7

Generative AI Tools for Deriving Insights

7.1 Introduction to Generative AI Tools

- **Overview of Generative AI Techniques:** Explores autoencoders, GANs, and VAEs, enabling understanding of how these techniques generate new data samples, images, and text.
- **Hands-on Exercise for Various Gen AI Tools:** Explore Gen AI tools through hands-on exercises, covering Plotly, Seaborn, and more for effective data visualization and analysis.

7.2 Applications of Generative AI

- **Application in Data Synthesis, Augmentation, and Anomaly Detection:** Explores how generative AI techniques are applied in creating synthetic data, augmenting datasets, and detecting anomalies effectively.

Module 8

Machine Learning Refresher

8.1 Introduction to Supervised Learning Algorithms

- **Simple Linear Regression:** Understand and apply the basic concepts of linear regression to analyze relationships between two variables.
- **Multiple Linear Regression:** Extend regression analysis to multiple predictors, mastering techniques for modeling complex relationships in datasets.
- **Polynomial Regression:** Learn to capture nonlinear relationships between variables using polynomial regression models.
- **Logistic Regression:** Dive into binary classification problems, mastering logistic regression techniques for predicting categorical outcomes.
- **K-Nearest Neighbors (KNN):** Explore the KNN algorithm for classification and regression tasks, leveraging proximity-based learning.

- **Decision Tree:** Understand decision tree structures and algorithms, mastering techniques for classification and regression tasks.
- **Support Vector Machine (SVM):** Learn to classify data points by finding the optimal hyperplane, mastering SVM for both linear and nonlinear classification problems.
- **Naive Bayes Classification:** Explore the probabilistic Naive Bayes algorithm for classification tasks, mastering techniques for text and categorical data.

8.2 Introduction to Unsupervised Learning

- **Types of Unsupervised Learning:** Uncover various unsupervised learning methods, understanding their applications and differences in extracting patterns from data without explicit guidance.

8.3 Different Algorithms for Clustering

- **K – Means Clustering:** Master K-Means Clustering, a popular unsupervised learning algorithm, to partition data into clusters based on similarity, ideal for segmentation and pattern recognition tasks.
- **Hierarchical Clustering:** Understand hierarchical grouping methods, analyzing data structures through dendrogram visualization, and applying clustering techniques to uncover patterns in datasets.

8.4 Association Rule Learning

- **Hands On:** Experience hands-on exercises in Association Rule Learning, delving into practical applications and techniques for deriving valuable insights from data sets.

Module 9

Advance Machine Learning

9.1 Ensemble Learning Techniques

- **Overview of Ensemble Learning:** Introduces ensemble learning techniques, combining multiple models for improved prediction accuracy and robustness in various machine learning tasks.
- **Bagging (Bootstrap Aggregating):** Explores bagging, a technique to reduce variance by training multiple models on bootstrap samples of the dataset.
- **Random Forest with Code and Real-life Example:** Dives into Random Forest, a popular ensemble method, with implementation and real-life case studies, demonstrating its effectiveness.
- **Boosting (AdaBoost and Gradient Boosting):** Covers boosting algorithms such as AdaBoost, Gradient Boosting, and XGBoost, enhancing model performance through sequential model training and correction.
- **XGBoost with Code and Real-life Example:** Focuses on XGBoost, a powerful gradient boosting algorithm, with hands-on coding exercises and real-world applications to illustrate its capabilities.
- **Stacking:** Examines stacking, a meta-ensemble technique, combining predictions from multiple base models to improve overall performance.
- **Ensemble Learning Applications and Case Studies:** Explores practical applications of ensemble learning across domains, showcasing case studies to illustrate its effectiveness in solving real-world problems.

9.2 Dimensionality Reduction

- **Introduction to Dimensionality Reduction:** Provides an overview of dimensionality reduction techniques, aiming to reduce the number of features while preserving important information.
- **Principal Component Analysis (PCA):** Explores PCA, a linear dimensionality reduction technique, to transform high-dimensional data into a lower-dimensional space while preserving variance.

- **Linear Discriminant Analysis (LDA):** Covers LDA, a supervised dimensionality reduction technique, focusing on maximizing class separability for classification tasks.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** Investigates t-SNE, a non-linear dimensionality reduction technique, emphasizing visualization of high-dimensional data clusters in lower-dimensional space.
- **Autoencoders for Dimensionality Reduction:** Explores autoencoders, neural network-based models, for unsupervised dimensionality reduction, learning efficient representations of input data.
- **Applications of Dimensionality Reduction in Machine Learning:** Examines real-world applications of dimensionality reduction techniques, demonstrating their role in improving model performance and interpretability in machine learning tasks.

9.3 Advanced Optimization Techniques

- **Gradient Descent Optimization Algorithms (SGD, Adam, RMSprop):** Explores gradient descent variants, including SGD, Adam, and RMSprop, enhancing understanding of optimization techniques in machine learning model training.
- **Learning Rate Schedulers:** Investigates learning rate schedulers, dynamically adjusting learning rates during training to improve model convergence and performance.
- **Momentum-Based Optimization:** Covers momentum-based optimization algorithms, enhancing gradient descent efficiency by introducing momentum terms to accelerate convergence and reduce oscillations.
- **Second-Order Optimization (Newton's Method, Quasi-Newton Methods):** Explores second-order optimization methods like Newton's method and quasi-Newton methods, offering faster convergence and robustness compared to first-order methods.
- **Meta-Learning and Hyperparameter Optimization:** Examines meta-learning and hyperparameter optimization techniques, automating the process of tuning model parameters for improved performance.
- **Practical Tips for Efficient Model Training and Optimization:** Provides practical insights and strategies for efficient model training and optimization, optimizing workflows and avoiding common pitfalls in machine learning projects.

Module 10

Data-Driven Decision-Making

10.1 Introduction to Data-Driven Decision Making

- **Understanding the Importance of Data-Driven Decision Making:** Explores the significance of leveraging data to inform decision-making processes, driving organizational efficiency and effectiveness.
- **Overview of the Decision-Making Process:** Provides a comprehensive understanding of the decision-making process, from problem identification to evaluation and implementation of solutions.
- **Role of Data in Decision Making:** Investigates how data serves as a crucial factor in informing decisions, enabling organizations to make informed choices based on empirical evidence.
- **Benefits and Challenges of Data-Driven Decision Making:** Examines the advantages and obstacles associated with adopting a data-driven approach to decision-making, highlighting opportunities for improvement and innovation.

10.2 Open Source Tools for Data-Driven Decision Making

- **Apache Superset:** Apache Superset is an open-source data exploration and visualization platform developed by Airbnb. It offers a rich set of visualization options, interactive dashboards, and the ability to connect to various data sources.
- **Redash:** Redash is an open-source data visualization and dashboarding tool that allows users to connect to various data sources and create interactive dashboards. It supports SQL queries and offers a wide range of visualization options.
- **Pentaho:** Pentaho is an open-source business intelligence suite that offers data integration, analytics, and reporting capabilities. It provides a visual interface for creating reports and dashboards and supports integration with big data platforms.

10.3 Deriving Data-Driven Insights from Sales Dataset

- **Introduction to Case Study:** Introduces a case study analyzing Adidas sales data, providing context for subsequent courses in data analysis.
- **Data Collection and Preparation: Understanding the Adidas Sales Dataset:** Covers techniques for collecting and preprocessing the Adidas sales dataset, ensuring data quality and readiness for analysis.
- **Exploratory Data Analysis (EDA): Identifying Key Patterns and Trends in Adidas Sales:** Explores exploratory data analysis methods to uncover insights and patterns in Adidas sales data, informing subsequent analysis.
- **Building Predictive Models: Using Machine Learning Algorithms for Sales Forecasting and Customer Segmentation:** Applies machine learning algorithms to predict sales and segment customers based on Adidas sales data.
- **Visualization and Interpretation: Creating Interactive Dashboards in Power BI to Derive Insights from Adidas Sales Data:** Demonstrates creating interactive dashboards in Power BI to visualize and interpret Adidas sales data effectively.
- **Decision Making: Using Data-Driven Insights to Make Informed Business Decisions for Adidas:** Explores using data-driven insights from the Adidas sales dataset to make informed business decisions, optimizing operations and strategies.
- **Case Study Discussion:** Facilitates discussion and hands-on exercises based on the Adidas sales case study, reinforcing learning and practical application of data analysis techniques.

Module 11

Data Storytelling

11.1 Understanding the Power of Data Storytelling

- **Introduction to Data Storytelling:** Introduces the concept of data storytelling, exploring techniques to effectively communicate insights derived from data analysis.
- **Importance of Storytelling in Data Analysis and Communication:** Examines the role of storytelling in data analysis, emphasizing its power to engage and persuade audiences effectively.
- **The Psychology of Storytelling: Why Stories Resonate with Audiences:** Explores the psychological principles behind storytelling, highlighting why narratives resonate and influence audience perceptions.
- **Real-Life Examples of Successful Data Stories and Their Impact on Decision Making:** Analyzes real-world data stories, demonstrating their impact on decision-making processes and organizational outcomes.

11.2 Identifying Use Cases and Business Relevance

- **Identifying Data-Driven Use Cases in Business:** Teaches methods for recognizing opportunities where data can drive business value, fostering innovation and strategic decision-making.
- **Understanding the Business Context:** Goals, Challenges, and Opportunities: Explores the business landscape to identify objectives, obstacles, and potential areas for data-driven solutions.
- **Selecting Relevant Data Sources and Metrics for the Use Case:** Guides in selecting appropriate data sources and metrics aligned with business goals, ensuring meaningful analysis and actionable insights.
- **Defining the Audience:** Stakeholders, Decision Makers, and End Users: Examines the stakeholders involved in the data-driven initiative, ensuring alignment of insights with user needs and organizational objectives.

11.3 Crafting Compelling Narratives

- **Structuring a Data Story: Beginning, Middle, End:** Guides in structuring data narratives effectively, ensuring coherence and engagement through a well-defined storyline.
- **Developing a Clear Message and Storyline:** Teaches crafting a compelling message and storyline, conveying data insights in a coherent and impactful manner.
- **Incorporating Data Insights into the Narrative Flow:** Explores integrating data insights seamlessly into the narrative flow, enhancing storytelling effectiveness and audience understanding.

- **Engaging the Audience:** Emotion, Connection, and Call to Action: Focuses on engaging the audience emotionally, fostering connection and motivating action through data storytelling techniques.

11.4 Visualizing Data for Impact

- **Choosing the Right Visualizations for the Story:** Guides in selecting appropriate visualizations to effectively convey data insights and enhance storytelling impact.
- **Data Visualization Best Practices: Clarity, Simplicity, and Relevance:** Covers principles for creating clear, simple, and relevant visualizations to communicate data insights effectively.
- **Creating Engaging Visuals: Charts, Graphs, Maps, and Infographics:** Explores techniques for designing visually appealing charts, graphs, maps, and infographics to engage and inform audiences.
- **Using Interactive Elements to Enhance Understanding and Engagement:** Demonstrates incorporating interactive elements into visualizations to facilitate exploration, understanding, and audience engagement with data stories.

Module 12

Capstone Project - Employee Attrition Prediction

12.1 Project Introduction and Problem Statement

- **Introduction to the Capstone Project: Employee Attrition Prediction:** Introduces the capstone project focusing on predicting employee attrition, applying data analysis techniques to HR datasets.
- **Overview of the Problem Statement:** Predicting Employee Attrition Rates: Provides an overview of the project goal, predicting employee attrition rates, and its significance in organizational management.
- **Understanding the Business Context:** Impact of Employee Turnover on Organizational Performance: Explores the impact of employee turnover on organizational performance, emphasizing the importance of reducing attrition rates.
- **Identifying Data Sources:** HR Records, Employee Surveys, Performance Metrics: Guides in identifying relevant data sources such as HR records, surveys, and performance metrics for the project.
- **Formulating Hypotheses:** Factors Influencing Employee Attrition and Their Impact: Helps in formulating hypotheses about factors influencing employee attrition and their potential impact on organizational outcomes.

12.2 Data Collection and Preparation

- **Collecting and Acquiring Relevant Data:** HR Databases, Survey Responses, Performance Reports: Guides in gathering pertinent data sources including HR databases, surveys, and performance reports for analysis.
- **Data Cleaning and Preprocessing Techniques:** Handling Missing Values, Outliers, and Inconsistencies: Covers methods for preparing data by addressing missing values, outliers, and inconsistencies to ensure accuracy.
- **Exploratory Data Analysis (EDA):** Understanding Employee Demographics, Job Roles, and Attrition Patterns: Analyzes employee data to identify patterns in demographics, job roles, and attrition, guiding subsequent analysis.
- **Feature Engineering:** Creating Relevant Variables such as Employee Tenure, Satisfaction Scores, and Performance Ratings: Explores techniques for creating new variables like tenure and satisfaction scores to enhance predictive modeling accuracy.

12.3 Data Analysis and Modeling

- **Choosing Suitable Data Analysis and Modeling Techniques:** Classification Algorithms for Predictive Modeling
- **Building Predictive Models:** Employing Machine Learning Models like Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting Machines
- **Evaluating Model Performance:** Using Metrics like Accuracy, Precision, Recall, and F1-Score
- **Interpreting Model Results:** Identifying Key Factors Contributing to Employee Attrition and Their Relative Importance

12.4 Data Storytelling and Presentation

- **Crafting a Compelling Narrative Around the Project Findings:** Highlighting the Business Impact of Employee Attrition
- **Visualizing Key Insights Using Data Storytelling Techniques:** Creating Interactive Dashboards and Visualizations
- **Creating Interactive Dashboards and Reports to Showcase the Project Results:** Displaying Attrition Trends, Predicted Attrition Rates, and Impact Analysis
- **Presenting the Capstone Project to Peers and Stakeholders:** Communicating Findings, Recommendations, and Actionable Insights for Mitigating Employee Attrition