CERTIFICATE IN DATA SCIENCE

The Certificate in Data Science is designed to cultivate curiosity and equip participants with the skills needed for data-driven decision-making.

Throughout the program, you'll gain a thorough understanding of the data lifecycle, learning to acquire, clean, structure, store, manipulate, analyze, and visualize data from diverse sources to address complex business questions. This certificate goes beyond the basics, teaching you to identify patterns and leverage machine learning and analytical tools applicable across various domains, including marketing, product development, project management, and general business administration.

The study plan consists of ten courses, plus a practical final project.

COURSES:

Subject Name	Theory sessions	ECTS
Introduction to Data Science	2	0
Python Coding	15	3
Probability and Statistics	15	3
Data Wrangling	5	1
Data acquisition in SQL	10	2
Exploratory Data Analysis and Data Visualization	5	1
Introduction to Machine Learning in Python	15	3
Introduction to Deep Learning	5	1
Introduction to Big Data	5	1
Ethics in Data Science	2	0
Capstone Project (Workgroup*)	15	0
TOTAL:	79 + 15*	15

1) **Introduction to Data Science (2 sessions):**
    a. **Description**: This section could provide a high-level overview of the field of Data Science and its applications in various industries.
    b. **Content**: Students could learn about the different stages of the Data Science process, including data acquisition, cleaning, analysis, and visualization, as well as the tools and techniques used in each stage. The section could also introduce students to common use cases for Data Science, such as predictive modeling, customer segmentation, and recommendation systems.
    c. **Goal**: By the end of this section, students should have a good understanding of what Data Science is and how it is used in practice.
2) **Python Coding (15 sessions):**

a. **Description**: This section could provide a comprehensive introduction to the Python programming language, including its syntax, data types, variables, control structures, functions, and libraries.
b. **Content**: Students could learn how to write basic Python code, manipulate data using built-in data structures such as lists and dictionaries, and use functions to perform common tasks. This section could also introduce students to the scientific computing libraries in Python, such as NumPy and Pandas, which are widely used in Data Science. These libraries provide tools for working with arrays and data frames, and allow students to perform tasks such as data manipulation, aggregation, and visualization more easily.
c. **Goal**: By the end of this section, students should have a solid understanding of Python programming and be able to use it to perform data analysis tasks.

3) **Probability and Statistics (15 sessions)** :
   a. **Description**: This section could cover basic concepts of probability, descriptive statistics, and inferential statistics.
   b. **Content**: Students could learn about the different types of probability distributions, such as normal, binomial, and Poisson, and how to use them to model real-world phenomena. This section could also introduce students to hypothesis testing, regression analysis, and other statistical techniques used in Data Science.
   c. **Goal**: By the end of this section, students should have a solid understanding of probability and statistics, and be able to apply statistical techniques to real-world data sets.
4) **Data Wrangling (5 sessions):**
   a. **Description**: This section could cover the process of cleaning and transforming raw data into a format that can be used for analysis.
   b. **Content**: Students could learn about the various challenges involved in working with real-world data, such as missing values, inconsistent data formats, and outliers. They could learn how to use Python and its libraries, such as Pandas, to perform tasks such as data cleaning, normalization, and transformation. This section could also introduce students to the basics of regular expressions, which are used to match and manipulate text patterns in data.
   c. **Goal**: By the end of this section, students should have a good understanding of the data wrangling process and be able to apply it to real-world data sets.
5) **Data acquisition in SQL (10 sessions)**:
   a. **Description**: This section could cover the basics of SQL, a language used to manage and query data stored in relational databases.
   b. **Content**: Students could learn how to write SQL statements to extract data from databases, perform aggregations, and join data from multiple tables. This section could also introduce students to database management systems, such as MySQL, and how to connect to databases using Python and its libraries, such as SQLAlchemy.

   c. **Goal**: By the end of this section, students should be able to write basic SQL queries and understand how to use SQL in a data science context to prepare data for analysis.

6) **Exploratory Data Analysis and Data Visualization (5 sessions)**:
   a. **Description**: This section could cover the process of analyzing and visualizing data to gain insights and identify patterns.
   b. **Description**: Students could learn how to use Python and its libraries, such as Pandas and Matplotlib, to perform tasks such as data summarization, visualization, and hypothesis testing. This section could also introduce students to various data visualization techniques, such as bar charts, line charts, scatter plots, etc., and how to use them effectively to communicate insights. The section could also cover the basics of statistics, such as measures of central tendency, variability, and correlation, and how to apply them to data sets.
   c. **Goal**: By the end of this section, students should be able to perform basic data analysis tasks, such as finding relationships between variables, identifying outliers, and generating visualizations to communicate insights. They should also have a good understanding of data visualization and be able to create basic visualizations, customize their appearance, and understand how to use visualizations to communicate insights.
7) **Introduction to Machine Learning (15 sessions)**:
   a. **Description**: This section could cover the basics of machine learning, including supervised and unsupervised learning algorithms, and how to evaluate their performance.
   b. **Content**: Students could learn about popular algorithms such as linear regression, k-nearest neighbors, decision trees, and k-means clustering, and how to apply them to real-world problems. This section could also introduce students to the basics of model evaluation, including accuracy, precision, recall, and F1 score, and how to use them to compare the performance of different algorithms.
   c. By the end of this section, students should have a good understanding of machine learning, and be able to apply supervised and unsupervised learning algorithms to real-world data sets.
8) **Introduction to Deep Learning (5 sessions)**:
   a. **Description**: This section could cover the basics of deep learning, including artificial neural networks, and how to build and train them for various tasks.
   b. **Content**: Students could learn about popular deep learning frameworks, such as TensorFlow and PyTorch, and how to build and train neural networks for tasks such as image classification, natural language processing, and time series forecasting. This section could also introduce students to advanced deep learning techniques, such as convolutional neural networks and recurrent neural networks, and how to apply them to solve more complex problems.
   c. **Goal**: By the end of this section, students should have a good understanding of deep learning and be able to build and train neural networks for various tasks.

9) **Introduction to Big Data (5 sessions)**:
    a. **Description**: This section could cover the challenges of working with large data sets, and techniques for processing and analyzing big data, such as MapReduce and Apache Spark.
    b. **Content**: Students could learn about the differences between big data and traditional data, and the types of problems that big data can solve. This section could also introduce students to the basics of distributed computing, and how to use tools such as Apache Spark to process and analyze big data in a scalable and efficient manner.
    c. **Goal**: By the end of this section, students should have a good understanding of big data and be able to apply big data processing techniques to real-world problems.
10) **Ethics in Data Science (2 sessions)**:
    a. **Description**: This section could cover ethical considerations in Data Science, such as privacy, bias, and fairness.
    b. **Content**: Students could learn about the ethical implications of working with sensitive data, such as personal information and financial data, and how to ensure that data is used in a responsible and secure manner. This section could also introduce students to the challenges of algorithmic bias, and how to ensure that models are fair and unbiased in their predictions.
    c. **Goal**: By the end of this section, students should have a good understanding of the ethical considerations involved in working with data, and be able to apply best practices to ensure that data is used in a responsible and ethical manner.
11) **Capstone Project**:
    a. **Description**: This section could be a hands-on project where students apply the concepts they have learned to a real-world problem, and present their findings to the collaborating companies.
    b. **Content**: This section could provide students with an opportunity to work on a project from start to finish, including data acquisition, cleaning, analysis, and visualization. Students could work in teams, and could use the tools and techniques learned in the previous sections to complete the project.
    c. **Goal**: By the end of this section, students should have a good understanding of the entire Data Science process and be able to apply it to real-world problems. The capstone project could also serve as a portfolio piece that students can showcase to potential employers or use as a foundation for further study in the field.

Detailed content:

1) **Introduction to Data Science:**
    a. Overview of Data Science
        i. Definition of Data Science
        ii. Importance of Data Science

iii. Applications of Data Science

b. The Data Science Process
   i. Data Acquisition
   ii. Data Cleaning
   iii. Data Analysis
   iv. Data Visualization
   v. Modeling
   vi. Deployment

c. Skills and Tools Required for Data Science
   i. Programming languages (Python, SQL, etc.)
   ii. Data analysis and visualization tools (Pandas, Matplotlib, etc.)
   iii. Statistics
   iv. Data modeling
   v. Machine learning libraries (scikit-learn, TensorFlow, etc.)

d. Types of Problems Solved by Data Scientists
   i. Predictive modeling
   ii. Identifying patterns
   iii. Making decisions based on data
   iv. Optimizing processes

2) **Python Coding:**

a. Overview of Python
   i. History of Python
   ii. Advantages of Python for Data Science

b. Basic Syntax
   i. Variables
   ii. Data types (strings, integers, floats, etc.)
   iii. Operators (arithmetic, comparison, etc.)
   iv. Control structures (if-else statements, for loops, etc.)
   v. Functions
   vi. Object-oriented programming

c. Data Structures
   i. Lists
   ii. Tuples
   iii. Dictionaries
   iv. Sets

d. Python Libraries for Data Science
   i. Pandas
   ii. NumPy
   iii. Matplotlib

3) **Probability and Statistics**

a. Overview of Probability and Statistics
   i. Definition of Probability and Statistics
   ii. Importance of Probability and Statistics in Data Science
   iii. Applications of Probability and Statistics in Data Science

b. Descriptive Statistics

i. Measures of Central Tendency (Mean, Median, Mode)
ii. Measures of Dispersion (Range, Variance, Standard Deviation)
iii. Box Plots and Outliers
iv. Skewness and Kurtosis

c. Probability Distributions
i. Discrete Distributions (Bernoulli, Binomial, Poisson)
ii. Continuous Distributions (Uniform, Normal, Exponential)
iii. Joint and Marginal Distributions
iv. Conditional Probability

d. Hypothesis Testing
i. Overview of Hypothesis Testing
ii. Null and Alternative Hypotheses
iii. One-Sample T-Test
iv. Two-Sample T-Test
v. Analysis of Variance (ANOVA)
vi. chi-square Test

e. Regression Analysis
i. Simple Linear Regression
ii. Multiple Linear Regression
iii. Logistic Regression
iv. Model Evaluation (R-squared, Confidence Intervals, Hypothesis Tests)

4) **Data Wrangling**:

a. Overview of Data Wrangling
i. Definition of Data Wrangling
ii. Importance of Data Wrangling
iii. Applications of Data Wrangling

b. Dealing with Missing Values
i. Types of missing values
ii. Methods for dealing with missing values (drop missing values, fill with mean or median, etc.)

c. Dealing with Outliers
i. Definition of outliers
ii. Methods for dealing with outliers (remove outliers, fill with mean or median, etc.)

d. Data Normalization
i. Overview of data normalization
ii. Methods for data normalization (min-max normalization, z-score normalization, etc.)

e. Data Aggregation
i. Overview of data aggregation
ii. Methods for data aggregation (grouping by, pivot tables, etc.)

f. Pandas Library for Data Wrangling
i. Overview of the Pandas library
ii. Importing data into Pandas

iii. Data Wrangling tasks in Pandas (dealing with missing values, outliers, normalization, aggregation, etc.)

5) **Data acquisition in SQL:**
   a. Overview of SQL
      i. Definition of SQL
      ii. Importance of SQL
      iii. Applications of SQL
   b. Basic Syntax
      i. Data types (strings, integers, floats, etc.)
      ii. Operators (arithmetic, comparison, etc.)
      iii. Keywords (select, from, where, etc.)
   c. Data Manipulation Commands
      i. Insert
      ii. Update
      iii. Delete
   d. Data Aggregation Commands
      i. Group by
      ii. Having
      iii. Sum
      iv. Count
      v. Average
   e. Relational Databases
      i. Overview of relational databases
      ii. Tables
      iii. Primary keys
      iv. Foreign keys
      v. Relationships between tables
   f. SQL Environment
      i. Overview of SQL environment (SQLite, MySQL, etc.)
      ii. Connecting to a database
      iii. Running SQL commands in the environment

6) **Exploratory Data Analysis and Data Visualization:**
   a. Overview of EDA and Data Visualization
      i. Definition of EDA and Data Visualization
      ii. Importance of EDA and Data Visualization
      iii. Applications of EDA and Data Visualization
   b. Descriptive Statistics
      i. Mean
      ii. Median
      iii. Mode
      iv. Standard Deviation
      v. Variance
   c. Data Distributions
      i. Normal Distribution
      ii. Skewness

iii. Kurtosis

d. Hypothesis Testing
   i. Overview of Hypothesis Testing
   ii. One-Sample T-Test
   iii. Two-Sample T-Test
   iv. ANOVA
   v. chi-square test

e. Data Visualization Techniques
   i. Line Plots
   ii. Scatter Plots
   iii. Bar Plots
   iv. Histograms
   v. Box Plots
   vi. Heat Maps
   vii. Pie Charts

f. Matplotlib Library for Data Visualization
   i. Overview of the Matplotlib library
   ii. Creating various types of plots and charts in Matplotlib
   iii. Customizing plots and charts in Matplotlib
   iv. Saving plots and charts as images

7) **Introduction to Machine Learning in Python**:

a. Overview of Machine Learning
   i. Definition of Machine Learning
   ii. Importance of Machine Learning in Data Science
   iii. Applications of Machine Learning in Data Science

b. Supervised Learning
   i. Linear Regression
      1. Overview of Linear Regression
      2. Simple Linear Regression
      3. Multiple Linear Regression
      4. Model Evaluation Metrics (R-squared, Mean Squared Error, etc.)
      5. Model Selection Techniques (Stepwise Regression, Best Subset Selection, etc.)
      6. Regularization Techniques (Ridge Regression, Lasso Regression, etc.)
      7. Non-Linear Regression
   ii. Logistic Regression
      1. Overview of Logistic Regression
      2. Model Evaluation Metrics (Accuracy, Precision, Recall, etc.)
      3. Regularization Techniques (Ridge Regression, Lasso Regression, etc.)
      4. Model Selection Techniques (Stepwise Regression, Best Subset Selection, etc.)
      5. Multi-Class Logistic Regression

iii. Decision Trees
    1. Overview of Decision Trees
    2. How Decision Trees Work
    3. Model Evaluation Metrics (Accuracy, Precision, Recall, etc.)
    4. Overfitting and Pruning Techniques
    5. Bagging and Random Forests
iv. Random Forests
    1. Overview of Random Forests
    2. How Random Forests Work
    3. Model Evaluation Metrics (Accuracy, Precision, Recall, etc.)
    4. Feature Importance
    5. Advantages and Limitations of Random Forests

c. Model Evaluation
    i. Overview of Model Evaluation
    ii. Confusion Matrix
    iii. Precision and Recall
    iv. F1 Score
    v. ROC Curve
d. Model Selection
    i. Overview of Model Selection
    ii. Cross-Validation
    iii. Hyperparameter Tuning

8) **Introduction to Deep Learning:**
    a. Overview of Deep Learning
        i. Definition of Deep Learning
        ii. Importance of Deep Learning in Data Science
        iii. Applications of Deep Learning in Data Science
    b. Artificial Neural Networks (ANN)
        i. Overview of ANN
        ii. Perceptrons
        iii. Multi-Layer Perceptrons (MLP)
        iv. Activation Functions
        v. Gradient Descent
    c. Convolutional Neural Networks (CNN)
        i. Overview of CNN
        ii. How CNN Works
        iii. Convolutional Layers
        iv. Pooling Layers
        v. Fully Connected Layers
    d. Recurrent Neural Networks (RNN)
        i. Overview of RNN
        ii. How RNN Works
        iii. Long Short-Term Memory (LSTM)
        iv. Gated Recurrent Unit (GRU)
    e. Autoencoders

i. Overview of Autoencoders
ii. How Autoencoders Work
iii. Convolutional Autoencoders
iv. Recurrent Autoencoders

f. Generative Adversarial Networks (GAN)
i. Overview of GAN
ii. How GAN Works
iii. Generator and Discriminator
iv. Training GAN

g. Transfer Learning
i. Overview of Transfer Learning
ii. Fine-Tuning Pretrained Networks
iii. Using Pretrained Networks as Feature Extractors

9) Introduction to Big Data

a. Overview of Big Data
i. Definition of Big Data
ii. Characteristics of Big Data (Volume, Variety, Velocity, Veracity, and Value)
iii. Importance of Big Data in Data Science

b. Big Data Processing Technologies
i. Overview of Big Data Processing Technologies
ii. MapReduce
iii. Hadoop Distributed File System (HDFS)
iv. Apache Spark
v. Apache Flink

c. NoSQL Databases
i. Overview of NoSQL Databases
ii. Document Databases (MongoDB)
iii. Columnar Databases (Cassandra)
iv. Key-Value Databases (Redis)
v. Graph Databases (Neo4j)

d. Big Data Analytics
i. Overview of Big Data Analytics
ii. Descriptive Analytics
iii. Predictive Analytics
iv. Prescriptive Analytics
v. Real-Time Analytics

10) **Ethics in Data Science**

a. Overview of Ethics in Data Science
i. Definition of Ethics in Data Science
ii. Importance of Ethics in Data Science
iii. Ethical Considerations in Data Science

b. Privacy and Security
i. Overview of Privacy and Security
ii. Data Privacy Laws (GDPR, CCPA, etc.)

iii. Security Measures (Encryption, Firewall, etc.)
iv. Ethical Considerations of Data Collection and Storage

c. Bias and Fairness
    i. Overview of Bias and Fairness
    ii. Algorithmic Bias
    iii. Mitigating Bias in Data Science
    iv. Ethical Considerations of Decision-Making with Data
d. Explainability and Transparency
    i. Overview of Explainability and Transparency
    ii. Model Interpretability
    iii. Explainable Artificial Intelligence (XAI)
    iv. Ethical Considerations of Artificial Intelligence