MOOC BASED SEMINAR REPORT



on

**INTRODUCTION TO DATA SCIENCE**

Submitted in partial fulfilment of the requirement for the V semester

**Bachelor of Technology**

By

**ANUJ BORA**

**University Roll No 2161088**

Under the Guidance of

## Mrs. Senam Pandey

**Assistant Professor Deptt. of CSE**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING GRAPHIC ERA HILL UNIVERSITY, BHIMTAL CAMPUS SATTAL ROAD, P.O. BHOWALI DISTRICT-**

**NAINITAL-263132 2023-2024**

1

## CERTIFICATE



THIS IS TO CERTIFY THAT **ANUJ BORA** HAS SATISFACTORILY PRESENTED MOOC BASED SEMINAR ON THE COURSE TITLE **INTRODUCTION TO DATA SCIENCE** IN PARTIAL FULLFILLMENT OF THE SEMINAR PRESENTATION REQUIREMENT IN **5th** SEMESTER OF BTECH DEGREE COURSE PRESCRIBED BY GRAPHIC ERA HILL UNIVERSITY DURING THE ACADEMIC SESSION 2023-2024.

Campus MOOCS -Coordinator and Mentor

2

# TABLE OF CONTENTS: -



|  |  |  |
| --- | --- | --- |
| **S.NO** | **CONTENT** | **PAGE NO.** |
| **1** | ACKNOWLEDGEMENT | 7 |
| **2** | INTRODUCTION | 8 |
| **3** | WEEK 1 | 10 |
| **4** | WEEK 2 | 11 |
| **5** | WEEK 3 | 14 |
| **6** | WEEK 4 | 16 |
| **7** | WEEK 5 | 19 |
| **8** | WEEK 6 | 24 |
| **9** | CERTIFICATE OF COMPLETION | 29 |

3

## ACKNOWLEDGEMENT



I take this opportunity to express my profound gratitude and deep regards to my guide **Mrs. Senam Pandey** for his exemplary guidance, monitoring and constant encouragement throughout the course.

The blessing, help and guidance given by him time to time helped me throughout the project. The Success and outcome of this course required a lot of guidance and assistance from many people, And I am extremely privileged to have got this all along the completion of my report. All that I have Done is only due to such supervision and assistance and I would not forget to thank them. I am Thankful to and fortunate enough to get constant encouragement, support and guidance from all the People around me which helped me in successfully completing my online course.

4

# INTRODUCTION



This course is about learning data science, its basic definition, importance, increasing popularity, basic process, prerequisites to learn data science and become a Data Scientist.

It is a 6-week course.

**Week 1:** Introduction To Data Science

**Week 2:** Data Science Process

**Week 3:** Python For Data Science

**Week 4:** Mathematics And Statistics For Data Science **Week 5:** Data Visualization

**Week 6:** Machine Learning For Data Science

## Week 1



In this module, I learnt about data and data science. It started by looking in detail about the basic definition of data and data science, its importance etc. Then we learnt about why data science is becoming so powerful nowadays.

## Data-

1. The unorganized facts and figures that do not convey any message on its own is called data
2. Data needs to be processed to convert it into some valuable information.

3. Rising growth in internet adoption and rapid technological advances in device connectivity are driving the flow of data at an exponential rate.

**Data Science:**

Data Science is an interdisciplinary field that uses algorithms, procedures and processes to examine large amount of data to uncover hidden patterns , generate insights and direct decision making

**Importance Of Data Science:**

Data science is important because it combines tools, methods, and technology to generate meaning from data.

• Data Science helps brands to understand their customers in a much enhanced and empowered manner.

• Its findings and results can be applied to almost any sector like travel, healthcare, education etc.

• Data Science is accessible to almost all sectors.

## WEEK 2



In the second week I learnt about process of Data Science.

The data science process is a systematic approach to extracting valuable insights and knowledge from data. It typically involves several key stages, each with its own set of tasks and activities. The specific steps can vary depending on the framework or methodology used, but here is a generalized outline of the data science process:

**1. Problem Definition:** - Clearly define the problem or question you want to address.

- Understand the business context and objectives.

**2. Data Collection:**

- Identify and gather relevant data sources.

- Collect raw data from various repositories.

**3. Data Cleaning and Preprocessing:**

- Clean and handle missing data.

- Remove outliers and irrelevant information.

- Transform and preprocess data for analysis.

**4. Exploratory Data Analysis (EDA):**

- Explore and visualize the data to understand patterns and relationships.

- Identify potential insights or trends.

**5. Feature Engineering:**

- Create new features from existing ones.

- Select and transform features to enhance model performance.

**6. Model Selection:**

- Choose appropriate models based on the nature of the problem (classification, regression, clustering, etc.).

- Split the data into training and testing sets.

**7. Model Training:**

- Train the selected models using the training dataset.

- Optimize model parameters for better performance.

**8. Model Evaluation:**

- Evaluate model performance using the testing dataset.

- Use metrics such as accuracy, precision, recall, F1-score, etc.

**9. Model Deployment:**

- Deploy the model for real-world use, integrating it into production systems.

**10. Monitoring and Maintenance:**

- Monitor the model's performance in real-world scenarios.

- Update and retrain the model as needed to maintain accuracy.

**11. Communication of Results:**

- Present findings and insights to stakeholders.

- Create visualizations and reports to communicate results effectively.

## WEEK 3



In the third week I learnt about Python programming language. Python is very easy and most popular language for Data Science.

Python is a widely used programming language in the field of data science due to its versatility and rich ecosystem of libraries. Here's how Python is leveraged throughout the data science process:

**1. Libraries:**

**-NumPy**: Enables efficient manipulation of large arrays and matrices.

- **Pandas:** Provides data structures like DataFrames for easy data handling and analysis.

**- Matplotlib and Seaborn**\* Used for creating various types of visualizations.

-**Scikit-learn:** Offers a range of machine learning algorithms for tasks like classification and regression.

2. **Jupyter Notebooks**:

- Jupyter Notebooks allow for interactive data exploration, combining code, text, and visualizations in a single document.

3**. Data Manipulation**:

- Pandas simplifies tasks like cleaning, transforming, and aggregating data, making it a crucial tool for data manipulation.

**4. Data Visualization:**

- Matplotlib and Seaborn are commonly employed for creating plots and

charts, aiding in the interpretation of data patterns.

**5.Machine Learning:**

- Scikit-learn facilitates the implementation of machine learning models for tasks such as classification, regression, and clustering.

**6. Deep Learning:**

- For more complex tasks, Python leverages libraries like TensorFlow and PyTorch, allowing the development of deep learning models.

**7. Community and Ecosystem:**

- Python boasts a robust and active community, providing a wealth of resources, tutorials, and support for data science practitioners.

## WEEK 4



Mathematics and statistics are fundamental to data science, providing the theoretical foundation for various algorithms, models, and analytical techniques. Here's an overview of key mathematical and statistical concepts in the context of data science:

**Mathematics:**

**1. Linear Algebra:**

- Essential for understanding and working with vectors and matrices, which are common data structures in machine learning.

- Used in tasks like regression, dimensionality reduction, and neural networks.

**2. Calculus:**

- Differential calculus is crucial for understanding rates of change, gradients, and optimization.

- Integral calculus is used in areas like probability and aggregating data.

**3. Probability:**

- Fundamental for modeling uncertainty and randomness in data.

- Probability distributions are central to understanding and modeling data variability.

**4. Optimization:**

- Used in machine learning for finding the best parameters or coefficients that minimize or maximize a certain function (e.g., cost function).

**Statistics**:

**1. Descriptive Statistics:**

- Measures of central tendency (mean, median, mode) and dispersion (variance, standard deviation) help summarize and describe datasets.

**2. Inferential Statistics:**

- Allows making inferences about a population based on a sample.

- Confidence intervals and hypothesis testing are common inferential techniques.

**3. Probability Distributions:**

- Understanding and working with distributions such as normal, binomial, and Poisson is crucial for statistical modeling.

**4. Statistical Testing:**

- Hypothesis testing is used to make decisions or draw conclusions about populations based on sample data.

- Common tests include t-tests, chi-square tests, and ANOVA.

**5. Regression Analysis:**

- Linear and nonlinear regression models are used to establish relationships between variables and make predictions.

**6. Time Series Analysis:**

- Techniques for analyzing time-dependent data, including autoregressive models, moving averages, and seasonality analysis.

**7. Bayesian Statistics:**

- Involves updating probabilities based on new evidence and is applied in machine learning, especially in probabilistic models.

**8. Multivariate Statistics:**

- Techniques for analyzing relationships between multiple variables, including covariance and correlation matrices.

## WEEK 5



In the fifth week I studied about Data Visualization.. Data visualization is a key aspect of the data science process, enabling the communication of complex information and patterns in a visually intuitive way. Effective data visualization helps to uncover insights, identify trends, and convey findings to a wider audience. Here are some important aspects of data visualization:

**Principles of Data Visualization**:

**1. Simplicity:**

- Keep visualizations simple and uncluttered to enhance understanding.

**2. Clarity:**

- Ensure that the message of the visualization is clear and easily interpretable.

**3. Accuracy:**

- Represent data accurately and avoid misleading visualizations.

**4. Relevance:**

- Include only relevant information and features in your visualizations.

**5. Consistency:**

- Use consistent colors, scales, and labeling across visualizations for better comparison.

**Common Types of Data Visualizations:**

**1. Bar Charts:**

- Used to compare quantities across different categories.

**2. Line Charts:**

- Display trends and changes over a continuous interval.

**3. Scatter Plots:**

- Show relationships between two numerical variables.

**4. Histograms:**

- Represent the distribution of a single variable.

**5. Pie Charts:**

- Display parts of a whole, useful for illustrating percentages.

**6. Heatmaps:**

- Visualize the magnitude of a phenomenon as color in two dimensions.

**7. Box Plots:**

- Summarize the distribution of a dataset, indicating the median, quartiles, and outliers.

**8. Bubble Charts:**

- Extend scatter plots by adding a third dimension, using the size of markers to represent additional information.

**9. Treemaps:**

- Represent hierarchical data structures using nested rectangles.

**10. Network Diagrams:**

- Show relationships between entities in a network.

**Tools for Data Visualization:**

**1. Matplotlib:**

- A versatile and widely-used plotting library for creating static, animated, and interactive visualizations in Python.

**2. Seaborn:**

- Built on top of matplotlib, Seaborn provides a high-level interface for creating attractive statistical graphics.

**3. Plotly:**

- A powerful library for creating interactive visualizations in Python and other programming languages.

**4. Tableau:**

- A popular data visualization tool that allows for the creation of interactive and shareable dashboards.

**5. Power BI:**

- Microsoft's business analytics tool for creating interactive reports and dashboards.

**Considerations for Effective Data Visualization:**

**1. Audience:**

- Tailor visualizations to the intended audience, considering their background and level of expertise.

**2. Storytelling:**

- Create a narrative with your visualizations to guide the viewer through the data story.

**3. Interactivity**:

- Incorporate interactive elements to allow users to explore and interact with the data.

**4. Aesthetics:**

- Pay attention to design elements such as color, font, and layout to enhance visual appeal.

**5. Accessibility**:

- Ensure that visualizations are accessible to individuals with disabilities, i

## WEEK 6

Machine learning is a subset of artificial intelligence that focuses on developing algorithms and models that enable computers to learn patterns from data and make predictions or decisions without explicit programming. In the context of data science, machine learning plays a crucial role in extracting meaningful insights and building predictive models. Here are key aspects of machine learning for data science:

**Types of Machine Learning:**

**1. Supervised Learning:**

- Involves training a model on a labeled dataset, where the algorithm learns the relationship between input features and corresponding target labels.

- Common algorithms: Linear Regression, Decision Trees, Support Vector Machines, Neural Networks.

**2. Unsupervised Learning:**

- Deals with unlabeled data, aiming to discover patterns or structure within the data.

- Common algorithms: K-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA).

**3. Semi-Supervised Learning:**

- Combines elements of both supervised and unsupervised learning, often used when only a fraction of the data is labeled.

**4. Reinforcement Learning:**

- Involves training an agent to make sequential decisions by interacting with an environment and receiving feedback in the form of rewards or penalties.

- Common algorithms: Q-Learning, Deep Reinforcement Learning.

**Steps in Applying Machine Learning to Data Science:**

**1. Problem Definition:**

- Clearly define the problem you want to solve and determine whether it's a regression, classification, clustering, or other types of problems.

**2. Data Collection and Preprocessing:**

- Gather relevant data and preprocess it to handle missing values, outliers, and ensure it is suitable for training models.

**3. Feature Engineering:**

- Select and transform features to improve model performance.

**4. Data Splitting:**

- Divide the dataset into training and testing sets to evaluate the model's performance on unseen data.

**5. Model Selection:**

- Choose a suitable machine learning algorithm based on the nature of the problem and data.

**6. Model Training:**

- Train the selected model using the training dataset.

**7. Model Evaluation:**

- Assess the model's performance on the testing dataset using appropriate evaluation metrics.

**8. Hyperparameter Tuning:**

- Fine-tune the model by adjusting hyperparameters to achieve better performance.

**9. Model Deployment:**

- Deploy the trained model to make predictions on new, unseen data in a real-world setting.

**Common Machine Learning Algorithms:**

**1. Linear Regression:**

- Used for predicting a continuous variable based on one or more input features.

**2. Logistic Regression:**

- Suitable for binary classification problems, predicting outcomes as either 0 or 1.

**3. Decision Trees**

- Tree-like models that make decisions based on input features.

**4. Random Forest:**

- Ensemble method that combines multiple decision trees for improved accuracy and robustness.

**5. Support Vector Machines (SVM):**

- Used for both classification and regression tasks, finding the hyperplane that best separates data into classes.

**6. K-Nearest Neighbors (KNN):**

- Classifies data points based on the majority class of their k nearest neighbors.

**7. Neural Networks:**

- Deep learning models with interconnected layers of neurons, suitable for complex tasks.

**8. K-Means Clustering:**

- Unsupervised algorithm for partitioning data into clusters.

**Challenges and Considerations:**

**1. Overfitting and Underfitting:**

- Striking the right balance to avoid overfitting (model memorizing the training data) or underfitting (model not capturing patterns).

**2. Data Quality:**

- The quality and relevance of the data directly impact model performance.

**3. Interpretability:**

- Understanding and interpreting the decisions made by complex models, especially in sensitive applications.

**4. Scalability:**

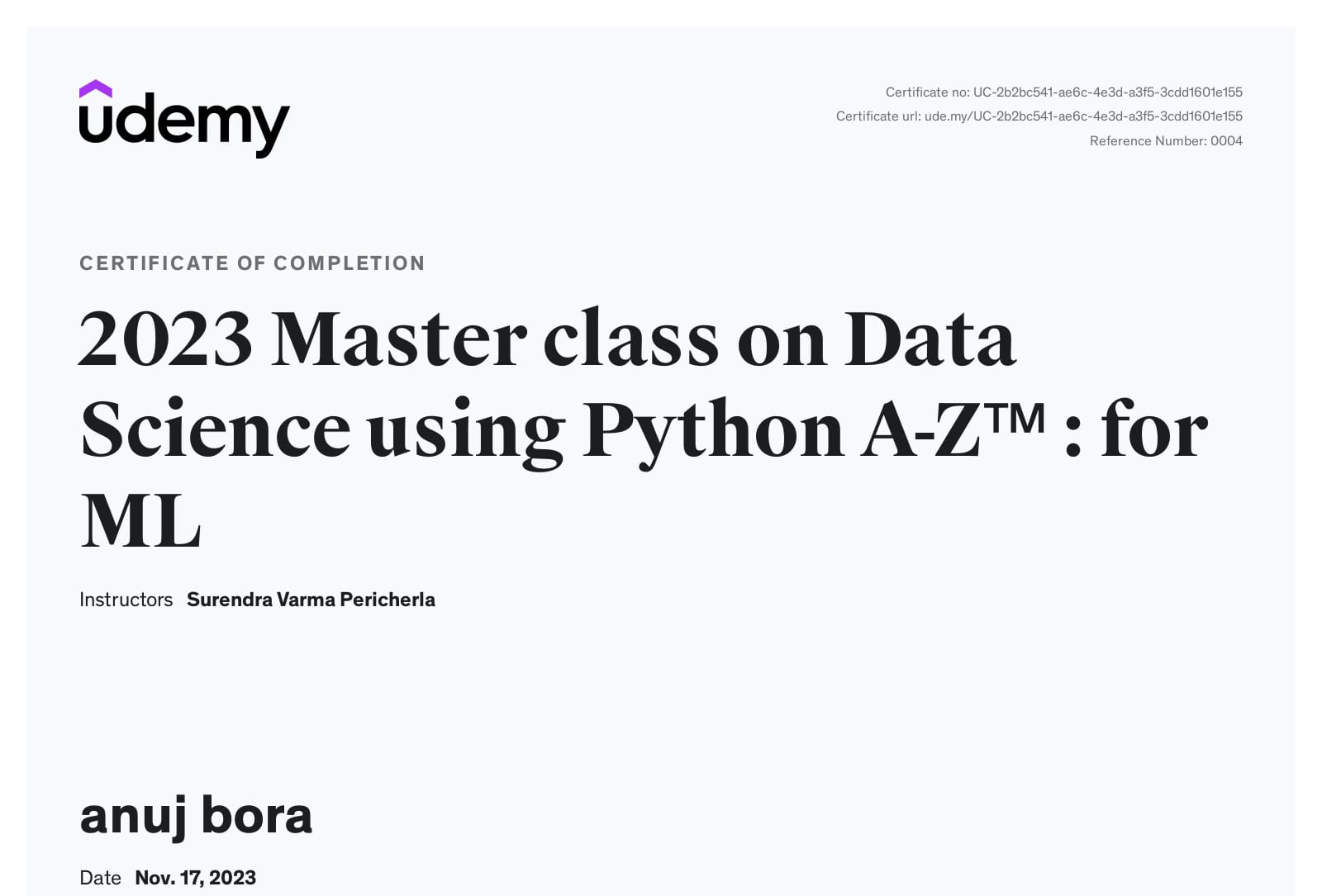
- Ensuring that the model can handle large datasets efficiently.

**5. Ethical Considerations:**

- Being aware of biases in the data and addressing ethical concerns, especially in areas like fairness, accountability, and transparency.



**CERTIFICATE OF COMPLETION:-**

****