ANJUMAN-I-ISLAM'S
KALSEKAR TECHNICAL CAMPUS

Micromasters Program

in

Statistics and Data Science

for

Bachelor of Engineering

Third Year w.e.f. A.Y. 2023-24

Under Inspiration of

MIT Boston's MITx MicroMasters Initiative

(As a preparatory course for the

MicroMasters Program in Statistics and Data Science)

Syllabus for Approval

Title of the Course	:	Micromasters in Statistics and Data Science (for undergraduate Engineering students)
Eligibility for Admission	:	After Passing Second Year Engineering with maximum 2 live KTs
Passing Marks	:	40%
No. of Years / Semesters	:	2 years / 4 semesters
Level	:	UG
Pattern	:	Semester
Status	:	Revised 2023
To be implemented from Academic Year	:	With effect from Academic Year: 2023-2024

Syllabus Authored by:

Prof. Tabrez Khan
I/c Head, CO, AIKTC

Dr. Fauwaz Parkar
I/c Head, CE, AIKTC

Prof. Salim Shaikh
Asst. Prof., CO, AIKTC

Dr. Shivaji Pawar
Associate Prof., CO, AIKTC

Syllabus Approved by:

Dr. Shariq Syed
I/c Dean, SoP, AIKTC

Dr. Rajendra Magar
I/c Dean, SoET, AIKTC

Dr. Ramjan Khatik
I/c Director, AIKTC

1. Overview

Anjuman - I -Islam’s
Kalsekar Technical Campus

School of Engineering &Technology

(With effect from 2023-24)

Mircomasters* in Statistics and Data Science

Year & Sem	Course Code and Course Title	Teaching Scheme Hours / Week			Examination Scheme and Marks					Credit Scheme
		Theory	Seminar/Tutorial	Pract.	Internal Assessment	End Sem Exam	Term Work	Oral/ Pract	Total	Credits
TE Sem V	**MSDS501: Statistics for Data Science**	04	--	--	40	60	25	--	125	04
	Total	**04**	**-**	**--**	**100**		**25**		**125**	**04**
Total Credits = 04										
TE Sem VI	**MSDS601: Fundamentals of Machine Learning Algorithms for Data Science**	04	--	--	40	60	25	--	125	04
	Total	**04**	**-**	**--**	**100**		**25**		**125**	**04**
Total Credits = 04										
BE Sem VII	**MSDS701: Fundamentals of Deep Learning Algorithms for Data Science**	04	--	--	40	60	25	--	125	04
	Total	**04**	**-**	**--**	**100**		**25**		**125**	**04**
Total Credits = 04										
BE Sem VIII	**MSDS801: Natural Language Processing and Visualization Techniques for Data Science**	04	--	--	40	60	25	--	125	04
	Total	**04**	**-**	**--**	**100**		**25**		**125**	**04**
Total Credits = 04										

The interdisciplinary nature of statistics and data science is a huge advantage. It bridges the gap between traditional engineering disciplines and the emerging field of data-driven decision-making. This makes it a must-have skill for any aspiring engineer, regardless of their specialization. Basic objective behind this curriculum student should learn real-world applicability in terms of optimizing manufacturing processes, designing efficient systems, or tackling complex engineering challenges. The curriculum is not just about theory; it's about solving tangible problems that have a direct impact on industries, society, and innovation.

AIKTC is offering this curriculum known as "MicroMasters program in Statistics and Data Science" which is effective from July 2023. The program is divided into four semesters:

1. **Statistics for Data Science**
2. **Fundamentals of Machine Learning Algorithms for Data Science**
3. **Fundamentals of Deep Learning algorithms for Data Science**
4. **Natural language processing and Visualization Techniques for Data Science**

The program currently consists of two courses spread over 2 semesters (5th and 6th semester), and a proctored cumulative exam at the end of the 5th and 6th semesters:

Proctor Cumulative Exam

1. **Statistics for Data Science** (5th Sem)
2. **Fundamentals of Machine Learning Algorithms for Data Science** (6th Sem)
3. **Fundamentals of Deep Learning algorithms for Data Science** (7th Sem)
4. **Natural Language Processing and Visualization Techniques for Data Science** (8th Sem)

Each course will be taught by AIKTC faculty and is at a similar pace and level of rigor as an on-campus course at AIKTC. This is roughly equivalent to honors degree course of 2 years at Mumbai University.

2 Prerequisites

While there are no formal pre-requisites for the program, only those students having maximum **2 live KTs in 3rd and/or 4th semester** are allowed to register for the program. Learners are expected to be good at multi-variable calculus, math reasoning and basic linear algebra. Prerequisites for individual courses are listed in respective syllabi of courses.

Semester V

Course Code	Course Name	Credits
MSDS501	Statistics for Data Science	04

Contact Hours			Credits Assigned			
Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
4	-		4	-	-	4

Theory				Term Work/Practical/Oral			Total
Internal Assessment		End Sem Exam	Duration of End Sem Exam	Term Work	Pract.	Oral	
Final project	Total						
40	40	60	02 Hrs.	25	-	-	125

Rationale

Statistics forms the foundation of data analysis, enabling data scientists to draw meaningful insights, make informed decisions, and communicate results effectively.

Objectives

To prepare students for micro-master's program of MIT. This course offers an introduction to probabilistic modelling and statistical inference, which are the keys to analysing data and making scientifically sound predictions. Following are the key objectives of the course:

1. **Understand Fundamental Concepts:** Develop a solid understanding of foundational statistical concepts, including measures of central tendency, variability, and distributions.
2. **Explore Data Descriptively:** Learn to summarize and visualize data using descriptive statistics and graphical representations. Interpret histograms, box plots, scatter plots, and other visualizations to gain insights into data patterns**.**
3. **Handle Data Quality Issues:** Acquire skills to identify and address data quality issues such as missing values, outliers, and inconsistencies. Use statistical techniques to clean and preprocess data effectively**.**
4. **Perform Inferential Analysis:** Learn hypothesis testing techniques to make inferences about population parameters from sample data. Understand concepts like p-values, confidence intervals, and significance levels**.**

5. **Build Predictive Models:** Apply regression analysis to model relationships between variables and predict outcomes. Understand the concepts of correlation, causation, and multicollinearity.
6. **Explore Probability Theory:** Gain a solid foundation in probability theory to understand uncertainty and randomness in data. Apply probability concepts to decision-making and risk assessment.
7. **Understand Sampling Techniques:** Explore different sampling methods and understand their implications for data analysis. Learn about random sampling, stratified sampling, and cluster sampling.
8. **Implement Statistical Software:** Gain hands-on experience with statistical software packages such as R or Python libraries (e.g., NumPy, pandas, statsmodels). Learn how to perform various statistical analyses using these tools.

Detailed Syllabus		
Module	**Sub- Modules/Contents**	**Periods**
I	**Introduction to Statistics and Data** ● Importance of statistics in data science ● Types of data: categorical, numerical, discrete, continuous ● Levels of measurement: nominal, ordinal, interval, ratio ● Data collection and sampling techniques	05
II	**Descriptive Statistics** ● Measures of central tendency: mean, median, mode. ● Measures of variability: range, variance, standard deviation ● Percentiles and quartiles ● Exploratory data analysis (EDA) using graphical representations.	06
III	**Non-Parametric Tests and ANOVA** ● Wilcoxon signed-rank test and Mann-Whitney U test. ● Kruskal-Wallis test for multiple groups. ● Analysis of Variance (ANOVA) and post-hoc tests. ● Introduction to categorical data analysis.	05
IV	**Sampling and Estimation** ● Sampling distributions and sampling error ● Point estimation and confidence intervals ● Margin of error and confidence level ● Hypothesis testing: null and alternative hypotheses	06

V	**Inferential Statistics / Hypothesis Testing** ● One-sample t-tests and z-tests ● Two-sample t-tests and z-tests ● Chi-squared tests for categorical data ● Type I and Type II errors, p-values, and significance levels	08
VI	**Regression and Correlation** ● Simple linear regression: model, assumptions, interpretation ● Multiple regression: multiple predictors, interaction terms ● Correlation analysis and coefficient of determination (R-squared) ● Residual analysis and model diagnostics	05
VII	**Probability and Distributions** ● Fundamentals of probability theory ● Probability distributions: discrete and continuous ● Normal distribution and its properties ● Central Limit Theorem and its implications	05

Course Outcomes

On completion of this course, learners will be able to:

1. Apply Descriptive Statistics and Data Exploration to solve given problem.
2. Analyse Probability Distributions of a given problem.
3. Perform Sampling distribution and its Estimation:
4. Apply Hypothesis Testing Techniques for given situation.
5. Apply Regression and Correlation for complex problem.
6. Perform Non-Parametric Tests and ANOVA on given problem.

Internal Assessment

Assessment Rubric: Statistics for Data Science - Final Project (40 Marks)

1. **Project Understanding and Scope (10 Marks)**

   Excellent (8-10 marks): Demonstrates a comprehensive understanding of the project's scope, objectives, and data requirements. Clearly defines the problem statement and sets appropriate goals for the project.
   Proficient (5-7 marks): Displays a good grasp of the project's scope and objectives. Defines the problem statement adequately and sets reasonable goals.
   Basic (2-4 marks): Shows a limited understanding of the project's scope and objectives. Problem statement and goals may lack clarity or specificity.
   Limited (0-1 marks): Fails to grasp the project's scope, objectives, or problem statement.

2. **Data Collection and Preparation (10 Marks)**

   Excellent (8-10 marks): Collects and prepares data meticulously, demonstrating a deep understanding of data cleaning, transformation, and handling missing values. Data is in a suitable format for analysis.
   Proficient (5-7 marks): Collects and prepares data effectively, addressing most issues related to data quality and missing values. Data is mostly in a suitable format for analysis.
   Basic (2-4 marks): Collects and prepares data but with some issues related to data quality and missing values. Data format for analysis might need improvement.
   Limited (0-1 marks): Data collection and preparation are significantly flawed, with evident data quality and format issues.

3. **Data Analysis and Interpretation (10 Marks)**

   Excellent (8-10 marks): Applies a wide range of advanced statistical techniques to analyze the data. Demonstrates exceptional skill in interpreting results and drawing meaningful insights.
   Proficient (5-7 marks): Applies appropriate statistical techniques for data analysis. Interprets results reasonably well, though there might be minor errors or limitations.
   Basic (2-4 marks): Applies basic statistical techniques with limited depth. Interpretations are shallow or may not fully capture the insights.
   Limited (0-1 marks): Data analysis and interpretations are either incorrect or missing.

4. **Visualizations and Communication (6 Marks)**

   Excellent (5-6 marks): Presents results using a variety of insightful visualizations. Communicates findings clearly and effectively, demonstrating a deep understanding of the data.
   Proficient (3-4 marks): Provides adequate visualizations to support findings. Communication is generally clear but might lack some detail or organization.
   Basic (1-2 marks): Attempts to visualize results but with limited clarity or depth. Communication lacks effective organization.
   Limited (0 marks): Fails to effectively visualize or communicate results.

5. **Overall Project Quality (4 Marks)**

   Excellent (3-4 marks): Produces a high-quality project with thorough documentation, proper formatting, and attention to detail. Demonstrates professionalism in the presentation.
   Proficient (2 marks): Submits a well-organized project with minor documentation or formatting issues.
   Basic (1 mark): Presents a project with noticeable documentation or formatting problems that slightly impact its quality.
   Limited (0 marks): Submits a project with significant documentation or formatting issues that greatly affect its quality.

Term work (25 Marks)

4 assignments to be given each of 5 marks. Further, 5 marks are allotted for attendance. If attendance is above 90%, 5 marks, if between 80-90%, then 4 marks, and if between 75-80%, then 3 marks. Below, 75%, no marks to be given for attendance.

End Semester Examination (60 Marks)

Weightage of each module in end semester examination will be proportional to number of respective lecture hours mentioned in the curriculum.

1 Question paper will comprise of **total five questions, each carrying 20 marks.**

2 Question 1 will be compulsory and should cover **maximum contents of the curriculum.**

3 **Remaining questions will be mixed in nature** (for example if Q.2 has part (a) from module 3 then part (b) will be from any module other than module 3).

4 **Only three questions need to be solved.**

Recommended Books: -

1. **Probability and Statistics for Engineers** –Miller, Freund-Hall, Prentice India Ltd.
2. **Practical Statistics for Data Scientists -** Peter Bruce, Andrew Bruce, and Peter Gedeck
3. **Introduction to Probability and Statistics for Engineers and Scientists**- Sheldon M. Ross.
4. **Statistical Methods for Data Science-** Asim Roy and S. M. Srinivasan
5. **The Art of Data Science-** Roger D. Peng and Elizabeth Matsui

Semester VI

Course Code	Course Name	Credits
MSDS601	Fundamentals of Machine Learning Algorithms for Data Science	04

Contact Hours			Credits Assigned			
Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
4	-	-	4	-	-	4

Theory				Term Work/Practical/Oral			Total
Internal Assessment		End Sem Exam	Duration of End Sem Exam	Term Work	Pract.	Oral	
Final Project	Average						
40	40	60	02 Hrs.	25	-	-	125

Rationale

The rationale for studying fundamental machine learning algorithms in the context of data science is rooted in the crucial role these algorithms play in extracting valuable insights and making informed decisions from data.

Objectives

To prepare students for micro-master's program of MIT, the objectives for the course on "Fundamental Machine Learning Algorithms for Data Science" should encompass a comprehensive understanding of essential concepts and techniques that form the backbone of machine learning in the context of data science. The key objectives are as follows:

1. **Conceptual Understanding**: Develop a clear and foundational understanding of fundamental machine learning concepts, including supervised learning, unsupervised learning, and reinforcement learning.

2. **Algorithm Familiarity**: Gain familiarity with a range of fundamental machine learning algorithms, such as linear regression, logistic regression, decision trees, k-nearest neighbors, clustering algorithms, and dimensionality reduction techniques like PCA.

3. **Application of Algorithms**: Learn how to apply these algorithms to real-world data science problems, including tasks like classification, regression, and clustering. Understand the appropriate use cases and limitations of each algorithm.

4. **Data Pre-processing:** Develop skills in data pre-processing, including handling missing data, normalizing features, dealing with categorical variables, and feature engineering. Recognize the impact of data quality on model performance.

5. **Model Evaluation:** Learn how to evaluate the performance of machine learning models using relevant metrics for different types of tasks (e.g., accuracy, precision, recall, F1-score, mean squared error). Understand techniques for avoiding over fitting and under fitting.

6. **Interpretability:** Understand the importance of model interpretability, especially in fields where explainability is critical. Learn how to interpret the results of linear models, decision trees, and other interpret-able algorithms.

7. **Practical Implementation**: Develop hands-on experience by implementing machine learning algorithms using popular programming libraries such as scikit-learn, Tensor Flow, or PyTorch. Gain the ability to code and experiment with different algorithms on real datasets.

8. **Problem-Solving Skills**: Develop the ability to identify suitable machine learning approaches for different data science problems. Learn how to frame a problem as a machine learning task and select the most appropriate algorithms.

Detailed Syllabus

Module	Sub- Modules/Contents	Periods
I	**Introduction to Machine Learning** 1.1 Machine learning, Its essentials for data science. 1.2 Types of machine learning: supervised, unsupervised, and reinforcement learning. 1.3 Role of machine learning in data-driven decision making. 1.4 Overview of the data science process	04
II	**Regression Models** 2.1 Linear regression: concepts, assumptions, and interpretation. 2.2 Regularization techniques: L1 (Lasso) and L2 (Ridge) regularization. 2.3 Logistic regression for classification problems. 2.4 Non linear regression. 2.4 Hands-on: Implementing linear regression and logistic regression using Python and scikit-learn.	06

III	**Decision Trees and Ensemble Methods** 3.1 Decision tree fundamentals: construction, splitting criteria, and pruning. 3.2 Bagging and random forests for improved model performance. 3.3 Boosting algorithms (AdaBoost, Gradient Boosting) and their advantages. Hands-on: Building decision trees and using ensemble methods for better predictive accuracy.	06
IV	**Clustering and Dimensionality Reduction** 4.1 K-means clustering: algorithm, initialization, and choosing the number of clusters. 4.2 Hierarchical clustering and its applications. 4.3 Principal Component Analysis (PCA) for dimensionality reduction. 4.4 Hands-on: Implementing clustering algorithms and PCA on real datasets.	08
V	**Support Vector Machines (SVM)** 5.1 SVM for binary classification: hyper plane, margins, and kernel tricks. 5.2 Soft-margin SVM and handling non-linearly separable data. 5.3 Kernel functions: linear, polynomial, radial basis function (RBF). 5.4 Hands-on: Implementing SVM with different kernels, tuning hyper parameters.	08
VI	**Model Evaluation and Optimization** 6.1 Model evaluation metrics: accuracy, precision, recall, F1-score, ROC curve, and AUC. 6.2 Techniques to handle over fitting and under fitting. 6.3 Cross-validation for robust model assessment. 6.4 Hyperparmeter tuning and grid search. 6.5 Hands-on: Evaluating models, optimizing hyperparmeter, and improving model performance.	08

Course Outcomes

On completion of this course, learners will be able to:

- Apply machine learning concepts like supervised and unsupervised learning, model training, validation, and testing.
- Choose the appropriate machine learning algorithm for a given problem.
- Perform data pre-processing tasks, such as handling missing data, normalizing features, dealing with categorical variables, and feature engineering.
- Develop machine learning models for various tasks such as classification, regression, and clustering.

Internal Assessment

Assessment Rubric: Final Project - (40 Marks)

1. **Project Understanding and Problem Definition (8 Marks)**

Excellent (6-8 marks): Demonstrates a clear and comprehensive understanding of the project’s problem statement and its significance in the context of machine learning for data science.
Proficient (4-5 marks): Shows a good understanding of the project's problem statement but might have some minor gaps or ambiguities.
Basic (2-3 marks): Has a limited understanding of the problem statement, and the significance of the problem might not be well-addressed.
Limited (0-1 marks): Struggles to define the problem and its relevance in the context of machine learning.

2. **Algorithm Selection and Implementation (10 Marks)**

Excellent (8-10 marks): Chooses appropriate machine learning algorithms and implements them accurately, demonstrating a deep understanding of the algorithms' concepts and mechanics.
Proficient (5-7 marks): Selects and implements machine learning algorithms correctly, but there might be minor issues or gaps in the implementation.
Basic (3-4 marks): Chooses and implements algorithms, but the understanding and implementation might be basic or lack depth.
Limited (0-2 marks): Algorithm selection and implementation are flawed, with significant inaccuracies or errors.

3. **Experimentation and Model Evaluation (12 Marks)**

Excellent (9-12 marks): Conducts thorough experiments, hyperparameter tuning, and model evaluation. Demonstrates an advanced ability to interpret and analyze the results.
Proficient (6-8 marks): Conducts experiments and model evaluation appropriately, though there might be some areas for improvement in experimentation or result interpretation.
Basic (3-5 marks): Attempts to conduct experiments and evaluation but with limited depth or understanding of the results.
Limited (0-2 marks): Experimentation and model evaluation lack coherence or accuracy.

4. **Innovation and Adaptation (6 Marks)**

Excellent (5-6 marks): Demonstrates innovative thinking by proposing creative modifications or adaptations to improve model performance or address specific challenges.
Proficient (3-4 marks): Suggests reasonable adaptations or improvements to the models, even if they are not highly innovative.
Basic (1-2 marks): Provides basic suggestions for improvement without much creativity or depth.
Limited (0 marks): Fails to propose any meaningful adaptations or improvements.

5. **Presentation and Documentation (4 Marks)**

Excellent (3-4 marks): Presents the project with exceptional clarity, organization, and professionalism. Documentation is thorough, well-structured, and easy to understand.
Proficient (2 marks): Presents the project clearly and meets documentation requirements, but there might be minor issues in organization or clarity.
Basic (1 mark): Presents the project with noticeable organization or clarity issues, affecting its overall quality.
Limited (0 marks): Presents the project with significant documentation or presentation issues that hinder its comprehension.

Term work (25 Marks)

4 assignments to be given each of 5 marks. Further, 5 marks are allotted for attendance. If attendance is above 90%, 5 marks, if between 80-90%, then 4 marks, and if between 75-80%, then 3 marks. Below, 75%, no marks to be given for attendance.

End Semester Examination (60 Marks)

Weightage of each module in end semester examination will be proportional to number of respective lecture hours mentioned in the curriculum.

1 Question paper will comprise of **total Five questions, each carrying 20 marks.**

2 Question 1 will be compulsory and should cover **maximum contents of the curriculum.**

3 **Remaining questions will be mixed in nature** (for example if Q.2 has part (a) from module 3 then part (b) will be from any module other than module 3).

4 **Only Three questions need to be solved.**

Recommended Books: -

1. Pattern Recognition and Machine Learning" by Christopher M. Bishop
2. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensor Flow" by Aurélien Géron
3. "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili
4. Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido

Semester VII

Course Code	Course Name	Credits
MSDS701	**Fundamentals of Deep Learning algorithms for Data Science**	04

Contact Hours			Credits Assigned			
Theory	**Practical**	**Tutorial**	**Theory**	**Practical**	**Tutorial**	**Total**
4	-	-	4	-	-	4

Theory				Term Work/Practical/Oral			Total
Internal Assessment		**End Sem Exam**	**Duration of End Sem Exam**	**Term Work**	**Pract.**	**Oral**	
Final Project	**Total**						
40	40	60	02 Hrs.	25	-	-	125

Rationale

The rationale for a course on "Fundamental Deep Learning Algorithms for Data Science" is rooted in the significant impact deep learning has had on the field of data science, providing powerful tools for extracting intricate patterns and insights from complex datasets.

Objectives

Upon completing this course, students should be able to:

1. **Comprehensive Understanding of Deep Learning**: Develop a solid grasp of the core concepts and principles of deep learning, including neural network architectures, activation functions, loss functions, optimization algorithms, and the backpropagation algorithm.
2. **Knowledge of Key Deep Learning Models:** Gain familiarity with essential deep learning models, such as feedforward neural networks, convolutional neural networks (CNNs) for image analysis, recurrent neural networks (RNNs) for sequential data, and transformer architectures for natural language processing (NLP).
3. **Practical Implementation Skills**: Learn how to implement deep learning models using popular frameworks like Tensor Flow or PyTorch. Gain hands-on experience in building, training, and evaluating deep learning models on real-world datasets.
4. **Feature Learning and Data Representation**: Understand the concept of feature learning in deep learning, and how deep neural networks can automatically extract relevant features from raw data.

5. **Hyperparmeter Tuning and Regularization:** Develop skills in hyperparmeter tuning to optimize model performance. Learn about regularization techniques, dropout, and other strategies to prevent overfitting in deep learning models.
6. **Transfer Learning and Pretrained Models:** Explore transfer learning by leveraging pretrained deep learning models, such as using CNNs pretrained on large image datasets.
7. **Advanced Deep Learning Topics**: Introduce advanced topics in deep learning, such as generative adversarial networks (GANs), auto encoders, and deep reinforcement learning. Provide students with an overview of the capabilities and applications of these advanced techniques.
8. **Real-World Applications:** Showcase real-world applications of deep learning in data science across various domains, highlighting the impact of deep learning in fields such as computer vision, natural language processing, healthcare, finance, and more.

Detailed Syllabus

Module	Sub- Modules/Contents	Periods
I	**Introduction to Deep Learning** ● Deep Learning Fundamentals. ● Historical context and evolution of neural networks. ● Key concepts: Neurons, activation functions, layers, architectures ● Gradient descent and backpropagation	04
II	**Convolutional Neural Networks (CNNs)** ● Introduction to image data ● Convolutional layers ● Pooling layers ● Architecture of CNNs (e.g., LeNet, AlexNet, VGG, ResNet) ● Hands on: Applications of CNNs in computer vision or healthcare or Agriculture	08
III	**Recurrent Neural Networks (RNNs)** ● Introduction to sequential data ● Basic RNN architecture ● Long Short-Term Memory (LSTM) networks ● Gated Recurrent Units (GRUs) ● Hands on: Applications of RNNs in natural language processing, time series analysis, etc.	06

IV	**Advanced Deep Learning Models** ● Generative Adversarial Networks (GANs) ● Variational Auto encoders (VAEs) ● Transfer learning and fine-tuning ● Attention mechanisms	08
V	**Deep Learning for Data Science Applications** ● Introduction to TensorFlow or PyTorch (hands-on exercises) ● Building and training deep learning models ● Evaluation metrics for deep learning ● Ethical considerations in deep learning	08
VI	**Final Project work** Students may work on a deep learning project along with trainer related to data science, where they apply the concepts learned throughout the course to a real-world problem.	06

Course Outcomes

On completion of this course, learners will be able to:

- Develop and Apply CNN Architecture on a given problem.
- Apply appropriate RNN architecture to solve real world application.
- Perform data pre-processing tasks on different types of images, such as noise removal, image enhancement, Data augmentation, and feature engineering.
- Develop deep learning models for various tasks such as Image classification, object detection.
- Validate Deep Learning Algorithms for Data Science Application.

Assessment Rubric: Final Project - (40 Marks)

1. **Project Understanding and Scope (8 Marks)**

Excellent (6-8 marks): Demonstrates a clear and comprehensive understanding of the project's scope, problem statement, and the relevance of deep learning algorithms for addressing the problem.

Proficient (4-5 marks): Shows a good understanding of the project's scope and problem statement, but there might be minor gaps or ambiguities.

Basic (2-3 marks): Has a limited understanding of the project's scope and problem statement, and the relevance of deep learning might not be well-addressed.

Limited (0-1 marks): Struggles to define the project's scope and problem statement, and the relevance of deep learning is unclear.

2. **Algorithm Selection and Implementation (10 Marks)**

Excellent (8-10 marks): Selects and implements appropriate deep learning algorithms accurately, demonstrating an in-depth understanding of their architecture, training process, and implementation nuances.

Proficient (5-7 marks): Chooses and implements deep learning algorithms correctly, but there might be minor issues or gaps in the implementation.

Basic (3-4 marks): Chooses and implements algorithms, but the understanding and implementation might be basic or lack depth.

Limited (0-2 marks): Algorithm selection and implementation are flawed, with significant inaccuracies or errors.

3. **Experimentation and Model Evaluation (12 Marks)**

Excellent (9-12 marks): Conducts thorough experiments, hyperparameter tuning, and model evaluation. Demonstrates advanced ability to interpret and analyze deep learning model results.

Proficient (6-8 marks): Conducts experiments and model evaluation appropriately, though there might be some areas for improvement in experimentation or result interpretation.

Basic (3-5 marks): Attempts to conduct experiments and evaluation but with limited depth or understanding of the results.

Limited (0-2 marks): Experimentation and model evaluation lack coherence or accuracy.

4. **Innovation and Adaptation (6 Marks)**

Excellent (5-6 marks): Demonstrates innovative thinking by proposing creative modifications or adaptations to improve model performance or address specific challenges within deep learning.

Proficient (3-4 marks): Suggests reasonable adaptations or improvements to the deep learning models, even if they are not highly innovative.

Basic (1-2 marks): Provides basic suggestions for improvement without much creativity or depth.

Limited (0 marks): Fails to propose any meaningful adaptations or improvements.

5. **Presentation and Documentation (4 Marks)**

Excellent (3-4 marks): Presents the project with exceptional clarity, organization, and professionalism. Documentation is thorough, well-structured, and easy to understand.

Proficient (2 marks): Presents the project clearly and meets documentation requirements, but there might be minor issues in organization or clarity.

Basic (1 mark): Presents the project with noticeable organization or clarity issues, affecting its overall quality.

Limited (0 marks): Presents the project with significant documentation or presentation issues that hinder its comprehension.

Term work (25 Marks)

4 assignments to be given each of 5 marks. Further, 5 marks are allotted for attendance. If attendance is above 90%, 5 marks, if between 80-90%, then 4 marks, and if between 75-80%, then 3 marks. Below, 75%, no marks to be given for attendance.

End Semester Examination (60 Marks)

Weightage of each module in end semester examination will be proportional to number of respective lecture hours mentioned in the curriculum.

1 Question paper will comprise of **total Five questions, each carrying 20 marks.**

2 Question 1 will be compulsory and should cover **maximum contents of the curriculum.**

3 **Remaining questions will be mixed in nature** (for example if Q.2 has part (a) from module 3 then part (b) will be from any module other than module 3).

4 **Only three questions need to be solved.**

Recommended Books:

1. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron
2. "Pattern Recognition and Machine Learning" by Christopher M. Bishop
3. "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
4. "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andrew G. Barto.

Semester VIII

Course Code	Course Name	Credits
MSDS801	**Natural Language Processing and Visualization Techniques for Data Science**	04

Contact Hours			Credits Assigned			
Theory	**Practical**	**Tutorial**	**Theory**	**Practical**	**Tutorial**	**Total**
4	-	-	4	-	-	4

Theory				Term Work/Practical/Oral			Total
Internal Assessment		**End Sem Exam**	**Duration of End Sem Exam**	**Term Work**	**Pract.**	**Oral**	
Final Project	**Total**						
40	40	60	02 Hrs.	25	-	-	125

Rationale

NLP and visualization techniques broadens the scope of data science by allowing professionals to extract insights and make informed decisions from text data, which was previously challenging to analyze using traditional data analysis techniques. Its applications span across industries and domains, contributing significantly to data-driven decision-making and innovation.

Objectives

After completion of this course the student will be equipped to work with diverse types of data, extract meaningful insights, and present those insights in ways that are accessible and impactful for various stakeholders. Following are the key are objectives for studying the subjects.

1. **Text data understanding and sentimental analysis:** Gain a deep understanding of the techniques and tools used to process, analyze, and extract insights from text data, enabling you to work with unstructured textual information. Learn how to analyze sentiments, opinions, and emotions expressed in text, which is crucial for understanding customer feedback, social media trends, and public sentiment.
2. **Text Classification and Categorization and Named Entity Recognition (NER):** To learn how to automatically classify text data into predefined categories, which has applications in spam filtering, topic modeling, and content recommendation. Develop

skills to identify and extract entities such as names, locations, organizations, and dates from text, important for information extraction and knowledge management.

3. **Text Generation:** Gain the ability to generate coherent and contextually relevant text, useful for tasks like chatbots, content creation, and language translation.
4. **Language Understanding and Information Retrieval**: Understand how to build systems that can understand the nuances and context of human language, aiding in building effective communication channels with users. Learn techniques for retrieving relevant information from large text datasets, helping in search engines and recommendation systems.
5. **Data Representation and Data Exploration**: Learn how to effectively represent complex data in visual forms such as charts, graphs, and dashboards, making it easier to grasp patterns and insights. Develop skills to explore and discover patterns, trends, and outliers in datasets through interactive visualizations.
6. **Interactive Dashboards:** Gain proficiency in designing interactive dashboards that allow users to explore data dynamically and make informed decisions.
7. **Geospatial and Temporal Data Visualization**: Learn techniques to visualize geographical data, enabling insights in fields like geography, urban planning, and logistics. Explore methods to visualize time-series data, helping in understanding trends, seasonality, and anomalies.
8. **Tools and Technologies:** Gain hands-on experience with popular data visualization tools and libraries like Matplotlib, Seaborn, Tableau, and D3.js.

Detailed Syllabus

Module	Sub- Modules/Contents	Periods
I	**Introduction to NLP** ● Introduction to NLP and its applications ● Basics of text data preprocessing: tokenization, stemming, lemmatization ● Part-of-speech tagging and named entity recognition. ● Introduction to NLP libraries: NLTK, spaCy.	06
II	**Text Representation** ● Bag-of-words model and its limitations ● TF-IDF (Term Frequency-Inverse Document Frequency) representation ● Word embeddings: Word2Vec and GloVe ● Sentiment analysis: determining sentiment from text.	06

III	**NLP Techniques** ● Text classification: Naive Bayes, SVM, and neural networks ● Topic modeling: Latent Dirichlet Allocation (LDA) ● Sequence labeling: Hidden Markov Models (HMM) and Conditional Random Fields (CRF) ● Language generation: basics of text generation using RNNs.	06
IV	**Data Visualization Fundamentals** ● Importance of data visualization in data science ● Data visualization principles and best practices ● Introduction to data visualization libraries: Matplotlib and Seaborn	06
V	**Advanced Data Visualization** ● Interactive visualization using Plotly. ● Geographic data visualization with GeoPandas. ● Visualizing large datasets: techniques for handling and displaying big data ● Visualization with D3.js (introductory concepts)	08
VI	**Advanced Visualization Techniques** ● Time series data visualization ● Network visualization and graph analysis ● 3D visualization and virtual reality (introductory concepts) ● Dashboard creation using tools like Tableau or Power BI ● Final Project and Wrap-up: Students work on a final project combining NLP analysis and data visualization. ● Project presentations and peer feedback.	08

Course Outcomes

On completion of this course, learners will be able to:

1. Apply Text Representation Techniques for sentiment analysis.
2. Apply NLP Techniques for given problems.
3. Analyze different Data Visualization techniques.
4. Implement Advanced Data Visualization Techniques.
5. Design and Develop data science application for real world problem.

Internal Assessment
Assessment Rubric: Final Project - (40 Marks)

1. **Project Understanding and Problem Definition (8 Marks)**
Excellent (6-8 marks): Demonstrates a clear and comprehensive understanding of the project's problem statement, emphasizing the role of natural language processing and visualization techniques.
Proficient (4-5 marks): Shows a good understanding of the project's problem statement, but there might be minor gaps or ambiguities in explaining the significance of NLP and visualization.
Basic (2-3 marks): Has a limited understanding of the project's problem statement and might not fully address the importance of NLP and visualization.
Limited (0-1 marks): Struggles to define the problem statement and does not adequately address the relevance of NLP and visualization.

2. **Data Preprocessing and NLP Techniques (10 Marks)**

Excellent (8-10 marks): Performs comprehensive data preprocessing and applies advanced natural language processing techniques accurately and effectively.
Proficient (5-7 marks): Conducts suitable data preprocessing and applies NLP techniques accurately, with minor issues or omissions.
Basic (3-4 marks): Performs basic data preprocessing and NLP techniques but with limited depth or understanding.
Limited (0-2 marks): Data preprocessing and NLP techniques are flawed or significantly incomplete.

3. **Visualization and Communication (8 Marks)**

Excellent (6-8 marks): Creates insightful visualizations that effectively communicate complex linguistic patterns. Demonstrates a deep understanding of visualization principles.
Proficient (4-5 marks): Generates suitable visualizations that support the findings but might have minor issues in clarity or insightfulness.
Basic (2-3 marks): Attempts to visualize results but lacks depth or creativity in visualization choices.
Limited (0-1 marks): Visualizations are missing or fail to communicate meaningful insights.

4. **Experimentation and Analysis (10 Marks)**

Excellent (8-10 marks): Conducts rigorous experiments involving NLP techniques, interprets results skillfully, and derives meaningful insights.
Proficient (5-7 marks): Conducts experiments and interprets results appropriately, but there might be minor gaps or limitations in analysis.
Basic (3-4 marks): Attempts experimentation and analysis, but lacks depth in interpreting results or deriving insights.
Limited (0-2 marks): Experimentation and analysis are either incorrect or insufficient.

5. **Presentation and Documentation (4 Marks)**

Excellent (3-4 marks): Presents the project with exceptional clarity, organization, and professionalism. Documentation is thorough, well-structured, and easy to understand.

Proficient (2 marks): Presents the project clearly and meets documentation requirements, but there might be minor issues in organization or clarity.
Basic (1 mark): Presents the project with noticeable organization or clarity issues, affecting its overall quality.
Limited (0 marks): Presents the project with significant documentation or presentation issues that hinder its comprehension.

Term work (25 Marks)

4 assignments to be given each of 5 marks. Further, 5 marks are allotted for attendance. If attendance is above 90%, 5 marks, if between 80-90%, then 4 marks, and if between 75-80%, then 3 marks. Below, 75%, no marks to be given for attendance.

End Semester Examination (60 Marks)

Weightage of each module in end semester examination will be proportional to number of respective lecture hours mentioned in the curriculum.

1 Question paper will comprise of **total Five questions, each carrying 20 marks.**

2 Question 1 will be compulsory and should cover **maximum contents of the curriculum.**

3 **Remaining questions will be mixed in nature** (for example if Q.2 has part (a) from module 3 then part (b) will be from any module other than module 3).

4 **Only three questions need to be solved.**

Recommended Books:

1. **Natural Language Processing with Python** by Steven Bird, Ewan Klein, and Edward Loper
2. **Speech and Language Processing** by Dan Jurafsky and James H. Martin
3. **Foundations of Statistical Natural Language Processing** by Christopher D. Manning and Hinrich Schütze
4. **Storytelling with Data: A Data Visualization Guide for Business Professionals** by Cole Nussbaumer Knaflic.
5. **Python for Data Analysis** by Wes McKinney
6. **Data Visualization: A Practical Introduction** by Kieran Healy
7. **Information Dashboard Design: Displaying Data for At-a-Glance Monitoring** by Stephen Few