Savitribai Phule Pune University
(Formerly University of Pune)
DEPARTMENT OF TECHNOLOGY

Foundation Course in Data Science

Savitribai Phule Pune University
(Formerly University of Pune)
DEPARTMENT OF TECHNOLOGY

1. **About the course:**

Brief Job Description: This job role is responsible for assisting in preparation of data science related documents and initiate data science awareness campaigns within the organization and with associates. The main duties consist of assisting superiors in the preparation of data science policy documents, preparing training material for data science awareness campaigns, coordinating with departments in matters of data science issues and assisting superiors in the conduct of data science analytics.

Personal Attributes: This job may require the individual to work independently in the collection of data and make decisions for his/her own area of work. The individual should have a high level of documentation drafting capability, passion for data science and attention for detail, should be ethical, compliance and result- oriented, should also be able to demonstrate interpersonal skills, along with willingness to undertake desk-based job with long working hours.

2. **Aim & Objectives of the Course:**

- **Understanding Data**: To help students understand the fundamentals of data, including its types, sources, and formats.
- **Data Analysis Techniques**: To teach students various techniques and methods for analyzing data, including statistical analysis, machine learning algorithms, and data visualization.
- **Programming Skills**: To develop students' proficiency in programming languages commonly used in data science, such as Python, R, and SQL.
- **Data Preprocessing**: To train students in data preprocessing techniques, including cleaning, transformation, and normalization of data.
- **Machine Learning and Predictive Modeling**: To provide students with a strong foundation in machine learning concepts and techniques for building predictive models from data.
- **Data Visualization**: To enable students to effectively communicate insights from data through visualization techniques and tools.
- **Big Data Technologies**: Depending on the course, there may be an objective to introduce students to big data technologies and platforms like Hadoop, Spark, and NoSQL databases.
- **Ethical Considerations**: To instill an understanding of the ethical considerations and responsibilities associated with working with data, including privacy, security, and bias issues.
- **Real-world Applications**: To expose students to real-world applications of data science across various industries, including healthcare, finance, marketing, and social sciences.
- **Problem-Solving Skills**: To foster critical thinking and problem-solving skills necessary for tackling complex data-related challenges.
- **Teamwork and Collaboration**: To promote teamwork and collaboration skills through group projects and activities.
- **Continuous Learning**: To encourage a mindset of continuous learning and adaptation to new tools, techniques, and technologies in the rapidly evolving field of data science.

3. **Title of the course: Foundation course in Data science**

4. Abbreviation of the Course:

CDS

5. Academic year in which course is to be initiated

Academic year	Open to	Examination
2024-25	All Students	End of academic year

6. Eligibility criteria for admission to the course:

12th Standard OR Equivalent.

7. Structure of the course (course duration):

(45 hours)

Modules	Mandatory / Optional	Estimated size (learning hours)
CDS – Module 1 Fundamentals of Data science	Mandatory	1
CDS – Module 2 Fundamentals of Statistics for Data Science	Mandatory	6
CDS – Module 3 Programming Fundamentals	Mandatory	10
CDS – Module 4 Exploratory Data Analysis (EDA)	Mandatory	8
CDS – Module 5 Machine Learning	Mandatory	10
CDS – Module 6 Feature Engineering	Mandatory	5
CDS – Module 7 Model Evaluation and Validation	Mandatory	5
Total		45

8. **Fee Structure:** The tuition fees and laboratory fees and other fees must be paid at the time of admission to the course. Students can opt for any one of the three learning modes. The course fees are as follows.

Blended Online Learning: _____

9. **Teaching scheme of the course (mode of teaching and learning):**

Blended Online Learning- Self Paced

10. **Examination system:**

On successful completion of examination, students will be awarded Certificate in Data science by the University and Skills Factory Learning Private Limited jointly. The examination pattern for this Course is as follows:

Code	Title	Type	Distribution of marks			Credits
			Internal	Univ.	Total	
CDS	Certificate in Data science	Theory, Practical cum Assignments	30	70	100	2

Mandatory Internal and External Evaluation: The course evaluation has two components: Internal assessment for 30 marks that may contain assignment/oral/viva/internal test, etc and the External examination will online mode for 70 marks.

11. **Procedure for conducting External and Internal assessment:**

Online assignment submission and online examination as per University Rules and schedule.

12. **Grade System**

- The examination outcome will not affect the regular academic examinations.
- Students will be awarded grades on the basis of marks achieved.
- The examination will be conducted for all modules at the end of the course.

- Criteria for assessment will be created by DIVY IT Solution (DIS). Each performance criterion@ will be assigned Theory and Skill Based Practical marks proportional to its importance in the module.

  The assessment will be conducted online through DIVY IT Solution (DIS).

- Format of questions will include a variety of styles suitable to the Performance Criteria (PC) being tested such as multiple-choice questions, programming questions, True and False, situational judgment, etc.

@Performance Criteria are statements that together specify the standard of performance required when carrying out an exam task.

13. Rules for Performance Improvement examination

- A Performance Improvement examination will be held by DIS after one month from the date of declaration of result.
- This Re-examination fees will be charged at _____

14. Award of grades and credit allocation:

Marks	Grade
80 and above	A+
70 to 79	A
60 to 69	B
41 to 59	C
40 and less	D

External (Online) evaluation comprises of the aforesaid modules on the basis of following components:

Module	Examination Pattern	Type of questions	Marks
CDS – Module 1 Fundamentals of Data science	Online Examination	Objective questions, MCQs	5
CDS – Module 2 Fundamentals of Statistics for Data Science	Online Examination	Objective questions, MCQs	10
CDS – Module 3 Programming Fundamentals	Online Examination	Objective questions, MCQs	15
CDS – Module 4 Exploratory Data Analysis (EDA)	Online Examination	Objective questions, MCQs	20
CDS – Module 5 Machine Learning	Online Examination	Objective questions	20

CDS – Module 6 Feature Engineering	Online Examination	Objective questions	15
CDS – Module 7 Model Evaluation and Validation	Online Examination	Objective questions	15
		Total	**100**

15. Basis for allocation of marks:

Course-related practical work will be entirely based on the skills to be developed in the students.It would include the topics as has been prescribed in the syllabi of every module

Internal Assessment: 30 marks

Practical components may be based on laboratory work, project, presentation etc. unless otherwise clearly specified in the syllabi of the modules

University Examination (External Evaluation): 70 marks

The marks would be clubbed with the internal assessment for the award of grades.

The Certificate in Data science course is divided into 2 credits:

Sr. No.	Module No.	Module Name	No. of Credits	No of Hours	Marks
1	Module 1, 2, 3	Fundamentals of Data science	One credit for theory part and one credit for practical part totalling two credits as per syllabus	17	20
		Fundamentals of Statistics for Data Science			
		Programming Fundamentals			
2	Module 4,5	Exploratory Data Analysis (EDA)		18	20
		Machine Learning			

4	Module 6	Feature Engineering		10	20
		Model Evaluation and Validation			

Savitribai Phule Pune University

(Formerly University of Pune)

DEPARTMENT OF TECHNOLOGY

Syllabus

16. Model Syllabus for Certificate in Data science:

Sl. No.	Module Name	Topics	Objective	Theory	Practical cum Assignment
				15	30
1	Fundamentals of Data science	a) Definition of Data science and its important in present scenario	a. Explain the fundamentals of Data science and the relevance of Data science	1 Hrs	
		b) Life Cycle of Data science	b. Discuss different disciplines covered under Data science.		
		c) Different disciplines of Data science	c. Discuss differences between technical branches and Data science.		
		d) Roles in Data Science	d. Explain the various posts/profiles in Companies.		
2	Fundamentals of Statistics for Data Science	a) Introduction to Statistics	a. Overview of statistics and its importance in data science b. Descriptive vs. inferential statistics c. Types of data: categorical, numerical (discrete vs. continuous)	6 Hrs	
		b) Data Visualization and Summary Statistics	a. Graphical methods for data visualization (histograms, box plots, scatter plots) b. Measures of central tendency (mean, median, mode) and variability (range, variance, standard deviation)		
		c) Probability Basics	a. Introduction to probability theory b. Probability rules and laws (addition rule, multiplication rule, conditional probability) c. Probability distributions (discrete and continuous)		

		d) Sampling Distributions	a. Introduction to sampling distributions b. Central Limit Theorem and its implications for statistical inference c. Confidence intervals and hypothesis testing		
		e) Statistical Inference	a. Point estimation and interval estimation b. Hypothesis testing: null and alternative hypotheses, significance level, p-values c. Type I and Type II errors		
		f) Parametric Tests	a. One-sample and two-sample t-tests b. Analysis of variance (ANOVA) for comparing means of multiple groups c. Assumptions and limitations of parametric tests		
		g) Non-parametric Tests	a. Wilcoxon signed-rank test and Mann-Whitney U test for comparing two groups b. Kruskal-Wallis test for comparing multiple groups c. Introduction to chi-square tests for categorical data		
		h) Regression Analysis	a. Simple linear regression: model formulation, estimation, interpretation b. Multiple linear regression: model extension to multiple predictors, model diagnostics c. Introduction to logistic regression for binary classification		
3	Programming Fundamentals	a) Introduction to Python Basics	a. Python syntax and structure b. Variables, data types, and type conversion c. Basic input/output operations	6 Hrs	4 Hrs

		b) Control Structures and Functions	a. Control flow: if statements, loops (for, while) b. Functions: defining functions, parameters, return values c. List comprehensions for concise data manipulation		
		c) Essential Libraries for Data Science	a. Introduction to NumPy: arrays, array creation, indexing, operations b. Introduction to Pandas: Series, DataFrame, data manipulation, indexing c. Basic data cleaning and manipulation tasks using Pandas		
		d) Data Visualization with Matplotlib	a. Introduction to Matplotlib: basic plots (line, scatter, bar), customization b. Simple data visualization exercises using Matplotlib c. Introduction to Seaborn for higher-level statistical visualization (optional)		
4	Exploratory Data Analysis (EDA)	a) Introduction to Exploratory Data Analysis	Overview of EDA and its importance in the data science process Principles of data exploration: understanding the data, identifying patterns, and formulating hypotheses	6 Hrs	4 Hrs
		b) Data Cleaning and Preprocessing	Techniques for handling missing values, outliers, and inconsistencies in data Data normalization and scaling for better analysis Feature engineering: creating new features and transforming existing ones		
		c) Bivariate Analysis	Exploring relationships between variables: correlation, covariance Visualizing bivariate data: scatter plots, pair plots, heatmaps		
		d) Multivariate Analysis	Techniques for analyzing relationships among multiple variables Dimensionality reduction techniques: principal component analysis (PCA), t-SNE		

			Visualizing multivariate data: parallel coordinates, Andrews curves		
		e) Advanced Topics in EDA	Time series analysis: exploring temporal patterns and trends Spatial analysis: visualizing geographical data and spatial patterns Interactive data exploration using tools like Plotly or Bokeh		
5	Machine Learning	a) Introduction to Machine Learning	Overview of machine learning concepts and types of machine learning algorithms Supervised, unsupervised, and semi-supervised learning Applications of machine learning in various domains	5 Hrs	5 Hrs
		b) Data Preprocessing	Data cleaning techniques: handling missing values, outliers, and inconsistencies Data transformation and scaling for better model performance Feature engineering: creating new features and selecting relevant ones		
		c) Supervised Learning Algorithms	Linear regression for continuous target variables Logistic regression for binary classification Decision trees and ensemble methods (bagging, boosting, random forests)		
		d) Evaluation Metrics and Model Selection	Common evaluation metrics for regression and classification tasks (e.g., RMSE, MAE, accuracy, precision, recall, F1-score) Cross-validation techniques for assessing model performance Hyperparameter tuning and model selection strategies		
		e) Unsupervised Learning Algorithms	K-means clustering for partitioning data into clusters Hierarchical clustering for creating dendrograms Dimensionality reduction techniques: principal component analysis (PCA), t-SNE		
		f) Neural Networks and Deep Learning	Introduction to artificial neural networks (ANNs) Deep learning architectures: convolutional neural networks (CNNs), recurrent neural networks (RNNs) Training neural networks using backpropagation and gradient descent		

		g) Advanced Topics in Machine Learning	Support vector machines (SVMs) for classification and regression tasks Time series analysis and forecasting techniques Introduction to reinforcement learning and its applications		
6	Feature Engineering	a) Introduction to Feature Engineering	Overview of feature engineering and its importance in machine learning Principles of feature selection and feature extraction Common challenges and considerations in feature engineering	3 Hrs	2 Hrs
		b) Data Exploration and Understanding	Exploratory Data Analysis (EDA) techniques for understanding relationships between features and target variables Identifying relevant features using statistical analysis and visualization Handling missing data and outliers in feature engineering		
		c) Feature Transformation and Creation	Techniques for transforming numerical features: scaling, normalization, log transformation Encoding categorical variables: one-hot encoding, label encoding, target encoding Feature creation methods: polynomial features, interaction features, binning		
		d) Advanced Feature Engineering Techniques	Handling date and time features: extracting temporal information, creating lag features Handling text data: text preprocessing techniques, feature extraction from text Feature selection techniques: filter methods, wrapper methods, embedded methods		

7	Model Evaluation and Validation	a) Introduction to Model Evaluation and Validation	Overview of model evaluation and validation in the machine learning pipeline Importance of proper evaluation metrics and validation techniques Bias-variance tradeoff and its implications for model performance	3 Hrs	2 Hrs
		b) Evaluation Metrics for Classification Models	Common evaluation metrics for classification tasks: accuracy, precision, recall, F1-score, ROC-AUC Interpreting confusion matrices and ROC curves Choosing appropriate metrics based on business objectives and class imbalance		
		c) Evaluation Metrics for Regression Models	Common evaluation metrics for regression tasks: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), R-squared Understanding residual plots and their interpretation Limitations and considerations for different regression metrics		
		d) Cross-Validation Techniques	K-fold cross-validation and its variants: stratified K-fold, repeated K-fold Leave-one-out cross-validation (LOOCV) and leave-p-out cross-validation Time series cross-validation for temporal data		
		e) Model Selection and Hyperparameter Tuning	Strategies for model selection: grid search, random search, Bayesian optimization Hyperparameter tuning techniques to optimize model performance Practical considerations for hyperparameter tuning and computational resources		
		f) Advanced Topics in Model Evaluation	Model interpretability and explainability techniques Model evaluation for ensemble methods: bagging, boosting, stacking Handling model deployment and monitoring in production environments		

17. **Course Outcome:**

- **Proficiency in Data Handling**: Students will gain proficiency in handling and manipulating data using various tools and libraries such as Python, R, Pandas, and NumPy.
- **Statistical Analysis Skills**: Students will develop a solid understanding of statistical concepts and techniques necessary for data analysis, including descriptive statistics, inferential statistics, and hypothesis testing.
- **Machine Learning Knowledge**: Students will learn the principles and techniques of machine learning, including supervised and unsupervised learning algorithms, model evaluation, and hyperparameter tuning.
- **Data Visualization Proficiency**: Students will be able to effectively communicate insights from data through data visualization techniques using libraries such as Matplotlib, Seaborn, and Plotly.
- **Problem-Solving Skills**: Students will develop problem-solving skills necessary to tackle real-world data science challenges, including identifying appropriate methodologies and algorithms for given problems.
- **Project Experience**: Students will work on hands-on projects throughout the course, applying their knowledge and skills to solve real-world data science problems.
- **Ethical Awareness**: Students will gain an understanding of the ethical considerations and responsibilities associated with working with data, including privacy, security, and bias issues.
- **Effective Communication**: Students will learn to effectively communicate their findings and insights from data analysis to stakeholders through written reports and presentations.
- **Continuous Learning Mindset**: Students will develop a mindset of continuous learning and adaptation to new tools, techniques, and technologies in the rapidly evolving field of data science.
- **Preparation for Data Science Careers**: By the end of the course, students will be well-prepared for careers in data science, equipped with the necessary skills and knowledge to succeed in various industries and roles.

18. **Book Recommendation:**

Certificate in Data science textbook by DIVY IT Solution (DIS).