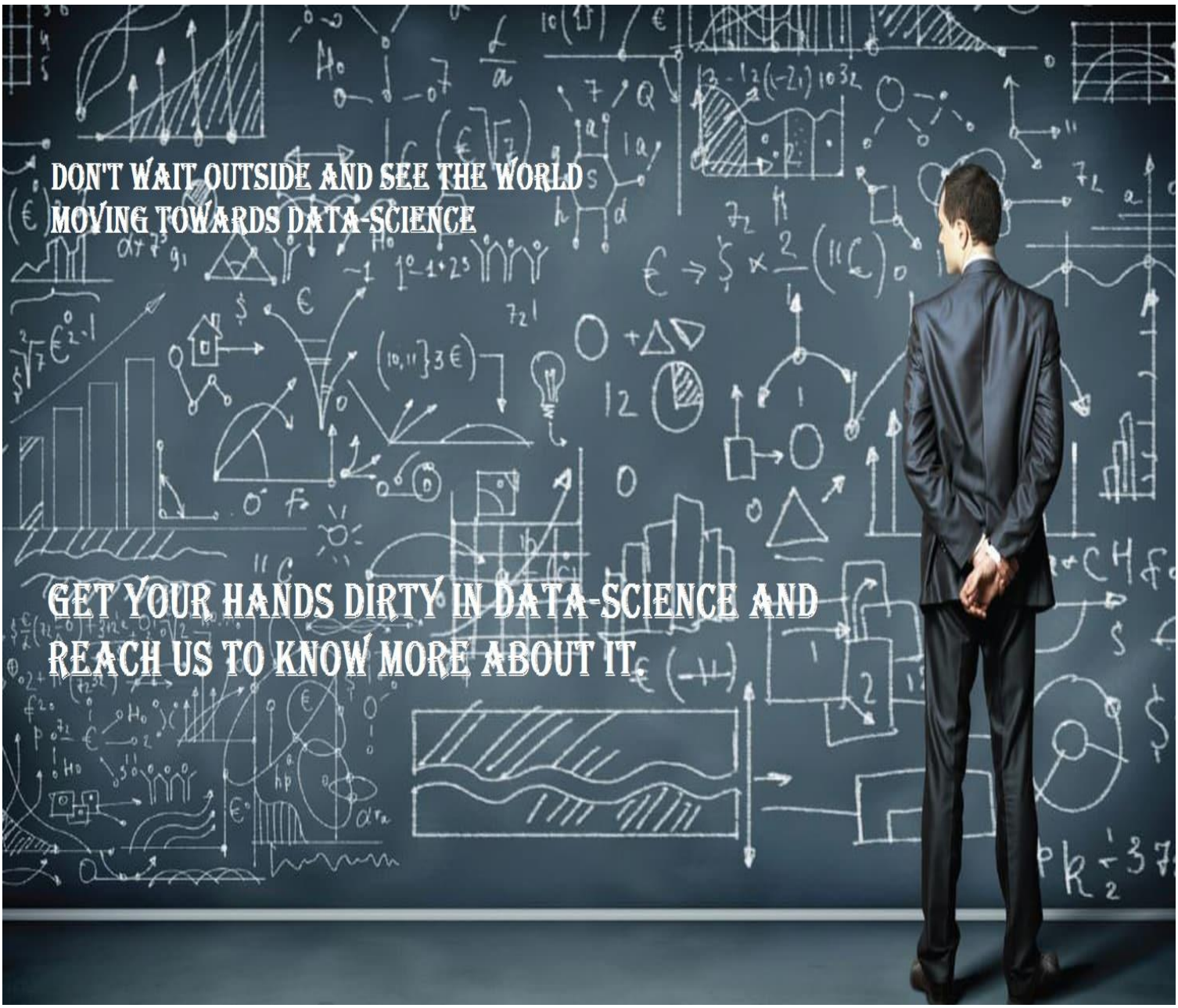

Diploma Program in

'Data Science'

Duration: 32 weeks (160 hours)

Sr. No.	Name of the Programme	Details About the Programme	Remarks
1.	**Diploma course in 'Data Science'**	**Mission of AIIITS:** The mission of AIIITS is to advance knowledge and educate students in information Technology and other areas that will best serve the nation. **Vision of AIIITS:** To position Pinnacle Infotrain as a premier institute responsive to emerging needs of industry. To produce high skilled graduates and contribute towards sustainable development of the industry and nation.	
		1. Introduction to the Programme: Data is everywhere. In fact, the amount of digital data that exists is growing at a rapid rate, doubling every two years, and changing the way we live. According to IBM, 2.5 billion gigabytes (GB) of data was generated every day in 2012. Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, uncertainty quantification, computational science, data mining, databases, and visualization The course aims at utilizing fully the capabilities of the free and open source software. The participants will get hands-on training. **2. Objectives of the Programme:** The course aims at imparting relevant programming abilities, develop capabilities for statistical analysis of data, develop ability to build and assess data-based models, develop machine learning algorithms, learn to visualize data which will ultimately help to take strategic decision and also for forecasting. Also Data Science skill is one of the most demanding skills looked for in industry. According to Forbes" Data Scientist Is the Best Job in America According	

		Glassdoor's 2018 Rankings" and the same demand is being seen in Indian industries **<u>Objectives:</u> the objectives of the proposed course are to impart knowledge on the following:** 1. Developing Programming skills of the relevant programming languages for data analysis and data visualization. 2. Developing machine learning algorithm. 3. Develop skills to manage unstructured and Big-data. 4. Learn tools which will help to take strategic decisions. 5. Learn skills which can give a very promising career.	
		3. Target Group of Learners: Data Science is an amalgamation of Science and Technology, it uses different tools and technologies to extract information out of a data which help in taking decisions be it in business, healthcare or finance. Applications are invited from those who wants to learn different tools and technologies to analyze data and infer meaningful information out of it. It also helps to build a promising career in every aspects.	
		4. Instructional Design: The course will consist of live lectures and assignments for every modules. After every module doubt-clearing sessions will be arranged where students will be free to discuss their doubts. Sessions for Interview preparation will be covered after the completion of all the modules.	
		5. Instructional Design: **i. <u>Duration of the Programme:</u> 8** months **ii. <u>Course delivery</u>** The course will be entirely delivered ***<u>online or offline.</u>*** There are six modules in the course which will be conducted online by expert faculties in the respective areas. Each week's menu will cover the following: 1. **<u>Interactive lectures:</u>** This online sessions will be conducted either on Microsoft team or Google meet or Zoom. The session link will be shared with the students.	

		2. **Lecture(s):** the theoretical and applied parts of the topic will be covered in lectures. 3. **Exercises and data:** Assignments will be allocated to the participants which they need to complete and submit and assessments on any particular module will be done based on the assignments. 4. **Doubt-clearing:** There will be an interactive forum as a platform to interact with each other and with the resource persons**.** Here the participants can discuss their difficulties, can ask questions and get the doubts clarified.	
		6. Eligibility: Students with any bachelor's/Master's degree in Science/ Mathematics/ Statistics/ Information Technology/Computer Science/ Engineering/ from any recognized universities in India or other countries.	
		7. Scheme and Evaluation: There would be three types of assessment for evaluating the performance of the participants - short and long answer questions, multiple type questions and practical exercises. Each participant will be given assignments and projects. After completion of the training, online examination will be conducted and Certificate will be jointly issued by Pinnacle Infotrain and Jain University only after completion of all the assignments, project and after qualifying the exam.	

8. Procedure for admission, Curriculum transaction and evaluation:

Admission will be based on prerequisite degree of any recognized universities in India. The Course is affiliated to the Jain University.

9. Fee structure:

The fees should be paid in one installment only, before the commencement of the course*.

Rs. 58,000/-

*Installment options can be provided with additional 8% on the course fees. The students can pay it in 4 easy installments

10. Syllabus:

Topic 1: Sql-

Introduction to Basic Database Concepts, E-R Modelling and Diagram, Normalization, Introduction to SQL, DDL and DML Statements, Working with Queries (DQL), Aggregate Functions, Joins and Set Operations, Implementation of Data integrity, Working with Constraints, Implementing Views, Data Control language (DCL), Working with Indexes, Writing Transact-SQL (T-SQL), Working with Stored Procedures and Functions, Implementing Triggers

(Duration-2 Weeks)

Topic 2: Python-

Core Python, Python Introduction, Environment, Getting Started , String Handling, Operators, Flow Controllers, Collections, Functions, Modules, Packages, File Handling, Advanced Python, Oops Concepts Regular Expressions, Database Access, Introduction to RDBMS, Installation of MySQL Python Modules, Multi-Threading, Working with csv , xml and Json files, Data Analytics, Introduction to Numpy, Computation on Numpy arrays, Numpy Structured Arrays, Introduction to Pandas, Pandas objects, Operating on data, Handling missing data, Working with time series, MatPlotLib

(Duration- 7 weeks)

Topic 3: R-

R Base Software, R Studio The IDE, Basic Operations, Operators and Types, R Functions, Logistic Regression in R, Reason for Logistic Regression, The Logistic Transform, Logistic Regression Modelling, Model

		Optimisation, Understanding ROC Curve, Default Modelling using Logistic Regression in R, Decision Trees Theory of Entropy & Information Gain, Cross Validations for Overfitting Problem, Ensemble Learning, Bootstrap Aggregation, Random Forests, Intrusion Detection in IT Network, Linear Regression in R, Covariance and Correlation, Multivariate Analysis, Hypothesis Testing Limitations of Regression, Loss Given Default using Linear Regression, Support Vector Machine, Classification as a Hyper Plane Location Problem , Motivation for Linear Support Vectors, Quadratic Optimization, Non Linear SVM, Kernel Functions, Default Modelling using SVM in R, Predictive Modelling, Decision Trees, Neural Networks, Predictive Modelling with Decision Trees, Neural Networks, Back Propagation Revision of Key Concepts, Parameter Estimation, Hypothesis testing, Bayesian Analysis, Identifying the best estimator, Other Statistical Theory, Model fitting Linear Regression, Non-linear Regression, Categorical Data Analysis, Time Series & Longitudinal Analysis Machine Learning, ANOVA/ Regression Analysis, Analysis of Variance & Covariance, Analysis of Variance Examine Regression Results, Regression Analysis, Linear and Logistic Regression, Tree and Bayesian Network Models, Decision Trees, Bagging, Random Forests, Boosted Trees, Bayesian Classification Models **(Duration-6 weeks)** **Topic 4: BigData Hadoop-** Big Data Introduction, Introduction to Hadoop, Hadoop Distributed File System (HDFS) Storage, HDFS Design and concepts, HDFS Architecture, Read and Write Architecture, Cluster setup, Adding New Data Node dynamically, High Availability, Zookeeper leader election algorithm, HDFS commands, MAP Reduce, Basics and Its architecture, Map Reduce Job Run, Legacy Architecture, Shuffling and Sorting, Hands on word count in Map/Reduce, Distributed Cache, Optimization Techniques, Map Side Joins, YARN Concepts, NOSQL, ACID in RDMBS, BASE IN NoSQL, CAP Theorem, Hbase Database in Detail, Hbase operations through shell, HIVE, Hive Introduction and Architecture, Hive Service , Shell , server, Working with Tables and different file formats, Partitions , Bucketing, External Partitioned tables, Order By , DISTRIBUTED By , Sorty by	

		differences, PIG, Execution Types, Grunt Shell, PigLatin, Data Processing , Schema on Read, Primitive Data types, Complex Data types, Data Loading , Storing , Filtering, Grouping & Joining, SPLITS and JOINS, HCATALOG, Introduction to Hcatalog, Hcatalog with PIG , HIVE and MR, SQOOP, Import data, Incremental Import, Export Data, FLUME, Introduction to Flume, Flume Agnets : Sources, Channels and Sinks, Flume Commands, Use cases, OOZIE, Workflow, How to schedule sqoop job, HIVE , MR , PIG **(Duration-9 weeks)** **<u>Topic 5:</u> Machine Learning-** Introduction, What is ML Problems, Data, Tools, Linear regression, SSE, Gradient descent, Closed form, Normal equations, Features, Overfitting and complexity, Training, validation, test data, Classification problems, Decision boundaries, Nearest neighbour methods, Probability and classification, Bayes optimal decisions, Naive Bayes, Gaussian class-conditional distribution, Linear classifiers, Bayes' Rule, Naive Bayes Model, Logistic regression, Online gradient descent, Neural Networks, Decision tree, Ensemble methods, Bagging, Random forests, Boosting, A more detailed discussion on Decision Tree and Boosting Unsupervised learning, Clustering, K-means, Hierarchical agglomeration, Advanced discussion on clustering and EM, Latent space methods, PCA, Text representations, naive Bayes, multinomial models, clustering, latent space models **(Duration-6 weeks)** **<u>Topic 6:</u> Statistics-** Mean, Mode, Median, Standard deviation, Probability theory, Measures of location (or central tendency) and dispersion, Random variables, Expectation of random variable and its properties, standard discrete probability distributions, Standard continuous probability distribution, bivariate and multivariate distributions, correlation and regression, Combination, Data Modeling **(Duration- 2 weeks)**	

		Key Reference Books 1. Python Crash Course by Eric Mathews. 2. Learning Python, 5th Edition by Mark Lutz 3. Data Analystics using Python by Bharti Motwani 4. SQL the complete reference, 3rd Edition by James R Groff and Paul N Weinberg 5. R for Data Science by Garrett Grolemund and Hadley Wickham 6. Learning Machine learning: From theory to Algorithms by Shai Ben-David and Shai Shalev-Shwartz 7. Machine Learning by Anuradha Srinivasaraghavan 8. Hadoop: The definitive Guide by Tom White 9. Big Data Analytics, Introduction to Hadoop, Spark and Machine Learning by Raj Kamal and Preeti Saxena. 10. The Elements of Statistical Learning by Jerome H Friedman, Robert Tibshirani abd Trevor Hastie	
		11. Quality Assurance: IQAC (Internal Quality Assurance Cell) is in place to oversee the Programme delivery mechanism and suggest changes specific to industry requirements. The quality of the programme will be ensured through strict monitoring by an executive committee including the Co-ordinator of the programme, the subject experts, Director. The Co-ordinator of the programme shall ensure the regular student feedback of courses, teachers and programme in the prescribed format towards the end of the semester and the same shall be analyzed to draw conclusions for effecting improvement. Periodical review meetings on the programme efficacy will be held in which the remarks of teachers on curriculum, syllabi and methods of teaching and evaluation will be given due importance. Moreover, the progress and the quality of the programme will be monitored by the Internal Quality Assurance Cell of Pinnacle Infotrain from the outcome and feedback of the learners as well as the proper documentation maintained in the Centre.	
		12. SLM: Self-Learning Material is available in English	