P.G. DIPLOMA IN DATA SCIENCES

Syllabus

Session (2020-2021)

KHALSA COLLEGE
AMRITSAR
-An Autonomous College

Note:

1. Copy rights are reserved. Nobody is allowed to edit it in any form. Defaulters will be prosecuted.

2. Subject to change in the syllabi at any time. Please visit the Khalsa College website time to time.

P.G. DIPLOMA IN DATA SCIENCES

SEMESTER – I

Sr. No.	Paper	Paper Name	Theory	Internal	Practical	Total	Page no.
1	**Paper-I**	Introduction to Python	**75**	**25**	**-**	**100**	**192-193**
2	**Paper-II**	Introduction to AI and Data Science	**75**	**25**	**-**	**100**	**194**
3	**Paper-III**	Big Data Analytics	**75**	**25**	**-**	**100**	**195**
4	**Paper-IV**	Programming Laboratory (Based on Python)	**-**	**13**	**37**	**50**	**196**
5	**Paper-V**	Programming Laboratory(Based on AI and Data Science)	**-**	**13**	**37**	**50**	**197**

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – I

Paper I: Introduction to Python

Time: 3 Hours **Max. Marks: 100**
Theory Marks:75
Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.
Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.
Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks.**
Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction to Python: Process of Computational Problem Solving, Python Programming Language
Data and Expressions: Literals, Variables and Identifiers, Operators, Expressions, Statements and Data Types
Control Structures: Boolean Expressions (Conditions), Logical Operators, Selection Control, Nested conditions, Debugging
Lists: List Structures, Lists (Sequences) in Python, Iterating Over Lists (Sequences) in Python
Functions: Fundamental Concepts, Program Routines, Flow of Execution, Parameters & Arguments
Iteration: While statement, Definite loops using For, Loop Patterns, Recursive Functions, Recursive Problem Solving, Iteration vs. Recursion

UNIT-II

Dictionaries: Dictionaries and Files, Looping and dictionaries, Advanced text parsing
Files: Opening Files, Using Text Files, String Processing, Exception Handling
Objects and Their Use: Introduction to Object Oriented Programming
Modular Design: Modules, Top-Down Design, Python Modules
Using Databases and SQL: Database Concepts, SQLite Manager Firefox Add-on, SQL basic summary, Basic Data modeling, Programming with multiple tables

References:

1. Python for Informatics, Charles Severance, version 0.0.7
2. Introduction to Computer Science Using Python: A Computational Problem-Solving Focus, Charles Dierbach, Wiley Publications, 2012, ISBN : 978-0-470-91204-1
3. Introduction To Computation And Programming Using Python, GUTTAG JOHN V, PHI, 2014, ISBN-13: 978-8120348660
4. Introduction to Computating& Problem Solving Through Python, Jeeva Jose and Sojan P. Lal,Khanna Publishers, 2015, ISBN-13: 978-9382609810
5. Introduction to Computing and Programming in Python, Mark J. Guzdial, Pearson Education, 2015, ISBN-13: 978-9332556591
6. Fundamentals of Python by Kenneth Lambert, Course Technology, Cengage Learning , 2015
7. Learning Python by Mark Lutz, 5th Edition, O'Reilly Media, 2013

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – I

Paper II: Introduction to Artificial Intelligence and Data Science

Time: 3 Hours **Max. Marks: 100**

Theory Marks:75

Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.

Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.

Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks.**

Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction to Artificial Intelligence: Definitions of AI, Intelligent Agents, Problem solving.

Knowledge, Reasoning and Planning: Logical Agents, Classical Planning, Knowledge Representation and Reasoning.

Learning: Learning from examples, Knowledge in learning.

Communicating, Perceiving and Acting:

Communication, Natural Language Processing, Perception, Robotics.

UNIT-II

Introduction to Data Science: Data Science-a discipline, Landscape-Data to Data science, Data Growth-issues and challenges, data science process. foundations of data science.

Data Exploration and Preparation: Messy data, Anomalies and artifacts in datasets. Cleaning data.

Data Representation and Transformation: Forms of data-tabular, text data, graph-based data. Modern databases- text files, spreadsheets, SQL databases, NoSQL databases, distributed databases, live data streams.

Representation of data of special types-acoustic, image, sensor and network data.

Computing with Data: Overview of R, Python and Julia.

Data Modeling: Basics of Generative modeling and Predictive modeling.

Data Visualization and Presentation: Charts-histograms, scatter plots, time series plots etc. Graphs, 3D Visualization and Presentation

References:

1.S.J. Russell and P.Norving: “Artificial Intelligence: A Modern Approach”, Pearson.

2. Sinan Ozdemir, “Principles of Data Science”, Packt Publishing.

3.E.Rich, K.Knight, S.B. Nair: “Artificial Intelligence”, Tata McGraw Hill Ed Pvt Ltd.

4.Joel Grus: “Data Science from Scratch”, O’Reilly.

5.Foster Provost & Tom Fawcett: “Data Science for Business” O’Reilly

6. Roger D. Peng & Elizabeth Matsui: “The Art of Data Science” Lean Publishing.

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – I

Paper III: Big Data Analytics

Time: 3 Hours **Max. Marks: 100**

Theory Marks:75

Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.

Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.

Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks.**

Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction to Big data : Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis – Nature of Data - Analytic Processes and Tools - Analysis vs Reporting.

UNIT-II

Mining data streams : Introduction To Streams Concepts – Stream Data Model and Architecture - Stream Computing - Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating Moments – Counting Oneness in a Window – Decaying Window - Real time Analytics Platform(RTAP) Applications - Case Studies - Real Time Sentiment Analysis- Stock Market Predictions.

Reference Books:

1. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.
2. Tom White “Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 2012.
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Publishing, 2012.
4. Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CUP, 2012.
5. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley& sons, 2012.
6. Glenn J. Myatt, “Making Sense of Data”, John Wiley & Sons, 2007.
7. Pete Warden, “Big Data Glossary”, O’Reilly, 2011.
8. Jiawei Han, Micheline Kamber “Data Mining Concepts and Techniques”, 2 nd Edition, Elsevier, Reprinted 2008.
9. Da Ruan, Guoquing Chen, Etienne E.Kerre, Geert Wets, “Intelligent Data Mining”, Springer, 2007.
10. Paul Zikopoulos, Dirkde Roos, Krishnan Parasuraman, Thomas Deutsch, James Giles , David Corrigan, “Harness the Power of Big Data The IBM Big Data Platform”, Tata McGraw Hill Publications, 2012.
11. Arshdeep Bahga, Vijay Madisetti, “Big Data Science & Analytics: A HandsOn Approach “,VPT, 2016
12. Bart Baesens “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications (WILEY Big Data Series)”, John Wiley & Sons,2014

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – I

Paper-IV
Programming Laboratory
(Based on Python)

Time: 3Hrs **Max. Marks: 50**
Practical Marks: 37
Practical Internal Assessment Marks: 13

Programs based on Python

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – I

Paper-V
Programming Laboratory
(Based on AI and Data Science)

Time: 3Hrs **Max. Marks: 50**
Practical Marks: 37
Practical Internal Assessment Marks: 13

Practical based on AI and Data Science

P.G. DIPLOMA IN DATA SCIENCES

SEMESTER – II

Sr. No.	Paper	Paper Name	Theory	Internal	Practical	Total	Page no.
1	**Paper-I**	Introduction to R	**75**	**25**	**-**	**100**	**199-200**
2	**Paper-II**	Data Preparation and Analysis	**75**	**25**	**-**	**100**	**201**
3	**Paper-III**	Introduction to Hadoop	**75**	**25**	**-**	**100**	**202-203**
4	**Paper-IV**	Programming Laboratory (Based on R Language)	**-**	**13**	**37**	**50**	**204**
5	**Paper-V**	Programming Laboratory(Based On Hadoop)	**-**	**13**	**37**	**50**	**205**

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – II

Paper I: Introduction to R

Time: 3 Hours **Max. Marks: 100**

Theory Marks:75

Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.

Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.

Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks.**

Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**

The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction: Learn to use help() function. Understand data types in R (logical, numeric, etc.) .Convert data types .Create, find, and remove data (vector, matrix, data frame) in R .Read external data into R (.txt, .csv) .Write R data into external files (.txt, .csv) .Understand and manipulate strings (e.g. substr(), scan()).Understand indexing of data in vectors, matrices, and data frames. Graphing techniques to visualize data selection.

Operators: Learn about operators (mathematics, logical, miscellaneous).Learn about basic math functions (e.g. sum()).Use operators and math functions on variables Learn about ifelse() function .Use ifelse() function on vectors and matrices. Use graphs to show the results.

UNIT-II

Loops: Understand how loops work in R. Create your own loop for vectors. Create a series of graphs with loop functions. Learn to use break and next statements in loops. Use loops to create and change data in vectors, matrices, and arrays. Use loops to create data as a list. Learn about double loops. Create your own double loops for matrix. Use operators and functions in single and double loops. Understand if else statement. Use if else statement for data manipulation. Compare if else statement with ifelse() function. Use ifelse() function in loops .Combine loops and if else statement. Represent your results with graphs. Use math functions in loops. Use math functions in if else statement. Show your results with graphs.

Functions: Understand advanced functions such as apply() and by().Use apply() and by()to calculate descriptive statistics. Create graphs for the calculated descriptive statistics. Understand customized functions. Interpret customized functions. Compare customized functions and build-in functions. Understand global parameters for graphing. Understand specific parameters in graph functions. Learn different ways to save your graphs. Learn to combine loops and customized functions. Learn to use customized functions in customized functions. Learn to save your functions and reuse them whenever needed.

Reference Books:

1. A First Course in Statistic Programming with R by Braun & Murdoch
2. A Beginner’s Guide to R by Zuur
3. R in a Nutshell by Adler
4. An introduction to R by Venables & Smith
5. Machine Learning with R by Brettlantz

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – II

Paper II: Data Preparation and Analysis

Time: 3 Hours **Max. Marks: 100**
Theory Marks:75
Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.
Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.
Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks.**
Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction: Source of Data, Process for Making Sense of Data

Describing Data: Observations and Variables, Types of Variables, Central Tendency, Distribution of the Data, Confidence Intervals, Hypothesis Tests

Preparing Data Tables: Cleaning the Data, Removing Observations and Variables, Generating Consistent Scales Across Variables, New Frequency Distribution, Converting Text to Numbers, Converting Continuous Data to Categories, Combining Variables, Generating Groups, Preparing Unstructured Data

Unit-II

Understanding Relationship: Visualizing Relationship Between Variables, Calculating Metrics About Relationships

Identifying and Understanding Groups: Clustering, Association Rules, Leaning Decision Trees from Data

Building Models From Data: Linear Regression, Logistic Regression, k-Nearest Neighbors, Classification and Regression Trees

References:

1. Making sense Of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, by GlennJ.Myatt and Wayne P.Johnson
2. The Visual Display of Quantitative Information, by Edward R.Tufte
3. Visualizing Data: exploring and Explaining Data with the Processing environment,by Ben Fry
4. Exploratory Data Mining and Data Cleaning, by Tamraparni Dasu

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER – II

Paper III: Introduction to Hadoop

Time: 3 Hours **Max. Marks: 100**
Theory Marks:75
Theory Internal Assessment Marks:25

Note: The question paper covering the entire course shall be divided into three sections.
Section A: It will have question No.1 consisting of 10 very short answer questions from the entire syllabus. Students will attempt 6 questions. Each question will carry **2.5 marks** with answer to each question up to 10 lines in length. The total weightage being **15 marks**.
Section B: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 2, 3, 4 and 5 will be set by the examiner from Unit-I of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks.**
Section C: It will consist of essay type/numerical questions up to five pages in length. Four questions numbering 6, 7, 8 and 9 will be set by the examiner from Unit-II of the syllabus. The students will be required to attempt any two questions. Each question will carry **15 marks.**
The total weightage of this section shall be **30 marks**.

UNIT-I

Introduction: History of Hadoop, The Hadoop Distributed File System, Components of Hadoop, Analysing the Data with Hadoop, Scaling Out, Hadoop Streaming, Design of HDFS, Java interfaces to HDFS Basics ,Developing a Map Reduce Application, How Map Reduce Works, Anatomy of a Map, Reduce Job run, Failures, Job Scheduling, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features Hadoop environment.

UNIT-II

Frameworks: Applications on Big Data Using Pig and Hive, Data processing operators in Pig, Hive services, HiveQL, Querying Data in Hive, fundamentals of HBase and ZooKeepe, IBM InfoSphere BigInsights and Streams. Predictive Analytics, Simple linear regression, Multiple linear regression, Interpretation 5 of regression coefficients. Visualizations - Visual data analysis techniques, interaction techniques, Systems and applications.

References:

1. Michael Berthold, David J. Hand, “Intelligent Data Analysis”, Springer, 2007.
2. Tom White “Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 2012.
3. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Publishing, 2012.
4. Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CUP, 2012.
5. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, John Wiley& sons, 2012.
6. Glenn J. Myatt, “Making Sense of Data”, John Wiley & Sons, 2007.
7. Pete Warden, “Big Data Glossary”, O’Reilly, 2011.
8. Jiawei Han, Micheline Kamber “Data Mining Concepts and Techniques”, 2 nd Edition, Elsevier, Reprinted 2008.
9. Da Ruan, Guoquing Chen, Etienne E.Kerre, Geert Wets, “Intelligent Data Mining”, Springer, 2007.

10.Paul Zikopoulos, Dirkde Roos, Krishnan Parasuraman, Thomas Deutsch, James Giles , David Corrigan, “Harness the Power of Big Data The IBM Big Data Platform”, Tata McGraw Hill Publications, 2012.
11. Arshdeep Bahga, Vijay Madisetti, “Big Data Science & Analytics: A HandsOn Approach “,VPT, 2016
12. Bart Baesens “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications (WILEY Big Data Series)”, John Wiley & Sons,2014

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER-II

Paper-IV
Programming Laboratory
(Based on R Language)

Time: 3Hrs

Max. Marks: 50
Practical Marks: 37
Practical Internal Assessment Marks: 13

Programs based on R Language

P.G. DIPLOMA IN DATA SCIENCES
SEMESTER-II

Paper-IV
Programming Laboratory
(Based on Hadoop)

Time: 3Hrs **Max. Marks: 50**

Practical Marks: 37

Practical Internal Assessment Marks: 13

Practical based on Hadoop