Savitribai Phule Pune University
Third Year Computer Engineering (2019 Course)
310251: Data Science and Big Data Analytics

Teaching Scheme:	Credit	Examination Scheme:
TH: 03 Hours/Week	03	In-Sem (Paper): 30 Marks
		End-Sem (Paper): 70 Marks

Prerequisite Courses: Discrete Mathematics (210241), Database Management Systems (310341)

Companion Course: Data Science and Big Data Analytics Laboratory (310256)

Course Objectives:

1. To understand the need of Data Science and Big Data
2. To understand computational statistics in Data Science
3. To study and understand the different technologies used for Big Data processing
4. To understand and apply data modelling strategies
5. To learn Data Analytics using Python programming
6. To be conversant with advances in analytics

Course Outcomes:

After completion of the course, learners should be able to

CO1: Analyze needs and challenges for Data Science Big Data Analytics

CO2: Apply statistics for Big Data Analytics

CO3: Apply the lifecycle of Big Data analytics to real world problems

CO4: Implement Big Data Analytics using Python programming

CO5: Implement data visualization using visualization tools in Python programming

CO6: Design and implement Big Databases using the Hadoop ecosystem

Unit I: Introduction 07 Hours

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data Science Life Cycle, Data: Data Types, Data Collection. Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, and Data Discretization.

Unit II Statistical Inference 07 Hours

Need of statistics in Data Science and Big Data Analytics**,** Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t- test.

.**Unit III Big Data Analytics Life Cycle 07 Hours**

Introduction to Big Data, sources of Big Data, Data Analytic Lifecycle: Introduction, Phase 1: Discovery,

Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operationalize

Unit IV **Predictive Big Data Analytics with Python** **07 Hours**

Introduction, Essential Python Libraries, Basic examples. Data Preprocessing: Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types: Predictive, Descriptive and Prescriptive. Association Rules: Apriori Algorithm, FP growth. Regression: Linear Regression, Logistic Regression. Classification: Naïve Bayes, Decision Trees. Introduction to Scikit-learn, Installations, Dataset, matplotlib, filling missing values, Regression and Classification using Scikit-learn.

Unit V **Big Data Analytics and Model Evaluation** **07 Hours**

Clustering Algorithms: K-Means, Hierarchical Clustering, Time-series analysis. **Introduction to Text Analysis**: Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis.**Model Evaluation and Selection**: Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time- series analysis using Scikit-learn, sklearn.metrics, Confusion matrix, AUC-ROC Curves, Elbow plot.

Unit VI: Data visualization and Hadoop **07 Hours**

Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. **Data Visualization using Python:** Line plot, Scatter plot, Histogram, Density plot, Box- plot.

Text Books:

1. David Dietrich, Barry Hiller, “Data Science and Big Data Analytics”, EMC education services, Wiley publication, 2012, ISBN0-07-120413-X.

2. Jiawei Han, MichelineKamber, and Jian Pie, “Data Mining: Concepts and Techniques” Elsevier Publishers Third Edition, ISBN: 9780123814791, 9780123814807.

Reference Books :

1. EMC Education Services, "Data Science and Big Data Analytics- Discovering, analyzing Visualizing and Presenting Data"
2. DT Editorial Services, "Big Data, Black Book", DT Editorial Services, ISBN: 9789351197577, 2016 Edition.
3. Chirag Shah, "A Hands-On Introduction To Data Science", Cambridge University Press, (2020), ISBN : ISBN 978-1-108-47244-9.
4. Wes McKinney, "Python for Data Analysis" O' Reilly media, ISBN: 978-1-449-31979- 3
5. "Scikit-learn Cookbook", Trent hauk,Packt Publishing, ISBN: 9781787286382
6. Jenny Kim, Benjamin Bengfort, "Data Analytics with Hadoop", OReilly Media, Inc., ISBN: 9781491913703.
7. Venkat Ankam, "Big Data Analytics", Packt Publishing, ISBN: 9781785884696

e-Books:

- An Introduction to Statistical Learning by Gareth James
- https://www.ime.unicamp.br/~dias/Intoduction%20to%20Statistical%20Learning.pdf
- Python Data Science Handbook by Jake VanderPlas
- https://tanthiamhuat.files.wordpress.com/2018/04/pythondatasciencehandbook.pdf
- Introducing Data Science by Davy Ciele, Manning Publications
- Introducing Data Science [PDF]
- Handbook for visualizing : a handbook for data driven design by Andy krik
- A Handbook for Data Driven Design
- An introduction to data Science: https://docs.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit?pli=1
- Hadoop Tutorial: https://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf?utm_source=7_&utm_medium=affiliate&utm_content=5f34cd37cdf1050001b09537&utm_campaign=Admitad&utm_term=761c575424fc4a6b48d02f72157eb578
- Learning with Python; How to think like a computer scientist: http://openbookproject.net/thinkcs/python/english3e/
- Python for everybody:
- http://do1.dr-chuck.com/pythonlearn/EN_us/pythonlearn.pdf
- Scikit Learn Tutorial
- https://scikit-learn.org/stable/

MOOCs Courses links:

- Computer Science and Engineering - NOC:Data Science for Engineers
- Computer Science and Engineering - NOC:Python for Data Science
- Computer Science and Engineering - NOC:Data Mining
- Computer Science and Engineering - NOC:Big Data Computing
- Big Data Computing - Course

Sinhgad Technical Education Society's

RMD SINHGAD SCHOOL OF ENGINEERING, PUNE

Department of Computer Engineering

TEACHING PLAN

Academic Year: 2023-24 (Semester: VI)

<table>
<tr><td colspan="2">Course Title:
Data Science and Big Data Analytics</td><td>Subject Code: 310251</td><td colspan="2">Class: T.E.</td><td>Division: A and B</td></tr>
<tr><td>Term: II</td><td colspan="2">Date of commencement of classes:08/01/2024</td><td colspan="3">Date of conclusion of teaching: 19/04/2024</td></tr>
<tr><td rowspan="2">Lecture Schedule:
3Hrs/ Week</td><td rowspan="2">Practical/Tutorial Schedule:
4 Hrs/Week</td><td colspan="4">Examination Scheme</td></tr>
<tr><td rowspan="2">Theory:100 M
In Sem:30 M (1 Hr)
End Sem: 70 (2Hrs.30 min.)</td><td>Term Work</td><td>Practical</td><td>Oral</td></tr>
<tr><td>50 M</td><td>25 M</td><td>NA</td></tr>
<tr><td rowspan="2">Subject Teacher</td><td rowspan="2">Mrs. Jyoti S. Raghatwan</td><td rowspan="2">Previous 3 Years University Result</td><td>2021-22</td><td>2022-23</td><td>2023-24</td></tr>
<tr><td>97%</td><td>97%</td><td></td></tr>
</table>

UNIT – I: Introduction to Data Science and Big Data

Syllabus:

Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 5 V's of Big Data, Relationship between Data Science and Information Science, Business intelligence versus Data Science, Data Science Life Cycle, Data: Data Types, Data Collection. Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction, Data Transformation, Data Discretization.

PLAN			**ACTUAL**		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
1		Basics and need of Data Science and Big Data, Applications of Data Science			
2		Data explosion, 5 Vs of Big Data			
3		Relationship between Data Science and Information Science			
4		Business intelligence versus Data Science, Data Science Life Cycle			
5		Data: Data Types, Data Collection.			
6		Need of Data wrangling, Methods: Data Cleaning, Data Integration, Data Reduction			
7		Data Transformation, Data Discretization.			
Make up Classes					
Contents Beyond syllabus					

UNIT –II: Statistical Inference					
Syllabus: Need of statistics in Data Science and Big Data Analytics**,** Measures of Central Tendency: Mean, Median, Mode, Mid-range. Measures of Dispersion: Range, Variance, Mean Deviation, Standard Deviation. Bayes theorem, Basics and need of hypothesis and hypothesis testing, Pearson Correlation, Sample Hypothesis testing, Chi-Square Tests, t- test.					
PLAN			**ACTUAL**		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
8		Need of statistics in Data Science and Big Data Analytics			
9		Measures of Central Tendency: Mean, Median, Mode			
10		Measures of Dispersion: Range, Variance, Mean Deviation			
11		Standard Deviation, Bayes theorem			
12		Basics and need of hypothesis and hypothesis testing			
13		Pearson Correlation, Sample Hypothesis testing			
14		Chi-Square Tests, t- test			
Make up Classes					
Contents Beyond syllabus					

UNIT –III: Big Data Analytics Life Cycle					
Syllabus: Introduction to Big Data, sources of Big Data, Data Analytic Lifecycle: Introduction, Phase 1: Discovery, Phase 2: Data Preparation, Phase 3: Model Planning, Phase 4: Model Building, Phase 5: Communication results, Phase 6: Operationalize					
PLAN			**ACTUAL**		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
15		Introduction to Big Data			
16		Sources of Big Data			
17		Data Analytic Lifecycle: Introduction			
18		Phase 1: Discovery, Phase 2: Data Preparation			
19		Phase 3: Model Planning, Phase 4: Model Building			
20		Phase 5: Communication results			
21		Phase 6: Operationalize			
Make up Classes					

Contents Beyond syllabus					

UNIT –IV: Predictive Big Data Analytics with Python					
Syllabus: Introduction, Essential Python Libraries, Basic examples. Data Preprocessing: Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data. Analytics Types: Predictive, Descriptive and Prescriptive. Association Rules: Apriori Algorithm, FP growth. Regression: Linear Regression, Logistic Regression. Classification: Naïve Bayes, Decision Trees. Introduction to Scikit-learn, Installations, Dataset, matplotlib, filling missing values, Regression and Classification using Scikit-learn.					
PLAN			**ACTUAL**		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
22		Introduction, Essential Python Libraries, Basic examples. Data Preprocessing			
23		Removing Duplicates, Transformation of Data using function or mapping			
24		Replacing values, Handling Missing Data, Analytics Types: Predictive, Descriptive and Prescriptive			
25		Association Rules: Apriori Algorithm, FP growth. Regression: Linear Regression, Logistic Regression			
26		Classification: Naïve Bayes, Decision Trees			
27		Introduction to Scikit-learn, Installations, Dataset, matplotlib			
28		Regression and Classification using Scikit-learn			
Make up Classes					
Contents Beyond syllabus					

UNIT –V: **Big Data Analytics and Model Evaluation**
Syllabus: Clustering Algorithms: K-Means, Hierarchical Clustering, Time-series analysis. **Introduction to Text Analysis**: Text-preprocessing, Bag of words, TF-IDF and topics. Need and Introduction to social network analysis, Introduction to business analysis.**Model Evaluation and Selection**: Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling, Parameter Tuning and Optimization, Result Interpretation, Clustering and Time- series analysis using Scikit-learn, sklearn.metrics, Confusion matrix, AUC-ROC Curves, Elbow plot.

PLAN			ACTUAL		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
29		Clustering Algorithms: K-Means, Hierarchical Clustering, Time-series analysis			
30		Introduction to Text Analysis: Text-preprocessing, Bag of words			
31		TF-IDF and topics. Need and Introduction to social network analysis			
32		Introduction to business analysis.Model Evaluation and Selection			
33		Metrics for Evaluating Classifier Performance, Holdout Method and Random Subsampling			
34		Parameter Tuning and Optimization, Result Interpretation, Clustering and Time- series analysis using Scikit-learn			
35		Sklearn.metrics, Confusion matrix, AUC-ROC Curves, Elbow plot			
Make up Classes					
Contents Beyond syllabus					

UNIT –VI: Data Visualization and Hadoop
Syllabus: Introduction to Data Visualization, Challenges to Big data visualization, Types of data visualization, Data Visualization Techniques, Visualizing Big Data, Tools used in Data Visualization, Hadoop ecosystem, Map Reduce, Pig, Hive, Analytical techniques used in Big data visualization. **Data Visualization using Python:** Line plot, Scatter plot, Histogram, Density plot, Box- plot.

PLAN			ACTUAL		
Lect. No.	Date	Topics	Date	Topics covered	Reasons for Deviation
36		Introduction to Data Visualization, Challenges to Big data visualization			
37		Types of data visualization, Data Visualization Techniques			
38		Visualizing Big Data, Tools used in Data Visualization			

39		Hadoop ecosystem, Map Reduce, Pig, Hive			
40		Analytical techniques used in Big data visualization			
41		Data Visualization using Python: Line plot			
42		Scatter plot, Histogram, Density plot, Box- plot			
Make up Classes					
Contents Beyond syllabus					

SUMMARY

No. of lectures allotted by university	42
Total no. of lectures conducted	
Percentage of syllabus covered	
Total no. of makeup classes	

Date **Subject Teacher** **HOD COMP**

UNIT WISE QUESTION BANK

Unit I: Introduction to Data Science and Big Data

CO1: Analyze needs and challenges for Data Science Big Data Analytics

Question No.	Questions	Marks
1.	What are data Types and need of Data wrangling?	7
2.	Differentiate between analysis and analytics? Discuss the importance of big data analytics?	8
3.	Define big data analytics? Identify three areas or domains in which data science is being used and describe how.	9
4.	What is data science? Identify three areas or domains in which data science is being used and describe how.	9
5.	How data science relates to and differs from Information Science?	6
6.	What is Data Explosion?	5
7.	In the context of Hospital Management System, describe five characteristics of Big Data in detail.	8
8.	List and Explain Sources of Big Data.	6
9.	How data science relates to and differs from Business intelligence?	5
10.	Differentiate Structured and Unstructured Data.	6
11.	Discussed the obstacles in unstructured data in detail.	5
12.	List and Explain various stages of Data Science Lifecycle.	6
13.	List out characteristic of Open Data.	5
14.	List the data Storage Format and Explain any two of them.	6
15.	Select and discuss the reason of the appropriate data storage format for personal digital assistants (PDAs), and smart watches, Comment on your answer in detail.	8
16.	Why data wrangling is important for data science?	5
17.	Draw DIKW Pyramid and Explain. Draw Data Analytics Life Cycle and give Briefly Explain its phase.	9
18.	Explain Type of Missing Data with Example.	6
19.	Discuss 4 conventional Methods to deal with missing data.	5
20.	Compare Next Observation Carried Backward and Linear Interpolation techniques to deal with missing data	7

21.	Discuss 3 Steps of Data Integration.	5
22.	Write a short on Data Normalization.	5
23.	Explain dimensionality reduction with relation to forward and backward selection.	6
24.	Write short notes (a) Data Cleaning, (b) Data Integration, (c) Data Reduction, (d) Data Transformation, (e) Data Discretization	9
25.	Compare BI Vs. Data science	6
26.	Explain Data Analytic Life cycle.	6
27.	What is relationship between Data Science and Information Science,	7
28.	Explain Business intelligence in details	4
29.	What is mean by Data Science? Explain in details	6
30.	Explain Data Science Life Cycle.	6

UNIT 2: Statistical Inference

CO2: Apply statistics for Big Data Analytics

Question No.	Questions	Marks
1.	What is Chi-Square Test?	5
2.	What is population and how it is different from a sample?	6
3.	Explain Nominal Data with Example. Also mention the Data Visualization techniques for nominal data.	8
4.	With reference to Absolute zero, Compare and differentiate Interval and Ratio scale data type of attributes.	8
5.	How to decide the type of data attributes? You can use flow chart to support your comments.	6
6.	A pizza outlet overview its weekly sales. They sold 57 cheese pizzas, 63 pasta pizzas, 53 veggies pizzas, 68 cottage cheese pizzas, and 56 max cheese pizzas. Find the mean of all the pizzas sold by them.	9
7.	With reference skewness of data, explain the empirical relation between mean, mode and median.	6
8.	With the suitable example, Comment on the statement "The range is influence by outliers"	6
9.	Herearethe19scoreslistedout. 5, 7, 10, 15,19,21, 21,22, 22,23, 23, 23, 23,23, 24,24,24, 24, 25 Calculate 1.5*IQR for below the first quartile and above the third quartile. How many data points can low outliers or above outliers?	9
10.	For the given numbers find out variation and standard deviation. Also discuss that how variation and standard deviation is related to each other? 4, 34, 11,12, 2,and 26	8
11.	What is the need of statistics in Data Science and Big Data Analytics?	6
12.	Explain the Measures of Central Tendency.	5
13.	What are the Measures of Dispersion?	5
14.	Explain the following terms 1) Range, Variance, 2) Mean Deviation, 3) Standard Deviation	9
15.	State and explain Bayes theorem.	6
16.	What is need of hypothesis and hypothesis testing?	5
17.	What is Pearson Correlation?	5
18.	A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is a four. Apply Bayes Theorem and Find the probability that the number obtained is actually a four.	6

19.	Define Join Probability and Conditional Probability.	5
20.	Given that the card randomly draws from a pack of 52 cards is as Face card, what is the probability that it is a Queen?	7
21.	Define hypothesis. What is hypothesis testing?	5
22.	What is the difference between Null and Alternative hypothesis. Give one example of each.	6
23.	Distinguish one tail or two tail hypotheses; draw the diagram to support your answer.	7
24.	Define Type I and Type II Error. Give example to differentiate between the two types of error.	8
25.	Explain the significance of critical values line when deciding Type-I and Type-II Errors.	6
26.	What is Chi-square test? Explain its significance in data analytics.	6
27.	What is T-test? What are the types of tests? Explain by the number of variables, degree of free demand means of examples.	8
28.	What is P-Value? Explain the significance of P Value in hypothesis testing.	6
29.	Grapefarm owner wants to compare the two farms to see if there is any weight difference in the bunch of grapes. From farm A, randomly collected 15 bunch with an average weight of 86 gms, and the standard deviation is 7. From farm B, collected10 bunches with an average weight of 80 gms and standard deviation of 8. With a95% confidence level (α=0.05), is there any difference in the farms? 1. Define Null and Alternative Hypothesis 2. Calculate T Statistic using appropriate T Test method. 3. Find out critical value (t score table) and degree of Freedom. 4. Interpret the results base don't Statistic and critical value.	9
30.	Memory capacity of 9 students was tested before and after training. State at 5 percent level of significance whether the training was effective from the following scores: Mean of Differences is (-0.778) Standard deviation of differences is(-1.361) 1. Define Null and Alternative Hypothesis 2. Calculate T Statistic using appropriate T Test method. 3. Find out critical value (t score table) and degree of Freedom. 4. Interpret the results based on T Statistic and critical value. Hint: To state the hypothesis, compare the performance before and after training. To prove the effective training The mean before training is less then after training.	9

Unit: III Big Data Analytics Life Cycle

CO3: Apply the lifecycle of Big Data analytics to real world problems

Question No.	Questions	Marks
1	What is driving data deluge? Explain with one example	6
2	What is data science? Differentiate between Business Intelligence and Data Science.	6
3	What are the sources of Big Data? Explain model building phase with example.	7
4	Explain big data analytics architecture with diagram. What is data discovery phase? Explain with example.	8
5	What is Model Building elaborate this phase of data analytics with the help of suitable example.	7
6	Explain any three sources of Big Data. Differentiate BI versus Data Science.	8
7	What are the three characteristic of Big Data and what are the main Consideration in processing Big Data.	8
8	Explain Descriptive, Diagnostic, Predictive analytics.	9
9	Discuss the following in detail a. Conventional challenges in big data b. Nature of Data	6
10	Describe any five characteristics of Big Data.	6
11	Define the different inferences in big data analytics.	5
12	Define big data. Why is big data required? How does traditional BI environment differ from big data environment?	8
13	What are the challenges with big data?	5
14	What are the three characteristics of big data? Explain the differences between Bl and Data Science.	6
15	Describe the current analytical architecture for data scientists.	6
16	Describe the Challenges of Big Data.	5
17	What is big data analytics? Explain in detail with its example.	6
18	Describe the Challenges of Big Data.	5

UNIT IV: Predictive Big Data Analytics with Python

CO4: Implement Big Data Analytics using Python programming

Question No.	Questions	Marks
1	Explain various data pre-processing steps. Discuss essential python libraries for preprocessing.	6
2	What are association rules? Explain Apriori Algorithm in brief	6
3	Explain the following i) Linear Regression ii) Logistic Regression	6
4	Explain scikit-learn library for matplotlib with example.	7
5	Explain why decision tree are used. Draw a sample decision tree and explain its parts.	8
6	How Apriori Algorithm works, explain with suitable example?	8
7	What is data preprocessing? Explain in details about handling missing data and transformation of data.	8
8	Explain Naïve Bayes' classifier and it applications.	6
9	Discuss the Looping Statements with an example. (i) while (ii) for (iii) range	6
10	Write a Python function to sum of the numbers in a list	6
11	Write the features of Python. Give the advantages & disadvantages of it.	8
12	What is the difference between a module and a package?	6
13	Explain in detail about python operators?	6
14	Write python program to illustrate variable length keyword arguments?	7
15	Write python program to perform linear search?	5
16	What is linear regression?	6
17	Define FP growth. Explain in detail.	8
18	Explain apriori algorithm.	6
19	Explain logistic regression.	6
20	Describe Naive Bayes Algorithm.	6
21	What are the types of naïve byes model?	6
22	Explain decision tree.	6
23	Explain scikit-learn libraries.	6
24	Write short note on matplotlib.	5
25	Write difference between regression and classification.	8

UNIT V: Big Data Analytics and Model Evaluation

CO5: Implement data visualization using visualization tools in Python programming

Question No.	Questions	Marks
1.	What is text processing? Explain TF-IDF with example	9
2.	With suitable example, explain the steps involved in k-means algorithm.	8
3.	Define following terms with respect to confusion matrix : [8] i) Accuracy ii) Precision iii) Recall iv) AUC-ROC	9
4.	Explain k-fold Cross Validation & Random Sub sampling.	6
5.	Write short note on i) Time series Analysis ii) TF - IDF.	8
6.	What is clustering? With suitable example explain the steps involved in k - means algorithm	8
7.	Write short note on i) Confusion matrix ii) AVC - ROC curve	8
8.	Discuss Holdout method and Random Sub Sampling methods	6
9.	What is K-means algorithm	6
10.	What is hierarchical clustering	5
11.	Write the difference between hierarchical clustering and K-means algorithm.	7
12.	Explain time series analysis.	5
13.	Define bag of words.	5
14.	What is TF-IDF?	6
15.	What is social network analysis and Business analysis?	8
16.	What are the types of model selection?	6
17.	How to evaluate models?	5
18.	Define holdout method.	2
19.	What is random sub-sampling method?	5
20.	Write short notes on confusion matrix.	5
21.	Define AUC-ROC curve.	5
22.	Describe Elbow-plot.	4

UNIT VI: Data visualization and Hadoop

CO6: Design and implement Big Databases using the Hadoop ecosystem

Question No.	Questions	Marks
1	With a suitable example, draw a Histogram, boxplot and explain its usages.	8
2	Describe the data visualization tool Tableau. List of data visualization tools.	8
3	What is Data Visualization? Describe the challenges of data visualization.	8
4	Explain architecture of Apache-Pig.	6
5	Write short note on i) Confusion matrix ii) AVC - ROC curve	8
6	Discuss Holdout method and Random Sub Sampling methods.	6
7	With a suitable example explain Histogram and explain its usages.	7
8	Describe the Data visualization tool “Tableau”. Explain its applications in brief.	8
9	With a suitable example explain and draw a Box plot and explain its usages.	6
10	Describe the challenges of data visualization. Draw box plot and explain its usages.	6
11	Introduce data visualization.	5
12	What are the challenges in data visualization?	5
13	What are the types of data visualization?	5
14	How to visualize big data?	5
15	What kind of tools is used in data visualization?	5
16	Explain Hadoop ecosystem.	6
17	Define map reduce.	2
18	Describe Pig?	5
19	Write difference between pig and Map-reduce	5
20	Explain Hive.	5
21	What are analytical techniques used in big data visualization?	6
22	What is the line plot?	4
23	Write about scatter plot	4
24	Define Histogram.	2
25	What is density Plot?	4

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

SET-A

Department of Computer Engineering

A.Y. 2023-24 (Semester-II)

UNIT TEST I EXAM

Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 1Hr **Maximum Marks: 30**

Instructions to Candidates:

1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4.
2. Neat diagrams must be drawn wherever necessary.
3. Assume suitable data, if necessary.

Q. No.		Questions	Marks	CO
1	a	Draw DIKW Pyramid and Explain. Draw Data Analytics Life Cycle and give Briefly Explain its phase.	5	CO1
	b	Select and discuss the reason of the appropriate data storage format for personal digital assistants (PDAs), and smart watches, Comment on your answer in detail.	5	CO1
	c	Define big data analytics? Identify three areas or domains in which data science is being used and describe how.	4	CO1
		OR		
2	a	Write short notes i. Data Cleaning, ii. Data Integration	4	CO1
	b	Why data wrangling is important for data science?	5	CO1
	c	Write a short on Data Normalization	5	CO1
3	a	Explain the following terms 1) Range, Variance, 2) Mean Deviation, 3) Standard Deviation	6	CO2
	b	With reference skewness of data, explain the empirical relation between mean, mode and median.	5	CO2
	c	Here are the 19 scores listed out. 5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25. Calculate 1.5*IQR for below the first quartile and above the third quartile. How many data points can low outliers or above outliers?	5	CO2
		OR		
4	a	Explain Nominal Data with Example. Also mention the Data Visualization techniques for nominal data.	6	CO2
	b	What is T-test? What are the types of Test? Explain by the number of variable, degree of freedom and means of examples.	5	CO2
	c	Define hypothesis. What is hypothesis testing?	5	CO2

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

Department of Computer Engineering
A.Y. 2023-24 (Semester-II)

SET- B

UNIT TEST I EXAM
Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 1Hr **Maximum Marks: 30**

Instructions to Candidates:
1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4,
2. Neat diagrams must be drawn wherever necessary
3. Assume suitable data, if necessary

Q. No		Questions	Marks	CO
1	a	List the data Storage Format and Explain any two of them.	5	CO1
	b	Draw Data Analytics Life Cycle and give Briefly Explain its phase.	5	CO1
	c	Why data wrangling is important for data science?	4	CO1
OR				
2	a	Explain dimensionality reduction with relation to forward and backward selection.	5	CO1
	b	Differentiate Structured and Unstructured Data. What is Data Explosion?	4	CO1
	c	Explain 5 Vs of Big Data	5	CO1
3	a	Define Join Probability and Conditional Probability. Given that the card randomly draws from a pack of 52 cards is as Face card, what is the probability that it is a Queen?	5	CO2
	b	What is Chi-square test? Explain its significance in data analytics.	5	CO2
	c	Define Type I and Type II Error. Give example to differentiate between the two types of error.	6	CO2
OR				
4	a	Explain the Measures of Central Tendency and Measures of Dispersion.	5	CO2
	b	What is P -Value? Explain the significance of P Value in hypothesis testing.	5	CO2
	c	What is the difference between Null and Alternative hypothesis. Give one example of each.	6	CO2

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

SET- A

Department of Computer Engineering

A.Y. 2023-24 (Semester-II)

UNIT TEST II EXAM

Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 1Hr **Maximum Marks: 30**

Instructions to Candidates:

1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4.
2. Neat diagrams must be drawn wherever necessary.
3. Assume suitable data, if necessary.

<table>
<tr><th>Q. No</th><th></th><th>Questions</th><th>Marks</th><th>CO</th></tr>
<tr><td rowspan="3">1</td><td>a</td><td>What are the three characteristic of Big Data and what are the main consideration in processing Big Data.</td><td>4</td><td>CO3</td></tr>
<tr><td>b</td><td>Explain any three sources of Big Data. Differentiate BI versus Data science.</td><td>5</td><td>CO3</td></tr>
<tr><td>c</td><td>Explain in detail how the model building phase is built by team in data analytics life cycle?</td><td>5</td><td>CO3</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td rowspan="3">2</td><td>a</td><td>Write short note on the following:
1) ETL
2) Common tools for the model building.
3) Model selection for data analytics</td><td>5</td><td>CO3</td></tr>
<tr><td>b</td><td>Explain Descriptive, Diagnostic, Predictive analytics.</td><td>5</td><td>CO3</td></tr>
<tr><td>c</td><td>List and explain the steps in data preparation phase of data analytics life cycle.</td><td>4</td><td>CO3</td></tr>
<tr><td rowspan="3">3</td><td>a</td><td>Calculate the support and confidence value for all the possible item sets.
<table>
<tr><td>Transaction ID</td><td>Items bought</td></tr>
<tr><td>1</td><td>Onion, Potato, Cold drink</td></tr>
<tr><td>2</td><td>Onion, Burger, Cold drink</td></tr>
<tr><td>3</td><td>Eggs, Onion, Cold drink</td></tr>
<tr><td>4</td><td>Potato, Milk, Eggs</td></tr>
<tr><td>5</td><td>Potato, Burger, cold drink, Milk eggs</td></tr>
</table></td><td>6</td><td>CO4</td></tr>
<tr><td>b</td><td>What are the types of analytics in big data? Explain in brief.</td><td>5</td><td>CO4</td></tr>
<tr><td>c</td><td>How Apriori Algorithm works, explain with suitable example?</td><td>5</td><td>CO4</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td rowspan="3">4</td><td>a</td><td>Explain Naïve Bayes' classifier and it applications.</td><td>5</td><td>CO4</td></tr>
<tr><td>b</td><td>Write short note on the following: i) Removing duplicates from data set. ii) Handling missing data iii) Data transformation.</td><td>6</td><td>CO4</td></tr>
<tr><td>c</td><td>Explain why decision tree are used. Draw a sample decision tree and explain its parts.</td><td>5</td><td>CO4</td></tr>
</table>

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

SET -B

Department of Computer Engineering

A.Y. 2023-24 (Semester-II)

UNIT TEST II EXAM

Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 1Hr **Maximum Marks: 30**

Instructions to Candidates:

1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4.
2. Neat diagrams must be drawn wherever necessary.
3. Assume suitable data, if necessary.

Q. No.		Questions	Marks	CO
1	a	What is Model Building? Elaborate this phase of data analytics with the help of suitable example	5	CO3
	b	What are the three characteristic of Big Data and what are the main consideration in processing Big Data.	4	CO3
	c	Explain Descriptive, Diagnostic, Predictive analytics.	5	CO3
OR				
2	a	i) Draw the diagram of data analytics life cycle in big data and briefly explain its phases	5	CO3
	b	What are the three characteristics of big data? Explain the differences between BI and Data Science.	5	CO3
	c	Discuss the following in detail a. Conventional challenges in big data b. Nature of Data	4	CO3
3	a	What is data preprocessing? Explain in details about handling missing data and transformation of data.	5	CO4
	b	Explain why decision tree are used. Draw a sample decision tree and explain its parts.	6	CO4
	c	Discuss the Looping Statements with an example. (i) while (ii) for (iii) range	5	CO4
OR				
4	a	Explain Naïve Bayes' classifier and it applications.	6	CO4
	b	Write the features of Python. Give the advantages & disadvantages of it.	5	CO4
	c	Define FP growth. Explain in detail	5	CO4

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

Department of Computer Engineering
A.Y. 2023-24 (Semester-II)

SET- A

PRELIM EXAMINATION
Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 2(1/2) Hr **Maximum Marks: 70**

Instructions to Candidates:
1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4, Q.5 OR Q.6, Q.7 OR Q.8
2. Neat diagrams must be drawn wherever necessary
3. Assume suitable data, if necessary

Q. No.	Question	Marks	CO
1	a. What is driving data deluge? Explain with one example	6	CO3
	b. What is Model Building elaborate this phase of data analytics with the help of suitable example.	5	CO3
	c. Draw the diagram of data analytics life cycle in big data and briefly explain its phases.	6	CO3
	OR		
2	a. What are the three characteristics of big data? Explain the differences between BI and Data Science.	5	CO3
	b. Explain Descriptive, Diagnostic, Predictive analytics.	6	CO3
	c. Write short note on the following: i) ETL ii) Common tools for the model building. iii) Model selection for data analytics.	6	CO3
3	a. Explain Naïve Bayes' classifier and it applications.	6	CO4
	b. What are the types of analytics in big data? Explain in brief.	6	CO4
	c. What is data preprocessing? Explain in details about handling missing data	6	CO4
	OR		
4	a. List and explain the steps in data preparation phase of data analytics life cycle.	6	CO4
	b. Write short note on: i. Apriori algorithm ii. FP growth	6	CO4
	a. Define the following i) Linear Regression ii) Logistic Regression.	6	CO4

Q. No.	Question	Marks	CO
5	a. Explain the following text analysis steps with suitable example. i) Part of speech (POS) tagging ii) Lemmatization iii) Stemming	6	CO5
	b. Explain the TF/IDF (term frequency-inverse document frequency) terms in text analysis with suitable example.	6	CO5
	c. With suitable example, explain the steps involved in k-means algorithm.	5	CO5
	OR		
6	a. Define following terms with respect to confusion matrix : iii) Recall iv) AUC-ROC	5	CO5
	b. What is clustering? With suitable example explain the steps involved in k – means	6	CO5
	c. Explain Confusion matrix in detail.	6	CO5
7	a. Describe the data visualization tool Tableau. List of data visualization tools.	6	CO6
	b. Explain in detail the Hadoop Ecosystem with suitable diagram	6	CO6
	c. Write short on i. Map reduce ii. Pig iii. Hive	6	CO6
	OR		
8	a. With a suitable example, explain and draw a Box plot and explain its usages.	6	CO6
	b. With a suitable example explain Histogram and explain its usages.	6	CO6
	c. Explain architecture of Apache-Pig.	6	CO6

Sinhgad Technical Education Society's
RMD Sinhgad School of Engineering, Warje, Pune-58

SET -B

Department of Computer Engineering
A.Y. 2023-24 (Semester-II)

PRELIM EXAMINATION
Class TE

Subject: Data Science and Big Data Analytics (310251) **Date:**

Time: 2(1/2) Hr **Maximum Marks: 70**

Instructions to Candidates:

1. Attempt Questions Q.1 OR Q.2, Q.3 OR Q.4, Q.5 OR Q.6, Q.7 OR Q.8
2. Neat diagrams must be drawn wherever necessary
3. Assume suitable data, if necessary

Q No.	Question	Marks	CO
1	a. What are the three characteristics of big data? Explain the differences between BI and Data Science.	5	CO3
	b. Discuss the following in detail i. Communication Results ii. Operationalize	6	CO3
	c. Explain in detail sources of big data.	6	CO3
	OR		
2	a. Explain Data Science Life Cycle.	5	CO3
	b. What is relationship between Data Science and Information Science?	6	CO3
	c. Write short notes i. Data Cleaning ii. Data Integration iii. Data Reduction	6	CO3
3	a. Explain Linear Regression.	6	CO4
	a. How Apriori Algorithm works, explain with suitable example?	6	CO4
	b. Explain about Removing Duplicates, Transformation of Data using function or mapping, replacing values, Handling Missing Data.	6	CO4
	OR		
4	a. Write short note on the following: i) Removing duplicates from data set. ii) Handling missing data iii) Data transformation.	6	CO4
	b. Differentiate between data analytics types.	6	CO4

Q No.	Question	Marks	CO
	c. Explain Linear Regression with diagram	6	CO4
5	a. Describe: i. K-means clustering ii. Hierarchical Clustering	6	CO5
	b. Explain k-fold Cross Validation & Random Subsampling.	5	CO5
	c. Explain Bag of words and TF-IDF.	6	CO5
	OR		
6	a. What is time series analysis? Explain in detail.	5	CO5
	b. Write short note on: i. Holdout Method iii. Random Subsampling	6	CO5
	c. Explain Elbow method in detail.	6	CO5
7	a. Explain the types of data visualization	6	CO6
	b. Explain Data Visualization Techniques	6	CO6
	c. What are Analytical techniques used in Big data visualization	6	CO6
	OR		
8	a. Explain Data visualization and Challenges to Big data visualization.	6	CO6
	b. Describe: i. line plot ii. Density plot iii.Histogram	6	CO6
	c. What is Map reduce, pig and Hive?	6	CO6