

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination Data science and its applications (21AD62)

TIME: 03 Hours

Max. Marks: 100

- Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.
03.
04.

Module -1			*Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is data Visualization? Explain bar chart and line chart	L3	1	8
	b	Write a note probability theory as applicable to data science.	L3	1	8
	c	Write a note on normal distribution	L2	1	4
OR					
Q.02	a	Explain the following i) vector addition ii) vector sum iii) vector mean iv) vector multiplication	L3	1	8
	b	Explain the following statistical techniques i) mean ii) median iii) mode iv) interquartile range	L3	1	8
	c	Explain Simpson's Paradox	L2	1	4
Module-2					
Q. 03	a	Explain gradient descent approach in detail with relevant example	L3	2	7
	b	Explain in detail on hypothesis testing with example	L3	2	7
	c	How to get data using stdin and stdout ?	L2	2	6
OR					
Q.04	a	Explain the methodologies to extract the data from web scrapping	L3	2	7
	b	Explain data cleaning, data munging and manipulating Data	L2	2	7
	c	Explain dimensionality reduction in detail	L3	2	6
Module-3					
Q. 05	a	Write a program to train a regularized logistic regression classifier on the iris dataset using sklearn. Train the model with the following hyper parameter $C = 1e4$ and report the best classification accuracy.	L3	3	8
	b	What is machine learning? Explain underfitting and overfitting in detail.	L2	3	6
	c	Explain Naive Bayes as really Dumb Spam Filter.	L3	3	6

OR					
Q. 06	a	Write a program to train an SVM classifier on the iris dataset using sklearn. Try different kernels and the associated hyper parameters. Train model with the following set of hyper parameters RBF kernel, gamma=0.5, one-vs-rest classifier, no-feature-normalization. Also try C=0.01,1,10C=0.01,1,10. For the above set of hyper parameters, find the best classification accuracy along with total number of support vectors on the test data	L3	3	8
	b	Explain regression model in detail for predicting the numerical values.	L3	3	6
	c	How support vector machine is used to classify the data explain.	L3	3	6
Module-4					
Q. 07	a	Illustrate the working of decision tree and hence explain importance of entropy in decision tree.	L3	4	7
	b	What is feed forward neural network? Explain the backpropagation method to train neural networks.	L3	4	7
	c	How deep learning is different from machine learning?	L2	4	6
OR					
Q. 08	a	Illustrate the working of Artificial neural network.	L3	4	7
	b	What is clustering and explain K-means clustering in detail.	L3	4	7
	c	Consider the dataset spiral.txt. The first two columns in the dataset corresponds to the co-ordinates of each data point. The third column corresponds to the actual cluster label. Compute the rand index for the following methods: • K – means Clustering • Single – link Hierarchical Clustering • Complete link hierarchical clustering. • Also visualize the dataset and which algorithm will be able to recover the true clusters.	L3	4	6
Module-5					
Q. 09	a	Explain Gibbs Sampling and Topic Modeling	L2	5	7
	b	Write a note on Recurrent Neural Networks	L3	5	7
	c	Explain Word Clouds and n-Gram Language Models	L2	5	6
OR					
Q. 10	a	Write a note on betweenness centrality and eigenvector centrality	L2	5	7
	b	Write a note on recommender systems	L2	5	7
	c	Explain item-based collaborative filtering and matrix factorization.	L2	5	6

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-I with effect from 2022

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E Degree Examination Data Science and its Applications (21AD62)

TIME: 03 Hours

Max. Marks: 100

Note: Answer any **FIVE** full questions, choosing at least **ONE** question from each module.

<i>Module -1</i>			<i>Marks</i>
Q.01	a	Explain Data Visualization and recognize its use. Sketch Python code segment to visualize line chart and scatterplot with example	06
	b	Summarize dispersion. Using Python code snippet explain the various measures of dispersion	07
	c	Briefly summarize the difference between variance and covariance. Write Python code for finding covariance	07
OR			
Q.02	a	Describe vectors in Data Science and explain any three operations on vectors with Python routine for each operation.	06
	b	Describe Normal Distribution with a Python routine for PDF and CDF	07
	c	Explain Simpson’s paradox with an example	07
Module-2			
Q. 03	a	Interpret the importance of <i>power</i> and <i>significance</i> in Statistical Hypothesis Testing with suitable Python routine.	10
	b	Sketch the use of csv.reader, csv.DictReader and csv.writer in processing Delimited Files	10
OR			
Q.04	a	Articulate the role of BeautifulSoup in Web scraping using Python snippet	10
	b	Summarize Stochastic and Minbatch Gradient Descent.	10
Module-3			
Q. 05	a	Explain K-Nearest Neighbors Algorithm using Iris dataset	10
	b	Summarize the concept of overfitting and underfitting with example and explain how it can be resolved?	10
OR			
Q. 06	a	Explain Naïve Bayes Algorithm in the context of classification with functions	10
	b	Explain the various parameters used in checking the correctness of prediction of Machine Learning Model	10
Module-4			
Q. 07	a	Describe Decision Tree. Interpret with Python routine the process of creating Decision Tree.	10

	b	Illustrate the importance of perceptron in Artificial Neural Network	10
OR			
Q. 08	a	Describe the basic idea behind clustering algorithm using color quantization as example	10
	b	Explain Neural Network as a sequence of Layers with functions	10
Module-5			
Q. 09	a	One of the recent advances in NLP is representing words as low-dimensional vectors .How word vectors helps to accomplish this task?	10
	b	Paraphrase bigram, trigram and ngram language models used in Natural Language Processing	10
OR			
Q. 10	a	Summarize Topic Modeling with reference to topic -word distribution and document-topic distribution	10
	b	Explain Word cloud approach in data visualization using Python code snippet	10

<i>Table showing the Bloom's Taxonomy Level, Course Outcome and Program Outcome</i>				
<i>Question</i>		<i>Bloom's Taxonomy Level attached</i>	<i>Course Outcome</i>	<i>Program Outcome</i>
Q.1	(a)	L3	CO 01	PO03
	(b)	L2	CO 01	PO 02
	(c)	L2	CO 01	PO 02
Q.2	(a)	L3	CO 01	PO 03
	(b)	L2	CO 01	PO 02
	(c)	L1	CO 01	PO 01
Q.3	(a)	L3	CO 02	PO 03
	(b)	L2	CO 02	PO 02
Q.4	(a)	L2	CO 02	PO 02
	(b)	L1	CO 02	PO 01
Q.5	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.6	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.7	(a)	L3	CO 04	PO 03
	(b)	L1	CO 04	PO 01
Q.8	(a)	L2	CO 04	PO 02
	(b)	L2	CO 04	PO 02
Q.9	(a)	L3	CO 05	PO 03
	(b)	L1	CO 05	PO 01
Q.10	(a)	L2	CO 05	PO 02
	(b)	L3	CO 05	PO 03

<i>Bloom's Taxonomy Levels</i>	<i>Lower order thinking skills</i>		
	<i>Remembering (knowledge): L₁</i>	<i>Understanding (Comprehension): L₂</i>	<i>Applying (Application): L₃</i>
	<i>Higher order thinking skills</i>		
	<i>Analyzing (Analysis): L₄</i>	<i>Valuating (Evaluation): L₅</i>	<i>Creating (Synthesis): L₆</i>

Model Question Paper-I with effect from 2022

USN									
-----	--	--	--	--	--	--	--	--	--

Sixth Semester B.E Degree Examination Data Science and its Applications (21AD62)

TIME: 03 Hours

Max. Marks: 100

Note: Answer any **FIVE** full questions, choosing at least **ONE** question from each module.

<i>Module -1</i>			<i>Marks</i>
Q.01	a	Explain Data Visualization and recognize its use. Sketch Python code segment to visualize line chart and scatterplot with example	06
	b	Summarize dispersion. Using Python code snippet explain the various measures of dispersion	07
	c	Briefly summarize the difference between variance and covariance. Write Python code for finding covariance	07
OR			
Q.02	a	Describe vectors in Data Science and explain any three operations on vectors with Python routine for each operation.	06
	b	Describe Normal Distribution with a Python routine for PDF and CDF	07
	c	Explain Simpson's paradox with an example	07
Module-2			
Q. 03	a	Interpret the importance of <i>power</i> and <i>significance</i> in Statistical Hypothesis Testing with suitable Python routine.	10
	b	Sketch the use of csv.reader, csv.DictReader and csv.writer in processing Delimited Files	10
OR			
Q.04	a	Articulate the role of BeautifulSoup in Web scraping using Python snippet	10
	b	Summarize Stochastic and Minbatch Gradient Descent.	10
Module-3			
Q. 05	a	Explain K-Nearest Neighbors Algorithm using Iris dataset	10
	b	Summarize the concept of overfitting and underfitting with example and explain how it can be resolved?	10
OR			
Q. 06	a	Explain Naïve Bayes Algorithm in the context of classification with functions	10
	b	Explain the various parameters used in checking the correctness of prediction of Machine Learning Model	10
Module-4			
Q. 07	a	Describe Decision Tree. Interpret with Python routine the process of creating Decision Tree.	10

	b	Illustrate the importance of perceptron in Artificial Neural Network	10
OR			
Q. 08	a	Describe the basic idea behind clustering algorithm using color quantization as example	10
	b	Explain Neural Network as a sequence of Layers with functions	10
Module-5			
Q. 09	a	One of the recent advances in NLP is representing words as low-dimensional vectors .How word vectors helps to accomplish this task?	10
	b	Paraphrase bigram, trigram and ngram language models used in Natural Language Processing	10
OR			
Q. 10	a	Summarize Topic Modeling with reference to topic -word distribution and document-topic distribution	10
	b	Explain Word cloud approach in data visualization using Python code snippet	10

<i>Table showing the Bloom's Taxonomy Level, Course Outcome and Program Outcome</i>				
<i>Question</i>		<i>Bloom's Taxonomy Level attached</i>	<i>Course Outcome</i>	<i>Program Outcome</i>
Q.1	(a)	L3	CO 01	PO03
	(b)	L2	CO 01	PO 02
	(c)	L2	CO 01	PO 02
Q.2	(a)	L3	CO 01	PO 03
	(b)	L2	CO 01	PO 02
	(c)	L1	CO 01	PO 01
Q.3	(a)	L3	CO 02	PO 03
	(b)	L2	CO 02	PO 02
Q.4	(a)	L2	CO 02	PO 02
	(b)	L1	CO 02	PO 01
Q.5	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.6	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.7	(a)	L3	CO 04	PO 03
	(b)	L1	CO 04	PO 01
Q.8	(a)	L2	CO 04	PO 02
	(b)	L2	CO 04	PO 02
Q.9	(a)	L3	CO 05	PO 03
	(b)	L1	CO 05	PO 01
Q.10	(a)	L2	CO 05	PO 02
	(b)	L3	CO 05	PO 03

<i>Bloom's Taxonomy Levels</i>	<i>Lower order thinking skills</i>		
	<i>Remembering (knowledge): L₁</i>	<i>Understanding (Comprehension): L₂</i>	<i>Applying (Application): L₃</i>
	<i>Higher order thinking skills</i>		
	<i>Analyzing (Analysis): L₄</i>	<i>Valuating (Evaluation): L₅</i>	<i>Creating (Synthesis): L₆</i>

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination Data science and its applications (21AD62)

TIME: 03 Hours

Max. Marks: 100

- Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.
03.
04.

Module -1			*Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is data Visualization? Explain bar chart and line chart	L3	1	8
	b	Write a note probability theory as applicable to data science.	L3	1	8
	c	Write a note on normal distribution	L2	1	4
OR					
Q.02	a	Explain the following i) vector addition ii) vector sum iii) vector mean iv) vector multiplication	L3	1	8
	b	Explain the following statistical techniques i) mean ii) median iii) mode iv) interquartile range	L3	1	8
	c	Explain Simpson's Paradox	L2	1	4
Module-2					
Q. 03	a	Explain gradient descent approach in detail with relevant example	L3	2	7
	b	Explain in detail on hypothesis testing with example	L3	2	7
	c	How to get data using stdin and stdout ?	L2	2	6
OR					
Q.04	a	Explain the methodologies to extract the data from web scrapping	L3	2	7
	b	Explain data cleaning, data munging and manipulating Data	L2	2	7
	c	Explain dimensionality reduction in detail	L3	2	6
Module-3					
Q. 05	a	Write a program to train a regularized logistic regression classifier on the iris dataset using sklearn. Train the model with the following hyper parameter $C = 1e4$ and report the best classification accuracy.	L3	3	8
	b	What is machine learning? Explain underfitting and overfitting in detail.	L2	3	6
	c	Explain Naive Bayes as really Dumb Spam Filter.	L3	3	6

OR					
Q. 06	a	Write a program to train an SVM classifier on the iris dataset using sklearn. Try different kernels and the associated hyper parameters. Train model with the following set of hyper parameters RBF kernel, gamma=0.5, one-vs-rest classifier, no-feature-normalization. Also try C=0.01,1,10C=0.01,1,10. For the above set of hyper parameters, find the best classification accuracy along with total number of support vectors on the test data	L3	3	8
	b	Explain regression model in detail for predicting the numerical values.	L3	3	6
	c	How support vector machine is used to classify the data explain.	L3	3	6
Module-4					
Q. 07	a	Illustrate the working of decision tree and hence explain importance of entropy in decision tree.	L3	4	7
	b	What is feed forward neural network? Explain the backpropagation method to train neural networks.	L3	4	7
	c	How deep learning is different from machine learning?	L2	4	6
OR					
Q. 08	a	Illustrate the working of Artificial neural network.	L3	4	7
	b	What is clustering and explain K-means clustering in detail.	L3	4	7
	c	Consider the dataset spiral.txt. The first two columns in the dataset corresponds to the co-ordinates of each data point. The third column corresponds to the actual cluster label. Compute the rand index for the following methods: • K – means Clustering • Single – link Hierarchical Clustering • Complete link hierarchical clustering. • Also visualize the dataset and which algorithm will be able to recover the true clusters.	L3	4	6
Module-5					
Q. 09	a	Explain Gibbs Sampling and Topic Modeling	L2	5	7
	b	Write a note on Recurrent Neural Networks	L3	5	7
	c	Explain Word Clouds and n-Gram Language Models	L2	5	6
OR					
Q. 10	a	Write a note on betweenness centrality and eigenvector centrality	L2	5	7
	b	Write a note on recommender systems	L2	5	7
	c	Explain item-based collaborative filtering and matrix factorization.	L2	5	6

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination DATA SCIENCE AND ITS APPLICATIONS

TIME: 03 Hours

Max. Marks: 100

- Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.
03. Write Equations and Diagrams where ever necessary

Module -1			*Bloom's Taxonomy Level	COs	Marks															
Q.01	a	What are the main measures of central tendency? Describe each one. How do you represent a vector in Python using libraries like NumPy?	L2	CO1	7															
	b	What are measures of dispersion, and why are they important?	L2	CO1	6															
	c	What is the Standard Normal distribution. Explain how to use the Z-score to standardize a normal random variable.	L2	CO1	7															
OR																				
Q.02	a	What is Simpson's Paradox, and how is it defined? Describe a situation or provide an example where Simpson's Paradox might occur.	L2	CO1	7															
	b	Explain the difference between correlation and causation. Why is it incorrect to infer causation from correlation alone? Describe an example where correlation does not imply causation.	L2	CO1	7															
	c	Discuss the Central Limit Theorem and its significance in relation to the Normal distribution.? How is the Normal distribution used in hypothesis testing?	L2	CO1	6															
Module-2																				
Q. 03	a	Predict the genre of 'barbie' movie with IMDB=7.4 and duration 114 using KNN, consider k=3 <table border="1" style="margin-left: 20px;"> <thead> <tr> <th>IMDB rating</th> <th>Duration</th> <th>Genre</th> </tr> </thead> <tbody> <tr> <td>8.0 (KGF)</td> <td>160</td> <td>Action</td> </tr> <tr> <td>6.2 (Gadar)</td> <td>170</td> <td>Action</td> </tr> <tr> <td>7.2(Rocky)</td> <td>168</td> <td>Comedy</td> </tr> <tr> <td>8.2 (OMG)</td> <td>155</td> <td>Comedy</td> </tr> </tbody> </table>	IMDB rating	Duration	Genre	8.0 (KGF)	160	Action	6.2 (Gadar)	170	Action	7.2(Rocky)	168	Comedy	8.2 (OMG)	155	Comedy	L3	CO2	8
IMDB rating	Duration	Genre																		
8.0 (KGF)	160	Action																		
6.2 (Gadar)	170	Action																		
7.2(Rocky)	168	Comedy																		
8.2 (OMG)	155	Comedy																		
	b	What is Simple linear Regression? How is error calculated in Linear Regression model? How would u detect overfitting in linear model	L3	CO2	7															
	c	What is an A/B test, and why is it used in data science? Describe the steps involved in designing and running an A/B test.	L2	CO2	5															
OR																				
Q.04	a	Brief out Bootstrapping . Explain how manipulation of data is done and brief out what is name tuples	L2	CO2	7															
	b	What is gradient descent, and why is it important in machine learning? Explain the difference between gradient descent and stochastic gradient descent.	L2	CO2	7															
	c	Explain the mathematical intuition of Multiple linear Regression. Explain	L3	CO2	6															

		the steps			
Module-3					
Q. 05	a	What is a model in the context of machine learning? Explain the difference between supervised and unsupervised learning models.	L2	CO3	6
	b	What is the bias-variance tradeoff in machine learning? How do bias and variance affect model performance?	L2	CO3	7
	c	What is feature extraction, and why is it important in machine learning? Explain the difference between feature extraction and feature selection.	L2	CO3	7
OR					
Q. 06	a	Describe the Iris dataset and its significance in machine learning. What are the features and target variable in the Iris dataset? How is the Iris dataset typically used to demonstrate classification algorithms?	L3	CO3	7
	b	What is the curse of dimensionality, and why is it important in machine learning? How does the curse of dimensionality affect distance-based algorithms like k-NN?	L2	CO3	7
	c	Discuss methods for tuning regularization parameters, such as cross-validation.	L3	CO3	6
Module-4					
Q. 07	a	Can decision trees handle continuous data? If so, how is entropy used to handle continuous data in decision trees? What are the limitations of decision trees	L2	CO4	6
	b	Describe the basic architecture of a Feed-Forward Neural Network. Explain the concept of a loss function.	L2	CO4	7
	c	What is deep learning, and how does it differ from traditional machine learning methods? Describe the general architecture of a deep learning model.	L2	CO4	7
OR					
Q. 08	a	What is an optimization algorithm, and what role does it play in training deep learning models? Explain the concept of gradient descent and its variants	L2	CO4	7
	b	What is clustering, and what is its primary purpose in data analysis? How does clustering differ from classification and regression tasks? ?	L2	CO4	7
	c	What is a loss function, and why is it important in deep learning? Describe common loss functions used for different	L2	CO4	6
Module-5					
Q. 09	a	What is an n-gram in the context of language modeling? In detail give the differences between unigrams, bigrams, and trigrams?	L2	CO5	7
	b	Explain the steps involved in making a prediction using user-based CF?	L3	CO5	7
	c	What are the benefits and limitations of using word clouds for summarizing textual data?	L2	CO5	6
OR					
Q. 10	a	Describe the function of a recurrent layer in a recurrent neural network (RNN). What are the benefits and limitations of using word clouds for summarizing textual data?		CO5	7
	b	What are Recurrent Neural Networks (RNNs), and how do they differ from feedforward neural networks? What are Long Short-Term Memory (LSTM) networks.	L2	CO5	7
	c	What is Item-Based Collaborative Filtering, and how does it differ from User-Based Collaborative Filtering?	L2	CO5	6

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021 (CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination Data Science and Its Applications

TIME: 03 Hours

Max. Marks: 100

Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

Module -1			*Bloom's Taxonomy Level	Cos	Marks
Q.01	a	Define Data Science Explain the Venn diagram of Data Science.	L2	CO1	6
	b	Illustrate Central limit theorem with a neat diagram.	L2	CO1	8
	c	Discuss Random variables with an example in detail.	L2	CO1	6
OR					
Q.02	a	Describe the statement "Correlation is not Causation" with an example in detail.	L2	CO1	6
	b	Write a Python Program to add Two Vectors and Multiply a Vector by Scalar.	L2	CO1	6
	c	Describe Baye's Theorem in detail with an example.	L2	CO1	8
Module-2					
Q. 03	a	Illustrate tqdm Library functions with an example.	L2	CO2	10
	b	Explain the way how Gradient descent is used to fit Parameterized models.	L2	CO2	10
OR					
Q.04	a	Describe A/B test with an example.	L2	CO2	10
	b	Discuss Bayesian Inference in detail.	L2	CO2	10
Module-3					
Q. 05	a	Discuss Bias-Variance tradeoff in detail.	L2	CO3	8
	b	Explain the Standard errors of Regression Coefficients.	L2	CO3	8
	c	Compare Overfitting and Underfitting the training data in Machine Learning.	L2	CO3	4
OR					
Q. 06	a	Discuss the need for fitting the model in Multiple Regression.	L2	CO3	10
	b	Explain Digression in detail.	L2	CO3	10
Module-4					
Q. 07	a	Explain Feedforward Neural Networks with neat diagram.	L2	CO4	10
	b	Compute Tensors in Deep Learning by implementing the concepts in Python.	L3	CO4	10
OR					
Q. 08	a	Construct Linear layers with implementation in Python.	L3	CO4	10
	b	Write a Python Program to train a Network that can compute XOR.	L2	CO4	10
Module-5					
Q. 09	a	Describe n-Gram language models in detail.	L2	CO5	10
	b	Interpret the role of Directed Graph and Page rank in Natural Language processing.	L2	CO5	10
OR					
Q. 10	a	Discuss User-based Collaborative Filtering.	L2	CO5	10
	b	Explain Matrix Factorization with an example.	L3	CO5	10

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination Data Science and its Applications

TIME: 03 Hours

Max. Marks: 100

Note: 01. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

Module -1			*Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is Matplotlib? Write Python program to plot Line chart by assuming your own data and explain the various attributes of line chart.	L3	CO1	8
	b	Describe dispersion and variance and write the python code to compute the variance.	L3	CO1	6
	c	Illustrate Simpson's paradox with an example.	L2	CO1	6
OR					
Q.02	a	Explain standard deviation and interquartile range and write python code to compute standard deviation and interquartile range.	L3	CO1	8
	b	Describe the statement "correlation is not causation" with an example in detail.	L2	CO1	6
	c	Discuss random variables with an example in detail.	L2	CO1	6
Module-2					
Q. 03	a	Explain null and alternative hypothesis by considering the example for a flipping coin. Write a Python program to flip the coin 1000 times and count the number of heads and tails. Based on the results, determine if the coin is fair.	L3	CO2	8
	b	Explain how gradient descent is used to fit parameterized models.	L2	CO2	6
	c	Explain cleaning and munging of data with an example.	L2	CO2	6
OR					
Q.04	a	What is P-Hacking? Describe A/B test with an example	L2	CO2	8
	b	Explain Confidence Intervals with an example.	L2	CO2	6
	c	Consider an html file. Write python program to Scrap the page extract values associated with tags and properties.	L3	CO2	6

Module-3					
Q. 05	a	Explain the simple linear regression model in detail and Write a Python program to illustrate gradient descent for a simple linear regression model.	L3	CO3	8
	b	Describe the role of correctness in machine learning	L2	CO3	6
	c	Explain the use of logistic function in logistic regression in detail.	L2	CO3	6
OR					
Q. 06	a	Describe theoretically the Naive Bayes theorem to model a sophisticated spam filter and write Python program to classify whether a message contains spam or not using Naive Bayes theorem.	L3	CO3	8
	b	Explain over fitting and under fitting in detail with an example.	L2	CO3	6
	C	Explain the goodness of fitting in multiple regression models.	L2	CO3	6
Module-4					
Q. 07	a	Discuss decision tree in detail. Write Python program to create a decision tree.	L3	CO4	10
	b	Explain bottom-up hierarchical clustering approach with an example in detail.	L2	CO4	10
OR					
Q. 08	a	Discuss the role of back propagation algorithm in training neural network.	L2	CO4	10
	b	Explain layer abstraction in deep learning. Write Python program to compute loss and optimization in deep learning.	L3	CO4	10
Module-5					
Q. 09	a	How does item-based collaborative filtering generate recommendations in a recommendation system?	L2	CO5	10
	b	Describe n-Gram language model and how they function in natural language processing	L2	CO5	10
OR					
Q. 10	a	Discuss word clouds and write Python program to generate word clouds.	L3	CO5	10
	b	Explain how grammars are used in modeling languages.	L2	CO5	10

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-I with effect from 2022

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E Degree Examination Data Science and its Applications (21AD62)

TIME: 03 Hours

Max. Marks: 100

Note: Answer any **FIVE** full questions, choosing at least **ONE** question from each module.

Module -1			Marks
Q.01	a	Explain Data Visualization and recognize its use. Sketch Python code segment to visualize line chart and scatterplot with example	06
	b	Summarize dispersion. Using Python code snippet explain the various measures of dispersion	07
	c	Briefly summarize the difference between variance and covariance. Write Python code for finding covariance	07
OR			
Q.02	a	Describe vectors in Data Science and explain any three operations on vectors with Python routine for each operation.	06
	b	Describe Normal Distribution with a Python routine for PDF and CDF	07
	c	Explain Simpson's paradox with an example	07
Module-2			
Q. 03	a	Interpret the importance of <i>power</i> and <i>significance</i> in Statistical Hypothesis Testing with suitable Python routine.	10
	b	Sketch the use of csv.reader, csv.DictReader and csv.writer in processing Delimited Files	10
OR			
Q.04	a	Articulate the role of BeautifulSoup in Web scraping using Python snippet	10
	b	Summarize Stochastic and Minbatch Gradient Descent.	10
Module-3			
Q. 05	a	Explain K-Nearest Neighbors Algorithm using Iris dataset	10
	b	Summarize the concept of overfitting and underfitting with example and explain how it can be resolved?	10
OR			
Q. 06	a	Explain Naïve Bayes Algorithm in the context of classification with functions	10
	b	Explain the various parameters used in checking the correctness of prediction of Machine Learning Model	10
Module-4			
Q. 07	a	Describe Decision Tree. Interpret with Python routine the process of creating Decision Tree.	10
	b	Illustrate the importance of perceptron in Artificial Neural Network	10

OR			
Q. 08	a	Describe the basic idea behind clustering algorithm using color quantization as example	10
	b	Explain Neural Network as a sequence of Layers with functions	10
Module-5			
Q. 09	a	One of the recent advances in NLP is representing words as low-dimensional vectors .How word vectors helps to accomplish this task?	10
	b	Paraphrase bigram, trigram and ngram language models used in Natural Language Processing	10
OR			
Q. 10	a	Summarize Topic Modeling with reference to topic -word distribution and document-topic distribution	10
	b	Explain Word cloud approach in data visualization using Python code snippet	10

Table showing the Bloom's Taxonomy Level, Course Outcome and Program Outcome				
Question		Bloom's Taxonomy Level attached	Course Outcome	Program Outcome
Q.1	(a)	L3	CO 01	PO03
	(b)	L2	CO 01	PO 02
	(c)	L2	CO 01	PO 02
Q.2	(a)	L3	CO 01	PO 03
	(b)	L2	CO 01	PO 02
	(c)	L1	CO 01	PO 01
Q.3	(a)	L3	CO 02	PO 03
	(b)	L2	CO 02	PO 02
Q.4	(a)	L2	CO 02	PO 02
	(b)	L1	CO 02	PO 01
Q.5	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.6	(a)	L2	CO 03	PO 02
	(b)	L1	CO 03	PO 01
Q.7	(a)	L3	CO 04	PO 03
	(b)	L1	CO 04	PO 01
Q.8	(a)	L2	CO 04	PO 02
	(b)	L2	CO 04	PO 02
Q.9	(a)	L3	CO 05	PO 03
	(b)	L1	CO 05	PO 01
Q.10	(a)	L2	CO 05	PO 02
	(b)	L3	CO 05	PO 03

Bloom's Taxonomy Levels	Lower order thinking skills		
	Remembering (knowledge): L ₁	Understanding (Comprehension): L ₂	Applying (Application): L ₃
	Higher order thinking skills		
	Analyzing (Analysis): L ₄	Valuating (Evaluation): L ₅	Creating (Synthesis): L ₆

Model Question Paper-1/2 with effect from 2021 (CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination

Data Science and Its Applications – 21AD62

TIME : 03 Hours

Max. Marks: 100

Note : 1. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

Module - 1			Bloom's Taxonomy Level	COs	Marks
Q.01	a	What is Data Science? Write a short note on data visualization.	02	01	06
	b	With matplotlib explain simple line chart and bar chart.	03	01	08
	c	Illustrate Simpson's paradox with an example.	03	01	06
OR					
Q.02	a	Explain vectors with a code to find the distance between two vectors.	03	01	06
	b	Write a code for the friend counts into a histogram using Counter and plt.bar, describing a single set of data.	03	01	08
	c	State and explain Baye's theorem.	02	01	06
Module - 2					
Q.03	a	What is p-hacking? Write a short note running an A/B testing.	02	02	08
	b	Write a program that counts the lines it receives and then writes out the count.	02	02	06
	c	Explain and write a code using "NamedTuple" class.	03	02	06
OR					
Q.04	a	Compute code to explain the beta distributions.	03	02	06
	b	Write a short note on beautiful soup library.	02	02	06
	c	Illustrate cleaning and munging with suitable code.	03	02	08
Module - 3					
Q.05	a	Explain overfitting and underfitting with suitable code.	03	03	06
	b	What is Iris Dataset? Build a model that can predict the class from the first four measurements.	03	03	08
	c	Write a short note on Maximum Likelihood Estimation.	03	02	06
OR					
Q.06	a	Explain logistic function in detail.	03	03	08
	b	Write a note on simple linear regression using gradient descent.	03	03	06
	c	Write a short note on feature extraction and selection.	02	03	06
Module - 4					
Q.07	a	Define Entropy and write a code for entropy calculation.	03	03	06
	b	Write o function to compute the gradients for backpropagation.	03	03	08
	c	Illustrate K-Nearest Neighbors with the code	03	03	06
OR					
Q.08	a	Write a code to find the minimum entropy partition for the whole dataset.	03	03	08
	b	Write a code to train a network that can compute XOR using newframework.	03	03	06
	c	Write a code that generates any number of clusters by performing the appropriate number of unmerges.	03	03	06

Module - 5					
Q.09	a	Write a code to generate sentences using bigrams.	03	05	06
	b	Write a code to find an eigenvector using matrix_times_vector.	03	05	06
	c	With an example explain the DataSciencester network sized by betweenness centrality.	03	05	08
OR					
Q.10	a	Define and explain Gibbs Sampling with a example.	03	05	08
	b	Write a code to explain the DataSciencester network sized by PageRank.	03	05	06
	c	Write a code to find the interests most similar to Big Data (interest 0)using Item based collaborative filtering.	03	05	06

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination

Subject Title: DATA SCIENCE AND ITS APPLICATIONS

TIME: 03 Hours

Max. Marks: 100

Note: 01. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

		Module -1	*Bloom's Taxonomy Level	COs	Marks																																								
Q.01	a	Explain with example explain the matplotlib library in python.	L2	CO1	6 Marks																																								
	b	Consider the following employees data: <table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th>Name</th> <th>John</th> <th>Mike</th> <th>Sara</th> <th>Tom</th> <th>Alex</th> <th>Nina</th> <th>David</th> </tr> </thead> <tbody> <tr> <td>Dept.</td> <td>IT</td> <td>Marketing</td> <td>HR</td> <td>IT</td> <td>Finance</td> <td>IT</td> <td>Marketing</td> </tr> <tr> <td>Salary</td> <td>50000</td> <td>60000</td> <td>45000</td> <td>55000</td> <td>60000</td> <td>52000</td> <td>58000</td> </tr> <tr> <td>Age</td> <td>25</td> <td>Nan</td> <td>30</td> <td>28</td> <td>Nan</td> <td>32</td> <td>Nan</td> </tr> <tr> <td>Hire_Date</td> <td>01-01-2015</td> <td>02-01-2016</td> <td>01-01-2017</td> <td>03-01-208</td> <td>01-04-2019</td> <td>01-01-2020</td> <td>02-01-2021</td> </tr> </tbody> </table> <ol style="list-style-type: none"> Find the standard deviation of salary of employees in each dept. of a company and identify the department with the highest standard deviation. Find the mean and median salary of employees in each department of the company. Find out the above using python code without using built-in functions.	Name	John	Mike	Sara	Tom	Alex	Nina	David	Dept.	IT	Marketing	HR	IT	Finance	IT	Marketing	Salary	50000	60000	45000	55000	60000	52000	58000	Age	25	Nan	30	28	Nan	32	Nan	Hire_Date	01-01-2015	02-01-2016	01-01-2017	03-01-208	01-04-2019	01-01-2020	02-01-2021	L3	CO1	7 Marks
Name	John	Mike	Sara	Tom	Alex	Nina	David																																						
Dept.	IT	Marketing	HR	IT	Finance	IT	Marketing																																						
Salary	50000	60000	45000	55000	60000	52000	58000																																						
Age	25	Nan	30	28	Nan	32	Nan																																						
Hire_Date	01-01-2015	02-01-2016	01-01-2017	03-01-208	01-04-2019	01-01-2020	02-01-2021																																						
	c	State and illustrate the Central Limit Theorem with a python code using a suitable example.	L3	CO1	7 Marks																																								
OR																																													
Q.02	a	What is Data Science? With example explain the role of a data scientist.	L2	CO1	8 Marks																																								
	b	Find the probability of the given events? <ol style="list-style-type: none"> A single letter is selected at random from the word 'MACHINE LEARNING'. The probability that it is a consonant. The probability of rolling 2 dice to get a sum of 4 or 7. Lottery tokens are numbered from 1 to 25. What is the probability that a token drawn is multiple of 5 or 7? The probability of getting a face in 52 cards. 	L2	CO1	8 Marks																																								
	c	Draw the scatter plot to illustrate the relationship between number of friends and the number of minutes spend on every day. friends = [70, 65, 72, 63, 71, 64, 60, 64, 67] minutes = [175, 170, 205, 120, 220, 130, 105, 145, 190] labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i']	L2	CO1	4 Marks																																								
Module-2																																													
Q.03	a	Explain null and alternative hypothesis by considering the example for a flipping coin.	L3	CO2	7 Marks																																								
	b	Write a python code for scrapping an HTML document with an example	L2	CO2	7 Marks																																								
	c	Illustrate tqdm library by considering an example.	L2	CO2	6 Marks																																								
OR																																													
Q.04	a	Illustrate the difference between named tuples and Data classes with example.	L2	CO2	6 Marks																																								

	b	Explain how gradient descent is use to fit the models.	L3	CO2	7 Marks																																																																						
	c	Explain in detail dimensionality reduction with example.	L3	CO2	7 Marks																																																																						
Module-3																																																																											
Q. 05	a	Describe theoretically the Naive Bayes theorem to model a sophisticated spam filter.	L2	CO3	6 Marks																																																																						
	b	Explain the working of Simple Linear Regression using Gradient descent with an example.	L3	CO3	7 Marks																																																																						
	c	Write Python program to build a K-nearest neighbor model that can predict the class from the IRIS dataset.	L3	CO3	7 Marks																																																																						
OR																																																																											
Q. 06	a	Describe theoretically the Naive Bayes theorem to model a sophisticated spam filter.	L2	CO3	6 Marks																																																																						
	b	Explain in detail the regularization technique in machine learning.	L2	CO3	7 Marks																																																																						
	c	Explain Support Vector Machines in detail.	L3	CO3	7 Marks																																																																						
Module-4																																																																											
Q. 07	a	Consider the following dataset. Write a program to demonstrate the working of the decision tree based ID3 algorithm. <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Price</th> <th>Maintenance</th> <th>Capacity</th> <th>Airbag</th> <th>Profitable</th> </tr> </thead> <tbody> <tr><td>Low</td><td>Low</td><td>2</td><td>No</td><td>Yes</td></tr> <tr><td>Low</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>Low</td><td>Low</td><td>4</td><td>No</td><td>Yes</td></tr> <tr><td>Low</td><td>Med</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Low</td><td>High</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Med</td><td>Med</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Med</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>Med</td><td>High</td><td>2</td><td>Yes</td><td>No</td></tr> <tr><td>Med</td><td>High</td><td>5</td><td>No</td><td>Yes</td></tr> <tr><td>High</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>high</td><td>Med</td><td>2</td><td>Yes</td><td>Yes</td></tr> <tr><td>High</td><td>High</td><td>2</td><td>Yes</td><td>No</td></tr> <tr><td>high</td><td>High</td><td>5</td><td>yes</td><td>Yes</td></tr> </tbody> </table>	Price	Maintenance	Capacity	Airbag	Profitable	Low	Low	2	No	Yes	Low	Med	4	Yes	Yes	Low	Low	4	No	Yes	Low	Med	4	No	No	Low	High	4	No	No	Med	Med	4	No	No	Med	Med	4	Yes	Yes	Med	High	2	Yes	No	Med	High	5	No	Yes	High	Med	4	Yes	Yes	high	Med	2	Yes	Yes	High	High	2	Yes	No	high	High	5	yes	Yes	L3	CO4	7 Marks
Price	Maintenance	Capacity	Airbag	Profitable																																																																							
Low	Low	2	No	Yes																																																																							
Low	Med	4	Yes	Yes																																																																							
Low	Low	4	No	Yes																																																																							
Low	Med	4	No	No																																																																							
Low	High	4	No	No																																																																							
Med	Med	4	No	No																																																																							
Med	Med	4	Yes	Yes																																																																							
Med	High	2	Yes	No																																																																							
Med	High	5	No	Yes																																																																							
High	Med	4	Yes	Yes																																																																							
high	Med	2	Yes	Yes																																																																							
High	High	2	Yes	No																																																																							
high	High	5	yes	Yes																																																																							
	b	Explain feedforward neural network in detail with a neat diagram.	L3	CO4	7 Marks																																																																						
	c	Define Clustering. Explain K-means clustering algorithm with an example.	L3	CO4	6 Marks																																																																						
OR																																																																											
Q. 08	a	Explain layer abstraction in deep learning.	L3	CO4	7 Marks																																																																						
	b	Define neural network? With the diagram explain the implementing AND function using perceptron algorithm.	L3	CO4	7 Marks																																																																						
	c	Explain bottom-up hierarchical clustering approach with an example in detail.	L3	CO4	6 Marks																																																																						
Module-5																																																																											
Q. 09	a	Describe n-Gram language models in detail.	L3	CO5	8 Marks																																																																						
	b	Explain Recurrent Neural Network in detail.	L3	CO5	7 Marks																																																																						
	c	Explain item based collaborative filtering.	L3	CO5	5 Marks																																																																						
OR																																																																											
Q. 10	a	Explain how grammars are used in modeling languages.	L3	CO5	7 Marks																																																																						
	b	Explain eigen vector centrality in detail.	L3	CO5	7 Marks																																																																						
	c	Discuss matrix factorization in detail.	L3	CO5	6 Marks																																																																						

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

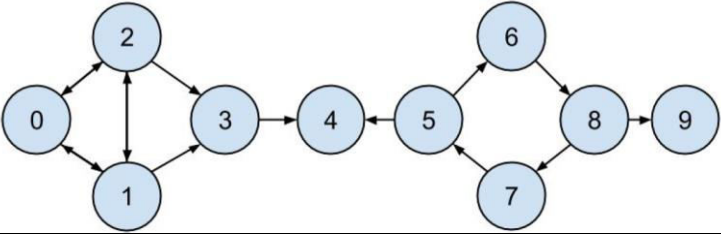
Sixth Semester B.E. Degree Examination Data Science and its Applications

TIME: 03 Hours

Max. Marks: 100

- Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.
03.
04.

Module -1			*Bloom's Taxonomy Level	COs	Marks																
Q.01	a	Develop a python program to plot a bar chart for the following data: Movies=["Annie Hall","Ben-Hur","Casablanca", "Gandhi", "West Side Story"] num_oscars = [5, 11, 3, 8, 10]. Also, draw the bar chart and label x and y axes.	L3	1	10																
	b	(i) Explain Simpson Paradox with an example. (ii) Develop python function for computing range of any given data and define dispersion.	L3	1	10																
	c	Explain probability density function and cumulative distribution function with example with relevant equations.	L2	1	10																
OR																					
Q.02	a	Develop a python program to plot a line chart for the given data. Explain the various attributes of the line chart. Draw the line chart. <table border="1" style="margin-left: 20px;"> <tr> <td>Year</td> <td>1950</td> <td>1960</td> <td>1970</td> <td>1980</td> <td>1990</td> <td>2000</td> <td>2010</td> </tr> <tr> <td>GDP</td> <td>300.2</td> <td>543.3</td> <td>1075.5</td> <td>2862.5</td> <td>5979.6</td> <td>10289.7</td> <td>14958.3</td> </tr> </table>	Year	1950	1960	1970	1980	1990	2000	2010	GDP	300.2	543.3	1075.5	2862.5	5979.6	10289.7	14958.3	L3	1	10
Year	1950	1960	1970	1980	1990	2000	2010														
GDP	300.2	543.3	1075.5	2862.5	5979.6	10289.7	14958.3														
	b	Develop python functions for computing the components of central tendencies with explanation.	L3	1	10																
	c	Describe the following probability concepts: (i) Conditional Probability (ii) Bayes Theorem (iii) Central Limit Theorem (iv) Normal Distribution (v) Random Variables	L2	1	10																
Module-2																					
Q. 03	a	Explain statistical hypothesis testing with examples.	L2	2	10																
	b	Develop a python code for counting number of lines and counting 10 most repeated words in the given file using stdin and stdout and regular expression.	L3	2	10																
	c	Illustrate 1D,2D and multi-dimensional data with example.	L2	2	10																
OR																					
Q.04	a	Explain A/B testing with an example, with relevant equation.	L2	2	10																
	b	Consider a HTML file and build a python program to scrap the page, extract values associated with tags and properties.	L3	2	10																
	c	Explain with an example the concept of rescaling.	L2	2	10																

Module-3					
Q. 05	a	Describe the following: (i)Overfitteng and Underfitting (ii)Bias-Variance Tradeoff (iii)Correctness	L2	3	10
	b	Develop a python program to train Logistic Regression Classifier on Iris Dataset using sklearn.	L3	3	10
	c	Explain Naïve Bayes algorithm.	L2	3	10
OR					
Q. 06	a	Explain KNN Classifier with example.	L2	3	10
	b	Develop a python program to train an SVM classifier on the iris dataset using sklearn. Try different kernels and the associated hyperparameters. Train model with the following set of hyperparameters RBFkernel, gamma=0.5, one-vs-rest classifier, no-feature-normalization. Also try C=0.01,1,10C=0.01,1,10. For the above set of hyperparameters, find the best classification accuracy along with total number of support vectors on the test data .	L3	3	10
	c	Describe regularization in detail.	L2	3	10
Module-4					
Q. 07	a	Demonstrate the role of entropy and entropy partition in creating decision tree with explanation and python code.	L3	4	10
	b	Illustrate K means clustering with example.	L2	4	10
	c	Explain the following terms of Neural Network : (i)Feed-Forward Neural Networks, (ii)Backpropagation	L2	4	10
OR					
Q. 08	a	Describe bottom up hierarchical clustering with example .	L2	4	10
	b	Build relevant code and explain Random forests algorithm.	L3	4	10
	c	Illustrate the working of Perceptron using OR Gate and AND Gate as example.	L2	4	10
Module-5					
Q. 09	a	Illustrate the following basic terminologies of Natural language processing (i) Gibbs Sampling, (ii)n- Gram model	L2	5	10
	b	Build with relevant python code and explain topic modeling for natural language processing.	L3	5	10
	c	Explain the following techniques used for recommender system (i)User based collaborative filtering (ii) Item based collaborative filtering	L2	5	10
OR					
Q. 10	a	Illustrate the following basic terminologies of Natural language processing (i)Word Cloud , (iv)Grammar	L2	5	10
	b	Explain the following metrics used for Network Analysis: (i) Betweenness centrality,(ii)Closeness centrality, (iii)Eigen vector centrality	L2	5	10
	c	Develop for a python function for Page Rank algorithm for the following Directed Graph 	L3	5	10

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination

Subject Title: DATA SCIENCE AND ITS APPLICATIONS

TIME: 03 Hours

Max. Marks: 100

Note: 01. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

		Module -1	*Bloom's Taxonomy Level	COs	Marks																																								
Q.01	a	Explain with example explain the matplotlib library in python.	L2	CO1	6 Marks																																								
	b	Consider the following employees data: <table border="1" style="margin: 5px 0;"> <thead> <tr> <th>Name</th> <th>John</th> <th>Mike</th> <th>Sara</th> <th>Tom</th> <th>Alex</th> <th>Nina</th> <th>David</th> </tr> </thead> <tbody> <tr> <td>Dept.</td> <td>IT</td> <td>Marketing</td> <td>HR</td> <td>IT</td> <td>Finance</td> <td>IT</td> <td>Marketing</td> </tr> <tr> <td>Salary</td> <td>50000</td> <td>60000</td> <td>45000</td> <td>55000</td> <td>60000</td> <td>52000</td> <td>58000</td> </tr> <tr> <td>Age</td> <td>25</td> <td>Nan</td> <td>30</td> <td>28</td> <td>Nan</td> <td>32</td> <td>Nan</td> </tr> <tr> <td>Hire_Date</td> <td>01-01-2015</td> <td>02-01-2016</td> <td>01-01-2017</td> <td>03-01-208</td> <td>01-04-2019</td> <td>01-01-2020</td> <td>02-01-2021</td> </tr> </tbody> </table> <ol style="list-style-type: none"> Find the standard deviation of salary of employees in each dept. of a company and identify the department with the highest standard deviation. Find the mean and median salary of employees in each department of the company. Find out the above using python code without using built-in functions.	Name	John	Mike	Sara	Tom	Alex	Nina	David	Dept.	IT	Marketing	HR	IT	Finance	IT	Marketing	Salary	50000	60000	45000	55000	60000	52000	58000	Age	25	Nan	30	28	Nan	32	Nan	Hire_Date	01-01-2015	02-01-2016	01-01-2017	03-01-208	01-04-2019	01-01-2020	02-01-2021	L3	CO1	7 Marks
Name	John	Mike	Sara	Tom	Alex	Nina	David																																						
Dept.	IT	Marketing	HR	IT	Finance	IT	Marketing																																						
Salary	50000	60000	45000	55000	60000	52000	58000																																						
Age	25	Nan	30	28	Nan	32	Nan																																						
Hire_Date	01-01-2015	02-01-2016	01-01-2017	03-01-208	01-04-2019	01-01-2020	02-01-2021																																						
	c	State and illustrate the Central Limit Theorem with a python code using a suitable example.	L3	CO1	7 Marks																																								
OR																																													
Q.02	a	What is Data Science? With example explain the role of a data scientist.	L2	CO1	8 Marks																																								
	b	Find the probability of the given events? <ol style="list-style-type: none"> A single letter is selected at random from the word 'MACHINE LEARNING'. The probability that it is a consonant. The probability of rolling 2 dice to get a sum of 4 or 7. Lottery tokens are numbered from 1 to 25. What is the probability that a token drawn is multiple of 5 or 7? The probability of getting a face in 52 cards. 	L2	CO1	8 Marks																																								
	c	Draw the scatter plot to illustrate the relationship between number of friends and the number of minutes spend on every day. friends = [70, 65, 72, 63, 71, 64, 60, 64, 67] minutes = [175, 170, 205, 120, 220, 130, 105, 145, 190] labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i']	L2	CO1	4 Marks																																								
Module-2																																													
Q.03	a	Explain null and alternative hypothesis by considering the example for a flipping coin.	L3	CO2	7 Marks																																								
	b	Write a python code for scrapping an HTML document with an example	L2	CO2	7 Marks																																								
	c	Illustrate tqdm library by considering an example.	L2	CO2	6 Marks																																								
OR																																													
Q.04	a	Illustrate the difference between named tuples and Data classes with example.	L2	CO2	6 Marks																																								

	b	Explain how gradient descent is use to fit the models.	L3	CO2	7 Marks																																																																						
	c	Explain in detail dimensionality reduction with example.	L3	CO2	7 Marks																																																																						
Module-3																																																																											
Q. 05	a	Describe theoretically the Naive Bayes theorem to model a sophisticated spam filter.	L2	CO3	6 Marks																																																																						
	b	Explain the working of Simple Linear Regression using Gradient descent with an example.	L3	CO3	7 Marks																																																																						
	c	Write Python program to build a K-nearest neighbor model that can predict the class from the IRIS dataset.	L3	CO3	7 Marks																																																																						
OR																																																																											
Q. 06	a	Describe theoretically the Naive Bayes theorem to model a sophisticated spam filter.	L2	CO3	6 Marks																																																																						
	b	Explain in detail the regularization technique in machine learning.	L2	CO3	7 Marks																																																																						
	c	Explain Support Vector Machines in detail.	L3	CO3	7 Marks																																																																						
Module-4																																																																											
Q. 07	a	Consider the following dataset. Write a program to demonstrate the working of the decision tree based ID3 algorithm. <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Price</th> <th>Maintenance</th> <th>Capacity</th> <th>Airbag</th> <th>Profitable</th> </tr> </thead> <tbody> <tr><td>Low</td><td>Low</td><td>2</td><td>No</td><td>Yes</td></tr> <tr><td>Low</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>Low</td><td>Low</td><td>4</td><td>No</td><td>Yes</td></tr> <tr><td>Low</td><td>Med</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Low</td><td>High</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Med</td><td>Med</td><td>4</td><td>No</td><td>No</td></tr> <tr><td>Med</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>Med</td><td>High</td><td>2</td><td>Yes</td><td>No</td></tr> <tr><td>Med</td><td>High</td><td>5</td><td>No</td><td>Yes</td></tr> <tr><td>High</td><td>Med</td><td>4</td><td>Yes</td><td>Yes</td></tr> <tr><td>high</td><td>Med</td><td>2</td><td>Yes</td><td>Yes</td></tr> <tr><td>High</td><td>High</td><td>2</td><td>Yes</td><td>No</td></tr> <tr><td>high</td><td>High</td><td>5</td><td>yes</td><td>Yes</td></tr> </tbody> </table>	Price	Maintenance	Capacity	Airbag	Profitable	Low	Low	2	No	Yes	Low	Med	4	Yes	Yes	Low	Low	4	No	Yes	Low	Med	4	No	No	Low	High	4	No	No	Med	Med	4	No	No	Med	Med	4	Yes	Yes	Med	High	2	Yes	No	Med	High	5	No	Yes	High	Med	4	Yes	Yes	high	Med	2	Yes	Yes	High	High	2	Yes	No	high	High	5	yes	Yes	L3	CO4	7 Marks
Price	Maintenance	Capacity	Airbag	Profitable																																																																							
Low	Low	2	No	Yes																																																																							
Low	Med	4	Yes	Yes																																																																							
Low	Low	4	No	Yes																																																																							
Low	Med	4	No	No																																																																							
Low	High	4	No	No																																																																							
Med	Med	4	No	No																																																																							
Med	Med	4	Yes	Yes																																																																							
Med	High	2	Yes	No																																																																							
Med	High	5	No	Yes																																																																							
High	Med	4	Yes	Yes																																																																							
high	Med	2	Yes	Yes																																																																							
High	High	2	Yes	No																																																																							
high	High	5	yes	Yes																																																																							
	b	Explain feedforward neural network in detail with a neat diagram.	L3	CO4	7 Marks																																																																						
	c	Define Clustering. Explain K-means clustering algorithm with an example.	L3	CO4	6 Marks																																																																						
OR																																																																											
Q. 08	a	Explain layer abstraction in deep learning.	L3	CO4	7 Marks																																																																						
	b	Define neural network? With the diagram explain the implementing AND function using perceptron algorithm.	L3	CO4	7 Marks																																																																						
	c	Explain bottom-up hierarchical clustering approach with an example in detail.	L3	CO4	6 Marks																																																																						
Module-5																																																																											
Q. 09	a	Describe n-Gram language models in detail.	L3	CO5	8 Marks																																																																						
	b	Explain Recurrent Neural Network in detail.	L3	CO5	7 Marks																																																																						
	c	Explain item based collaborative filtering.	L3	CO5	5 Marks																																																																						
OR																																																																											
Q. 10	a	Explain how grammars are used in modeling languages.	L3	CO5	7 Marks																																																																						
	b	Explain eigen vector centrality in detail.	L3	CO5	7 Marks																																																																						
	c	Discuss matrix factorization in detail.	L3	CO5	6 Marks																																																																						

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination DATA SCIENCE AND ITS APPLICATIONS

TIME: 03 Hours

Max. Marks: 100

- Note: 01. 02. Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.
03.
04.

Module -1			*Bloom's Taxonomy Level	COs	Mark s
Q.01	a	Describe the process of creating a bar chart using matplotlib. What information is typically conveyed by a bar chart?	L2	1	4
	b	Describe Bayes's Theorem and its significance in statistical inference. How can Bayes's Theorem be applied to improve classification models?	L2	2	8
	c	Explain the concept of correlation and its significance in data analysis. Discuss Simpson's Paradox and other correlational caveats with examples.	L2	1	8
OR					
Q.02	a	Discuss the concept of vectors and matrices in Linear Algebra. Provide examples of how they are used in data manipulation and machine learning.	L2	1	10
	b	Explain the Central Limit Theorem and its importance in statistics. Provide an example to demonstrate how it is used in practical data analysis.	L3	1	10
Module-2					
Q. 03	a	What are p-values and confidence intervals in the context of hypothesis testing? Discuss their significance and how they are used in making statistical inferences.	L2	2	10
	b	Describe the steps involved in obtaining data from various sources such as stdin, stdout, reading files, web scraping, and using APIs. Provide a detailed example of using the Twitter API to gather data.	L2	2	10
OR					
Q.04	a	Describe the process of statistical hypothesis testing. Using the example of flipping a coin, explain how you would determine if a coin is fair or biased.	L2	2	10
	b	What is gradient descent? Explain the idea behind gradient descent and how it is used to fit models. Discuss the differences between batch gradient descent, minibatch gradient descent, and stochastic gradient descent.	L2	2	10
	c				
Module-3					
Q. 05	a	Define machine learning and discuss the difference between overfitting and underfitting. How can these issues be mitigated in model training?	L2	3	10
	b	Discuss the process of simple linear regression. Explain how gradient descent and maximum likelihood estimation are used to fit the model.	L2	3	10
OR					
Q. 06	a	Illustrate the process of feature extraction and selection in machine learning. Why is this step important, and what techniques are commonly used?	L3	3	7
	b	Explain the k-nearest neighbors (k-NN) algorithm with an example	L2	3	7

	c	What is the Naive Bayes algorithm? Illustrate its application with an example of a spam filter.	L3	3	6
Module-4					
Q. 07	a	Explain the concept of a decision tree and describe how entropy is used to create a decision tree. Provide an example to illustrate the process.	L3	4	8
	b	Describe the basic structure and function of a perceptron. How is it used as a building block in feed-forward neural networks?	L2	4	6
	c	Describe the concept of clustering and its applications in data science.	L2	4	6
OR					
Q. 08	a	Discuss the concept of random forests and how they improve upon individual decision trees. Explain the advantages and potential drawbacks of using random forests	L2	4	10
	b	Compare and contrast different activation functions used in neural networks. Discuss their impact on model performance with examples.	L2	4	10
Module-5					
Q. 09	a	Explain the concept of word clouds and their use in natural language processing with an example	L2	5	6
	b	Illustrate the PageRank algorithm and its application in directed graphs. How does it work and what is its significance in network analysis?	L3	5	7
	c	What is matrix factorization in the context of recommender systems? Discuss how it is used to improve recommendation accuracy and provide an example.	L2	5	7
OR					
Q. 10	a	Discuss n-gram language models and their application in NLP. How do these models help in understanding the context within a text? Provide an example.	L2	5	10
	b	Discuss the architecture and function of recurrent neural networks (RNNs). Provide an example of using a character-level RNN in a text generation task.	L2	5	10

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.

Model Question Paper-1/2 with effect from 2021(CBCS Scheme)

USN

--	--	--	--	--	--	--	--	--	--

Sixth Semester B.E. Degree Examination Data Science and its Applications

TIME: 03 Hours

Max. Marks: 100

Note: Answer any **FIVE** full questions, choosing at least **ONE** question from each **MODULE**.

Module -1			*Bloom's Taxonomy Level	COs	Marks
Q.01	a	What are random variables? State Bayes's theorem in detail with an example.			
	b	Describe matrices. Compute a function to create a matrix given its shape and a function for generating its elements and use the same function to generate to 5 x 5 identity matrix.			
	c	Illustrate central limit theorem.			
OR					
Q.02	a	Who is a Data scientist? Draw the data science life cycle in detail.			
	b	Describe Dispersion. Compute code to compute standard deviation.			
	c	Illustrate normal distribution and continuous distribution in detail.			
Module-2					
Q. 03	a	Write a note on confidence intervals in detail.			
	b	Compute code to estimate the gradient.			
	c	Illustrate A/B test with an example.			
OR					
Q.04	a	Write a note on null and alternative hypothesis by considering the example for a flipping coin.			
	b	Compute code to explain the beta distributions.			
	c	Illustrate p-Values with an example.			
Module-3					
Q. 05	a	Write a short note on digression with code.			
	b	Explain logistic function in detail.			
	c	Illustrate standard errors of regression coefficients.			
OR					
Q. 06	a	Write a note on simple linear regression using gradient descent.			

	b	Explain super vector machines in detail.			
	c	Illustrate regularization.			
Module-4					
Q. 07	a	What is entropy in the context of information theory and how it is used to measure the impurity of a dataset in decision tree algorithms with a mathematical formula and explain its components?			
	b	Explain clustering and its importance in unsupervised learning and explain the k-means clustering algorithm with a numerical example.			
OR					
Q. 08	a	Explain the process of training a neural network on the MNIST dataset with the architecture, input preprocessing, evaluation metrics and provide a summary of the network's performance.			
	b	Illustrate the backpropagation algorithm and its importance in training neural networks and explain with an example how gradients are computed and weights are updated.			
Module-5					
Q. 09	a	Write a note on word cloud and its utility in visualizing textual data. Provide an example of a word cloud generated from the following text: "The quick brown fox jumps over the lazy dog."			
	b	Describe Gibbs sampling and its application in machine learning or statistical modeling. Provide an example of using Gibbs sampling to estimate parameters in a Bayesian model.			
	c	Describe the architecture of a recurrent neural network (RNN) and its application in sequential data modeling. Implement a simple character-level RNN using Python and train it on a text dataset.			
OR					
Q. 10	a	Define network analysis and how two centrality measures are used to evaluate node importance in a network. Calculate the degree centrality and betweenness centrality of nodes in a small social network graph.			
	b	Explain PageRank with Hypertext Induced Topic Selection algorithm in terms of their underlying principles and use cases.			
	c	Compare singular value decomposition with probabilistic matrix factorization in terms of their suitability for recommendation systems.			

*Bloom's Taxonomy Level: Indicate as L1, L2, L3, L4, etc. It is also desirable to indicate the COs and POs to be attained by every bit of questions.