(b) How will you determine the number of clusters in a clustering algorithm? Determine the number of cluster required for the below table,

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

Explain the result with proper graphs / diagram.

(3 + 3) + (3 + 3) = 12

7. (a) Hierarchical clustering is a powerful technique that allows us to build tree structures from data similarities. Explain with Python code or Give a Case Study how can Dendrogram help in Hierarchical clustering?

(b) Define Visual Analytics. Why and when do we use Graph database?

8 + 4 = 12

Group – E

8. (a) Name some Deep Learning Frameworks.

(b) How Naive Bayes classifier works? In Fig 1, we have a training data set of weather with corresponding target variable 'Play'.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Fig.1

Calculate the prior probability for given class labels. Find Likelihood probability with each attribute for each class. Put these value in Bayes Formula and calculate posterior probability to prove Players will play if weather is sunny.

4 + 8 = 12

9. (a) One of the formal definition of visualization says "... is the process of extracting salient features from the sets of data and displaying the features in an intuitive and expressive way". Comment on all the italicized terms in the definition.

(b) (i) What are Mackinlay's design criteria?
(ii) Give examples for each of the following:
(a) Structural visualization (b) Temporal visualization (c) Geospatial visualization (d) Multidimensional visualization.

5 + (3 + 4) = 12

DATA SCIENCE
(CSEN 5141)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Which of the following step is performed by data scientist after acquiring the data?
(a) Data Replication (b) Data Integration
(c) Data Cleansing (d) None of (a), (b) and (c).

(ii) Methods to treat missing values are
(a) list wise deletion and pair wise deletion and
(b) imputation for missing values
(c) prediction
(d) outlier detection and (c) and (b).

(iii) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.
Which of the following statement is true in following case?
(a) Feature F1 is an example of nominal variable
(b) Feature F1 is an example of ordinal variable
(c) It doesn't belong to any of the above category
(d) None of the above.

(iv) Correlation coefficients are used in statistics to measure
(a) How strong a relationship is between two variables
(b) How strong a relationship is between two values
(c) Experiments performed in the EDA process
(d) Required in Percentile based outlier removal.

(v) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?
1) In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.

2) In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.
3) In GD, you either use the entire training data to update a parameter in each iteration.
(a) Only (1) (b) (1) and (3) (c) (1), (2) and (3) (d) Only (3).

(vi) Which of the following statements is/are true about "Type-1" and "Type-2" errors?
1) Type1 is known as false positive and Type2 is known as false negative.
2) Type1 is known as false negative and Type2 is known as false positive.
3) Type1 error occurs when we reject a null hypothesis when it is actually true.
(a) Only (1) (b) (2) and (3) (c) (1) and (3) (d) Only (3)

(vii) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively.
Now, you have added 2 in all values of X (i.e. new values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 & D3 respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?
(a) D1= C1, D2 < C2, D3 > C3 (b) D1 = C1, D2 = C2, D3 = C3
(c) D1> C1, D2 < C2, D3 =C3 (d) D1 < C1, D2 < C2, D3 < C3.

(viii) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?
(a) 1 (b) 2 (c) 8 (d) 4.

(ix) Tableau is a tool used for complex visualization and simplification of complex data because it is
(a) Beautiful and Interactive Dashboard (b) Speed and easy to Use
(c) Has Predictive Analytical Capabilities (d) Both (a) and (b).

(x) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?
(a) K-NN (b) Random Forest
(c) Linear Regression (d) All of (a), (b) and (c).

Group – B

2. (a) Define Data Science. What are the different categories of data and explain any four of them?

(b) What are the methods of Data Exploration? Give 2 mechanisms of Data Exploration.

(2 + 4) + (3 + 3) = 12

3. (a) What is big data? How it is different from traditional data? What is the importance of big data with respect to an Industry?

(b) What do you gain by studying Different Distributions? Name 4 types of Distribution. Explain – Binomial Distribution.

(2 + 2 + 2) + (2 + 2 + 2) = 12

Group – C

4. (a) Explain Logistic Regression. Write a python code for Prediction using any Regression technique.

(b) Define Precision, Recall and Accuracy.

(4 + 4) + 4 = 12

5. (a) (i) Write the difference between structured data and rectangular data? How can you represent structured data into rectangular data using python?

(ii) What is the sampling distribution of the sample mean? Why does a small sample size cause problems?

(iii) What is Central Limit Theorem and when it is needed?

(b) Assume that 15% of students at a university wear contact lenses. We randomly pick 200 students. What is the standard deviation of the proportion of students in this group who may wear contact lenses?

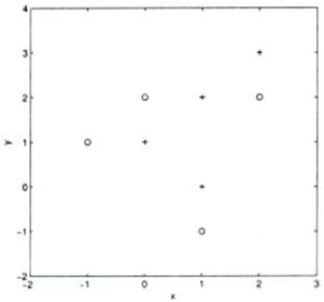
[(2 + 1) + (2 + 2) + 2] + 3 = 12

Group – D

6. (a) Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

Below is a scatter plot which shows the above data in 2D space.



Suppose, you want to predict the class of new data point x = 1 and y =1 using Euclidian distance in 3-NN. In which class this data point belong to? What change do you observe if try to predict the class using 6-NN? Give proper explanation.