SRN | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE17CS322

ESA B.TECH. 5th SEMESTER - December 2019

UE17CS322- DATA ANALYTICS (Dr. GS)

Time: 180 minutes | Answer All Questions | Max Marks: 100

There are four pages to this question paper with FIVE questions for a total of 100 marks. Read the questions carefully and answer all questions briefly and to the point.

Question 1 [20 marks]

(a) [6 marks (2+2+2)] *Are people mindful of the C-footprint they leave on a daily basis? Are they mindful of the usage of water or their use of plastic?* You are tasked to collect demographical and attitudinal factors of your fellow citizens to answer these and other questions relating to the environment and climate change.

Suggest **(i)** any two ways through which you can collect this data and the format in which you would store the data to facilitate further analysis, **(ii)** any one input you would collect that can be validated from a secondary source; also identify the secondary source used for validation **(iii)** any other interesting insight (that is unrelated to climate change) that you can get from this data.

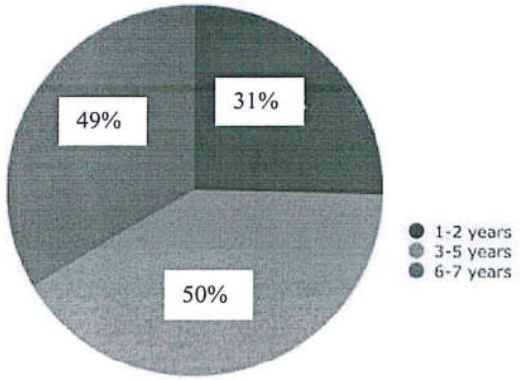
(b) [8 marks ((3+3)+2)] **(i)** In the table given below, the length (in km) of various types of roads during 2007 - 2015 have been presented. What is the exploratory data analysis (EDA) you would perform on this data? Suggest any two inferences you could arrive at based on the EDA.

(ii) Does any entry in the table appear to be anomalous? Justify briefly. (Hint: Look at the entries in 2014, 2015)

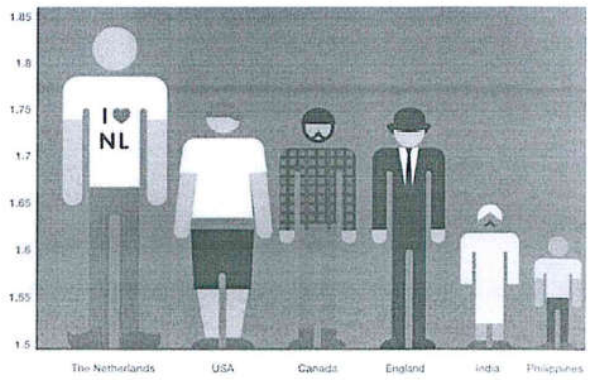
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
National Highways	66590	66754	70548	70934	70934	76818	79116	91287	97991
State Highways	152235	154522	158497	160177	163898	164360	169227	170818	167109
Other PWD Roads	835003	863241	962880	977414	998895	1022287	1066747	1082267	1101178
Panchayat Raj Roads	1372867	1388750	1514952	1518205	1530366	1587787	1725318	1800747	1831043

(c) [6 marks (4+2 marks)] Identify any two errors in each of the two visualizations presented on the following page. Suggest an alternative visualization technique for the data in each of these figures.

SRN | | | | | | | | | | | |



(i) Years of experience required by employers



(ii) Average male height (in meters)

Question 2 [20 points]

(a) **[6 marks** (2+2+2)] Answer the following with a suitable justification:
If the Pearson's correlation coefficient for two variables is 0.85, can we conclude:
(i) the two variables are necessarily correlated?
(ii) the two variables do not share a cause-effect relationship? (because correlation does not imply causation)
(iii) the two variables can be modeled using linear regression?

(b) **[8 marks** (2+2+4)] The data given below depicts the time taken to read and the number of sentences in articles that are either technical (T) or nontechnical (N).

Time taken to read an article (in hrs)	2.7	1.4	3.3	1.3	3	7.6	5.9	6.9	8.6	7.7	1.9
Number of sentences ($*10^3$)	2.5	2.3	2.4	1.8	3	2.7	2.2	1.8	3.5	3.5	3.1
Type of article	N	N	N	N	N	T	T	T	T	T	?

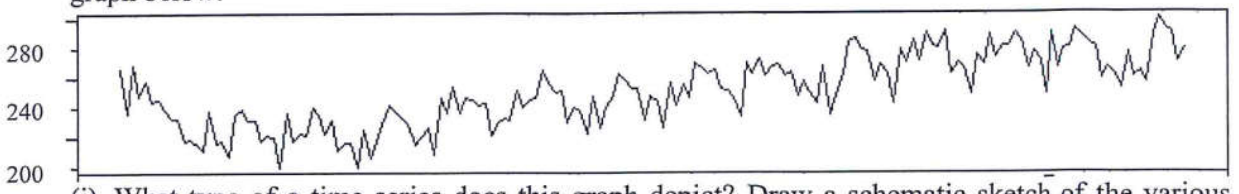
Given: $\beta_1 = 0.41444855$ (the slope corresponding to the number of hours to read the article) and $\beta_2 = -0.2486209$ (the slope corresponding to the number of sentences in the article) (i) suggest a method to compute β_0, the intercept (note: do actually compute it) (ii) How can regression be used to label the article with 3100 words and that takes 1.9 hours to read? (iii) What is the difference between Ridge and Lasso regression and how can one of these be used for feature selection?

(c) **[6 marks** (3+3)] (i) You have been tasked by Google to select interns for the GSoc program. The data comprises ten features (scores in eight relevant courses and two entrance tests). On modeling this data using multiple regression, you find the R2 value for model A = 0.81 and the R2 value for model B=0.78. Does this mean model A is better than model B? Why or why not? (ii) In the context of logistic regression for binary classification, how is an 'RoC curve' drawn? When there are multiple classification models available, how are the corresponding RoC curves compared to select a model that is most suitable for the given problem?

SRN | | | | | | | | | | | | |

Question 3 [20 marks]

(a) **[6 marks** (4+2)**]** The price of 1kg of onions over the past couple of months has been depicted as a graph below:



(i) What type of a time series does this graph depict? Draw a schematic sketch of the various components of this data.

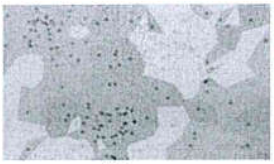
(ii) If the price of onions is Rs. 210/ kg in October, Rs. 220/ kg in November and Rs. 230/ kg in December, what would the price per kg be in January if we were to use a simple smoothing filter of size 3 for forecasting?

(b) **[8 marks** (2+2)+4)] (i) How do we determine whether a signal is stationary or not and why is stationarity important for a time series? Suggest any two methods that can be used to remove (or minimize) the nonstationarities in a time series signal. (ii) What are the ACF and PACF? Briefly explain how these functions can be used to estimate model parameters for AR and MA models.

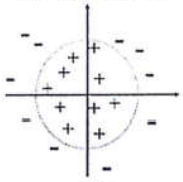
(c) **[6 marks** (3+3)**]** For each of the models below, write the expression for computing the output at time t and describe the nature of the model (or the characteristics of the underlying time series):
(i) ARIMA(0,0,1) and (ii) ARIMA(1,1,0)

Question 4 [20 marks]

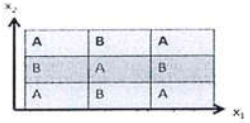
(a) [6 marks (2+2 +2)**]** Which supervised learning method is most suited for the spread of data given below? Briefly justify your answer for each case:



(i)



(ii)



(iii)

(b) [8 marks (2+2+4)**]** For each of the situations below, identify the type of recommendation system (collaborative filtering, content based filtering or knowledge based filtering) that must have been used and briefly explain your rationale for the answer:

(i) You have just purchased a book on Amazon and immediately it suggests 10 other books you may be interested in.

(ii) You create an account on Coursera to sign up for a certification course. Even before you search for which you would like to register, Coursera suggests some courses you may find interesting.

(iii) For the data given below:

Customer id	1	2	3	4	5
Items bought	{a,b,c,d,e}	{a,b,c,d,e}	{b,c,d,e}	{a,b,c,d}	{a,b,d,e}

Which is the stronger association rule? {e}→{b,d} or {b,d}→{e}. Briefly explain.

(c) [6 marks (4+2)**]** There have been a number of projects related to Machine Learning, AI, etc., in the recent years. The Department is keen to group the projects into a small number of categories such as 'theory', 'NLP', 'computer vision', etc., to facilitate easy indexing and search. Given the digital copies of the project reports (i) suggest the steps you would take to cluster similar documents and tag them with the appropriate category labels (ii) Given a new project report, how would you determine which class it belongs to?

Question 5 [20 marks]

(a) **[6 marks** (4+2)**]** (i) In a study on determining the relationship between activity level and weight gained, how can we determine whether age is a confounding variable? If age is found to be a confounding factor, how can we improve our analysis? (ii) Suggest any two conditions that help us rule out a variable as a confounding factor.

(b) **[8 marks** (4+(2+2))**]** (i) An insurance company classifies drivers as low-risk if they are accident-free for one year. Past records indicate that 98% of the drivers in the low-risk category (L) will remain in that category the next year, and 78% of the drivers who are not in the low-risk category (L') one year will be in the low-risk category the next year. Find the transition matrix P for this problem and prove that S = [0.975 0.025] is the steady state matrix.

(ii) We have studied the concept of bull, bear and stagnant markets in class. Assuming the transition probabilities are as described in the matrix **M** below, and the state in the current week is described by matrix C = [0 1 0], what state is C most likely to be in **one week from now** and **five weeks from now**?

M =

From \ To	**Bull**	**Bear**	**Stagnant**
Bull	0.9	0.075	0.025
Bear	0.15	0.18	0.05
Stagnant	0.25	0.25	0.5

(c) [6 marks (3+3)**]** (i) How is sparse PCA different from PCA and when is sparse PCA used? (ii) Why is Latent Semantic Analysis (LSA) preferred when finding similar documents or keywords over PCA?

*****