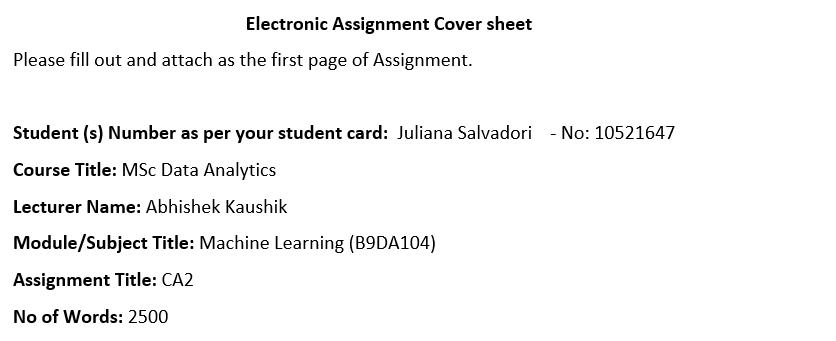
****

**Q1**

● Define Data sampling and its steps with examples.

Data sampling is a statistical method used to help data analysts to discover patterns in a large set of data without the need to analyse the whole set of data, which would be time-consuming and costly. It means it is possible to find meaningful information about a population by analysing a sample of that population. For example, to know if a cake is good or not, you do not need to eat the whole cake, you just need to taste a slice of the cake. Steps:

1) define population to be analysed - cakes from bakery X

2) define sampling frame, that is, where the sample will be selected from - you want to analyse one cake or the entire cake production for a day.

3) define method of data sampling to be applied. The method can be a probability or non-probability sampling, and there are various methods for each of these types.

4) define sampling size - it will depend on the population size, to analyse one cake, one slice of it would be enough, to analyse the entire cake production for a day, it will probably necessary to take several slices of cake during that day.

5) and then, perform the sample collection and analyse.

References:

Cavello, M. (2020) ‘How to See the Bigger Picture with Data Sampling’, Learning Hub. Available at: <https://learn.g2.com/data-sampling#:~:text=What%20is%20data%20sampling%3F,often%20it%20should%20be%20collected>

(Accessed: 01 Nov 2020)

Rouse, M. (2018) ‘Data Sampling’, Search Business Analytics. Available at:

<https://searchbusinessanalytics.techtarget.com/definition/data-sampling> (Accessed: 01 Nov 2020)

Analytics help, Sampled data and (other), About data sampling (2020) Google. Available at: <https://support.google.com/analytics/answer/2637192?hl=en> (Accessed: 01 Nov 2020)

Hessing, T. (2019) ‘Data Sampling Techniques & Uses’, Six Sigma Study Guide. Available at: <https://sixsigmastudyguide.com/data-sampling/> (Accessed: 01 Nov 2020)

Akhouri, A. (2018) ‘Sampling Process’. Available at: <https://www.amritaakhouri.com/single-post/2018/01/23/Sampling-process> (Accessed: 01 Nov 2020)

● Define Decision Tree, Information gain and Entropy.

Decision tree is one of the supervised learning methods used in decision analysis to solve regression problems and mostly classification problems. It works through a set of rules in the form of conditional statements to show and classify the data according to the defined conditions, in order to help in making decision about the target variable.

Information gain is used as a split criterion for decision trees to decide which is the predictor that should be picked up as root for the decision tree at each step by showing which predictor is the most significant.

Entropy is the measure of uncertainty, which is used to calculate the Information gain, and therefore, impacts on how the decision tree chooses to split the data. Thus, we can say:

* greater probability and less information mean more uncertainty, and therefore, less information gain.
* less probability and more information mean less uncertainty, and therefore, greater information gain.

Accordingly, when we have several predictors, how to know which one is the most significant? The one which gives you the biggest information gain, and consequently, the less uncertainty.

References:

Honari, B. (2020) ‘Decision Tree’, Data Mining. Available at: <https://elearning.dbs.ie/pluginfile.php/1122510/mod_resource/content/1/Decision%20Tree%20Lecture%20Notes.pdf> (Accessed: 01 Nov 2020)

Gupta, Prashant (2017) ‘Decision Trees in Machine Learning’, Towards Data Science. Available at: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052> (Accessed: 01 Nov 2020)

Sayad, S. (2020), ‘Decision Tree – Classification’, An Introduction to Data Science. Available at:

<https://www.saedsayad.com/decision_tree.htm#:~:text=The%20information%20gain%20is%20based,%2C%20the%20most%20homogeneous%20branches).&text=The%20result%20is%20the%20Information%20Gain%2C%20or%20decrease%20in%20entropy> (Accessed: 01 Nov 2020)

Sujan, N. (2018) ‘What is Entropy and why Information gain matter in Decision Trees?’, Medium. Available at: <https://medium.com/coinmonks/what-is-entropy-and-why-information-gain-is-matter-4e85d46d2f01> (Accessed: 01 Nov 2020)

Loaiza, S. (2020), ‘Entropy and Information Gain’, Towards Data Science. Available at: <https://towardsdatascience.com/entropy-and-information-gain-b738ca8abd2a> (Accessed: 01 Nov 2020)

Alex (2020) ‘Decision Trees, Entropy, and Information Gain’, BOOSTEDML Articles on Statistics and Machine Learning for Healthcare. Available at: <https://boostedml.com/2020/06/decision-trees-entropy-and-information-gain.html>

(Accessed: 01 Nov 2020)

● Define Chinese restaurant algorithm and Agglomerative Hierarchical Clustering with example.

Chinese restaurant process (CRP) is an unsupervised learning method used for clustering of data. It is a random process comparable to a Chinese Restaurant that has an infinite number of tables (clusters) and the customers walk in and chose to seat at an empty or occupied table accordingly with a set of probabilities applied by the algorithm. The algorithm will calculate the probability to select an empty or occupied table based on the total of customers already at each table (partition).

Hierarchical clustering is one of the clustering techniques which groups objects by their similarity aiming to build a hierarchy of clusters. There are two types of hierarchical clustering: agglomerative and divisive, the agglomerative being the most popular of them. In the agglomerative clustering, each object starts as an individual cluster, the next step is to calculate the proximity between them and then each the object (cluster) is joined to the nearest one until there is only one big cluster or K clusters are created. Calculate the proximity or similarity between each object is an important step to help on the merging process, this can be done using several methods.

References:

Singh, M. (2015) ‘Chinese Restaurant Process’, SlideShare. Available at: <https://www.slideshare.net/MohitdeepSingh/chinese-restaurant-process> (Accessed: 01 Nov 2020)

Stoltzfus, J. (2020) ‘How can the Chinese restaurant process and other similar machine learning models apply to enterprise AI?’, Techopedia. Available at:

<https://www.techopedia.com/how-can-the-chinese-restaurant-process-and-other-similar-machine-learning-models-apply-to-enterprise-ai/7/33191> (Accessed: 01 Nov 2020)

Kaustubh, N. (2017) ‘Chinese Restaurant Process’, Medium. Available at: <https://medium.com/somx-labs/chinese-restaurant-process-1bf036b5bf0f> (Accessed: 01 Nov 2020)

‘HIERARCHICAL CLUSTERING IN R: THE ESSENTIALS’ (2020). Data Novia. Available at: <https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/> (Accessed: 01 Nov 2020)

Manning, C., Raghavan, P. and Schütze, H. (2008) Introduction to Information Retrieval. Cambridge University Press. Available at: <https://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html> (Accessed: 01 Nov 2020)

Patlolla, C. (2018) ‘Understanding the concept of Hierarchical clustering Technique’, Towards Data Science. Available at:

<https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec> (Accessed: 01 Nov 2020)

‘Agglomerative Clustering’ (2020). RapidMiner Studio Core. Available at: <https://docs.rapidminer.com/latest/studio/operators/modeling/segmentation/agglomerative_clustering.html> (Accessed: 01 Nov 2020)

**Q2 (1800- 1900 words)**

Classification

Why Many Americans Don't Vote

<https://github.com/fivethirtyeight/data/tree/master/non-voters>

Regression

Do Americans Support Impeaching Trump?

<https://github.com/fivethirtyeight/data/tree/master/impeachment-polls>

<https://projects.fivethirtyeight.com/impeachment-polls/>

References

Linear regression

<https://widdowquinn.github.io/Teaching-Data-Visualisation/exercises/two_variables/continuous_x_continuous_y.html> (Accessed: 7 Nov 2020)

<https://medium.com/better-programming/data-science-modeling-how-to-use-linear-regression-with-python-fdf6ca5481be> (Accessed: 7 Nov 2020)

<https://towardsdatascience.com/predicting-missing-values-with-python-41e581511117> (Accessed: 7 Nov 2020)

Logistic regression

<https://www.kaggle.com/faressayah/logistic-regression-tutorial-predict-ad-click> (Accessed: 7 Nov 2020)

<https://colab.research.google.com/github/diannekrouse/LRPython/blob/master/LogisticRegression.ipynb#scrollTo=j8qGhoAt-h5-> (Accessed: 7 Nov 2020)

<https://towardsdatascience.com/how-to-do-logistic-regression-in-python-and-r-b7225ea678e8> (Accessed: 7 Nov 2020)

<https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python> (Accessed: 7 Nov 2020)

K Nearest Neighbour

<https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/> (Accessed: 7 Nov 2020)

<https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55> (Accessed: 7 Nov 2020)

<https://towardsdatascience.com/knn-in-python-835643e2fb53> (Accessed: 7 Nov 2020)

<https://medium.com/data-science-analytics/naive-bayes-or-k-nn-for-classification-60a4d92e7bab> (Accessed: 10 Nov 2020)

Others

<https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35> (Accessed: 14 Nov 2020)

<https://www.kaggle.com/shashankasubrahmanya/missing-data-imputation-using-regression> (Accessed: 14 Nov 2020)

Bilogur, (2018). Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547, <https://doi.org/10.21105/joss.00547> (Accessed: 14 Nov 2020)

<https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/> (Accessed: 14 Nov 2020)