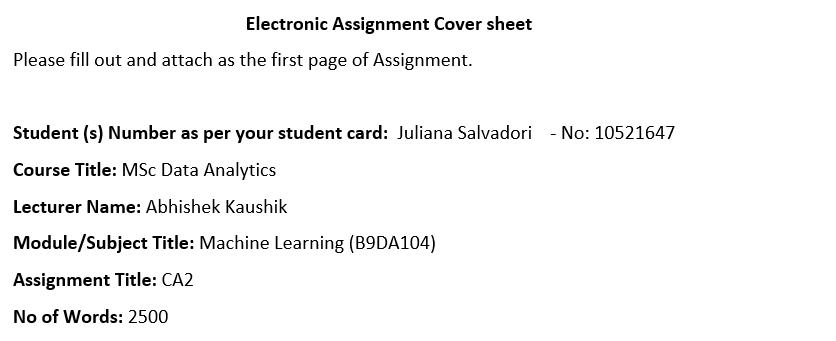
****

**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)**

**Introduction**

In this task I went through two different datasets in order to apply some of the machine learning and data analysis techniques studied on this module.

The first dataset chosen was *Why Many Americans Don't Vote* [1], which has only categorical variables and, therefore, it was suitable to build only the classification model. Thus, I had to select a different dataset *The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women* [11] with the view to build the regression model required.

**Classification**

Classification models are used to predict discrete values, that is, it uses a set of independent variables (that can be categorical or continuous) to the predict a categorical dependent variable [2][3]. Basically, it will classify the data into two or more categories according to the features available to train the model.

There are different approaches for classification and, therefore, different types of algorithms that can be used according to the dataset to be utilised. On this task, I will use K-Nearest Neighbour for classification as it is simple and one of the most used in machine learning for this purpose.

**KNN algorithm**

K-Nearest Neighbour algorithm is one of the predictive supervised learning methods applied in machine learning for classification. It is a simple algorithm and also called as a lazy learner as it actually does not learn from the training dataset and stores a model that will be applied to classify any new data, instead it uses the training set to compare any new data to similar data points and then allocate the new data to the most similar category [6][7][8]. For this reason, it falls into the category of non-parametric models, which do not make strong assumptions, as they are free to form any function form but require more data for training [9] and also will not suit well for real time predictions.

**About the dataset chosen for classification**

*Why Many Americans Don't Vote* is a survey that aims to map the profile of American voters and non-voters and understand what are the main reasons that lead them to choose not to vote. The survey had 8,327 respondents but only nearly 6,000 of them could be used on this analysis by the authors as they were able to match the answer to the respondents voting history and, therefore, the respondents were categorized in three main groups: people who rarely or never vote, people who sometimes vote and those ones who almost always vote [1]. And that categorization is what the classification algorithm built will try to predict based on a subset of the questions asked.

**Classification algorithm implementation**

**Regression**

Regression models are also used to predict values, however, the output for this type of algorithm is numerical (continuous), that is, it uses a set of independent variables (that can be categorical or continuous) to the predict a continuous dependent variable. The idea on this algorithm is to estimate the output (continuous dependent variable) by building the relationship between the input variables in the format of a function y = f(X) [10][3][23].

In the same way as classification models, there are also different approaches for regression and, therefore, different types of algorithms that can be used according to the dataset to be utilised. On this task, I will use Linear regression as it is simple to understand, easy to apply and one of the most used methods for prediction in machine learning.

**Linear regression**

Linear regression algorithm is one of the predictive supervised learning methods applied in machine learning for regression. It is a simple and commonly used algorithm that builds the linear relationship between the dependent variable (response) and one or more independent variables (features) aiming to find the best fit line, that is, minimize the error between predicted and actual values [3][10][17]. Linear regression falls into the category of parametric models which make very strong assumptions as they have a fix set of variables and equations, and they also can have a bias effect. This type of model does not need a huge amount of data for training, but they can nevertheless lead to a poor fit problem [9].

**About the dataset chosen for regression**

*The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women* uses two dataset to observe the association between the importance of women in a film and that film’s budget and gross profits. The BechdelTest.com which contains the list of movies that were analysed and verified if they pass the Bechdel test and The-Numbers.com that contains the financial information on these movies such as box office and budget data [11]. Thus, the regression algorithm built will verify if there is a linear relationship between the movie’s budget (response variable) and the indicator that the movie passed/failed the Bechdel test and its gross profit (international and domestic).

**Classification algorithm implementation**

References

[1] Thomson-DeVeaux, A., Mithani, J., Bronner, L. and Lannes, L. (2020) ‘Why Many Americans Don't Vote’, FiveThirtyEight. Available at: <https://projects.fivethirtyeight.com/non-voters-poll-2020-election/> (Accessed: 03 Nov 2020)

[2] ‘Linear Regression vs Logistic Regression’ (2020), Java T Point. Available at: <https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning> (Accessed: 15 Nov 2020)

[3] ‘Regression vs. Classification in Machine Learning’ (2020). Java T Point. Available at: <https://www.javatpoint.com/regression-vs-classification-in-machine-learning> (Accessed: 15 Nov 2020)

K Nearest Neighbour

[4] Brownlee, J. (2020) ‘Develop k-Nearest Neighbors in Python From Scratch’, Machine Learning Mastery. Available at: <https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/> (Accessed: 07 Nov 2020)

[5] Maklin, C. (2019) ‘K Nearest Neighbor Algorithm In Python’, Towards Data Science. Available at: <https://towardsdatascience.com/k-nearest-neighbor-python-2fccc47d2a55> (Accessed: 07 Nov 2020)

[6] Zoltan, C. (2018) ‘KNN in Python’. Towards Data Science. Available at: <https://towardsdatascience.com/knn-in-python-835643e2fb53> (Accessed: 07 Nov 2020)

[7] Srikanth, P. (2014) ‘Naive Bayes or K-NN for classification?’, Medium. Available at: <https://medium.com/data-science-analytics/naive-bayes-or-k-nn-for-classification-60a4d92e7bab> (Accessed: 10 Nov 2020)

[8] ‘K-Nearest Neighbor(KNN) Algorithm for Machine Learning’ (2020), Java T Point. Available at: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning> (Accessed: 15 Nov 2020)

[9] Kaushik, A. (2020) ‘Introduction to ML’, Dublin Business School. Available at: <https://elearning.dbs.ie/mod/resource/view.php?id=921696>

[10] Hoare, T. (2020) ‘Data Mining: Supervised Learning Classification – Ensemble Models - Regression - Deep Learning’, Dublin Business School. Available at: <https://elearning.dbs.ie/pluginfile.php/1157117/mod_resource/content/0/Revision%20Lecture%201%20-%20Supervised%20Learning%20Classification%20-%20Ensemble%20Models%20-%20Regression%20-%20Deep%20Learning%20%20.pdf>

[11] Hickey, W. (2014) ‘The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women’, FiveThirtyEight. Available at: <https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/> (Accessed: 15 Nov 2020)

Linear regression

[12] Sublett, D. (2019) ‘Data Science Modeling: How to Use Linear Regression with Python’, Medium. Available at: <https://medium.com/better-programming/data-science-modeling-how-to-use-linear-regression-with-python-fdf6ca5481be> (Accessed: 07 Nov 2020)

[13] Pierre, S. (2019) ‘Predicting Missing Values with Python’, Towards Data Science. Available at: <https://towardsdatascience.com/predicting-missing-values-with-python-41e581511117> (Accessed: 07 Nov 2020)

[14] Moreno, A. (2019) ‘Simple and multiple linear regression with Python’, Towards Data Science. Available at: <https://towardsdatascience.com/simple-and-multiple-linear-regression-with-python-c9ab422ec29c?gi=929a4ebc8b98> (Accessed: 15 Nov 2020)

[15] ‘Multivariate Linear Regression in Python with scikit-learn Library’ (2020), Finance Train. Available at: <https://financetrain.com/multivariate-linear-regression-in-python-with-scikit-learn-library/> (Accessed: 15 Nov 2020)

[16] ‘Multiple Linear Regression’ (2020), Java T Point. Available at:

<https://www.javatpoint.com/multiple-linear-regression-in-machine-learning> (Accessed: 15 Nov 2020)

[17] ‘Simple Linear Regression in Machine Learning’ (2020), Java T Point. Available at: <https://www.javatpoint.com/simple-linear-regression-in-machine-learning> (Accessed: 15 Nov 2020)

[18] Loukas, J. (2020) ‘Is your model overfitting? Or maybe underfitting? An example using a neural network’, Towards Data Science. Available at: <https://towardsdatascience.com/is-your-model-overfitting-or-maybe-underfitting-an-example-using-a-neural-network-in-python-4faf155398d2> (Accessed: 18 Nov 2020)

Others

[19] Subrahmanya, S. (2018), ‘Missing Data Imputation using Regression’, Kaggle. Available at: <https://www.kaggle.com/shashankasubrahmanya/missing-data-imputation-using-regression> (Accessed: 14 Nov 2020)

[20] Pierre, S. (2020), ‘Predicting Missing Values with Python’, Towards Data Science. Available at: <https://towardsdatascience.com/predicting-missing-values-with-python-41e581511117> (Accessed: 14 Nov 2020)

[21] Bilogur, A. (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)

[22] Brownlee, J. (2020) ‘How to Use StandardScaler and MinMaxScaler Transforms in Python’, Machine Learning Mastery. Available at: <https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/> (Accessed: 14 Nov 2020)

[23] Nelson, D. (2020) ‘Multiple Linear Regression with Python’, Stack Abuse. Available at: <https://stackabuse.com/multiple-linear-regression-with-python/> (Accessed: 15 Nov 2020)