**Course: Artificial Intelligence and Data Science**

**Module Leader: Mr. Sahan Priyanayana**

**CM2604 Machine Learning Assignment Type: Individual**

**Student Name: W.A.C.I.V Bandara**

**IIT ID-20222453**

**RGU ID-2311689**

**Table of content**

**Corpus preparation**

The UCL repository is a website which contain huge selection of datasets for machine learning tasks. Adult data set was selected form UCL repository and this dataset contain 48842 data records and 15 attributes and indicate that given person Census Income is more than 50K or not.

Data cleaning was one of the pre-processing steps that was carried out. A few of the steps included removing outliers, rejecting duplicate values, and looking for and removing null values.

Random Forest and Naïve Bayes are the algorithms that used for classify weather the salary exceeds 50K or not. After cleaning and preprocessing phase 70% of data is used for training while remaining amount used for testing purposes.

**Solution methodology**

Random forest and Naïve Bayes algorithms are used for classify salary. Summery of two algorithms that used were given below.

**Random forest** - This is a popular and adaptable ensemble learning technique for tasks involving regression and classification. It's a decision tree extension in which several decision trees are trained using various training data subsets and then combined to provide predictions.

* Multiple subsets are created by bootstrapping, which involves sampling random subsets of the training data with replacement.
* There is a decision tree constructed for every data subset. Nevertheless, a random subset of features is chosen for splitting at each node of the tree rather than all of the features. The trees are better decorated because of this unpredictability.
* To arrive at the final forecast, predictions from individual trees are aggregated by voting (for classification) or averaging (for regression).

Random forests are renowned for their high accuracy, scalability to huge datasets, and resilience against overfitting. When working with multi-feature, high-dimensional data, they are especially useful.

**Naïve Bayes** – This algorithm relies on the premise of predictor independence and is a probabilistic classification method that applies the Bayes theorem.

* Using the Bayes theorem, it determines the likelihood of each class given a collection of input features.
* The "naive" assumption of Naive Bayes simplifies probability calculations by assuming that the existence of a given characteristic in a class is independent of the existence of any other feature (i.e., features are analyzed independently).
* The class with the highest probability is then chosen to be the input data's predicted class.

In order to estimate the required parameters, naive Bayes classifiers require very minimal training data and are quick and simple to deploy. However, feature independence may not always hold true, which in some circumstances may result in less than ideal performance.