Abstract No: 211 **ICT, Mathematics and Statistics**

GLOBAL TRENDS IN DATA SCIENCE INDUSTRY

A.G.S. Polgolla[1*] and S.P. Abeysundara[1,2]

[1]Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka
[2]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka
[*]sankhig@gmail.com

The hiring of employees depends on several factors, such as academic qualifications and social factors. Recognising the patterns of these factors of employees is important for companies when recruiting new employees. The main objective of this study was to explore and identify varying clustering patterns among employees around the world. A publicly available dataset on a survey conducted by Kaggle during 2017 – 2020 was considered. As missing values were present in this dataset, a data imputation method for categorical data was used to process a complete dataset. The most appropriate variables for a cluster analysis were selected using the forward selection method. Considering the Gower distance as the dissimilarity measure, the k-medoids algorithm was used to partition the dataset into clusters. Preliminary analysis revealed that most companies prefer to accommodate male employees between 22 – 29 years old, which has been a consistent factor over the years. The propensity to hire individuals with Master's or Doctoral degrees has declined over time, and now the companies tend to hire individuals with Bachelors or professional degrees. In addition, the number of students and the number of Indians/Asians working in the industry has increased dramatically over the years. These results are an indication of companies trying to increase their workforce for a low cost. Eight clusters were identified in cluster analysis for each year separately. Each year, a cluster of 18 – 21 years old individuals, with most of them being males who possess a Bachelor's degree, is present. Also, almost all the employees with a Doctoral degree are 30 – 39 years old or older where many of them are males. Most of the female data scientists with a Master's degree were between 22 – 29 years of age. The 30 – 39 years old employees with a Master's degree were clustered together, with the majority being males, but only in 2019, they were scattered in higher age categories. Most of the data scientists above 40 years were either Bachelor's or Master's or other professional qualification holders. However, in 2020, a considerable number of Doctoral degree holders were added to this cluster. The insights provided by this study would mainly be useful to companies in recruiting Data scientists based on their academic qualifications and demographic characteristics.

Keywords: Academic qualification, Cluster analysis, Data imputation, Gower distance