Upon finishing an undergraduate degree in Mathematics and Statistics and Operations Research, I started working as an Insight Analyst in the gaming and banking industries in Malta. My working experience inspired me to expand my knowledge further and delve deeper into the practical aspects of statistics.

As a result, I decided to embark on a two-year journey to obtain a Masters degree that combined my passion for both statistics and computing and ultimately opted for a degree in Data Science at the University of Dundee. This Masters degree provided me with so much knowledge and proved to be a great introduction to the world of data science.

Several modules were tackled during the degree, ranging from dimensional modelling to big data and of course, machine learning. The degree was meant to take the form of blended learning, in which the lectures of every module take place in person at the university during a week of lectures, while the assignments and personal reading can be done online. Unfortunately, due to Covid-19 restrictions, most lectures were ultimately done online.

The machine learning and data mining modules were particularly interesting as the foundation of these disciplines is statistics and mathematics. Several programming tools such as Python can be used in order to implement different techniques and algorithms including clustering, regression, classification and neural networks. The latter is a relatively newer concept which was explained in detail during one of the modules.

The masters program consisted of a final project for which I chose to look into clustering algorithms by making use of a large data set known as the Hipsci dataset which consists of the stem cell data of different participants. My project involved implementing different clustering techniques to ultimately assist in understanding which protein groups and peptides are correlated with specific diseases. This type of machine learning is known as unsupervised machine learning, where the data is untagged or unlabelled and the goal of the algorithm is to identify patterns in the data. Several clustering methods were implemented on the dataset and the generated results were compared to identify the most optimal algorithm. Further to this, dimensionality reduction was also required in order to decrease the number of variables in the dataset prior to implementing any clustering algorithms.

Apart from a teaching me a lot about data science, and several tools such as Python and R, the past two years also taught me conduct my own independent research, which is an invaluable skill in this day and age. Since the masters was done on a part time basis while I was still working full time, efficient time-management was crucial in order balance both priorities. Furthermore, this course has allowed me to further my statistical and computing skills, thus, additionally enhancing my abilities as an analyst.

The degree disclosed in this publication is funded by the Tertiary Education Scholarships Scheme.