Name of the Researcher: Rakesh Kumar Muraleedharan

Course Name: Master of Science – Data Analytics

College Name: CCT College, Westmoreland St, Dublin 2

Thesis Title:

Predicting the screening colonoscopy numbers across Ireland using machine learning.

Objectives:

1. To identify the variations in colonoscopies for each gender and defined age groups, across different months of a year in order to better plan the colonoscopy capacity in different units using descriptive statistics, furthermore do a hypothesis testing to test if the colonoscopy counts for male population are less compared to females in order to assist the colonoscopy units better.
2. To identify the features that influence the colonoscopy counts, by undertaking appropriate co-relation studies between the features, and furthermore identify if there are any seasonal factors affecting the colonoscopy numbers.
3. To predict the colonoscopy numbers across Ireland by building an appropriate machine learning model, making use of the historical colonoscopy records as the secondary data source, and the eligible population including the extended age group from the latest census as the primary data source.

SAMPLING STRATEGY:

It is proposed to use three different sampling strategies as part of the research, this includes both probabilistic (simple and stratified) sampling, this is primarily used while attaining the objective stated above. Non-probabilistic sampling (judgemental) is done to support the research and get inputs on the factors influencing the research.

This research project proposes to use two separate populations for achieving the above objectives. The first population are the males who have undergone colonoscopy across different age groups. The second population are the females who have undergone colonoscopy across different age groups. Stratified sampling is applied on these two populations and descriptive statistics are applied on these samples. Also, in the sample the population is further grouped for each age group.

The third population that is proposed to be used in the research is the number of eligible clients based on the census 2022 data. *Probabilistic* sampling since is done on this population, since all the clients are equally eligible for colonoscopy. We will merge the census data with the actual colonoscopy numbers and then apply suitable machine learning models on the sample records, we will use 70% of the data to train the model and the remaining 30% to evaluate the model.

Additionally, in order to better understand the dependencies of the data used in analysis a mixture of judgement and convenience sampling is applied. In-depth interview will be done as, detailed in the primary research section below.

PRIMARY RESEARCH METHODOLOGY

This research proposes to use in-depth interview as a primary research methodology. The studies have suggested statistically that men participate less in the Bowel screening programme. Evidence from literature demonstrates, there are impacts of screening participation due to the socio-economic status, age, seasons having high/less colonoscopy counts. In depth understanding of all these factors would be achieved in the in-depth interview.

(Detailed Questions/ Agenda will be shared separately)