National College of Ireland

A machine learning prediction-based analysis for the implementation of general practitioner E-health and Fintech services in Ireland

MSc Research Project
MSc in FinTech

Reena Pillay Rajagopalan
Student ID: 18186807

School of Computing
National College of Ireland

Supervisor: Victor del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Reena Pillay Rajagopalan

Student ID: 18186807

Programme: MSc. in FinTech **Year:** 2019/2020

Module: Research Project

Supervisor: Victor del Rosal

Submission Due Date: First submission - 17th August 2020 & Final submission – 28th September 2020

Project Title: **"A machine learning prediction-based analysis for the implementation of a general practitioner E-health and Fintech services in Ireland."**

Word Count: 6030 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ReenaPillay

Date: 28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A machine learning prediction-based analysis for the implementation of a general practitioner E-health and Fintech services in Ireland

Reena Pillay Rajagopalan
18186807

Abstract

The public health service in Ireland has been facing ongoing pressure in dealing with patient waiting lists. The wait times to see a general practitioner also contributes to the overall waiting list statistics of Ireland. This paper conducts a quantitative analysis of Dublin and County Meath by analysing the population, number of GP available, GP visitation lists, unemployed age groups, smart gadget and broadband users based on recent Census, Data.gov.ie and HSE 2016 data. The purpose of this study is to provide quantitative insights on the existing general practitioner systems and the likelihood of implementing e-health and fintech services. CRISP-DM framework was applied to structure the research methodology and to apply machine learning techniques accordingly. In conducting this study; Correlation, Naïve Bayes, Data Mining and Visualisation tools were selected to gather information on the demographics and the tech savvy environment of both counties. About 12% of the population belonging to 16 to 24 and 25 to 34 age group were unemployed in both counties indicating the higher dependency for public healthcare. The tech savvy analysis shows the population of Dublin has 78% smart gadget and broadband users and Meath has 22%. These quantitative results reveal that while there are many factors contributing to the healthcare, there is a 75% likelihood of E-health and Fintech services being implemented in Dublin and County Meath to achieve a more efficient and effective General Practitioner system.

1 Introduction

Ireland's healthcare system is governed by the Health Act 2004. It offers a free public service funded by the Irish government and costs the country 7% of the Gross Domestic Product (GDP).[1] The Health Service Executive (HSE) manages the public healthcare system and the procedures required to be followed by the residents of Ireland. A medical card system is in place since the 1970s and enables residents; generally, those receiving welfare payments or earn low income to have access to a range of free health services. The latest statistics of Ireland's population is 4.904 million which is double the population since the 1970s.[2] As the

[1] https://www.statista.com/statistics/429208/healthcare-expenditure-as-a-share-of-gdp-in-ireland/
[2] https://data.gov.ie/

population in Ireland increases the Medical Card utilisation increases resulting to a 'Trolley crisis' with over 9,500 patients waiting on hospital trolleys for assessment.[3]

A subsequent system, GP visit Card scheme was also introduced in 2006 which allowed residents to visit their general practitioner (GP) for free which averagely cost between €40 to €60. A referral from the GP is the prerequisite to further their entitlement of the public healthcare service. Although a system and medical cards are available, according to HSE's statistics, there are over 30% of the population have a medical card for public healthcare while the remaining are on a private health scheme.[4] Overcrowding in the healthcare centres and long waiting time to visit the GP is a trendy topic in Ireland. However, there are still gap of understanding the existing GP system affecting the wait time for public healthcare.

This study will conduct a prediction-based analysis on the existing general practitioner system in Dublin and County Meath for the implementation of e-health and Fintech services. The justification of this approach is provided below in section 1.4.

1.1 Background & Motivation

The primary health care in Ireland starts from the GP who provides the initial consultation and necessary referral for the next course of action within the healthcare system. Figure 1 (OECD, 2019) below provides a general flow of the healthcare system in Ireland.

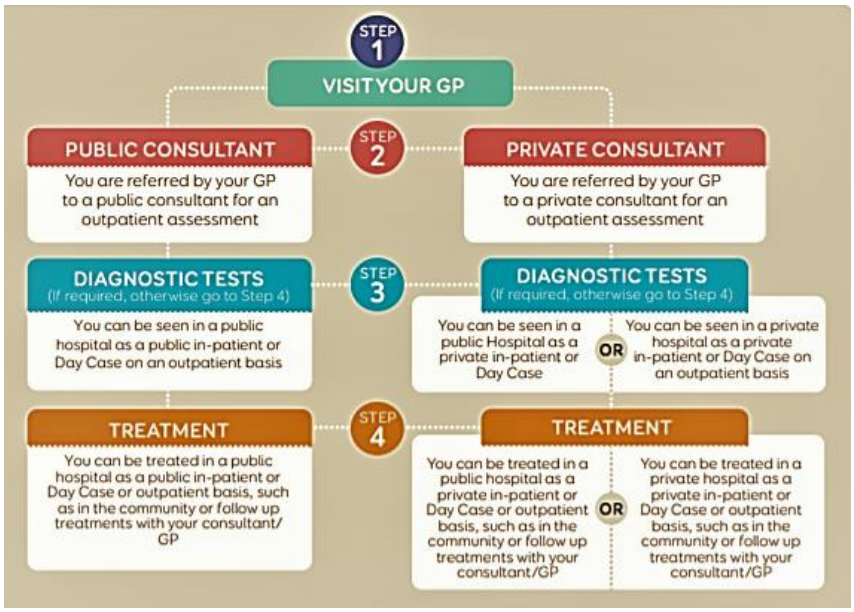

Figure 1: Ireland Healthcare System process

Kilroy and Refaie (2020) research suggested that approximately 31% of Irish residents visited the GP without an appointment while approximately 38% manage to book an appointment on the same day. The authors are of the view point that 69% of the residents are gaining the free GP visit benefits within the public healthcare system. On the flip side, findings from Flynn and Lynam (2020) research indicated that approximately 32% or 1.52

[3] https://www.statista.com/topics/3418/healthcare-system-in-ireland/
[4] https://www.hse.ie/eng/services/find-a-service/eligibility.html

million people are on a public healthcare waiting list. To reiterate the authors viewpoint, the National Association of General Practitioners (NAGP) reported that average wait time for a GP visit was initially 34 hours[5] is now taking up to 1,008 hours (6 weeks).[6]

Although there are existing literatures surrounding a broader approach on the healthcare system in Ireland, there are limitations on the research around the GP system practiced within the counties in Ireland. Every town within the county has several GP allocated for the residents. The GP can operate as a sole trader or among other GPs in health centres. There are 188 GPs within Dublin and 39 GPs within Meath. E-health, a platform governed by HSE, established in 2015 works as a platform for sharing information and strategy among diverse players within the healthcare industry. However, there is no information addressing the GPs manual process. Moreover, after a comprehensive search there were no quantitative analyses focused on GP visitation statistics in the counties aligning to e-health development.

This understanding persuaded the study to utilise the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework during the quantitative analysis. This robust and efficient framework breaks the quantitative analysis into six major phases to generate useful insights. A justification of this framework and machine learning technique is explained in section 1.4.

1.2 Research Question

Following the motivation outlined for this study, the key research question can be coined as; *What is the likelihood of E-health and Fintech services being implemented in Dublin and County Meath to achieve a more efficient and effective General Practitioner system using the CRISP-DM framework?*

1.3 Objectives

- √ The primary aim of this research is to provide a quantitative and investigative insight on the probability of a GP e-health and fintech services being implemented in Dublin and County Meath. This aim supports this study's objective to contribute to existing literature on Ireland's healthcare system. It achieves this by analysing data on Dublin and County Meath's population, number of GP available, GP visitation list, unemployed residents, smart gadget users and broadband users.
- √ In conducting this study, quantitative methodologies are incorporated within the CRISP-DM framework. Correlation, classification, data mining and visualisation machine learning tools will be used to determine fixed or random significant effects from the variables in the data. The aim is to gather accurate insights to contribute to

[5] https://www.irishtimes.com/news/ireland/irish-news/waiting-time-for-doctor-appointments-now-34-hours-1.2405701#:~:text=The%20NAGP%20said%20the%20waiting,three%2Dday%20period%20in%20October.
[6] https://www.independent.ie/irish-news/health/urgent-gp-appointment-could-take-a-week-while-patients-are-facing-six-week-wait-for-regular-visit-warn-doctors-36739895.html

technology experts and information communication technology (ICT) business developments.

√ This study also intends to attain valuable insights for other researcher's future work.

In order to achieve the objectives in alignment to the primary research question, this study will strive to answer the following five secondary questions:

1. Is the population in both counties compatible to the number of GP available?
2. How is the population, smart gadget users and broadband users spread in both counties?
3. What is the age group in both counties that were unemployed and dependent on free GP service?
4. What is the average population of smart gadget users and broadband users per month in both counties?
5. In both counties, do gender have an effect on the usage of smart gadgets?

1.4 Justification

The current pandemic, Covid-19 has affected over 26,000 people in Ireland as per August 2020 statistics. [7] An approximate increase of 26% to the public healthcare waiting list. This triggered a surge in extending the GP waiting period for residents thus only causing a chain effect in obtaining referrals to continue within the public healthcare. Prior to Covid-19, there are no research focused solely on the GP system in Ireland and now with the pandemic, the existing system is raising many red flags of inefficiencies. This concern justifies the importance of this research to provide quantitative insights for the probability of implementing a more efficient and effective e-health and fintech services general practitioner system. As for the machine learning techniques, correlation is used as an indication to identify changes between two variables in the health data. Naïve Bayes calculates the probability assumption using prior probabilities of various data within the health data. Visualisation tool clearly communicate insights from the health data through graphical plots. These machine learning tools supports the CRISP-DM framework built to provide insights of the GP system and demographic of Dublin and County Meath in relation to the topic.

The rest of this paper is structured as follows: Section 2 will detail out some related work that were used to guide this research. Section 3 will describe the research methodology in a quantitative approach using the CRISP-DM framework to answer the secondary questions and evaluate the results. Section 4 will describe the design specification and implementation of the machine learning techniques. Section 5 will be discussions based on personal views of this research. The study concludes in Section 6 including limitations for future work.

[7] https://covid19ireland-geohive.hub.arcgis.com/

2 Related Works

The healthcare system has attracted the need for change with the rise of population, waiting list, global pandemic, Fintech and machine learning techniques. A plethora of research have introduced machine learning techniques on specific medical areas and of recently healthcare blockchain implementation studies. However, there were limited studies focusing on the general practitioners' system and burning issue. The lack of this area of research triggered the motivation of this paper's research question.

A notable example from a paper by Kilroy and Refaie (2020) conducted a qualitative and quantitative study of GPs assessment and management methods when dealing with patients with tinnitus in Ireland. The authors used neural network algorithm to demonstrate the results by integrating the human intuition to the machine learning input. This study utilised the issues addressed and statistical knowledge on the manual GP system in Dublin. Though neural networks were proposed, this study did not find it compatible as there is no qualitative measure nor specialised area of medical expertise used in this study. Flynn and Lynam (2020) presents a novel approach of examining the health and welfare history of Ireland in order to improve the social, political and economic standards. The authors recommended using a Foucauldian historical genealogical model to study the historical problem rather than the period of the problem. This paper was inspired by the authors' framework in examining events as contingencies which helped to outline the objectives and breaking down the business understanding using historical facts.

Ben-Assuli and Vest (2020) utilised the Latent Dirichlet allocation (LDA) model as the primary machine learning technique to extract information related to the research topic surrounding emergency department revisits. The authors carried out forecasting measures and speculative analysis on the future revisits' figures. This study was intrigued to carry out the LDA model surrounding Ireland's healthcare, however the trending topic currently is revolving around Covid-19, politics and e-commerce business. Though the LDA approach did not narrow down towards the research question of this paper, the knowledge and statistical thought process conducted by the authors helped to guide the quantitative structure of this paper. In an earlier paper, Morrissey, et.al (2008) compares the demographic accessibility of residents from urban and rural areas in Ireland. The authors simulated a micro-level healthcare data using simple logit models to derive a probability understanding of the age group owning a medical card. This study utilised the authors research on demographics of Dublin and Meath and included age group analysis as one of the secondary questions.

Another classification technique was proposed by Soleimani-Roozbahani, et.al (2019) that uses Naïve Bayes classifier to make predictions on the accuracy of the authors meta-analysis of published healthcare research for the period of 2008 to 2018. The classifier narrowed down the accuracy to eight main words used in external healthcare research papers and the authors visually presented them. This paper adopted the Naïve Bayes model to due to the fast learning and performance of the model as suggested by the authors. Azadeh-Fard, et.al (2019) referred to the CRISP-DM framework to study in-patient length of stay (LOS) in hospitals. The authors proposed a new framework using the combination of CRISP-DM and

Six Sigma to monitor the LOS data. The CRISP-DM framework methodology approach was opted in this paper to enable a good modelling structure to align with the objectives.

The Naive Bayes model was also one of the approaches taken by Kumar, et.al (2020) to compare the performance with other classification models in predicting malignant and benign breast cancer using global data from UCI repository. This paper analysed the data preparation measures taken when using the Naïve Bayes classifier to ensure a smooth deep learning application. Ogbuabor and La (2018) conducted an analysis using Artificial Neural Networks (ANN) to understand the human activity when using smartphones. The authors applied the confusion matrix testing to monitor the performance of ANN. This study applied the knowledge on the smart phone statistical insights and confusion matrix testing on Naïve Bayes performance.

O'Doherty, et.al, (2020) selected a random number function of 100 patients in two years from 40 general practices in Ireland to investigate the prevalence and management of adults with psychological condition. This study extracted the understanding on demographics, free public health eligibility and policies in Ireland to support this research. Friedemann, et.al (2019) took a fintech investigative and quantitative approach in analysing the use and outcome of one stop clinics on potential cancer diagnosis. Although the authors' research was based around UK, this study applied the fintech knowledge on the data mining techniques and used the database recommended by the authors.

3 Research Methodology

The CRISP-DM framework is used in this section to guide the analysis of this study. This framework was launched by Daimler Chrysler in 1999 to aid all levels of data mining expertise to comprehensively mine data using the six-phase process (Shafique and Qaiser, 2014). Figure 2 shows the breakdown of the six phases to conduct mining methodology accordingly.

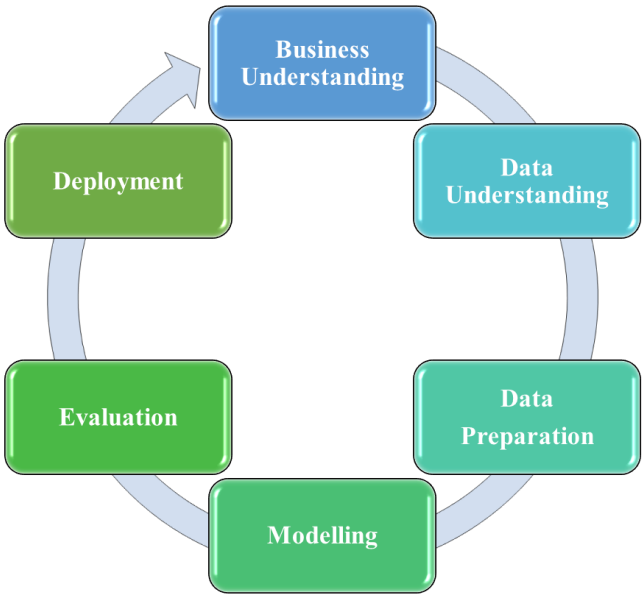

Figure 2: CRISP-DM Framework

3.1 Business Understanding

The business understanding of this framework helps the study to define the problem it wishes to resolve by breaking down into smaller parts. Based on external research conducted, the waiting list is already identified as a problem in Ireland's healthcare. This waiting list occurs from the initial stage, the general practitioners (Cibilis, et.al, 2016). GP waiting list of averagely 34 hours have now increased to 6 weeks.[8]

This research project chose to narrow down the topic to focus on Dublin and County Meath. Potential ICT implementation aligned to the research question will be analysed to gain overall project understanding. Hence this section will be answering the following secondary questions as previously stated in section 1.3:

1. Is the population in both counties compatible to the number of GP available?
2. How is the population, smart gadget users and broadband users spread in both counties?
3. What is the age group in both counties that were highly unemployed and dependent on free GP service?
4. What is the average population of smart gadget users and broadband users per month in both counties?
5. In both counties, do gender have an effect on the usage of smart gadgets?

3.2 Data Understanding

The data understanding of this framework focuses on the quality checking and important information obtain from the data to help narrow down the variable's selection. The data used in this study for analysis was obtained from Central Statistics Office (CSO/Census), Data.gov.ie and HSE. All the data narrows down to Dublin and County Meath and are for the year 2016 to suit the latest Census report. The machine learning software used in this paper is RStudio, a programming language known for statistical, computing and graphical analysis. The Census data shown in figure 3 below consisted of 73 columns 2869 rows.[9] Important columns such as county, total population breaking down into age and gender, smart gadget users, broadband users, GP visits, unemployment rate and broadband users were identified.

[8] https://www.thejournal.ie/hospital-waiting-lists-plans-4536838-Mar2019/
[9] http://census2016.geohive.ie/datasets/population-by-general-health-sex-nuts-3-census-2016-theme-12-3-ireland-2016-cso-osi

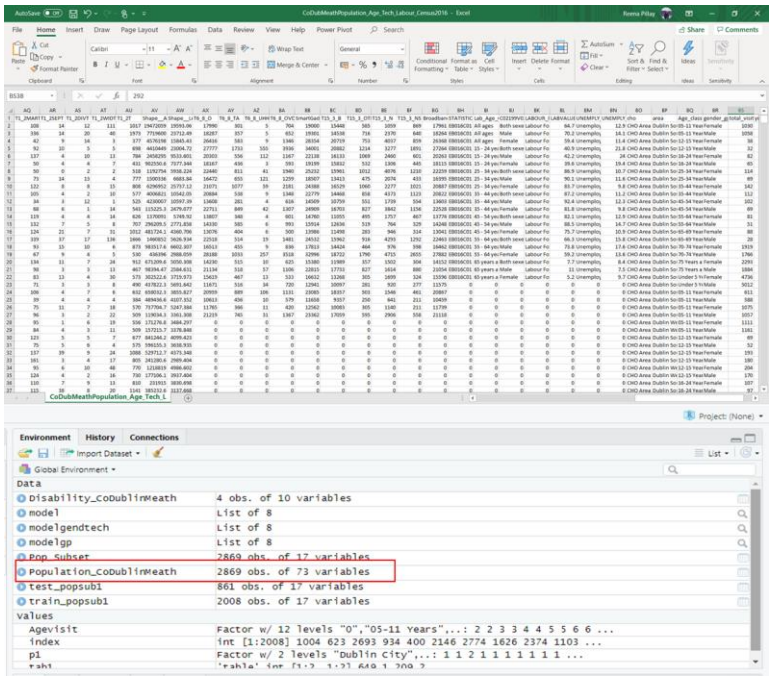

Figure 3: Co. Dublin and Meath population data

The data from Data.gov.ie refers to Census reporting consisted of 60 columns and 254 rows where most columns were weekly numbers and identification of residents with disability and the carer.[10] The important column shown in figure 4 of the data used for this study was the total number of persons with disability consisted of 10 columns and 4 rows.

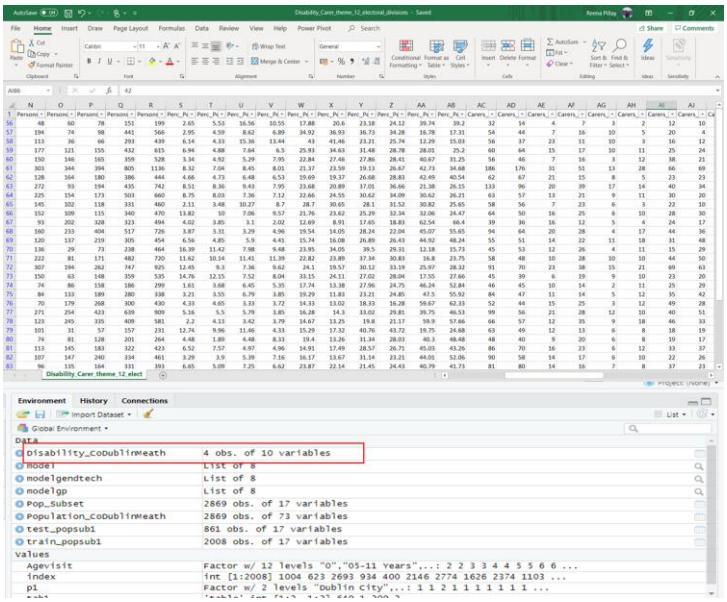

Figure 4: Co. Dublin and Meath disability and carer data

[10] https://data.gov.ie/dataset/disability-and-carers-t12-ed

The data gathered from HSE was the number of GP available in Dublin and County Meath.[11] Both counties have a total of 227 GPs which will be used as part of the comparison understanding when attempting the secondary questions. The data used to conduct machine learning tool does not seem to have empty columns but do contain columns with zero counted as reported numbers.

3.3 Data Preparation

In this phase the framework focuses on checking, cleaning and transforming the new data as the final data set (Shafique and Qaiser, 2014). The two main data from Census and Data.gov.ie was checked for duplication and missing value resulted to zero values as shown in figure 5.

```
> sum(duplicated(Population_CoDublinMeath))
[1] 0
> sum(is.na(Population_CoDublinMeath))
[1] 0
> sum(duplicated(Disability_CoDublinMeath))
[1] 0
> sum(is.na(Disability_CoDublinMeath))
[1] 0
```

Figure 5: Nil duplicated and missing values in datasets

However, upon conducting a glimpse to the dataset, there were superfluous variables identified which were not useful for this study. To ensure these data were superfluous, a sum testing was conducted referring to Census's column descriptions. The sum columns of T6_8_OO, T6_8_TAA, T6_8_UHHH and T6_8_OVDD as well as T15_3_B, T15_3_OTH, T15_3_N, and T15_3_NS totalled up to the SmartGadget_Owners_T6_8_T and Broadband_T15_3_T column accordingly. Based on this understanding and testing, data selection process was done to drop these unnecessary columns and change two of the column names to smoothen the modelling process. The new dataset now contains 17 columns with the same number of rows of 2869 and 'Town' and 'Population' names given to column 3 and 4 as shown in figure 6 below.

[11] https://www2.hse.ie/services/find-a-gp/

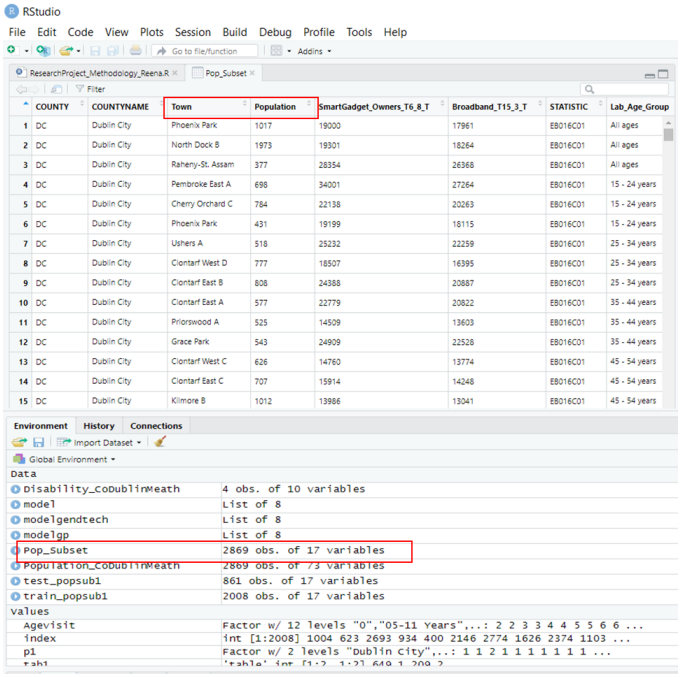

Figure 6: Variable selection and renaming column titles

Henceforth, this study will focus on population, total GP visits, unemployed residents, smart gadget users, broadband users, county, age group, gender and total number of persons with disability variables.

3.4 Modelling

In this section, the selection and application of modelling techniques are identified with suitable parameters for the mining of the data.[12] Upon understanding and preparing the data, the following machine learning techniques are carried out for data analytical testing on the transformed data.

- **Correlation**

According to Roy, et.al, (2016) this technique is useful to detect the relation of two variables and the results will contribute to the decision making of further testing to the more important variable. This study narrowed down to three pairing of important variables from the secondary questions to gain business understanding. A decision was made to test the correlation between population and total GP visits, population and unemployment rate, age and GP visit. The strength of the relationship of these pairings will be examined.

The value of correlation coefficient ranges between -1 and 1 (Roy, et.al, 2016). When the value is -1, it indicates that the paired variables have a strong negative correlation. When it is 0 it indicates that the variables have no relationship to influence the movement of their results. A correlation of 1 shows a strong positive correlation between the paired variables.

[12] https://www.sv-europe.com/crisp-dm-methodology/

There are many measures for correlation testing, but this study chose the 'Spearman' and 'Pearson' methods as both measures have similar characteristics in evaluating linearly related variables that are ordinal or ratios. The formula for both correlation measures further explained below:

Spearman:

$$\rho = \frac{S_{xy}}{S_x\, S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)\cdot\left(\left(R(y_i) - \overline{R(y)}\right)\right.}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}} \quad \textbf{(1)}$$

In the Spearman formula, n represents 2869 as the number of observations in the dataset. The first correlation testing when *xi* is the population of both counties and *yi* is the total GP visits while the second correlation testing, *xi* is age group of both counties and *yi* is the total GP visits.

Pearson:

$$P_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad \textbf{(2)}$$

In the Pearson formula, n represents the same 2869 observations. Correlation testing for when *xi* represents population of both counties and *yi* represents the unemployment value in both counties.

- **Classification**

The classification technique is useful to make categorical analysis and deduce predictions of the dataset to gain business understanding aligned to the research topic (Gupta, et,al, 2011). This study selected to utilise the Naïve Bayes technique as the main machine learning tool for its probabilistic classification ability to learn quickly from a training data. According to Bayes' theorem, this technique is useful to make predictive relationships between two events and their conditional probabilities (Jiang, et.al, 2019). This classifier basically assumes that the absence or the presence of the chosen variable in a class has no relation to other absent or present variables (Jiang, et.al, 2019). This machine learning classifier gathers the consideration that these properties do contribute independently to the probability assumption when conducting the prediction output.

The formula below shows the conditional probability where event C happens given that event A already happened:

Conditional probability:

$$P(C|A) = \frac{P(A \cap C)}{P(A)} \quad \textbf{(3)}$$

To enhance the testing of this probability function, a training data is recommended to gather a clearer understanding. Since this research revolves around a quantitative understanding and discussion, the consideration of training data is applied. This machine learning technique was

chosen to understand resident's tech savvy behaviour in both counties in order to carry out the likelihood of ICT implementation analysis. The age group of unemployed residents and the influence of gender on smart gadget user's data will use this technique to gain deeper insights and assumption of the demographic of both counties as well as the existing technology usage.

This Naïve Bayes technique will be used to classify the smart gadget and broadband usage in Dublin and Meath with reference to the population. The average number of users will be derived to gain a probabilistic tech savvy understanding in both counties. The model will also classify the age group that are highly unemployed using the unemployment values to gain insights of the age group that are more dependent on free GP services. This model also will classify the gender from the GP visits that uses smart gadget in both counties. This is to gain insight on the effect that gender classification has on smart gadget users as well as to make assumption on the type of individuals visiting the GP.

- **Visualization**

According to Villanueva and Chen (2019) refers to Wickham (2016) describing data visualization technique as a powerful tool to represent information and insights from the data in a visual form such as graphs and charts. These visuals are able to demonstrate the message of the analyses aptly to the audience. This study chose the GGplot statistical package to plot the data in order to discuss the relationship of the variables and make statistical assumptions to answer the secondary questions.

3.5 Results and Evaluation

In this section of the framework, the evaluation on the results from the models are conducted. According to Cibilis, et.al (2016), assessment is done on the algorithm results to evaluate if the models provided answers to business understanding and meets the objectives.

This study plotted the variables in the data to understand the selected variables spread in both counties. The total population of both counties based on this 2016 report sums up to 1,526,281 residents shown in figure 7. The plot in figure 8 shows that Dublin is more populated than County Meath indicating that there are more residents living in the capital county in Ireland. According to Census 2016 reporting, 1,331,237 is Dublin's population and 195,044 is Meath's population. Figure 9 results show the smart gadget users in both counties and Dublin having more smart gadget users with 530,753 residents while Meath with 70,649 residents summing up to 601,402 residents. Figure 10 shows the plot of broadband users in both counties, Dublin has more with 479,159 broadband users while Meath with 63,861 summing up to a total of 543,020 residents.

```
> sum(Pop_Subset$Population)
[1] 1526281
> sum(Pop_Subset$SmartGadget_Owners_T6_8_T)
[1] 601402
> sum(Pop_Subset$Broadband_T15_3_T)
[1] 543020
```

Figure 7: Sum of selected variables

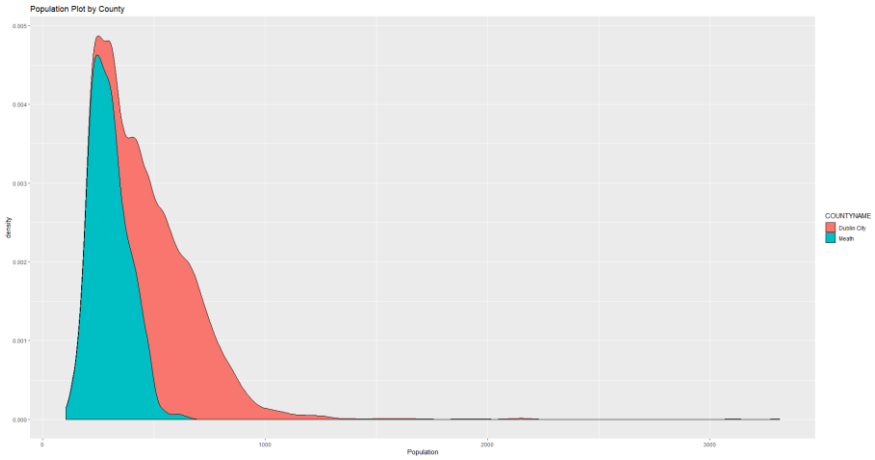

Figure 8: Population spread in Co. Dublin and Meath

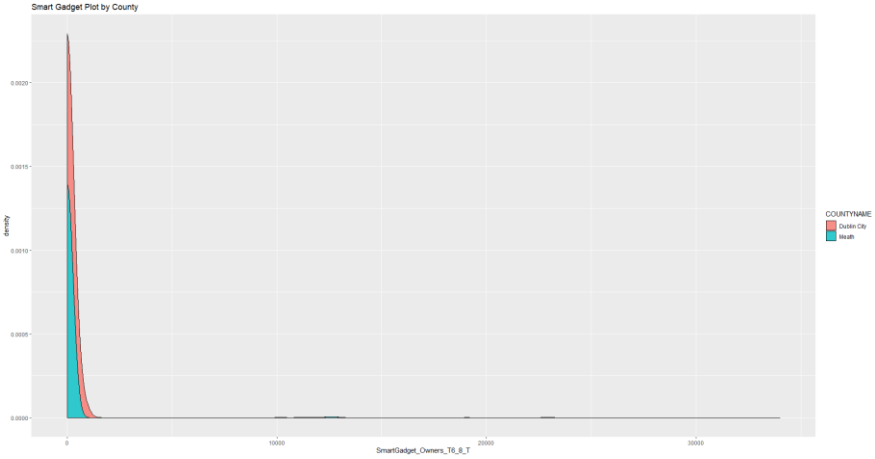

Figure 9: Smart Gadget users spread in Co. Dublin and Meath

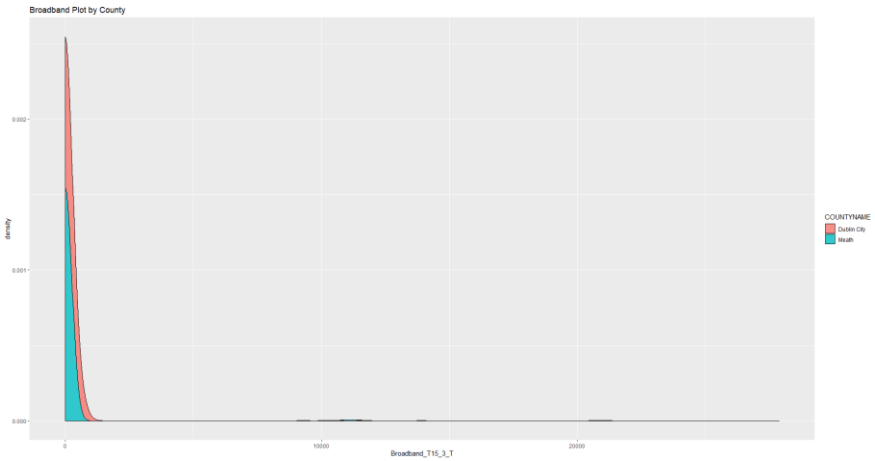

Figure 10: Broadband users spread in Co. Dublin and Meath

The results from figure 9 and 10 reflects on figure 8 can be assumed due to Dublin being more populated. The analysis of this plot answers the second question from the secondary questions. Ireland statistics indicated that 39 GPs were available in Meath[13] and 118 in Dublin.[14] The European standards practiced by Irish College of General Practitioner (ICGP), the number of patient visits by a GP should be averagely 1,000 per week.[15] Taking this calculation of having a standard 40-hour weekly work shift, GP would be expected to consult approximately 25 patients per hour (2.4 minutes per patient). These average weekly numbers were challenging targets to be achieved by most towns in Dublin averaging between 442 to 665 patients and relying fully on human labour through the end to end process.[16] Taking the midpoint of 442 and 665 (554 patients), a GP is consulting approximately 14 patients per hour (4.3 minutes per patient). In terms of Meath, the population of 195,044 is serviced by 39 GPs which is 5,000 patients per GP. According to the Irish Examiner, Irish people visit their GP on average 3 times per year.[17] Taking a 40-hour work week, this equates to 7 patients per hour (8.3 minutes per patient) for GPs in Meath. The current pandemic affects the statistics where GPs not only needing to resolve waiting time issues but need to also give priority to Covid-19 related symptom patients (Kelly, 2020). The increase in population causes a chain effect on GP working hours, average time with each patient and the health of residents including the GP (Lonergan, et.al, 2020). Thus, to answer the first question from the secondary questions, this can only derive that the populations of Dublin and Meath are not compatible to the number of GPs available as the average time spent for a GP consultation is 14.1 minutes.[18] This means that Dublin would require approximately 205 number of GP and approximately 55 number of GP for Meath without considering the population increase. Calculation breakdown shown below in figure 11.

[13] https://www.irelandstats.com/gp-in-meath-county/
[14] https://www.ehealthireland.ie/Strategic-Programmes/Open-Data-for-Health/
[15] https://www.icgp.ie/
[16] https://www.thejournal.ie/gp-breakdown-country-3971888-Apr2018/
[17] https://www.irishexaminer.com/news/arid-30832240.html
[18] https://www.irishexaminer.com/news/arid-30973434.html

Irish Examiners (IE) Survey	
Average time for 1 patient	14.1
Hrs per week	40.0
Mins per week	2,400.0
Patients per week	170
Calculation estimation using the above	
GPs in Dublin	118
Dublin Population	1,331,237
Patients per GP	11,282
Visits per year estimation	3
Total patients per GP	33,845
Visits per week per GP	651
Additional patients compared to IE report	481
Percentage of these add on patients	74%
Actual number of GPs in Dublin needed if following IE average time (estimation)	**205**
GPs in Meath	39
Dublin Population	195,044
Patients per GP	5001
Visits per year estimation	3
Total patients per GP	15003
Visits per week per GP	289
Additional patients compared to IE report	118
Percentage of these add on patients	41%
Actual number of GPs in Meath needed if following IE average time (estimation)	**55**

Figure 11: Calculation breakdown to estimate actual GPs needed

In testing the correlation between population and total visits using the Spearman method, it resulted to RHO of 0.27 shown in figure 12. This indicates that there is a positive correlation between these two variables, but it is weak and would not affect the analysis of the study that much.

```
        Spearman's rank correlation rho

data:  Pop_Subset$Population and Pop_Subset$total_visit
s = 2889722860, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2657975
```

Figure 12: Correlation results of Population and Total GP visits

Figure 13 shows the correlation testing between population and unemployment rate using the Pearson method resulted to RHO of 0.015. This also indicates that both variables are positively correlated but is extremely weak and should be independently evaluated.

```
        Pearson's product-moment correlation

data:  Pop_Subset$Population and Pop_Subset$UNEMPLY_VALUEPERC
t = 0.80597, df = 2867, p-value = 0.4203
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02155570  0.05161678
sample estimates:
       cor
0.01505069
```

Figure 13: Correlation results of Population and Unemployment rate

Figure 14 shows the correlation testing between the age groups that visit the GP and total visits of GP. This resulted to RHO of 0.72 where both variables are positively correlated and are influenced by any movement or changes. These correlation understandings supported the study to conduct further analysis on the dataset.

```
        Spearman's rank correlation rho

data:  Pop_Subset$Age_classification_Gpvisit and Pop_Subset$total_visit
S = 1101861657, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.720046
```

Figure 14: Correlation results of Age group visiting the GP and Total GP visits

To conduct Naïve Bayes, the dataset was split into a 70% training and 30% testing data to better understand the characteristics of the data. Upon testing the combination of the county name, population, smart gadget and broadband users it resulted in Dublin having a more populated smart gadget and broadband user base of 78.4% and Meath of 21.6% as shown in figure 15. The Naïve Bayes also detailed out the results that the average number of smart gadget users in Dublin is 17,001 per year and Meath with 6,382 per year. Broadband users in Dublin is 13,632 and Meath is 5,807 per year. To answer the fourth question from the secondary questions, this result helps the study to assume that Dublin has a higher average monthly user of smart gadget of 1,417 and broadband of 1,136. Meath has an average monthly user of 532 for smart gadgets and 484 for broadband.

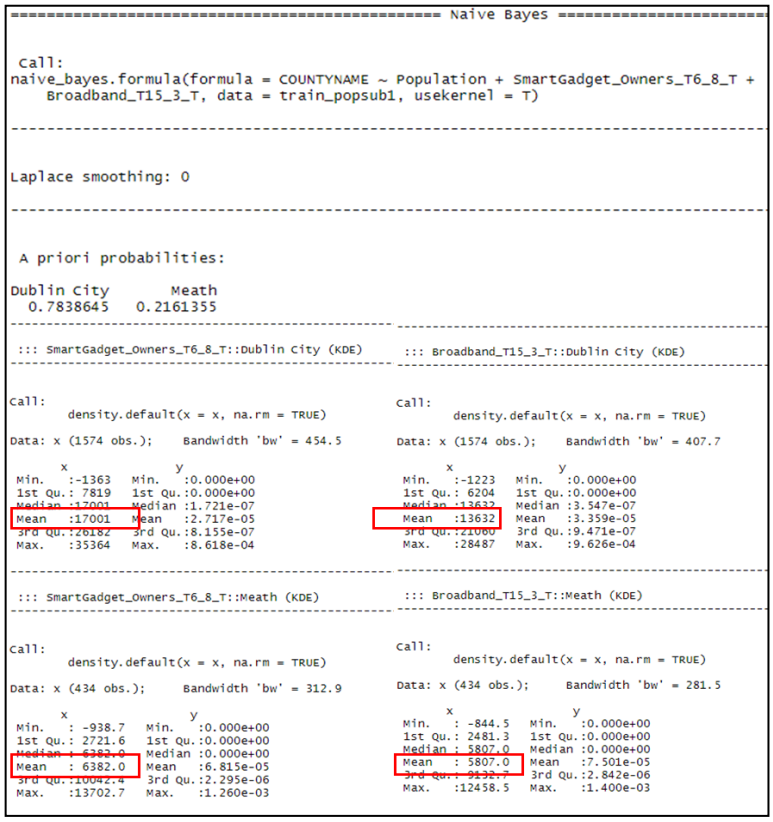

Figure 15: NB of Smart Gadget and Broadband users against Population

The model was also applied on age group column which was a similar classification for GP visit against unemployment value based on county. To answer the third question from the secondary question, the model resulted in a higher percentage of between 16 to 24 and 45 to 54 years age groups that were unemployed in Dublin while in Meath, the age group of 25 to 34 and 65 to 69 were unemployed as shown in figure 16. The retirement age in Ireland is on average at 65.[19] This study can assume that while the free public healthcare is for all residents, there are a higher percentage of unemployed residents from those age groups that are much more dependent on the free GP services.

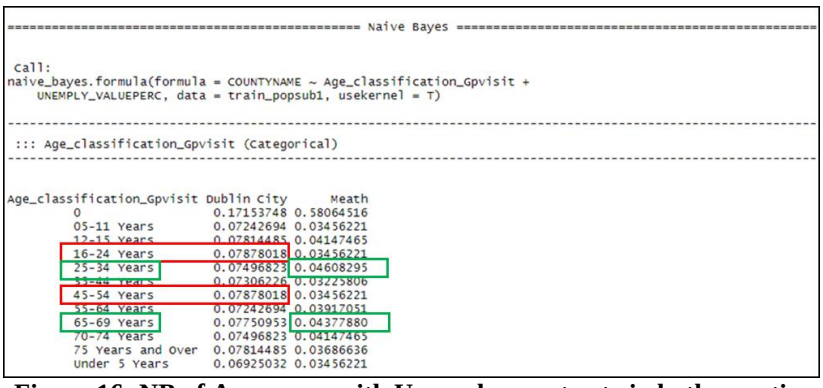

Figure 16: NB of Age group with Unemployment rate in both counties

The model then tested the gender using the same classification from the GP visits against smart gadget users based on county. The results in figure 17 did indicate that smart

[19] https://www.citizensinformation.ie/en/employment/retirement/

gadget users based on gender classification were higher in Dublin than Meath. There was not a big difference in percentage between the genders, but male was slightly dominating the smart gadget user's category by only 0.005% in both counties. To answer the fifth secondary question, this study has evaluated from this data that gender does have an effect to the smart gadget user category as it was able to split the usages into percentage by gender class.

```
============================================= Naive Bayes =============================================


 Call:
naive_bayes.formula(formula = COUNTYNAME ~ gender_gpvisit + SmartGadget_Owners_T6_8_T,
    data = train_popsub1, usekernel = T)

-------------------------------------------------------------------------------------------------------

gender_gpvisit Dublin City     Meath
        0       0.1715375 0.5806452
        Female  0.4116900 0.2073733
        Male    0.4167726 0.2119816
```

Figure 17: NB of Gender class with Smart gadget users in both counties

A confusion matrix was used to check the performance of Naïve Bayes model for this dataset using the testing data of 30% and it resulted in 75% accuracy with a low kappa of 12%. This could be due to the split of testing data containing more zero values. This research objective was to analyse and gain as much insights from the existing data in which the model was able to do so.

The disability data contained the number of residents that have disabilities was plot using the GGplot visualisation tool shown in figure 18 below.

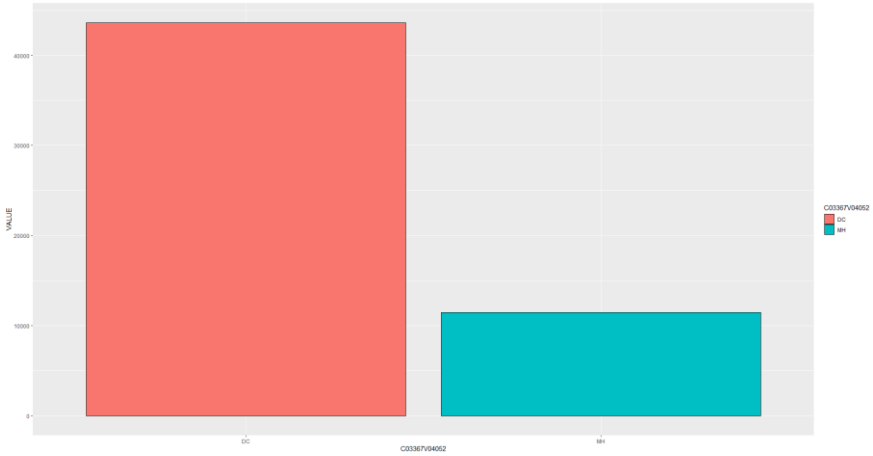

Figure 18: GGplot of Residents with Disability in both counties

Dublin has more residents with disability of 81,502 with 37,883 being males and 43,619 being females. Meath has a total of 22,608 residents with disability where 11,150 being males and 11,458 being females. This information may be useful to the selection and user interface of ICT implementation to suit the demographic of the population in both counties.

In summary, the table below of figure 19 reflects the overall results obtained from machine learning techniques above.

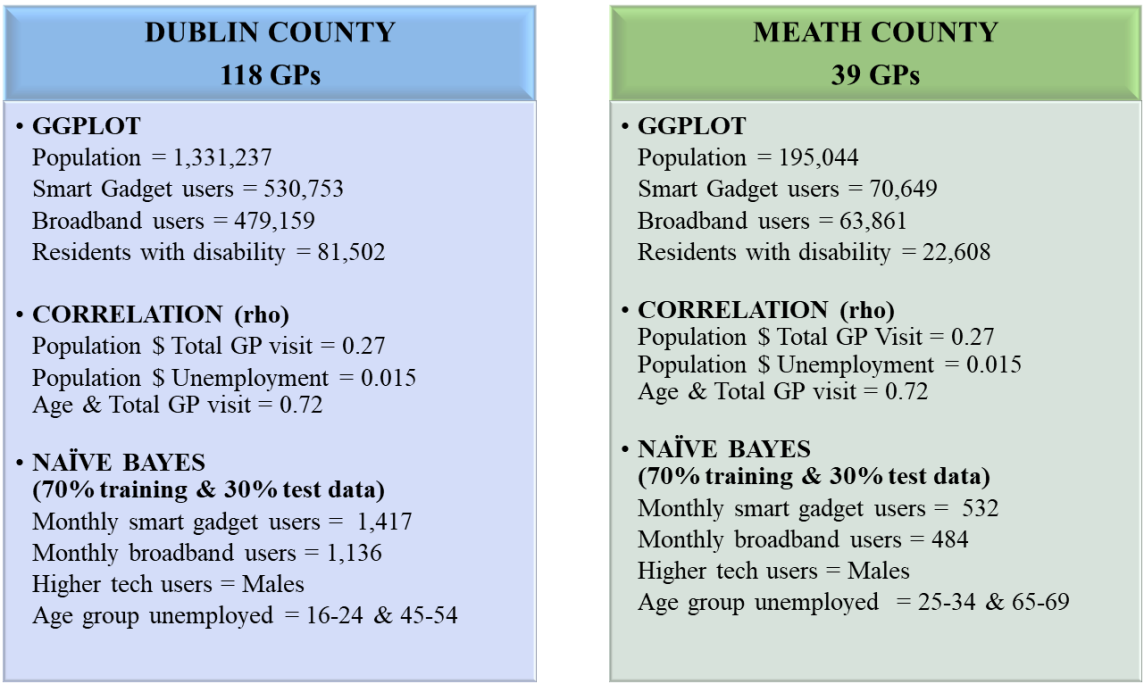

Figure 19: Summary of Modelling results

CRISP-DM framework in this section was helpful to guide the quantitative analysis to deploy a good business understanding from the data. The machine learning techniques were able to answer the secondary questions providing insights on the existing GP system, demographic of both counties and the tech savvy characteristics of the residents.

4 Design Specification and Implementation

This section provides an overview of the algorithm implemented for the purpose of this study. RStudio an integrated programming software that uses R programming language to conduct statistical, computing and graphical analysis. The operating system that RStudio used for this research is Windows. This programming software supports Naïve Bayes classifier which is the main algorithm used for this study. According to Majka, the naïve bayes package is maintained based on three principles which are efficient, user friendly and written in Base R.[20]

This classifier uses the Bayes Theorem where it uses an algorithm to make a prediction of the group of probabilities for each class based on a given data point for each class (Saritas and Yasar, 2019). The class with the highest probability also known as the Maximum A Posteriori (MAP) is taken as the likelihood of the class (Jiang, et.al, 2019). Based on the MAP understanding, this study implemented the classifier to test out the selected variables from Dublin and County Meath population and health data which initial had an assumption that all features in the dataset are unrelated to each other. Similar to the logic of flipping the coin, the Bayes Rule uses the algorithm choosing the probability of that selected variable in the training dataset, P(X|Y) to find the probability of the training dataset from that selected

[20] https://majkamichal.github.io/naivebayes/

variable, P(Y|X) (Saritas and Yasar, 2019). This classifier algorithm was implemented to derive insights for the third, fourth and fifth secondary questions.

5 Discussions

On review of the findings in this paper, it demonstrates that Dublin, the highest populated county has higher smart gadget and broadband users with millennials being the primary age group. The County Meath analysis has a higher age bracket in regard to being tech savvy. This suggests that Ireland already has an existing tech savvy population. These quantitative insights were supported with the fact that Ireland has the fastest growing-tech worker population since 2016 hence earning the global hub of the technology ecosystem recognition.[21] The gender classification testing resulted to males being slightly more of tech savvy user based on the analysis on the smart gadget users. However, the difference in the percentage was very minimal that it was challenging for the study to derive any assumption. On the contrary, Accenture reported that the gender composition of Ireland corporate boards with technology experience are mostly females of 16% when compared with males of only 9%.[22] This variable was selected to best tested in the hope to contribute insights to the type of user interface development when implementing e-health and fintech services.

However, while this analysis gathered useful technology information, there were limitations in identifying a significant correlation between the variables particularly the significance on the number of residents visiting the GP in relation to the population. This could be implied that the size of the data was not big enough since it was focused on a year load of data of two counties only. The accuracy of the data was reflected as 75% though the kappa value was low, but the analysis was aligning to the objective of this paper. The accuracy of the results could have been improved by obtaining data on residents' online activities, fintech businesses and healthcare start-ups. This would have provided a stronger discussion and deeper insights in terms of understanding the tech savvy ecosystem and the comforts of residents to these digital changes.

6 Conclusion and Future Work

This research conducted a quantitative analysis of Dublin and County Meath data on population, number of GP available, GP visitation list, unemployed residents, smart gadget users and broadband users. The purpose of this analysis was to gather insights on the likelihood of e-health and fintech services being implemented in both counties to achieve a more efficient and effective General Practitioner using CRISP-DM framework. The CRISP-DM framework was useful as a reference point when applying correlation, Naïve Bayes and visualisation techniques accordingly. This resulted in quantitative insights of the population in both counties was not compatible to the number of GP available. Dublin has the highest

[21] http://www.comit.ie/latest-news-from-comit/how-ireland-has-become-the-leading-tech-hub-of-europe-.776.html

[22] https://irishtechnews.ie/female-board-members-far-more-likely-than-male-board-members-to-have-professional-technology-experience-accenture-research-finds/

spread of population, smart gadget users and broadband users. The highly unemployed age groups were 16 to 24 and 45 to 54 in Dublin while they were 25 to 34 and 65 to 69 in Meath. The study assumed that these unemployed residents were more dependant on the free public healthcare system and are part of the waiting list. On average there are 1,949 smart gadget users and 1,620 broadband users who are mostly males in both counties. There were 3 key objectives in conducting this paper as outlined in section 1.3. Based on the results and external research, it can be concluded that all objectives were achieved and there is a 75% likelihood of E-health and Fintech services being implemented in Dublin and County Meath to achieve a more efficient and effective General Practitioner system.

Although this research provided quantitative analysis and machine learning applications, there is still room for improvement. For future analysis, several other data sources can be included to provide more insights of Ireland's general practitioner system and public healthcare. The duration of the data could be extended, and qualitative approach could be included to gain a holistic understanding of the healthcare waiting list issue. Furthermore, a more sophisticated machine learning technique could be applied such as K-means clustering model and artificial neural networks for both qualitative and quantitative data to improve the overall results of this research.

References

Azadeh-Fard, N., Megahed, F.M., and Pakdil, F. (2019) 'Variations of length of stay: a case study using control charts in the CRISP-DM framework', *International Journal of Six Sigma and Competitive Advantage*, 11, pp. 204.

Ben-Assuli, O., and Vest, J. R. (2020) 'Data mining techniques utilizing latent class models to evaluate emergency department revisits', *Journal of biomedical informatics*, pp. 101.

Ciblis, A. S., Butler, M.L., Quinn, C., Clare, L., Bokde, A.L.W., Mullins, P.G., and McNutty, J.P. (2016) 'Current Practice in the Referral of Individuals with Suspected Dementia for Neuroimaging by General Practitioners in Ireland and Wales', *PLoS ONE*, 11(3), pp. 1-13.

Flynn, A. V. and Lynam, J. M. (2020) 'Using a historical genealogical approach to examine Ireland's health care system', *Nursing Inquiry*, 27(1), pp. 1-9.

Friedemann, S.C., Tompson, A., Holtman, G.A., Bankhead, C., Gleeson, F., Lasserson, D. and Nicholson, B.D. (2019) 'General practitioner referrals to one-stop clinics for symptoms that could be indicative of cancer: a systematic review of use and clinical outcomes', *Family practice*, 36(3), pp.255-261.

Gupta, S., Kumar, D. and Sharma, A. (2011) 'Performance analysis of various data mining classification techniques on healthcare data', *International journal of computer science & Information Technology* (IJCSIT), 3(4), pp.155-169.

Jiang, L., Zhang, L., Yu, L. and Wang, D. (2019) 'Class-specific attribute weighted naive Bayes', *Pattern recognition 88*, pp.321-330.

Kelly, B.D. (2020) 'Emergency mental health legislation in response to the Covid-19 (Coronavirus) pandemic in Ireland: Urgency, necessity and proportionality', *International Journal of Law and Psychiatry*, p.101564.

Kilroy, N. and El Refaie, A. (2020) 'Tinnitus management in Ireland: a pilot study of general practitioners', *Irish journal of medical science*, pp. 1-11.

Kumar, V., Mishra, B.K., Mazzara, M., Thanh, D.N. and Verma, A. (2020) 'Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications', *In Advances in Data Science and Management*, Springer, Singapore, pp. 435-442.

Lonergan, P.E., Logan, J., Diver, S., Nugent, C., Hegarty, I., Plunkett, O., Mealy, K., Hyland, J.M., McNamara, D.A. and Rogers, E. (2020) 'Does clinical validation and the implementation of new models of outpatient service delivery have the potential to reduce waiting lists? - a pilot study in Letterkenny University Hospital', *Irish Journal of Medical Science*, pp.1-6.

Morrissey, K., Clarke, G., Ballas, D., Hynes, S. and O'Donoghue, C. (2008) 'Examining access to GP services in rural Ireland using microsimulation analysis', *Area*, 40(3), pp.354-364.

O'Doherty, J., Hannigan, A., Hickey, L., Meagher, D., Cullen, W., O'Connor, R., and O'Regan, A. (2020) 'The prevalence and treatment of mental health conditions documented in general practice in Ireland', *Irish journal of psychological medicine*, 37(1), pp.24-31.

OECD-European Observatory on Health Systems and Policies (2019), 'Ireland: Country Health Profile 2019, State of Health in the EU', *OECD Publishing Paris/European Observatory on Health Systems and Policies*.

Ogbuabor, G. and La, R. (2018) 'Human activity recognition for healthcare using smartphones', *In Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 41-46.

Roy, S., Mondal, S., Ekbal, A. and Desarkar, M.S. (2016) 'CRDT: Correlation Ratio Based Decision Tree Model for Healthcare Data Mining', *IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 36-43.

Saritas, M.M. and Yasar, A. (2019) 'Performance analysis of ANN and Naive Bayes classification algorithm for data classification', *International Journal of Intelligent Systems and Applications in Engineering*, 7(2), pp.88-91.

Shafique, U. and Qaiser, H. (2014) 'A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)', *International Journal of Innovation and Scientific Research*, 12(1), pp.217-222.

Soleimani-Roozbahani, F., Ghatari, A.R. and Radfar, R. (2019) 'Knowledge discovery from a more than a decade studies on healthcare Big Data systems: a scientometrics study', *Journal of Big Data*, 6(1), pp.1-15.

Villanueva, R.A.M. and Chen, Z.J. (2019) 'ggplot: Elegant graphics for data analysis measurements', *Interdisciplinary Research and Perspectives*, 17 (3), 160-167.

Wickham, H., 2016. 'ggplot2: elegant graphics for data analysis', *Journal of Statistical Software, 35(1).*