National College of Ireland

The Significance of External Factors on an Individual's Health Determination

MSc Research Project
MSc. in Data Analytics

Richard Burke
Student ID: 15034097

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Richard Burke		
Student ID:	15034097		
Programme:	MSc in Data Analytics	**Year:**	2020
Module:	Research Project		
Supervisor:	Dr. Catherine Mulwa		
Submission Due Date:	17/08/2020		
Project Title:	The Significance of External Factors on an Individual's Health Determination		
Word Count:	6093	**Page Count**	21

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:	Richard Burke
Date:	16/08/2020

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):	

The Significance of External Factors on an Individual's Health Determination

Abstract

Determining health outcomes in the general population is critical to the appropriate development of public policy. Identifying the key drivers of public health outcomes is crucial to effective understanding and implementation of these policies. This research explores the key socioeconomic and locality-based explanatory factors that influence the self-determined health outcomes of the residents in an area. Furthermore, it demonstrates the necessity of quality inputs and technical expertise into the development of public policy. Data was captured from the Ireland Central Statistics Office, Ireland's open data portal, and the Ireland deprivation index. The data was processed to generate health classifications by Small Area in Fingal, County Dublin. Five classification models (Gradient Boosting, Random Forest, Naïve Bayes, Multinomial Regression, and AutoML's automatic model) were trained on sample data to predict the health outcome for the remaining holdout test set. The results demonstrate the capacity to accurately predict the aggregated health determination of a small area from publicly available data. Multinomial Regression was the best performing model with an accuracy scored of 49% compared to a no-information rate of 20%. Furthermore, it uncovers the interrelationship between the availability of local facilities and self-health determination, with the areas identified as having poorer health outcomes having greater access to the facility factors incorporated into this research. On this basis, it is recommended that policymakers ensure data capture and model development is a key part of their policy decision process. Further research is required to identify additional factors that could strengthen the effectiveness of these models and to complete comprehensive model optimisation.

1 Introduction

Living in an increasingly divided society with ever-widening gaps in wealth and prosperity. Globally, healthcare has become one of the leading markets and is expected to be valued at USD $10.5 trillion by 2022[1].

As healthcare becomes increasingly expensive day by day, the health and wellbeing of those who can ill-afford the increased cost are disproportionally impacted. It can be argued that one can measure the success of a society on the ability of those who are less well-off to create a healthy and balanced lifestyle. This becomes an increasingly critical problem as the world enters another period of global instability, and questions arise about how to inform public policy best moving forward. This research focuses on the identification of critical factors that impact one's self-assessed health, doing so through the categorisation and prediction of an area's health outcome from the aggregation of publicly available data points available to public policymakers.

[1] https://www2.deloitte.com/ie/en/pages/life-sciences-and-healthcare/articles/global-health-care-sector-outlook.html

1.1 Motivation and Background of Research

The application of research into public health is critical in effective planning for society. It is estimated that Ireland spent €21.1 billion on health in 2017[2]. This represents 11.7% of Ireland's Gross National Index (GNI). Effective planning of resources would enable minimisation of this cost through the promotion of healthy living and lifestyles at increasing levels of effectiveness. In order to undertake this, it is imperative to first understand the drivers of health and wellbeing.

Since the 2008 crash, Ireland has gone through a significant period of growth before ultimately reaching near-full employment in late 2019[3]. However, the initial effects of the global Covid-19 pandemic were truly felt in Q1 2020 and immediately changed this upward trajectory and public landscape, creating a greater need than ever for active public policy development and implementation. The challenge faced by many governments in striving to create equality in outcomes and recoveries will lead to a raft of public policy initiatives. Ensuring decisions are appropriately informed and based on robust research will drive great long-term benefits and positive outcomes for society. Key to creating these outputs and outcomes are enhanced data capture, integration, and model developments.

The general public tends to downplay material, structural, and environmental causes of poor health (Blaxter, 1997), and this informs many societal choices with health outcomes given insufficient consideration in strategic decisions. Identification and incorporation of the key factors that drive health outcomes will increase awareness and understanding of the implications of specific policy implementation. However, as stated by Hassani, Saporta, and Silva (2014), National Statistics Institutes responsible for data collection rarely use Machine Learning (ML) or novel techniques to analyse their data sets and instead rely on more traditional methods. Increasing complexity in data capture and data modelling has the potential to drive more significant health outcomes.

The primary motivation for this research is to demonstrate the value of greater data capture to deliver more comprehensive data sets and display its potential for data model development/integration. Identifying areas prone to poorer/greater health outcomes will deliver an understanding of the key drivers and offer potential solutions to assist in the generation of public policy. Current identification of critical factors is limited by data availability and inefficient data capture.

[2] https://www.cso.ie/en/statistics/statisticalyearbookofireland/
[3] https://www.cso.ie/en/releasesandpublications/er/lfs/labourforcesurveylfsquarter32019/

1.2 Research Question

RQ**:** "*Using existing available factors, can self-assessed public health outcomes be accurately derived and predicted using machine learning models (Gradient Boosting, Random Forest, Naïve Bayes, Multinomial Regression, and AutoML's automatic model)?"*

Accurate prediction of areas with positive/adverse health outcomes will help drive targeted public policy to ensure those with the greatest need will be recognised and receive the required level of support. In this analysis, the outcome will involve the inspection of available factors that measure the health and prosperity of a population before establishing what impact socioeconomic, demographic, geographic, and various other amenity factors have on the health of the general population. Increased research must be performed to examine the effect that these geographic and amenity factors have on the health outcomes of a population. Successful application of the input variables to demonstrate key drivers of health outcomes will inform the debate on public policy and expenditure and exhibit key factors that prompt the separation of areas into different health categories.

Sub RQ**:** *"Which currently available factors are significant in determining the health outcome of an area?"*

A secondary element of this analysis will involve the application of various methods and models to test the validity of these factors to predict public health outcomes and improve their performance anticipating the overall health status of the community. Model variation will be crucial to the successful implementation of future policy, and more research specific to public health is recommended. Advancements in the processing of available data points and the utilisation of emerging infrastructure are critical to the enduring success of policy implementation.

1.3 Research Objectives

Objective 1: *Investigate the state-of-the-art surrounding health determination and public health prediction.*
Objective 2: *Design a technical framework to underpin the implementation of research into public health determination and prediction.*
Objective 3: *Implement a solution that drives an understanding of public health determination and prediction using 5 ML models.*
Objective 4: *Evaluate, interpret, and critique the results and output of the various ML model implementations.*

The report is structured to meet the objectives outlined above. Chapter 2 introduces an investigation of the existing literature in the determination of public health outcomes and the factors that drive these outcomes. Chapter 3 presents the research methodology and design principles underpinning the implementation outlined in chapter 4. Chapter 5 evaluates the outputs and presents a critical analysis of the results.

2 Literature Review

Highlighting the existing theory on health reporting and the various deterministic factors that impact the response will guide this study. There are three main factors within the existent literature that will be considered and explored:

1. Socioeconomic and demographic
2. Local services and amenities
3. State-of-the-art architecture

The interaction of these variables, along with their direct impact on the response, will be explored in the context of the existing research on data analytics and health determination.

2.1 Socioeconomic and Demographic Factors

Existing theory in relation to socioeconomic factors and health determination is prevalent but with limited qualitative studies across a general population that deliver quantitative data. The consensus is that lower socioeconomic status is a direct driver of poor health. In studies that have already been completed, specific unhealthy mediating behavioural factors, such as smoking, unhealthy diet, and lack of exercise, were strongly linked to lower socioeconomic statuses and an increased morbidity/ill-health Stringhini, Sabia, Shipley, Brunner, Nabi, Kivimaki, & Singh-Manoux (2010). Furthermore, the differential effects of physical activity at work and during leisure time may be an important driver of socioeconomic health differences according to Pampel, Krueger, and Denney (2010). As one engages in manual labour during their employment, there is an absence of more beneficial, high-intensity activities (which have been shown to have the most significant health benefits).

Interestingly, social isolation and emotional loneliness have also been significantly correlated with lower household incomes (Macdonald & Nixon, 2018). For example, Kahneman and Riis (2005) found a correlation coefficient $r = 0.85$ in the 18 OECD countries analysed. Happier people tend to lead more active lives (Lathia, Sandstrom, Mascolo, & Rentfrow, 2017) and live longer.

One of the larger studies undertaken on socioeconomic factors identified a more significant association between poor cardiovascular health and socioeconomic problems other than illness. (Winkleby, Jatulis, Frank, & Fortmann, 1992). It also identified that a lower occupational position (a proxy for socioeconomic position) was shown to have a greater impact on health than obesity or hypertension. Another health indicator, frailty, has been found to be more prevalent among people in lower socioeconomic classes and can be considered a mediator of socioeconomic inequalities (Hoogendijk, Heymans, Deeg, & Huisman, 2018).

Clearly, socioeconomic policy is a complex topic with multi-faceted aspects. Examining just one aspect (i.e., dietary/behavioural) provides an incomplete picture as one's health and

wellbeing is comprised of many factors (Moor, Spallek, & Richter, 2017), which provides the inspiration for further data incorporation.

It is surmised that blue-collar workers, clerks and service workers are more likely to be exposed to negative working conditions (Niedhammer, Chastang, David, & Kelleher, 2008). There are a number of existing deprivation indices that will feature heavily in this report. In their research for the Trinity Deprivation Index[4], the authors included four key indicators to define deprivation.

1. Unemployment
2. Low social class
3. Local authority rented housing
4. No car

Moreover, (Rigby, Boyle, Brunsdon, Charlton, Dorling, French, and Pringle 2017) identified three contributing factors to health inequality, namely structural, behavioural, and contextual levels. The author found higher levels of mortality in the major cities and in isolated rural areas. Building on these results with localised analysis is critical to furthering the discussion on public health policy.

2.2 Local Services and Amenities

Limited research has been conducted on the impact of the local environment on the public's health and wellbeing. The environment in this context could relate to the natural surroundings, such as parks and trees, or to the cultural surroundings and available amenities. Despite the scarce research on this topic, some studies have linked community activities to improved physical and mental health in comparison to individuals who are less involved with other people (Thompson & Kause, 1998). Initiatives such as these community activities that empower people to decide what's most important lead to greater self-awareness and greater social impact (Plane & Klodawski, 2013).

For example, one study related to neighbourhood demonstrated that a sense of belonging in a neighbourhood was associated with better physical and mental health, lower stress, and other positive health factors (Young, Russell, & Powers, 2004). Similarly, studies have concluded that adolescents who had access to a recreation centre had an increased level of physical activity by 22% (An, Yang, & Li, 2017). This demonstrates the value of local planning to promote activity for youths.

It is also well documented that stress levels tend to be reduced with greater access and time spent in green spaces (Hazer, Formica, Dieterlein, Morley, 2018; Ulrich, Simons, Losito, & Fiorito, 1991). It was found that the greater accessibility to a range of social infrastructure services promoted the SWB (subjective wellbeing) of residents (Davern et al., 2017).

[4] https://www.tcd.ie/medicine/public_health_primary_care/research/deprivation/

Interestingly, Cole's research (2019) indicates that the health benefits of green space in a neighbourhood are not felt by all social classes equally in New York City. Those with a lower socioeconomic status see little or no health benefits from open green spaces while those with higher incomes were seen to benefit the most (Cole, 2019). As West asserts, parkland density is correlated with physical activity and negatively correlated with obesity (West, Shores, & Mudd, 2012). Greenspace has proven to be beneficial in areas, but not to all residents, and not equally. Identification of additional amenity factors will help deliver enhancements of the existing research and improve its applicability.

2.3 State-of-the-Art Architecture

Censuses are an important tool in public planning and population planning. Therefore, they feature strongly in research as a decision support tool. Data mining is the extraction of previously obscured information from large data sets requiring specialised techniques. However, data collected on a census level is often aggregated at wider levels and has therefore presented problems when applied to ML algorithms to draw meaningful insights.

For example, Bin Sheng (2010) used classification and regression tree (CART) algorithms to classify Chengyang and Laixi inhabitants into classes based on income and consumption levels. They achieved consistent results using four attributes for every person from the census data. CART is often chosen for its simplicity and interpretability.

More recently, Alvarez-Galvez (2016) used Bayesian networks to decipher the relationship between socioeconomic status and health. They argue that standard regression models are not sufficiently sophisticated to handle the level of complexity of the interrelationship between their variables. Bayesian networks provide the ability to graphically depict the complex relationship between the variables rather than specific associations.

Similarly, in their classification of British entrepreneurs, (Montebruno, Bennett, Smith, & Van Lieshout, 2020) utilise an ensemble of ML algorithms alongside more classical logit regression (LR) and argue for a multidisciplinary approach to data classification. They determined that ten common and optimised ML algorithms performed better than LR in almost every case. This is consistent with the find from Alvarez-Galvez (2016) discussed above.

In his study on the use of big data in the census, Chatfield (2010) identified several challenges facing the CSO in Ireland with the adoption of big data and more modern methodologies. These include access, capability, and reputational risk. However, it's clear from recent developments, such as its collaboration with the Economic and Social Research Institute (ESRI) in Ireland, that the CSO is keen to expand the census use. This, in turn, increases the potential for more in-depth analytical solutions, and it's clear that additional research is required in this area.

3 Research Methodology and Design

3.1 Introduction

The basis of this work was to combine quantitative research on a local level with the power of ML to develop a framework to understand the key drivers of health self-determination in Irish Society. The chosen methodology was Crisp-DM as the problem requires a quantitively detailed and technological solution and can be viewed in Figure 1. This methodology was the chosen implementation due to the nature of the problem statement and contains the following steps: (i) Business Understanding: develop an indicative measure of health (ii) Data Understanding: identify, extract, and understand relevant data sets (iii) Data Preparation: transform the data set into a spatial ready form (iv) Data Modelling: model analysis and performance evaluation (v) Evaluation: discover the key drivers and decipher their relationship and measure through predicted health outcomes (vi) Deployment**:** visualise and rank model performance and factor influence on results.

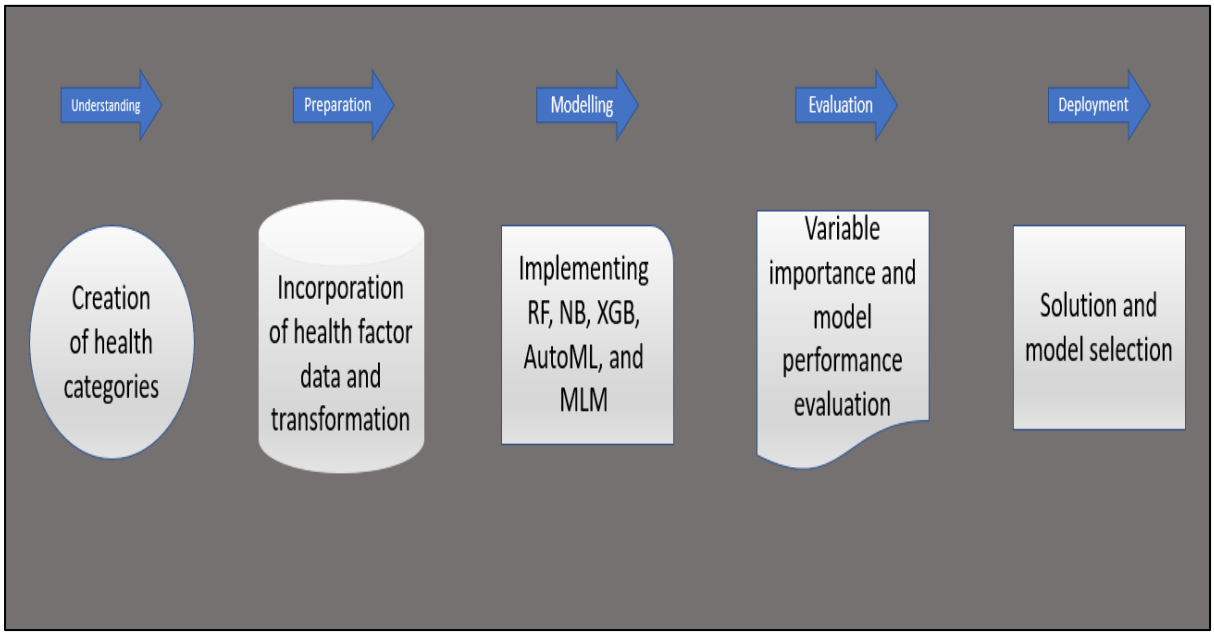

Figure 1: Methodology of public health factor identification and determination.

3.2 Data Understanding

In the most recent iterations of the Ireland census, a question asked respondents to self-assess their general health. This was divided into five categories (very good, good, fair, bad, and very bad). To gauge the sense of an area, a ratio of the sum of respondents 3:5/total was created and split into equal buckets based on the result. This prompted the generation of a factor with five levels (higher, above, middle, below, and poor) that reflects the problem statement: Which factors prompt the separation of areas into these buckets?

3.3 Data Preparation

Fingal County Council has a rich and varied set of data relating to the local area available on data.gov.ie. Seven relevant data sets were selected: streetlights, trees, parks, schools, playgrounds, health centres, and leisure sites. Each data set contributed to representing the local amenities and geographical features, as highlighted in the literature review. Fingal County Council is made up of 964 small areas with a wide mix of city and country style living which is representative of Ireland as a whole. The CSO also provides a shapefile to visualise these small areas. The chosen method for this project was to plot as polygons utilising the standard geographic coordinates systems World Geodetic System (WGS).

The area covered by Fingal County Council has been chosen for this project as it represents a useful example of a cross-section of Irish society. In a prosperous nation that is 10th on the world's prosperity index[5], with a healthcare system ranked 51st on the same index, Fingal contains both a significant urban and rural population and includes a diverse range of amenities that is broadly reflective of society in Ireland. Its population at the last census was 296,214, with its inhabitants residing in locations ranging from densely populated areas to rural townlands.

To augment this data, the deprivation index created by Trutz Hasse and Jonathan Pratchke in conjunction with Pobal is utilised with permission granted for the small area data set[6]. This contains a score of each small area based on several measures across car ownership to education and social status. This provides a rich cross-sectional societal view to determine the affluence and deprivation of each small area, providing an opportunity to augment the data set with local area level enrichment data. However, it was collated using the 2011 mapping for small areas, of which circa 3% have been realigned. These will be excluded to ensure a complete data set is available with comparative reference data.

3.4 Modelling

Five methods were selected: Random Forest; XGBoost; AutoML; Multinomial Regression; Naïve Bayes. These were chosen for their varied implementations and balance in interpretability. Benchmarks of performance can be gauged through the application of the traditional methods, such as Regression and Random Forest, as discussed in the literature review. These also have the advantage of greater interpretability and thus a greater ability to inform discussions on the key factors that drive health determination.

[5] https://www.prosperity.com/
[6] http://trutzhaase.eu/services/hp_deprivation_index/

3.5 Evaluation

Variable importance was considered for each of the models to determine the key drivers of the predictions. Precision and kappa were used to evaluate model performance to identify and isolate the most performant models; namely, which models were most accurately able to predict the health categorisation of small areas through the enrichment factors.

3.6 Deployment

The data was presented as a visualisation in show areas by their derived health category and the various contributing factors. The results of the models were gauged by their effectiveness in solving the problem, accuracy, and interpretability in order to determine the optimal solution.

3.7 Design Flow

The design process delivering this methodology is based on a three-tiered structure (data tier, business tier, and logic tier), as shown in Figure 2. The data tier consists of the data capture, cleansing, integration, and factor derivation using geo-based distance measures. The logic tier contains the health category generation and the various model implementations using R. The presentation tier contains the model output depiction and underlying data visualisation.

Figure 2: Design of public health factor identification and determination.

4 Implementation

The response variable is obtained from the 2016 Ireland Census, which asked respondents to assess their health from a level of 1–4 (categorically very poor to very good). The input variables come from publicly available or documented sources, such as the census, deprivation index, and Fingal County Council through the Ireland open data portal. The amenity factor data sets were extracted as geographic coordinates (longitude and latitude or easting and northing), which were then projected onto the shapefile to indicate which polygon (i.e., small area) each factor belonged to. The output represented the existence of each point in every small area in Fingal. These were subsequently aggregated to give a count of the number of points in each polygon.

Due to the sparse nature of some of the data (i.e., parks, schools, playgrounds, health centres, and leisure sites), the distance from each polygon to the nearest amenity point (in metres) was calculated to represent the accessibility of each amenity. These were derived by transforming the data and projecting it onto a planar coordinate system to allow for the calculation of the distance.

This data was merged with the amenity data using the small area key through the R programming language. Various packages were used to process the data, measure the distance, and assign the categories (e.g., sf, rgdal, maptools, rgeos, sp, and data.table). R was also used to model the data and transform the variables.

Base models: Naïve Bayes
Parameter tuned model: XGBoost, Random Forest, Multinomial Regression
Automated optimisation: AutoML

The package Caret in R was used to implement all the models (save for Auto ML) and to develop optimised parameters for XGBoost and Random Forest. Classification accuracy, precision, and recall were used as measures of a model's performance on a random 25% holdout set.

Moreover, variable importance (where available) was used to demonstrate the relative impact each variable had on the model's outcome and to determine which variables were the most significant drives of one's self-determined health.

AutoML was utilised to demonstrate the effectiveness of auto-tuned models. This was undertaken using the h2o package in R to automate the supervised model[7]. AutoML trains and cross-validates a random forest, an extremely randomised forest, a random grid of gradient boosting machines (GBMs), and a random grid of deep neural networks, and then trains a

[7] https://www.rdocumentation.org/packages/h2o/versions/3.30.0.1/topics/h2o.automl

stacked ensemble using all the models. Finally, confusion matrices were created for all models to compare the output and ascertain the best model performance.

5 Evaluation and Results

5.1 Results

Each response variable was aggregated and reviewed by its explanatory variable to deliver a high-level understanding of the breakdown of the health determinations. A plot of Fingal County Council can be seen in Figure 3 depicting the varied levels of health determination across the region and small areas. The figure depicts each area coloured by its respective outcome.

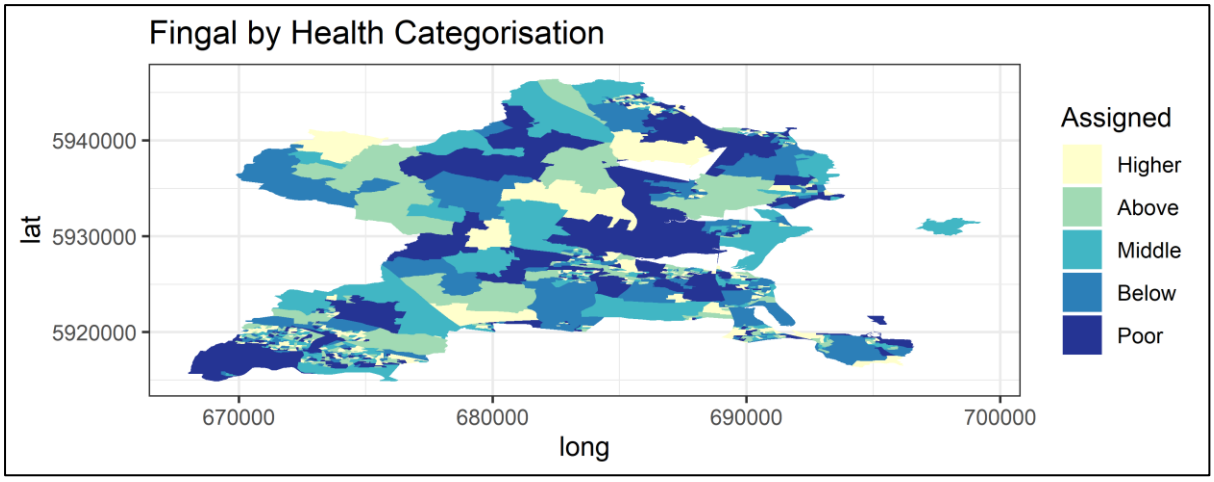

Figure 3: Map of Fingal by self-determined health.

While every area and electoral district has a mix of outcomes, a pattern can be seen within the map, with the same outcomes often crossing area boundaries. These areas were aggregated into electoral districts to discern the areas with the greatest prevalence of poor health outcomes. This can be seen from Table 1, where Electoral District (ED) Blanchardsotwn-Corduff has 83.33% of areas in its electoral district assigned to the lowest bucket.

Table 1: Lowest-scoring electoral districts

Assigned	*ED Name*	*Poor Heath Areas*	*Total Areas*	*Percentage*
Poor	Blanchardstown-Corduff	10	12	83.33%
Poor	Balbriggan Urban	15	28	53.57%
Poor	Blanchardstown-Roselawn	3	6	50.00%
Poor	Blanchardstown-Tyrrelstown	3	6	50.00%
Poor	Kilsallaghan	3	6	50.00%

Visually, it is readily apparent which factors contribute to a lower rating of health. Looking at the top (Figure 5) and bottom six areas (Figure 4) by assigned health indicates a dramatic difference to the mean in distribution across the factors. As illustrated in Figure 3, areas with high levels of education scored especially well on their self-assessed health.

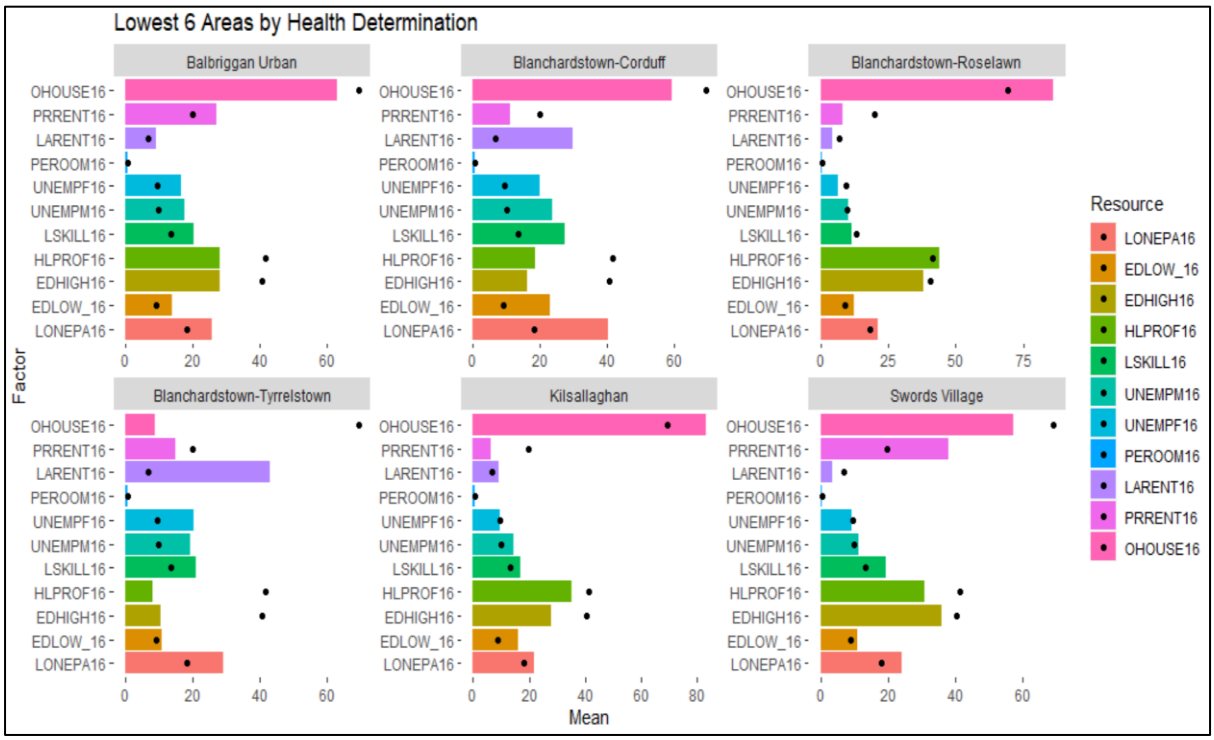

Figure 4: Lowest six areas by self-determined health.

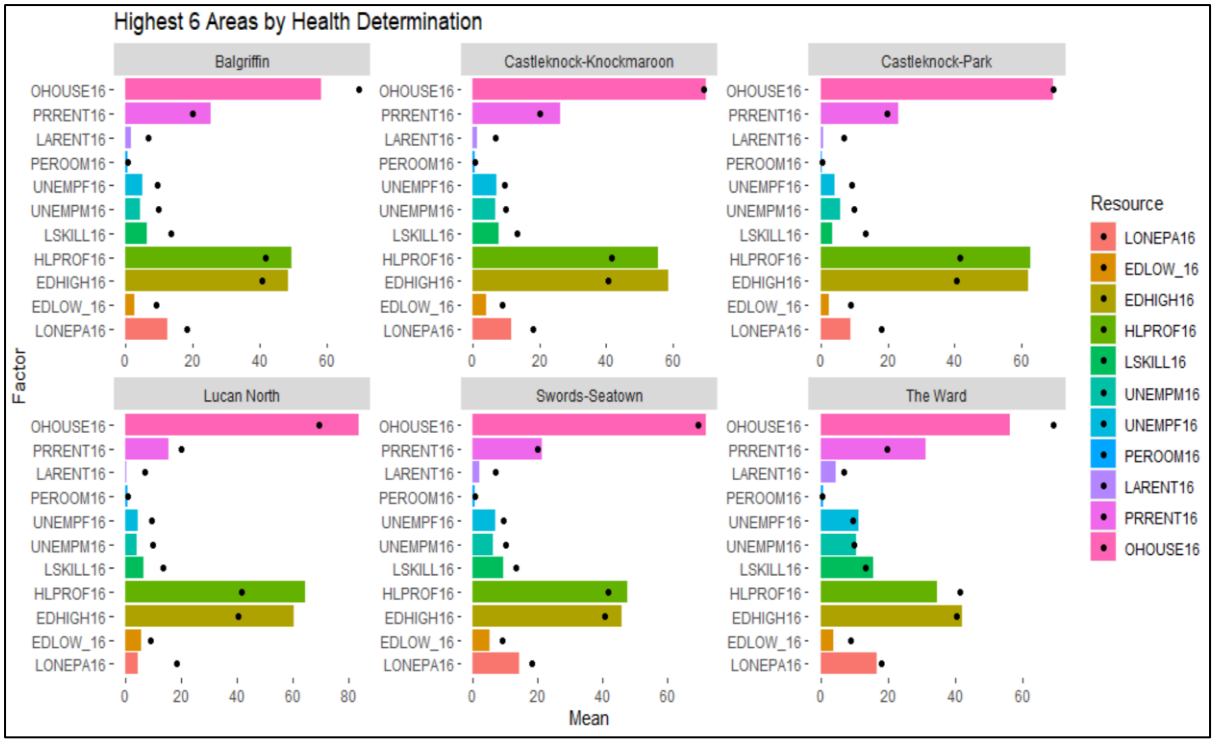

Figure 5: Highest six areas by self-determined health.

These factors were collated and then split into five models. The health classification response was equally divided across the regions, so a no-information accuracy score of 20% is to be assumed in each model.

All models were implemented using the caret package in R and allowed for random tuning to ensure that bias was not present in the comparisons. Data sets were split into training (75% constituting of 705 rows and 28 input variables) and test (25% constituting of 232 rows and 28 input variable). Every model was implemented, tested, and (where applicable) the hyperparameters were adjusted to improve results. An expanded grid search was created for RF, XGB, and MLM. The models were selected using 10-fold cross-validation to ensure a robust model was delivered and avoid overfitting the data. Furthermore, AutoML from the package h2o was implemented using a local cluster. This followed a similar methodology and set up to the aforementioned caret implementations.

The results for each model can be viewed below. An accuracy rate of 49% was best achieved using multinomial Regression.

Table 2: Model results

Model Results Summary

Model	***Accuracy (%)***	***Kappa (%)***	***AccuracyLower (%)***	***AccuracyUpper (%)***	***AccuracyNull (%)***	***AccuracyPValue (Probability)***
Random Forest	0.47	0.34	0.40	0.53	0.20	0.00
Multinomial Regression	0.49	0.36	0.42	0.455	0.20	0.00
XGBoost	0.43	0.29	0.37	0.50	0.20	0.00
Naive Bayes	0.42	0.35	0.35	0.48	0.20	0.00
h2o	0.38	0.22	0.32	0.45	0.275	0.048

Furthermore, variable importance was captured for the models, as illustrated below in Table 3.

Table 3: Variable importance

Predictor	*Random Forest*	*Multinomial Regression*	*XGBoost*
AGEDEP16	44.28	2.32	0.94
EDHIGH16	64.03	1.14	11.56
EDLOW_16	100.00	13.69	100.00
Health_Area	0.00	63.51	0.00

HLPROF16	56.15	1.99	4.11
HP2016abs	70.23	26.22	16.51
HP2016rel	76.89	23.17	42.53
LARENT16	39.07	3.32	0.29
Leisure_Area	4.63	35.04	0.00
LONEPA16	47.53	1.36	3.41
LSKILL16	49.46	1.24	4.00
min_dist_health	51.98	0.05	3.52
min_dist_leisure	42.28	0.05	0.51
min_dist_park	41.25	0.00	2.41
min_dist_playg	40.16	0.00	0.00
min_dist_schools	46.65	0.04	4.68
OHOUSE16	42.52	1.84	0.00
Parks_Area	1.11	49.97	0.00
PEROOM16	26.37	100.00	0.00
Play_Area	0.64	37.78	0.00
POPCHG16	18.42	12.42	2.31
PRRENT16	44.76	2.21	3.39
School_Area	3.86	17.99	0.00
Streetlight_Area	40.26	0.34	0.89
TOTPOP16	47.59	0.07	2.56
Tree_Area	31.15	0.15	0.73
UNEMPF16	39.11	5.05	0.00
UNEMPM16	48.24	1.13	0.00

5.2 Critical Analysis

It is clear from the results above that the deprivation index is an indicator of health outcomes. All results delivered were shown to be significant, as demonstrated by the accuracy P value in Table 2.

This combined factor created from the derivation index proved to be one of the most critical variables in determining the health outcomes in all of the models. However, the most important

factor in the most accurate model was the number of people per room in a household. As can be seen in Table 3, the enrichment factors played a crucial role in generating accurate results.

Areas with the lowest health outcomes scored strikingly on the individual factors that made up the deprivation index. As can be seen in Figure 6, these areas had a significantly higher proportion of single parents (LONEPA16), unemployment (UNEMPM16 and UNEMPF16), low or semi-skilled manual workers (LSKILL16), local authority housing (LARENT16) while having the lowest ratio of residents reaching higher-level education (EDHIGH16). The vast majority of the factors resulted in the 'Poor' category being either the highest or the lowest (e.g., low skilled worker percentage for 'Poor' area was 20% - the highest across health outcomes). A sliding scale is evident across nearly all factors from 'Poor Health' to 'Higher Health'. It can be clearly interpreted from these results that health and socioeconomic status are intrinsically linked.

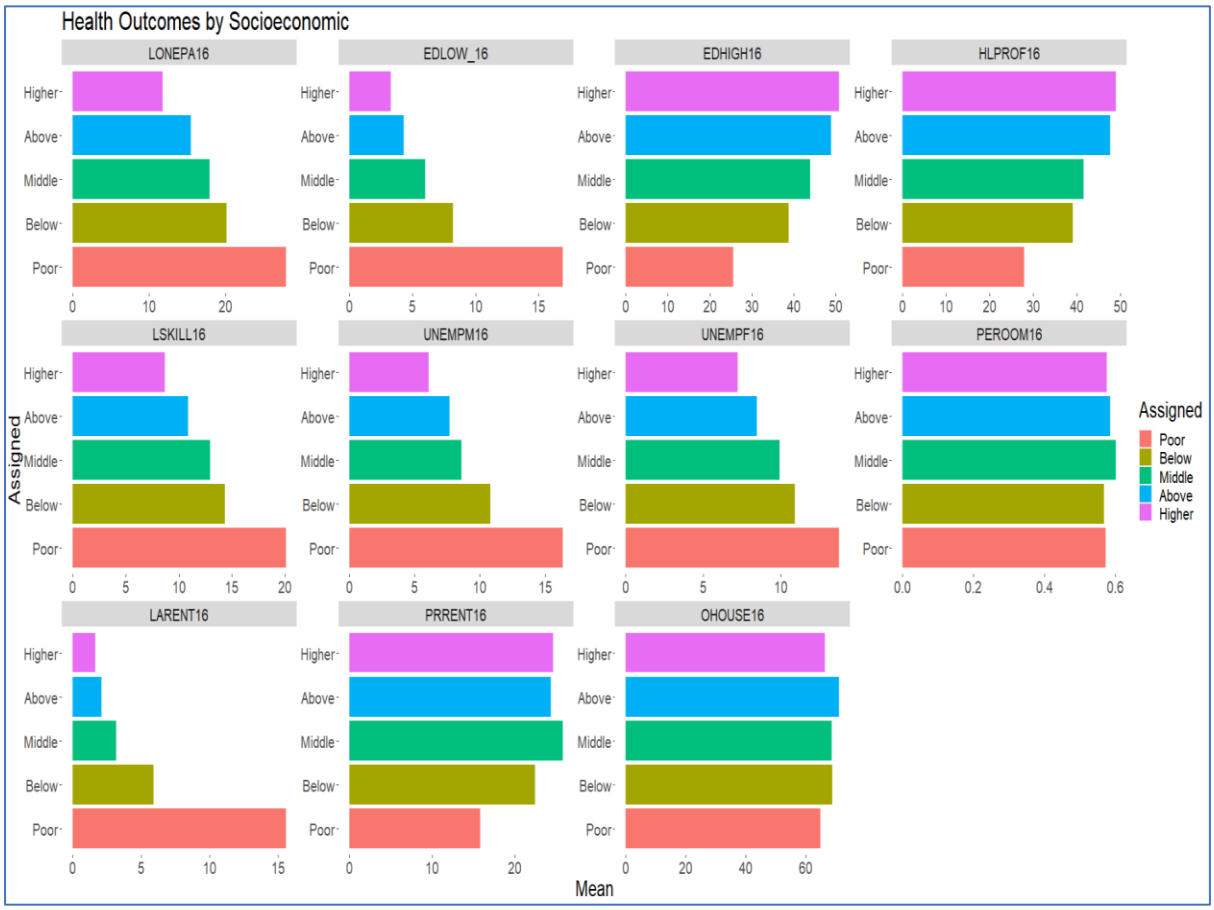

Figure 6: Socioeconomic by Assigned.

As in Figure 4, the electoral district with the lowest level of self-determined health, namely Blanchardstown-Corduff, has the highest level of residents with primary education only in all of Fingal. This provides evidence for the hypothesis established in related works and indicates that being of a lower socioeconomic class is likely to lead to poorer health outcomes.

This study's second point of analysis was to use various methods and models to explore whether available facilities can make a significant difference in health outcomes. It is clear from the variable importance displayed in Table 3 that these factors were significant in delivering accurate predictions from the models. It demonstrates that the availability of local facilities is predictive of the general wellbeing of an area's inhabitants.

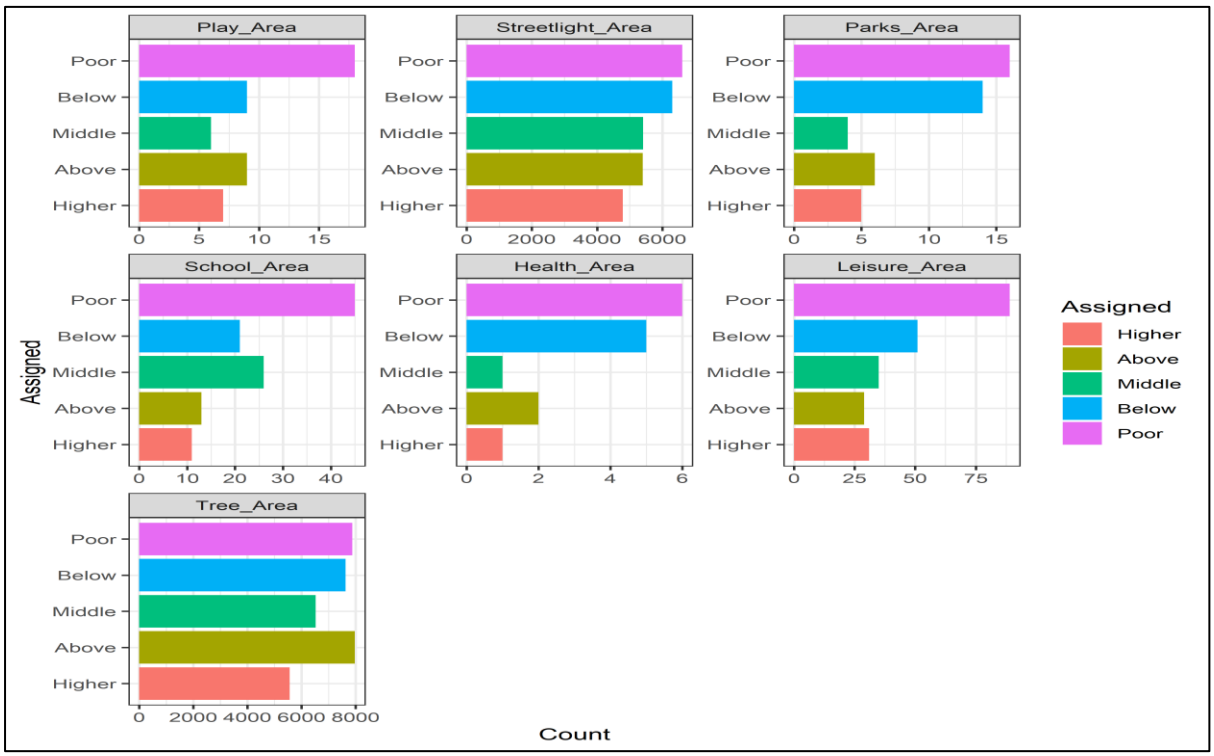

Figure 7: Resources by derived health bucket.

In the most accurate model (i.e., multinomial Regression), the number of parks in an area was the explanatory variable with the greatest significance. Surprisingly, the availability of these public amenities was significant in each model by their mere presence. In Figure 6, it demonstrates that these amenities were much more prevalent in areas with lower health outcomes rather than improved outcomes.

While this finding demonstrates that these facilities are available in areas that could be deemed to have a higher need, no correlating positive health outcome has been realised. In the long term, it would be interesting to examine whether benefits brought about by the availability of these amenities can help change the pattern of health in these areas. It would be relevant to include the creation date of the amenity and track the changes in health outcomes over time. A limitation of this study results from both the lack of available data and barriers to obtaining relevant data. This study and future work can be further enhanced with greater data capture and integration.

Table 4: Resources by derived health bucket

Assigned	*Play_Area*	*Streetlight_Area*	*Parks_Area*	*School_Area*	*Health_Area*	*Leisure_Area*	*Tree_Area*
Higher	7	4787	5	11	1	31	5564
Above	9	5401	6	13	2	29	7978
Middle	6	5408	4	26	1	35	6523
Below	9	6307	14	21	5	51	7630
Poor	18	6614	16	45	6	89	7887

Multinomial Regression proved to be the strongest performer, delivering the highest accuracy and kappa value. Evidently, the models performed well, utilising the data provided. Augmenting the deprivation data with the local amenity data significantly enhanced the model's output and demonstrated the importance of capturing the enrichment data.

In terms of performance, the models scored relatively similarly using the baseline caret implementation. There was a worsening performance in the utilisation of the h2o and the AutoML functions versus the automatic parameter tuning. An 11-percentage point gap from the highest performer (MLM) to the lowest (h2o) demonstrates the value of utilising multiple models in order to derive the best fit. Chatfield (2010) discussed the capability challenge in terms of producing the best models on public resources. Interpretability across the models was consistent (apart from AutoML's ensemble). Interestingly, the results somewhat differ from the results outlined above by (Montebruno, Bennett, Smith, & Van Lieshout, 2020) in terms of the regression performance with MLM being the top performer.

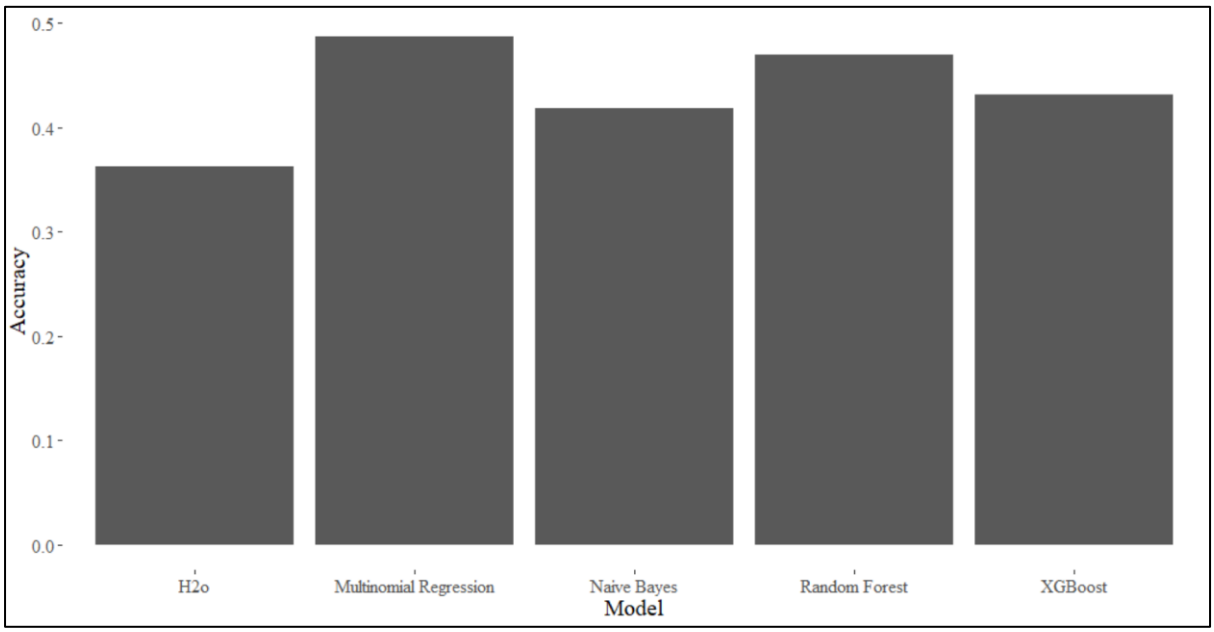

Figure 8: Model Accuracy

While this study didn't develop bespoke ensembles, the difference in implementation plans taken by these models indicates that an ensemble model might be more proficient in determining the health outcomes. This should be explored for future developments while weighing up the impact on the interpretability of the outcome.

6 Conclusion and Future Work

Throughout this research, the focus was to explore the effect socioeconomic factors had on determining the course of an individual's health. Through the aggregation of available data from the census, deprivation indices, and open governmental data, it was proven that these factors are significant drivers of health. Therefore, investment in underprivileged areas is paramount to changing these health outcomes and saving potential costs of the health service.

Moreover, it was demonstrated that the availability of public amenities is predictive in determining health outcomes in the Fingal area, and from which accurate ML models can be built. Interestingly, the amenities included tended to negatively correspond with the overall self-determined health of the local population. Further research on whether the health outcomes changed with the development of these resources should be conducted. Clarity as to why these facilities were located where they are and an understanding of their development history would further inform the research conducted in this paper.

Finally, public policy needs robust decision-making tools. Therefore, investing in data capture and data analysis would enable more precise and sophisticated model development to drive impactful policy decision making. This analysis has elucidated that the collection of relevant enrichment data can drive greater model performance with an accuracy score of 49% vs a non-information rate of 20%. Furthermore, the development of a range of models delivers improved results, and these considerations should be at the centre of any policy initiative development and implementation.

7 Acknowledgements

I would like to thank my supervisor, Dr. Catherine Mulwa, for her assistance. I thank my partner for her support throughout this project and always. I'd like to express my gratitude to my family, especially my parents, for their constant guidance and motivation.

8 References

Alvarez-Galvez, J. (2016). Discovering complex interrelationships between socioeconomic status and health in Europe: a case study applying Bayesian networks. *Social Science Research, 56,* 133–143.

An, R., Yang, Y., & Li, K. (2017). Residential neighborhood amenities and physical activity among US children with special health care needs. *Maternal and Child Health Journal*, *21*(5), 1026–1036.

Bin Sheng, S. G. (2010). *Data Mining in census data with CART*. Chengdu, China: ICACTE.

Blaxter, M. (1997). Whose fault is it? People's own conceptions of the reasons for health inequalities. *Social Science & Medicine*, *44*(6)747–756.

Chatfield, A. T.-C. (2010). Census big data analytics use: International cross case analysis. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data* Age, 1-10.

Cole, H., Triguero-Mas, M., Connolly, J., & Anguelovski, I. (2019). Determining the health benefits of green space: Does gentrification matter? *Health & Place*, *57*, 1–11.

Davern, M., Gunn, L., Whitzman, C., Higgs, C., Giles-Corti, B., Simons, K., … Badland, H. (2017). Using spatial measures to test a conceptual model of social infrastructure that supports health and wellbeing. *Cities & Health*, *1*(2), 194–209.

Hassani, H. G., Saporta, G., & Silva, E. (2014). Data mining and official statistics: the past, the present and the future. *Big Data*, *2*(1), 34–43.

Hazer, M., Formica, M., Dieterlen, S., & Morley, C. (2018). The relationship between self-reported exposure to greenspace and human stress in Baltimore, MD. *Landscape and Urban Planning*, *169*, 47–56.

Hoogendijk, E., Heymans, M., Deeg, D., & Huisman, M. (2018). Socioeconomic inequalities in frailty among older adults: results from a 10-year longitudinal study in the Netherlands. *Gerontology*, *64*(2), 157–164.

Kahneman, D., & Riis, J. (2005). Living, and Thinking about It: Two Perspectives on Life. *The Science of Well-Being*, *1*, 285–304.

Lathia, N., Sandstrom, G., Mascolo, C., & Rentfrow, P. (2017). Happier People Live More Active Lives: Using Smartphones to Link Happiness and Physical Activity. *PLoS one*, *12*(1), e0160589.

Macdonald, S. J., & Nixon, J. (2018). 'Loneliness in the city': examining socio-economics, loneliness and poor health in the North East of England. *Public Health*, *165*, 88–94.

Montebruno, P., Bennett, R. J., Smith, H., & Van Lieshout, C. (2020). Machine learning classification of entrepreneurs in British historical census data. Information Processing & Management, 57(3), 102210.

Moor, I., Spallek, J., & Richter, M. (2017). Explaining socioeconomic inequalities in self-rated health: a systematic review of the relative contribution of material, psychosocial and behavioural factors. *J Epidemiol Community Health*, *71*(6), 565–575.

Niedhammer, I. C., Chastang, J. F., David, S., & Kelleher, C. (2008). The contribution of occupational factors to social inequalities in health: findings from the national French SUMER survey. *Social Science & Medicine*, *67*(11), 1870–1881.

Pampel, F., Krueger, P., & Denney, J. (2010). Socioeconomic disparities in health behaviours. *Annu Rev Sociol*, *36*, 349–70.

Plane, J., & Klodawsky, F. (2013). Neighbourhood amenities and health: Examining the significance of a local park. *Social Science & Medicine*, *99*, 1–8.

Rigby, J., Boyle, M., Brunsdon, C., Charlton, M., Dorling, D., French, W., & Pringle, D. (2017). Towards a geography of health inequalities in Ireland. *Irish Geography*, *50*(1), 1–27.

Stringhini, S., Sabia, S., Shipley, M., Brunner, E., Nabi, H., Kivimaki, M., & Singh-Manoux, A. (2010). Association of socioeconomic position with health behaviors and mortality. *JAMA*, *303*(12), 1159–1166.

Thompson, E., & Kause, N. (1998). Living alone and neighborhood characteristics as predictors of social support in late life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *53*(6), 354–36.

Ulrich, R., Simons, R., Losito, B., & Fiorito, E. (1991). Stress recovery during exposure to natural and urban environments. *Journal of Environmental Psychology*, *11*(3), 201–230.

West, S., Shores, K., & Mudd, L. (2012). Association of available parkland, physical activity, and overweight in America's largest cities. *Journal of Public Health Management and Practice*, *18*(5), 423–430.

Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. American journal of public health, 82(6), 816-820

Young, A. F., Russell, A., & Powers, J. R. (2004). The sense of belonging to a neighbourhood: can it be measured and is it related to health and well being in older women? *Social Science & Medicine*, *59*(12), 2627–2637.