ON USING DEMOGRAPHIC DATA WITH DEPRIVATION INDEX FOR PREDICTING CHRONIC DISEASES

OLUGBENGA IYIOLA

Master's Program in Computer Science

APPROVED:

Monika Akbar, Ph.D., Chair

Mahmud Shahriar Hossain, Ph.D.

Rigoberto Delgado, Ph.D.

Leopoldo Gemoets, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

ON USING DEMOGRAPHIC DATA WITH DEPRIVATION INDEX FOR PREDICTING CHRONIC DISEASES

by

OLUGBENGA IYIOLA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2021

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Monika Akbar, for her support and guidance towards the successful completion of this project. Her encouragement as well as critique were instrumental to helping me navigate through the various challenges I faced in the course of the research.

I would also like to really appreciate all my committee members for their contributions. First, I want to appreciate Dr Mahmud Shahriar Hossain for helping me to understand all the data mining concepts that I needed for this research. Next, I want to thank Dr Rigoberto Delgado and Dr Leopoldo Gemoets for giving me the opportunity to be a member of their research team. This tremendously helped me develop my research skills and the subject matter for this study was as a result of what I learned as a member of the team. Their influence and assistance also helped me access all the datasets from the Department of Public Health that I needed for the work.

I am using this opportunity to also appreciate my family and friends for their encouragement and show of goodwill towards the success of my graduate program. Above all, to Jesus the lover of my soul, I am grateful for the gift of life and for always being there for me.

Abstract

Researchers have worked on modeling and predicting the likelihood of developing chronic diseases, such as diabetes and high blood pressure, using medical data (e.g., heart-rate, blood sugar). However, many of these diseases demonstrate strong links with demographics and socio-economic status (e.g., race, gender, income). It is also less time-consuming to retrieve demographic and socio-economic data, some of which are publicly available through US Census Bureau, than to carry out medical tests. Hence, demographic data can give a quicker estimate of the susceptibility of a person to a chronic disease.

In this work, we study the effect of using medical vs. demographics data for modelling and predicting two chronic diseases: diabetes and high blood pressure. We proposed an updated deprivation index to build disease models that consider demographic data. Our results indicate demographic data are as good or better indicators for predicting chronic diseases.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

There are many studies [24, 38, 56] on disease prognosis using machine learning algorithms. These diseases can be broadly categorized under two groups: chronic and infectious. Significant progress has been made on the predictive accuracies of many machine learning algorithms for detecting chronic diseases. As a consequence, it is now possible to detect some of these diseases earlier than before, thus significantly reducing their mortality rates [33].

Studies mostly rely on diagnostic test results (or, patient's vital information) for determining the likelihood of developing chronic diseases [39]. We observe that in such studies, demographic data has been used less, although such data hold much promise as they contain valuable information about common trends seen across communities. Accordingly, a few researchers have developed deprivation indices integrating different types of demographic data [16]. These indices are useful in measuring health outcomes in a geographical area. Such indices mostly use generic demographic attributes such as unemployment rate, household size, and number of vehicles [71] [64]. However, for communities that demonstrate different traits - such as, communities on the border of a country - these indices fall behind in accurately capturing the vulnerability towards certain diseases. In this paper, we extend an existing deprivation index to include community-specific demographics data, such as citizenship information and preferred choice of language. We retrieved these data from the US Census Bureau [17]. The data is available at multiple levels of

abstraction (e.g. zipcode, census tract etc.). We used the zipcode-level census (i.e., the ZCTAs*) data. In particular, we investigated if a person's area of residence along with the modified deprivation index score, is a better indicator for predicting his/her vulnerability towards chronic diseases. We used two chronic diseases as our case studies which were diabetes and high blood pressure(HBP). Before we proceed, it is important we discuss these two diseases and their impacts.

1.2 What is a Chronic Disease?

According to Centers for Disease Control and Prevention(CDC), chronic diseases are medical conditions with a life span of one year or more that require ongoing treatment or/and limit people from performing their daily activities. Two of the most prevalent chronic diseases in the US are diabetes and HBP and they constitute the leading causes of death and disability.

Diabetes is caused by high blood glucose and this disease disproportionately affects minority populations and the elderly [6]. El Paso county, with a dominant population of Hispanics, has about 12% diabetic adult patients excluding pregnant women diagnosed with the disease during their pregnancy period. HBP on the other hand, is caused when the force of the blood pushing against the walls of the blood vessels is consistently too high. This disease can lead to heart attack, kidney failure etc. About 32% [5] of El Paso's adult population have been diagnosed to have blood pressure 140/90 mm Hg, which is considered high. This condition is also common among minorities, especially blacks, elderly people above 65 years, alcoholics, and obese people.

Early detection of these two diseases has been proven to be able to significantly reduce their effects and mortality rate. This has been the main motivation of

*ZIP Code Tabulation Areas (ZCTAs)are generalized areal representations of United States Postal Service ZIP Code service areas

many studies that seek to design accurate machine learning models that can be used to predict an individual's vulnerability to these diseases. For example, for a chronic disease like cancer, we are either interested in the prediction of susceptibility, recurrence and/or survivability [37]. Early detection, managing and controlling chronic kidney disease will also significantly help to increase its survival rate [36].

Clark [35] pointed out that the successful management and treatment of a patient with a chronic disease is determined to a great extent by his social and environmental factors. He noted that the patient's actions determine the outcome of the chronic disease control efforts.

Hence, there is a need to have an in-depth knowledge of how demographic factors can increase the susceptibility to chronic diseases and why such factors could prove to be very valuable for machine learning models.

1.3 Demographic and Health Related Data

Demographics consist of a population's socioeconomic information such as income, age, marital status, sex etc. These data are collected by the U.S. Census Bureau yearly via the American Community Survey (ACS) and decennial through a comprehensive count of every American household.

However, health data is any data on the health status, reproduction, mortality causes, and quality of life for an individual or population. Usually, these data are collected when individuals interact with health care systems such as going for a medical check up. There are two types of health data, structured and unstructured. Structured health data is standardized and it includes information such as patient names, contacts, demographics, laboratory values, financial details etc. However, unstructured health data is not standardized and examples of this are emails, audio recordings, physician notes for a patient, [19] etc. Apparently, demographic data can give a quicker and less costly estimate of the susceptibility of a person for a

chronic disease than health data. Therefore, a machine learning model that uses only demographic data for its prediction would be a preferred choice. It would be useful to understand how a deprivation index can enhance the predictive performance of such machine learning model.

1.4 Understanding Deprivation Index

According to Peter Townsend [83], deprivation is a state of observable and demonstrable disadvantage relative to the local community or the wider society or nation. Measuring deprivation in a targeted area has been found to be very useful in determining how to use scarce resources to prevent, diagnose, and effectively manage chronic conditions within vulnerable populations [64]. The measurement of socio-economic deprivation is usally done through the use of composite indices. A popular composite index that is frequently used by researchers [52, 68, 78] is the Townsend deprivation index, which was developed in the United Kingdom. It is a small-area deprivation index with four census-based variables which are unemployment rate, home ownership, household overcrowding, and vehicle availability [82]. It can be constructed for any geographical area for which census data are available. This index has been used to study the relationship between deprivation in a geographical area and various health outcomes such as bacteremic pneumonia [30], tuberculosis [60, 63], sexually transmitted diseases [60], motor vehicle fatalities [50] and infant mortality [48].

However, a major limitation of such indices is that they are usually sensitive to urban–rural differences [28]. According to Ana et al. [64], a particular composite index may not necessarily be suitable to meet the health needs in all geographic regions or across diverse population groups. To our knowledge, there is no existing deprivation index that fully captures the deprivation conditions in border cities with high population of immigrants.

In this study, we use an approach similar to that of the IMD [53] and Townsend [83] to construct a customized deprivation index suitable for communities on the border of a country such as El Paso. This updated deprivation index is used as a feature in the machine learning models that consider mainly demographic data. We would now describe specifically the problem that we want to address.

1.5 Problem Context

In this research, our aim is to build a machine learning model with mainly demographic information combined with a novel deprivation index to determine the vulnerability of a person to both diabetes and high blood pressure.

According to Center for Medicare and Medicaid, it is estimated that about 12% of the adult population in El Paso, Texas are currently living with diabetes while about 31.6% in the city are living with high blood pressure [5] which consequently have a negative economic impact on the county. CDC estimated that HBP costs the health care system $214 billion per year and results in $138 billion lost productivity on the job [3]. Similarly, the total estimated cost of diagnosed diabetes was $327 billion in medical costs and lost productivity in 2017.

It has been established that delayed diagnosis and poor management of chronic diseases are common which mainly contribute to adverse effects on the patients and society at large. This has led to several works that have been carried out on early detection with most of them using health data and a few combining health data with minimal demographic data of patients. This research aims to establish that prevalent socio-economic conditions contribute largely to many chronic diseases especially for border towns with large concentration of immigrants and economically disadvantaged people.

So, let us assume we have a set of the two chronic diseases, CD $= \{CD_d, CD_{hbp}\}$, where CD_d is diabetes disease and CD_{hbp} is high blood pressure disease. Also,

let Y be a two-dimensional dataset, where every row in it represents a vector v with information of respondents, such that $v = \{x_{demo}, x_{health}\}$, where x_{demo} are the demographic features and x_{health} are the health features. Lastly, assume for every zip code in a reference area e.g. county, state, country etc., we have a deprivation index, DI.

Now that we have our problem formulated, let us discuss our approach to address it.

1.6 Research Objectives and Novelty Statement

Our goal then, is to build a machine learning model M with only the demographic features i.e., $v = \{x_{demo}\}$, to predict occurrence of CD. We also want to know how M performs when $v = \{x_{demo}, DI\}$, especially in border towns like El Paso.

The novelty and the contributions of this research are hinged on the fact that it provides answers to the following two questions.

1. What is the effect of the proposed deprivation index on the performances of our selected machine learning algorithms in predicting the occurrence of chronic diseases?

2. What is the effect of using demographic data for predicting chronic diseases compared to health related data? [39, 46, 49]

In addition, we would also like to know how the index compares to existing index like the Townsend index. The proposed approach in this study to enable us answer the above questions includes the following;

- Selecting appropriate datasets and gaining a good understanding of them through exploratory data analysis.
- Efficiently cleaning the datasets and replacement of the missing values in them.

- Employing feature engineering on the datasets and selection of the best features for the algorithms.
- Construction of the proposed deprivation index [79] and adding it as a feature to the dataset.
- Using classification algorithms to train the models and comparing their performances with and without the proposed deprivation index.
- Study whether demographics or health data provide better insights into patients' susceptibilities to the two chronic diseases.

1.7 Organization

The rest of the paper is organized as follows;

Chapter 2 presents some of the relevant work in this research area. It provides relevant theoretical frameworks on prediction of chronic diseases with machine learning and helps us to understand the previous usage of deprivation index in this field.

Chapter 3 gives a detailed description of the methodology adopted in this work. It includes a detailed description of all the machine learning models employed for all the predictions. It also explains the concepts for constructing the new deprivation index introduced in the study.

Chapter 4 describes all the steps taken to process the datasets.

Chapter 5 summarizes the results from the selected machine learning algorithms. It answers the two main research questions in the study.

Chapter 6 concludes the research suggesting possible future directions.

Chapter 2

Literature Reviews

Many past studies have been done on disease prognosis and arguably significant progress has been made on the predictive accuracies of several machine learning algorithms that have been considered. In this chapter, we provide a detailed summary of previous related works that will help us understand the novelty of the model that we have developed in this research and how it performs better.

2.1 Impact of Chronic Diseases

Several works have been done to understand the impact of chronic diseases in a community. Aside from a chronic disease negatively affecting the sick individuals, it can also have severe consequences on the lives of their spouses and other family members [25]. Murray and Lopez [70] pointed out that the impact of a chronic disease can be measured by the disability-adjusted life years (DALY), the years of life lost (YLL) and the years lived with disability (YLD).

Usually, chronic diseases develop slowly, but lasts for a long time and require medical treatment [65]. They deteriorate the overall health of patients resulting in a significant decline in their ability to live well, productivity and health related quality of life. Patrick and Erickson [42] define health-related quality of life as the value assigned to duration of life as modified by the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy. In a similar vein, Devins et al. [40], postulated that chronic disease disrupts an individual's life and that this disruption can be measured by the impact it has on