Disease and ancestry

Abdelrehim Sabri (500998232)

2019-10-14

Introduction

Every ethnic group has a different diet and lifestyle. When people moved to North America, they brought their eating habits with them. The large size and highly diverse population resulted in new types of deceases was not known in North America.

“Diagnosis has important implications for patient care. When a diagnosis is accurate and made in a timely manner, a patient has the best opportunity for positive health outcome because clinical decision making well be tailored to a correct understanding of the patient's health problem.” [1]

My key research question is to define what factors are the relevant in predicting the disease the new patient might have. Also, to predict the type of medications different cities should have to provide them for their patients.

After exploring the data, regression analysis will be conducted. Regression analysis provides detailed insight that can be applied to further improve services. The process of performing a regression allows to confidently determine which factor matter most, which factors can be ignored, and how these factors influence each other.

The source code for this project is available on github at https://github.com/abdelrehim-sabri/capstone

References:

[1] Holmboe and Durning, 2014 https://www.ncbi.nlm.nih.gov/books/NBK338593/#

Literature Review

Several publications were reviewed with the emphasis being placed on determining potential factors which may have significant effects on the type of disease a person might carry based on his/her profile including ancestry.

BMC Medical Genomics noted that it is becoming increasingly difficult to keep information about genetic ancestry separate from information about health, and consumers of genetic ancestry tests are becoming more aware of the potential health risks associated with particular ancestral lineages. Because some of the proposed associations have received little attention from oversight agencies and professional genetic associations, scientific developments are currently outpacing governance regimes for consumer genetic testing. [2]

The journal about "Ethnicity & disease" talks about how ethnicity and disease provides a comprehensive source of information for common illnesses through the study of ethnic patterns of disease.[3]

From the dataset I wanted to create a set of classification rules that make a decision, or predict behavior. To start, a training data is chosen that contains a certain set of attributes as well as their likely outcome. The article "The Use of Classification in Data Mining" discusses the job of the classification algorithm in discovering how the used attributes reaches its conclusion.[4]

Finally, the article "Linear Regression Analysis" explain the importance of regression analysis for the analysis of medical data.[5]

References:

[2] BMC Medical Genomics https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5223458/

[3] https://www.researchgate.net/journal/1049-510X_Ethnicity_disease

[4] https://www.lifewire.com/classification-1019653

[5] PMC, US National Library of Medicine, National Institutes of Health https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018/

Dataset

The data used for this project can be found at https://www.kaggle.com/karimnahas/medicaldata.

The attributes used include: gender, data of birth, employment status, education, marital status, zipcode, children, ancestry, average commute, daily internet use, available vehicles, military service, disease. The attribute that is removed from the study is patient id.

The following table describes them:

No	Attribute	Description	Data Type
1	id	The patient id	Nominal
2	gender	Male or Female	Nominal
3	zipcode	Area code where the patient lives	Numeric
4	Employment_status	retired, employed, unemployed, or student	Nominal
5	education	bachelors, phd/md, masters, or highschool	Nominal
6	Marital_status	married or single	Nominal
7	children	Total number of children the patient has	Numeric
8	ancestry	Original country the patient came from	Nominal
9	Avg_commute	The amount of time the patient uses the commute	Numeric
10	Daily_internet_use	The amount of time the patient uses the internet	Numeric

11	Available_vehicles	Does the patient have vehicles?	Nominal
12	Military_service	Did the patient serve in the military?	Nominal
13	disease	Name of the disease	Nominal

Table 1 Dataset Description

The following chart shows the total number of patients for each disease:

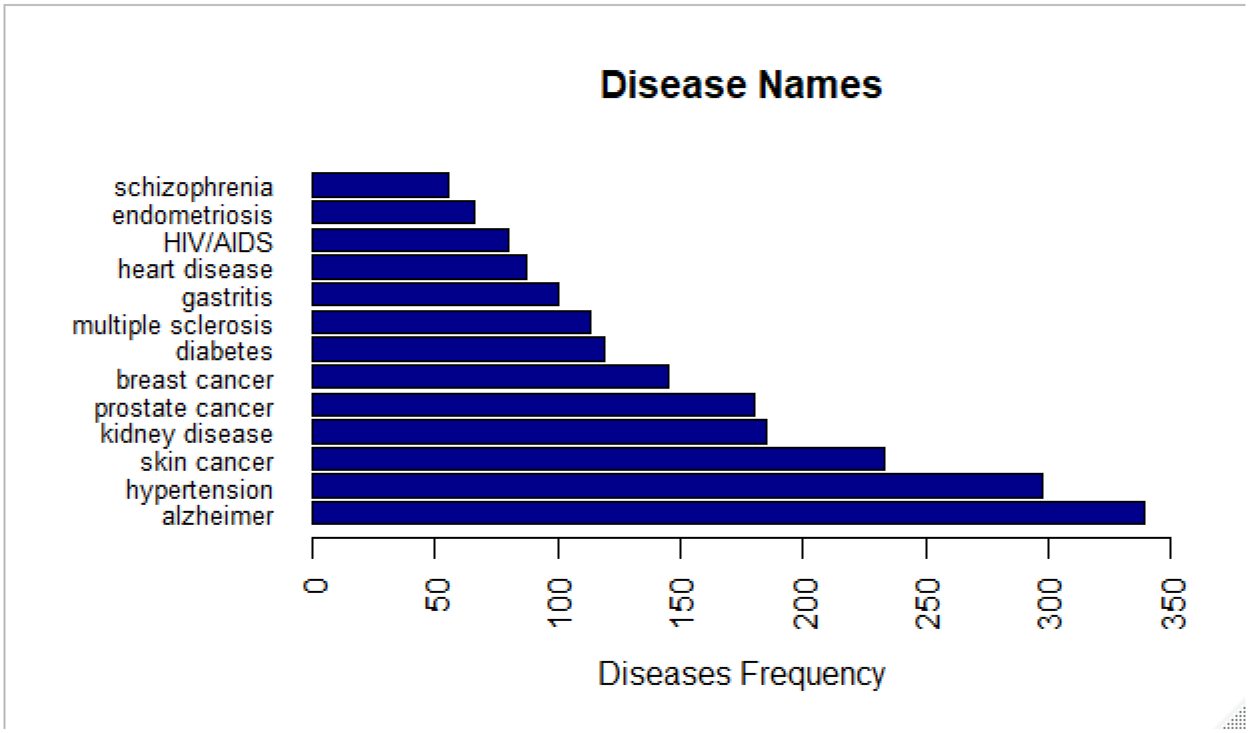
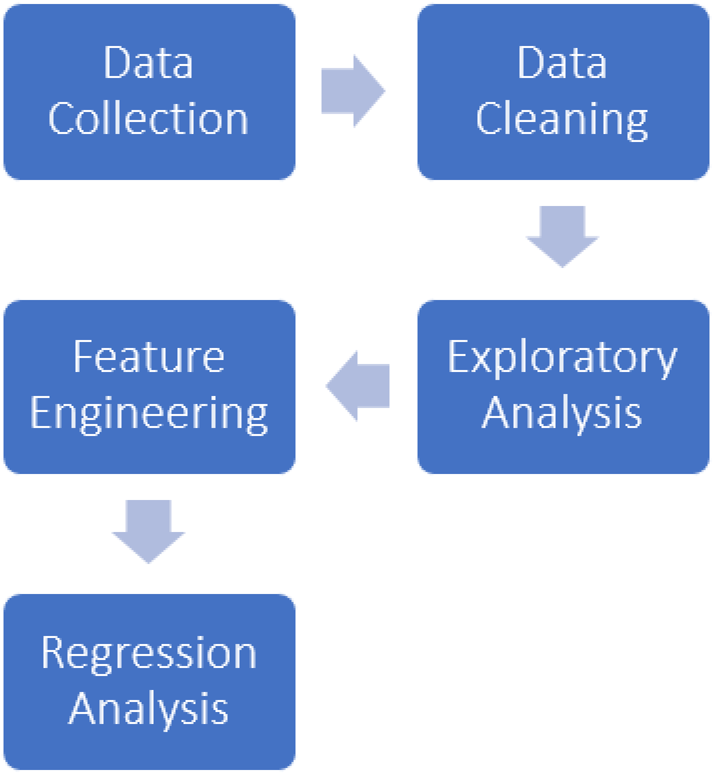


Figure 1: Counts of the various disease names

Approach

The approach to be taken is shown in the graph below, and is described in the following subsections.



Step 1: Data Collection

Import the data into R, specifically into data-frames.

Step 2: Data Cleaning

The data collected is real world data and containing misspelled values that need to be fixed to match field categories. Need to calculate the age through date of birth and group the ages into categories. Also, split the diseases into multiple columns and convert value to binary for easy analysis.

Step 3: Exploratory Analysis

Once the data is cleaned, exploratory analysis can take place. This will include searching for existing correlations in the data as well as identification of attributes that will likely be useful in the regression analysis.

Step 4: Feature Engineering

If required, attributes may need to be re-factored or engineered to provide better inputs to the regression.

Step 5: Regression Analysis

Regression will be performed to build a model to predict diseases given the inputs identified in the previous steps.