Disease Recurrence

Abdelrehim Sabri (500998232)

2019-10-21

Introduction

Every ethnic group has a different diet and lifestyle. When people moved to North America, they brought their eating habits, level of education, and life style with them. The large size and highly diverse population resulted in different types of diseases that effected different age and ethnic groups.

“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]

The research question is to define what factors are relevant to predicting the disease the patient might have. Also, who might be susceptible to which type of disease.

After exploring the data, there is a possibility that a patient might have multiple diseases. Therefore, each instance can be assigned with multiple categories, as such, this type of problem needs to use multi-label classification.

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

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 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 provide a comprehensive source of information for common illnesses through the study of ethnic patterns of disease.[3]

The dataset will be used to create a set of classification rules to predict behavior and support the decision making process. 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]

The book goes over the machine learning terminology, and how to interpret the dataset using different models and algorithms. [5]

Since the dataset is dealing with multi-class classification, this tutorial will show some tips and tricks to improve the multi-class classification results. [6]

Once the model is selected, to measure the effectiveness and performance of the model confusion matrix is used. [7]

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	dob	Date of birth	Nominal
4	zipcode	Area code where the patient lives	Numeric
5	Employment_status	retired, employed, unemployed, or student	Nominal
6	education	bachelors, phd/md, masters, or highschool	Nominal
7	Marital_status	married or single	Nominal
8	children	Total number of children the patient has	Numeric
9	ancestry	Original country the patient came from	Nominal
10	Avg_commute	The amount of time the patient uses the commute	Numeric
11	Daily_internet_use	The amount of time the patient uses the internet	Numeric
12	Available_vehicles	Does the patient have vehicles?	Nominal
13	Military_service	Did the patient serve in the military?	Nominal
14	disease	Name of the disease	Nominal

Table 1 Dataset Description

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

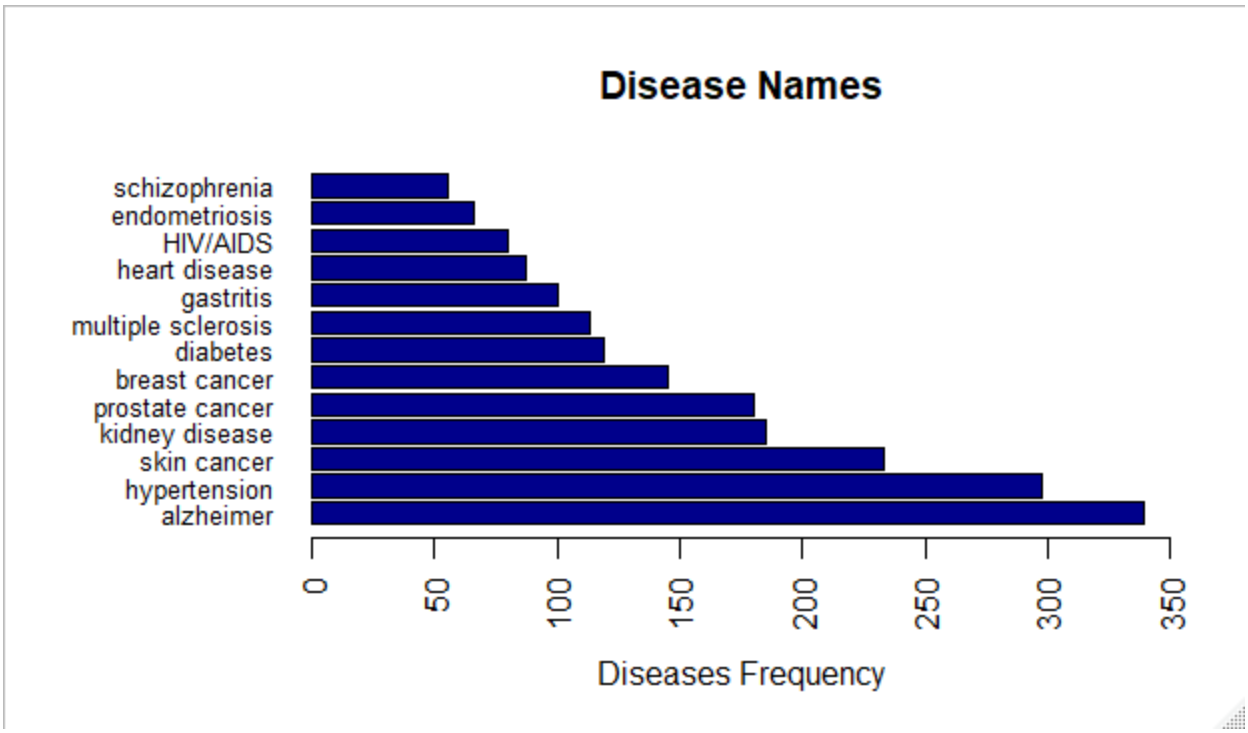

Figure 1: Counts of the various disease names

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.

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 machine learning models.

Step 4: Transformation

Transform the multi-label problem into single-label problem by applying binary relevance, classifier chains and label powerset.

Step 5: Machine Learning Models

Naïve Bayes, k-nearest neighbors, and Random Forest will be performed to build a model to predict diseases given the inputs identified in the previous steps.

Step 6: Conclusion and Recommendations

Pick the best algorithm for the analysis and recommend other attributes and recourses to enhance the study.

References:

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

[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] Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, by Christoph Molnar, 2019-09-18, https://christophm.github.io/interpretable-ml-book/

[6] Tips and Tricks for Multi-Class Classification, https://medium.com/@b.terryjack/tips-and-tricks-for-multi-class-classification-c184ae1c8ffc

[7] Understating Confusion Matrix, https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62