Final Project Proposal: How is Data Science Transforming Health Care and Insurance?

Team KMeans-Plotter

Sujata Vitala

Indira Vadlapatla

Ronald Martis

Satya Madakshira

Ashish Shrivastava

Scope

Use Case	How
Preventative care… Predicting disease outbreaks on both the individual and the community level	• Classification Analysis based on Gender, BMI, Age, RACE for Diabetes
Diagnostic care… Probability model to predict a disease	• Model to predict if person with certain characteristics would likely to have certain disease
Insurance… Adjusting insurance premiums based on publicly available risk factors and predicting insurance premiums based on different attributes of insurance applicants thus doing a better risk analysis. Optimal pricing also ensures more people would buy insurance thus benefiting the general public and companies.	• Apply ML to perform clustering

Available Datasets

- https://www.data.gov/health/ (Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) - FY2011)

Columns in this Dataset

Column Name	Description	Type	
DRG Definition	Code and description identifying the DRG. DRGs are a class...	Plain Text	T
Provider Id	Provider Identifier billing for inpatient hospital services.	Number	#
Provider Name	Name of the provider.	Plain Text	T
Provider Street Address	Street address in which the provider is physically located.	Plain Text	T
Provider City	City in which the provider is physically located.	Plain Text	T
Provider State	State in which the provider is physically located.	Plain Text	T
Provider Zip Code	Zip code in which the provider is physically located.	Number	#
Hospital Referral Region Description	HRR in which the provider is physically located.	Plain Text	T
Total Discharges	The number of discharges billed by the provider for inpatie...	Number	#
Average Covered Charges	The provider's average charge for services covered by Med...	Money	#
Average Total Payments	The average of Medicare payments to the provider for the ...	Money	#
Average Medicare Payments		Money	#

Show Less

Available Datasets

▶

Large Health Data Sets

Air Quality Statistics from EPA Data - findthedata.org

BRFSS - Behavioral Risk Factor Surveillance System (US federal)

Birtha - Vitalnet software for analyzing birth data (Business)

CDC Wonder - Public health information system (US federal)

CMS - The Centers for Medicare and Medicaid Services. US federal.

DATA2010 - Healthy People 2010 monitoring system. US federal.

data.world - Requires password. Seems like small data sets.

Data Warehousing Information Center - Decision support systems.

Epigram - Vitalnet software for analyzing mortality data. Business.

Eurocat - European Surveillance of Congenital Anomalies.

Ferret - Access large data sets. US federal.

Florida CHARTS - Florida Community Health Assessment Resource Tool Set.

HCUP - Healthcare Cost and Utilization Project. Discharge data. US federal.

HCUPnet - Statistics and trends about hospital stays. US federal.

Health Information Tennessee (HIT) - Tennessee health data.

Health Data All Stars - Health Data Consortium

HIW - Health Indicators Warehouse (DHHS)

Human Mortality Database - Mortality and population data.

ICD-10 - International Classification of Diseases.

IPUMS - Minnesota Population Center

MassCHIP - Massachusetts Community Health Information Profile.

MedTrend - Vitalnet software for analyzing hospital discharge data. Business.

Missouri MICA - Missouri Information for Community Assessment.

NAACCR - North American Association of Central Cancer Registries. Non-profit.

NAHDO - National Association of Health Data Organizations. US non-profit.

Available Datasets

▶ https://datacatalog.worldbank.org/dataset/health-nutrition-and-population-statistics (HealthStats provides key health, nutrition and population statistics gathered from a variety of international sources. Themes include population dynamics, nutrition, reproductive health, health financing, medical resources and usage, immunization, infectious diseases, HIV/AIDS, DALY, population projections and lending. HealthStats also includes health, nutrition and population statistics by wealth quintiles.)

Overview	Data & Resources	Visualization

API
Health Nutrition and Population Statistics
Resource Type: **API Documentation**

Go to Resource

CSV file
Health Nutrition and Population Statistics
Resource Type: **Download**   Format: **CSV ZIP**

Go to Resource

Databank
Health Nutrition and Population Statistics
Resource Type: **Query Tool**

Go to Resource

Excel file
Health Nutrition and Population Statistics
Resource Type: **Download**   Format: **EXCEL ZIP**

Go to Resource

Dataset for Risk Analysis & Prediction

- https://www.kaggle.com/c/prudential-life-insurance-assessment/data

- The dataset is provided as a part of the Prudential Life Assessment competition on Kaggle.

- By developing a predictive model that accurately classifies risk using a more automated approach, both public and company can benefit from correct premiums.

Implementation Options

- Sci-Kit Learn, ML or Deep Learning learning libraries

- **Other Considerations**
  - Python Pandas
  - Python Matplotlib
  - JavaScript Plotly
  - JavaScript Leaflet
  - MySQL Database
  - Tableau