Case Studies

In

Healthcare Analytics

FORE School of Management, New Delhi

Contents

About this Document

Today, Healthcare Sector is using Machine Learning and Deep Learning on a very large scale. These technologies handover to their masters, predictive capabilities with accuracy that could not have been imagined a few years back. Given the data and given the knowledge of these techniques and related tools, Healthcare providers are using them in ways that are totally innovative. A few case studies in this compilation provide a bird's eye-view of how some of these institutions are striving to utilise the analytical powers of Machine Learning to their advantage. Compilation lists a number of problems that these industries posed to data science community for solution at various times. Both, the problems posed and the dataset released are explained. Such massive use of this technology is essential for them to continue to serve society in timely and effective manner.

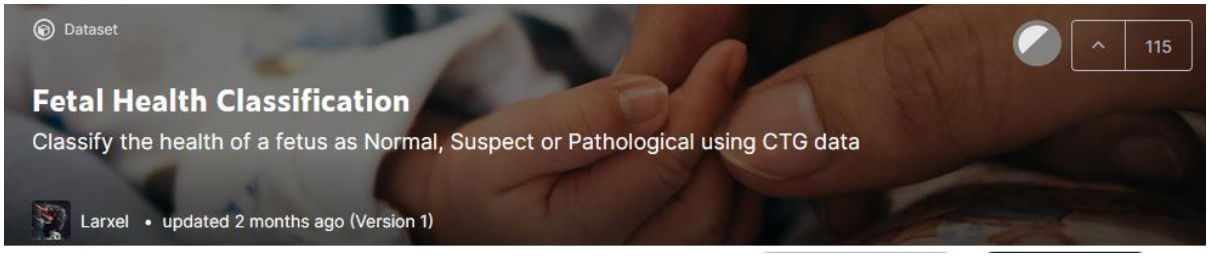
Datasets are from multiple sources: some from Kaggle, some from Open Access Series of Imaging Studies (OASIS), some from Stanford University, Centre for Artificial Intelligence in Medicine & Imaging, some from UCI, Machine Learning Repository, a few from personal collection and other sources.

In our course we will be going through the challenges mentioned in this Case Study Document either as class-work, or as part of students' exercises or as Capstone Project.

Last amended: 5th May, 2021
My folder: C:\Users\Administrator\OneDrive\Documents\health_care

1.Classify fetal health in order to prevent child and maternal mortality

Ref: https://www.kaggle.com/andrewmvd/fetal-health-classification



Context

Reduction of child mortality is reflected in several of the United Nations' Sustainable Development Goals and is a key indicator of human progress.
The UN expects that by 2030, countries end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce under-5 mortality to at least as low as 25 per 1,000 live births.

Parallel to notion of child mortality is of course maternal mortality, which accounts for **295 000 deaths** during and following pregnancy and childbirth (as of 2017). The vast majority of these deaths **(94%)** occurred in low-resource settings, and most **could have been prevented**.

In light of what was mentioned above, **Cardiotocograms (CTGs)** are a simple and cost accessible option to assess fetal health, allowing healthcare professionals to take action in order to prevent child and maternal mortality. The equipment itself works by sending ultrasound pulses and reading its response, thus shedding light on fetal heart rate (FHR), fetal movements, uterine contractions and more.

Data

This dataset contains **2126** records of features extracted from Cardiotocogram exams, which were then classified by three expert obstetritians into **3 classes**:

- Normal
- Suspect
- Pathological

Problem

- Create a multiclass model to classify CTG features into the three fetal health states.

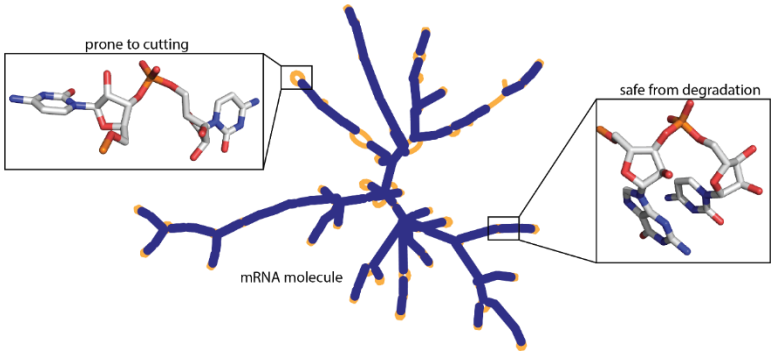
2. COVID-19 mRNA Vaccine Degradation Prediction

Ref: https://www.kaggle.com/c/stanford-covid-vaccine

Winning the fight against the COVID-19 pandemic will require an effective vaccine that can be equitably and widely distributed. Building upon decades of research has allowed scientists to accelerate the search for a vaccine against COVID-19, but every day that goes by without a vaccine has enormous costs for the world nonetheless. We need new, fresh ideas from all corners of the world. Could online gaming and crowdsourcing help solve a worldwide pandemic? Pairing scientific and crowdsourced intelligence could help computational biochemists make measurable progress.

mRNA vaccines have taken the lead as the fastest vaccine candidates for COVID-19, but currently, they face key potential limitations. One of the biggest challenges right now is how to design super stable messenger RNA molecules (mRNA). Conventional vaccines (like your seasonal flu shots) are packaged in disposable syringes and shipped under refrigeration around the world, but that is not currently possible for mRNA vaccines.

Researchers have observed that RNA molecules have the tendency to spontaneously degrade. This is a serious limitation--a single cut can render the mRNA vaccine useless. Currently, little is known on the details of where in the backbone of a given RNA is most prone to being affected. Without this knowledge, current mRNA vaccines against COVID-19 must be prepared and shipped under intense refrigeration, and are unlikely to reach more than a tiny fraction of human beings on the planet unless they can be stabilized.



The Eterna community, led by Professor Rhiju Das, a computational biochemist at Stanford's School of Medicine, brings together scientists and gamers to solve puzzles and invent medicine. Eterna is an online video game platform that challenges players to solve scientific problems such as mRNA design through puzzles. The solutions are synthesized and experimentally tested at Stanford by researchers to gain new insights about RNA molecules. The Eterna community has previously unlocked new scientific principles, made new diagnostics against deadly diseases, and engaged the world's most potent intellectual resources for the betterment of the public. The Eterna community has advanced

biotechnology through its contribution in over 20 publications, including advances in RNA biotechnology.

In this competition, we are looking to leverage the data science expertise of the Kaggle community to develop models and design rules for RNA degradation. Your model will predict likely degradation rates at each base of an RNA molecule, trained on a subset of an Eterna dataset comprising over 3000 RNA molecules (which span a panoply of sequences and structures) and their degradation rates at each position. We will then score your models on a second generation of RNA sequences that have just been devised by Eterna players for COVID-19 mRNA vaccines. These final test sequences are currently being synthesized and experimentally characterized at Stanford University in parallel to your modeling efforts -- Nature will score your models!

Improving the stability of mRNA vaccines was a problem that was being explored before the pandemic but was expected to take many years to solve. Now, we must solve this deep scientific challenge in months, if not weeks, to accelerate mRNA vaccine research and deliver a refrigerator-stable vaccine against SARS-CoV-2, the virus behind COVID-19. The problem we are trying to solve has eluded academic labs, industry R&D groups, and supercomputers, and so we are turning to you. To help, you can join the team of video game players, scientists, and developers at Eterna to unlock the key in our fight against this devastating pandemic.

Problem

> We will be predicting the degradation rates at various locations along RNA sequence. There are multiple ground truth values provided in the training data. While the submission format requires all 5 to be predicted, only the following are scored: `reactivity`, `deg_Mg_pH10`, and `deg_Mg_50C`.

Files

- **train.json** - the training data
- **test.json** - the test set, without any columns associated with the ground truth.
- **sample_submission.csv** - a sample submission file in the correct format

3.Predict a biological response of molecules from their chemical properties

Ref: https://www.kaggle.com/c/bioresponse

The objective of the problem is to help build as good a model as possible so that one can, as optimally as this data allows, relate molecular information, to an actual biological response.

The data has been shared in the comma separated values (CSV) format. Each row in this data set represents a molecule. The first column contains experimental data describing an actual biological response; the molecule was seen to elicit this response (1), or not (0). The remaining columns represent molecular descriptors (d1 through d1776), these are calculated properties that can capture some of the characteristics of the molecule - for example size, shape, or elemental constitution. The descriptor matrix has been normalized.

The data is in the comma separated values (CSV) format. Each row in this data set represents a molecule. The first column contains experimental data describing a real biological response; the molecule was seen to elicit this response (1), or not (0). The remaining columns represent molecular descriptors (d1 through d1776), these are caclulated properties that can capture some of the characteristics of the molecule - for example size, shape, or elemental constitution. The descriptor matrix has been normalized.

The data is in the comma separated values (CSV) format. Each row in this data set represents a molecule. The first column contains experimental data describing a real biological response; the molecule was seen to elicit this response (1), or not (0). The remaining columns represent molecular descriptors (d1 through d1776), these are caclulated properties that can capture some of the characteristics of the molecule - for example size, shape, or elemental constitution. The descriptor matrix has been normalized.

4.Help develop safe and effective medicines by predicting molecular activity

Ref: https://www.kaggle.com/c/MerckActivity/overview
Ref: https://github.com/CathyQian/Data_Science_Projects/tree/master/Predicting_Merck_Molecular_Activity
Ref: https://towardsdatascience.com/predicting-molecular-activity-using-deep-learning-in-tensorflow-f55b6f8457f9

Help enable the development of safe, effective medicines.

When developing new medicines it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. The objective of this competition is to identify the best statistical techniques for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures

The challenge is based on 15 molecular activity data sets, each for a biologically relevant target. Each row corresponds to a molecule and contains descriptors derived from that molecule's chemical structure.

Data Description

> The Training and Test Sets each consist of 15 biological activity data sets in comma separated value (CSV) format. Each row of data corresponds to a chemical structure represented by molecular descriptors.

Training and Test Files

The training files are of the form:

- **Column 1:** Molecule ID
- **Column 2:** Activity. Note that these are raw activity values and different data sets can have activity measured in different units.
- **Column 3-end:** Molecular descriptors/features

The test files are in the same format with Column 2 removed.

Molecule IDs and descriptor names are global to all data sets. Thus, some molecules will appear in multiple data sets, as will some descriptors.

The challenge is to predict the activity value for each molecule/data set combination in the test set. To keep predictions for molecules unique to each data set, a data set identifier has been prepended to each molecule ID (e.g., "ACT1_" or "ACT8_").

Data Set Creation

For each activity, the training/test set split is done by dates of testing. That is, the training set consists of compounds assayed by a certain date, and the test set consists of compounds tested after that date. Therefore, it is expected that the distribution of descriptors will not necessarily be the same between the training and test sets.

5. Can you accurately predict medical insurance costs?

Ref: https://www.kaggle.com/mirichoi0218/insurance

This is a problem in regression, given following data:

About data

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

6. Predict lung function decline—Pulmonary Fibrosis Progression

Ref: https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/data

Imagine one day, your breathing became consistently labored and shallow. Months later you were finally diagnosed with pulmonary fibrosis, a disorder with no known cause and no known cure, created by scarring of the lungs. If that happened to you, you would want to know your prognosis. That's where a troubling disease becomes frightening for the patient: outcomes can range from long-term stability to rapid deterioration, but doctors aren't easily able to tell where an individual may fall on that spectrum. Your help, and data science, may be able to aid in this prediction, which would dramatically help both patients and clinicians.

Current methods make fibrotic lung diseases difficult to treat, even with access to a chest CT scan. In addition, the wide range of varied prognoses create issues organizing clinical trials. Finally, patients suffer extreme anxiety—in addition to fibrosis-related symptoms—from the disease's opaque path of progression.

Open Source Imaging Consortium (OSIC) is a not-for-profit, co-operative effort between academia, industry and philanthropy. The group enables rapid advances in the fight against Idiopathic Pulmonary Fibrosis (IPF), fibrosing interstitial lung diseases (ILDs), and other respiratory diseases, including emphysematous conditions. Its mission is to bring together radiologists, clinicians and computational scientists from around the world to improve imaging-based treatments.

In this problem, you'll predict a patient's severity of decline in lung function based on a CT scan of their lungs. You'll determine lung function based on output from a spirometer, which measures the volume of air inhaled and exhaled. The challenge is to use machine learning techniques to make a prediction with the image, metadata, and baseline FVC as input.

If successful, patients and their families would better understand their prognosis when they are first diagnosed with this incurable lung disease. Improved severity detection would also positively impact treatment trial design and accelerate the clinical development of novel treatments.

7. Predict Possibility of Heart Attack

Ref: https://www.kaggle.com/imnikhilanand/heart-attack-prediction

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no/less chance of heart attack and 1 = more chance of heart attack

Attribute Information

1) age
2) sex
3) chest pain type (4 values)
4) resting blood pressure
5) serum cholestoral in mg/dl
6)fasting blood sugar > 120 mg/dl
7) resting electrocardiographic results (values 0,1,2)
8) maximum heart rate achieved
9) exercise induced angina
10) oldpeak = ST depression induced by exercise relative to rest
11)the slope of the peak exercise ST segment
12) number of major vessels (0-3) colored by flourosopy
13) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14) target: 0= less chance of heart attack 1= more chance of heart attack

8. Classify Pulmonary Embolism cases in chest CT scans

Ref: https://www.kaggle.com/c/rsna-str-pulmonary-embolism-detection/overview

If every breath is strained and painful, it could be a serious and potentially life-threatening condition. A pulmonary embolism (PE) is caused by an artery blockage in the lung. It is time consuming to confirm a PE and prone to overdiagnosis. Machine learning could help to more accurately identify PE cases, which would make management and treatment more effective for patients.

Currently, CT pulmonary angiography (CTPA), is the most common type of medical imaging to evaluate patients with suspected PE. These CT scans consist of hundreds of images that require detailed review to identify clots within the pulmonary arteries. As the use of imaging continues to grow, constraints of radiologists' time may contribute to delayed diagnosis.

The Radiological Society of North America (RSNA®) has teamed up with the Society of Thoracic Radiology (STR) to help improve the use of machine learning in the diagnosis of PE.

In this problem, you'll detect and classify PE cases. In particular, you'll use chest CTPA images (grouped together as studies) and your data science skills to enable more accurate identification of PE. If successful, you'll help reduce human delays and errors in detection and treatment.

With 60,000-100,000 PE deaths annually in the United States, it is among the most fatal cardiovascular diseases. Timely and accurate diagnosis will help these patients receive better care and may also improve outcomes.

9. Healthcare Provider Fraud Detection Analysis

Ref: https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

Provider Fraud is one of the biggest problems facing Medicare. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

Rigorous analysis of Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambiguous diagnosis code is used to adopt costliest procedures and drugs. Insurance companies are the most vulnerable institutions impacted due to these bad practices. Due to this reason, insurance companies increased their insurance premiums and as result healthcare is becoming costly matter day by day.

Healthcare fraud and abuse take many forms. Some of the most common types of frauds by providers are:

a) Billing for services that were not provided.
b) Duplicate submission of a claim for the same service.
c) Misrepresenting the service provided.
d) Charging for a more complex or expensive service than was actually provided.
e) Billing for a covered service when the service actually provided was not covered.

Problem Statement

> The goal of this project is to " predict the potentially fraudulent providers " based on the claims filed by them. Along with this, we will also discover important variables helpful in detecting the behaviour of potentially fraud providers. further, we will study fraudulent patterns in the provider's claims to understand the future behaviour of providers.

10. Detect Malaria through Infected Cell Images

Ref: https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria

Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes. It is preventable and curable.

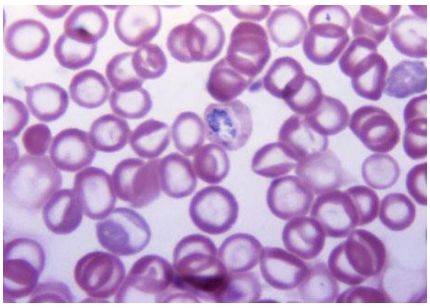
- In 2017, there were an estimated 219 million cases of malaria in 90 countries.
- Malaria deaths reached 435 000 in 2017.
- The WHO African Region carries a disproportionately high share of the global malaria burden. In 2017, the region was home to 92% of malaria cases and 93% of malaria deaths.

Malaria is caused by Plasmodium parasites. The parasites are spread to people through the bites of infected female Anopheles mosquitoes, called "malaria vectors." There are 5 parasite species that cause malaria in humans, and 2 of these species – P. falciparum and P. vivax – pose the greatest threat.

Diagnosis of malaria can be difficult:

- Where malaria is not endemic any more (such as in the United States), health-care providers may not be familiar with the disease. Clinicians seeing a malaria patient may forget to consider malaria among the potential diagnoses and not order the needed diagnostic tests. Laboratorians may lack experience with malaria and fail to detect parasites when examining blood smears under the microscope.
- Malaria is an acute febrile illness. In a non-immune individual, symptoms usually appear 10–15 days after the infective mosquito bite. The first symptoms – fever, headache, and chills – may be mild and difficult to recognize as malaria. If not treated within 24 hours, P. falciparum malaria can progress to severe illness, often leading to death.

Microscopic Diagnosis



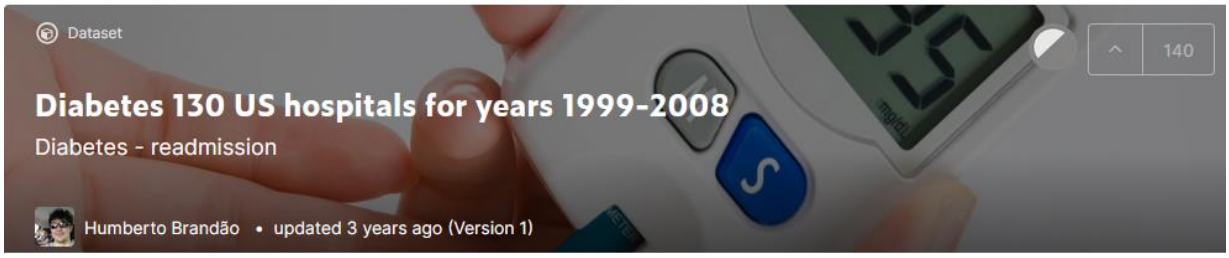
Malaria parasites can be identified by examining under the microscope a drop of the patient's blood, spread out as a "blood smear" on a microscope slide. Prior to examination, the specimen is stained to give the parasites a distinctive appearance. This technique remains the gold standard for laboratory

confirmation of malaria. However, it depends on the quality of the reagents, of the microscope, and on the experience of the laboratorian.

Malaria infected cells can be distinguished from uninfected cells. The problem is to build a trained model built using a collection of infected and un-infected cells. The dataset is from Kaggle but originally from official NIH Website.

11. Predict hospital readmission for diabetes patients

Ref: https://www.kaggle.com/brandao/diabetes



Basic Explanation

It is important to know if a patient will be readmitted in some hospital. The reason is that you can change the treatment, in order to avoid a readmission. In this database, you have 3 different outputs:

1. No readmission;
2. A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);
3. A readmission in more than 30 days (this one is not so good as well the last one, however, the reason can be the state of the patient.

In this context, you can see different objective functions for the problem. You can try to figure out situations where the patient will not be readmitted, or if their are going to be readmitted in less than 30 days (because the problem can the the treatment), etc. Make your choice and let's help them creating new approaches for the problem.

Content

"The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

Problem relates to predicting whether a diabetic patient will seek re-admissions.

12. Predict length of stay in hospital

Ref: https://github.com/vtien/Predicting-Hospital-Stay-Project
Ref: https://www.kaggle.com/jonasalmeida/2015-deidentified-ny-inpatient-discharge-sparcs

Length of stay is a critical indicator of the efficiency of hospital management. Hospitals have limited resources, requiring efficient use of beds and clinician time. For these reasons and more, it is in the best interest of patients, hospitals, and public health to limit any hospital stay to no longer than necessary and to have an idea of how long a given inpatient may need to stay. In this way, the ability to predict how long a patient will stay as soon as they enter the hospital and are diagnosed can have many positive effects for a hospital and its efficiency. Beyond benefits to the hospital, predicting patient's length of stay also greatly benefits the patients and patient's families themselves, as they will have an idea of how long they are expected to stay on day 1 of their visit.

A model that could predict patient length of stay could allow hospitals to better analyse factors such as the procedures, demographics of patients, and others that influence length of stay the most. Such analysis could pave the path for reductions in the length of inpatient stay, which could in turn have the effect of decreased risk of infection and medication side effects, improvement in the quality of treatment, and increased hospital profit with more efficient bed management.

The dataset used in this project can be found on Kaggle here. It contains various descriptions for inpatients at hospitals in the NY area in 2015. Patient data has been de-identified according to HIPAA regulations.

N.Y State: Hospital Inpatient Discharges 2015

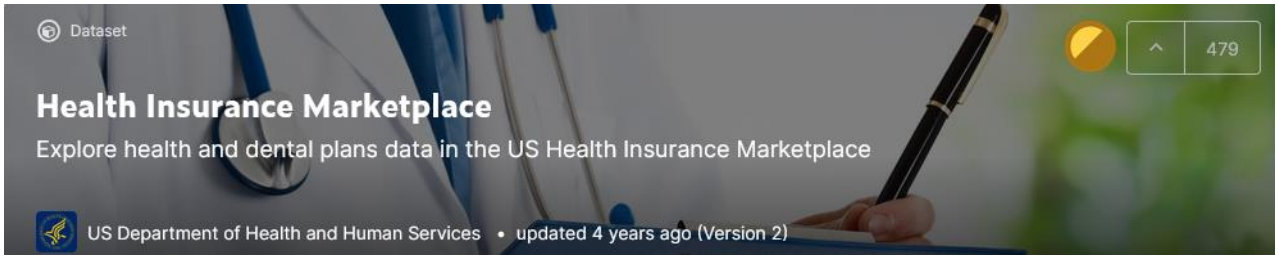
https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8

> The State-wide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. The de-identified data file does not contain data that is protected health information (PHI) under HIPAA. The health information is not individually identifiable; all data elements considered identifiable have been

redacted. For example, the direct identifiers regarding a date have the day and month portion of the date removed.

13. Explore health and dental plans data in the US Health Insurance Marketplace

Ref: https://www.kaggle.com/hhs/health-insurance-marketplace



Description

The Health Insurance Marketplace Public Use Files contain data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace.

Exploration Ideas

To help get you started, here are some data exploration ideas:

- How do plan rates and benefits vary across states?
- How do plan benefits relate to plan rates?
- How do plan rates vary by age?
- How do plans vary across insurance network providers?

See this forum thread for more ideas, and post there if you want to add your own ideas or answer some of the open questions!

Data Description

This data was originally prepared and released by the Centers for Medicare & Medicaid Services (CMS). Please read the CMS Disclaimer-User Agreement before using this data.

Here, we've processed the data to facilitate analytics. This processed version has three components:

1. Original versions of the data

The original versions of the 2014, 2015, 2016 data are available in the "raw" directory of the download and "../input/raw" on Kaggle Scripts. Search for "dictionaries" on this page to find the data dictionaries describing the individual raw files.

2. Combined CSV files that contain

In the top level directory of the download ("../input" on Kaggle Scripts), there are six CSV files that contain the combined at across all years:

- **BenefitsCostSharing.csv**
- **BusinessRules.csv**
- **Network.csv**
- **PlanAttributes.csv**
- **Rate.csv**
- **ServiceArea.csv**

Additionally, there are two CSV files that facilitate joining data across years:

- **Crosswalk2015.csv** - joining 2014 and 2015 data
- **Crosswalk2016.csv** - joining 2015 and 2016 data

3. SQLite database

The "database.sqlite" file contains tables corresponding to each of the processed CSV files. The code to create the processed version of this data is available on GitHub.

14. Identify acute intracranial hemorrhage and its subtypes

Ref: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/data

Intracranial hemorrhage, bleeding that occurs inside the cranium, is a serious health problem requiring rapid and often intensive medical treatment. For example, intracranial hemorrhages account for approximately 10% of strokes in the U.S., where stroke is the fifth-leading cause of death. Identifying the location and type of any hemorrhage present is a critical step in treating the patient.

Diagnosis requires an urgent procedure. When a patient shows acute neurological symptoms such as severe headache or loss of consciousness, highly trained specialists review medical images of the patient's cranium to look for the presence, location and type of hemorrhage. The process is complicated and often time consuming.

In this problem, the challenge is to build an algorithm to detect acute intracranial hemorrhage and its subtypes.

You'll develop your solution using a rich image dataset provided by the Radiological Society of North America (RSNA®) in collaboration with members of the American Society of Neuroradiology and MD.ai.

If successful, you'll help the medical community identify the presence, location and type of hemorrhage in order to quickly and effectively treat affected patients.

15. Predict the onset of diabetes based on diagnostic measures

Ref: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Content

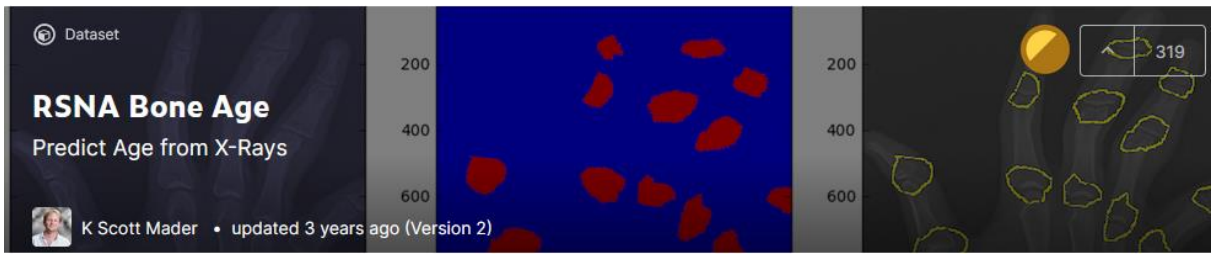
The datasets consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Problem

Can you build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

16. Predict Age from X-rays

Ref: https://www.kaggle.com/kmader/rsna-bone-age



Context

At RSNA 2017 there was a contest to correctly identify the age of a child from an X-ray of their hand. This is the dataset on Kaggle making it easier to experiment with and do educational demos. Additionally maybe there are some new ideas for building smarter models for handling X-ray images.

Content

A number of folders full of images (digital and scanned) with a CSV containing the age (what is to be predicted) and the gender (useful additional information)

Data sets used in the Pediatric Bone Age Challenge have been contributed by Stanford University, the University of Colorado and the University of California - Los Angeles.

The MedICI platform (built CodaLab) used for the challenge is provided by Jayashree Kalpathy-Cramer, supported through NIH grants (U24CA180927) and a contract from Leidos.

Problem

- Can you predict with better than 4.2 months accuracy?
- Is identifying the joints an important step?
- What algorithms work best?
- What do the algorithms focus on?
- Is gender a necessary piece of information or can it be automatically derived from the image?

17. Predict if an infant is likely to develop autistic tendencies

Ref: https://www.kaggle.com/fabdelja/autism-screening-for-toddlers

Domain: Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of toddlers that contained influential features to be utilised for further analysis especially in determining autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioural features (Q-Chat-10) plus other individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science.

Data Type: Predictive and Descriptive: Nominal / categorical, binary and continuous
Task: Classification but can be used for clustering and association or feature assessment
Attribute Type: Categorical, continuous and binary
Area: Medical, health and social science
Missing values? No
Number of Instances (records in your data set): 1054
Number of Attributes (fields within each record): 18 including the class variable
Attribute Information: For Further information about the attributes/feature see doc file.

18. Predict severity of epileptic seizure

Ref: https://www.kaggle.com/chaditya95/epileptic-seizures-dataset

The original dataset from the reference consists of 5 different folders, each with 100 files, with each file representing a single subject/person. Each file is a recording of brain activity for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Each data point is the value of the EEG recording at a different point in time. So we have total 500 individuals with each has 4097 data points for 23.5 seconds.

We divided and shuffled every 4097 data points into 23 chunks, each chunk contains 178 data points for 1 second, and each data point is the value of the EEG recording at a different point in time. So now we have 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y {1,2,3,4,5}.

- The response variable is y in column 179, the Explanatory variables X1, X2, …, X178
- y contains the category of the 178-dimensional input vector. Specifically y in {1, 2, 3, 4, 5}:
    - 5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open
    - 4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed
    - 3 - Yes they identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area
    - 2 - They recorder the EEG from the area where the tumor was located
    - 1 - Recording of seizure activity

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure. Our motivation for creating this version of the data was to simplify access to the data via the creation of a .csv version of it. Although there are 5 classes most authors have done binary classification, namely class 1 (Epileptic seizure) against the rest.

19. Detect Autism from a facial image

Ref: https://www.kaggle.com/gpiosenka/autistic-children-data-set-traintestvalidate

Context

The dataset was collected by Gerry, as an individual effort. He is a retired Director Satcom at General Dynamics, Scottsdale, Arizona, United States. This is what he states about data collection:

> "After reading medical journals that provide research on facial morphology and its correlation to Autistic diagnosis I became interested in the subject of Autism. The methods used by researchers involved physical measurement of facial features and then data analysis to try to correlate these measurements with the presence of Autism. This physical measurement process is very time consuming and thus expensive and is prone to measurement errors.Consequently that approach is of no practical diagnostic value. I speculated that a much better approach would be to gather a data set of images of children with Autism and without Autism. Thus began a long and laborious process of gather the needed images. I searched for existing databases and was not able to find any. I then e-mailed many Autism based groups to solicit their assistance to gather the needed images. I was unsuccessful in gaining any cooperation. These organizations are reluctant to participate with private individuals that are not associated with medical research organizations or university sponsored projects. As a lone independent researcher my requests were rejected. I was then forced to attempt to gather the images via internet searches. To do this I visited many websites and Facebook pages associated with Autism and if images were available of Autistic children I downloaded those images. I was able to gather about 1500 such images. Unfortunately they are not of the best quality or consistency with respect to the facial alignment and perspective or image size, To the extent possible I processed these images and developed a python program to automatically crop the images to only include to the extent possible a facial image. I also tried to somewhat align these images but was not successful.
> For the non-autistic images I simply download random images of children and likewise cropped the images. Data shows that roughly 1% of children have some degree of Autism. Since there is no way to search for"non-autistic" children images I assume that in the images I collected as non-Autistic there are in fact some images of children with Autism. At a 1% rate I doubt these errors have much impact on the results.
> Then I set about creating a CNN classifier on the data set. I had already developed a General Purpose CNN with multiple model options. The model I used is derived from transfer learning using MobileNet V1. The best accuracy I have achieved to date is about 93% accuracy on the test set. My minimum goal is to achieve 95% accuracy with a low false alarm rate. I believe this would be achievable with a large and higher

quality data set. I put the data set and kernel on Kaggle with the hope others would become interested,(hopefully including some medical researchers) and perhaps contribute to improving the data set.
If I can achieve the required accuracy my plan is to develop an on line web application whereby parents could submit one (or better yet several) images of their child and receive a returned probability of the potential of Autism. This used in association with existing diagnostic questionnaires would provide a high accuracy screening for Autism. Parents could then be motivated to seek a complete medical diagnostic analysis.
Early detection of Autism is extremely important to the child's development and I hoped development of this classifier with it's
associated data set would enhance that happening."

Content

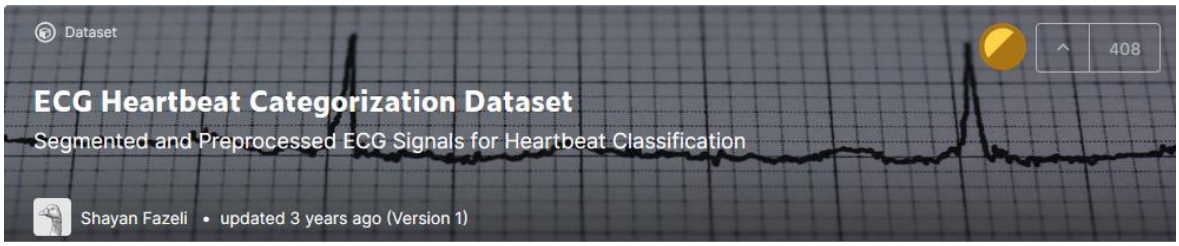
The data set is provided to be used in two ways, both standard for machine learning tasks. One standard method is that I have divided the data into Training, Test and Validation sets. The training set is labeled as train. It consists of two sub directories, Autistic and Non*Autistic. The Autistic sub directory consists of 1667 facial images of Autistic children in 224 X 224 X 3, jpg format. The Non*Autistic sub directory contains 1667 images of children I "assume" to not have Autism also in 224 X 224 X 3, jpg format.
A note for those who use Keras-flow from directory, The images in all sub directories are sequentially numerically labeled (001.jpg, 002.jpg etc). The "zeros" is provides so that when flow-from-directory is used the file order is preserved. This makes it fairly easy to correlate a test image prediction to the associated image file.
The validation images are located in the valid directory. It likewise is separated into 50 images of autistic children and 50 images of non-autistic children in the same format as for the training set.. An important note here, My CNN monitors the validation loss and save the model with the lowest validation loss and used that model to conduct predictions on the test set. Since model selection is based on the validation images I selected the validation images to be those of the "best" quality in terms of image fidelity, facial orientation etc. The test data is located in the test directory similarly divided into 100 images of autistic children and 100 images of non-autistic children.
The second way the data is provided is in the consolidated directory. This directory also has the two sub directories of Autistic and Non-Autistic. It represents the consolidation of the files from the train, test and valid directories into a single set. Users then can partition the consolidated data into their own train, test and validation sets.

Problem

One is of course interested in kernels that can achieve accuracy above the 95% accuracy. Equally import is if you find any errors in the data set such as duplication between training images and testing images. Most importantly, it is hoped to find others interested sufficiently that they might contribute to augment the data set with additional images primarily of children clinically diagnosed as Autistic.

20. Can you identify myocardial infarction?

Ref: https://www.kaggle.com/shayanfazeli/heartbeat



Abstract

This dataset is composed of two collections of heartbeat signals derived from two famous datasets in heartbeat classification, the MIT-BIH Arrhythmia Dataset and The PTB Diagnostic ECG Database. The number of samples in both collections is large enough for training a deep neural network.

This dataset has been used in exploring heartbeat classification using deep neural network architectures, and observing some of the capabilities of transfer learning on it. The signals correspond to electrocardiogram (ECG) shapes of heartbeats for the normal case and the cases affected by different arrhythmias and myocardial infarction. These signals are pre-processed and segmented, with each segment corresponding to a heartbeat.

Content

Arrhythmia Dataset

- Number of Samples: 109446
- Number of Categories: 5
- Sampling Frequency: 125Hz
- Data Source: Physionet's MIT-BIH Arrhythmia Dataset
- Classes: ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4]

The PTB Diagnostic ECG Database

- Number of Samples: 14552
- Number of Categories: 2
- Sampling Frequency: 125Hz
- Data Source: Physionet's PTB Diagnostic Database

Remark: *All the samples are cropped, downsampled and padded with zeroes if necessary to the fixed dimension of 188.*

Data Files

This dataset consists of a series of CSV files. Each of these CSV files contain a matrix, with each row representing an example in that portion of the dataset. The final element of each row denotes the class to which that example belongs.

Problem

Can you identify myocardial infarction?

21. Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adults

Ref: https://www.kaggle.com/jboysen/mri-and-alzheimers

Context:

The Open Access Series of Imaging Studies (OASIS) is a project aimed at making MRI data sets of the brain freely available to the scientific community. By compiling and freely distributing MRI data sets, we hope to facilitate future discoveries in basic and clinical neuroscience. OASIS is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI)(at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).

Content:

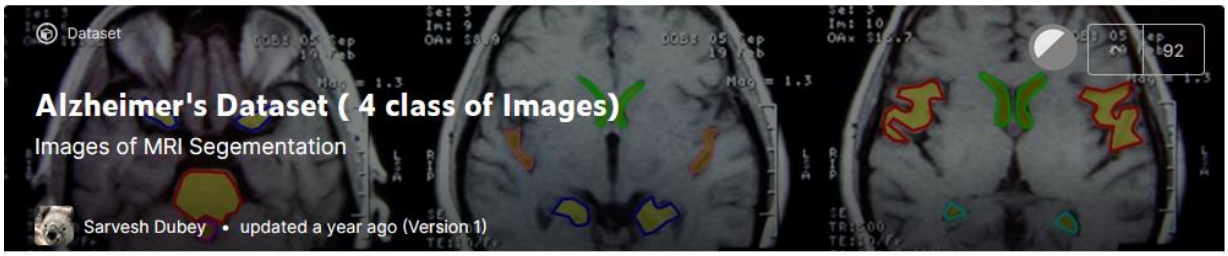
- **Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults**: This set consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). Additionally, a reliability data set is included containing 20 nondemented subjects imaged on a subsequent visit within 90 days of their initial session.
- **Longitudinal MRI Data in Nondemented and Demented Older Adults**: This set consists of a longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Problem

Can you predict dementia? Alzheimer's?

22. Can you create an accurate model to predict the stage of Alzheimers

Ref: https://www.kaggle.com/tourist55/alzheimers-dataset-4-class-of-images



Context

The Data is hand collected from various websites with each and every labels verified.

Content

The data consists of MRI images. The data has four classes of images both in training as well as a testing set:

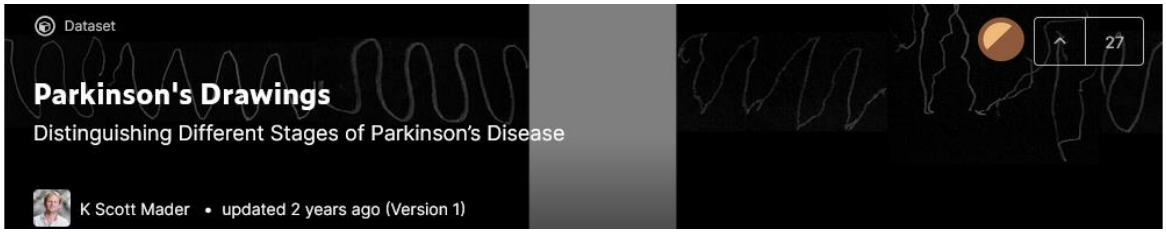
1. Mild Demented
2. Moderate Demented
3. Non Demented
4. Very Mild Demented

Problem

The main inspiration behind sharing this Dataset is to make a very highly accurate model predict the stage of Alzheimers.

23. Distinguishing Different Stages of Parkinson's Disease

Ref: https://www.kaggle.com/kmader/parkinsons-drawings



Content

Images of healthy and patients with Parkinsons drawing spirals and waves. The images are further divided into training and testing groups for comparing (or reproducing) the results of the original publication.

Acknowledgements

The data came from the paper: Zham P, Kumar DK, Dabnichki P, Poosapadi Arjunan S and Raghav S (2017) Distinguishing Different Stages of Parkinson's Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral. Front. Neurol. 8:435. doi: 10.3389/fneur.2017.00435

Problem

- What patterns are present for Parkisons patients?
- Are there small, robust features that can be used for diagnosis?

24. Predict medical insurance costs?

Ref: https://www.kaggle.com/mirichoi0218/insurance

Context

Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

Content

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Acknowledgements

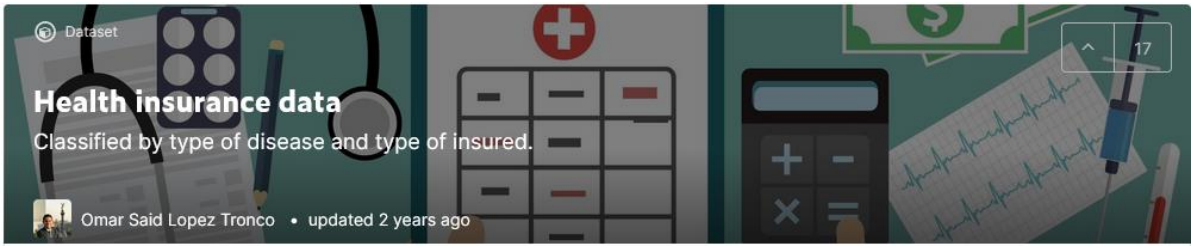
The dataset is available on GitHub here.

Problem

Can you accurately predict insurance costs?

25. Explore Health Insurance Data for costs

https://www.kaggle.com/omartronco/health-insurance-data



Context

This dataset has been created and provided for Actuarial Science students, at ITAM (Mexico).

Content

The data is related to health insurance for a group of insurance policies. This dataset includes also accidents coverage. Even though diseases are hard to classify, this dataset is divided into 3 types: acute, sub-acute and chronic. The group has 5 types of insureds, each group with different coverage modifications.

Comments

The data has been simulated based on expert judgment and designed for easily fitting actuarial models.

26. Forecast sales of drugs using store, promotion, and competitor data

Ref: https://www.kaggle.com/c/rossmann-store-sales

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

In their first Kaggle competition, Rossmann is challenging you to predict 6 weeks of daily sales for 1,115 stores located across Germany. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation. By helping Rossmann create a robust prediction model, you will help store managers stay focused on what's most important to them: their customers and their teams!

Data Files

File Name	Available Formats
sample_submission.csv	.zip (55.25 kb)
store.csv	.zip (8.33 kb)
test.csv	.zip (143.25 kb)
train.csv	.zip (5.66 mb)

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

Files

- **train.csv** - historical data including Sales
- **test.csv** - historical data excluding Sales
- **sample_submission.csv** - a sample submission file in the correct format
- **store.csv** - supplemental information about the stores

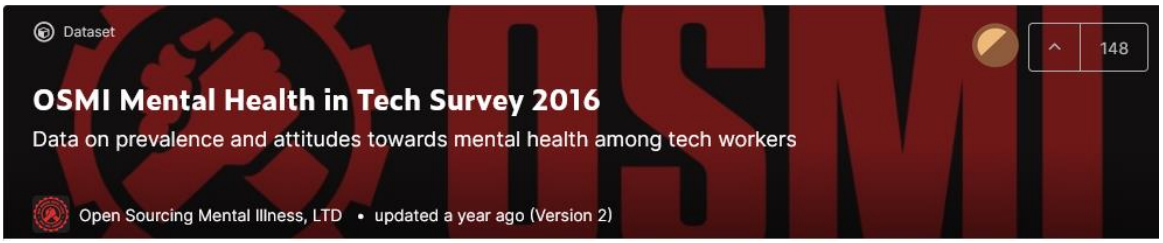
Data fields

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

- **Id** - an Id that represents a (Store, Date) duple within the test set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what you are predicting)
- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

27. Prevalence and attitudes towards mental health among tech workers

https://www.kaggle.com/osmi/mental-health-in-tech-2016



OSMI Mental Health in Tech Survey 2016

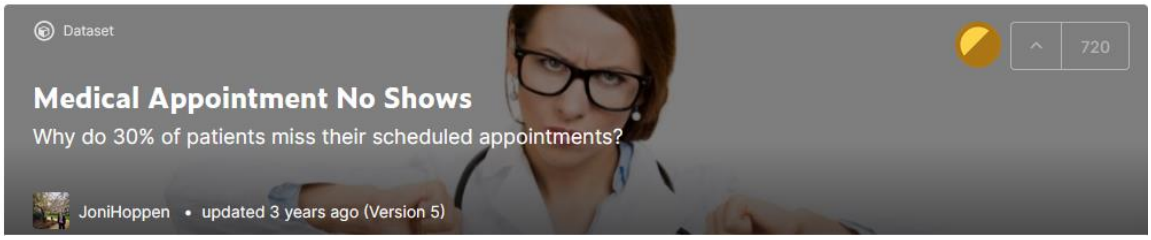
Currently over 1400 responses, the ongoing 2016 survey aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers.

How Will This Data Be Used?

We are interested in gauging how mental health is viewed within the tech/IT workplace, and the prevalence of certain mental health disorders within the tech industry. The Open Sourcing Mental Illness team of volunteers will use this data to drive our work in raising awareness and improving conditions for those with mental health disorders in the IT workplace.

28. Can you predict if a patient will keep his appointment?

https://www.kaggle.com/joniarroba/noshowappointments



Context

A person makes a doctor appointment, receives all the instructions and no-show. Who to blame? Can you predict in advance if a patient will turn up?

Content

110.527 medical appointments its 14 associated variables (characteristics). The most important one if the patient show-up or no-show to the appointment. Variable names are self-explanatory

Data Dictionary

PatientId

- Identification of a patient

AppointmentID

- Identification of each appointment

Gender

- Male or Female . Female is the greater proportion, woman takes way more care of they health in comparison to man.

DataMarcacaoConsulta

- The day of the actuall appointment, when they have to visit the doctor.

DataAgendamento

- The day someone called or registered the appointment, this is before appointment of course.

Age

- How old is the patient.

Neighbourhood

- Where the appointment takes place.

Scholarship

- True of False . Observation, this is a broad topic, consider reading this article https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia

Hipertension

- True or False

Diabetes

- True or False

Alcoholism

- True or False

Handcap

- True or False

SMS_received

- 1 or more messages sent to the patient.

No-show

- True or False.

Problem

What if that possible to predict someone to no-show an appointment?

########## ##################