|  |
| --- |
| **# 1: Classifying clinically actionable genetic mutations**  Problem Statement:   * Once sequenced, a cancer tumor can have thousands of genetic mutations. * The challenge is distinguishing which mutations contribute to tumor growth from those that do not * Currently this interpretation of genetic mutations is done manually by a clinical pathologist and is extremely time consuming, because of the manual review based on clinical literature * Problem statement: can we train a classifier that classifies genetic variations based on an expert-annotated knowledge base of mutation annotations? And how can existing classifications and related medical terms help?   What I will actually be doing:   * **Use NLP to train multiple classifiers (e.g. Naïve Bayes, KNN, SVM) that will perform multinomial classification into 9 classes of genetic mutations, based on interpretation of unstructured text** * Use API: Use the US National Centre for Biotechnology Information (NCBI)’s Entrez (molecular biology database) API (<https://www.ncbi.nlm.nih.gov/clinvar/docs/maintenance_use/#web>) to retrieve **disease** and **severity** information * Use Word2Vec: to build a word vector “dictionary” (?) based on training set corpus * Use web scraping: of the online Biomedical Entity Search Tool (BEST) (<http://best.korea.ac.kr/>) to find entities (diseases, drugs, targets, transcription factors, miRNAs) related to specific genes and variants as inputs; we then apply weights to these entities that are then used for model training   Success Metric(s):   * Best sensitivity and AUC scores   Data Source(s):   * Kaggle training set contains two files with 3,322 rows each   + One file contains the knowledge base, where each row can have more than 80,000 words (>200 MB)   + The other file contains gene, variant and class (target) * Kaggle testing set: 5669 rows; excludes class * Cleaning/munging effort: significant   Challenges and how I plan to overcome them:   * Difficult to find a corpus related to genetic mutations   + Mitigation: build my own from the training set using Word2Vec * Pre-processing of medical acronyms may lead to ambiguity   + Mitigation: compare model performance with and without pre-processing   Is this project reasonable given the time I have: Yes  Example of text:  “Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin.”  Examples: breast-ovarian cancer, renal cell cancer, stomach cancer, colorectal cancer |

|  |
| --- |
| **#2: Understanding mental health factors in the workplace**  Problem Statement:   * To understand the relationship between the frequency of mental health illness and attitudes towards mental health based on the Open Sourcing Mental Illness (OSMI) surveys and the World Happiness Report, so that the WHO and government authorities can shape public policy to improve the mental health of its residents.   What I will actually be doing:   * Combine 4 years’ worth of OSMI data and World Happiness Report data, based on country of participant * My primary aim is to train an appropriate classifier that can predict mental illness based on the OSMI survey answers and World Happiness Report indicators * My secondary aim is to discover any interesting trends   Success Metric(s):   * Sensitivity and AUC   Data Source(s):   * OSMI surveys (yearly data exists from 2016 to 2019) (<https://osmihelp.org/research>) – examples of questions:   + Average of ~740 participants per year   + Questions cover the attitudes of participants, employers and colleagues’ attitudes towards mental health, as well as demographic data such as age, gender, and country of residence * World Happiness Report data on Kaggle (<https://www.kaggle.com/unsdsn/world-happiness>) (2015 to 2019)   + Happiness Score, GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices, Generosity, Perceptions of corruption * Cleaning/munging effort: manageable   Challenges and how I plan to overcome them:   * There is some evidence suggesting that **socio-economic and demographic factors affect subjective well-being in very different ways across cultures and countries at various levels of economic development.**   + Mitigation: The World Happiness Report data should help to account for this |

|  |
| --- |
| **#3: Underlying patterns behind the COVID-19 outbreak in Singapore**  Problem Statement:   * To use contact tracing data to uncover patterns behind the spread of COVID-19 in Singapore, and to compare them with Romania, so that government authorities can take measures to establish travel restrictions and local containment procedures   What I will actually be doing:   * Explore various clustering techniques (e.g. K-Means) to determine if there are any patterns (clusters) within the data * Explore if age has any bearing in the spread of COVID * Explore the use of Neo4J **graph database** to model the data (people and places) to find interesting relationships * Explore the use of network analysis (cf. Zachary’s Karate Club)   Success Metric(s):   * Use Silhouette Coefficient and other unsupervised learning success metrics   Data Source(s):   * Wikipedia page on COVID pandemic in Singapore (<https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Singapore>)   + 200 rows representing confirmed cases as of 13 Mar 2020   + Case Number, date announced, date discharged, age, gender, nationality, case relation, place of stay, hospital admitted to, been to highly affected areas, source * Wikipedia page on COVID pandemic in Romania (population: 20M) (<https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Romania>)   + 123 rows   + Case number, date, status (recovered, hospitalised), source of infection (local vs imported), origin of infection (country), detection location, sex, age, notes (case relation), source * Wikipedia page on COVID pandemic in Ireland (population: 4.83M) * Cleaning/munging effort: manageable   Challenges and how I plan to overcome them:   * Contact tracing dataset is small   + Mitigation: continual update of dataset based on available data (<https://www.gov.sg/article/covid-19-cases-in-singapore>) |

|  |
| --- |
| #4: Predicting dengue cases in Singapore  Problem Statement:   * Can we build a model to predict the number of dengue cases in Singapore for the rest of 2020, based on weather and past patterns of dengue outbreaks, so that government authorities can take pre-emptive measures to control the spread of dengue?   What I will actually be doing:   * Building monthly dataset for weather * Building monthly datasets for Singapore, Malaysia, Indonesia * Build calendar of special events/festivals in Singapore * Develop a predictive model based on regression techniques (Lasso, Ridge, Elastic Net) * Use Epiweeks (<https://pypi.org/project/epiweeks/>) Python package to calculate epidemiological weeks using CDC and ISO week numbering systems   Success Metric(s):   * R2, RMSE   Data Source(s): data.gov.sg datasets:   * Rainfall monthly total (1982 to 2020) * Relative humidity monthly mean (1982 to 2020) * Surface air temperature monthly mean (1982 to 2020) * Vector control data dengue outbreak statistics: yearly dengue incidents (2007 to 2015) * Weekly infection cases for different diseases (e.g. dengue, zika, hand foot mouth disease) from 2012 to 2019 * Yearly cases for Indonesia and Malaysia (Project Tycho <https://www.tycho.pitt.edu/data/>) * Special festivals and events in Singapore   + <https://www.holidify.com/pages/festivals-in-singapore-223.html>   + <https://www.guidemesingapore.com/business-guides/immigration/get-to-know-singapore/singapore-holidays-and-festivals-guide> * Cleaning/munging effort: manageable   Challenges and how I plan to overcome them:   * Working with time series data * Determining if statistics for other insect-based infectious diseases (perhaps from overseas) can be used |

|  |
| --- |
| #5: When can we expect the COVID-19 pandemic to end?  Problem Statement:   * To use available H1N1 and COVID-19 data to predict when the COVID-19 pandemic will peak and end, so that governments can mount an adequate response to contain the spread of the disease   Success Metric(s):  Background:   * An **epidemic** is an event in which a disease is actively spreading. In contrast, the term **pandemic** relates to geographic spread and is used to describe a disease that affects a whole country or the entire world.   Additional things to explore:   * How deadly is COVID-19 (a study about Case Fatality Rate, or death rate) * How infectious is COVID-19 (a study of R0, basic reproduction number) * How soon will people become immune? * Explore different approaches: e.g. SEIR (Susceptible, Exposed, Infected, Resistant)   Data Source(s):   * COVID dataset: <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset> * SARS dataset: [https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset/data#](https://www.kaggle.com/imdevskp/sars-outbreak-2003-complete-dataset/data)   Approach:   * Model R0 (difficult) based on SEIR model – product of:   + probability of infection given contact w infectious person (b)   + contact rate (k)   + infectious duration (d) * Model CFR (death rate) * Adopt smoothing * Use gamma distribution * Tune gamma distribution using GridSearch or RandomizedSearch   Challenges and how I plan to overcome them:   * May only be able to predict R0 for a single city (SEIR generally assumes the whole population has the same infection risk) * The available dataset only has number of suspected, confirmed, deaths and recovered cases over time; it does not show the cases broken down by severity (e.g. susceptible, exposed, resistant)   Sources:   * <https://www.livescience.com/how-deadly-is-coronavirus-covid-19.html> * <https://www.vox.com/science-and-health/2020/3/6/21161234/coronavirus-covid-19-science-outbreak-ends-endemic-vaccine> * Modelling the coronavirus epidemic in a city with Python (<https://towardsdatascience.com/modelling-the-coronavirus-epidemic-spreading-in-a-city-with-python-babd14d82fa2>) * The SIR epidemic model (<https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/>) * Infection Modeling — Part 1 (<https://towardsdatascience.com/infection-modeling-part-1-87e74645568a>) |