Design and Implementation of prediction for Medical Insurance

Problem Statement

The objective of proposed work is to predict the risk for medical insurance and identify those patients most at risk of being re-admitted. It means that patients can have greater support after discharged from hospital.There are some risks types which is required to be predicted for a patient.

- **Cost & Utilization:**The expected cost for a member in the next X years, Predicting emergency department visits. The likelihood of a hospital re-admission for recently discharged patients
- **Clinical:**The likelihood of a person suffering an acute event, such as an acute myocardial infection. The expected disease progression (cost, visit, and/or clinical severity) of someone with a mental illness. A person's expected risk of mortality.
- **Program-focused:**A patient's likelihood to engage in a mobile health program.

An illustration of how healthcare providers can take advantage of machine learning is being able to predict hospital re-admission for chronically ill patients.While the healthcare sector is being transformed by the ability to record massive amounts of information about individual patients, the enormous volume of data being collected is impossible for human beings to analyses. Machine learning provides a way to automatically find patterns and reason about data, which enables healthcare professionals to move to personalized care known as precision medicine. There are many possibilities for how machine learning can be used in healthcare, and all of them depend on having sufficient data and permission to use it. Previously, alerts and recommendations for a medical practice have been developed based on external studies, and hard-coded into their software. However, that can limit the accuracy of that data because they might be from different populations and environments.

Background

Electronic health record is fast becoming the most powerful tool in the medical toolkit. All the information will be stored in the cloud [1-3]. It will have to be because the size of the electronic file containing your complete patient record is estimated to be as much as six terabytes. That's a quarter of the whole of Wikipedia (24Tbs) [4]. A data file that large is required to enable the practice of precision medicine. This a new revolution in healthcare. It is the ability to target healthcare treatment specifically for an individual. In addition to improving health outcomes, precision

medicine will save much money because it is enabled by unique data insights that lead to more targeted treatments. Machine learning is when a computer has been taught to recognize patterns by providing it with data and an algorithm to help understand that data [5-7]. We call the process of learning 'training' and the output that this process produces is called a 'model'. A model can be provided new data and it can reason about this new information based on what it has previously learned.

Methodology

Step 1:Data collection: This will involve collection of student feedback in the form of structured data like the grades, enrollment data, progression rates as well as unstructured data like student opinions expressed through surveys, web blogs, twitter, Facebook etc.

Step 2: Data Preprocessing: In this phase, the data is prepared for the analysis purpose which contains relevant information. Pre-processing and cleaning of data are one of the most important tasks that must be one before dataset can be used for machine learning. The real-world data is noisy, incomplete and inconsistent. So, it is required to be cleaned.

Step 3:Extraction of Feature Set/Training Data

Feature set or training data can be prepared from the cleaned data by using any of the available techniques like bag of words, -gram, N-gram, POS, TOS tagging etc. The training data can also be prepared by providing them labels and then divide it into two classes like positive class and negative class. The feature sets and training set that has obtained by using any of the above methods will be used for the implementation of machine learning algorithms.

Step 4:Implementation of Machine Learning Algorithm on Feature Set/Training Data[7]

Classification: To determine a label or category – it is either one thing or another. We train the model using a set of labelled data. As an example, we want to predict if a person's mole is cancerous or not, so we create a model using a data set of mole scans from 1000 patients that a doctor has already examined to determine whether they show cancer or not. We also feed the model a whole bunch of other data such as a patient's age, gender, ethnicity, and place of residence. Then create a model which will enable us to present a new mole scan & decide if it depict cancer or not.

Regression: A Regression model is created when we want to find out a number – for example how many days before a patient discharged from hospital with a chronic condition such as diabetes will return.

Step 5: Testing of Data

Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

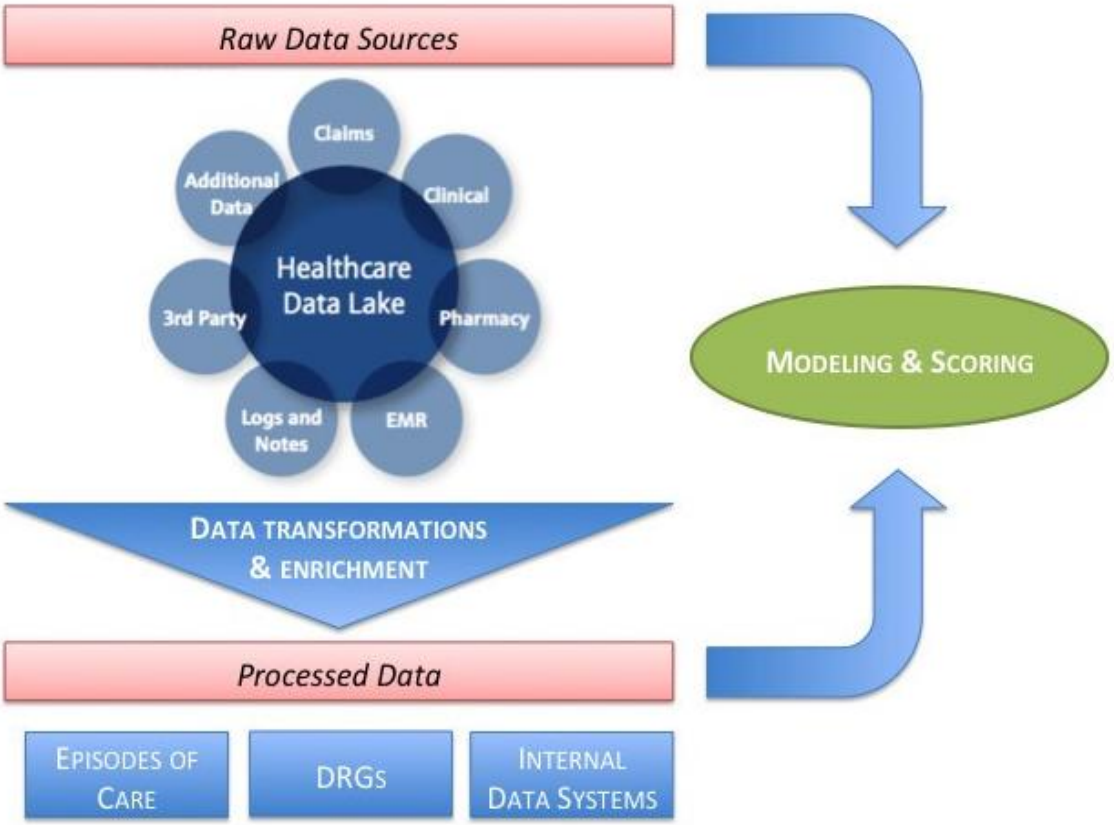

Figure 1: Prediction model for medical insurance [8]

Experimental Design

Dataset:

Links for the Datasets

https://archive.ics.uci.edu/ml/datasets.html?sort=nameUp&view=list

https://www.kaggle.com/c/ClaimPredictionChallenge

https://public.enigma.com/?gclid=EAIaIQobChMIyf35qdXm2gIV0AQqCh19iQYuEAMYASAAEgJduvD_BwE

Evaluation Measures:

- **Accuracy**: Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

- **Precision:** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

- **Recall:** It is the number of correct positive results divided by the number of **all** relevant samples (all samples that should have been identified as positive).

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Software & Hardware Requirements:

Python based Computer Vision and Deep Learning libraries will be exploited for the development and experimentation of the project. Tools such as Anaconda Python, and libraries such as Tensorflow, and Keras will be utilized for this process.

References

[1] Precision Medicine Initiative (NIH). https://www.nih.gov/precision-medicine-initiative-cohort-program.

[2] Lyman, Gary H., and Harold L. Moses. "Biomarker Tests for Molecularly Targeted Therapies--the Key to Unlocking Precision Medicine." The New England journal of medicine375.1 (2016): 4-6.

[3] Collins, Francis S., and Harold Varmus. "A new initiative on precision medicine." New England Journal of Medicine 372.9 (2015): 793-795.

[4] Xu, Rong, Li Li, and QuanQiu Wang. "dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text." BMC bioinformatics 15.1 (2014): 105.

[5] Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333.

[6]Tatonetti, Nicholas P., et al. "Data-driven prediction of drug effects and interactions." Science translational medicine 4.125 (2012): 125ra31-125ra31.

[7] Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." Nature Reviews Genetics 16.6 (2015): 321.

[8] Ganguli, Ishani, et al. "What Do High-Risk Patients Value? Perspectives on a Care Management Program." Journal of general internal medicine (2018): 1-8.