CS 6350 Machine Learning
Project Mid-term Report
Using Machine Learning to predict First Stage of Chronic Kidney Disease

March 13, 2018

1 Team Members

Name: Lama Albarqawi	**uNID:** u6011663
Name: Yulong Liang	**uNID:** u1143816

2 Brief Introduction

Chronic Kidney Disease (CKD) is a condition in which kidney function deteriorates, allowing blood waste to accumulate in and damage the body, CKD progresses gradually and causes a gradual loss of kidney function over time. Progression of CKD stages can be slowed down or prevented by early detection and control of risk factors, such as arterial hypertension and proteinuria, by tight blood pressure control and inhibition of the renin-angiotensin system. To this purpose, early referral to a nephrologist is important to identify patients at risk and provide individualized and comprehensive care aimed to slow disease progression and limit or prevent the occurrence of CKD advanced stages and related complications. CKD has five stages, individuals with the early stages often do not experience noticeable symptoms. However, if left untreated, the disease progresses to kidney failure, at which point the only treatment options are regular and costly dialysis, or kidney transplant.

3 Motivation

Advantages of early prediction includes

- Enhancing the quality of life for the individuals by trying to delay the progression of CKD (e.g. controlling Blood Pressure, losing weight, controlling protein in urine, Low potassium diet, water pill, fix acid levels) and preventing complications of the advanced stages.

- Preventing the need for Dialysis and/or Kidney Transplant. Both of these treatments can result in a huge burden on the patient, physically, psychologically and economically.
- CKD often has no symptoms in its early stages and can go undetected until it is very advanced. That said, early prediction can prevent the need for an emergency, unscheduled dialysis treatment at a hospital, which can cost around $9,900 for a single treatment. In case the patient was not diagnosed earlier and is not aware of his medical condition. Which if was predicted earlier, would dramatically decrease such costs on both individuals and insurance companies.

Goal of CKD Prediction
Raising a red flag for the individuals who are at high risk, informing them with the urgent need to go and see a nephrologist, in order to prevent further complications.

Main indicators for CKD and risk of progression, and tests that can be done to check kidney function status

- Blood Pressure, hypertension: Reduced urination (kidney cleans the blood of wastes and excess fluids). Excess fluids can cause high blood pressure
- Albumin: level of the protein albumin in urine, a high albumin level may indicate kidney disease.
- Diabetes: Blood Sugar level, and Fatigue, loss of appetite, edema
- Anemia, Hemoglobin level, and White and Red blood cells count
- Wastes in blood, and wastes in urine (levels of bacteria, creatinine, sodium, potassium) and Blood urea

Why to use Machine Learning techniques? Why not the traditional or existing methods?

Machine learning is all about developing mathematical and computational methodologies for learning and extracting insights from data and discovering patterns hidden within these data. The more data provided for the machine learning algorithm is the better. Hence, healthcare is a fertile ground for machine learning, since its very rich with patients data. Predicting diseases by just studying bunch of data features (e.g. vital signs and other measures) and trying to discover patterns by studying same features of the already diagnosed patients, is a task that humans or ordinary automated tools would struggle with. Another point is that humans or the programs that the human brain develop, will focus on what the human already knows, and will not search for other signs and indicators, in other words, they will not find new patterns and will only focus on a number of specific signs that they expect a patient at a risk would have.

4 What we have done so far

- Both of us were interested in doing medical related project, so we started searching for a disease that can be predicted by using Machine Learning algorithms, and also in which there would be enough data available for tackling it. We considered Diabetes, Autism, Cardiac arrest and heart failure, and Chronic Kidney Disease. Then we have done a brief research of each of them, for the data availability issue, we have chosen CKD. After that, we have done more focused research on the 5 stages of CKD, their symptoms and indicators, statistics for number of patients in order to measure the importance of such research area, in addition to a brief research on the costs of the treatments and medications. As we stated previously, we have discovered that an early prediction can save lives and cut a huge amount of costs.

- Since healthcare data is not so easy to access, and it needs a process for getting needed permissions to use them in research, we searched for data that are available online for the public, and a data that can be efficiently used for predicting Chronic Kidney Disease purposes. For such project, the data needed should be able to be binary classified, which means that the label space should be either has CKD or has no CKD. After a long search we have found a very interesting dataset provided by UCI, that can be accesses here: `http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease#`.

- We have done some research on what have been done in using Machine Learning for predicting CKD, in order to get an idea on how to start.

- The raw data is segmented with positive example in the first half and negative in second half. Thus, we shuffled the data and split the training and testing data with a rate of 5:3. Since the data is not too much, we might use one random test data strategy in the future.

- The most important thing to think about is how to deal with missing value. For the numeric data, we tried the following strategies:

  1. delete that instance;
  2. delete that instance;
  3. replace with the mean of all the examples;

4. replace with the mean of examples with the same label;

For the categorical data, we tried the following strategies:

1. delete that instance;
2. delete that attribute;
3. replace with the most common value of all the examples;
4. replace with the most common value of examples with the same label;
5. add another category representing missing value;

Since missing data are quite common in the dataset, deleting all the instances with missing value dramatically reduce the training examples. With 60% of the data deleted, the final result is inaccurate. Deleting attribute can be very tricky. We need more work to decide which attribute can be deleted without affecting the final result. Replacing with the mean/most common value of all the examples is as good as with those of examples with the same label. We think it is because this particular dataset. Adding another missing value category also give us a great result.

- For the categorical features, we encoded them using the famous OneHotEncoder method into binary features. For the numeric features, we did the feature scaling to normalize them so that for some classifier, the features will have equal effect.

- We implemented the algorithms that we have already learnt in the class.

  1. **Decision Tree** Decision Tree with no max-depth limit gave us an accuracy of 97-98%. We think the dataset is well collected so that there are few noises, which avoided overfitting.
  2. **Linear Classifier with SGD** Linear Classifier with Stochastic Gradient Descent performed really bad. It gave us an accuracy of 30-60%. We think the data is not linearly separable so that it cannot be well predicted with linear model.
  3. **Linear Classifier with Perceptron** Linear Classifier with Perceptron also performed bad with an accuracy of 30-60%. Thus, in the next step, we might not use linear classifier to build the model. We will use other models or use the kernel trick of SVM to do feature transformation before fit into a linear model.

About the Dataset

The dataset that is available online includes data of 400 patients, 250 of them diagnosed with CKD and the rest (150 patients) are diagnosed with noCKD. for each patient, it has values for 24 features, listed as follows: age,blood pressure, specific gravity, albumin, sugar, red blood cells count, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia. The label is of class CKD, notCKD.

The online dataset had all the records ordered in which patients with label ”ckd” are listed at first, and then it lists the ”noCKD” patients. Since we do not want this order to affect the ML algorithms, we have shuffled the dataset multiple times to get a random listing for patients with respect to their diagnosis. Then we have splitted the dataset again, to get separate datasets for training and testing purposes. We included 250 patients in the training dataset, and 150 for testing.

5 What we are planning to do next

- Explore other algorithms and study their results, after each algorithm implementation we should do a critical comparison to see which algorithms can give us more accurate results, what algorithms can learn the data more efficiently and give more precise predictions. More specifically,

  1. For Decision Tree, we are going to try the assemble version like random forest and xgboost to see if we can improve the accuracy.
  2. For Linear Classifier, since the data might not be linear separable, we are going to try the popular Support Vector Machines with kernel trick.
  3. For other machine learning techniques, we will try k-Nearest Neighbor, Naive Bayes and Neural Network.

- We are planning to study each algorithm by training it using the whole training dataset features and another time by eliminating some features, according to our research on the disease we have found that (based on a medical point of view) there are some specific indicators that have stronger indication for kidney malfunction than others, but we are excited to see the results of the algorithms and what can they learn out of the features. We will also take number of missing values into consideration when eliminate features.