**Machine Learning CS6350**

**Project Report #1**

**Team Members: Lama Albarqawi u6011663**

**Yulong Liang uNID**

**Project Title**: **Using Machine Learning to predict First Stage of Chronic Kidney Disease**

**What’s CKD? What are the advantages of early prediction?**

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 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.

**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 both for 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.

**Indicators for CKD and risk of progression, and tests that can be done to check kidney function status:**

* Blood Pressure and 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
* Anemia (not having enough Red Blood cells) , Hemoglobin level
* Wastes in blood, and wastes in urine (levels of bacteria, creatinine, sodium, potassium)
* Blood urea
* White blood cells count
* Fatigue, loss of appetite, edema

**The motivation. 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 it’s 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.

**What we have done so far**

* We both were interested in doing medical related project, so we started searching for a disease that can be predicted 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 get, 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# . More about this dataset later.
* 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.
* Talk about how we have splitted the dataset into training dataset and testing dataset, based on what?? Take into consideration the number of patients that have ckd and who have not.
* Include something about how we started the ML work , what algorithm we have used so far and why we have chosen it.

**About the Dataset**

Data for 400 patients was collected in an Indian hospital. It includes 250 patients that were already diagnosed with CKD and 150 individuals who do not have CKD. Please find below the details:

Feature Space:

1. Age(numerical): age in years
2. Blood Pressure(numerical): bp in mm/Hg
3. Specific Gravity(nominal): sg - (1.005,1.010,1.015,1.020,1.025)

4. Albumin(nominal): al - (0,1,2,3,4,5)

5. Sugar(nominal): su - (0,1,2,3,4,5)

6. Red Blood Cells(nominal): rbc - (normal,abnormal)

7. Pus Cell (nominal): pc - (normal,abnormal)

8. Pus Cell clumps(nominal): pcc - (present,notpresent)

9. Bacteria(nominal): ba - (present,notpresent)

10.Blood Glucose Random(numerical): bgr in mgs/dl

11.Blood Urea(numerical): bu in mgs/dl

12.Serum Creatinine(numerical): sc in mgs/dl

13.Sodium(numerical): sod in mEq/L

14.Potassium(numerical): pot in mEq/L

15.Hemoglobin(numerical): hemo in gms

16.Packed Cell Volume(numerical)

17.White Blood Cell Count(numerical): wc in cells/cumm

18.Red Blood Cell Count(numerical): rc in millions/cmm

19.Hypertension(nominal): htn - (yes,no)

20.Diabetes Mellitus(nominal): dm - (yes,no)

21.Coronary Artery Disease(nominal): cad - (yes,no)

22.Appetite(nominal) : appet - (good,poor)

23.Pedal Edema(nominal): pe - (yes,no)

24.Anemia(nominal): ane - (yes,no)

Label Space: ckd , notckd

**What are we 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.
* 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.
* Add some technical details