Project proposal: Medical prediction for patients of kidney disease

1. Motivation and problem definition

Almost a large number of the population worldwide is affected with a major health problem, chronic kidney disease. As a result, early detection and characterization are considered to be critical factors in the management and control of this long-lasting kidney disease. These tasks have been traditionally performed by well-trained healthcare professionals; however, they are still some of the most challenging work due to the subtle signs and difficult to detect symptoms hidden in data set. Herein, use of well-organized data mining techniques is shown to expose hidden information from clinical and laboratory patient data, which can be helpful to assist physicians in maximizing accuracy for identification of disease severity stage. The existing works without the use of the machine learning algorithms fail to provide the accuracy of prediction to the needed extent. So, our project will try to indicate that applying different machine learning algorithms provide better classification and prediction performance for determining whether one patient has chronic kidney disease. The project will try to predict the chronic kidney diseases of patients using systematic and automatic methodologies. Among the methodologies, the machine learning algorithm and feature selection are some of the very kinds.

1. Proposed solution

To give a solution for the challenge of medical prediction for kidney disease patients, we will work though 3 consecutive phases:

1. Choosing and preprocessing dataset

We choose to use the [Chronic\_Kidney\_Disease Dataset](https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease?fbclid=IwAR2bJXrFFo9VK) because it’s openly available online for public access on the Machine Learning Repository website[1]. The data was retrieved from an Indian hospital over approximately a 2-month period. It contains 25 attributes and 400 records. The first 24 attributes are the risk factors of chronic kidney disease while the 25th attribute is the classification of the disease (ckd or notckd). There are 11 numeric attributes and 14 nominal attributes. The numeric attributes include Age, Blood Pressure, Blood Glucose Random, Blood Urea, Serum Creatinine, Sodium, Potassium, Hemoglobin, Packed Cell Volume, White Blood Cell Count, Red Blood Cell Count. The nominal attributes include Specific Gravity, Albumin, Sugar, Red Blood Cells, Pus Cell, Pus Cell clumps, Bacteria, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, Class.

﻿The order of the attributes, the attributes’ name, the attributes’ datatype and attributes unit is given as follows:

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)

25.Class (nominal)

class - (ckd,notckd)

This dataset has rich resourceful information and each record is the statistic about medical situation of an individual. However, this dataset contains a considerable amount of missing values and mistyped characters, a preprocessing data phase need to be implemented to overcome this drawback.

1. Features extraction

Because the original dataset has 24 attributes and not all of them give a relevant information about medical situation of an individual. If we keep the irrelevant attributes, they may cause some noise to the predicted results. Also, because the considered attributes come from the patients, which reflect their health situation, there must be some correlations between some attributes. Keeping all the attributes will lead to the existence of redundant information. Another important reason to have the feature extraction/selection come from the explainable capacity of Machine Learning models. Medical expert needs to have a clear view about the strong predictors as the medical metrics so that they can analyze and explain the situation to the patients. As such, this phase is important.

At this time, we plan to use SelectKBest, PCA and Recursive Feature Elimination(RFE) to determine the best 10 attributes (strong predictors). ﻿The ExtraTreesClassifier is then used to determine the feature importance set.

1. Apply various machine learning models with trial-and-test strategy

After the preprocessing and features extraction phases, we will apply various machine learning models to determine which model will give the best result in predicting kidney disease patient. The model we will do the experiments are: ﻿Decision Tree Classifier, ﻿k-Nearest Neighbor Classifier, ﻿Logistic regression Classifier, ﻿Support Vector Machine Classifier, ﻿Nonlinear Support Vector Machine, ﻿Ensemble Methods

After getting the results, we will compare the efficiency of the models to determine which one is the best choice. Also, we probably propose some improvement for our approach.

﻿

﻿[1] Machine Learning Repository - Center for Machine Learning and Intelligent Systems.

Retrived from

<https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease?fbclid=IwAR2bJXrFFo9VK>