POLITECNICO DI TORINO

ICT For Smart Societies

2018/2019

ICT for Health (Lab 3)
Prof. Monica Visintin

DECISION TREE
&
CHRONIC KIDNEY DISEASE

5

GENNARO RENDE
S218951

1 Introduction

1.1 Chronic Kidney Disease

The kidneys are a pair of organs, with the shape of a bean, located on either side of the spine, below the ribs and behind the belly. Among their functions, filtering the blood is one of them: the kidneys filter over 180 liters per day producing an ultrafiltrate that goes through other processes such as reabsorption, secretion and excretion. In the end, this leads to the production of urine. The "chronic kidney disease" is the malfunction of the kidneys in carrying out their basic tasks; this can be caused by many different conditions such as type 1 or type 2 diabetes, high blood pressure, vesicoureteral or frequent kidney infections.

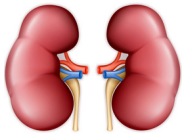


Figure 1: Kidneys

1.2 Glomerular Filtration Rate

The Glomerular Filtration Rate is a value that represents the stage of the disease; it can be calculated in many ways, for example by evaluating the amount of creatinine that is removed from the blood. If a patient removes 1440 mg in 24 h, this is equivalent to removing 1 mg/min. If the blood concentration is 0.01 mg/mL (1 mg/dL), then one can say that 100 mL/min of blood is being "cleared" of creatinine, since, to get 1 mg of creatinine, 100 mL of blood containing 0.01 mg/mL would need to be cleared.

Stage	**Description**	**GFR (mL/min per 1.73m^2)**
1	Kidney damage with normal or increased GFR	≥ 90
2	Kidney damage with mildly decreased GFR	60 - 89
3	Moderately decreased GFR	30 - 59
4	Severely decreased GFR	15 - 29
5	Kidney failure	$<$15 (or undergoing dialysis)

Table 1: Evaluation of Chronic Renal Disease

1.3 Our purpose

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. The discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation [1] and contributes to effective prediction, exploration, diagnosis and decision making. Machine learning techniques can provide tools to handle these processes. We aim to find out if patients, according to their ~~medical~~ values, have developed Chronic Kidney Disease or not, even in case the Data-Set is not totally complete. In order to perform analysis, prediction, diagnosis and decision making requires we need as much data as possible. Such data allow to perform **CART** algorithms (Classification And Regression Tree), available in the Python Scikit Learn library, more precisely. If some data are missing, we can obtain them with regression algorithms.

2 Regression on data

2.1 Cleaning and managing the data

It is crucial to have a Data-Set without typing or conversion errors: for this reason, the first task is to detect and delete all those elements that alter the file's structure. The second task is to fill all the "empty features" with the **NaN** (not a number). To perform the regression, we will use a Data-Set provided by the University of California [2]. We will only consider patient records containing at least 20 features. Since many features are nominal (such as the ones in the **Table 2**) we opted for a numerical notation as follows:

Nominal description	normal	abnormal	present	not present	yes	no	poor	good	CKD	NOT CKD
Numerical conversion	0	1	1	0	1	0	0	1	1	0

Table 2: Conversion of the descriptions in binary values

2.2 The Algorithm

We modeled our Data-Set as a matrix of **400** rows (number of patients) and **25** columns (features of the patients). The data matrix $\mathbf{X}$ contains only patients records with 25 registered features and with normalized values. The data train matrix is the full Data-Set $\mathbf{X}$ without the column of the first occurrence of a **NaN**, which is the column $\mathbf{y_{train}}$. We found the optimum weight vector $\mathbf{w}$ of linear regression by using Ridge Regression with a fixed **Lagrangian multiplier** $\lambda = \mathbf{10}$. Finally, we regressed the missing value (denormalized) in the row where it was initially missing. This method is then reiterated to regress more missing features.

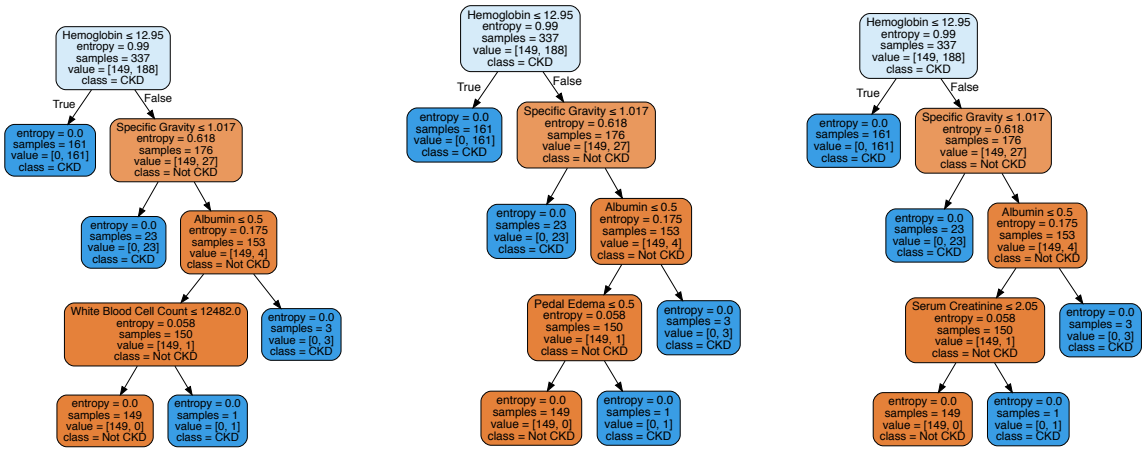
3 Results

3.1 Discussion on the results of the "Decision Tree"

The divisive hierarchical algorithm used generates alternatively 3 different binary trees, as it is shown in **Figure 2**. "*Hemoglobin*" is the feature that has a bigger importance in defying the state of the patient; its threshold is, in all the decision trees, **12.95 g/dL**, a reasonable value considering the expected one for a healthy adult (12 - 16 g/dL). In all the decisive trees created "*Specific Gravity*" and "*Albumin*" are respectively second and third in importance to determinate the presence or not of the disease. The less important feature in each tree changes: in the first one it is "*White Blood Cell Count*", in the second it is "*Pedal Edema*" while in the third it is "*Serum Creatinine*".

Feature	**Tree 1**	**Tree 2**	**Tree 3**
Hemoglobin	67,39%	67,39%	67,39%
Specific Gravity	24,59%	24,59%	24,59%
Albumin	5,41%	5,41%	5,41%
White Blood Cell Count	2,59%	0	0
Pedal Edema	0	2,59%	0
Serum Creatinine	0	0	2,59%

Table 3: Feature's importance



(a) *Tree 1* (b) *Tree 2* (c) *Tree 3*

Figure 2: Decision Trees for different random states

3.2 Conclusions

Available data are not sufficient for a precise account of a patient's health state since hemoglobin is considered as the most important element for the diagnosis of kidney failure. According to specialists, the value of albumin is the most important feature to be taken into consideration in the data set. The filtration of liquids that takes place in the kidney glomerulous allows smaller molecules to pass. Proteins, such as albumin, do not pass as they are normally bigger. If the opposite happens, kidneys are not working properly. In the second phase of filtration, filtered liquids except albumin are reabsorbed into the tubule. This causes a lack of albumin in the body. Albumin is necessary for oncotic pressure, which allows the re-absorption of liquids at the level of capillary venules. If this physiological process is altered, the level of liquids within tissues increases. Physicians can find evidence of alteration in case of a pedal edema. Another signal of kidney failure is an increase of serum creatinine over a certain threshold. However, this has to be put in context because excess of serum creatinine can be caused by physical over-training as well. Finally, it may be concluded that the algorithm does not work properly if data availability is poor. For this reason, the glomerular filtration rate is still the most reliable signal of kidney failure.

References

[1] `Prediction System Using Data Mining Techniques`
`http://paper.ijcsns.org/07_book/200808/20080849.pdf`

[2] `Chronic Kidney Disease Data Set - UCI`
`https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease`