POLITECNICO
DI TORINO

MASTER OF SCIENCE IN ICT FOR SMART SOCIETIES

Report on ICT for Health
Laboratory N°3

Decision tree and chronic kidney disease

Author:
Bruno VALENTE

a.a. 2018-2019

1 Introduction

The **kidneys** are two organs of the renal system responsible for the filtering of blood and the elimination of wastes and extra fluid through urine. This process also maintains the overall fluid balance of the human body. The production of hormones that regulate blood pressure, control the generation of red blood cells and promote bone health is also performed by the kidneys.

Because of the toxins they encounter, the kidneys are susceptible to various problems, like **chronic kidney disease (CKD)** - the progressive destruction of these organs. The prevalence of kidney disease is 10% in the adult population. Given this rate and the fact that this condition is irreversible, CKD is to be considered a worldwide public health problem with severe consequences on patients' lives. Millions of them die for kidneys failure or need expensive treatments like dialysis or a kidney transplant in order to save their lives.

There are many causes of chronic kidney disease. Diabetes and high blood pressure appear to be the most common ones. Obesity, cigarette smoking and a family history of kidney disease are also risk factors of CKD. The risk of the disease increases for people older than 50 years and occurs more frequently in African, Hispanic, Aboriginal and Asian populations. CKD can sometimes be predicted by getting regular screenings, especially if a number of risk factors occurs, and an early diagnosis can help to slow down the illness progression. Therefore, it could be useful for medical doctors to have ICT tools which help them doing their job.

The purpose of this project is to build a **decision tree** for the classification of chronic kidney disease data. The analysed dataset contains clinical features of 400 patients, collected in an Indian hospital. The number of columns is 25, out of which 24 are actually clinical parameters and one is the class (either CKD or not-CKD). Clinical features are numerical (e.g. blood pressure, potassium, haemoglobin, etc. have all numerical values, leaving aside their units of measurement) or categorical (e.g. appetite can be 'good' or 'poor', red blood cells are 'normal' or 'abnormal', etc.).

2 Preliminary phases

2.1 Data preparation

An important phase in machine learning problems is the initial preparation of the data, since files are not often "clean". It can be done by **manually** editing the file in which the dataset is stored. This first option could be taken into consideration only if the amount of data is not too large, otherwise it would require too much time.

A better option is to find **automatic** ways to prepare the data. This second method is the one used during this laboratory, exploiting appropriate Python libraries and methods. The expedients used while reading the CKD dataset are the following:

- The first 29 rows are skipped because they only contain information about the attributes;
- Since the dataset contains some typing errors (e.g. extra blanks or extra

commas), it is useful to find a pattern able to properly map and handle these errors. This is done by using a *regular expression* as attribute separator;

- Missing values, identified with '*?*' in the dataset file, are initially marked as *NaNs* (*NaN* stands for "not a number");
- Categorical features are mapped into numbers (e.g. 'yes' is mapped into 1, while 'no' into 0) because Python implementation of the hierarchical classification algorithm only supports numerical values.

2.2 Management of missing data

First of all, one has to consider that rows with too many unknown features could only introduce a larger uncertainty. Hence, records with 5 or more missing values have been removed from the dataset, that now consists of 304 patients.
Remaining unknown attributes have been predicted using **regression**, according to the following steps:

1. Patients who have no missing data have been chosen as training set
2. Data has all been normalized
3. A matrix which has the weight vectors of every feature in each column has been evaluated using regression on the training data. The algorithm performed in order to find the optimum weight vectors is the ridge regression with Lagrangian multiplier (λ) equal to 10
4. Each of the missing values has been predicted using the correct column of the matrix evaluated at the previous step. Before saving it, the regressed feature has been denormalized and rounded to the nearest acceptable value for the feature set it belongs to.

3 Decision tree for classification

One of the most popular algorithm to generate decision trees is the so called *C4.5* algorithm. The main concept used in this method is that of the **information entropy**. In a nutshell, the dataset is split into subsets until the entropy of the class for the considered subset is zero or all the attributes have been used. Each splitting is performed using the feature for which the **mutual information** between the class and the feature itself has the largest value.

3.1 Results on CKD data

Training the algorithm with the CKD data provided, the decision tree shown in Figure 1 is obtained. The feature which has the maximum entropy is the haemoglobin ('*hemo*'), with a value of 0.998. The threshold of this first attribute to split the dataset is 12.95. At the second step the feature taken into consideration is the specific gravity ('*sg*'), with an entropy of 0.55 and a threshold equal to 1.017. The last

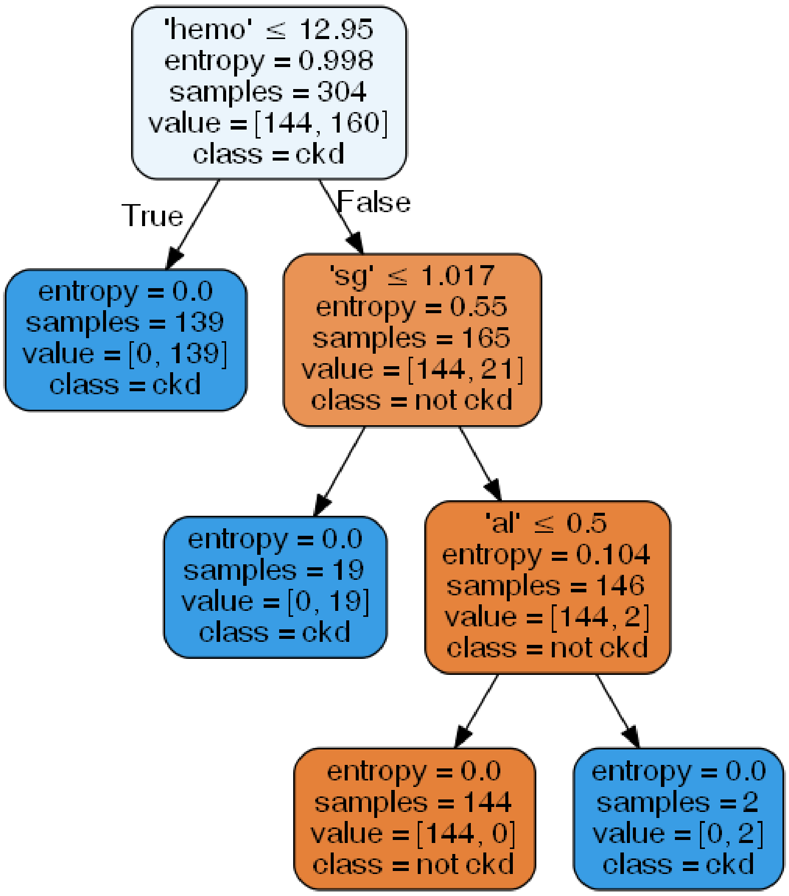

Figure 1: Decision tree on CKD data.

split is performed using the albumin (‘*al*’). Its threshold to decide whether a patient has a CKD or not is 0.5, while its entropy is equal to 0.104.
In Table 1 it is possible to notice another parameter describing the “importance” (also known as the Gini importance) of the three features above mentioned. As can be seen, haemoglobin has a lot of importance in the model. Anyway, maybe this is not the first parameter a medical doctor observes in order to diagnose CKD. This may be due to the small number of patients in the dataset and to the regressed missing values.

Since the provided dataset does not have a large number of patients, this laboratory does not include a testing phase. A more thorough analysis could be done in the future, measuring the accuracy of the classification tree on new data.

Feature	Importance
Haemoglobin	0.70093
Specific gravity	0.24882
Albumin	0.05025

Table 1: Importance of features.