Final report on the ICT for Health laboratories

Luca Gioacchini
s257076

A.A.2018/19

Contents

1 Chronic Kidney Disease

Kidney disease, or nephropathy, is a damage to kidneys, organs placed in the retroperitoneum which balance the water volume in body, filter and clean the blood and produce regulatory hormones. Chronic kidney disease causes a loss of these renal functions, which can be measured through the Glomerular Filtration Rate (GFR) and the Creatinine Clearance Rate (CrCl). The first one describes the amount of filtered fluid which flows through the kidney, while the second one is used to estimate the GFR by analyzing the volume of blood plasma cleared from creatinine, the breakdown product of creatine phospate in muscle, per unit time. An early Chronic Kidney Disease (CKD) state records a GFR value $\geq$ 90 mL/min per 1.73 m^2. Another CKD index is the elevated protein level in urine (proteinuria), such as albuminuria, related to albumin levels. Usually a kidney disorder is characterized by a proteinuria value $\geq$ 3.5 g/day.

10% of adults suffer from chronic kidney disease and treatments (dialysis or transplantation) are rarely affordable, so it is increasing the amount of data related to ckd which can leads to a disease prediction thanks to machine learning techniques, such as Decision Trees.

2 Classification and Regression Trees (CART)

Decision trees are decision support tool exploiting the tree-like model. They are used in data mining, it is easily understood and they are based on the entropy criterion and the information gain concept.

By considering a discrete random variable x which can assume x_i values with probabilities p_i ($i = 1...N$), the **entropy** of x is a measure of the variable predictability and it is described by:

$$\mathbb{H} = \sum_{i=1}^{N} p_i log_2 \frac{1}{p_i} \tag{1}$$

The maximum entropy is $\mathbb{H} = log_2 N$ and means that no information can be obtained by the observation of x, since all the values have the same probability. On the other hand, the minimum value of $\mathbb{H}$ (zero entropy) means that the random variable can be exactly predicted.

Regarding the information gain, it is a synonym for Kullback-Leibler (KL) divergence between two random variables described by their probability density functions p_i and q_i[1] and it measures the amount of information gained by the first one when the second one is observed.

Another criterion that can be used to generate decision trees is the Gini Impurity one:

$$G(\mathbf{x}) = 1 - \sum_{i=1}^{N} p_i^2 \tag{2}$$

By considering the label (or class) distribution of a dataset, the Gini Impurity measures how often a variable randomly chosen from the dataset can be incorrectly classified if the label is randomly assigned by observing its distribution.

[1] KL divergence: $\mathbb{KL}(p||q) = \sum_{i=1}^{N} p_i log \frac{p_i}{q_i}$

In light of the above, the CART algorithm allows the generation of regression trees, which are binaries, in which a single node is linked to at most two children. To represent through a binary tree a continue variable, with more than two values, it is possible to split the tree in more levels. Each levels contains binary nodes.

3 Experimental Results

The data used to generate a Decision tree for detecting a CKD are the ones provided in 2015 by the Algappa University(IND). The dataset consists of 25 attributes recorded from 150 healthy patients and 250 chronic kidney diseased ones. 11 attributes are numeric (such as sodium, potassium, etc.) and 14 are nominal (such as albumin, appetite, etc.). The 25th attribute is the class (CDK or notCDK).

3.1 Data Management

After having manually cleaned some data and translated the nominal features as numeric ones, the dataset is still missing some information because of the manual recording of it. To solve this problem the complete entries of the dataset has been used as training matrix for the ridge regression with the Lagrangian multiplier $\lambda = 10$. The estimated categorical data are obtained by rounding the ones resulting from the ridge regression algorithm.

After the data management it is possible to create a decision tree by exploiting the *scikit-learn* python library, which uses an optimized version of the CART algorithm[2]. The two criteria can be fed to the python library: the entropy one and the Gini one.

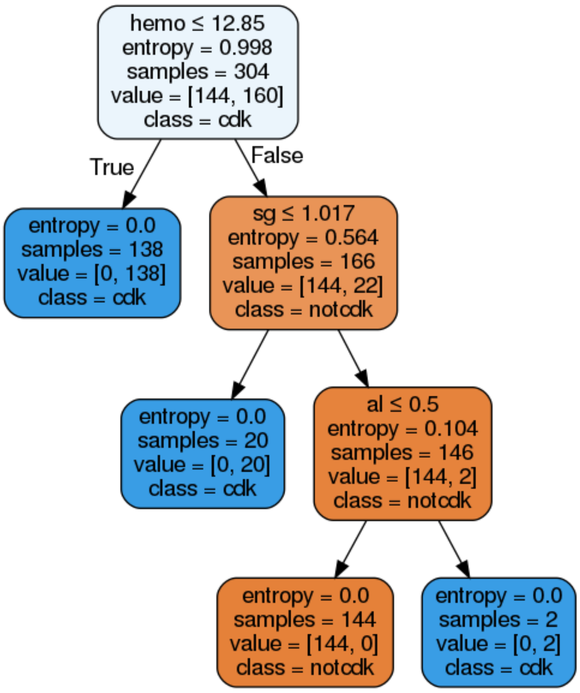

Figure 1: Decision Tree with Entropy criterion for CDK estimation

[2] https://scikit-learn.org/stable/modules/tree.html

Figure 1 shows the decision tree based on the dataset described in Section 3. The first node is the tree root, the *hemo* label refers to the hemoglobin, which is the feature which splits better the dataset. The amount of entropy introduced by this feature is 0.998 and 304 samples are used by the CART algorithm to state which is the root feature. If a subject has a hemoglobin value ≤ 12.85, it is classified as diseased (True branch), otherwise the *sg*, or Specific Gravity, is evaluated. In this case the threshold is set as sg value ≤ 1.017 and it is based on the observation of $304 - 138 = 166$ samples. This procedure continues until the amount of entropy introduced by the last feature is equal to 0, which means that no more information can be extracted from the dataset.

To obtain a clearer understanding of the tree, the feature labels are: *hemo* for hemoglobin, *sg* for specific gravity and *al* for albumin.

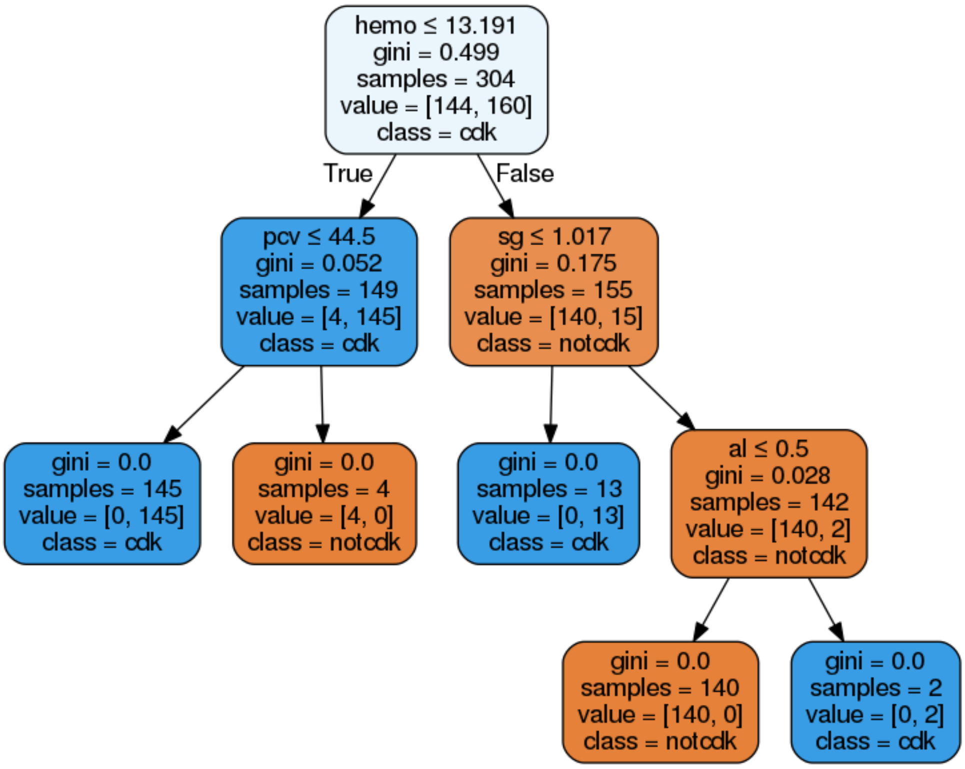

Figure 2: Decision Tree with Gini criterion for CDK estimation

According to what has been said regarding the Figure 1, Figure 2 shows the decision tree generated when the Gini Impurity criterion defined in Equation (2) is applied. The procedure is the same of the entropy criterion, but in this case the stopping condition is determined by the value of the Gini Impurity equal to 0. By observing the *value* field in the root node of both the decision trees, it is clear that the results are similar, since 144 patients are classified as diseased and 160 as healthy, even if the decision paths are different.

The feature labels are the same of Figure 1, except for the *pcv* one which refers to the packed cell volume.

Figure 3 shows the feature importance used by the CART algorithm to determine the decision path. The Hemoglobin importance is the greater one, according to the generated trees, which are rooted in this feature. The Packet Cell Volume is not taken into account

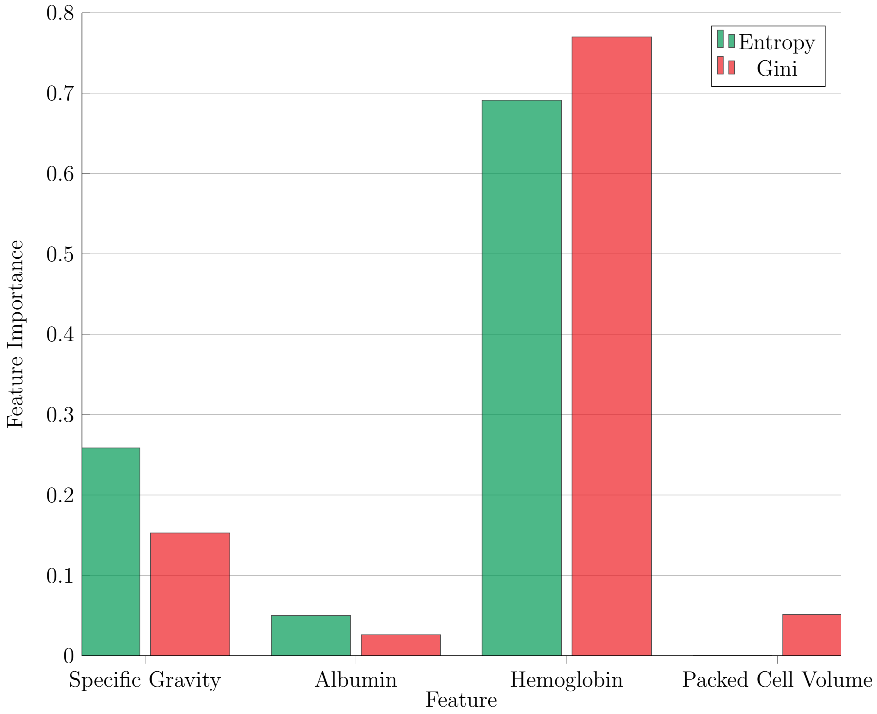

Figure 3: Comparison of feature importance when two different criteria are used in the scikit-learn python library

when the entropy criterion is used, so its importance is 0, on the other hand, by observing Figure 2, the last nodes level refers to this feature, so its importance is $\neq 0$.
Regarding the other features, their importance is proportional to the entropy or Gini Impurity value obtained for the considered feature.

4 Conclusions

By considering the medical information treated in Section 1, it is clear that the CART algorithm is not so accurate, since one of the most important feature to determine if a subject is diseased is the albumin one. This happen because of an incorrect data recording (the ridge regression applied to fill the missing data can introduce errors) and because of the limited amount of data.

Regarding the criteria used in the CART algorithm, experimental results show that even if the Gini Impurity takes into account one more feature (Packed Cell Volume) and different decision paths are used, the amount of diseased subjects is the same for both the obtained decision trees.