Report 5: Chronic Kidney Disease classification

Francesco Conforte, s277683,
ICT for Health attended in A.Y. 2020/21

December 31st, 2020

1 Introduction

Chronic kidney disease (CKD) derives from a gradual loss of kidney filtering capability over time, typically caused by high blood pressure and diabetes. Prevalence of the illness is around 10% in adult population, and its early detection avoids the dramatic consequence of complete kidney failure and necessity of kidney transplant.

Whilst a cure does not exist for CKD, treatments of kidney disease are available to reduce the symptoms, but they are expensive and impair the normal life of the affected subject (long dialysis sessions).

Kidney functionality can be assessed through the Glomerular Filtration Rate (GFR), calculated from the 24-hour collected urine or from the blood creatinine test.

A public dataset is available [1] to explore correlations between CKD and subject parameters. In particular, the dataset includes 24 features (see Table 1), among which 11 are numerical and 13 are categorical. Each of the 400 points of the dataset belongs either to class `ckd` (chronic kidney disease is present) or `notckd`. Unfortunately, some features are missing for some subjects (see Table 2) and must be replaced; on the contrary, there are no cases of missing class.

Object of the work is to use the dataset to build decision trees to classify new subjects as either healthy or affected by chronic kidney disease. Decision trees are all built using Python Scikit Learn class `DecisionTreeClassifier` using entropy criterion, but they differ on the method used to substitute the missing values.

2 Methods

The considered methods to manage missing values are the following:

1. Simple removal from the dataset of all the rows that contain one or more missing values.

2. Replacement of each missing value with the median of the corresponding feature.

	feature	meaning	type
1	age	age	numerical
2	bp	blood pressure (mm/Hg)	numerical
3	sg	specific gravity	categorical
4	al	albumin	categorical
5	su	sugar	categorical
6	rbc	red blood cells	categorical
7	pc	pus cell	categorical
8	pcc	ps cell clumps	categorical
9	ba	bacteria	categorical
10	bgr	blood glucose random (mg/dl)	numerical
11	bu	blood urea (mg/dl)	numerical
12	sc	serum creatinine (mg/dl)	numerical
13	sod	sodium (mEq/L)	numerical
14	pot	potassium (mEq/L)	numerical
15	hemo	haemoglobin (gms)	numerical
16	pcv	packet cell volume	numerical
17	wc	white blood cell count	numerical
18	rc	red blood cell count (million/cmm)	numerical
19	htn	hypertension	categorical
20	dm	diabetes mellitus	categorical
21	cad	coronary artery disease	categorical
22	appet	appetite	categorical
23	pe	pedal edema	categorical
24	ane	anaemia	categorical

Table 1: Features in the UCI kidney dataset

3. Substitution of the missing values to the linearly regressed ones, obtained with linear least square method and using only the complete rows as training data.

2.1 Removal of rows with missing values

As shown in Table 2, by removing all the rows that contain at least one missing value, only 158 rows remain in the dataset, out of which 43 are related to subjects affected by CKD. With respect to the original dataset, the reduced one has a much smaller ratio of positive cases (i.e. $43/158 = 0.27$ against $250/400 = 0.62$), which for sure has an impact on the results.

The obtained decision tree is shown in Figure 1. It has to be noted that albumin (”al”) is a categorical feature that takes values in the alphabet $\{0, 1, 2, 3, 4, 5\}$ where 0 means ”normal” and 5 means ”very abnormal/pathological” (i.e. very small quantities of albumin). It is therefore correct that a subject with categorical feature albumin less than 0.5 can be

m	number of rows with m missing values
0	158
1	45
2	33
3	37
4	31
5	33
6	12
7	20
8	8
9	12
10	4

Table 2: Missing values in the dataset.

considered healthy. Note again that serum albumin levels less than 3.80 g/dL are associated with increased odds of rapid kidney function decline and increased risk of incident chronic kidney disease, but here feature ”al” does not represent serum albumin quantities measured in g/dL, but degree of normality of the albumin quantity.

However, among the 116 subjects with ”al=0”, there is just one person affected by CKD, who is detected because of missing hypertension (”htn” equal to zero). Of course this result cannot be generalized, and actually the software generates different decision trees each time it is run, since it can take other equivalent features to isolate the only subject positive to CKD. Therefore, the decision tree obtained from the reduced dataset only allows to find the importance of albumin in the diagnosis of CKD.

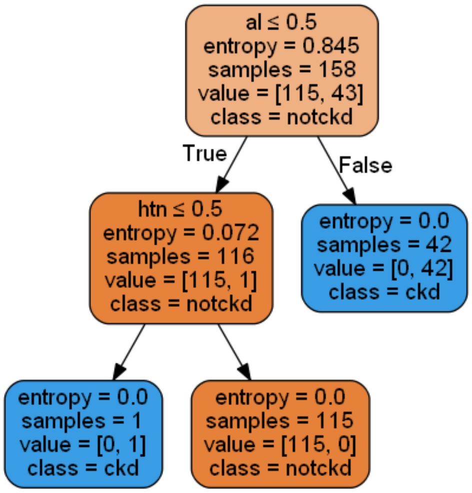

Figure 1: Decision tree obtained using only the rows without missing values.

2.2 Substitution with the median

In the second method, missing values are replaced by the median values. In particular, for each feature in the dataset, the median is evaluated using only the reduced dataset described in Sect. 2.1. The median is preferred to the mean because it does not suffer from the presence of outliers in the dataset.

With this method, the entire dataset can be used to build the decision tree, shown in Figure 2.

The root node of the classified decision tree is identified with the levels of Serum Creatinine (”sc”). Creatinine is produced continuously during normal muscle breakdown. The kidneys filter creatinine from the blood into the urine and reabsorb almost none of it. If the kidney is not functioning properly, it will not effectively eliminate the accumulated creatinine from the blood. This will result in an increased creatinine level. Therefore, it can be a useful feature for urologists in identifying CKD.

Actually, the quantity of ”sc” varies according to the gender of the person, to the race, to the age, etc..., but since the decision trees reported in this report are built by using data of Indian people, thus non-Hispanic patients, the identified level of ”sc”$\leqslant$1.25 mg/dL can be considered proper from a medical point of view [2].

The second-level decision tree is identified with the level of (urine) specific gravity (”sg”)[1]. This categorical feature can take values in the alphabet {1.005,1.010,1.015,1.020,1.025} and generally a normal specific gravity varies in the range[2] of 1.020 to 1.028. Therefore, a value of (urine) specific gravity lower than 1.017 can be considered a good indicator of CKD.

The third-level decision tree is identified with the level of haemoglobin (”hemo”). Actually, by observing the tree of Fig. 2, the number of people in the dataset having a value lower than or equal to 12.85 gms is just 9 over 164 people. This means that this result cannot be generalized, as it happened for the tree of Sec. 2.1. Indeed, the software generates different trees starting from the third-level every time it is run. Nevertheless, this tree is useful to stress the importance of serum creatinine, specific gravity and haemoglobin in the diagnosis of CKD.

2.3 Substitution with regressed value

The reduced dataset described in Sect. 2.1 is used as training dataset, corresponding to matrix $\mathbf{Z}_{tr}$. If only feature f is missing in row k of the original dataset, then matrix $\mathbf{X}_{tr}$ is defined equal to $\mathbf{Z}_{tr}$ in which column f has been removed, whereas column $\mathbf{y}_{tr}$ is column f of $\mathbf{Z}_{tr}$. Then matrix $\mathbf{X}_{tr}$ stores the regressors and vector $\mathbf{y}_{tr}$ has the role of regressand, within the training dataset. The optimum column vector of weights $\mathbf{w}$ that minimizes the square error between $\mathbf{X}_{tr}\mathbf{w}$ and $\mathbf{y}_{tr}$ is obtained, according to the LLS method, as

$$\mathbf{w} = \left(\mathbf{X}_{tr}^T\mathbf{X}_{tr}\right)^{-1}\mathbf{X}_{tr}^T\mathbf{y}_{tr}$$

[1] Specific gravity is the weight of urine compared to distilled water which has a specific gravity of 1.000.

[2] It is strictly related to gender and age. Generally, adults normal values are between 1.020 and 1.028.

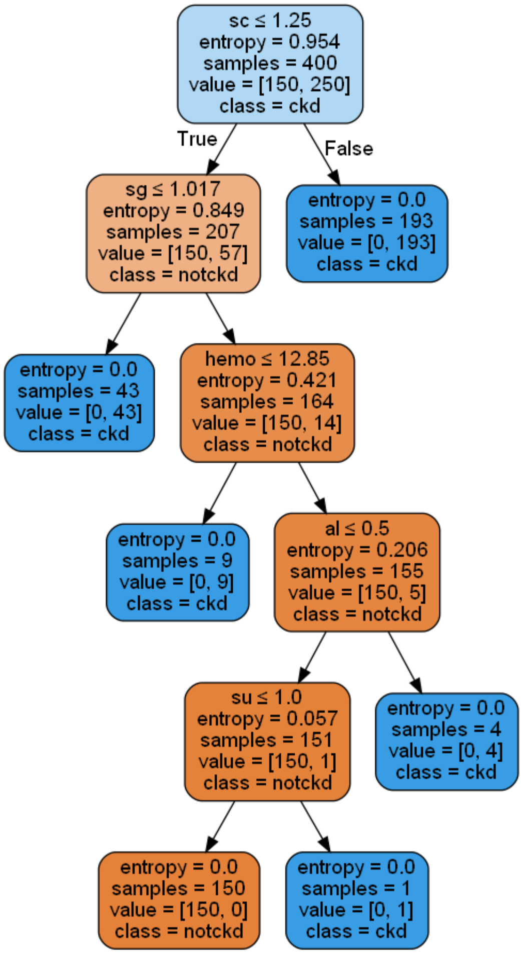

Figure 2: Decision tree obtained by replacing the missing values with the median.

Being[3] $\mathbf{x}^T$ row k without feature f, The missing value is found as

$$\hat{y}_f = \mathbf{x}^T \mathbf{w}.$$

If, in general, m features are missing in row k, then $\mathbf{X}_{tr}$ is defined equal to $\mathbf{Z}_{tr}$ in which all the m columns corresponding to the missing values have been removed, m vectors $\mathbf{w}$ are generated and each missing value is replaced.

Actually only the rows with up to 6 missing values are considered, considering that regression accuracy cannot be sufficient if more than one fourth of the data is missing. Therefore, the obtained dataset after the replacement of the missing values is made of 349 rows, with 199 positive cases (ratio of positive cases 0.57, more similar to the ratio 0.62 of the original dataset).

The obtained decision tree is shown in Fig. 3.

[3] Vector $\mathbf{x}$ is a column vector, $\mathbf{x}^T$ is a row vector.

In this case, the root node is identified with the level of haemoglobin (”hemo”). A study regarding the correlation between anaemia, which occurs when the level of haemoglobin is lower than the normal level (13.0 to 15.0 gms), and CKD [3] has confirmed that haemoglobin level is an important factor to consider in the diagnosis of CKD, because it is strongly predictive of complications and death from cardiovascular causes in patients with chronic kidney disease. Therefore, the first-level decision tree can be considered correct from a medical point of view.

By going ahead with second and third levels of the tree, two features already used by trees of Fig. 1 and Fig. 2 are selected by the software to make the decision. As shown in Fig. 3, they are (urine) specific gravity and albumin. Their importance from a medical point of view is explained in Sec. 2.1 and 2.2.

In the end, the whole decision tree can be considered correct from a medical point of view but nothing can be stated about its accuracy since no statistical analysis was made on it.

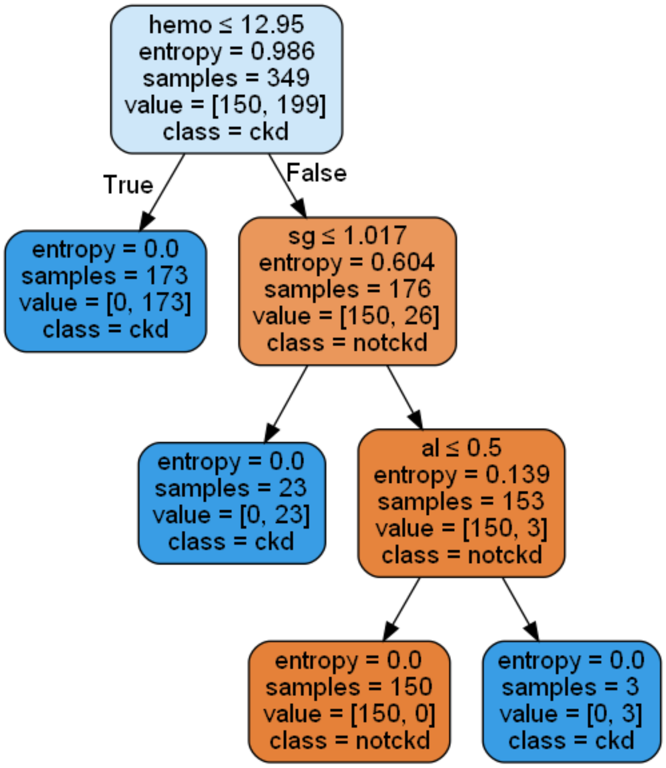

Figure 3: Decision tree obtained by replacing the missing values with the regressed values.

3 Accuracy, sensitivity, specificity

The decision tree of Sect. 2.1, obtained with the reduced dataset of 158 points, was used to classify all the 400 points of the dataset obtained in Sect. 2.2.

Accuracy, sensitivity and specificity were measured several times, using different seed states (from 0 to 199) for the generation of the decision tree (i.e. with a different second level of the decision tree). Results are given in Table 3.

It is clear that the obtained decision tree of Fig. 1 performs perfectly in detecting cases of people truly not affected by CKD. Indeed specificity is 100%. On the contrary, the decision tree has, on average, a low sensitivity, since its mean value is 74.1%. This means that the tree mainly states that a person is not affected by chronic kidney disease by reversing the truth of the matter. The result is a high number of false negatives.

	Mean value [%]	Max value [%]	Min value [%]
Accuracy	83.8	89.5	76.5
Sensitivity	74.1	83.2	62.4
Specificity	100	100	100

Table 3: Statistical results obtained with 200 measurements.

4 Conclusions

Results reported in Table 3 cannot be considered acceptable from a medical point of view, since human lives are directly involved. Accuracy of such a model is of high priority; hence accuracy of just 83.8% cannot be accepted.

Furthermore, chronic kidney disease is strictly related to mortality because early detection of CKD is very difficult. Indeed, detecting chronic renal failure is difficult until 25% of renal function has already been lost [4]. This observation states that false negatives cannot be accepted at all in detecting CKD. Therefore, sensitivity has to be increased.

Whilst obtained statistical results cannot be accepted from a medical point of view, the built decision trees may be beneficial in identifying which features should be considered the most and which features can be judged to be less important during the detection of CKD. In the end, it can be claimed that factors like albumin, haemoglobin, urine specific gravity and serum creatinine[4] are the most informative ones. Other useful factors for CKD detection not stated in this report were, instead, identified by other researchers in [4], like age, red blood cells and blood glucose random.

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease, 2015-07-03.

[2] https://www.emedicinehealth.com/creatinine_blood_tests/article_em.htm.

[4] Factors are not in order of importance.

[3] Tilman B. Drüeke, Francesco Locatelli, Naomi Clyne, Kai-Uwe Eckardt, Iain C. Macdougall, Dimitrios Tsakiris, Hans-Ulrich Burger, and Armin Scherhag. Normalization of hemoglobin level in patients with chronic kidney disease and anemia. *New England Journal of Medicine*, 355(20):2071–2084, 2006. PMID: 17108342.

[4] Chin-Chuan Shih, Chi-Jie Lu, Gin-Den Chen, and Chi-Chang Chang. Risk prediction for early chronic kidney disease: Results from an adult health examination program of 19,270 individuals. *International Journal of Environmental Research and Public Health*, 17(14), 2020.