Report 5: Chronic Kidney Disease classification

Francesco Donato, s304810,
ICT for Health attended in A.Y. 2021/22

December 30th, 2021

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 and measure the performance. Decision trees are all built using Python Scikit Learn class `DecisionTreeClassifier` [2] using entropy criterion; missing values are replaced using regression trees available in the same Python library [3].

2 Methods

2.1 Removal of rows with missing values

Table 2 shows that only 158 out of 400 rows have no missing values. If only these data are used, then most of the information is lost and the number of positive cases is 43, with a ratio 43/158 = 0.27, which is much less than in the original dataset 250/400 = 0.62. As a

	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	hemoglobin (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	anemia	categorical

Table 1: Features in the UCI kidney dataset

consequence, the decision tree [2] based on just these 158 rows used as training dataset, shown in Figure 1, might be not completely correct. Notice 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 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 subject affected by CKD, who is detected because of absence of 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.

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.

2.2 Substitution of missing with regressed values

The reduced dataset $\mathbf{Z}_{tr}$ with no missing values (described in Sect. 2.1) is used as training dataset to perform regression on the missing values. If only feature f is missing in row k, then the training regressor matrix $\mathbf{X}_{tr}$ is defined equal to $\mathbf{Z}_{tr}$ where column f is removed (158 rows and 23 columns), whereas the training regressand column $\mathbf{y}_{tr}$ is set equal to column f of $\mathbf{Z}_{tr}$. Matrix $\mathbf{Z}_{tr}$ and vector $\mathbf{y}_{tr}$ are used as inputs to train the tree regressor [3] and then the missing value in row k is substituted with the regressed value obtained by feeding the tree with the valid part of row k. If more than one feature is missing in row k, then exactly the same procedure is used, but the training regressand is a matrix instead of being a column.

Actually, only the rows with up to 6 missing values (191) were included in this process, 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 new dataset is randomly shuffled and, to have a fair comparison with the result obtained in Sect. 2.1, 158 rows are used to train the decision tree. The obtained decision tree is shown in Fig. 2.

It can be observed how for the decision tree shown in Fig. 2, the features determining if the patient is positive to CKD or not are hemoglobin (hemo) and specific gravity (sg), while for the decision tree shown in Fig. 1 are albumin (al) and hypertension (htn), hence in both cases only two features are sufficient to detect CKD. It also should be considered that the decision tree obtained after the random shuffle of the new dataset (Fig. 2) has an accuracy equal to 0.95, while the decision tree in Fig. 1 has an accuracy equal to 0.85. Regarding the analysis of the obtained confusion matrixes, the decision tree with the lower accuracy predicted 27 false negatives and 1 false positive, while the decision tree with the higher

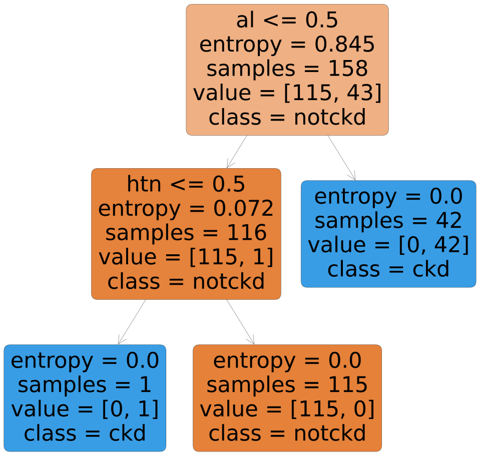


Figure 1: Decision tree obtained using only the rows without missing values.

accuracy predicted significantly fewer false negatives (which were 6) and 3 false positives.

Knowing that decision trees tend to overfit, shuffling was performed 6 times and 6 slightly different decision trees were obtained. Overall, in the decision trees generated, there was the presence of a total of 9 features, where the most frequent ones among the 6 trees were hemoglobin (hemo) and specific gravity (sg), which appeared in each one of them. The minimum number of features necessary to detect a patient positive to CKD was 2 and the maximum amount was 4. The features present in these trees can be considered relevant as they minimize the entropy which gradually reaches a value equal to zero (no uncertainty), meaning that the patient status can be exactly predicted, splitting the dataset into groups for effective classification. In terms of accuracy in predicting the class of patients, the least accurate tree had an accuracy of 0.95, while the most accurate reached 0.98, meaning that the decision trees were not significantly affected by overfitting.

3 Accuracy, sensitivity, specificity

The decision tree of Sect. 2.1, obtained with the reduced dataset of 158 points, was used to classify the 191 points of the dataset with missing values regressed as described in Sect. 2.2. The decision trees obtained in Sect. 2.2 were used to classify the 191 points not belonging to the training dataset.

Accuracy, sensitivity, and specificity were measured several times, using different state seeds in the generation of the decision tree [2], and several shuffles for the decision trees of Sect. 2.2. Results are given in table 3. The table shows the mean, the standard deviation, the maximum value and the minimum value of sensitivity, specificity and accuracy. In order to obtain these statistics, a total of 150 measurements was performed.

In order to directly give the doctor the output on if the patient has or not CKD, an algorithm was implemented in order to create a random forest of 1000 trees, which gave the

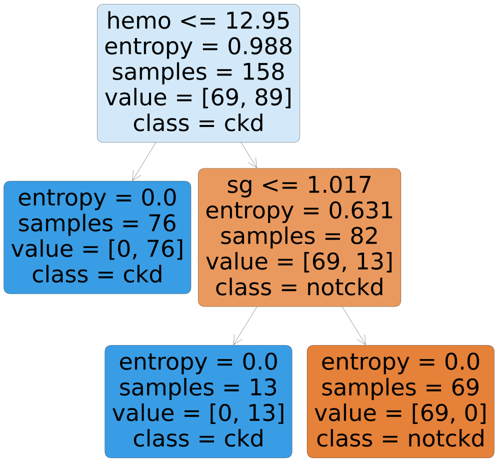
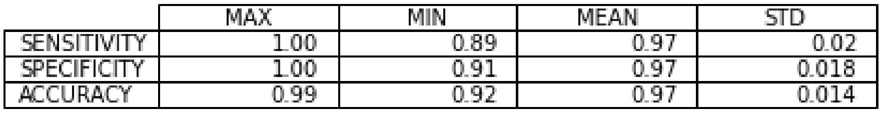


Figure 2: Decision tree obtained by replacing the missing values with the regressed values.



	MAX	MIN	MEAN	STD
SENSITIVITY	1.00	0.89	0.97	0.02
SPECIFICITY	1.00	0.91	0.97	0.018
ACCURACY	0.99	0.92	0.97	0.014

Table 3: Statistics of sensitivity, specificity and accuracy on test dataset.

following results: sensitivity equal to 1 (100%), specificity equal to 1 and accuracy equal to 1.

4 Conclusions

Kidneys are crucial organs of the body and life is at risk when kidney function does not work properly anymore. They balance the volume of water in the body, filter the blood and eliminate the waste products. Moreover, they produce hormones that regulate some important body functions (blood pressure, making of red blood cells, etc.). Chronic kidney disease (CKD) derives from a gradual loss of kidney filtering capability over time, typically caused by high blood pressure and diabetes. This disease is the most common cause of nephrotic syndrome, which is a kidney disorder that causes the body to pass too much protein in the urine.

Regarding the 6 decision trees obtained through shuffling in Sect. 2.2, the features which appeared in each of them are hemoglobin (hemo) and specific gravity (sg). Hemoglobin is the protein molecule in red blood cells that carries oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues back to the lungs [4], while a urine specific gravity test compares the density of urine to the density of water. This quick test can help determine how well kidneys are diluting the urine. Urine that is too concentrated could mean that kidneys are not functioning properly or that the patient is not drinking enough water [5].

In conclusion, taking advantage of decision trees is a solution for predicting the patient status. Analyzing the statistics in Table 3, the accuracy, the sensitivity (true positive rate) and the specificity (true negative rate) show that this predicting instrument should be sufficiently reliable for the doctor. An advantage for the doctor in exploiting a decision tree is that it allows him/her to learn how to make decisions, in fact, he/she could choose to rely on one specific decision tree which he/she personally considers to be more powerful than the others. However, knowing that decision trees tend to overfit, a random forest could be implemented in order to avoid this issue but as opposed to decision trees, a random forest gives directly the output with the opinion on if the patient has or not CKD (it is just an aid for the doctor), hence the doctor just decides to trust it or not.

References

[1] `https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease`

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

[3] `https://scikit-learn.org/stable/modules/tree.html#regression`

[4] `https://www.medicinenet.com/hemoglobin/article.htm`

[5] `https://www.healthline.com/health/urine-specific-gravity`