Report 5: Chronic Kidney Disease classification

Marco Colocrese, s301227,
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 KCD.

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. This example of tree obtained using shuffled data is clearly different from the previous one as it uses different features and has different numbers of values belonging to the two classes.

Knowing that decision trees tend to overfit, shuffling was performed other 1000 times and 1002 different decision trees were obtained. Overall, some trends could be detected, analyzing the use of features: the two histograms of Fig. 3 show how many time every feature is used in the construction of the 1002 analyzed trees and their importance. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance[4]. In both the histograms it can be

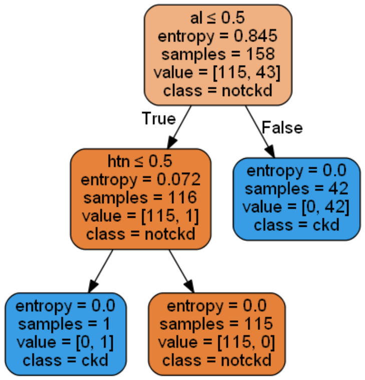


Figure 1: Decision tree obtained using only the rows without missing values.

observed that the most used and important features are *specific gravity, albumin, blood urea, hemoglobin and packet cell volume.* However, some features, such as *ps cell clumps, bacteria and coronary artery disease*, have no importance and could be deleted without influencing the goodness of the results.

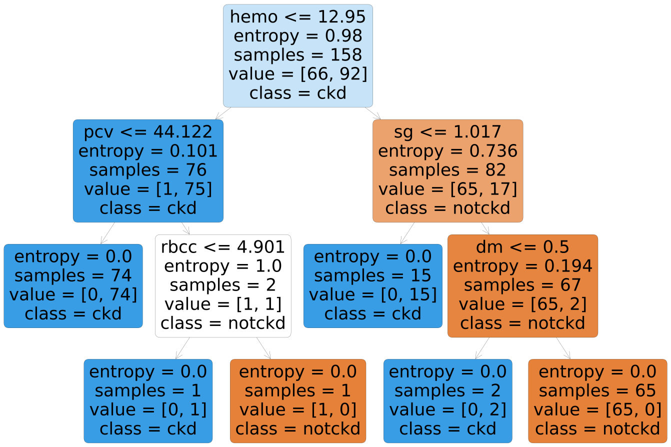
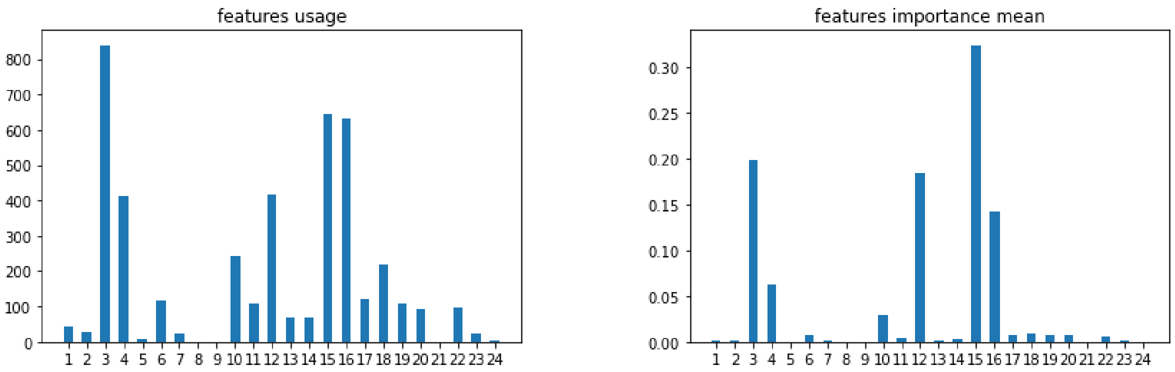


Figure 2: Decision tree obtained by replacing the missing values with the regressed values.

3 Accuracy, sensitivity, specificity

The decision tree of Sect. 2.1, obtained with the reduced dataset of 158 points, was used to classify 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 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. Overall, analyzing mean and standard deviation of



(a) Features usage.

(b) Features importance.

Figure 3: *Numbers can be associated to features using Tab.1.*

	Decision trees				Random forest
	Max	Min	Mean	Standard deviation	
Sensitivity	1.0	0.850	0.961	0.040	0.970
Specificity	1.0	0.850	0.970	0.037	1.0
Accuracy	0.995	0.874	0.965	0.034	0.984

Table 3: Statistical results

decision trees results, the results are good, reaching very high values of sensitivity, specificity and accuracy. However, it is important to stress minimum values that can represents a problem for the patient health: in particular sensitivity and accuracy should be close to 1. Comparing these results to the accuracy (0.853) and the antidiagonal of the confusion matrix (1 27) of the first decision tree (Fig. 1), the shuffled decision trees give better results, probably due to the regression errors mitigation. Those low values are symptoms of overfitting. Better results can was obtained using random forest, reaching accuracy 0.984, specificity 1 and sensitivity 0.970. However, both algorithms have been tested as random forest could be considered less useful by medical doctors that could want a decision tree that they can analyze to look at the features, instead of having directly the final classification.

4 Conclusions

As a cure does not exist for CKD, its detection assumes a huge importance, making possible to stop or to slow down the gradual loss of kidney filtering capabilities over time. Furthermore, CDK triggers other health issues (cardiovascular diseases) leading to premature death or disability. Moreover, treatments like dialysis and kidney transplantation are not affordable or extremely costly, burdening the public healthcare purse; for instance, the yearly economic costs of care for CKD and end-stage renal disease (ESRD, that occurs when CKD reaches an advanced state) in patients over age 65 are $60 billion, representing 24% of total Medicare expenditures in 2011 in the United States of America[5].

Kidneys perform key tasks such as excretion of waste, maintenance of fluid balance and hormones synthesis. Therefore it is really important to keep them healthy trough everyday simple habits. Some of the recommended ones are: get enough sleep, drink more than 1.5 liters of water par day, stop smoking, limit alcohol intake and make physical activity a part of the daily routine. Moreover it is really important to verify the presence of diabetes, high blood pressure, heart disease and a family history of kidney failure that are related to an higher probability of developing CKD[6].

As the healthcare industry is producing massive amounts of data, machine learning techniques can help, as attested by the reported results.

The developed software can also help in understanding the importance of the considered features. For instance, albumin resulted to be one of the most important features (Fig. 3). A countercheck can be given by the medical knowledge: the finding of albumin in the urine is considered the first sign of an otherwise silent kidney condition, as albumin is a protein found in the blood and a healthy kidney does not let albumin pass from the blood into the urine[7]. The albuminaria (that indicates the high quantity of albumin) can also be used to classify the risk of adverse outcomes related to kidney disease (KDIGO classification). The analysis showed that another important feature is the hemoglobin, the iron-rich protein that allows red blood cells to carry oxygen from lungs to the rest of the body.

However, these results must be interpreted and analyzed by medical doctors. For instance, the low importance of anemia probably depends on its strong correlation with hemoglobin and not on its real medical importance in CKD diagnosis. In fact, anemia (condition in which the amount of hemoglobin is lower than normal values) often gets worse as kidney disease progress and more kidney function in lost[6].

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://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[5] https://journals.physiology.org/doi/full/10.1152/ajprenal.00266.2016

[6] https://www.niddk.nih.gov/health-information/kidney-disease

[7] https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd/tests-diagnosis/albuminuria-albumin-urine