# Understanding and predicting Chronic Kidney Disease

Pablo Vales

# Abstract

In this work the Chronic Kidney Disease (CKD) dataset from UCI Machine Learning repository has been used to understand how different biological markers get modified when developing CKD and to build two models to predict whether someone has CKD or not. For the former we have found that older people seem to be more prone to develop it and hypertension is a prevalent marker in such cases. For the latter, the features with the most predictive power were found to be, among others, hemoglobin, glucose levels or hypertension. The models show great potential but, as we will discuss, a more extensive dataset is needed for a reliable accuracy measure.

# Motivation

Chronic Kidney Disease (CKD) is characterized by having damaged kidneys that drastically reduces these organ's filtering power. It affects approximately 9% of the world population. Kidneys do not have the ability to heal themselves and thus, it makes CKD a very dangerous condition.

It is thus very important for the **non-medical population** to be able to identify how particular markers get modified in the case someone develops CKD: **an early diagnosis improves the prognosis**. Conversely, **medically trained** individuals should have the means to rapidly and effectively detect if someone has CKD and to know what markers are the most important ones to predict this disease.

For these reasons I will (i) delve into how the features that can be **measured at home** by an average person are compared between healthy and sick individuals and (ii) build a classification model to check the dataset potential for diagnosis

# Dataset

The data used comes from the Chronic Kidney Disease (CKD) dataset, downloaded from the UCI Machine Learning Repository. The dataset contains **400 instances and 25 features** linked to a **binary target label** that defines whether someone has CKD or not. 11 features are numerical and 14 nominal. The dataset is imbalanced, that is, it has 250 negative (healthy) and 150 positive (CKD) instances.

When dropping rows with missing values, the number of instances goes down to 114 negatives (healthy) and 43 positives (CKD).

Some of the relevant features are, for instance: age, blood pressure, appetite, blood glucose, hemoglobin, ion blood concentration (Sodium and Potassium), among others.

# Data Preparation and Cleaning (1)

1. The dataset is provided in .artff format. In order to transform it to a format pandas can handle I wrote a simple RegEx-based function to build a .csv file.

2. All features had assigned an “Object” dtype. Therefore, another function was used to automatically change the feature types to: **float** if the value was a number and leave them as objects if the value was alphanumeric.

3. All missing values were labeled by default with the interrogation symbol “?”. A value replacement in the whole dataset was done from “?” to numpy NaN values.

4. Next step required checking if there were any mislabeled features. Exploration showed that there were indeed categorical values with different nomenclatures. Specific replacement was done accordingly.

# Data Preparation and Cleaning (2)

5. Missing value data imputation for model building was done as following:

    a. Categorical columns: missing values were imputed as the most frequent value for **each** target class

    b. Numerical columns: missing values were imputed as the corresponding mean for **each** target class

# Research Questions

1. How do the following markers behave when someone has CKD?:
    a. Age
    b. Diastolic blood pressure and hypertension
    c. Appetite
    d. Blood glucose and diabetes

2. How well can we predict if someone has CKD with Decision Trees and Logistic Regression? What are the most relevant features for the models to make this prediction?

**Important note**: this work does not delve into correlation and causality of CKD and the corresponding health markers. Abnormal values of the studied markers must be always consulted with a doctor for a correct assessment, as they may not be a consequence of an already developed CKD.

# Methods

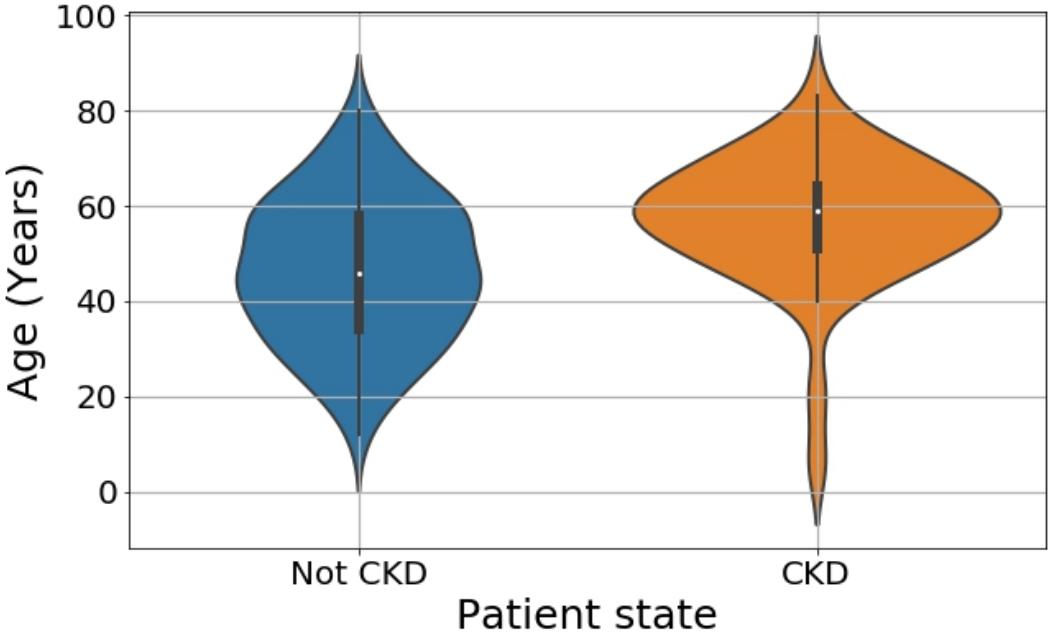
**Data exploration**: histograms, boxplots and/or violinplots were used for continuous numerical features, while bar plots were used for categorical features. Since this dataset is imbalanced, bar plots were done showing the relative % of people for each class instead of absolute values for appropriate comparison. This exploratory analysis was done in the reduced dataset i.e rows with missing values dropped.

**Model building**: One Hot Encoding was applied to the categorical variables to allow the model to treat string data types appropriately. Training data was 70% of the total. 5-fold Stratified cross validation was applied for accurate extraction of the model hyperparameters via GridSearchCV. The utilized metric score was based on simple accuracy score and confusion matrix to check the complete effectiveness of the model in terms of False Positives and False Negatives.

Prediction was first done on the **“reduced dataset”** i.e rows with missing values dropped, **which has 157 instances**. Finally, the model was applied to the **“complete dataset” (400 instances)** with missing value imputation as explained in Section “Data Preparation and Cleaning(2)”

# Part 1. Relating marker features with Chronic Kidney Disease

# CKD marker: Age

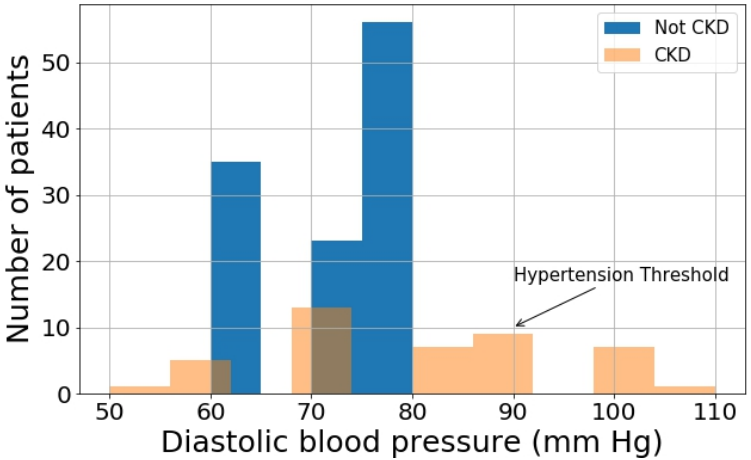


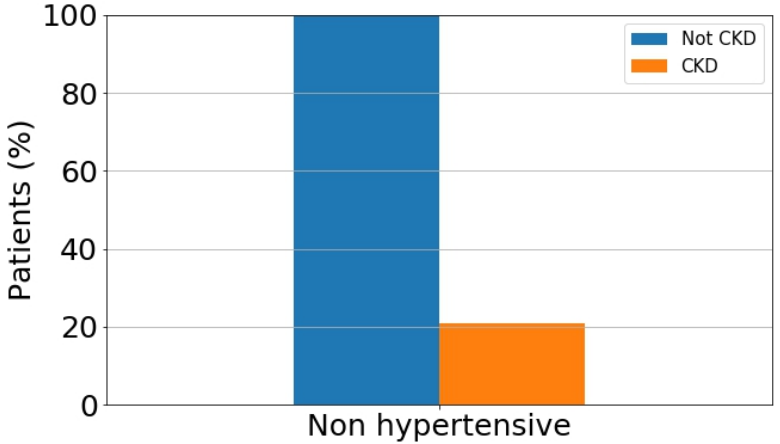

Mean age of healthy people: 57 years

Mean age of sick people: 46 years

We can see that age for healthy people is more evenly distributed while **people with CKD is more concentrated between 40 and 80 years old**

# CKD markers: blood pressure and hypertension



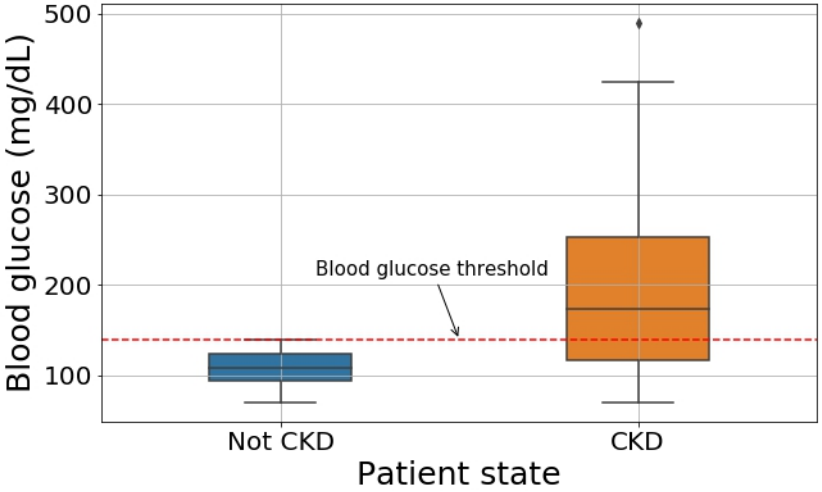

All healthy patients lie below the diastolic hypertension threshold, while 40% of people with CKD are above it.

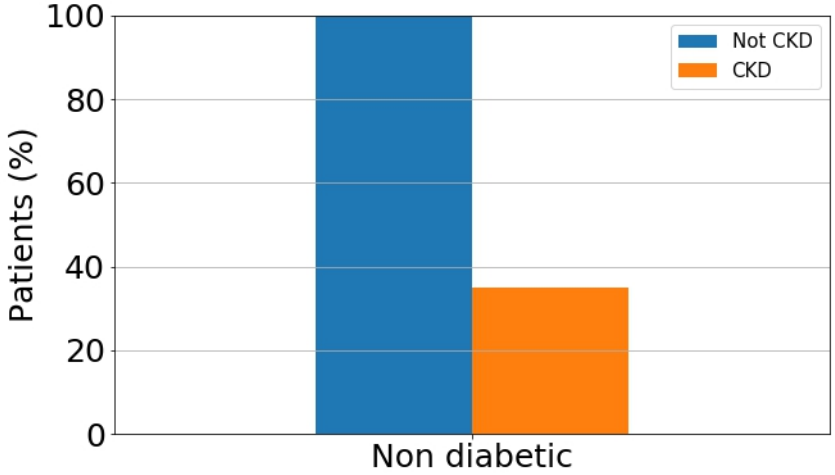
Does this mean only 40% of sick people is hypertensive?

Actually, **almost 80% of CKD patients are hypertensive**: if **either** the systolic or diastolic components lie above their thresholds of 140 and 90 mm Hg, respectively:

Hypertensive

# CKD markers: blood glucose and diabetes

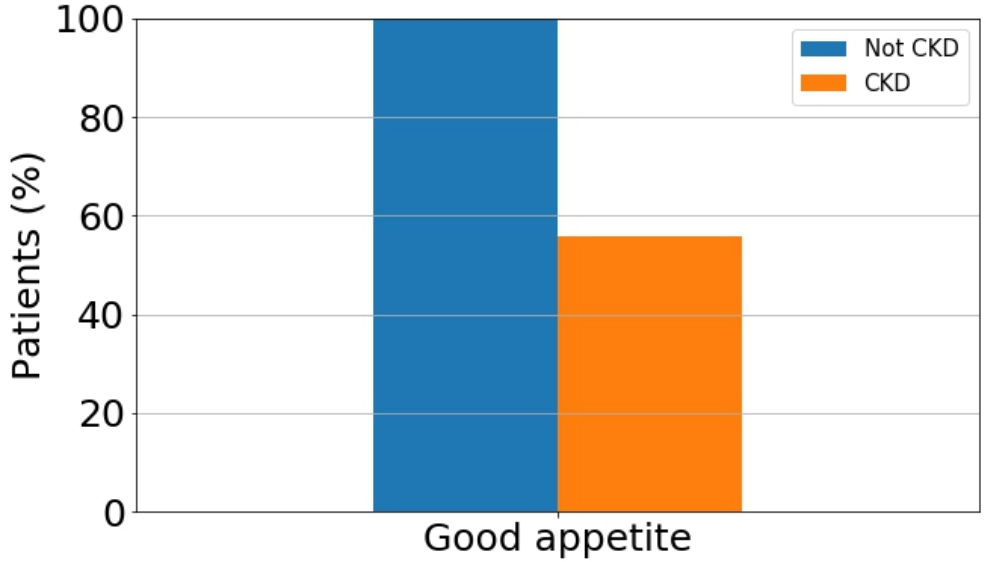



All healthy people lie below the 140 mg/dL diabetes-risk threshold. Meanwhile, **58% of CKD patients have values larger than 140 mg/dL**.

**65% of CKD patients display diabetes**, which is (within a reasonable error) in agreement with the blood glucose levels.

# CKD marker: appetite

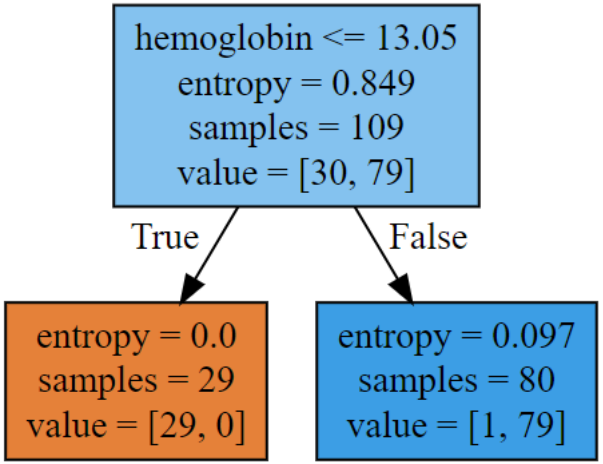



While all patients without CKD have a good appetite, **almost half of the patients with CKD show a poor appetite**.

# Part 2. Predicting Chronic Kidney Disease

# Reduced dataset: CKD prediction with decision tree

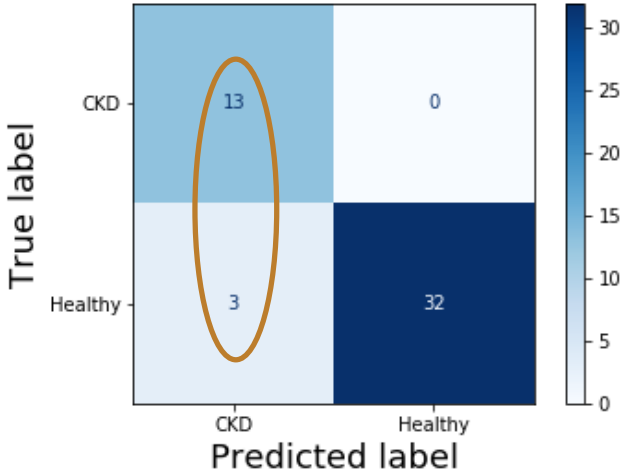
**Optimal tree geometry**




We can see that the most relevant feature, and the only one that was chosen by the model is **hemoglobin**.

**Test accuracy: 97%**

However, since our dataset is imbalanced, confusion matrix shows that ~18% of sick patients were mislabeled and thus, 97% accuracy does not reliably show the model's efficiency:



# Reduced dataset: CKD prediction with logistic regression

## Feature importance

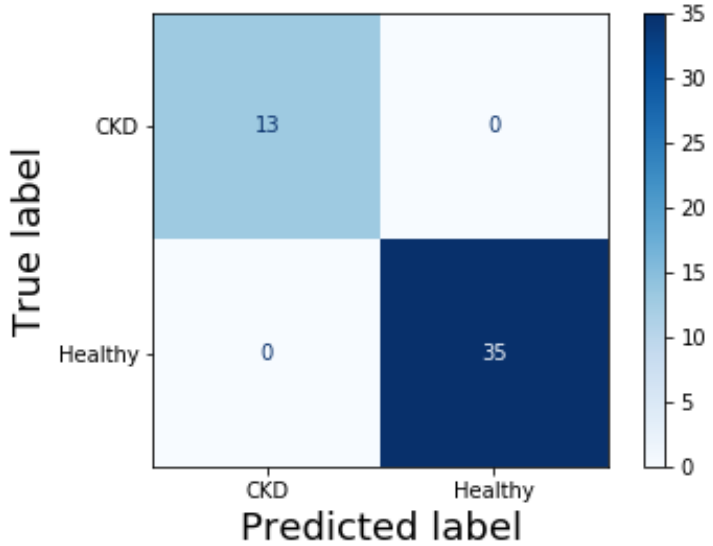
**y=Not CKD** top features

| Weight[?] | Feature |
|---|---|
| +0.477 | hypertension_no |
| +0.474 | red_blood_cells_normal |
| +0.348 | hemoglobin |
| +0.327 | Na |
| +0.311 | pack_cell_vol |
| +0.287 | specif_grav_1.02 |
| +0.272 | pus_cell_normal |
| +0.260 | diabetes_no |
| +0.224 | specif_grav_1.025 |
| … *7 more positive* … | |
| … *11 more negative* … | |
| -0.213 | ser_creatinine |
| -0.225 | blood_pres |
| -0.243 | age |
| -0.260 | diabetes_yes |
| -0.272 | pus_cell_abnormal |
| -0.295 | white_blood_cell_count |
| -0.346 | specif_grav_1.01 |
| -0.420 | specif_grav_1.015 |
| -0.474 | red_blood_cells_abnormal |
| -0.477 | hypertension_yes |
| -0.720 | albumin |

Optimal L2-regularization coefficient *C=1*

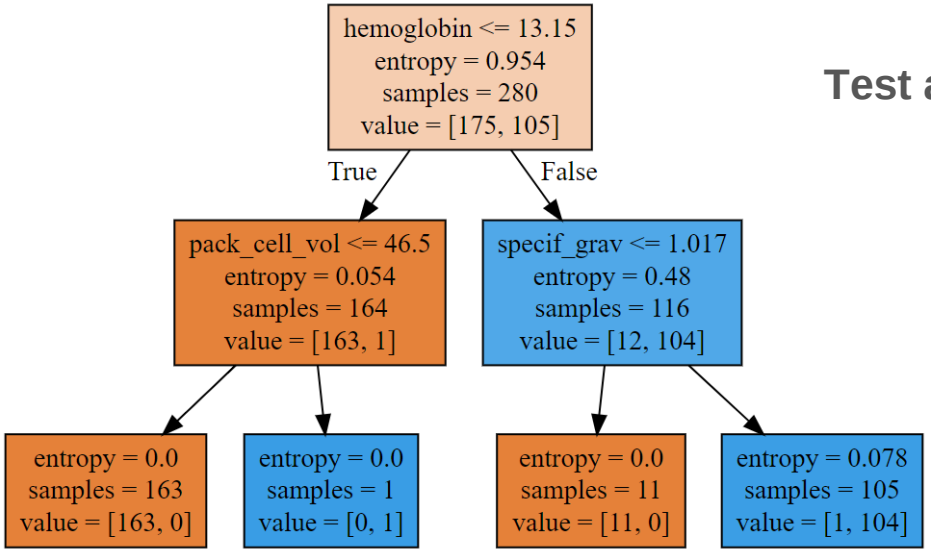
**Hypertension and albumin** amount are the strongest predictors.

**Test accuracy: 100%**



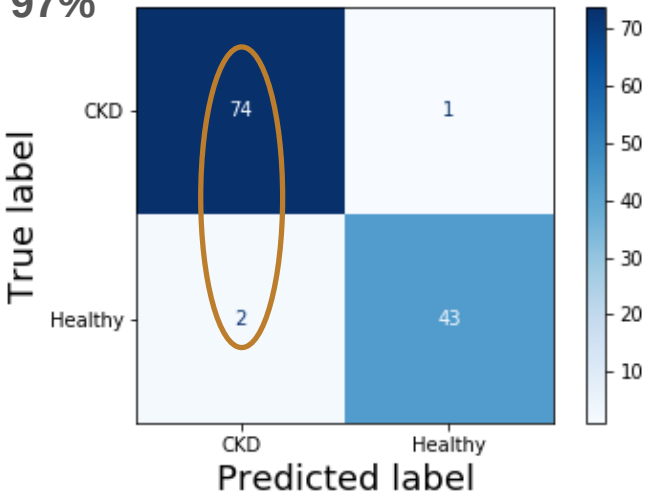
# Complete dataset CKD prediction with decision tree

**Optimal tree geometry**




Although it has the same accuracy as with the reduced dataset, **it is intrinsically better** as the % of false negatives is lower (2.6%):

**Test accuracy: 97%**




The most relevant feature is **hemoglobin,** as with the reduced dataset, followed by **packed cell volume and specific gravity.**

# Complete dataset: CKD prediction with logistic regression

## Feature importance

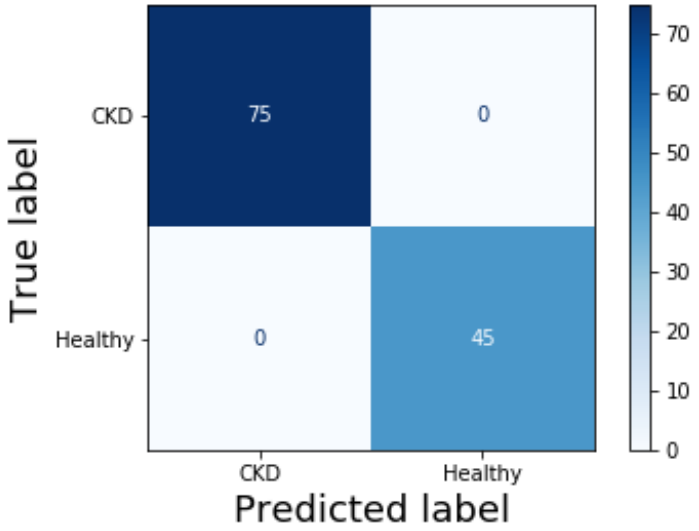
**y=Not CKD** top features

| Weight[?] | Feature |
|---|---|
| +1.738 | specif_grav |
| +1.304 | hemoglobin |
| +1.192 | red_blood_cell_count |
| +0.928 | pack_cell_vol |
| +0.618 | Na |
| +0.471 | hypertension_no |
| +0.465 | diabetes_no |
| +0.388 | red_blood_cells_normal |
| +0.234 | pus_cell_normal |
| +0.194 | pedal_edema_no |
| … 7 more positive … | |
| … 7 more negative … | |
| -0.194 | pedal_edema_yes |
| -0.234 | pus_cell_abnormal |
| -0.388 | red_blood_cells_abnormal |
| -0.431 | blood_pres |
| -0.465 | diabetes_yes |
| -0.471 | hypertension_yes |
| -0.595 | sugar |
| -0.729 | albumin |
| -0.808 | blood_gluc_random |
| -0.864 | ser_creatinine |

Optimal L2-regularization coefficient *C=1*

**Specific gravity and creatinine** amounts are the strongest predictors.

Test accuracy: 100%



# Limitations

All limitations in this work are concerned with the dataset size and properties:

1. Size is of only 400 instances: we may have a limited amount of information that can affect our data exploration and model building.
2. The dataset is imbalanced, with a 37.5% of positive target values and 62.5% of negative ones: the predictive model will tend to fit to the most numerous label and fail on the prediction for the other.

These characteristics make the extracted conclusions from this work an acceptable **preliminary** work to check the **general feature trend** and the **potential predicting power** of such features with different models.

As **future work**, a similar study with a much larger dataset is necessary to be able to extract accurate information on the marker's behavior as a function of the target label and to build a model that can be used in a real-world application. In case the dataset is still imbalanced, an imbalanced imputation method like SMOTE could be applied.

# Conclusions

1. Health markers exploration for home measurements:
   - CKD has a larger prevalence in people between **40-80 years old**
   - **Hypertension is a common** signal of CKD (80% of sick patients)
   - High glucose levels and **diabetes** are directly linked with CKD
   - A 45% of people with CKD show a **poor appetite**
2. Machine learning model:
   - **Model prediction shows great potential** with both decision trees and logistic regression. Logistic regression seems to yield better accuracies but, as already noted, a larger dataset is needed to correctly assess the model's real efficiency and accuracy.
   - Features with the most predictive power were found to be: **hemoglobin, hypertension, blood glucose, packed cell volume, creatinine and albumin levels and specific gravity**