Machine learning report

Chronic Kidney Disease

Sirine Dahech
Zeineb ben salem
Farah Saadaoui
Balkis Melki
Ahmed Jarraya
Baha Mestiri

Division Machine Learning
Department of Computer Science
ESPRIT

Abstract

Chronic kidney disease (CKD) is a global health issue with a high rate of morbidity and mortality and a high rate of disease progression. Because there are no visible symptoms in the early stages of CKD, patients frequently go unnoticed. The early detection of CKD allows patients to receive timely treatment, slowing the disease's progression. Due to its rapid recognition performance and accuracy, machine learning models can effectively assist physicians in achieving this goal. We propose a machine learning methodology for the CKD diagnosis in this paper. This information was completely anonymized. As a reference, the CRISP-DM® model (Cross industry standard process for data mining) was used.

1 Introduction

Chronic kidney disease (CKD) is one of the leading causes of death in recent years, according to a report by the Global Burden of Disease [1]. One in every seven persons has CKD, one of the undiscovered illnesses that have the greatest influence on patients' quality of life and increase the chance of death significantly. The general system of social security in health (SGSSS) has taken chronic kidney disease (CKD) into account [2], as a high-cost pathology for generating a powerful economic impact on the finances of the system, causing a dramatic effect on the quality of life of the patient and their family, including employment repercussions. To reduce the high mortality of CKD, research should be deepened and directed to the initial stages of the disease, analyzing its risk group, with the help of laboratory tests, seeking that patients do not reach the final stages such as dialysis, transplantation, or death [3]. Through automatic learning, the aim is to find a valuable contribution so that an early classification of the disease can be carried out in its initial stages through the results of clinical laboratories, taking advantage of the great potential of automatic learning in the analysis and classification of the data. It is necessary that the technical help tools that are based on data can support the decision-making process in the initial diagnoses quickly, with high precision, and at low cost. With them, the time required for diagnosis is reduced, allowing the patient to receive treatment for the disease before it progresses to a stage of no return.

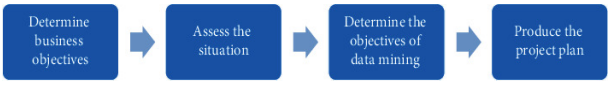
Machine learning is an important field today with mass availability of internet access, and with it the amount of context-specific data that could be analysed to optimize daily pracitices. In this project we will apply and analyse some machine learning algorithms that are bundled in WEKA (Waikato Environment for Knowledge Analysis), to a data set to design and implement a machine learning model that, based on data from clinical laboratories, predicting the possible diagnosis of CKD in its initial stages, helping reduce the mortality rate and costs for the health system

2 Methodology

In the development of this project, the CRISP-DM® model is used, which is the broadest reference guide used in the development of analytical and mining projects to data collected from clinical laboratories. For this, each of the proposed stages will be implemented. The basic methodology can be described by the following figure, which will be further elaborated on, in the following subsections.

2.1 Business Understanding

This phase is divided into four tasks that will help better understand the business, as shown in Figure 1.



2.1.1 Determination of Business Objectives

Chronic kidney disease (CKD) is a type of kidney disease in which there is gradual loss of kidney function over a period of months to years, Initially, there are generally no symptoms; later, symptoms may include leg swelling, feeling tired,

vomiting, loss of appetite, and confusion. Complications include an increased risk of heart disease, high blood pressure, bone disease, and anaemia. CKD is associated with a decrease in kidney function related to age and is accelerated in hypertension, diabetes, obesity, and primary kidney disorders. CKD is a global health problem with a high morbidity and mortality rate, and it induces other diseases. As there are no obvious symptoms during the early stages of CKD, patients often do not notice the disease, this being the main feature, eventually leading to a complete loss of kidney function. Early detection of CKD allows patients to receive timely treatment to improve the progression of this disease. As it has been proposed in the objectives of the work, the aim is to develop an automatic learning model for the prediction in the diagnosis of CKD and to contribute to the reduction of significant complications in the disease such as dialysis processes, kidney transplantation, or reaching death. The main criterion of success for this project, with the help of machine learning, is to identify the behaviors or behavior patterns in the initial stages of CKD to improve the quality of life of patients.

2.1.2 Assessment of the Situation

The idea for the approach of this project arises from the current situation regarding the increase in the confirmatory diagnosis of CKD, and lack of treatment or the user's ignorance of its pathologies leads to irreversible kidney failure in the final stages of CKD, such as dialysis for life, financially affecting the health system, as it is a costly treatment that generates the most significant amount of absorption of the resources available for health . This could be reduced by using tools such as machine learning. Although the application of machine learning in healthcare and other areas is favorable, the field of kidney disease has not yet exploited its full potential.

2.1.3 Determination of the Data Mining Objectives

As referenced in the general objective, the technical terms of this project are to design, implement, and deploy a machine learning model that, based on data from Apollo Hospitals in india, allows to classify the possibility of a diagnosis of CKD. Through the analysis of laboratory studies that are low-cost for health entities, these data reduce the mortality rate and costs of the health system. The medical history and the laboratory tests indicate identifying symptoms or signs that can be used as constitutive variables of the problem in CKD patients on a large scale since a large amount of data can be handled without inconvenience. With the initial data, a description and exploration of these are made, verifying that they can be used or have the minimum information to perform the classification, through the analysis of these data and obtain the patients with an incidence of CKD. With the data obtained, a training set is molded. Several tests are carried out that define or determine the most appropriate technique(s) for the classifier and that the results are practical and efficient. With the defined classifier, the predictive models are trained and validated to establish the model with the highest precision for the data, selecting the one that offers the best results. Predictive models often run calculations during ongoing transactions, for example, to assess the risk or opportunity for a particular patient in a way that provides insight into the treatment decision-making.

2.1.4 Production of the Project Plan

The project plan can be found in the schedule annex-project work plan. It describes all the necessary steps, from the problem statement and data collection to its analysis.

2.2 Study and Understanding of the Data

This section describes the initial data obtained, such as the number of records and fields per record and their identification, each field's meaning, and the initial format's description, as shown in Figure 2.

2.2.1 Collection of Starting Data

The data set used for this project was obtained thanks to Apollo Hospitals, India, and its Senior Consultant Nephrologist, Dr.P.Soundarapandian.M.D.,D.M . They allowed and authorized the treatment of these data. The dataset contains 400 samples (250 CKD, 150 notckd). In this data set, each sample has 27 variables or predictive characteristics.

2.2.2 Verification of the Quality of Data

In this section, data verification was performed to determine the consistency of the field values and the amount of distribution of the null values and to find out-of-range values that can generate noise for the process. This verification process was carried out on the entire data set received. In the fields where no records were found, the unknown fields were changed to a Null value.

2.3 Data Preparation

2.3.1 Data cleaning

By having the information from the data, the focus is on identifying the variables to be used. During the data review, a total of 27 variables were found. Within the data we don't have any null values but we have values which are indicated with "?" . We turned those values into missing values and then we replaced the numerical values with the mean and the categorical with the mode for any missing values.

2.3.2 Data construction

the next step was transforming the numeric and non-numeric attributes. We dropped the ID column. After that we encoded the attributes pcc , rbc , pc , ba , htn, dm, cad, pe and anen to 1 and 0. And for the attribute appetit we replaced ckd by 1 and nonckd by 0 while for the attribute class we replaced good by 1 and poor by 0.

2.3.3 Feature Selection

For the feature selection we made the heatmap to know the corrolation between attributes. From Heatmap and Scatterplot, we can easily observe that PCV and Hemoglobin are highly correlated with 88%. So we can remove anyone of these columns as it is acting like duplicate of another. Also , we can observe that RBC count and PCV are 76 % correlated and RBC count and hemoglobin are 75 % correlated while Blood Urea and Serum Creatinine are 69 % correlated. So we ended up by dropping the column hemo,rbc and sc. We got then 24 attributes. We used after that Recursive Feature Elimination with Cross-Validation which performs recursive feature elimination with cross-validation loop to extract the optimal features to find out 17 optimal attributes.

2.3.4 Data formatting

we used Standardization as a scaling technique where the values are centered around the mean with a unit standard deviation