Chronic disease analyze by the use of spark

Aisha Askarova [1]
Computer Science and Software
SDU
Kaskelen,Kazakhstan
180107112@stu.sdu.edu.kz

Bauyrzhan Taimanov [2]
Computer Science and Software
SDU
Kaskelen,Kazakhstan
180107196@stu.sdu.edu.kz

Adil Khamidullov [3]
Computer Science and Software
SDU
Kaskelen, Kazakhstan
180107260@stu.sdu.edu.kz

***Abstract*—chronic kidney disease (CKD) is one of the diseases with a high mortality rate. This is a disease caused by the loss of kidney function over a long period of time. At the initial stage, the disease does not show any symptoms. In the absence of medical treatment, a person may suffer from other complications, such as high blood pressure, anemia, malnutrition, increased risk of cardiovascular disease, cognitive impairment, and impaired physical function. Automated diagnostics using Big Data analysis methods using Spark is of interest to researchers. In this study, we found the correlation between these health attributes and chronic kidney disease.This way we can allow early detections that facilitate medical interventions.Identify key precursors to chronic kidney disease that can be used for machine learning. For the research, PySpark MLlib was used to build the more robust and scalable machine learning pipeline in distributed systems.**

Keywords—chronic kidney disease, spark ,big data, jupyter notebook,PySpark

I. Introduction

Chronic kidney disease (CKD)—or chronic renal failure (CRF), as it was historically termed—is a term that encompasses all degrees of decreased kidney function, from damage–at risk through mild, moderate, and severe chronic kidney failure [1].

Our kidney is involved in multiple key functions.

1) They help to maintain overall fluid balance
2) They help to regulate and filter minerals from blood
3) They help to filter wastes generated from medications, food and toxic substances
4) They then help in creating hormones that help produce red blood cells, promote bone health, and regulate blood pressure

So what happens if our kidney is damaged?

We have two kidneys. If one doesn't function well, the burdens get carried over to the second kidney. If the patient fails to take drastic measures to improve his/her condition, both kidneys will fail, leading to acute renal failure. This can be fatal without artificial filtering (dialysis) or a kidney transplant. Of course, this occurs at the advanced stage of chronic kidney disease and symptoms will only show up at a severe stage. Kidneys are responsible for filtering waste from the body. It also affects other organs to function well. Waste accumulates to blood, causing the person to feel sick. Side effects of CKD include high blood pressure, anemia, malnutrition, weak bones, and nerve damage [2]. Eventually, it will cause complications such as reduced glomerular filtration rate (GFR), heightened risk of getting cardiovascular disease, and lastly impaired cognitive and physical function[3]. Major causes of CKD include diabetes and high blood pressure. Early-stage CKD is hard to detect, it may not be easily diagnosed due to non-showing of symptoms. Chronic kidney disease prediction is one of the problems in medical decision-making because of difficulty in early-stage detection and high mortality rate. Thanks to the PySpark MLib libraries, we used different analysis methods to identify the correlation between the main attributes and disease incidence, analyze it, and apply the data to predict the diagnosis through classification algorithms in machine learning.

The automated diagnosis of different classes has attracted many researchers [4]. Factors that lead to the classification of patients suffering from chronic and non-chronic kidney disease will be discussed in the next section. Accurate diagnostics made by the classification algorithm will be a great help for medical practitioners, especially those in the field of nephrology.

II. Environments

The proponent used Jupyter Notebook for data analysis. Jupyter is a tool for interactive developing and presenting of data science projects. Libraries and packages used in the study include pandas, matplotlib, numpy, seaborn and pyspark.

- Pandas- is an open source library which provides easy to use and high performing data structures and data analysis tools
- Numpy- is a library for Python programming language, which supports large multidimensional array alongside with high level mathematical functions to operate on those arrays
- Matplotlib – plotting library for Python. Provides object oriented API for embedding plots into applications
- Seaborn- data visualization library for Python. Provides high interface for attractive and informative statistical graphics
- PySpark – Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark.
- SciPy - a free and open-source Python library used for scientific computing and technical computing.

III. Data Exploration

This paper collects Chronic Kidney DiseaseData Set from UCI machine learning repository.The dataset contains 400 numbers of records and each record is formulated as collection of 26 variables. Information regarding these 23 attributes is provided in Table 1.

The attribute 'classification' variable identifies whether the patient has CKD or not. This variable is kept as dependent or target variable during the classification process. The rest variables are fed as input to the classifier model in order to predict the target class.

Data Set Column Information:

Table 1: Parameters and its Description and Allowed Values

Parameters	**Description and Allowed Values**
Age	Discrete Integer Values, Range [2- 90] in the year
Blood Pressure	Discrete Integer Values , Range [50- 180] in mmHg
Specific Gravity	Nominal Values
Albumin	Nominal Values
Sugar	Nominal Values
Pus Cells	Nominal Values, with values of "Normal" and "Abnormal"
Pus Cells Clumps	Nominal Values
Bacteria	Nominal Values, with values of "Present" and "Not Present"
Blood Urea	Discrete Integer Values
Blood Glucose Random	Discrete Integer Values
Serum Creatinine	Numeric Values
Sodium	Discrete Integer Values
Hemoglobin	Numeric Values
Potassium	Numeric Values
Packed Cell Volume	Discrete Integer Values
White Blood Cell Count	Discrete Integer Values
Red Blood Cell Count	Numeric Values
Hypertension	Nominal Values (Yes, No)
Diabetes mellitus	Nominal Values (Yes, No)
Coronary Artery Disease	Nominal Values (Yes, No)
Appetite	Nominal Values (Good, Poor)
Pedal Edema	Nominal Values (Yes, No)
Anemia	Nominal Values (Yes,No)

Attribute Information:

We use 25 + class = 26 (12 numeric ,14 nominal)

Id(numerical) - Patient Id Age(numerical) - age in years

Blood Pressure(numerical) - bp in mm/Hg

Specific Gravity(nominal) - sg - (1.005,1.010,1.015,1.020,1.025)

Albumin(nominal) - al - (0,1,2,3,4,5)

Sugar(nominal) - su - (0,1,2,3,4,5)

Red Blood Cells(nominal) - rbc - (normal,abnormal)

Pus Cell (nominal) - pc - (normal,abnormal)

Pus Cell clumps(nominal) - pcc - (present,notpresent)

Bacteria(nominal) - ba - (present,notpresent)

Blood Glucose Random(numerical) - bgr in mgs/dl

Blood Urea(numerical) -bu in mgs/dl

Serum Creatinine(numerical) - sc in mgs/dl

Sodium(numerical) - sod in mEq/L

Potassium(numerical) - pot in mEq/L

Hemoglobin(numerical) - hemo in gms

Packed Cell Volume(numerical)

White Blood Cell Count(numerical) - wc in cells/cumm=

Red Blood Cell Count(numerical) - rc in millions/cmm

Hypertension(nominal) - htn - (yes,no)

Diabetes Mellitus(nominal) - dm - (yes,no)

Coronary Artery Disease(nominal) - cad - (yes,no)

Appetite(nominal) - appet - (good,poor)

Pedal Edema(nominal) - pe - (yes,no)

Anemia(nominal) - ane - (yes,no)

Class (nominal)- class - (ckd,notckd)

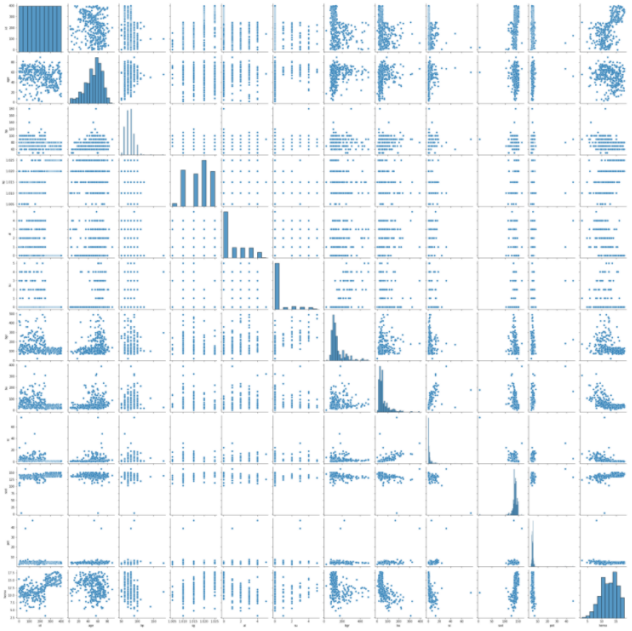

Fig 1.Visualization of dataset with seaborn.

IV. Dataset Preprocessing

Figure 2 shows the Spearman correlation matrix of the dataset. The dataset contains 400 samples with two different classes; "CKD" which is a label of chronic kidney disease and "NOTCKD" for non-chronic kidney disease.

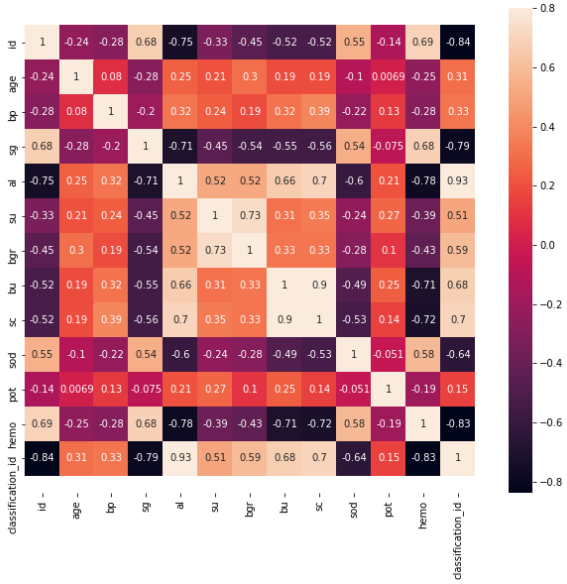

Fig 2. Correlation matrix

CKD has 250 cases while NOTCKD has 150 cases. There are 24 attributes, 11 which are numerical, and 13 are nominal values. From the correlation matrix, it is shown that variables hemoglobin, packed cell volume and red blood cell count has high positive correlations, reporting a Spearman coefficient of 0.8 for red blood cell and 0.9 for packed cell volume.

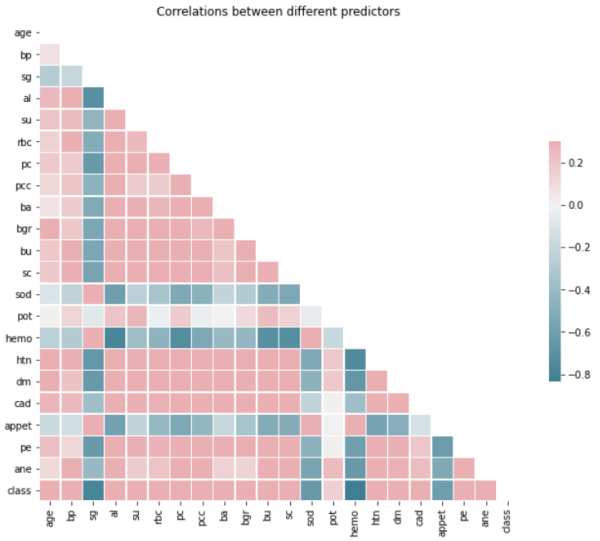

Fig 3. Visualization of the correlation of different predictors

A. Data cleaning

Figure 4 shows the number of missing values per features. The ratio of missing data exceeds the norm. Therefore, we decided to delete the missing data instead of replacing the data.

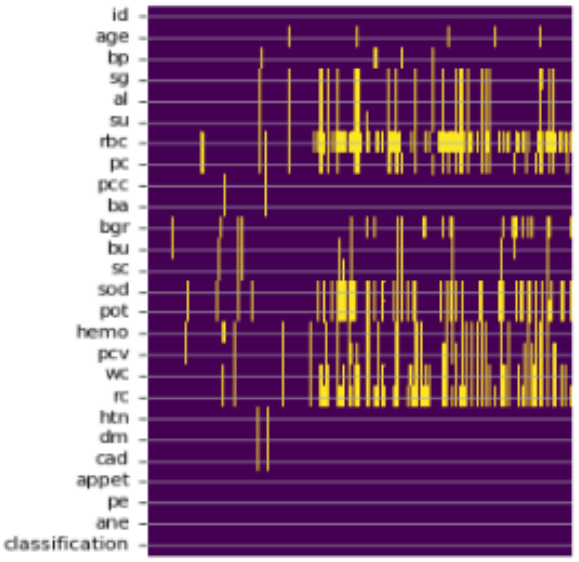

Fig 4. Missing values

V. Analyze by the use of spark

First of all, we import useful libraries to analyze by pyspark.

```
Data overview
root
 |-- id: integer (nullable = true)
 |-- age: integer (nullable = true)
 |-- bp: integer (nullable = true)
 |-- sg: double (nullable = true)
 |-- al: integer (nullable = true)
 |-- su: integer (nullable = true)
 |-- rbc: string (nullable = true)
 |-- pc: string (nullable = true)
 |-- pcc: string (nullable = true)
 |-- ba: string (nullable = true)
 |-- bgr: integer (nullable = true)
 |-- bu: double (nullable = true)
 |-- sc: double (nullable = true)
 |-- sod: double (nullable = true)
 |-- pot: double (nullable = true)
 |-- hemo: double (nullable = true)
 |-- pcv: string (nullable = true)
 |-- wc: string (nullable = true)
 |-- rc: string (nullable = true)
 |-- htn: string (nullable = true)
 |-- dm: string (nullable = true)
 |-- cad: string (nullable = true)
 |-- appet: string (nullable = true)
 |-- pe: string (nullable = true)
 |-- ane: string (nullable = true)
 |-- classification: string (nullable = true)
```

Fig 5. Data overview

Using spark function printSchema() we can overview our data, and see that we have in each column nullable values.

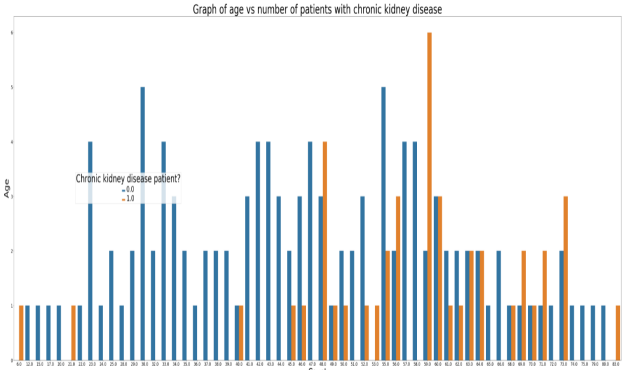

Fig 6. Correlation between age and disease

Correlation between age and whether a patient has chronic kidney disease. As people older as disease symptoms show higher.

We have correlation between red blood cell and whether the patient has chronic kidney disease

```
cont = pd.crosstab(changedTypedf.toPandas()['rbc'],changedTypedf.toPandas()['classification'])
scipy.stats.chi2_contingency(cont)
```

```
(50.26119039594268,
 1.345850460645332e-12,
 1,
 array([[ 13.10126582,   4.89873418],
        [101.89873418,  38.10126582]]))
```

Fig 7. Correlation between red blood and disease.

Our kidneys create an essential hormone called erythropoietin (EPO). EPO are chemical messengers that play a key role in the production of red blood cells. Patients with chronic kidney disease have low EPO, resulting in low levels of red blood cells. This will eventually lead to anemia.

Given that the red blood cell here is a nominal data, we will need to use Chi-square test to calculate correlation.

We will be using a 95% confidence interval (95% chance that the confidence interval you calculated contains the true population mean).

The null hypothesis is that they are independent.

The alternative hypothesis is that they are correlated in some way.

We have correlation between pus cell and whether a patient has chronic kidney disease

```
cont = pd.crosstab(changedTypedf.toPandas()['pc'],changedTypedf.\
                   toPandas()['classification'])
scipy.stats.chi2_contingency(cont)
```

```
(90.54654327784726,
 1.8067640363056176e-21,
 1,
 array([[21.10759494,  7.89240506],
        [93.89240506, 35.10759494]]))
```

Fig 8. Correlation between pus.cel and disease.

We performed the test and we obtained a p-value < 0.05 and we can reject the hypothesis of independence. There seems to be a correlation between the condition of pus cells and whether the patient has chronic kidney disease.

We have correlation between anemia and whether a patient has chronic kidney disease

```
cont = pd.crosstab(changedTypedf.toPandas()['ane'],changedTypedf.\
                   toPandas()['classification'])
scipy.stats.chi2_contingency(cont)
```

```
(43.61151522379983,
 4.004756257274093e-11,
 1,
 array([[103.35443038,  38.64556962],
        [ 11.64556962,   4.35443038]]))
```

Fig 9. Correlation between anemia and disease.

Anemia happens when there are insufficient red blood cells to carry out their duties. Our kidneys produce an important hormone called erythropoietin (EPO). This hormone tells your body to make red blood cells. For CKD patients, their kidneys cannot make enough EPO. Low EPO levels cause low red blood cell count, resulting in anemia.

Given that anemia here is nominal data, we will need to use Chi-square test to calculate correlation.

We will be using a 95% confidence interval (95% chance that the confidence interval you calculated contains the true population mean).

The null hypothesis is that they are independent.

The alternative hypothesis is that they are correlated in some way.

VI. Results

We observed strong correlation between CKD and the following:

- Correlation map between fields
- Red blood cell
- Pus Cell
- Anemia

We have explained most of the sightings, but there seems to be a weird observation for blood urea as we were expecting a positive correlation there. What we can say about that is that the CKD patients were informed about their issues and restricted their diet to reduce excessive nitrogen levels in their blood to a safe level.

VII. Conclusion

CKD classification and detection may be used to give a second opinion to doctors and pathologists. This paper proposed an approach for the analysis of chronic kidney disease Big Data techniques.By analyzing the correlations, you can understand the key attributes for identifying a diagnosis.

Acknowledgment

The researcher would like to express their gratitude to the CSS 439 Distributed Big Data Systems course of SDU Engineering Faculty and also the Azamat Serek for his support, understanding and for interesting material.

References

[1] *T.Anothaisintawee,S.Rattanasiri,A.Ingsathit,J.Attia,and A.Thakkinstian,"Prevalence of chronic kidney disease:A Systematic review and meta analysis, "Clin.Nephrol.,vol.71,no.3,pp.244–254, 2009,doi:10.5414/CNP71244.*

[2] *R. Pecoits-filho, V. Perkovic, M. J. Sarnak, S. W. Tobe, and C. R. V Tomson, "Blood pressure in chronic kidney disease: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference," no. Table 1, pp. 1027– 1036, 2019.*

[3] *C. Brown and J. A. Williams, "Clinical Practice Guideline Anaemia of Chronic Kidney Disease," no. June, 2017.*

[4] *P. S. Mun, H. N. Ting, S. M. Mirhassani, and T. A. Ong, "Prediction of chronic kidney disease using urinary dielectric properties and support vector machine," J. Microw. Power Electromagn. Energy, vol. 50, no. 3, 2016.*