kidney disease

Category	Causal Discovery
Description	The aim is to classify whether a patient has chronic kidney disease or not. The classification is based on a attribute named 'classification' which is either 'ckd'(chronic kidney disease) or 'notckd.
Priority to test	☑
URL	https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset

Introduction

In the realm of medical research, understanding the intricate relationships between various clinical parameters and diseases is paramount. This study is dedicated to exploring a comprehensive dataset containing an array of clinical attributes, with a primary focus on uncovering the factors that significantly contribute to kidney disease. Our objective is to analyze how these parameters not only correlate with the existence of kidney disease but also to assess their potential in predicting its occurrence. Through this analytical journey, we aim to discern the pivotal clinical markers that are most indicative of kidney health.

Dataset (400 rows and 25 features)

id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	48	80	1.02	1	0		normal	notpresen	notpresen	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notpresen	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notpresen	notpresen	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notpresen	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notpresen	notpresen	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
5	60	90	1.015	3	0			notpresen	notpresen	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68	70	1.01	0	0		normal	notpresen	notpresen	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
7	24		1.015	2	4	normal	abnormal	notpresen	notpresen	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
8	52	100	1.015	3	0	normal	abnormal	present	notpresen	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
9	53	90	1.02	2	0	abnormal	abnormal	present	notpresen	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50	60	1.01	2	4		abnormal	present	notpresen	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
11	63	70	1.01	3	0	abnormal	abnormal	present	notpresen	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68	70	1.015	3	1		normal	present	notpresen	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	68	70						notpresen	notpresen	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd

The dataset employed in this analysis was collected over a two-month period in India, incorporating data from patients both with and without kidney disease. Comprising 400 entries initially. Beacause the original dataset includes many missing values and outliers, so it refined down to 213 after data cleaning, this dataset encapsulates a diverse range of clinical attributes. These include fundamental health indicators such as patient age, blood pressure, and more specific markers like serum creatinine and urine specific gravity, among others. Each parameter provides a unique lens through which the state of kidney health can be examined, making this dataset a valuable resource for medical research.

- **Age**: Patient's age
- **Blood Pressure (bp)**: Blood pressure levels
- **Specific Gravity (sg)**: Density of urine
- **Albumin (al)**: Albumin levels in the blood
- **Sugar (su)**: Sugar content in urine
- **Red Blood Cells (rbc)**: Presence of red blood cells in urine
- **Pus Cell (pc)**: Presence of pus cells in urine
- **Pus Cell Clumps (pcc)**: Presence of clumps of pus cells in urine
- **Bacteria (ba)**: Presence of bacteria in urine
- **Blood Glucose Random (bgr)**: Random blood glucose levels
- **Blood Urea (bu)**: Urea levels in blood
- **Serum Creatinine (sc)**: Creatinine levels in serum
- **Sodium (sod)**: Sodium levels
- **Potassium (pot)**: Potassium levels
- **Hemoglobin (hemo)**: Hemoglobin count
- **Packed Cell Volume (pcv)**: Volume percentage of red blood cells
- **White Blood Cell Count (wc)**: Count of white blood cells
- **Red Blood Cell Count (rc)**: Count of red blood cells
- **Hypertension (htn)**: Presence of hypertension

- **Diabetes Mellitus (dm)**: Presence of diabetes mellitus
- **Coronary Artery Disease (cad)**: Presence of coronary artery disease
- **Appetite (appet)**: Appetite levels (good/poor)
- **Pedal Edema (pe)**: Presence of pedal edema
- **Anemia (ane)**: Presence of anemia
- **Class**: Diagnosis class (disease/no disease)

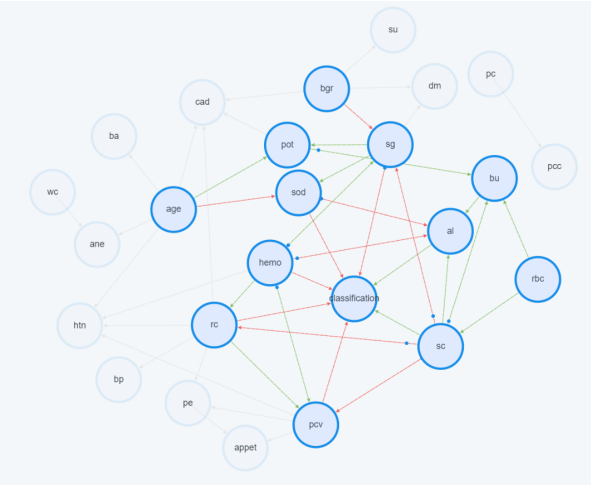
Problems we want to answer

In our analysis, we aim to address the following key questions:

1. How do different clinical parameters correlate with the presence of kidney disease?
2. What are the most significant predictors of kidney disease in the dataset?
3. Can we predict the presence of kidney disease based on these parameters?

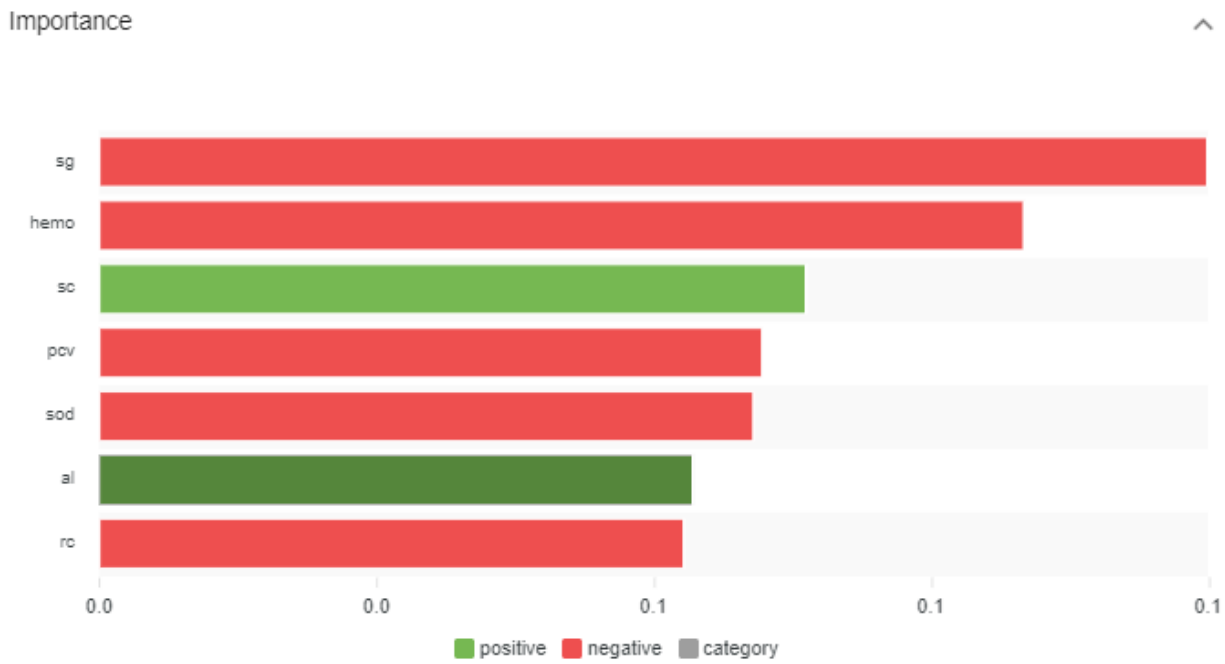
Analysis in Karma

After data cleaning(reduced to 213 row), We Used Karma 360 tool and visualize the relation of these factor. We could also see the information from every node.



Information_node

Features	Value
Name	classification
Datatype	integer
total	212
empty	0
valid	212
mean	0.5
std	0.5
min	0
25%	0
50%	0.5
75%	1
max	1



- **Correlational Analysis**: If you clicked the node in Karma 360, you can see all the causes of ckd(**chronic kidney disease**) from the karma.

Includes↓ (From most important to less important):

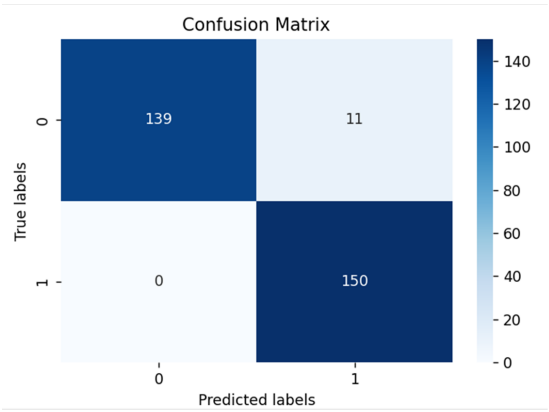
1. sg(**Specific Gravity): importance: -0.127**
2. hemo(**Hemoglobin): importance: -0.106**
3. sc(**Serum Creatinine): importance: 0.081**
4. pcv(**Packed Cell Volume): importance: -0.076**
5. sod(**Sodium): importance: -0.075**
6. al(**Albumin): importance: 0.068**
7. rc(**Red Blood Cell Count): importance: -0.067**

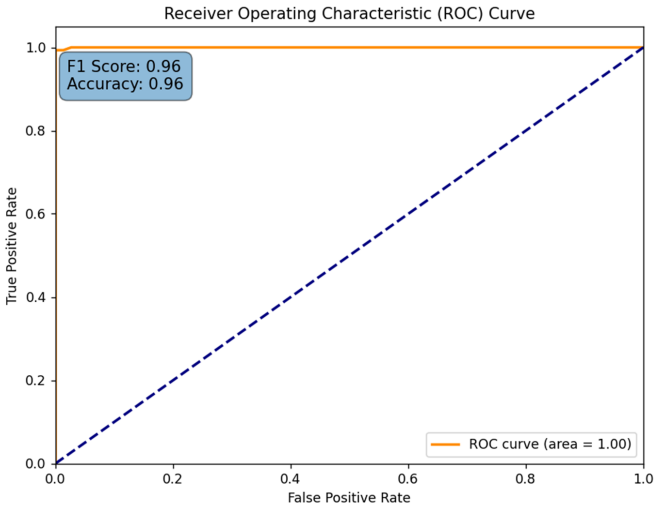
If the factor is red, that means it is inversely proportional to the node, which is that the lower values can cause higher value of the node. If it is green, means that it is inversely proportional to the node

- **Feature Importance**: The most crucial indicator, specific gravity, normal ranges between 1.005 to 1.030. Our analysis found that in individuals without

kidney disease, sg values consistently exceeded 1.020, aligning with our predictive findings- the lower sg means the higher ckd rate.

- **Predictive Modeling**: The predictive model demonstrated high accuracy, as evidenced by the confusion matrix and ROC curve. This performance underscores the model's capability to reliably predict kidney disease based on clinical parameters.





Result

Based on the causal graphs and predictive models generated:

- We highlight significant factors that is directly affected to ckd such as **Specific Gravity and Hemoglobin.**
- We discuss the predictive accuracy of our models and how they could be used in clinical settings to improve diagnostics.

Our comprehensive analysis has not only highlighted significant predictors of kidney disease but also enhanced our understanding of how clinical parameters interrelate with the disease. The predictive models developed could serve as crucial tools in clinical settings, potentially improving early diagnosis and treatment strategies. This study underscores the importance of multifactorial analysis in understanding complex health conditions like kidney disease, providing a blueprint for further research and application in healthcare.

Resource link

Karma 360: http://192.168.50.3:28081/

Dataset: https://www.kaggle.com/datasets/akshayksingh/kidney-disease-dataset/data

Glossary: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
Another CKD dataset: https://www.kaggle.com/datasets/abhia1999/chronic-kidney-disease/data