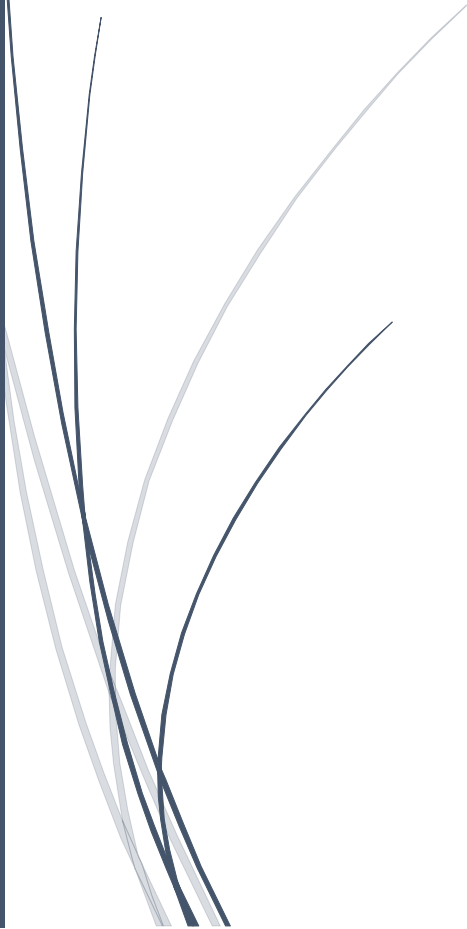


A dark blue vertical bar runs down the left side of the slide. A blue arrow points to the right from this bar, containing the date.

12/19/2021

## Chronic Kidney Disease using Machine Learning Prediction Models



# Index

<i>Abstract:</i>	-----	(2)
<i>Introduction:</i>	-----	(2)
<i>Goal:</i>	-----	(3)
<i>Related work:</i>	-----	(3)
<i>Methods:</i>	-----	(4)
<i>Analysis:</i>	---- -----	(5)
<i>Model Development:</i>	-----	(6)
<i>Result:</i>	-----	(6)
<i>Testing for Real World:</i>	-----	(8)
<i>Front-End:</i>	-----	(9)
<i>Discussion:</i>	-----	(10)
<i>Conclusions:</i>	-----	(10)
<i>Reference:</i>	-----	(11)
<i>Appendices:</i>	-----	(11)

# Chronic Kidney Disease using Machine Learning Prediction Models

## Abstract:

Because of the increasing number of patients, the high risk of progression to end-stage renal disease, and the poor prognosis of morbidity and mortality, chronic kidney disease (CKD) places a significant burden on the healthcare system. The goal of this research is to create a machine-learning model that can predict the early-stage chronic kidney diseases. A total of 400 people were chosen. The Random Forest model performed best with a 0.989 accuracy. The model proposed in this study may be useful for policymakers in predicting CKD trends in the population. The models can enable close monitoring of people at risk, early detection of CKD, better resource allocation, and patient-centred management.

**Keywords:** chronic kidney disease; machine learning.

---

## Introduction:

In this project, we will write a Python programme to predict and classify whether a patient has chronic kidney disease (CKD). utilizing various machine learning techniques

Chronic kidney disease, also known as chronic kidney failure, refers to the progressive loss of kidney function. Wastes and excess fluids in your blood are filtered by your kidneys and excreted in your urine. When chronic kidney disease progresses, dangerous levels of fluid, electrolytes, and wastes can accumulate in your body.

Chronic kidney disease treatment focuses on slowing the progression of kidney damage, usually by addressing the underlying cause. Chronic kidney disease can lead to end-stage kidney failure, which is fatal in the absence of artificial filtering (dialysis) or a kidney transplant.

The study was limited by numbers because it was based on data. As a result, it limits the model's generalizability to the global population or a different region. The presence of noise in the data due to human and technical errors that are difficult to identify may also have an impact on the model's performance.

**Goal:**

Our kidneys are responsible for filtering blood and removing waste in the form of urine. It also known as chronic kidney failure, Dangerous quantities of fluid, and wastes can build up in the body at advanced stages. Patients must then undergo dialysis or consider a transplant if this occurs.

Our goal in this experiment is to investigate if we can use 24 variables to predict whether a patient will develop chronic kidney disease or not. We may be able to recognize and help patients at risk of kidney failure if we can identify characteristics that have a strong influence on kidney failure.

These ML-based models, in the opinion of policymakers, could be efficiently used in resource management and initiating public health initiatives such as closely monitoring and early detection of CKD. Clearly, for such models to be used in clinical practise with individual patients, the feature set would need to be expanded to include laboratory measurements and possibly lifestyle information, which will be addressed in future research.

**Related work:**

**Research 1: Chronic Kidney Disease Prediction Using Machine Learning Methods by 2020 Moratuwa Engineering Research Conference (MERCon)**

**Link =**

[https://www.researchgate.net/publication/344319206\\_Chronic\\_Kidney\\_Disease\\_Prediction\\_Using\\_Machine\\_Learning\\_Methods](https://www.researchgate.net/publication/344319206_Chronic_Kidney_Disease_Prediction_Using_Machine_Learning_Methods)

**research 2: Fuzzy logic: A tool to predict the Renal diseases by Research Journal of Pharmacy and Technology**

**Link =**

[https://www.researchgate.net/publication/351923032\\_Fuzzy\\_logic\\_A\\_tool\\_to\\_predict\\_the\\_Renal\\_diseases](https://www.researchgate.net/publication/351923032_Fuzzy_logic_A_tool_to_predict_the_Renal_diseases)

**research 3: chronic kidney disease prediction using machine learning techniques by 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)**

**Link =**

[https://www.researchgate.net/publication/355762024\\_Chronic\\_kidney\\_disease\\_prediction\\_using\\_machine\\_learning\\_techniques](https://www.researchgate.net/publication/355762024_Chronic_kidney_disease_prediction_using_machine_learning_techniques)

**On above research we found a way to do kidney dialysis using machine learning methods and techniques with remarkable accuracies.**

## Methods:

### ➤ Dataset:

- Apollo Hospitals in Tamil Nadu, India, gathered the data for this study. Each patient's records consist of information about CKD and NON-CKD and their value of the different attributes is included in the research database. The Supplementary Materials contain additional analysis of the dataset.

### ➤ Dataset's Attribute Information:

- We use 24 + class = 25 (11 numeric ,14 nominal)
- 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
- Haemoglobin(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)
- Anaemia(nominal) = ane - (yes, no)

### ➤ Targeted Feature


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

➤ **Data pre-processing:**

  
REPLACE  
NULL VALUES

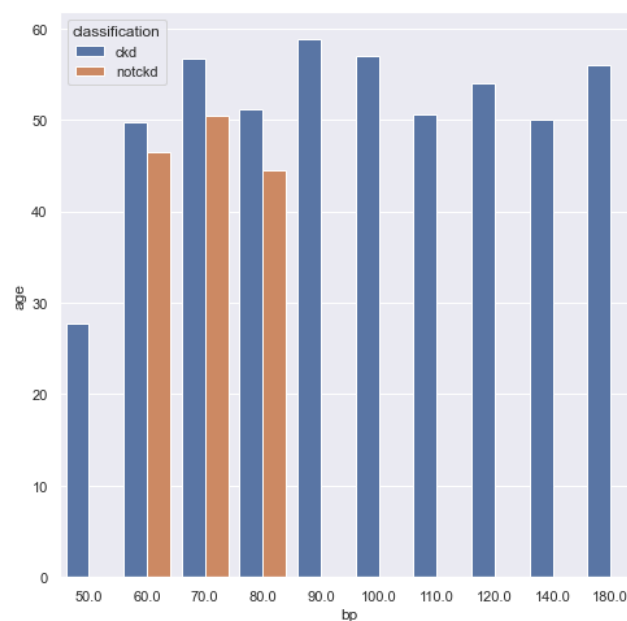
  
DROP THE  
UNNECESSARY  
COLUMNS

  
CHECK THE  
SKEWNESS

  
CORRELATION  
BETWEEN  
FEATURES

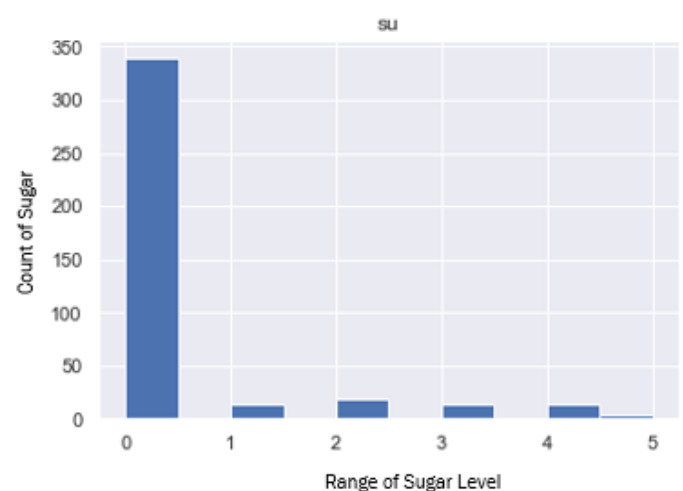
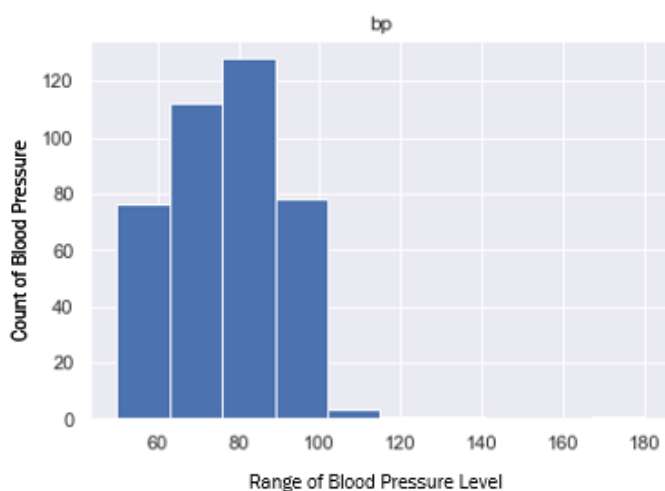
**Figure1. data pre-processing**

➤ **Analysis:**



**Figure 2. Chronic Kidney Disease effect by Age and Blood Pressure**

Here we observed that most of the cases of the non - CKD was discover from the age of 45 to 50. And between the blood pressure of 60 to 80.



**Figure 3. Distribution of Blood Pressure  
Distribution of Sugar**

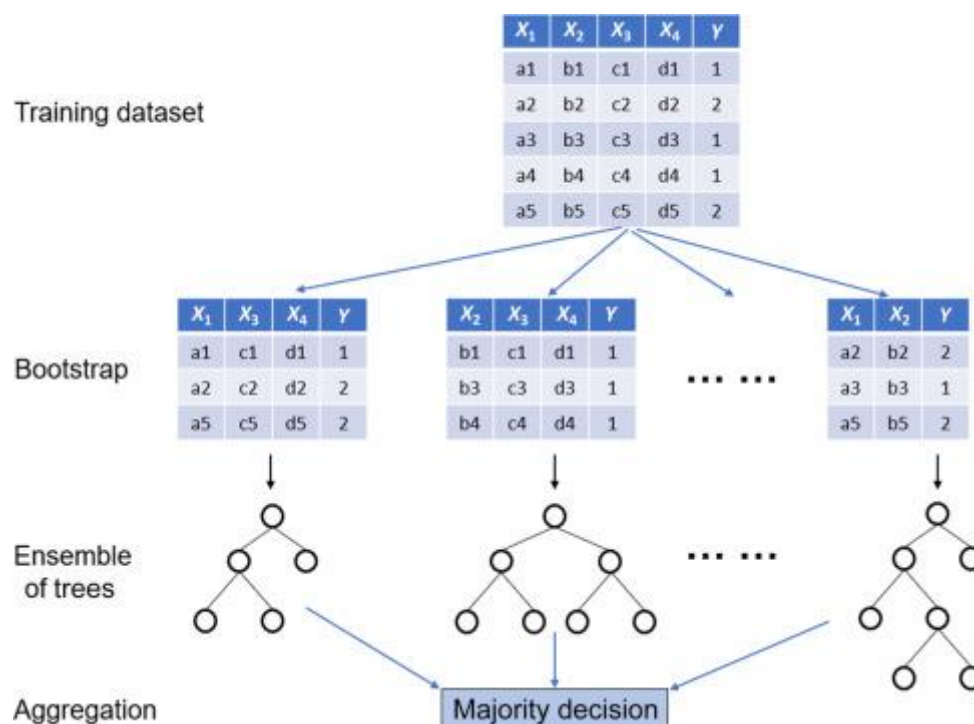
**Figure 4.**

### ➤ Model Development

We used various modeling algorithms from packages like Scikit-learn (decision tree, random forest), K-Nearest Neighbors Algorithm, Support Vector Machine, Multilayer perceptron.

### ➤ Result

Among the methods listed above, we used random forest as a final model. We split the dataset into an 70% training set and a 30% test set.



**Figure 5. random forest model architecture**

Then, we trained our models on the training set and reported their performance on the test set.

### ➤ Train Result:

Accuracy Score: 100.00%

Classification Report: Precision Score: 100.00%

Recall Score: 100.00%

F1 score: 100.00%

Confusion Matrix:

```
[[44  0]
 [ 0 76]]
```

➤ Test Result:

=====

Accuracy Score: 98.21%

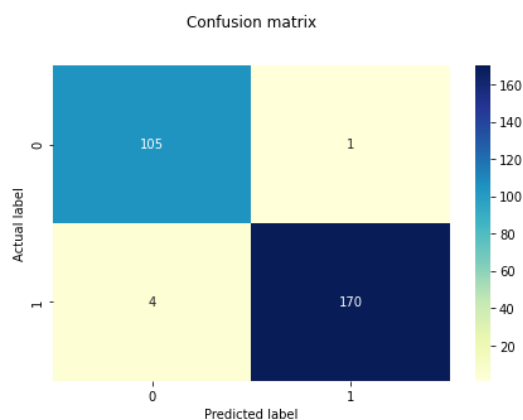
-----

Classification Report: Precision Score: 99.42%

Recall Score: 97.70%

F1 score: 98.55%

precision	recall	f1-score	support	
0.0	0.96	0.99	0.98	106
1.0	0.99	0.98	0.99	174
accuracy			0.98	280
macro avg	0.98	0.98	0.98	280
weighted avg	0.98	0.98	0.98	280



**Figure 6. random forest model confusion matrix**

Sensitivity (TPR) : [0.99056604 0.97701149]

Specificity (TNR) : [0.97701149 0.99056604]

Overall accuracy

Accuracy : [0.98214286 0.98214286]

Accuracy: 0.9821

Average Sensitivity: 0.9906

Average Specificity: 0.9770

=====

True Positives: 170

True Negatives: 105

False Positives: 1

False Negatives: 4

-----

Accuracy: 0.98



Mis-Classification: 0.02

Sensitivity: 0.98

Specificity: 0.99

Precision: 0.99

f\_1 Score: 0.98

### ➤ Other Machine Learning Models:

#### Split Ratio 70:30

Machine Learning Model	Accuracy Score	
	Train Dataset	Test Dataset
Decision Tree Classifier	100	0.917857
KNN	83.33	0.710714
SVM	100	0.621429
MLP	63.33	0.621429

**Figure 7. other Machine Learning Models with Accuracy**

### ➤ Testing the Model for Real World

#### Testing the model with demo data

```
In [280]: dt = {'age':[50], 'bp':[90], 'sg':[1.03], 'al':[4], 'rbc':[0], 'su':[2], 'pc':[0], 'pcc':[1], 'ba':[0], 'bgr':[106], 'bu':[56],
'sc':[2.7], 'sod':[142], 'pot':[3.4], 'hemo':[10.8], 'pcv':[16], 'wc':[9600], 'rc':[3.8], 'htn':[1], 'dm':[0], 'cad':[0],
'appet':[1], 'pe':[1], 'ane':[0]}

data_preds = pd.DataFrame(dt)

In [281]: data_preds.head()

Out[281]:
```

	age	bp	sg	al	rbc	su	pc	pcc	ba	bgr	...	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane
0	50	90	1.03	4	0	2	0	1	0	106	...	10.8	16	9600	3.8	1	0	0	1	1	0

```
1 rows x 24 columns

In [282]: y_pred=clf.predict(data_preds)

In [283]: label = {0:"no risk",1:"risk"}
label[y_pred[0]]

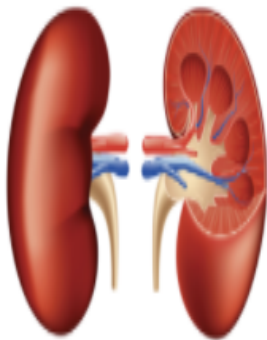
Out[283]: 'risk'
```

- **When we are putting the random values in the random forest model, we can accurately predict that the patient has a CKD or NOT CKD.**

➤ Front-End

Kidney Disease Prediction

[[ prediction\_level ]]



Age	Hypertension
<input type="text" value="Enter age"/>	<input type="checkbox"/> Yes
	<input type="checkbox"/> No
Blood pressure	Red blood cell
<input type="text" value="60 to 180"/>	<input type="checkbox"/> Abnormal
	<input type="checkbox"/> Normal
Specific gravity	Diabetes mellitus
<input type="text" value="between 1.000 and 1.035"/>	<input type="checkbox"/> Yes
	<input type="checkbox"/> No
Albumin	Coronary artery disease
<input type="text" value="between 1 to 5"/>	<input type="checkbox"/> Yes
	<input type="checkbox"/> No
Sugar	Appetite
<input type="text" value="between 1 to 5"/>	<input type="checkbox"/> Good
	<input type="checkbox"/> Poor
Blood glucose random	Pericard edema
<input type="text" value="22 mg/dL to 690 mg/dL"/>	<input type="checkbox"/> Yes
	<input type="checkbox"/> No
Blood urea	Anemia
<input type="text" value="6 to 25 mg/dl"/>	<input type="checkbox"/> Yes
	<input type="checkbox"/> No
Serum creatinine	Protein
<input type="text" value="0.4 to 76"/>	<input type="checkbox"/> Normal
	<input type="checkbox"/> Abnormal
Sodium	Protein clumps
<input type="text" value="4 to 163"/>	<input type="checkbox"/> present
	<input type="checkbox"/> Not Present
Potassium	Bacteria
<input type="text" value="2.5 to 4.7"/>	<input type="checkbox"/> Present
	<input type="checkbox"/> Not Present
Hemoglobin	Red blood cells
<input type="text" value="51 to 17.0"/>	<input type="checkbox"/> Abnormal
	<input type="checkbox"/> normal
Packed cell volume	
<input type="text" value="9 to 54"/>	
White blood cell	
<input type="text" value="2200 to 26400"/>	
<input type="button" value="Predict"/>	

### ➤ Discussion:

The random forest model outperforms the other machine-learning models. This is intriguing because they are not as well established in the prediction of health risks. Given that our networks were not particularly large or complex, the most likely explanation was that they made use of temporal information.

Unlike the commonly used approach, which processes laboratory data in addition to other patient data, we developed a method that does not require laboratory data and instead processes only patients' diagnoses, prescriptions, and basic demographic data (i.e., age and gender), because such data is typically available on a larger scale. This was markedly different from previous studies that relied on laboratory values.

Because CKD typically manifests in older people, it is reasonable to consider age to be a strong predictor. Diabetes mellitus, gout, chronic glomerulonephritis, and essential hypertension are all associated with decreased kidney function.

This clearly correlates with age, as older people have more conditions that necessitate polymedications, and CKD is more common in older age groups.

### ➤ Conclusions:

In this study, we created and tested a number of artificial intelligence-based models that took into account various variables such as gender, age, comorbidities, and medications.

These models forecast a patient's likelihood of developing chronic kidney disease. The Random Forest Classifier outperformed the other models tested, with an AUROC metric of 0.98.

We examined the RF model to determine which features are most important for prediction. Diabetes mellitus, age, blood pressure, and rbc were the most prominent features, all of which are reasonable considering CKD.

These ML-based models, in the opinion of policymakers, could be efficiently used in resource management and initiating public health initiatives such as closely monitoring and early detection of CKD.

Clearly, for such models to be used in clinical practise with individual patients, the feature set would need to be expanded to include laboratory measurements and possibly lifestyle information, which is beyond the scope of the future.

### ➤ Contributions

- Darshan N Siddhpura
  - Project Documentations, Presentations, Project Research
- Dhrumil Shah
  - Exploratory Data Analysis, Coding, Model Research
- Kushal Modi
  - Documentation, Data collections, Project Research

- **Priyank Patel**

- **Front-End Design, Model Testing, Model Implementation**

- **Reference:**

- **Dataset**

- [https://archive.ics.uci.edu/ml/datasets/chronic\\_kidney\\_disease](https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease)

- <https://dom-pubs.onlinelibrary.wiley.com/doi/full/10.1111/dom.14178>

- <https://www.sciencedirect.com/science/article/pii/S2468024919313877>

- <https://journals.sagepub.com/doi/full/10.1177/2054358118776326>

- <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250467#sec005>

- **Image**

- <https://ars.els-cdn.com/content/image/3-s2.0-B9780128177365000090-f09-17-9780128177365.jpg>

- **Reports**

- [https://www.researchgate.net/publication/344319206 Chronic Kidney Disease Prediction Using Machine Learning Methods](https://www.researchgate.net/publication/344319206_Chronic_Kidney_Disease_Prediction_Using_Machine_Learning_Methods)

- [https://www.researchgate.net/publication/351923032 Fuzzy logic A tool to predict the Renal diseases](https://www.researchgate.net/publication/351923032_Fuzzy_logic_A_tool_to_predict_the_Renal_diseases)

- [https://www.researchgate.net/publication/355762024 Chronic kidney disease prediction using machine learning techniques](https://www.researchgate.net/publication/355762024_Chronic_kidney_disease_prediction_using_machine_learning_techniques)

- **Appendices:**

**Testing the model with demo data**

```
In [280]: dt = {'age':[50], 'bp':[90], 'sg':[1.03], 'al':[4], 'rbc':[0], 'su':[2], 'pc':[0], 'pcc':[1], 'ba':[0], 'bgr':[106], 'bu':[56],
              'sc':[2.7], 'sod':[142], 'pot':[3.4], 'hemo':[10.8], 'pcv':[16], 'wc':[9600], 'rc':[3.8], 'htn':[1], 'dm':[0], 'cad':[0],
              'appet':[1], 'pe':[1], 'ane':[0]}

data_preds = pd.DataFrame(dt)
```

```
In [281]: data_preds.head()
```

```
Out[281]:
```

	age	bp	sg	al	rbc	su	pc	pcc	ba	bgr	...	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane
0	50	90	1.03	4	0	2	0	1	0	106	...	10.8	16	9600	3.8	1	0	0	1	1	0

1 rows x 24 columns

```
In [282]: y_pred=clf.predict(data_preds)
```

```
In [283]: label = {'0':"no risk",1:"risk"}
          label[y_pred[0]]
```

```
Out[283]: 'risk'
```

**Thank you**