|  |  |
| --- | --- |
| **Chronic\_Kidney\_Disease Data Set** *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/00336/), [Data Set Description](https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease)  **Abstract**: This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period. |  |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 400 | **Area:** | N/A |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 25 | **Date Donated** | 2015-07-03 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 147333 |

**Data Set Information:**

We use the following representation to collect the dataset  
age - age  
bp - blood pressure  
sg - specific gravity  
al - albumin  
su - sugar  
rbc - red blood cells  
pc - pus cell  
pcc - pus cell clumps  
ba - bacteria  
bgr - blood glucose random  
bu - blood urea  
sc - serum creatinine  
sod - sodium  
pot - potassium  
hemo - hemoglobin  
pcv - packed cell volume  
wc - white blood cell count  
rc - red blood cell count  
htn - hypertension  
dm - diabetes mellitus  
cad - coronary artery disease  
appet - appetite  
pe - pedal edema  
ane - anemia  
class - class

**Attribute Information:**

We use 24 + class = 25 ( 11 numeric ,14 nominal)  
1.Age(numerical)  
age in years  
2.Blood Pressure(numerical)  
bp in mm/Hg  
3.Specific Gravity(nominal)  
sg - (1.005,1.010,1.015,1.020,1.025)  
4.Albumin(nominal)  
al - (0,1,2,3,4,5)  
5.Sugar(nominal)  
su - (0,1,2,3,4,5)  
6.Red Blood Cells(nominal)  
rbc - (normal,abnormal)  
7.Pus Cell (nominal)  
pc - (normal,abnormal)  
8.Pus Cell clumps(nominal)  
pcc - (present,notpresent)  
9.Bacteria(nominal)  
ba - (present,notpresent)  
10.Blood Glucose Random(numerical)  
bgr in mgs/dl  
11.Blood Urea(numerical)  
bu in mgs/dl  
12.Serum Creatinine(numerical)  
sc in mgs/dl  
13.Sodium(numerical)  
sod in mEq/L  
14.Potassium(numerical)  
pot in mEq/L  
15.Hemoglobin(numerical)  
hemo in gms  
16.Packed Cell Volume(numerical)  
17.White Blood Cell Count(numerical)  
wc in cells/cumm  
18.Red Blood Cell Count(numerical)  
rc in millions/cmm  
19.Hypertension(nominal)  
htn - (yes,no)  
20.Diabetes Mellitus(nominal)  
dm - (yes,no)  
21.Coronary Artery Disease(nominal)  
cad - (yes,no)  
22.Appetite(nominal)  
appet - (good,poor)  
23.Pedal Edema(nominal)  
pe - (yes,no)  
24.Anemia(nominal)  
ane - (yes,no)  
25.Class (nominal)  
class - (ckd,notckd)

**problem faced:**

a large no of cells in data set were missing, we tried to fill those columns using average by considering age and other factors which can influence the value in that columns.

Since the data set deals with medical informations , we think that its not practical to fill those columns by taking the average. Because there may be a chance that the instances that the person is treated as a kidney patient could be as a result of a missing value, and thus its not practical to fill those columns by taking the average .About 90% of data in some columns were missing,anyhow we tried to create models in both ways... like filling missing values by taking average as well as eliminating those columns.

Then we found out that those columns which we filled were inappropriate and thus we get less accuracy, precision and f1 score. Thus we eliminate rows with missing values and then the data set with 400 rows get reduced to 156 and then we get 100% accuracy,precision and f1 score.

**GROUP MEMBERS**

VYSAGHAN.KM

ALEN JOSEPH

ANSHAD RAHMAN P