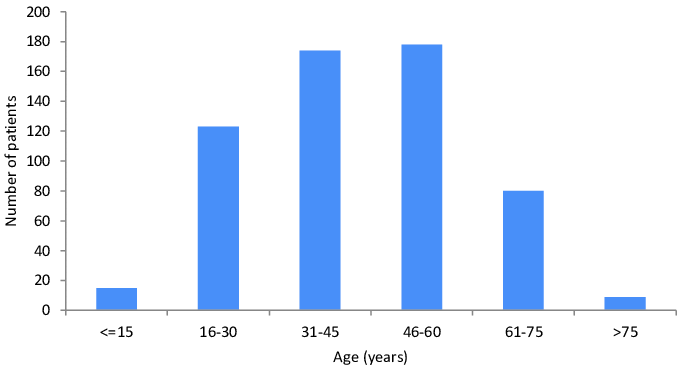
**PREDICTION OF CHRONIC KIDNEY DISEASES USING MACHINE LEARNING**

1. **Abstract:**

With increase in life expectancy and prevalence of life style there is an increase of Chronic kidney diseases. The failure or the malfunctioning of the kidney is known as Chronic Kidney Disease. Diabetes and hypertension account for more than a 60% of the cases in the world. This chronic kidney disease is majorly seen in aged people. Even though there where many precautions and preventive steps taken by the Indian government, there are around 800 per million population cases registered. Though this is not major cause compared to Cardiovascular diseases, this is also a cause of many deaths. Here in this paper we propose a prediction system which predicts the presence of Chronic Kidney disease using the patients or victims data using machine learning algorithm.

1. **Introduction:**

Machine learning is one of the applications of artificial intelligence (AI) that provides computers, the ability to learn automatically and improve from experience instead of explicitly programmed. It focuses on developing computer programs that can access data and use it to learn from themselves. The main aim is to allow computers to learn automatically without human intervention and also adjust actions accordingly. Chronic Kidney diseases had a great deal of attention in medical research. The diagnosis of kidney disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective.



1. **Literature Survey:**

Machine Learning is widely used in the medical field such as prediction of chronic kidney disease since it is a multidisciplinary field. Using machine leraning researchers are developing various techniques in-order to predict the chronic kidney diseases with high accuracy. Large no. of research work is carried out for medical diagnosis for various diseases

S.DilliArasu and Dr. R. Thirumalaiselvi [3] has worked on missing values in a dataset of chronic Kidney Disease. Missing values in dataset will reduce the accuracy of our model as well as prediction results. They find solution over this problem that they performed a recalculation process on CKD stages and by doing so they got up with unknown values. They replaced missing values with recalculated values.

Pinar Yildirim [8] searches the effect of class imbalance when we train the data by using development of neural network algorithm for making medical decision on chronic kidney disease. In this proposed work, a comparative study was performed using sampling algorithm. This study reveals that the performance of classification algorithms can be improved by using the sampling algorithms. It also reveals that the learning rate is a crucial parameter which significantly effect on multilayer perceptron.

Sahil Sharma, Vinod Sharma, and Atul Sharma [9], has assessed 12 different classification algorithm on dataset which having 400 records and 24 attributes. They had Compared their calculated results with actual results for calculating the accuracy of prediction results. They used assessment metrics like accuracy, sensitivity, precision and specificity. They find that the decision tree technique gives accuracy up to 98.6%, sensitivity of 0.9720, and precision of 1 and specificity of 1.

Charleonnan et al. [5] employ decision tree, logistic regression, support vector machine (SVM), and k-nearest neighbor (KNN) as classifiers for CKD detection using a datasetwith2classes,400 instances and 24 attributes. They use a CKD dataset from the UCI machine learning repository. The results show the SVM technique as the better detection technique for detection accuracy and sensitivity.

1. **Proposed System:**

**4.1Data Preprocessing:**

A significant-world data typically contains noises, lack of values and perhaps

in some unsuitable layout that is not generally applicable to machine learning

models.

**4.1.1Data Description**

The chronic kidney disease dataset is downloaded from the Kaggle repository which constitutes of 25 attributes with 400 records of data as listed below:

|  |  |  |
| --- | --- | --- |
| **Attribute Symbol** | **Attribute Type** | **Description** |
| age | age in years | Numerical |
| bp | Blood Pressure (bp in mm/Hg) | Numerical |
| sg | Specific Gravity (1.005,1.010,1.015,1.020,1.025) | Nominal |
| al | Albumin (0,1,2,3,4,5) | Nominal |
| su | Sugar (0,1,2,3,4,5) | Nominal |
| rbc | Red Blood Cells (normal, abnormal) | Nominal |
| pc | Pus Cell (normal, abnormal) | Nominal |
| pcc | Pus Cell clumps (present, not present) | Nominal |
| ba | Bacteria (present, not present) | Nominal |
| bgr | Blood Glucose Random (bgr in mgs/dl) | Numerical |
| bu | Blood Urea (bu in mgs/dl) | Numerical |
| sc | Serum Creatinine (sc in mgs/dl) | Numerical |
| sod | Sodium (sod in mEq/L) | Numerical |
| pot | Potassium (pot in mEq/L) | Numerical |
| hemo | Haemoglobin (hemo in gms) | Numerical |
| pcv | Packed Cell Volume | Numerical |
| wc | White Blood Cell Count (wc in cells/cumm) | Numerical |
| rc | Red Blood Cell Count (rc in millions/cmm) | Numerical |
| htn | Hypertension (yes, no) | Nominal |
| dm | Diabetes Mellitus (yes, no) | Nominal |
| cad | Coronary Artery Disease (yes, no) | Nominal |
| appet | Appetite (good, poor) | Nominal |
| pe | Pedal Edema (yes, no) | Nominal |
| ane | Anemia (yes, no) | Nominal |

4.1.2 Handling missing data:

After the import of the data from the repository there will be missing data in the dataset. So in order to process the prediction handling those missing values should be done. This can be done in two ways:

1. by deleting those particular rows
2. by replacing those values with mean, meadian or any particular value

In our model we replaced the missing values using the mean of the observations

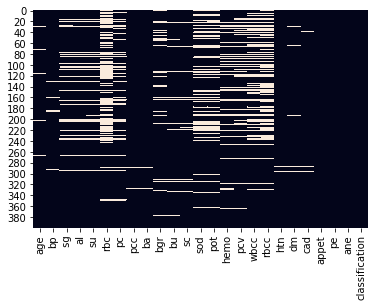
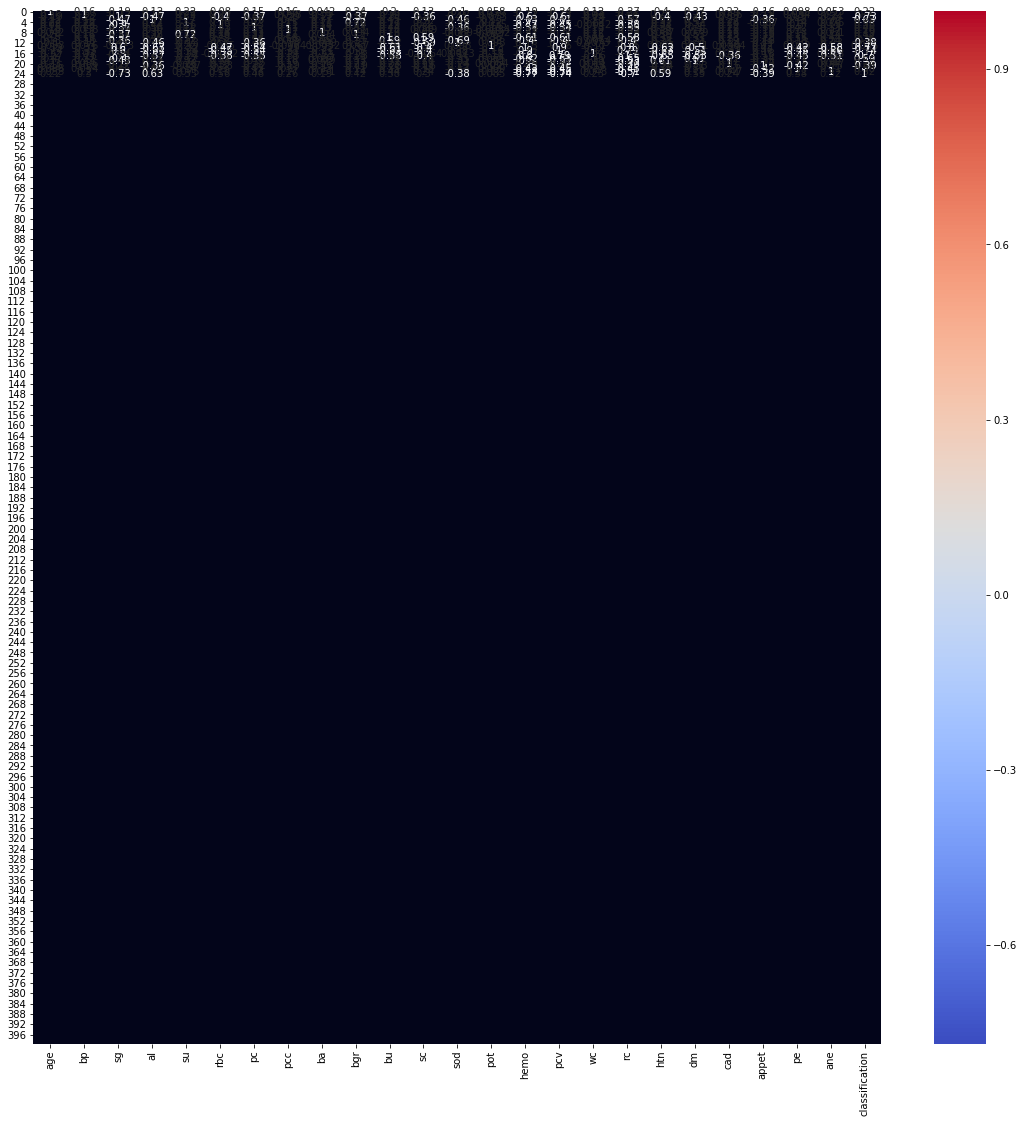


Fig: missing data visualization

After replacing the null values with the preprocessing with mean



**8. Conclusion:**

