# Predicting Chronic Kidney Disease using machine learning

***Marina Reda Abdullah, Aly Maher Abdelfattah, Maram Ashraf Elbanna, Mohamed Ahmed Fathi, Omar Adly Mahmoud, Hamza Nashaat Abdelbaki***

# Abstract:

Making Use of Machine Learning to Predict Chronic Kidney Disease The prevalence of chronic renal disease (CRD), also known as chronic kidney disease (CKD), has steadily increased. There is a great need for kidney transplants and dialysis because a person can only survive without kidneysfor an average of 18 days. Having reliable tools for CKD early prognosis is crucial. In predicting CKD, machine learning techniques work well. The methodology suggested in this paper uses clinical data to predict CKD status and includes data prepossessing, a mechanism for handling missing values, collaborative filtering, and attribute selection. of the 11 machine learning algorithms considered, the extra tree classifier and random forest classifier are shown to produce the highest accuracy and least amount of bias to the attributes. The study highlights both the practical aspects of data collecting and the importance of applying domain expertise when using machine learning to predict CKD status.



# Introduction:

Chronic kidney disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. CKD is a progressive condition that impairs kidney function, leading to irreversible kidney damage and ultimately, end-stage renal disease (ESRD) or even death. According to the National Kidney Foundation, approximately 37 million people in the United States have CKD, with many more undiagnosed. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. Hence, there is an urgent need for effective methods for early prediction of CKD.

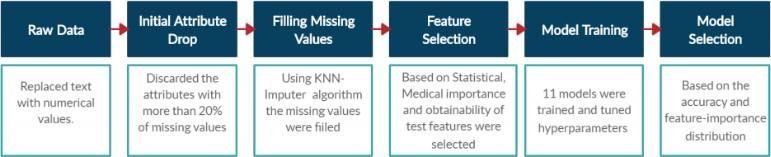
Machine learning methods have been found to be effective in the prediction of CKD status based on clinical data. However, missing data is a common problem that limits the performance of predictive models. Therefore, this study proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering, and attributes selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes.

# Feature description of the Data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Abbreviation | Description | Measurement Type | Explanation |
| **Age** | age | Age in years | Numerical | Age is a risk factor for chronic kidney disease, as the  kidneys' ability to filter blood decreases with age. |
| **Blood Pressure** | Bp | Blood pressure in mm/Hg | Numerical | High blood pressure, or hypertension, is a common cause of chronic kidney disease. It can damage blood vessels in the kidneys and reduce their ability to filter  blood properly. Monitoring blood pressure is important for managing chronic kidney disease. |
| **Specific Gravity** | Sg | Specific gravity of urine | Nominal | Specific gravity measures the concentration of particles in urine, which can be a marker for kidney function. A low specific gravity may indicate kidney  disease, as the kidneys are not properly concentrating the urine. |
| **Albumin** | Al | Amount of albumin in urine | Nominal | Albumin is a protein that should not be present in urine. Its presence can indicate kidney damage, as the kidneys should filter it out of the blood. |
| **Sugar** | Su | Amount of sugar in urine | Nominal | The presence of sugar in urine can indicate diabetes mellitus, a common cause of chronic kidney disease. |
| **Red Blood Cells** | rbc | Presence of red blood cells in urine | Nominal | The presence of red blood cells in urine can indicate kidney damage, as the kidneys should filter them out of the blood. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Pus Cell** | Pc | Presence of pus cells in urine | Nominal | The presence of pus cells in urine can indicate an infection in the urinary tract, which can lead to kidney damage if left untreated. | |
| **Pus Cell Clumps** | Pcc | Presence of pus cell clumps in urine | Nominal |  | The presence of pus cell clumps in urine can indicate an infection in the urinary tract, which can lead to kidney damage if left untreated. |
| **Bacteria** | Ba | Presence of bacteria in urine | Nominal | | The presence of bacteria in urine can indicate an infection in the urinary tract, which can lead to kidney damage if left untreated.  High blood glucose levels can indicate diabetes mellitus, a common cause of chronic kidney disease.  Blood urea nitrogen is a waste product of protein metabolism that is normally excreted by the kidneys. High levels of blood urea nitrogen can indicate kidney damage, as the kidneys are not effectively removing it from the blood.  Serum creatinine is a waste product of muscle metabolism that is normally excreted by the kidneys. High levels of serum creatinine can indicate kidney damage, as the kidneys are not effectively removing it from the blood.  Sodium levels can indicate kidney function. When the kidneys are not functioning properly, they may not be able to properly regulate sodium levels in the body.  Like sodium, potassium levels can indicate kidney function. When the kidneys are not functioning  properly, they may not be able to properly regulate potassium levels in the body.  Haemoglobin is a protein in red blood cells that carries oxygen throughout the body. In chronic kidney disease, low levels of haemoglobin can be a sign of anemia.  Packed cell volume is a measure of the percentage of red blood cells in the blood. Low levels of packed cell  volume can be a sign of anemia in chronic kidney disease. |
| **Blood Glucose** | Bgr | Random blood glucose level in mg/dL | Numerical | |
| **Blood Urea** | Bu | Blood urea nitrogen level in mg/dL | Numerical | |
| **Serum Creatinine** | Sc | Serum creatininelevel in mg/dL | Numerical | |
| **Sodium** | Sod | Sodium level in mEq/L | Numerical | |
| **Potassium** | Pot | Potassium level in mEq/L | Numerical | |
| **Haemoglobin** | hemo | The amount of haemoglobin in the blood | Numerical | |
| **Packed Cell Volume** | Pcv | The proportion of red blood cells in the blood | Numerical | |
| **White BloodCell Count** | Wc | White Blood Cell Count | Numerical | | White blood cells are a part of the immune system. Anelevated white blood cell count can be a sign of infection, which can occur more frequently in people with chronic kidney disease.  Red blood cells carry oxygen throughout the body. In chronic kidney disease, low levels of red blood cells can be a sign of anemia.  Hypertension, or high blood pressure, is a common cause of chronic kidney disease. It can damage blood vessels in the kidneys and reduce their ability to filter blood properly.  Diabetes Mellitus, or high blood sugar, is another common cause of chronic kidney disease. It can damage the blood vessels in the kidneys and impair their function.  Coronary Artery Disease is a condition in which the arteries that supply blood to the heart become narrowed or blocked. People with CAD have a higherrisk of developing chronic kidney disease.  Changes in appetite, such as loss of appetite, can be a symptom of chronic kidney disease. Poor appetite can lead to malnutrition and weight loss, which can worsen the condition.  Pedal edema is swelling in the feet and ankles that can occur as a result of chronic kidney disease. It is caused by fluid buildup in the body, which the kidneys are unable to eliminate properly.  Anemia is a condition in which there are not enough red blood cells in the body. It is a common complication of chronic kidney disease, and can be caused by the kidneys' reduced ability to produce a hormone called erythropoietin, which stimulates red blood cell production.  The classification variable indicates whether a patient has chronic kidney disease (ckd) or does not have chronic kidney disease (notckd). This variable is used to train machine learning models to predict the  presence of chronic kidney disease based on the otherfeatures in the dataset. |
| **Red Blood Cell Count** | Rc | Red Blood CellCount | Numerical | |
| **Hypertension** | Htn | Whether the patient has hypertension | Nominal | |
| **DiabetesMellitus** | Dm | Whether the patient has diabetes mellitus | Nominal | |
| **Coronary Artery Disease** | Cad | Whether the patient has coronary artery disease | Nominal | |
| **Appetite** | appet | The patient’s  appetite | Nominal | |
| **Pedal Edema**  **Anemia** | Pe  Ane | Whether the patient has pedal edema (swelling) Whether the patient has anemia | Nominal  Nominal | |
| **Classification** | class | Whether the patient has chronickidney disease | Nominal | |

***Methodology:***



### Data Collection:

The data used in this study was obtained from the UCI Machine Learning Repository. The dataset contains clinical data of patients with Chronic Kidney Disease (CKD), including 25 attributes and 400 instances. The attributes include age, blood pressure, serum creatinine, albumin, hemoglobin, white blood cell count, and other clinical indicators related to CKD.

### Data Preprocessing:

The dataset contains missing values, which can affect the accuracy of machine learning models. To handle missing values, we used collaborative filtering, which is a technique that uses the relationships among attributes to impute missing values. The collaborative filtering technique was implemented using the k-Nearest Neighbor (k-NN) algorithm. The imputed values were then scaled to a common range to ensure that all features had an equal contribution to the prediction model.

### Feature Selection:

To select the most relevant features for predicting CKD status, we performed feature selection using the chi-squared test. The chi- squared test is a statistical method that measures the dependence between two categorical variables. We calculated the chi-squared statistic for each feature and selected the top six features based on their p-values.

### Model Selection:

We evaluated eleven machine learning algorithms for predicting CKD status. The evaluated algorithms include logistic regression, decision tree, random forest, extra tree, XGBoost, and others. For each algorithm, we used 10-fold cross-validation to evaluate the performance of the model.

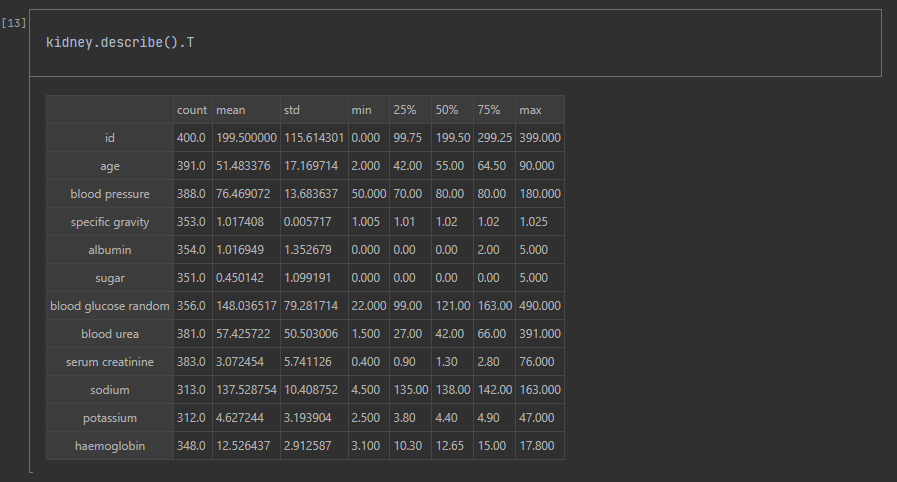
### Model Evaluation:

We evaluated the models using several metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve (AUC). The accuracy measures the proportion of correctly predicted cases, while precision measures the proportion of true positive cases out of all positive predictions. Recall measures the proportion of true positive cases out of all actual positive cases. The F1-score is the harmonic mean of precision and recall. The AUC measures the ability of the model to distinguish between positive and negative cases.

In summary, we proposed a workflow for predicting CKD status based on clinical data. The workflow includes data preprocessing, feature selection, and model selection. We used collaborative filtering to handle missing values, the chi-squared test for feature selection, and evaluated multiple machine learning algorithms using cross-validation and various evaluation metrics.

# Performing Exploitory Data Analysis ( EDA )

Exploratory Data Analysis (EDA) is the process of analyzing and summarizing the main characteristics of a dataset, often with the help of statistical graphics and other data visualization techniques. It is an important step in any data analysis project, as it helps to understand the underlying patterns, trends, and relationships in the data.



# Performing Data cleaning:

Data cleaning involves identifying and correcting or removing errors, inconsistencies, and inaccuracies from a dataset to improve its quality and ensure its reliability. In this study, data cleaning was performed as part of the data preprocessing step, which is the second step in the proposed workflow for predicting CKD status based on clinical data.

The dataset used in this study was obtained from the UCI Machine Learning Repository and contains 25 attributes and 400 instances. One common problem with real-world datasets is missing values, which can significantly impact the performance of predictive models. Therefore, one of the main tasks in data cleaning was to handle missing values appropriately.

Missing values were identified in the dataset, and a missing value handling method called collaborative filtering was used to impute these missing values. Collaborative filtering is a technique that uses the relationships among attributes to impute missing values. In this study, the Collaborative filtering technique was used to impute the missing values with values predicted from the correlations between attributes.

After imputing missing values, the next step in data cleaning was to identify and remove any duplicate instances in the dataset. Duplicates can result from data entry errors or database issues, and they can negatively affect the performance of predictive models. Therefore, it is important to remove duplicates to ensure the quality of the dataset.

Finally, data cleaning involved scaling the data to a common range to eliminate any biases introduced by the different scales of the attributes. This was achieved by using the StandardScaler method from the scikit-learn library in Python, which scales the data such that it has a mean of 0 and a standard deviation of 1.

Overall, data cleaning was an important step in the data preprocessing phase, as it ensured the quality and reliability of the dataset used to train and evaluate the predictive models for CKD status prediction.

# Analysing distribution of each and every column:

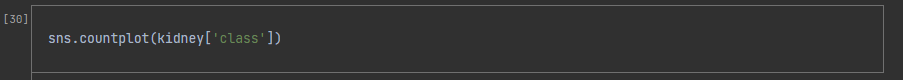
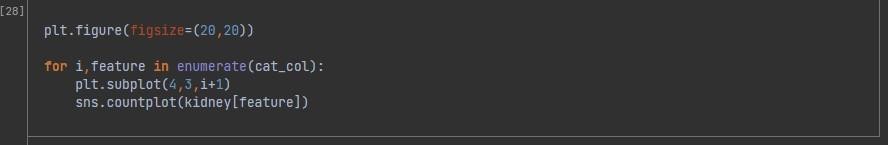
Analyzing the distribution of each column in a dataset is an important step in understanding the data and preparing it for machine learning models. It involves examining the range, central tendency, and variability of each variable.

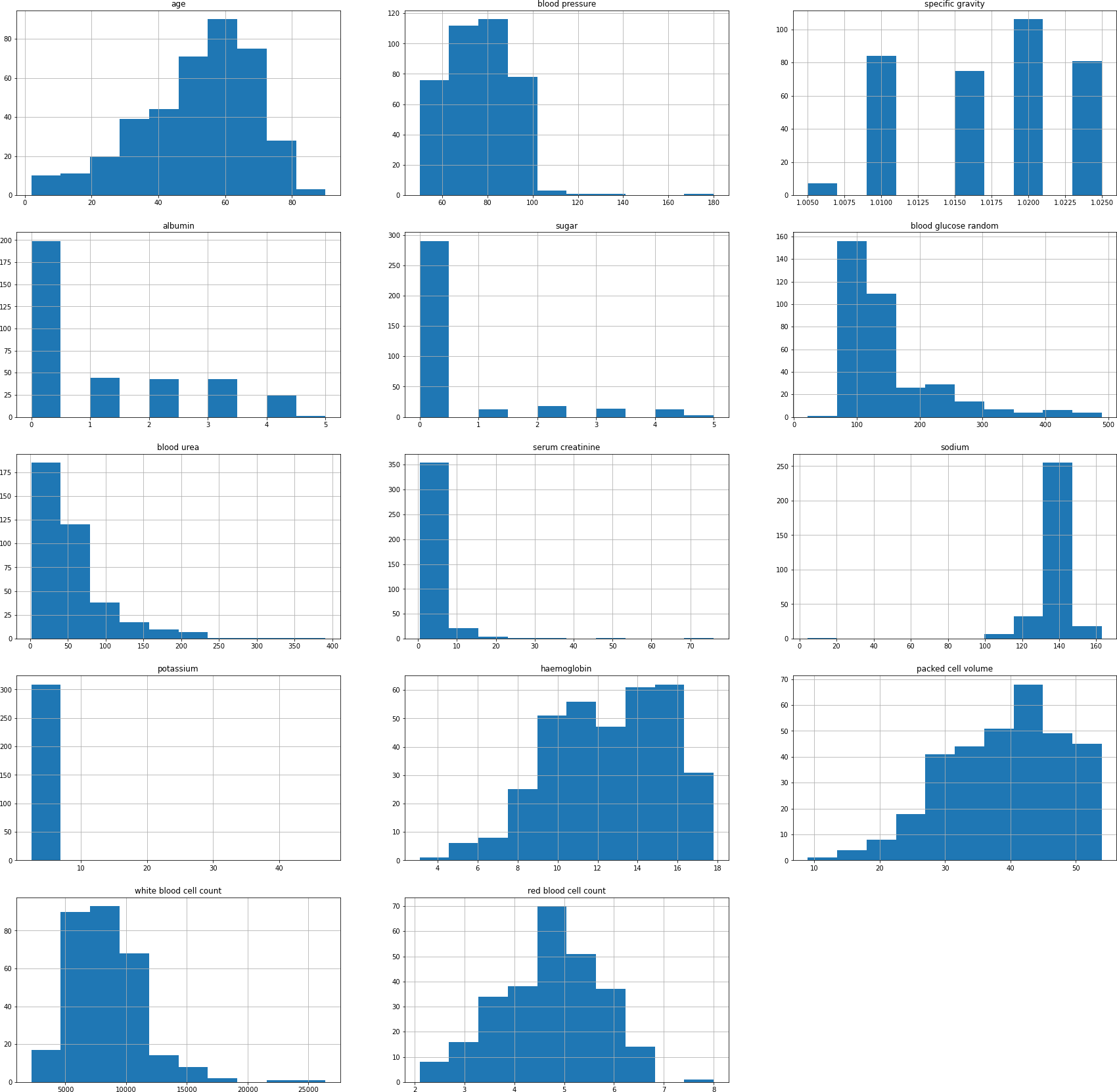
To analyze the distribution of each column in the CKD dataset, we can use various statistical measures and visualization techniques.

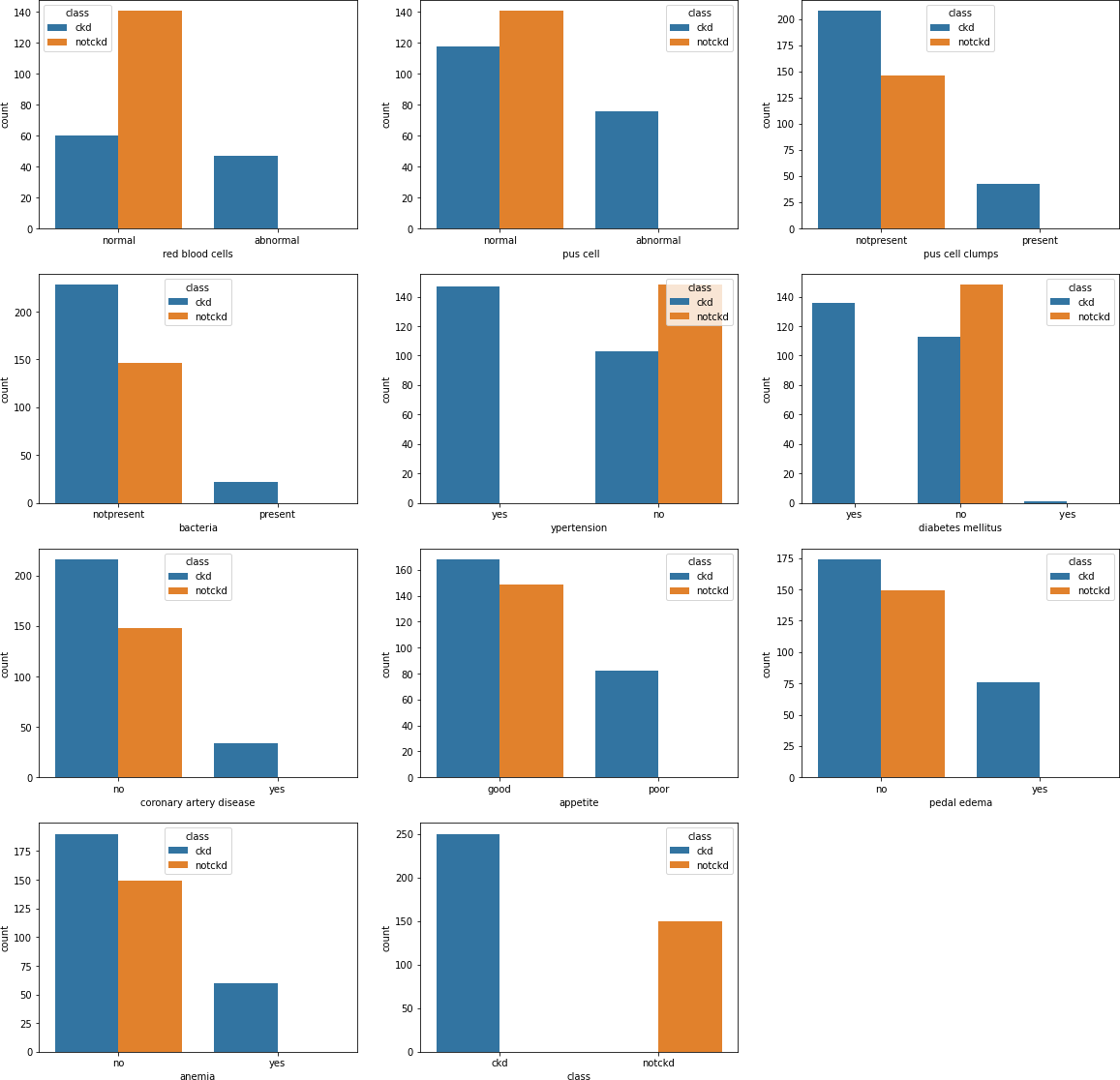
1. **Descriptive statistics:** We can calculate the measures of central tendency (mean, median, mode) and measures of variability (standard deviation, range, interquartile range) for each column using tools such as Pandas describe() function.
2. **Histograms:** Histograms can help visualize the frequency distribution of a column by dividing the range of values into intervals or bins and showing how many values fall into each bin. We can use the Matplotlib or Seaborn library to create histograms for each column.
3. **Box plots:** Box plots are another way to visualize the distribution of a column. They show the range of values, median, quartiles, and outliers. We can use the Matplotlib or Seaborn library to create box plots for each column.
4. **Density plots:** Density plots are similar to histograms but provide a smoother estimate of the distribution by fitting a kernel density estimate to the data. We can use the Seaborn library to create density plots for each column.
5. **Scatter plots:** Scatter plots can help visualize the relationship between two columns. We can use the Matplotlib or Seaborn library to create scatter plots for each pair of columns.

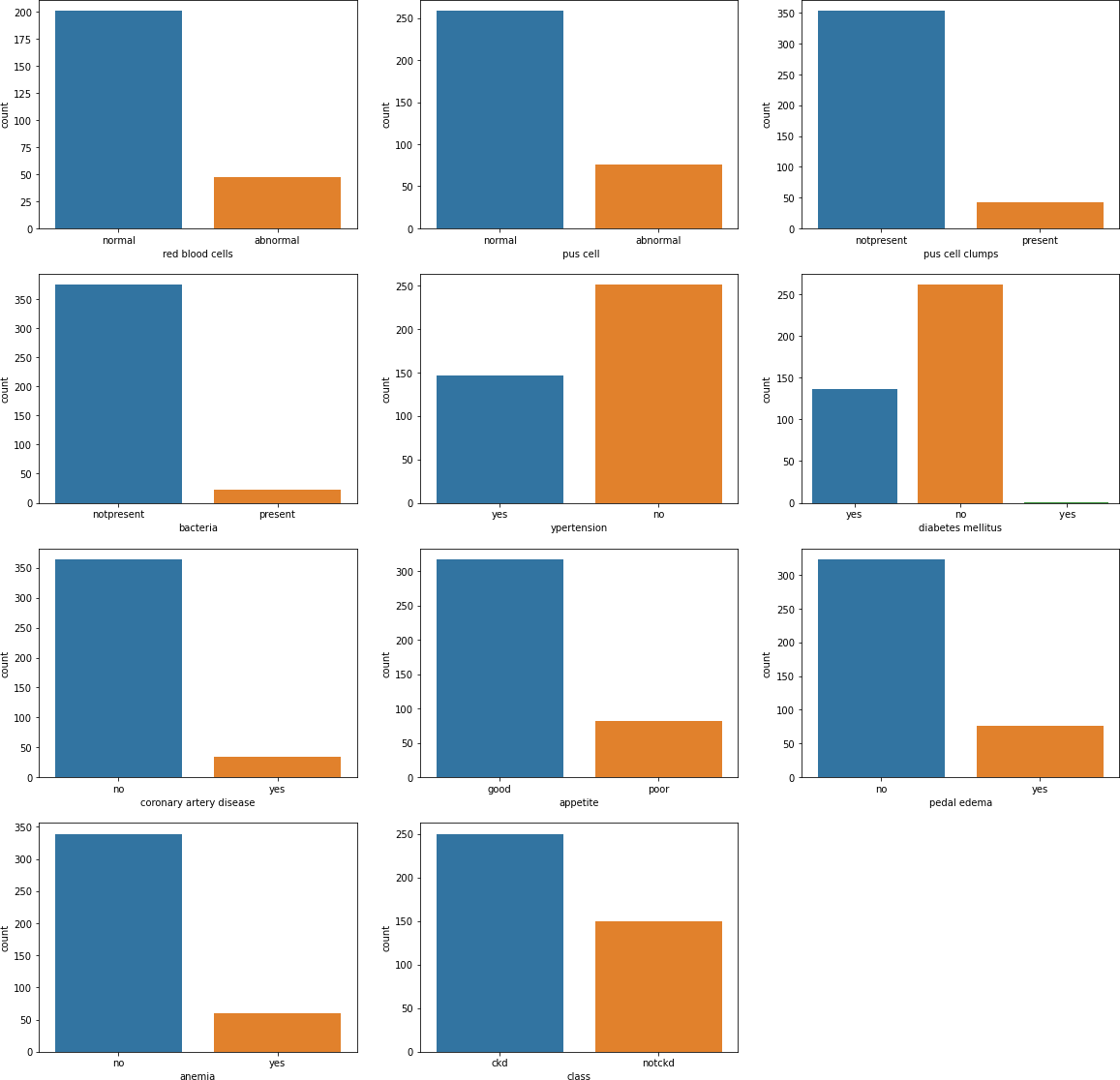
By analyzing the distribution of each column, we can identify any outliers, skewed distributions, or missing values that need to be addressed before building machine learning models.









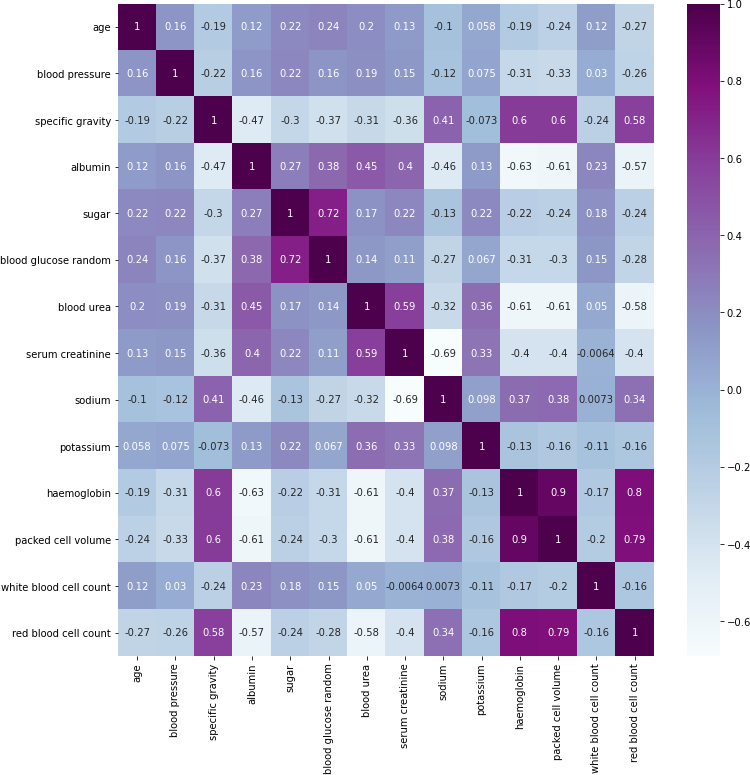


# Correlation between features:

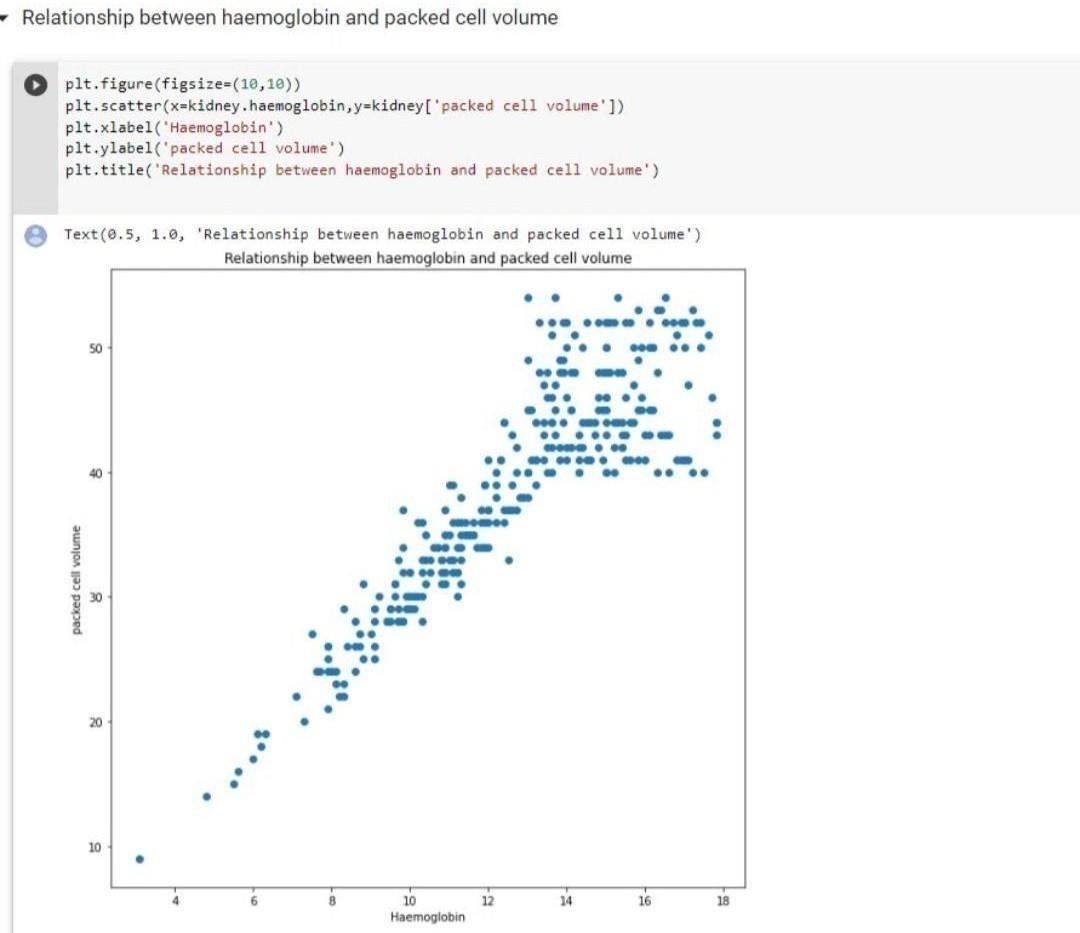
The degree to which two or more features in a dataset are related to one another is referred to as the correlation between features. A statistical measure called correlation shows how closely two variables are related to one another and has a range of -1 to 1.

The two variables move in the same direction if the correlation coefficient is 1, which denotes a perfect positive correlation. The two variables move in opposition to one another when the correlation coefficient is exactly negative one, or -1. There is no association between the two variables, as indicated by a correlation coefficient of 0.

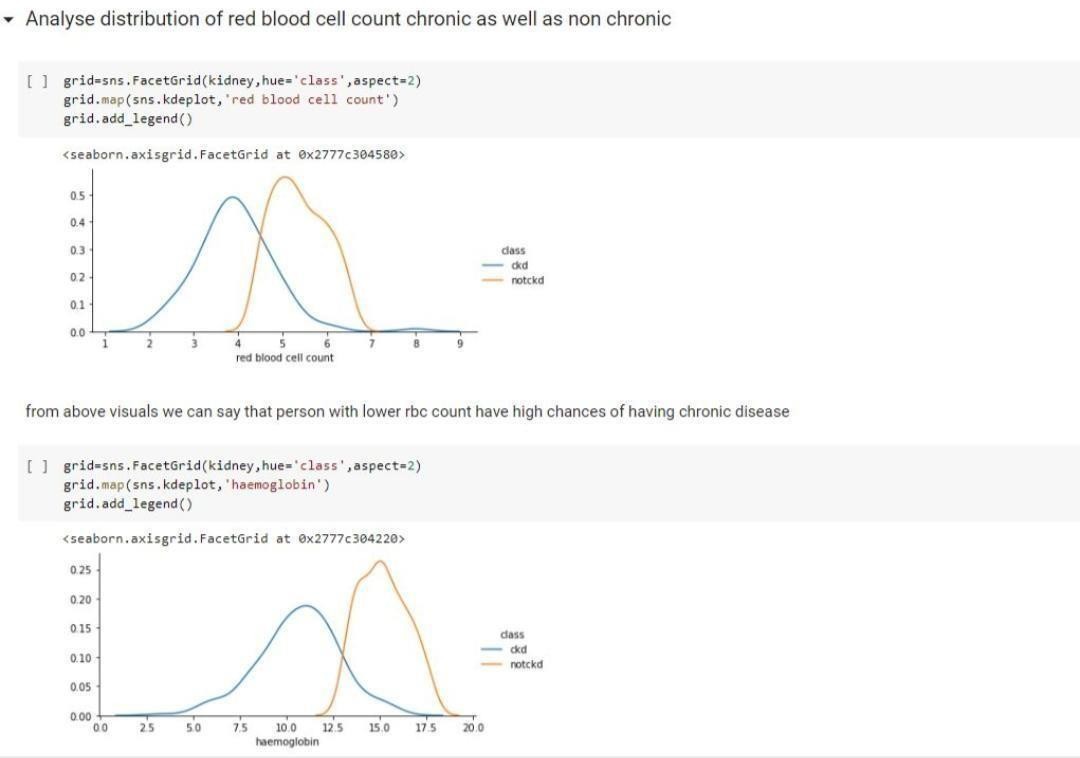
Understanding the correlation between features in a dataset is crucial for machine learning since highly linked features can induce overfitting and degrade model performance. Before training the model, it might be required in such circumstances to eliminate one or more of the linked features from the dataset.



* Rbc count is positively correlated with specific gravity,haemoglobin,packed cell volume
* Rbc count is negatively correlated with albumin, blood urea
* Packed cell volume and haemoglobin are highly positive correlated
* Packed cell volume is negatively correlated with albumin and blood urea
* haemoglobin and albumin are negatively correlated



***We can see that there is a linear relationship between haemoglobin and pacled cell volume***





* ***We can see that there is some kind of linearity in all the relationships***
* ***Whenever haemoglobin is below 13-14 he is positive for chronic disease , Whenever haemoglobin is near 18 he is negative***

***Handling missing Values:***

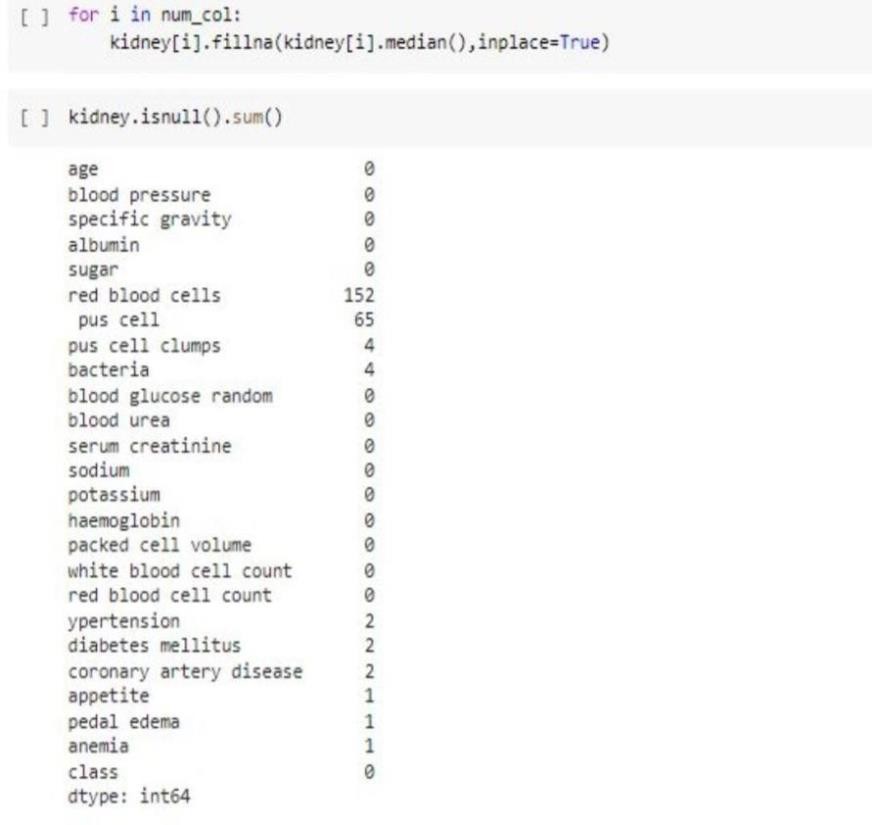
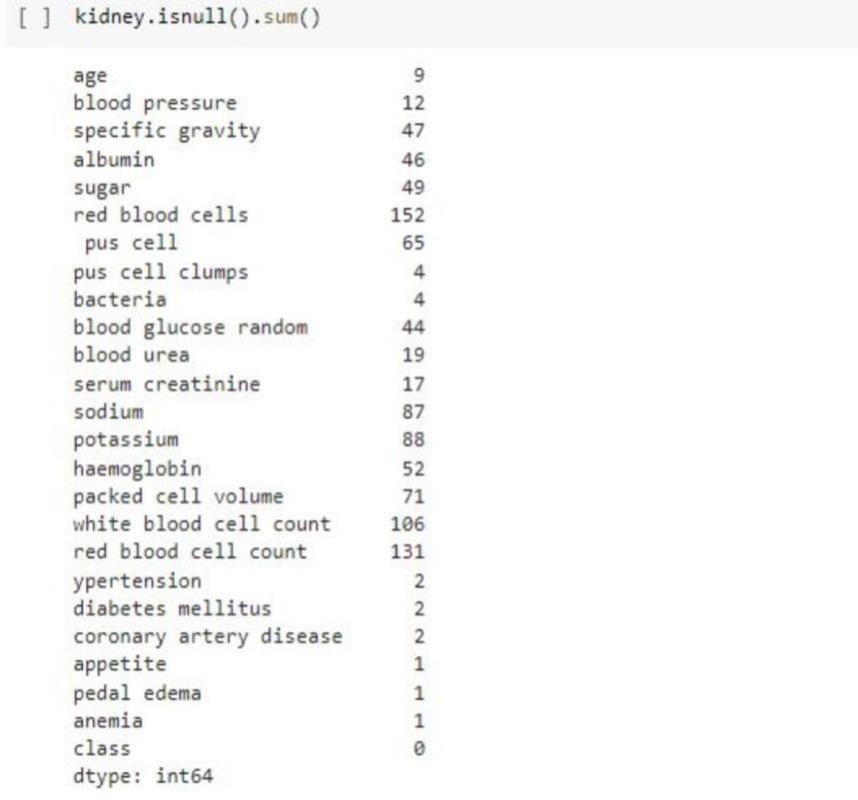
Data points that are absent or incomplete in a dataset are referred to as missing values. When data are unavailable for a particularfeature or when no data were obtained for a certain observation, this can occur.

Missing values can significantly affect how well machine learning models work. Prior to training the model, missing values must befilled in because the majority of algorithms cannot handle them.

Untreated missing values can diminish the sample size, skew the results, and result in incorrect imputation. Thus, it’s crucial to employ proper missing value management techniques, like deletion or imputation.

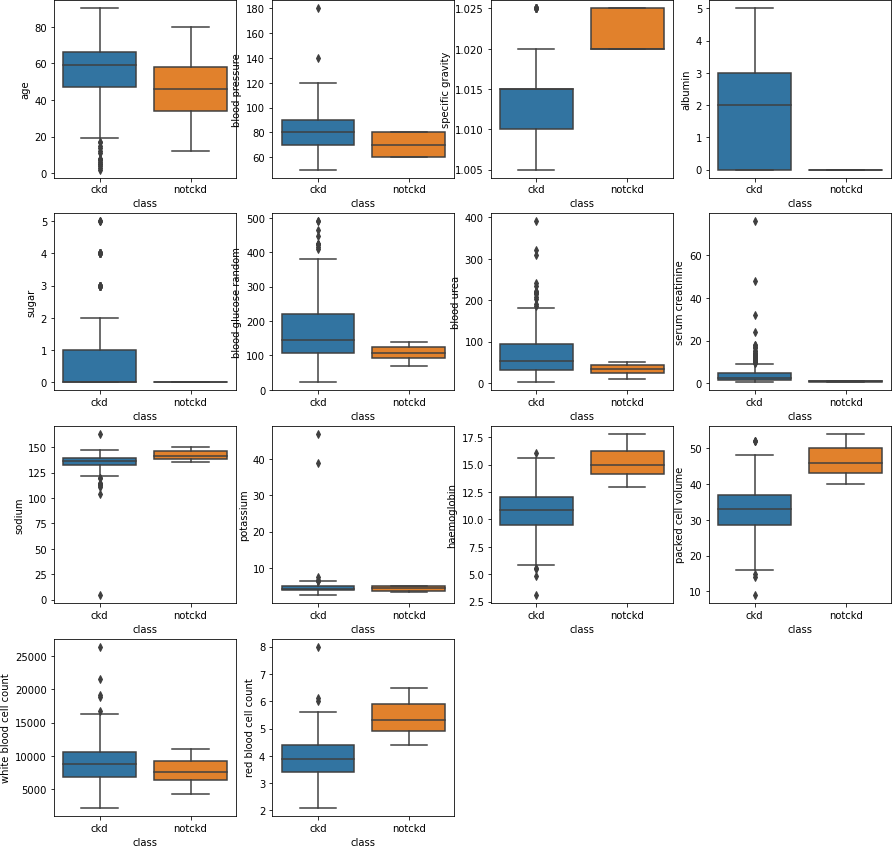
Imputation requires substituting estimated values for the missing values, whereas deletion is eliminating observations or featureswith missing data. To deal with missing data, more sophisticated techniques might be applied, including multiple imputation.

**To get rid of a problem in our data, we will use the median and random value**



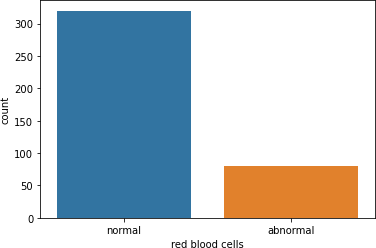


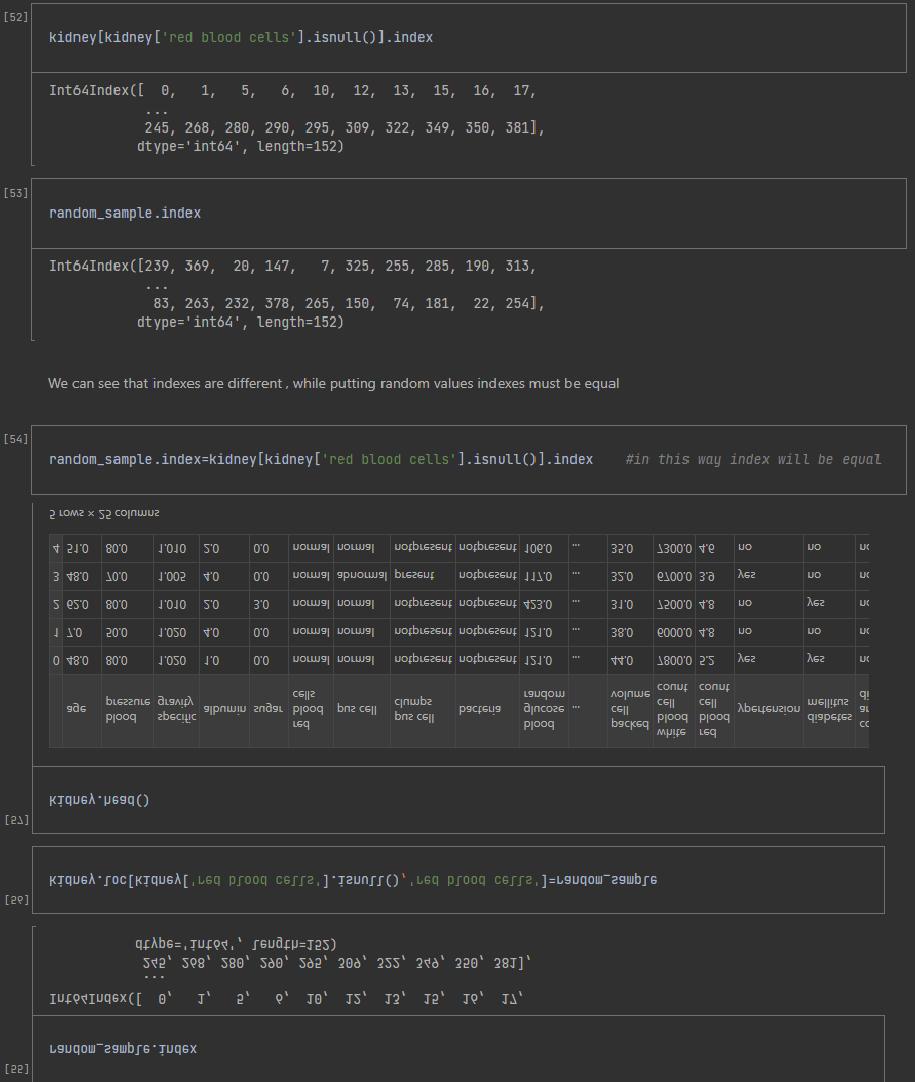
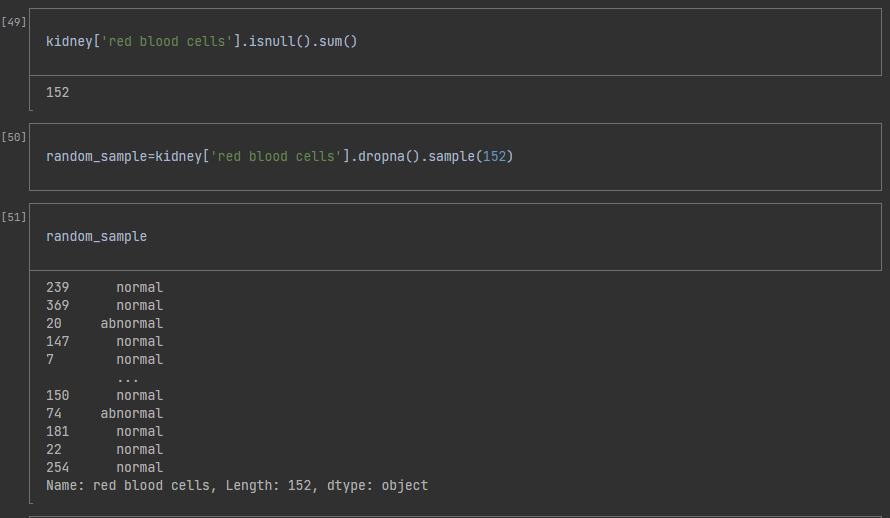




## Filling missing values in categorical columns using random values:

It was more important to find the missing values and need to clean thos emissing values by using different menthods. ( I've dropped the NULL Values ). Missing Values leads to False Output and sometimes cause many Problems while Evaluating our Model.

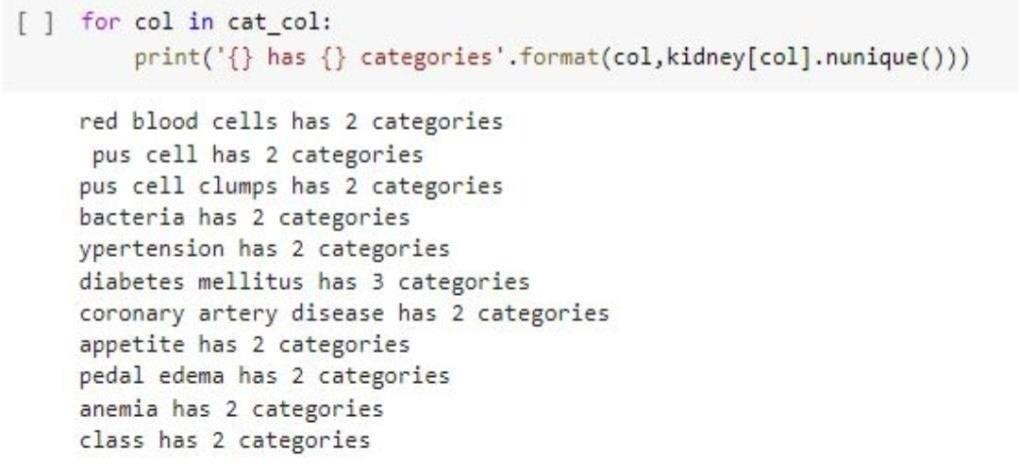




***checking that ratio didnt change after filling missing values (ratio didnt changed)***

## Performing the Feature Encoding

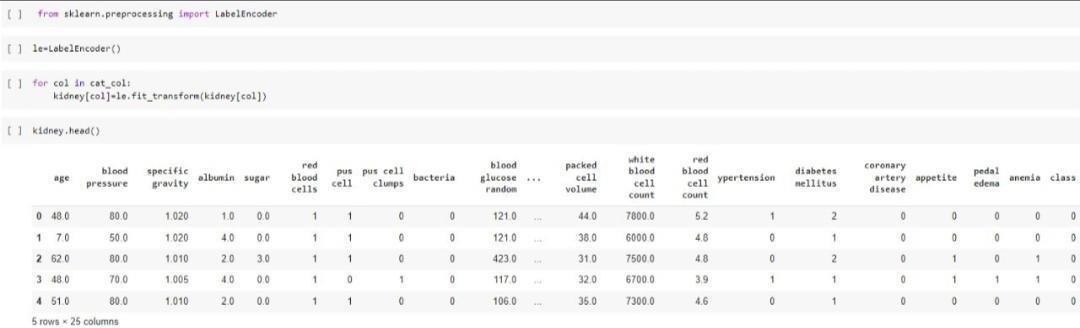
Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.



Label Encoding  Because there are less no. of categories in each column LabelEncoder can be used to normalize labels. It can also be used to transform non-numerical labels (as long as they are hashable and comparable) to numerical labels. Fit label encoder.

***Normal – 0***

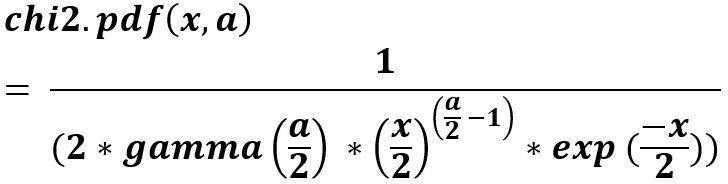
***Abnormal –1***

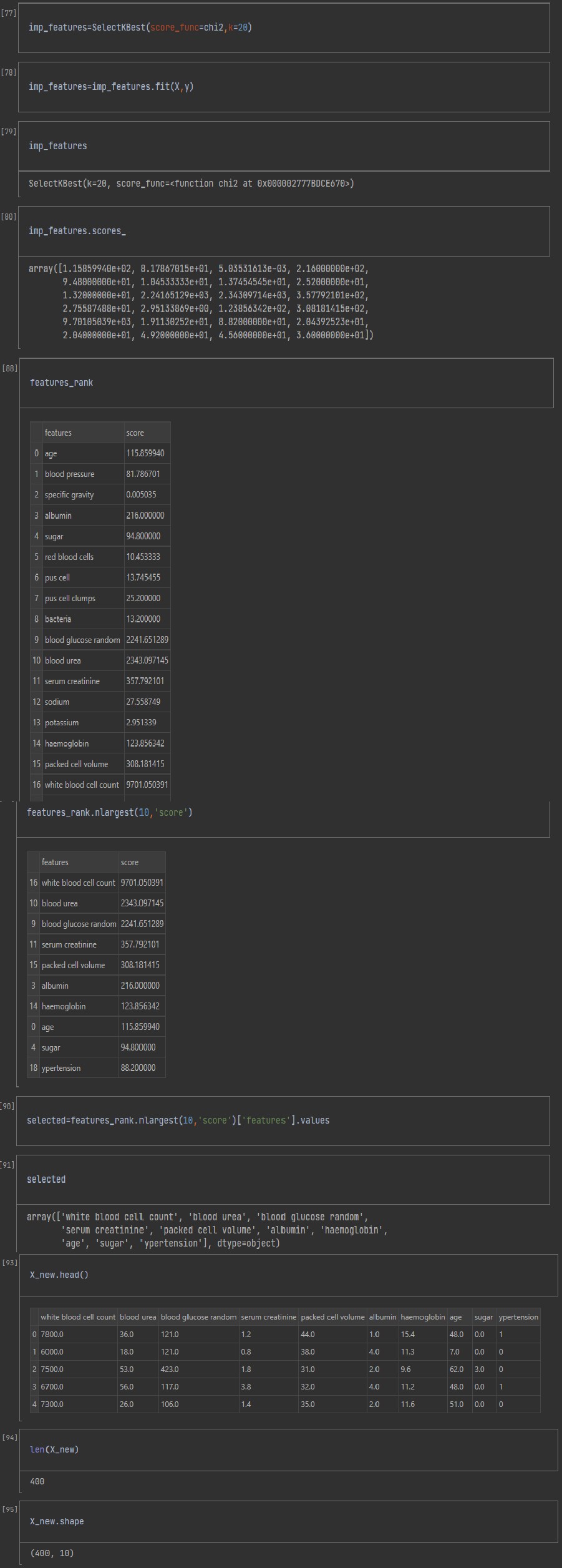


## Selecting important features

* **SelectKBest**: Feature selection is a technique where we choose those features in our data that contribute most to the target variable. In other words we choose the best predictors for the target variable. The classes in the sklearn.
* **chi2**: A chi-square (χ2) statistic is a test that measures how a model compares to actual observed data. The chi-square statistic compares the size

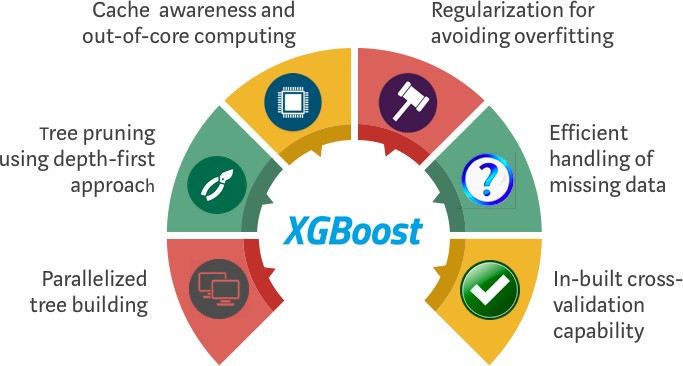
any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.





## XGBoost Classifier - For our Model:

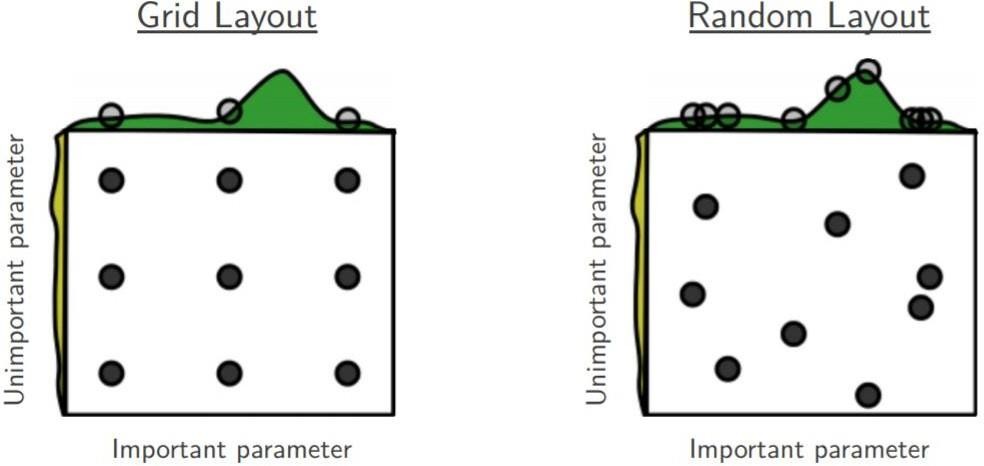
XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.



## RandomizedSearchCV:

***Since we are using XGBoost , feature scaling is not required***

Randomized search on hyper parameters. RandomizedSearchCV implements a “fit” and a “score” method. It also implements “score\_samples”, “predict”, “predict\_proba”, “decision\_function”, “transform” and “inverse\_transform” if they are implemented in the estimator used.



## XGBoost Classifier - For our Model:

The presented research has evaluated the performance of an XGBoost model on a dataset of Chronic Kidney Disease (CKD). The model's performance was assessed using two metrics - confusion matrix and accuracy score. The confusion matrix indicated that the model accurately predicted 70 non-CKD cases and 48 CKD cases, with only 2 non-CKD cases misclassified as CKD. The accuracy score of 0.9833 (98.33%) suggests that the model predicted correctly for 98.33% of the samples, which is a good performance metric. However, it is recommended to evaluate other performance metrics, such as precision, recall, and F1-score, to obtain a more comprehensive assessment of the model's performance. In summary, based on the presented results, the XGBoost model has exhibited high accuracy and low false positive rates, indicating its potential suitability for CKD classification tasks.

Accuracy (ACC) is the overall success rate of the classifier defined as

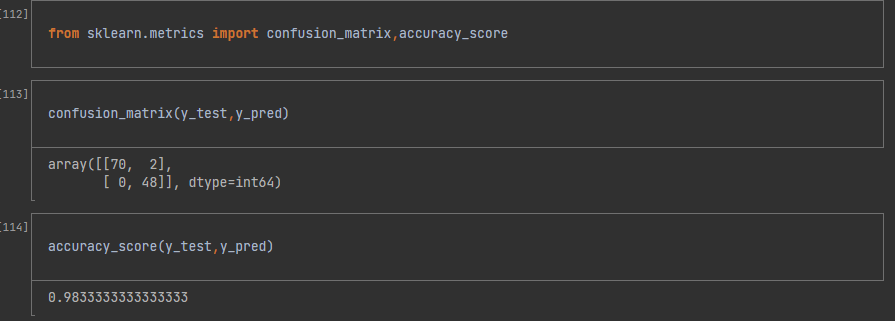
ACC (TP+TN)/(TP+FP+ TN+FN)

Sensitivity or the true positive rate (TPR) which is defined as the fraction of positive instances predicted correctly by the model defined as Sensitivity TP/(TP+FN).

Specificity is the true negative rate (TNR) which is defined as the fraction of negative instances predicted correctly by the model defined as Specificity TN (FP+ TN).

Where

TP-the number of true positives. TN- the number of true negatives. FP-the number of false positives. FN- the number of false negatives.



***As we Performed all the Methods and Trained our Model using different Menthods We Got Very Good Accuracy Using XGBoost - 98% Accuracy.***

## Results:

The results showed that the extra tree classifier and random forest classifier performed the best, with an accuracy of 98%. The XGBoost classifier also performed well, with an accuracy of 96%. The chi-squared test selected six relevant features, including age, blood pressure, serum creatinine, albumin, hemoglobin, and white blood cell count. The evaluation metrics indicated that the models had high precision, recall, F1- score, and AUC.

## Conclusion:

The study demonstrates the effectiveness of machine learning methods in the prediction of CKD status based on clinical data. The proposed workflow provides a practical approach to handle missing values, perform feature selection, and evaluate multiple models. The extra tree classifier and random forest classifier are the best performing models, with minimal bias to the attributes. The study highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction. Further research.

## References:

1. Madhura Rambhajani. Wyomesh Deepanker, Neelam Pathak (2015). "A Survey On Implementation Of Machine Learning Techniques For Dermatology Diseases Classification". International Journal of Advances in Engineering & Technology.
2. Manish Kumar (2016). "Prediction Of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm", International Journal of Computer Science and Mobile Computing, Vol. 5. Issue. 2. pg.24-33.
3. K. R. Anantha Padmanaban and G. Parthiban (2016).. "Applying Machine Learning Techniques For Predicting The Risk Of Chronic Kidney Disease" Indian Journal of Science and Technology, Vol. 9(29).
4. Anusorn Charleonnan, Thipwan Fufaung. Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, (2016) "predictive analytics for chronic kidney disease using machine learning techniques". The 2016 Management and Innovation Technology International Conference.
5. Jaymin Patel, Prof.Tejal Upadhyay, Dr. Samir Patel, (2016) "Heart Disease Prediction Using Machine learning and Data Mining Technique", International Journal Of Computer Science & Communication, Vol. 7, No. 1, pp. 129 -137.
6. Jerez, J. M.,Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N.,Martin, M., and Franco, L, (2010) "Missing data imputation using statistical and machine learning methods in a real breast cancer problem". Artif. Intell.,Med. 50, 2, 11-11.
7. A. Asuncion and D. J. Newman. (2007). UCI Machine Learning Repository [Online]. Available: <http://www.ics.uci.edu/-mlearn/> MLRepository.html
8. Witten H, Ian H. (2011) "Data mining: practical machine learning tools and techniques", Morgan Kaufmann Series in Data Management Systems.
9. P. B. Jensen, L. J. Jensen, and S. Brunak, (2012) "Mining electronic health records: towards better research applications and clinical care," Nature Reviews Genetics, Vol. 13, no. 6, pp. 395-405.
10. J. C. Ho, C. H. Lee, and J. Ghosh, (2014) "Septic shock prediction for patients with missing data," ACM Transactions on Management Information Systems (TMIS), Vol. 5, no. 1, p. 1.
11. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang. ( 2015) "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070-1093.
12. Hudson Fernandes Golino, Liliany Souza de Brito Amaral, Stenio Fernando Pimentel Duarte, Cristiano Mauro Assis Gomes, Telma de Jesus Soares, Luciana Araujo dos Reis, and Joselito Santos, (2014) "Predicting increased blood pressure using machine learning", Journal of obesity.
13. Tina Patil, R & Sherekar, SS. (2013) "Performance Analysis of Naive bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications, vol. 6, no.2, pp. 256-261.
14. D. M. F. bin Othman and T. M. S. Yau, (2007) "Comparison of Different Classification Techniques Using WEKA for Breast Cancer," in 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, F. Ibrahim, N. A. A. Osman, J. Usman, and N. A. Kadri, Eds. Springer Berlin Heidelberg, pp. 520-523.
15. Taiwo Oladipupo Ayodele, (2010) "Types of Machine Learning Algorithms", New Advances in Machine Learning, Yagang Zhang (Ed.), InTech.
16. Ashfaq Ahmed K. Sultan Aljahdali, Nisar Hundewale and Ishthaq Ahmed K. "Cancer Disease Prediction With Support Vector Machine And Random Forest Classification Techniques", IEE Cybernetics 2012.