**DEVELOPMENT OF A PREDICTIVE MODEL FOR EARLY DETECTION OF KIDNEY DISEASE USING ML MODEL**

The project topic is aimed at developing a machine learning model which will accurately predict the detection of kidney disease in an individual. The data used for this study was gotten from Kaggle, an open-source data warehouse.

The data contains 26 columns ad 400 rows. In order to prepare the data for machine learning procedures, different preprocessing techniques were carried out- Data Cleaning and transformation, Data Exploration and Data Scaling.

**DATA CLEANING:** The data consists of 26 columns, however, most of these rows have missing values (NaNs). The table showing the percentage of missing observations in each column is represented below;

|  |  |
| --- | --- |
| Variables | Percentage of Missing Observations/Values |
| ID | 0.00 |
| Age | 2.25 |
| Bp | 3.00 |
| Sg | 11.75 |
| Al | 11.50 |
| SU | 12.25 |
| RBC | 38.00 |
| PC | 16.25 |
| PCC | 1.00 |
| BA | 1.00 |
| BGR | 11.00 |
| BU | 4.75 |
| SC | 21.75 |
| SOD | 22.00 |
| POT | 13.00 |
| Hemo | 17.75 |
| Pcv | 26.50 |
| Wc | 32.75 |
| Rc | 0.50 |
| Htn | 0.50 |
| Dm | 0.50 |
| Cad | 0.25 |
| Appet | 0.25 |
| Pe | 0.25 |
| Ane | 0.00 |

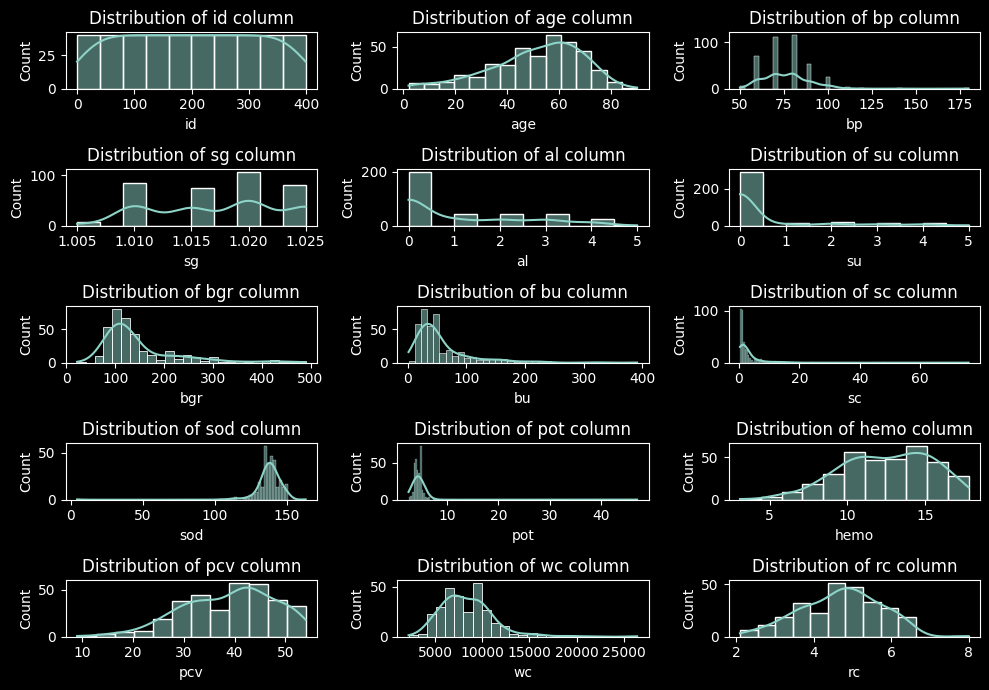
From the above table presentation, it is evident that the variable with the highest percentage of missing values is rbc with 38% which is less than the threshold (50%). These variables will be filled with either their mode, mean or median- depending on the data type.

METHODS OF FILLING NULL VALUES

1. Usage of Mode: the method of using mode is adopted in this project if the data type is object and categorical. This is carried out to ensure that the most occurring value in the variables are not tampered with.
2. Usage of Mean: mean method of filling null values is adopted when the variable is numeric. However, the mean method is critical as it could distort the distribution of variable if used for a variable which is not normally distributed. For the purpose of this project, the mean was used to fill the null values for just the rc column which is shown visually to be normally distributed.
3. Usage of Median: the method os using median for filling null values is also applicable to numerical data. However, if the data is not normally distributed, the median of the variable can be used to fill the null values so as not to tamper with the distribution of that particular variable. Most of the numerical variables in the dataset before transformation are not normally distributed- either positively skewed or negatively skewed. For these variables, the median method of filling null values was adopted.

**DISTRIBUTION PLOT OF NUMERICAL VARIABLES**

The distribution plot is a statistics plot which is used to show the distribution of variables- to detect whether a variable is normally distributed or not normally distributed. It is a univariate plot which can only be adopted when dealing with numerical variables. If the distribution plot of a variable is bell-shaped, the variable is considered to be normally distributed. If the distribution plot is not bell-shaped, it is referred to as non-normality and the shape is expected to either the skewed to the right (positively skewed) or skewed to the left (negatively skewed). The distribution plot can as well be shown using histogram plot and the kernel density (kde). The plots are presented below;



COUNT OF CATEGORICAL VARIABLES

The visualization below is aimed at showing the count of different categories in each categorical variable. Barplot was used to visualize it;



**MACHINE LEARNING**

The first procedure carried out was splitting the dataset into X and Y, where X contains all the explanatory features and Y contains the target feature (classification). The target feature has two categories, ckmd which represents the presence of kidney disease and notckmd which represents absence of kidney disease. Before proceeding to fitting the model, the data was scaled using Standard Scaling procedure in order to solve for outliers in the explanatory features.

Three machine learning models were adopted to fit the dataset- SVM, KNN and Random Forest. Model validation was as well carried out using accuracy score, confusion matrix, sensitivity, specificity and AUC score.

|  |  |
| --- | --- |
| ALGORITHM | ACCURACY SCORE |
| Support Vector Classifier | 98.75% |
| K-Nearest Neighbor | 98.75% |
| Random Forest Classifier | 100% |

These algorithms are developed using all the variables in the dataset. The best model is the random forest classifier with 100% accuracy score. Therefore, it predicted the test data accurately compared to support vector classifier and k-nearest neighbor.

**ELITISM GENETIC ALGORITHM**

The elitism genetic algorithm is an algorithm used for feature selection. This algorithm was integrated in this project to determine the best features which predict best the target feature. The selected features are then used to train the three models which we were required to use in fitting the dataset and the validation scores are presented below;

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ALGORITHM | ACCURACY SCORE | SENSITIVITY | SPECIFICITY | AUC SCORE |
| Support Vector Classifier | 86.25% | 0.8462 | 0.8929 | 0.9341 |
| K-Nearest Neighbor | 86.25% | 0.8462 | 0.8929 | 0.9680 |
| Random Forest Classifier | 100% | 1.0 | 1.0 | 1.0 |

From the table above, it is evident that the best algorithm for predicting kidney disease in individual is random forest classifier which has an accuracy score of 100%, sensitivity and specificity of 1.0 and an AUC score of 1.0.