PROJECT REPORT

"Chronic Kidney Disease Prediction Model using Stacking"

Submitted By

Name	Roll No.
Rohit Kumar	20118078
Dipraj Sonawane	20118036

6th Semester

Information Technology

National Institute of Technology, Raipur

Under the supervision of

Dr. G. P. Gupta
Assistant Professor

National Institute of Technology, Raipur

Date of Submission:-

Acknowledgement

I am grateful to **Dr. G.P. Gupta**, Assistant Professor, Department of Information Technology, NIT Raipur for his proficient supervision on the project on **"Chronic Kidney Disease Prediction using stacking ensemble learning model on different ML algorithms"**. I am very thankful to you Sir for your guidance and support. I am greatly indebted to professor sir and my project partner Dipraj Sonawane for their advice, constructive suggestions, positive and supportive attitude and continuous encouragement.

TABLE OF CONTENT

“Chronic Kidney Disease Prediction using Stacking Ensemble Learning method on different ML algorithms”

Abstract:-

Chronic kidney disease (CKD) is a medical condition in which the gradual loss of kidney functioning over an extended period occurs and it is a significant global health issue which affects millions of people worldwide. CKD occurs due to diabetes, hypertension, and many other disorders which affect the kidneys. Early detection and necessary precautions plays an important role in CKD diagnosis. Diagnosis tests like urine test, blood test etc. helps in identifying the condition of the patient. According to a study it affects 10% of the total population and in south africa due to lack of health care system 15% of the population is affected. In this report we have discussed the technology in the context of CKD which can help in prediction that someone is suffering from CKD or not with the help of some ML prediction algorithms. In this project we have developed an ML Model for CKD prediction with the stacking method of ensemble learning in which we stack different ML algorithms to get a highly efficient model.

Keywords: -

Chronic Kidney Disease,Machine Learning,Ensemble Learning, Stacking

Introduction:-

Due to lack of medical facilities a cheap and fast system for medical diagnosis is very beneficial to tackle diseases. Machine Learning algorithms such as Random forest,Logistic regression, Decision Tree, Xgboost, KNN, SVM etc are being used to treat several chronic diseases like cancer, TB and heart related disease. A medical system is designed using these above mentioned algorithms which detect and predict if someone is suffering from that disease or not.In this project we have developed a model with high efficiency for predicting Chronic Kidney Disease by stacking all of the above mentioned machine learning algorithms.

Ensemble Learning

The ensemble learning models in machine learning combine the results obtained from different learners/models and combine those results for better accuracy and improved decisions.These tables use the multiple learning models to produce one optimal predictive result.

Ensemble Methods:-

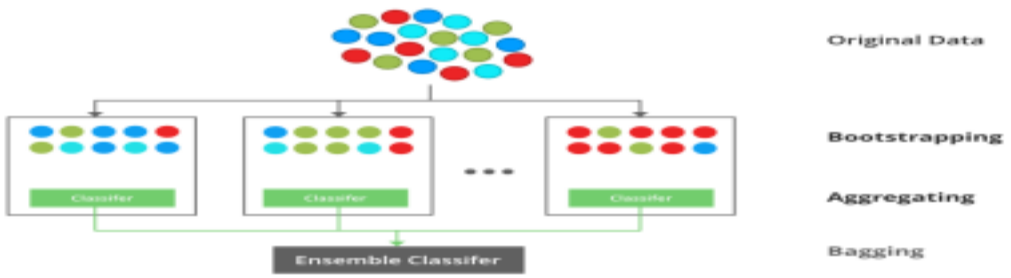
1) Bagging
2) Stacking
3) Boosting

Bagging:-

This technique is also known as bootstrapping and aggregating techniques.This algorithm is designed in such a manner which improves the stability and accuracy of the algorithms of machine learning which are used in classification and regression.

The implementation of bagging is as follows:-

Step 1: Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
Step 2: A base model is created on each of these subsets.
Step 3: Each model is learned in parallel with each training set and independent of each other.
Step 4: The final predictions are determined by combining the predictions from all the models.



Boosting:-

This is an ensemble modelling technique which is used to build a strong classifier from the number of weak classifiers by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

Implementation of boosting:-

Step 1: Initialise the dataset and assign equal weight to each of the data points.
Step 2: Provide this as input to the model and identify the wrongly classified data points.
Step 3: Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalise the weights of all data points.
Step 4:if (got required results)
Goto step 5
else
Goto step 2
Step 5: end

Stacking:-

It is an ensemble learning technique in which various weak learners are arranged in a parallel manner in such a way that by combining them with meta learners, we can predict better results.

The implementation of Stacking:-

Step 1: Split training data sets into n-folds using the **RepeatedStratifiedKFold** as this is the most common approach to preparing training datasets for meta-models.

Step 2: Now the base model is fitted with the first fold, which is n-1, and it will make predictions for the nth folds.

Step 3: The prediction made in the above step is added to the x1_train list.

Step 4: Repeat steps 2 & 3 for remaining n-1folds, so it will give an x1_train array of size n.

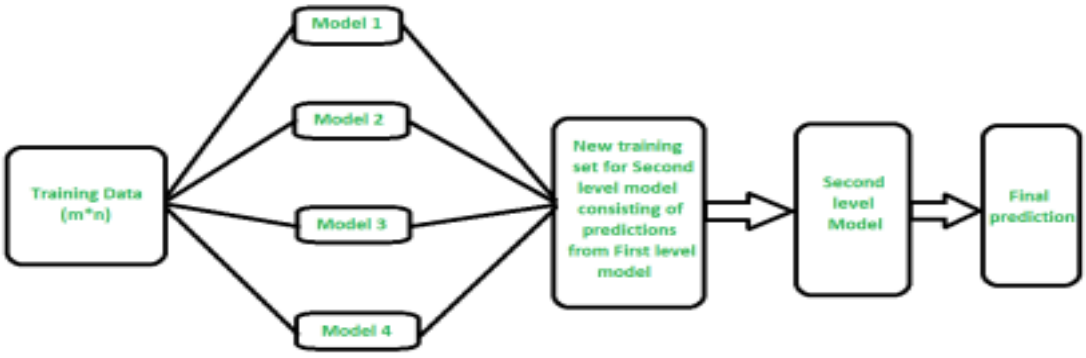
Step 5: Now, the model is trained on all the n parts, which will make predictions for the sample data.

Step 6: Add this prediction to the y1_test list.

Step 7: In the same way, we can find x2_train, y2_test, x3_train, and y3_test by using Model 2 and 3 for training, respectively, to get Level 2 predictions.

Step 8: Now train the Meta model on level 1 prediction, where these predictions will be used as features for the model.

Step 9: Finally, Meta learners can now be used to make a prediction on test data in the stacking model.



In this project we have used stacking ensemble learning method. We have stacked

different ML algorithms through logistic Regression which are KNN, Decision Tree, Random forest, Xgboost,SVM, Logistic regression.

Model Development:-

In the context of chronic kidney disease (CKD), the development of a predictive model

can play a crucial role in early detection and risk assessment. By utilising machine learning algorithms and statistical techniques, we can create a model that aids in identifying individuals at higher risk of developing CKD and predicting the progression of the disease.

The following steps outline the model development process:

1. **Data Collection**: We took the dataset from UCI(University of California). This dataset contains relevant patient information, such as age, gender, medical history, laboratory test results (e.g., serum creatinine levels, estimated glomerular filtration rate), blood pressure readings, and other clinical parameters. This dataset should include both CKD-positive and CKD-negative cases, covering various stages of the disease.

2. **Data Preprocessing:** Performed data cleaning, normalisation, and imputation of missing values. This step ensures that the dataset is of high quality and ready for analysis.

3. **Feature Selection:** Identify the most important features that contribute significantly to the prediction of CKD.In the dataset features like serial number does not play a role in CKD so that feature is removed from the data set .This step helps in reducing dimensionality and enhancing the model's performance. PCA was used to reduce the dimensionality of the data.

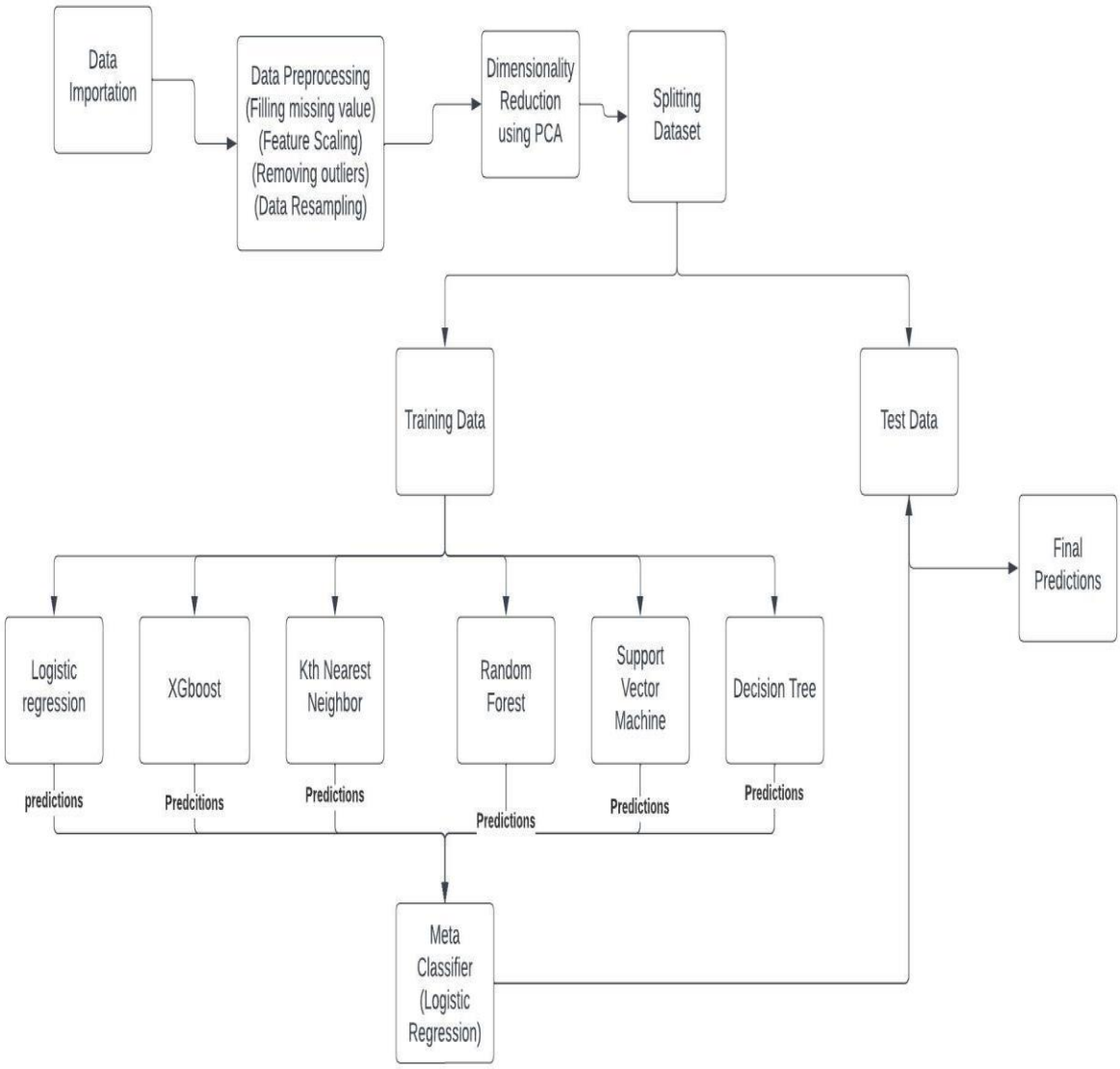
4. **Model Selection:** Machine learning algorithms, such as KNN, decision trees, random forests,Logistic regression,support vector machines, Xgboost were chosen.

5. **Model Training:** Split the dataset into training and testing sets to evaluate the model's performance accurately. Train the chosen machine learning algorithm using the training data.

6. **Model Evaluation:** Use the testing set to evaluate the model's accuracy, F1-score, and other relevant metrics. Compare the results against established benchmarks to assess the model's effectiveness.

By following these steps, the developed predictive model for CKD can aid healthcare professionals in making informed decisions, improve patient outcomes, and contribute to the early detection and management of this chronic condition.

Flow Chart:-



Model Explanation:-

Dataset Importation:

- First of all we imported all the necessary python libraries such as numpy for mathematical calculations and pandas for making data frames and scikit-learn for ml models and sea born for visual representation.
- After that we import the csv file through connecting google drive and google colab.
- Now, using the pd.read function of the pandas library we created a data frame and then we printed the data.

Data Preprocessing:

- Here the data preprocessing begins and we count the null values present in every column of the data set.

- We are using a simple Imputer to fill the missing values with the most frequent value(mode).We could not use mean strategy as mean could not be calculated as data is non numeric

- After that we check for the unique values or special characters present in the data set which occurred due to some typing mistake.

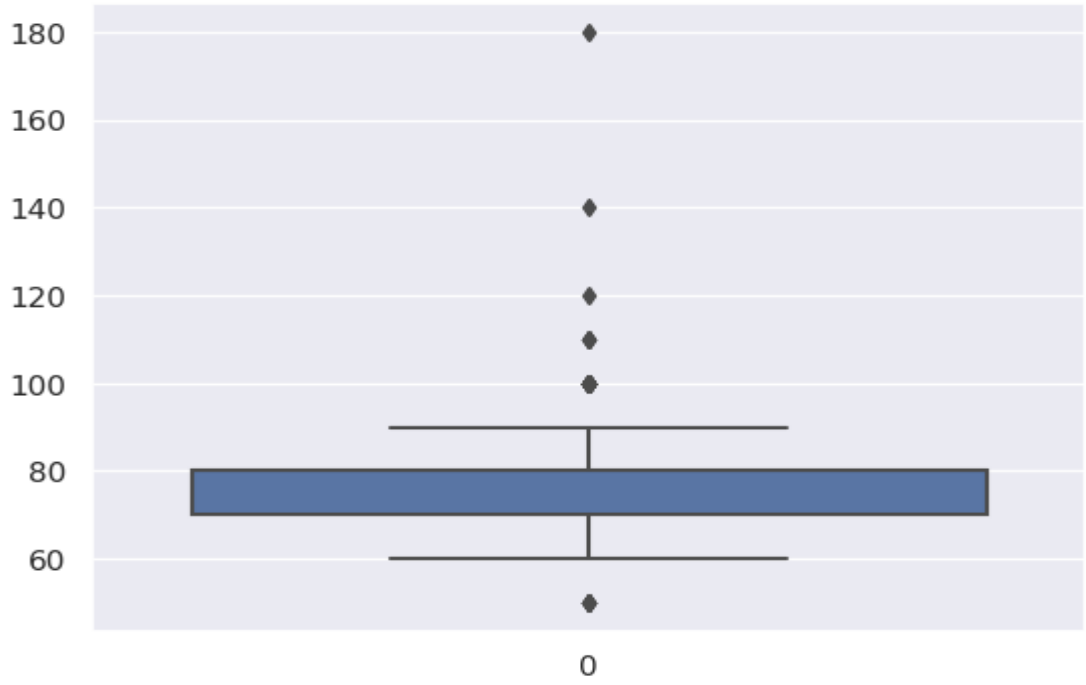
- Now, we replace those unique values with the mode values of the columns.

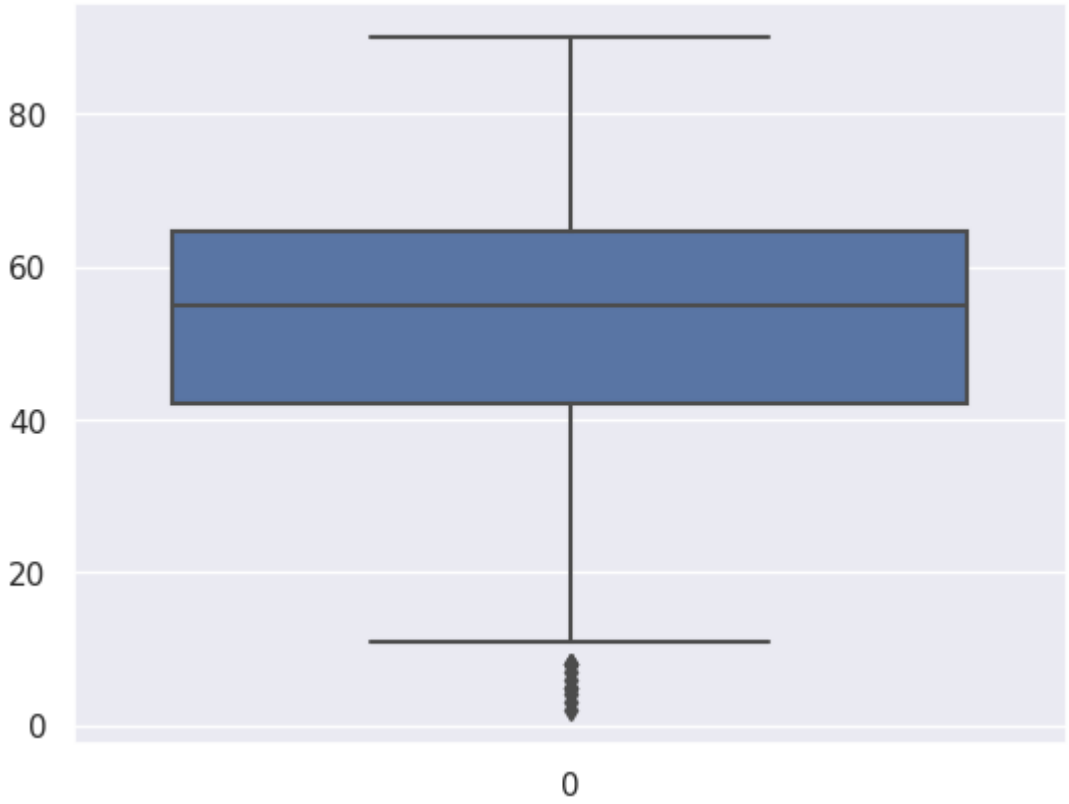
- After that we checked if the data set is balanced or not by graphical representation using pyplot and seaborn library.

- We plotted the different pairs of columns and observed the correlation between them graphically using seaborn.

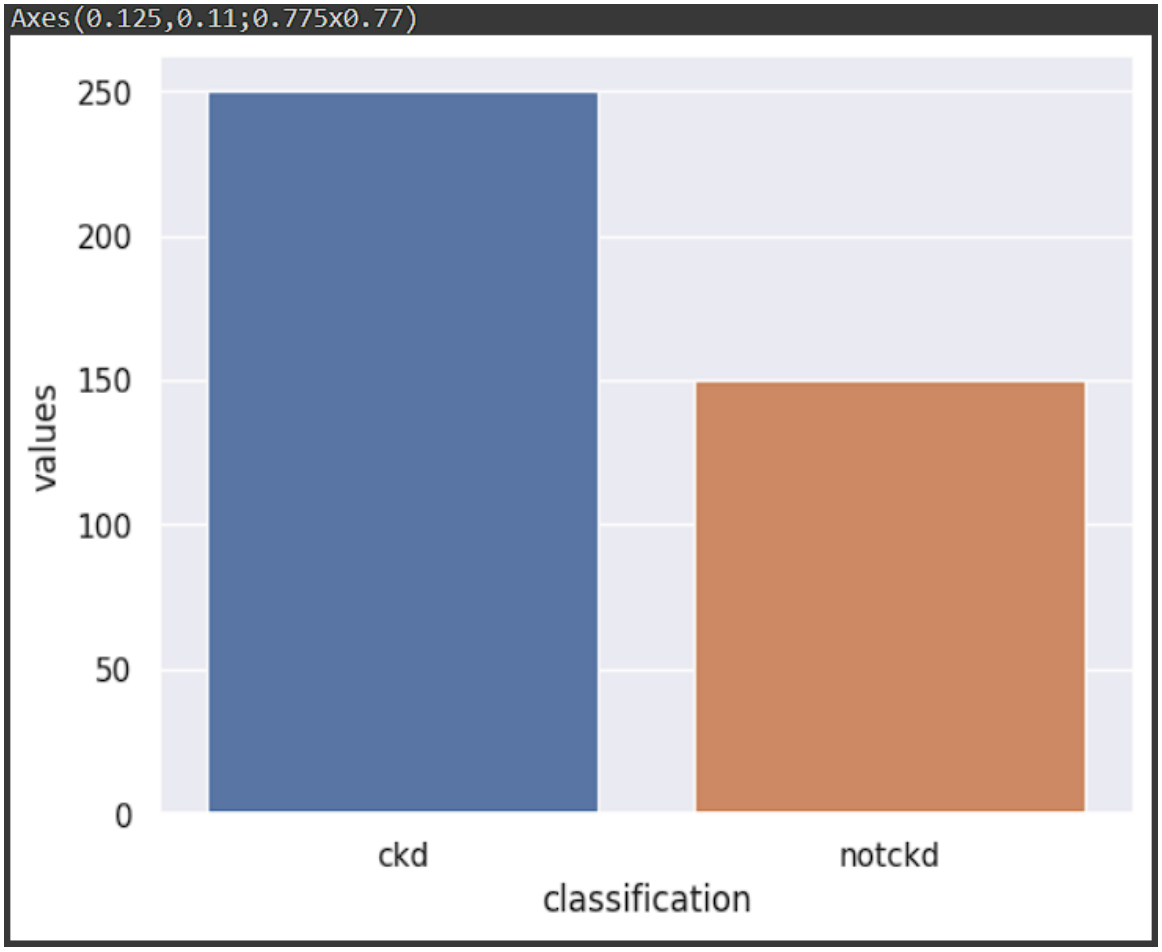
- Now, we checked the data distribution to find out what type of data it is.

- After that we find the outliers using the boxplot function which is present in the seaborn library and remove them.



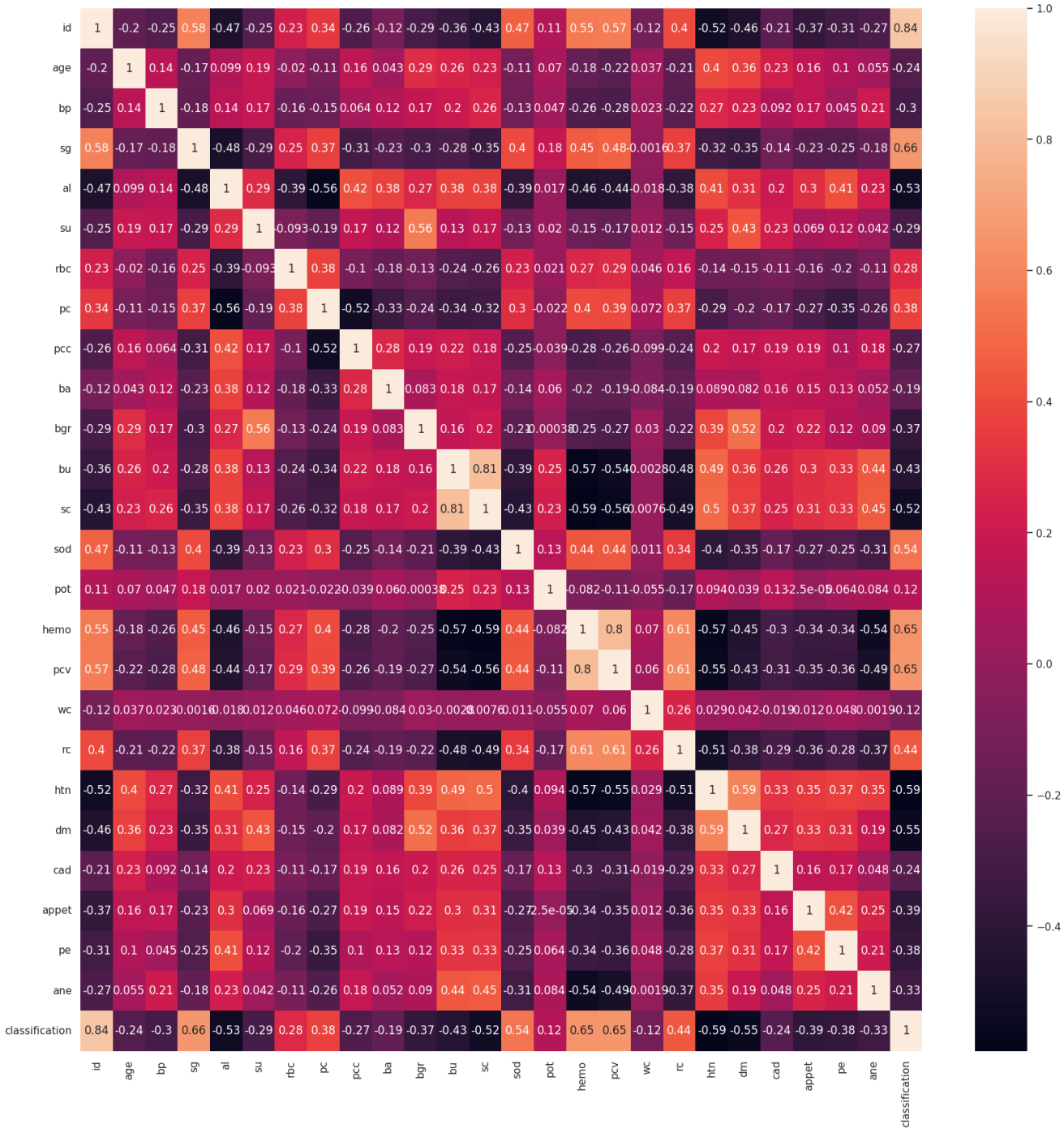


- We found our data to be unbalanced. We made it balanced. We could not use data augmentation as the data is related to the medical field. So, We did data resampling.



As, from above we can see that the data is not balanced. So, we did data resampling.

- We found the data to be highly correlated so we can apply PCA to reduce the data dimensionality.

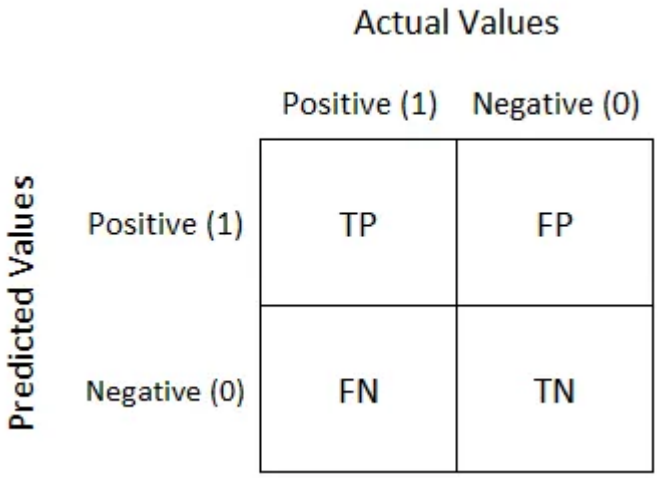


Building Classification Models:

- Now as our data is pre-processed, we build classification models like XGboost, SVM, KNN, Random forest,Logistic regression and Decision Tree.
- After that we build a stacked model of all these above models using a stacking ensemble learning method.

Building Confusion Matrix:-

Confusion Matrix is a performance measurement for machine learning classification. As this is a binary classification Problem we got a 2*2 confusion matrix




```
Confusion Matrix:
[[45  1]
 [ 0 54]]
```

We got the above confusion Matrix.
With the help of a confusion matrix we can calculate Precision , Recall, Accuracy and F1 score.

Recall:-

$$Recall = \frac{TP}{TP + FN}$$

Precision:-

$$Precision = \frac{TP}{TP + FP}$$

F-1 Score:-

It is basically the harmonic mean of precision and recall.

$$F - measure = \frac{2*Recall*Precision}{Recall + Precision}$$

Result:-

The project aimed to develop a predictive model for chronic kidney disease (CKD) using machine learning algorithms and clinical data. The model developed is found promising and an efficiency of 0.98 is recorded.

Code Link:-

GitHub Link:-

https://github.com/rdx-a/Chronic_kidney_disease-Prediction-using-stacking-ensemble-learning

Collab Link:-

https://colab.research.google.com/drive/1AhCYWjarmvITYxBwj0BnRQagia1ibCUU?usp=sharing

Conclusion:-

Through extensive data analysis and model development, we have successfully created a predictive model capable of accurately identifying individuals at risk of CKD. The model utilises crucial clinical features, such as serum creatinine levels, estimated glomerular filtration rate (eGFR), blood pressure readings, diabetes status, and hypertension status, to make reliable predictions. While the developed model shows promise, we acknowledge some limitations. The dataset used for development may not fully represent the entire population, and further external validation on larger and more diverse datasets is essential to assess its generalizability.

References:-

1. UCI Dataset
   https://archive.ics.uci.edu/

2. Adeola Ogunleye and Qing-Guo Wang, “XGBoost Model for Chronic Kidney Disease Diagnosis”, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 17, NO. 6, NOVEMBER/DECEMBER 2020