JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF CHRONIC KIDNEY DISEASES USING CLUSTERING TECHNIQUE IN MACHINE LEARNING

V.Sujatha[1]
, A. Naresh[2], M.Vasumathi Devi[3]
[1,2] Professor, Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur, Andhra Pradesh 522009

[3] Associate Professor, Vignan's Nirula Institute of Technology and Science for Women, Peda Palakaluru, Guntur, Andhra Pradesh 522009

ABSTRACT
In today's society, people are increasingly striving to maintain a higher level of awareness regarding their health. However, Chronic Kidney Disease (CKD) poses a unique challenge as it may not exhibit any symptoms or may require specific disease-related indicators, making it difficult to predict or anticipate its occurrence. Previous literature has utilized classification and regression techniques such as SVM and RF, but these methods have limitations due to the limited time frame considered. To address these limitations, we are implementing clustering techniques such as K-means in conjunction with Principal Component Analysis (PCA). The accuracy rate of estimating CKD using K-means is 99%. This study evaluates the prediction of renal disease using K-means in clustering. The results provide support for increased accuracy in the early categorization of chronic kidney disease.

Keywords: (CKD)chronic kidney disease, (SVM) Support Vector Machines, (RF) Random Forest.

I. INTRODUCTION:

Chronic kidney disease (CKD) is expressed as an abrupt repudiation in kidney function, resulting in the removal of waste and excess water from the blood and their excretion in urine. As CKD progresses, the body may retain water, electrolytes, and harmful substances. Initially, CKD may have minimal or no symptoms, and diagnosis may only occur at an advanced stage of the disease.

In the treatment of chronic kidney disorder, it is crucial to address the rudimentary cause in regulation to slow down the advancement of renal deterioration. Even if the cause of kidney disease is prohibited, the suffering will persist, eventually leading to end-stage renal failure without the need for dialysis or a kidney transplant.

The objective of this replica is to unfold and validate predictive replicas for the progression of chronic kidney disease. The vital focus is to assess the risk of renal failure, which is defined as the requirement for dialysis or renal replacement therapy. The model utilizes publicly available laboratory tests and other symptoms to predict normal renal function in patients with CKD stages 1 to 3. Additionally, the model provides guidance on leading a healthy lifestyle and enables physicians to visualize the risk and severity of the disease.

Clustering analysis enables the identification of patterns in data collection through k-means, allowing for the early prediction of future risks associated with specific diseases. The proposed model aims to predict whether a patient, based on their lifestyle, is currently experiencing CKD or may develop it in the future. This information can assist clinicians in making informed decisions regarding the appropriate course of action.

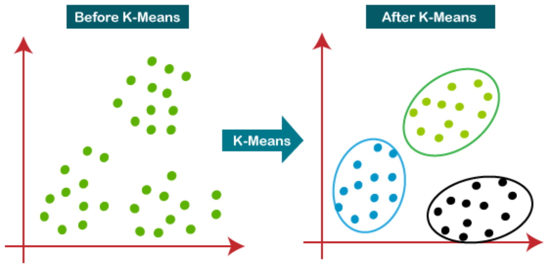

Fig1:K-Means

Using k-means alone also does not lead to high accuracy rates. Therefore, by combining PCA and K means we get higher and more accurate results. For example, PCA is used to capture most of the variance in the data.

The proposed sampling method involves the selection of the optimal algorithm for the intended purpose. Subsequently, the information is categorized as either CKD or NOT, encompassing both CKD and NOT. The eGFR value is then computed using the CKD-EPI equation in the event of CKD being present. By utilizing the eGFR test, the patient's present condition can be ascertained.

Our process consists of three steps one is preprocessing where we perform data correction due to errors and misclassification of data. The next step was exploratory data analysis in that we created diagrams showing the distribution and relationships of the various factors.

Didn't see much correlation, but hypothesis testing could be done to make it more convincing. The next step is clustering and forecasting, we thought if the clusters match the target categories, we have a forecasting model and we also have different clusters to analyze each category.

II. LITERATURE REVIEW:

This section elucidates the efforts undertaken prediction for chronic kidney failure through the utilization of ML techniques. In this study, we investigated the impact of dataset collection and its subsequent pre-treating on the precision of our research.

Two Members of the Legislative Assembly (MLAs) have received training to forecast the extent of Diabetic Kidney Disease (DKD) stages. SubsequentlyThe evaluation of their performance involved comparing it to the risk score employed by Centre for Disease Control and Prohibition in 2021, as employed by Marwa Almasoud and Tomas E Ward. The aim of the particular investigation is to create and authenticate ML algorithm that can accurately forecast the various stages of diabetic kidney disease (DKD) should be taken into consideration within a span of 5 years following the identification of type 2 diabetes mellitus. these algorithms demonstrated a remarkable performance in predicting any-stage DKD, the hold-out set yielded an area under the recipient operating criterion curve of 0.75, while the more severe endpoints exhibited an AUROC surpassing 0.82. In comparison, the CDC uncertain score yielded an AUROC of [insert value] [1].

In 2022, Saurabh Pal proposed the use of three ML classifiers, namely LR, DT, and SVM, to predict persistent kidney disease [2].The study aimed to generate and authenticate a prognostic replica for predicting the occurrence of persistent kidney failure.The UCI ML Repository was the source of the dataset on chronic kidney disease utilized by the researchers, which comprises 25 distinct features. The results were presented graphically through creating a python code to generate a boxplot. The precision of the logistic regression bagging ensemble method was improved from 93.28% to 94.53%. The decision tree demonstrated the highest level of optimality as a replica for predicting persistent kidney failure, achieving a precision of 97.23%.

The aim of the present study, authored by Saurabh Pal in 2023, is to predict the occurrence of CKD by utilizing definite and indefinite attributes through the implementation of SVM, RF, and ANN. [3] The objective of this study is to establish a framework for the timely identification of CKD, comprising three components: (a)The utilization of fundamental classifiers on attributes that are definite, (b)the utilization of onset classifiers on indefinite attributes, (c)The execution of onset classifiers on attributes, encompassing both definite and indefinite ones, is being carried out, and (d) the enhancement of outcomes by amalgamating the outcomes of the aforementioned three classifiers through a bulk vote. The Random Forest classifier demonstrated superior performance in terms of perfection, precision, recall, F-score, and ROC-AUC, achieving values of 0.92, 0.63, 0.55, 0.60, and 0.76, individually. Based on these outcomes, the superiority of the RF classifier over the SVM and ANN classifiers in the context of this persistent kidney disorder pattern can be inferred.

This article presents a proposal fora comparative analysis was conducted on ML techniques applied to persistent kidney disorder. The techniques to be utilized include the SVM, DT, NB, [4] and RF algorithms. This article is cited by Vineeta Gulati and Neeraj Raheja in 2021. For the implementation of the Support Vector Machine and Decision Tree algorithms, fourteen attributes from 400 patient records from the UCI repository were utilized. On the other hand, a total of 25 attributes were employed for the Random Forest algorithm. The Support Vector Machine model yields an accuracy of 96.7% with a mere 14 attributes, while the MLP model achieves an accuracy of 99.5% with 24 attributes. However, certain attributes, such as those influenced by prescribed medications that affect the patient's blood pressure, may increase the likelihood of incorrect predictions.

Logistic Regression, Naïve Bayes, and Random Forest algorithms were engaged by Qiong Bai, Chunyan Su, Wen Tang [5], and Yike Li to predict the incidence of end-stage kidney disease. The study utilized a dataset which consist of 748 patients who were diagnosed with persistent kidney disorder and patients were recruited from 2006 April to 2008 March, with a subsequent supplement period of 6.3 ± 2.3 years. Out of the total number of patients, 70 (9.4%) developed end-stage kidney disease. The results are presented as the mean accompanied by a 95% confidence interval. The study utilized the kidney failure risk equation (KFRE) and analyse the data by calculating the integral of the curve.

This proposal recommends the utilization of machine learning methodologies, specifically ACO and SVM classifiers, for the analysis and diagnosis [6] of persistent kidney disorder as reported in a 2020 article by Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A. The dataset used in this study is the publicly accessible CKD Dataset from the UCI warehouse, which consists of 400 selections belonging to two distinct classes. Among the 25 attributes, 11 are numeric, 13 are nominal, and one serves as the class attribute.Precision is ascertained by computing the quotient of true positives divided by the absolute sum of true positives and false positives.This determination of recall involves the computation of the proportion of true positives to the absolute sum of true positives and false negatives. The F-Means is calculated by considering both precision and recall.On the contrary, accuracy is determined by evaluating the rate of true positives and true negatives to the absolute sum of true positives, false positives, false negatives.

The prediction models utilized for persistent kidney disorder include RF, SVM, and DT was proposed by Dibaba Adeba Debal and Tilahun Melak Sitote in 2022. The analysis was conducted using a dataset comprising of 400 instances. Specifically,a dataset comprising 361 samples, encompassing 25 unique features, encompasses the pester class, was utilized. The accuracy assessment results reveal that the Probabilistic Neural Network (PNN) [7] achieved a score of 96.7%, while the Radial Basis Function (RBF) scored 87%. The SVM and Multilayer Perceptron (MLP) models achieved scores of 60.7% and 51.5%, respectively.The significance lies in acknowledging the PNN and RBF models, developed by Rady together with Anwar, attained the highest accuracy scores of 96.7% and 87%, respectively. Furthermore,the binary classification task was performed with utmost accuracy by the SVM model, achieving a score of 99.8%.

In 2022, Elias Dritsas and Maria Trigka utilized linear SVM, DT, and RF algorithms were utilized to predict the likelihood of developing persistent kidney disorder.Clinical and blood biochemical dimensions were obtained from a cohort of 551 patients who exhibited proteinuria. Various predictive models were evaluated, including RF, XGBoost, LR, ElasNet, [8] lasso and crest reversion, k-NN, SVM and ANN.The accuracies of the NB, RF, and LR models in forecasting the risk of chronic kidney disease were evaluated as 93.9%, 98.88%, and 94.76%, individually. The SVM replica with L2 penalty attained the topmost precision of 98.86%.

The purpose of this exploration is to utilizingML classification procedures, a dataset extracted from medical records can be analyzed to anticipate the numerous stages of CKD. This proposal was put forth by Hamida Ilyas, Sajid Ali, Mahvish Ponum, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Iftikhar, and Mubasher Hussain Malik. Specifically, the RF and J48 procedures were employed to unfold a reliable and practical replica for accurately identifying the various stages of CKD. Notably, [9] J48 outperformed random forest in predicting CKD across all stages, achieving an accuracy rate of 85.5%. This model holds potential for the development of an automated system to assess the severity of CKD.

III. ALGORITHM:

Action-1: Elect the value of K to consider the number of clusters.

Action-2: A set of K points or centroids can be elected in a random manner.

Action-3:Any data point must be allocated to the centroid that is closest to it, leading to the creation of the predetermined K clusters.

Action-4:Compute the variance and reposition the centroid of each cluster, taking into account the specified modifiers.

Action-5: proceed to come again with third step, every datapoint is reassigned to the closest centroid of its respective cluster.

Action-6: To make a prediction, it is recommended to employ a mixture of PCA and the K-Means clustering algorithm.

Action-7: In the event of any reassignment, please proceed to step 4. Otherwise, proceed to the FINISH stage.

Action-8: The model has been prepared and is now in a state of readiness.

IV. PROPOSED MODEL:

The text is divided into three separate sections that cover pre-processing, exploratory data analysis, clustering, and forecasting. The dataset used in this study consisted of 24 attributes, which were measured across 400 samples. Each sample contained multiple items. Out of the 24 attributes, 14 were numeric, while the remaining 10 were categorical. This information is relevant to the current dataset being analysed.

```
<class 'pandas. core. frame. DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 26 columns):
 #  Column    Non-Null Count Dtype
--- ------    -------------- -----
 0  ID        400 non-null   int64
 1  Age       391 non-null   float64
 2  sg        353 non-null   float64
 3  bp        388 non-null   float64
 4  Al        354 non-null   float64
 5  Su        351 non-null   float64
 6  pc        335 non-null   object
 7  Rbc       248 non-null   object
 8  Pcc       396 non-null   object
 9  BA        396 non-null   object
 10 bu        381 non-null   float64
 11 bgr       356 non-null   float64
 12 sca       383 non-null   float64
 13 sodi      313 non-null   float64
 14 pota      312 non-null   float64
 15 hemog     348 non-null   float64
 16 wc        295 non-null   object
 17 pcv       330 non-null   object
 18 rca       270 non-null   object
 19 Htn       398 non-null   object
 20 cad       398 non-null   object
 21 rm        398 non-null   object
 22 appe      399 non-null   object
 23 ane       399 non-null   object
 24 pe        399 non-null   object
 25 cls       400 non-null   object
dtypes: float64(11), int64(1), object (14)
memory usage: 81.4+ KBS
```

Pre-processing:

We started our process with data cleaning because we have a lot of misplaced values in the data. We can see that there is a typo that needs to be corrected. For example, there are no missing values in the value column and there are also 2 values for each section attribute. Many things have been misclassified. Some objects have several missing values. Therefore, we must use inference methods to discard them or pin them to appropriate values. I chose to pin them using KNNImputer considered from sklearn. Some things are outliers. Invalid outliers are the result of data entry errors, commonly known as noise in the data. Once passed, we need to classify the categorical and numeric attributes, and deal with the numeric attributes later. Now the data is clean. We begin to imagine segmented objects.

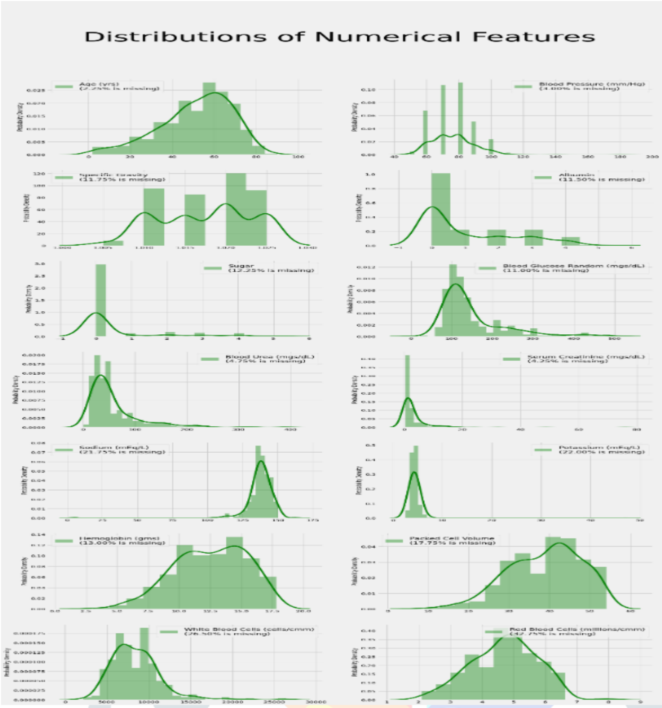

Fig:2:Distributions of Numerical Features

Some features contain very distant outliers. where others have independent values, but we can consider them as continuous values. Some items have such a high proportion of missing values that they cannot be assigned to a central trend. That would make classification difficult. Also, some features are subtle while most others are general. Some items have a very high percentage of missing values and some have none. Some differences seem to be common in these cases, such as diabetes and hypertension. Since we're only considering 400 samples, that means we're not doing any deductible statistics.

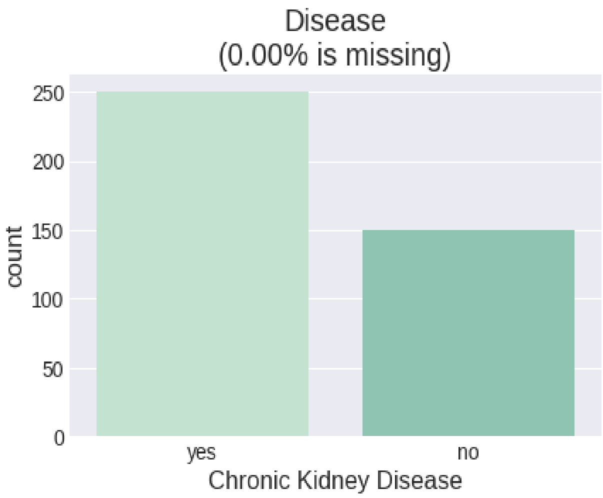

Fig:3: Disease Missing

It shows that this dataset represents neither a country nor even a group of ill people.

Now let's consider the missing values. As we proceed with imputation with KNN, we must first encode the categorical features. Therefore, it would be appropriate to use a single hot encoding because each categorical feature has only 2 values. Considering one-hot encoding, all NaN values are equal to 0, which is dangerous. KNN measures distance, which means we need to make sure our parts are on the same scale. Otherwise, the distances are mainly based on objects with large extinctions, meaning that the algorithm will pay close attention to those features. Pinned values are formatted data. But they will also be fed into predictive models. If the simulation does not represent assurance so well, then our model means that we learn useless information. The missing data will not be the same every time. Since there are many outliers, we use some nonlinear methods. A Quantile transform transforms a distribution to normal or equal. To make our data normal, the power transformer uses another non-linear transformation, so the “StandardScaler” standardizes our attributes. RobustScaler centers the data but, instead of splitting it by standard deviation, splits it by inter-quantile ranges that can be declared. The second complex scalar is defined in detail, and the first power transformer is normalized. It was the StandardScaler that kept the platelet distribution normal. RobustScaler has slightly increased the ‘0’ measure while nonlinear transforms have decreased it, especially the uniform quantile transform. I got good prediction results using Gaussian transformer data. I used a Gaussian quantile transformer for imputation and a StandardScaler for scaling the features for forecasting.

Exploratory Data Analysis:

We drew diagrams showing the distribution and relationships of the various elements. I didn’t find many synonyms. They also clearly showed that PCA made no significant difference, and the performance of Naïve Bayes was impressive as well.

In this case, we used the Sea Born Dark Web method. We identified the categorical data and then constructed the graph.We don’t see much correlation between factors.If they are then highly correlated with each other, we need to use more effective dimensionality reduction techniques.

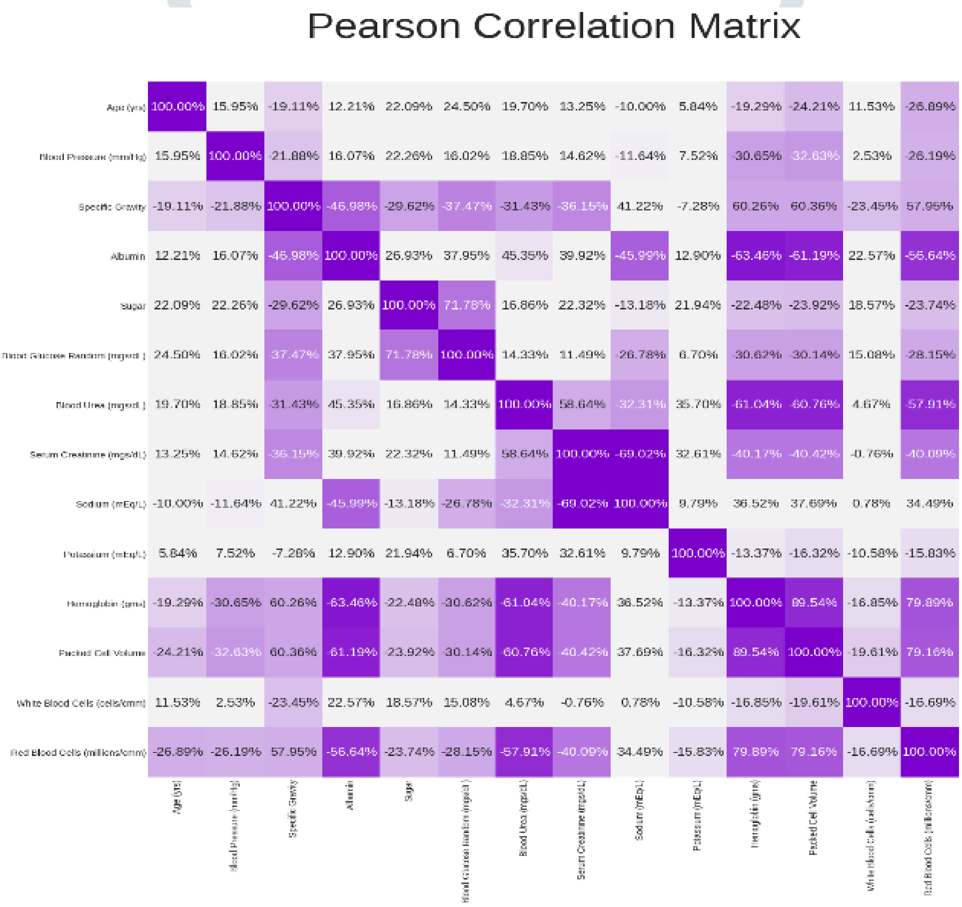

Fig:4: Correlation Matrix

Clustering & Prediction:

If each cluster is compatible with a singular category, then a predictive model is established. A comparable estimation can be made by measuring inertia for varying numbers of clusters. In order to make it applicable, it is necessary to identify an elbow point. Similar to KNN, K-Means is also influenced by the scaling of features. Therefore, it is imperative to standardize the features using sklearn's Standard Scaler.

Since there is no discernible elbow point, it is necessary to explore alternative options. There will always be a cluster where the majority do not have the disease. Other clusters may solely consist of individuals with the illness, which could be considered as overfitting.

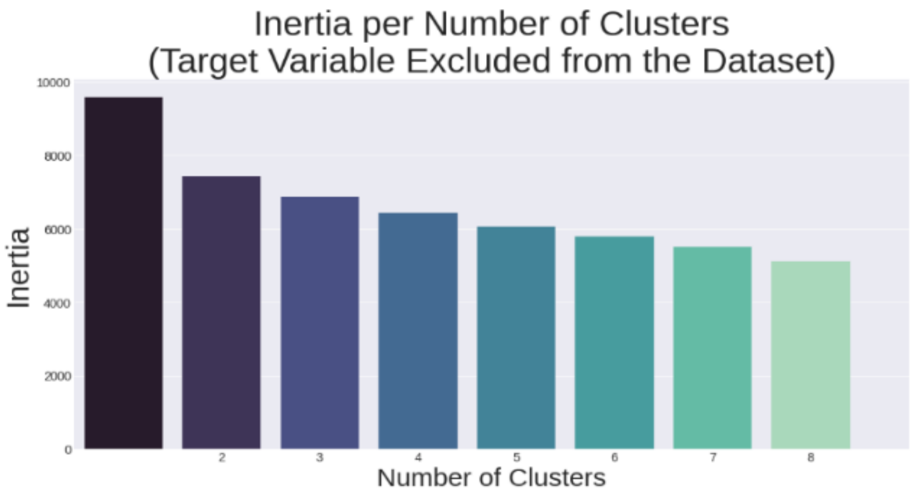

Fig:5: Cluster Classification

it is possible to achieve improved outcomes through the elimination of noise. The gradual decrease in the number of misclassified samples is observed as the number of clusters increases. This phenomenon may be attributed to the projection of K-Means clustering. To explore this further, we will attempt K-Means with PCA. However, before implementing PCA, it is advisable to examine the distribution of our data in space using T-SNE. Upon analysis, it is evident that the data is perfectly separated.

Upon segregating the categorical data and target values, it is observed that the categories are well separated, albeit not to the same extent as when the target variable was included. It is worth investigating the variation for each PCA component. This analysis reveals a distinct elbow point, indicating a significant detachment.

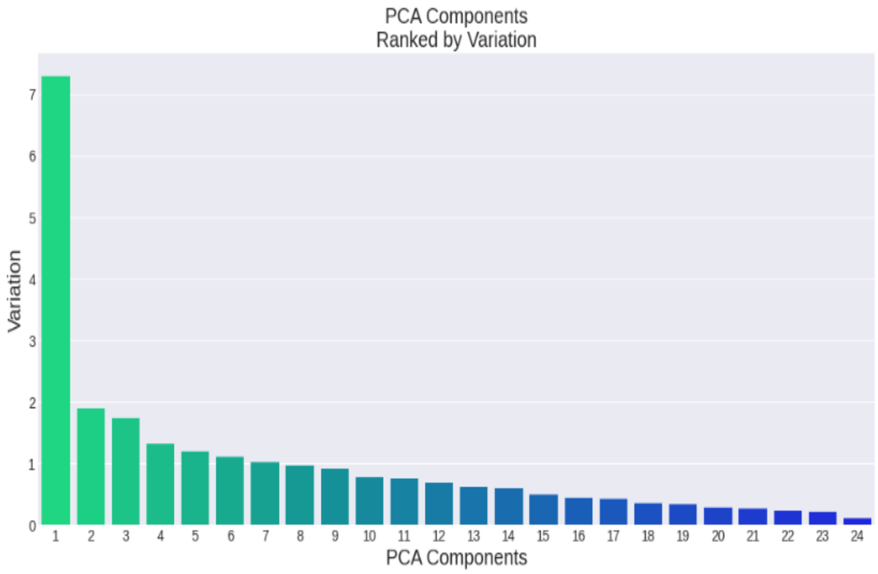

Fig:6: PCA components Variation

Certain components, particularly the initial ones, exhibit greater separation between categories. Conversely, the latter components introduce more noise at the edges of the data. It is important to note that T-SNE remains an approximate method, thus the certainty of its outcomes is uncertain.

V.RESULTS:

Our optimal outcome was achieved with the utilization of one principal component analysis (PCA) component and six clusters. The training accuracy was calculated to be 99% (238 out of 240). The inclusion of additional components beyond the first proved to be largely ineffective. As the number of clusters increased, the mixed cluster became more refined. While we could continue to increase the number of clusters to further refine the mixed cluster, there is a risk of overfitting the model. To accurately measure the model's performance, it is necessary to train it on a set of training data and then test it on different data. There are two methods for making predictions using a clustering algorithm:

1. Train the model and associate each cluster with the appropriate category, which is determined by the majority for mixed clusters. Then, assign each sample to be predicted to the nearest cluster center (centroid). In mixed clusters, the most represented values will be considered correct.

2. Train the model and associate each cluster with the appropriate category. Then, to ascertain a centroid for each category, the solution involves computing the weighted average of centroids for each category, with the weights being cluster sizes. Finally, assign each sample to be predicted to the nearest category centroid. This approach is likely to reduce variance but may increase bias.

Fig:7: Models Comparison

Excellent! Only one sample out of 160 has been misclassified, resulting in a testing accuracy of 99%. However, it should be noted that certain clusters may not be robust enough to withstand variations. In a different chronic kidney disease (CKD) dataset, these clusters may not be present or their structure may differ.

Ensuring that the data is well separated in space poses a significant challenge. Once this hurdle is overcome, the remaining tasks should be relatively straightforward. Principal Component Analysis (PCA) has proven to be effective in this case. However, there are other likely techniques that could be explored.

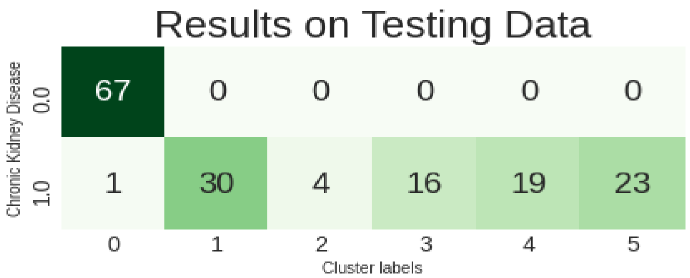

Fig:8: Testing Data Results

Given that the best results were obtained with one PCA component, it is reasonable to expect Linear Discriminant Analysis (LDA) to yield satisfactory results as well. At this point, it would be beneficial to conduct further cluster analysis. For instance, one could investigate correlations between clusters and features.

VI.CONCLUSION:

In this inspection, we examined the application of ML algorithms to detect persistent kidney disorder without any mandatory testing or features. Our approach involved utilizing machine learning clusters such, as k-means and PCA on a dataset consisting of 400 records. To ensure data accuracy we employed techniques like Naive Bayes, StandardScaler and RobustScaler. By analyzing two variables and applying a feature selection method we determined that factors like urine color, albumin levels and severity are crucial in predicting CKD.

We validated our classifiers using a 10 cross validation technique and achieved an impressive accuracy rate of 99% surpassing previous studies with fewer resources. This suggests that CKD can be detected using as 6 clusters or even just one cluster. Furthermore, our findings highlight the role of urine color, in identifying CKD.

Our goal is to extend the reliability of our conclusions by utilizing a more refined dataset or an equation that produces consistent results, rather than relying on other datasets with similar potential. Additionally, we intend to incorporate various features, such as CKD risk indicators like diabetes, hypertension, and type-break pedigree algorithms, to leverage the comprehensive dataset for predicting CKD in the future. This will offer valuable resources to reduce the prevalence of CKD.

VII.REFERENCES:

[1]. Allen A, Iqbal Z, Green-Saxena A, et al. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. BMJ Open Diab Res Care 2022;10: e002560. doi:10.1136/bmjdrc-2021-002560

[2]. Dibaba Adeba Debal and Tilahun Melak Sitote. Chronic kidney disease prediction using machine learning techniques Debal and Sitote Journal of Big Data (2022) 9:109 https://doi.org/10.1186/s40537-022-00657-5

[3]. Saurabh Pal. Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms. 30 March 2023 https://doi.org/10.1007/s11042-023-15188-1

[4]. Vineeta Gulati, Neeraj Raheja. Comparative analysis of machine learning techniques based on chronic kidney disease dataset. IOP Conf. Series: Materials Science and Engineering, 1131 (2021) 012010; 10.1088/1757-899X/1131/1/012010

[5]. Qiong Bai, Chunyan Su, WenTang1 & Yike Li. Machine learning to predict end stage kidney disease in chronic kidney disease. (2022) 12:8377; https://doi.org/10.1038/s41598-022-12316-z

[6]. Reshma S, Salma Shaji, S R Ajina, Vishnu Priya S R, Janisha A. Chronic Kidney Disease Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181IJERTV9IS070092; Vol. 9 Issue 07, July-2020.

[7]. Saurabh Pal. Chronic Kidney Disease Prediction Using Machine Learning Techniques. 16 August 2022. Biomedical Materials & Devices. https://doi.org/10.1007/s44174-022-00027-y

[8]. Dritsas, E.; Trigka, M. Machine Learning Techniques for Chronic Kidney Disease Risk Prediction. Big Data Cogn. Comput.2022, 6, 98. https://doi.org/10.3390/bdcc6030098

[9]. Hamida Ilyas, Sajid Ali, Mahvish Ponum, Osman Hasan, Muhammad Tahir Mahmood, Mehwish Iftikhar and Mubasher Hussain Malik. Chronic kidney disease diagnosis using decision tree algorithms. Ilyas et al. BMC Nephrology (2021) 22:273; https://doi.org/10.1186/s12882-021-02474-z.