International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

# AN EFFECTIVE PREDICTION OF CHRONIC KIDENY DISEASE USING DATA MINING CLASSIFIERS AND DIFFERENT DATA SAMPLING TECHNIQUES

**S.MAHESWARI[1], Dr. (Mrs) S. HEMALATHA[2]**

*M.Sc.,M.C.A.,M.Phil., Ph.D., (CS), Assistant Professor, Department of Computer Science. Master of Philosophy, Department of Computer Science. Shrimati Indira Gandhi College, Trichy-2, Tamil Nadu, India*

---

**Abstract:** Early prediction and proper treatments can possibly stop, or slow the progression of this chronic disease to end-stage, where dialysis or kidney transplantation is the only way to save patient's life. This paper is done by Chronic Kidney Disease dataset from UCImachinelearningrepository.CKDdatasetcontainsconta inimbalancedatasetfeatures.The problem of imbalanced classes arises when one set of classes dominate over another set of classes. The former is called majority class while the latter is called minority class. It causes the data mining model to be more biased towards majority class. It causes poor classification of minority classes. Hence, this problem throws the question of "accuracy" out of question. So that in this paper, split in to two phases. One is data sampling and other ones Prediction model. This paper is done by different data sampling methods like SMOTE, ADYSAN and SMOTE + Tome Links, K- means SMOTE. After getting modified data sampling dataset, to apply the different data mining algorithms .e Decision tree, Random Forest, SVM and KNN to predict the prediction of Chronic Kidney Disease in early stage. Based on accuracy, precision and sensitivity, specificity value from implemented tested data mining model to find out the best Sampling swell as Data mining algorithms.

## INTRODUCTION

It is all know that Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation? In simple words it's said that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of chronic kidney disease (CKD)per year in India. Inveterate kidney illness is additionally called renal failure. It may be a perilous infection of the kidney which produces continuous misfortune in kidney usefulness. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. In case CKD isn't recognized and cured in early arrange at that point persistent can appear taking after Indications: Blood Weight, weakness, week beans, destitute sustenance wellbeing and nerve harm, diminished resistant reaction since at progressed stages perilous levels of liquids, Electrolytes and squanders can construct up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by utilizing ancient CKD understanding information to prepare anticipating show. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of your blood keratinize, age, race, gender, and other factors. The earlier disease is detected the better chance of slowing or stopping its progression. Based upon GFR the renal damage severity by CKD is categorized into following five stages.

| STAGE | DESCRIPTION | GFR(mL/min) |
|---|---|---|
| - | At increased risk for CKD | >=90withriskfactors |
| 1 | Kidney damage with normal or increased GFR | >=90 |
| 2 | Mild decrease in GFR | 60-89 |
| 3 | Moderate Decrease in GFR | 30-59 |
| 4 | Severe decrease in GFR | 15-29 |
| 5 | Kidney Failure | <15ordialysis |

Table1.1CKDiscategorizedintofollowingfivestages

There are few studies related to the automatic diagnosis of CKD in the literature but they are not fully effective to help medical experts. For CKD, the authors usually use predictive analysis models to predict its progression and, in the best of scenarios, try to stop the disease. Information Mining procedures are connected on it to assess the execution in case of foreseeing whether the individual has kidney maladies or not. Kidney disease is a growing problem. For most individuals, kidney harm happens gradually over numerous a long times, frequently due to diabetes or tall blood weight. This is called chronic kidney disease. When somebody includes a sudden alters in kidney work since of illness, or damage, or has taken certain drugs usually called intense kidney harm. This can occur in a person with normal

kidneys or in someone who already has kidney problems. The main risk factors for developing kidney disease are: Diabetes, High blood pressure, cardiovascular (heart and blood vessel) disease, and family history of kidney failure. In 2010, a case study for predicting CKD in a local hospital in England was Presented in where two conditions were considered:

(a) Moderate to severe CKD
(b) End stage kidney failure.

## PROBLEM STATEMENT

Lesson Awkwardness could be a common issue in machine learning, particularly in classification issues. Imbalance data can hamper our model accuracy big time. Machine Learning algorithms tend to produce unsatisfactory classifiers when faced with imbalanced datasets. Most machine learning algorithms work best when the number of samples in each class is approximately equal. This is because most algorithms are designed to maximize accuracy and reduce errors, so a balanced dataset is needed in such cases.

## OBJECTIVE

The objective of this work is to explore data sampling techniques and to balance the imbalanced dataset to improve the performance of the input dataset.

## LITERATUREREVIEW

There are many researchers who work on prediction of CKD with the help of many different classification algorithms. And those researchers get expected output of their model.

Gunarathne W.H.S.D et.al. Has compared results of different models. And finally, they concluded that the Multiclass Decision Forest algorithm gives more accuracy than other algorithms which is around 99% for the reduced dataset of 14 attributes.

S.Ramya and Dr.N.Radha worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The proposed work deals with classification of different stages of CKD according to its gravity. By analyzing different algorithms like Basic Propagation Neural Network,RBF and RF. The analysis results indicates that RBF algorithm gives better results than the other classifiers and produces 85.3% accuracy.

S.DilliArasu and Dr. R. Thirumalaiselvi 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.

Asif salekin and john stankovic they use novel approach to detect CKD using machine learning algorithm. They get result on dataset which having 400 records and 25 attribute which gives result of patient having CKD or not CKD. They use k-nearest neighbors, random forest and neural network to get results. For feature reduction they use wrapper method which detects CKD with high accuracy.

Pinar Yildirim searches the effect of class imbalance when the data is trained 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 perception.

Sahil Sharma, Vinod Sharma, and Atul Sharma, has assessed 12 different classification algorithms 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.

BaisakhiChakraborty, 2019 proposed development of CKD prediction system using machine learning techniques such as K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine and Multi-Layer Perception Algorithm. These are applied and their performance is compared to the accuracy, precision, and recall results. Finally, Random Forest is chosen to implement this system.

Arif-Ul-Islam, 2019 proposed a system in which prediction of disease is done using Boosting Classifiers, Ant-Miner and J48 Decision Tree. The aim of this paper is twofold that is, analyzing the performance of boosting algorithms for detecting CKD and deriving rules illustrating relationships among the attributes of CKD. Experimental results prove that the performance of Gadabouts was less that of Logit Boost by a fraction. S.Belina V, 2018 proposed a system that uses extreme learning machine and ACO for CKD prediction. Classification is done using MATLAB tool and ELM has few constraints in the optimization. This technique is an improvement under the sigmoid additive type of SLFNs.

Siddheshwar Tekale, 2018 described a system using machine learning which uses Decision tree SVM techniques. By comparing two techniques finally. Concluded that SVM gives the best result. Its prediction process is less time consuming so that doctors can analyze the patients within a less time period.

The algorithms provide a basis to identify high risk patients who might benefit from more detailed assessment, closer monitoring or interventions to reduce their risk. Later, in 2011, Tangri et al. develop prediction models using demographic, clinical, and laboratory data.The most accurate model included age, sex, estimated Glomerular Filtration Rate (GFR), albuminuria, serum calcium, serum phosphate, serum bicarbonate, and serum albumin. In 2014, the use of data mining techniques for predicting kidney dialysis survival was presented, where three data mining techniques were used namely ANN, Decision tree and Logical Regression, being the first one the best among the three achieving an accuracy value of 93.852%, a sensitivity of 93.87% and a specificity of 93.87%.

In data mining, data is acquired from distinct sources; however, these datasets are often distorted. Majority of the real-world datasets are imbalanced, viz, majority of the instances are labeled to be belonging to one class called majority class while very few are labeled to be belonging to another class called minority class. This class misbalancing problem is highly prevalent in medical data and currently is the hot topic of research.With such sort of imbalanced datasets, classifiers built have a inclination of creating tall exactness for lion's share course and destitute expectation precision for minority classes.

For dealing with of imbalanced information, analysts have proposed arrangements both at information and algorithmic level, but the Destroyed procedure utilized in ponders has appeared way better execution within the accessible writing. This have been proposed in a decision support predictive model based on SMOTE dataset rebalancing algorithm and various data mining classification techniques. In utilized Destroyed to diminish course lopsidedness in our dataset within the first step and within the moment step, the rebalanced dataset was utilized with different information mining classifiers to find the best classifier for CKD disease.

## EXISTING SYSTEM

There is many research work done on imbalanced dataset.

• Dal Pozzolo A worked on incremental learning together with utilizing examining procedures and last expectation was done utilizing gathering of those models..They observed better results with random forest classifier along with Synthetic Minority Oversampling Technique (SMOTE) sampling technique.

• Drummond C and Holte RC provided a new insight on the two sampling techniques. It directs us to adapt the machine learning algorithms to imbalanced data by balancing it. They observed that under-sampling technique beats over-sampling. Their investigate implied the utilize of arbitrary timberland classifier with information adjusting, to realize way better forecast exactness.

## PROPOSED SYSTEM

• This paper is done by chronic kidney disease dataset from UCI machine learning repository. CKD dataset contains contain imbalance dataset features.

• The problem of imbalanced classes arises when one set of classes dominate over another set of classes. The former is called majority class while the latter is called minority class. It causes the data mining model to be more biased towards majority class. It causes poor classification of minority classes. Hence, this problem throws the question of “accuracy” out of question.

• So that in this paper, split in to two phases. One is data sampling and other one is Prediction model.

• This paper is done by different data sampling methods like SMOTE, ADASYN: Adaptive Synthetic Sampling Approach, SMOTE + Tomek Links, K- means SMOTE.

• After getting modified data sampling dataset, to apply the different data mining algorithms i.e Decision tree, Random Forest, SVM and KNN to predict the prediction of chronic kidney disease in early stage.

• Based on accuracy, precision and sensitivity, specificity value from implemented tested data mining model to find out the best Sampling as well as Data mining algorithms. This is used by Scikit Learn library to implement and testing our goal.

• And different classifier models will be trained with the balanced dataset to improve the performance.

## METHODOLOGY

This paper is done by different data sampling methods like SMOTE, ADASYN: Adaptive Synthetic Sampling Approach, SMOTE + Tome Links, K- means SMOTE. After getting modified data sampling dataset, to apply the different data mining algorithms Decision tree, Random Forest, SVM and KNN to predict the prediction of chronic kidney disease in early stage. This is used by Sickest Learn library to implement and testing our goal.

## DATAMININGOVERVIEW

The definition of Information Mining or Information Disclosure in Databases is the activity that extricates a few modern imperative data contained in expansive databases. The target of information mining is to discover startling characteristics, covered up highlights or other vague connections within the information based on techniques' combination. Today, many applications in a wide and various ranges of business founded and worked in this regulation. In 1996, U. Fayyad, G.Shapiro characterized the common information disclosure prepare as an intuitively and iterative prepare including more or less the taking after steps: understanding the application field, information selecting, preprocessing and cleaning data, integration of data, data reduction and transformation, selecting algorithms of data mining, interpretation and description of the results and using the discovered knowledge. In fact, the data mining can be classified into two categories descriptive and predictive. Really, within the later a long time, information mining is involving extraordinary position of consideration zone within the society of commerce or managing an account since its flexibility in working with a huge sum of information, and turning such information into clear data and knowledge Most of the people may be confused in understanding between the terms "knowledge discovery" and "data mining" in different areas. Information revelation in databases is the method of distinguishing substantial, novel, likely valuable and at long last reasonable patterns/models with information. On the other hand, data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that under some acceptable computational efficiency limitations, finds patterns or models in data.

## DATAMININGTASKS

Data mining used in different type of techniques to extract the knowledge from the data, the techniques are:

**Anomaly detection** (Outlier/ deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation. (Ex-bank fraud identification)

Three types of anomaly detections:

**Unsupervised Anomaly Detection techniques** detect anomalies in an unlabeled test data set under the assumption

**Supervised Anomaly Detection Techniques** require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier.

**Semi-Supervised Anomaly detection techniques** construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model

**Association rule learning (Dependency modeling**) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits.

Utilizing affiliation run the show learning, the grocery store can decide which items are habitually bought together and utilize this data for showcasing purposes. This is sometimes referred to as market basket analysis.

**Clustering** – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

**Classification** – is the errand of generalizing known structure to apply to unused information. For illustration, an mail program might endeavor to classify an email as "true blue" or as "spam".

**Regression** – attempts to find a function which models the data with the least error. ϖ**Summarization** – providing a more compact representation of the data set, including visualization and report generation. In this data mining techniques used to mining the different kind of data, the forms of data for mining applications are database data, data warehouse data, transactional data, data streams, graph or networked data, spatial data, text data, multimedia data, and the World Wide Web data.

**Prediction -** Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.

For illustration, forecast models in information mining are utilized by a showcasing director who anticipate that how much sum a specific client will spend amid a deal, so that up and coming sale amount can be planned accordingly.

## CLASSIFICATION TECHNIQUES

Classification may be a prepare revelation demonstrate (capacities) that portray and recognize classes of information or concept that points to be utilized to foresee the course of the protest which name. Class is unknown. Classification is part of data mining, where data mining is a term used to describe the knowledge discovery in databases. Information mining is additionally a handle that employments factual methods,

science, manufactured insights, and machine learning for extricating and distinguishing valuable data and important information from a assortment of large datasets. The classification process is based on four components

**Class** -Categorical subordinate variable within the frame that represents the 'label' contained within the question.

For example: heart disease risk, credit risk, customer loyalty, the type of earthquake.

**Predictor** -The independent variables are represented by characteristic (attribute) data. a. For example: smoking, drinking alcohol, blood pressure, savings, assets, salaries.

**Training dataset** -One data set that contains the value of both components above are used to determine a suitable class based on predictor. **Testing dataset** -Containing new data which will be classified by the model that has been Classification could be an information mining work that allocates things in a collection to target categories or classes. The objective of classification is to precisely predict the target course for each case within the information.

There is some classical data-mining classification algorithms listed below.

- Association rule mining
- Bayesian Classification
- Decision tree classification
- Nearest Neighbor
- Neural Networks(Back Propagation)
- Support Vector Machines(SVMs)

**This paper is done by using**

- Decision tree
- KNN
- SVM
- Random Forest

## DETAILSOFALGORITHMS

## DECISIONTREE

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression problems in which classification is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies different instances by sorting them based on the attribute (feature) values. Each and every node in a decision tree represents an attribute in an instance to be classified for given data set. All branches represent an outcome of the test; each leaf node holds the class label. The occasion is classified based on their include esteem. There are numerous methods for finding the feature that best divide the training data such as information gain, gain ratio, Gini indexetc. The foremost common way to construct choice trees by utilizing top-down eager strategy dividing, beginning with the preparing set and recursively finding a part include that maximizes a few neighborhood model. They are the three basic algorithms are widely used that are ID3, C4 .5 and CART.

## ID3ALGORITHM

ID3 is an iterative Dichotomies 3. It is primitive decision tree algorithm introduced by Quinlan Ross in 1986. The basic idea is to make a decision tree by using the top-down greedy approach. ID3 uses the information gain for selecting the best feature For characterizing the data pick the calculation must be done first of the Entropy (X): - Σ [P (I) log (2) P (I)]

Where – P (I) alludes to extent of S have a place to Lesson I, S are all the records.(instance), C refer as Class, Σ is over C i.e., Summation of all the classifier. Information Gain (X, A) = Entropy(X) – Σ ((|Xv|/|S|) Entropy (Xv)) Where A is feature for which gain will be calculated, V is all the Possible of the feature, Xv is the no of element for each V.

## C4.5ALGORITHM

C4.5 is the decision tree algorithm generated Quinlan. It is an extension of ID3 algorithm. TheC4.5 can be Refer as the statistic Classifier. This algorithm uses Gain radio for feature selection and to be constructing the decision tree. It handles both continuous and discrete features. C4.5 algorithm is widely used because of its quick classification and high precision rate.

The pickup radio “Normalized” the data pick up as takes after (Quinlan 1993).

Gain Radio (A, X) = information Gain (X, A) / Entropy (X, A).

## CARTALGORITHM

It is stand for Classification Regression Tree introduced by Bremen CART uses binary splitting that means the node has exactly two outgoing edges and splitting are done by the Gini index. Gini Index = 1 – Σ P^2 (I)

The properties of CART are that it is able to produce the relapse tree.

## MATHEMATICALFORMULATION

Given training vectors $x_i \in R^n$,i=1…, l and a label vector y $\in R^i$, and a label vector y ∈, a decision tree recursively

partitions the space such that the samples with the same labels are grouped together. Let the information at hub m be spoken to by Q. For each candidate split θ=(j,$t_m$)consisting ofa feature and threshold tm, partition the data into$Q_{left}$(θ) and$Q_{right}$(θ)subsets.

$$Q_{left}(\theta) = (x, y)|x_j <= t_m$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$

The impurity at m is computed using an impurity function H (), the choice of which depends on the task being solved (classification or regression)

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

Select the parameters that minimize the impurity.

$$\theta^* = \mathrm{argmin}_\theta \, G(Q, \theta)$$

Recuseforsubsets$Q_{le}$(θ*)and$Q_{right}$(θ*)untilthemaximumallowabledepthisreached,$N_m$<$Min_{samples}$or$N_m$=1.

## KNEARESTNEIGHBOURS (KNN)

K Closest Neighbors (KNN) could be a straightforward calculation that stores all accessible cases and classifies unused cases based on a similitude degree (e.g., separate capacities).A case is classified by a majority vote of its neigh bourse, with the case being assigned to the class most common amongst its KNN measured by a distance function. The Euclidean separate between two focuses x and y is given by the condition

TheEuclideandistancebetween two points x and y is given by the equation

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

The esteem of k (the positive numbers) is decided by assessing the information set. Cross- validation is another way to subsequently determine a good k value by using an independent data set to validate the k. Here the value taken are (k =1, 3, 5 and 10) and it produces good result at k =10.

This suggests that the k esteem gets bigger the result will be more exact. In most cases the optimal k value will be between 3 and 10.

## SUPPORTVECTORMACHINE (SVM)

SVM is a set of related supervised learning method used for classification and regression SVM is represented with the help of hyper plane. For example, given Set of points belong ingoted throne of the two classes, an SVM find shay per plane having the agree stops possible fraction of points of the same class on the same plane.

This separating hyper planes called the optima's pirating per plane (OSH)that maximizes the distance between the two parallel hyper planes and could minimized her is kormas classifying Examples of the given test dataset. It is shown in follow in figure. Given some training data Q, setoff point of the form:

Q={(X1,Y1),(X2,Y2)….(Xn,Yn)}

Here Yi=-1/1thatdenotestheclasstowhich data point Xnbelongs. Any hyper planecanbewrittenasthesetofpointssatisfying$W^T$.X+b=0HereWisnormaltohyper plane; b isaconstant.

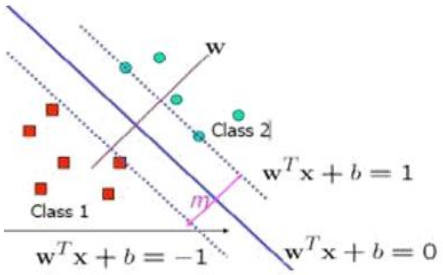



Figure SVM training with two classes

The following formula constraint is added: $W^T$. Xi + b<=-1 for xi of first class and$W^T$.Xi+b>=1forxiofsecondclassThiscanbewrittenas,The formula is minimized by ||w||Subject toYi($W^T$.xi+b)>=1,for alli.

## RBFKERNEL

Proper parameters setting might improve the SVM classification curacy for given dataset. RBF erne function is used as classifier. It is able to analyses eight dimensional data. The output of the kernels modified by the Euclidean distance. RBF Kernel function can be defined as:K (Xi,Xj)=EXP(-y||Xi- Xj||2)Here, yisKernelparameterandXi,XjareSupportvector and testing data point Xi.

## SOFTMARGIN

If the resists nosy per plane that can split the"yes"and"no"examples, the Soft Margin method will choose a hyper plane that splits the examples as cleanly

as possible. The method introduces nonnegative slack variables, which measure the degree of misclassification of the data.

$y_i(W.X_i - b) \geq 1 - \varepsilon_i 1 \leq i \leq n$

The objective function to be increased by a function which penalizes non-zero ξ, and the optimization becomes a trade-off between a large margin and as mall error penalty.

If the penalty function is linear, the optimization problem becomes:

$$Minimize \frac{1}{2} \quad \| w \|^2 + C \sum_{i=1}^{n} s_i$$

## RANDOMFORESTS

Arbitrary woodlands (RF) develop numerous person choice trees at preparing. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As they use a collection of results to make a final decision, they are referred to as Ensemble techniques.

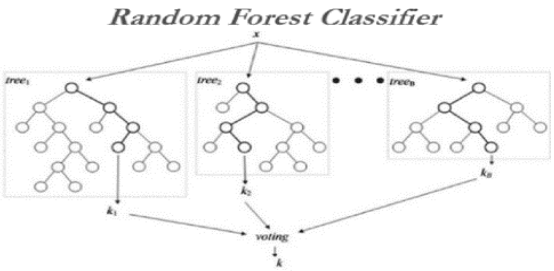



Figure Random Forest tree

## FEATUREIMPORTANCE

Include significance is calculated as the diminish in hub pollution weighted by the likelihood of coming to that hub. The hub likelihood can be calculated by the number of tests that reach the hub, separated by the entire number of tests. The higher the esteem the more critical they include.

## IMPLEMENTATIONINSCIKIT-LEARN

For each decision tree, Scikit-learn calculate a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Nisub(j)=theimportanceofnodej

W sub(j)=weightednumberofsamplesreachingnodej

Csub(j)=theimpurityvalueofnodej

left (j)=childnodefromleftsplitonnodej

right(j)=childnodefromrightsplitonnodej

*sub()isbeingusedassubscriptisn'tavailableinMedium*

Theimportanceforeachfeatureonadecisiontreeisthencalculatedas:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

Fisub(i)=theimportanceoffeaturei

Nisub(j)=theimportanceofnodej

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

The final feature importance, at the Random Forest level, is it's average over all the trees. The sum of the feature's importance value on each tree is calculated and divided by the total number of trees:

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

RFfisub(i)=theimportanceoffeatureicalculatedfromalltresesintheRandomForestmodel.

Normfisub(ij)=thenormalizedfeatureimportanceforiintreej

T=totalnumberoftrees.

## DATASETDESCRIPTION

The CKD dataset was collected from 400 patients from the University of California, Irvine Machine Learning Repository. The dataset comprises 24 features divided into 11 numeric features and 13 categorical features, in addition to the class features, such as "ckd" and "notckd" for classification. Features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood

urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell tally, ruddy blood cell tally, hypertension, diabetes mellitus, coronary course illness, craving, pedal edema, and frailty.The diagnostic class contains two values: ckd and notckd. All features contained missing values except for the diagnostic feature. The dataset is unbalanced because it contains 250 cases of "ckd" class by 62.5% and 150 cases of "notckd" by 37.5%.

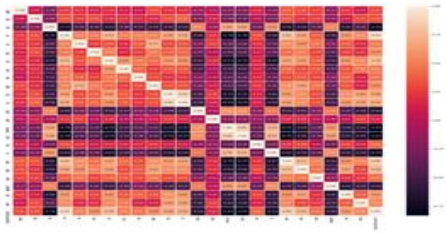


figure4.Datasetdescription

## EVALUATIONMEASURESFORCLASSIFICATION TECHNIQUES

The criteria taken for our comparison of classifier are Accuracy, Precision, F- Measure. For calculating these criteria, in utilized the disarray framework in our calculation prepares.

The general view of confusion matrix is given below.

| **Confusion Matrix** | | |
|---|---|---|
| Predicted Class<br>Yes no | | |
| Actual class | | |
| Yes | TP | FN |
| No | FP | TN |

Inconfusionmatrixthepredictedclassistheclassthatispredictedbytheclassifier andtheactualclassistheclassthatisgiveninthedataset.

**True positives (TP):** These refer to the positive tulles that were correctly labeled by the classifier. Let Tube the number of true positives.

**True negatives (TN):** These are the negative tulles that were correctly labeled by the classifier. Let TN be the number of true negatives.

**False positives (FP):** These are the negative topples that were incorrectly labeled as positive (e.g., tulles of class buy computer = no for which the classifier predicted buys computer= yes). Let FP be the number of false positives.

**False negatives (FN):** These are the positive tuples that were mislabeled asnegative(e.g.,tuplesofclassbuycomputer=yesforwhichtheclassifier predictedbuyscomputer=no).LetFNbethenumberoffalsenegatives.

Fromtheconfusionmatrixtheaccuracy,sensitivityandspecificity, precision arecalculatedas follows.

**Accuracy:**

Classificationaccuracyisthepercentageofinstancesthatare correctly classifiedbythemodel.Itiscalculatedasthesumofcorrectclassificationdivided by the total number ofsamples.

**Accuracy = (TP+TN)/ TP+FP+TN +FN**

**Sensitivity:** It is the measure of the ability of a classification model to select instances of certain class from the dataset. It is the proportion of actual positive which are predicted positive.

**Sensitivity =TP/TP+FN**

**Specificity:** This is a measure that is commonly used in two class problems where the focus son particular class. It is the proportion of the negative class that was predicted negative and it is also known as the true negative rate.

**Specificity =TN/FP +TN**

**Precision:** In an imbalance classification problem with two classes, precision is calculated as the number of true positives divided by the number of true positives and false positive

**Precision =TP/ TP+FP**

## DSIGNANDDEVOLPEMENT

This paper is done by using different data sampling me thuds' like SMOTE, ADYSAN and SMOTE +Tome Links, K-means SMOTE. After getting modified data sampling dataset, to apply the different data mining algorithms i.e Decision tree, Random Forest, SVM and KNN to predict the prediction of chronic kidney disease in early stage.

The above listed algorithm is implemented with help of Sickly-learn. Sciklit-learn is a free software machine learning library for the Python programming language. Sciklit-learn is designed to interoperate with the Python numerical and scientific libraries NumPy and Spicy. Those algorithms have been tested with CKD Dataset downloaded from UCI machine learning data repository. The performances of the algorithms have been compared in terms of Accuracy, precision, Sensitivity, and

Specificity. Based on Heart disease dataset, confusion matrix can be taken as,

**TP (True Positive):** The no. of people who actually suffer from 'Disease' among those who were diagnosed 'Disease'.

**TN (TrueNegative):** States the number of people who are 'healthy' among those who were diagnosed 'Disease'.

**FP(FalsePositive):** Depicts the number of persons who are unhealthy that is 'Disease'butwasdiagnosedas'healthy'.

**FN(FalseNegative):**Thenumberofpeoplefoundtobe'unhealthy'amongthosewhowere diagnosed as 'Disease'.

The performance of classification can be measure in the fallowing criteria.

- Sensitivitymusthavehighpercentage.
- Specificitymusthavelowpercentage.
- Accuracymusthavehighpercentage.
- Precisionmusthavehighpercentage.

ThefollowingDetailArchitectureofworkflowdiagramcanbedescribedentireprojectprocess.

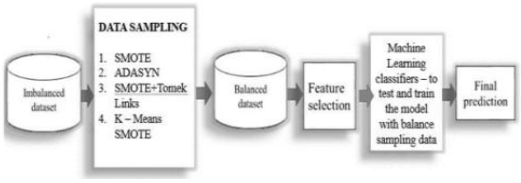



Figure5.1Detail Architecture of work flow diagram

## SAMPLINGTECHNIQUESDETAILS

### SMOTE(Synthetic Minority Over sampling Technique)

• SMOTE is an oversampling technique where the synthetic samples are generated for the minority class.

• This algorithm helps to overcome the over fitting problem posed by random oversampling.

• It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

### Working procedure

• At first the total no. of oversampling observations, N is set up.

For the most part, it is chosen such that the twofold course dispersion is 1:1. But that can be tuned down based on require.

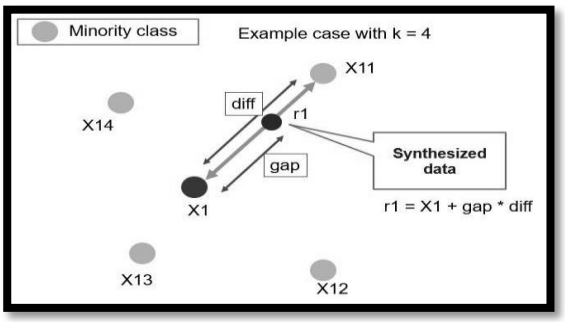
Then the iteration starts by first selecting a positive class instance at random. Next, the KNN's (by default 5) for that instance is obtained.

• At that point the emphasis begins by to begin with selecting a positive lesson occurrence at irregular Next; the KNN's (by default 5) for that instance is obtained.

• At last, N of these K instances is chosen to interpolate new synthetic instances.

To do that, utilizing any remove metric the distinction in separate between the include vector and its neighbors is calculated.

Now, this difference is multiplied by any random value in [0,1] and is added to the previous feature vector.




```
from imblearn.over_sampling import SMOTE

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using SMOTE
smt = SMOTE()
#X_train, y_train = smt.fit_resample(X_train, y_train)
X_train_sm, y_train_sm = smt.fit_resample(X_train, y_train)

counter = Counter(y_train_sm)
print('After',counter)
```

### ADASYN:AdaptiveSyntheticSamplingApproach

ADASYNisageneralizedformoftheSMOTEalgorithm.

This algorithm also aims to oversample the minority class by generatingsyntheticinstancesfor it.

But the difference here is it considers the density distribution,
***ri***whichdecidestheno.ofsyntheticinstancesgeneratedforsampleswhichdifficulttolearn.

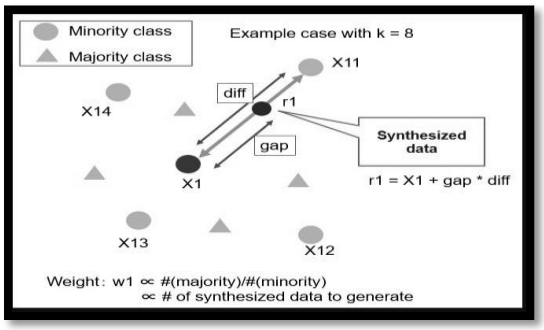
Due to this, it helps in adaptively changing the decision boundaries basedon the samples difficult to learn. This is the major difference compared toSMOTE.

### WorkingProcedure

Fromthedataset,thetotalno.ofmajority$\mathbf{N^-}$andminority$\mathbf{N^+}$arecapturedrespectively. The threshold value is presented by, $\mathbf{d^{th}}$for the maximumdegree of class imbalance. Total no. of synthetic samples to be

generated,G=($N^-$ –$N^+$) x *β.Here,*$\beta=(N^+/N^-)$.Foreveryminoritysample$x_i$,KNN'sareobtainedusingEuclideandistance, and ratio $r_i$is calculated as $\Delta i/k$ and further normalized as $r_X <= r_i / \sum r$.

Thereafter,thetotalsyntheticsamplesforeach$x_i$willbe,$g_i=r_x x G$.Nowiterate from *1 to* $g_i$to generate samples the same way as done inSMOTE.




```
from imblearn.over_sampling import ADASYN

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using ADASYN
ada = ADASYN(random_state=130)
X_train_ada, y_train_ada = ada.fit_resample(X_train, y_train)

counter = Counter(y_train_ada)
print('After',counter)
```

## SMOTE+TOMEK LINK

• SMOTE+TOMEK is such a hybrid technique that aims to clean overlapping data points for each of the classes distributed in sample space.

• After the oversampling is done by SMOTE, the class clusters may be invading each other's space.

• As a result, the classifier model will be over fitting. Now, Tomek links are the opposite class paired samples that are the closest neighbors to each other.

Subsequently, the lion's share of course perceptions from these joins are evacuated because it is accepted to extend the lesson partition close the choice boundaries

• Now, to get better class clusters, Tomek links are applied to oversampled minority class samples done by SMOTE.

• Thus, instead of removing the observations only from the majority class, we generally remove both the class observations from the Tomeklinks.

**Codesnippt:**

```
from imblearn.combine import SMOTETomek

counter = Counter(y_train)
print('Before',counter)
# oversampling the train dataset using SMOTE + Tomek
smtom = SMOTETomek(random_state=139)
X_train_smtom, y_train_smtom = smtom.fit_resample(X_train, y_train)

counter = Counter(y_train_smtom)
print('After',counter)
```

## K-MEANS SMOTE

K-Means SMOTE is an oversampling method for class-imbalanced data.

It helps classification by creating minority course tests in secure and significant regions of the input space. The strategy maintains a strategic distance from the era of commotion and successfully overcomes awkward nature between and inside classes.

• Cluster the entire input space using k-means. • Distribute the number of samples to generate across clusters:

• Filter out clusters which have a high number of majority class samples.

• Assign more synthetic samples to clusters where minority class samples are sparsely distributed.

• Oversample each filtered cluster using SMOTE.

## RESULTANDDISCUSSION

Theimplementationandtesting to be donebeforeusingdatasamplingandafterusingdatasamplingtechniquesusingCKDdataset. Thispaper is done byusingdifferentdatasamplingmethodslikeSMOTE,ADASYN: Adaptive Synthetic Sampling Approach, SMOTE + Tomek Links, K-means SMOTEusingCKDdataset. Aftergettingmodifieddatasamplingdataset, to apply the different datamining algorithmsi.eDecisiontree, RandomForest,SVM and KNN to predictthepredictionofchronic kidney disease nearly stage. This paper is done by using Accuracy, precision, Sensitivity, specificity and precision as performance of evaluation measure.

The detailed analysis showed in table and graph format.

| WITHOUT USING SAMPLING TECHNIQUES | | | | |
|---|---|---|---|---|
| ALGORITHM | ACCURACY | SENSITIVITY | SPECIFICITY | PRECISION |
| DECISION TREE | 100 | 100 | 100 | 100 |
| KNN | 72.5 | 0 | 100 | NAN |
| SVM | 100 | 100 | 100 | 100 |
| RANDOM FOREST | 100 | 100 | 100 | 100 |

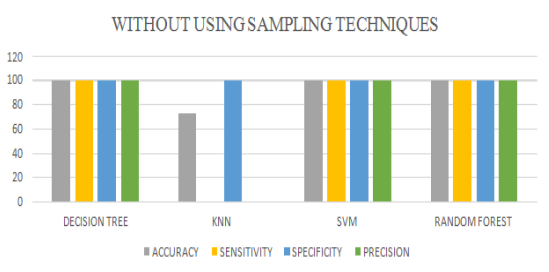



Figure Result of Without using sampling techniques

| SMOTE RESULT | | | | |
|---|---|---|---|---|
| ALGORITHM | ACCURACY | SENSITIVITY | SPECIFICITY | PRECISION |
| DECISION TREE | 100 | 100 | 100 | 100 |
| KNN | 78.125 | 22.22 | 100 | 100 |
| SVM | 100 | 100 | 100 | 100 |
| RANDOM FOREST | 100 | 100 | 100 | 100 |

Table SMOTE sampling techniques result

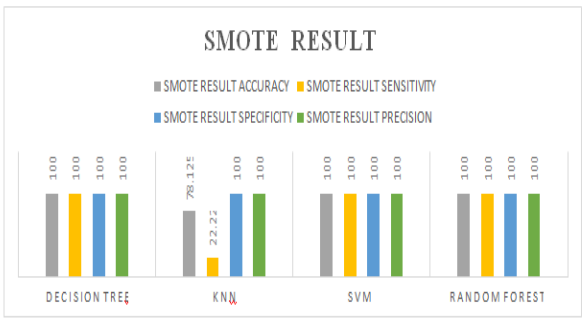



Figure SMOTE sampling techniques result

| ADASYN SAMPLING RESULT | | | | |
|---|---|---|---|---|
| ALGORITHM | ACCURACY | SENSITIVITY | SPECIFICITY | PRECISION |
| DECISION TREE | 100 | 100 | 100 | 100 |
| KNN | 50 | 20 | 63.64 | 20 |
| SVM | 100 | 100 | 100 | 100 |
| RANDOM FOREST | 100 | 100 | 100 | 100 |

Table ADSYN SMOTE sampling techniques result

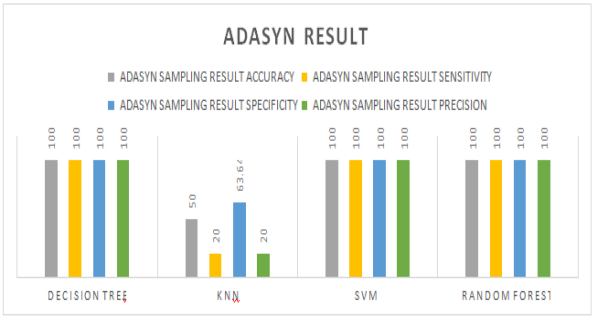



Figure ADASYN SMOTE sampling techniques result

| SMOTET-OMEK SAMPLING RESULT | | | | |
|---|---|---|---|---|
| ALGORITHM | ACCURACY | SENSITIVITY | SPECIFICITY | PRECISION |
| DECISION TREE | 68.85 | 67.74 | 70 | 70 |
| KNN | 65.57 | 61.29 | 70 | 67.86 |
| SVM | 73.77 | 80.65 | 66.67 | 71.43 |
| RANDOM FOREST | 77.05 | 77.42 | 76.67 | 77.42 |

Table SMOTE T- OMEK sampling techniques result

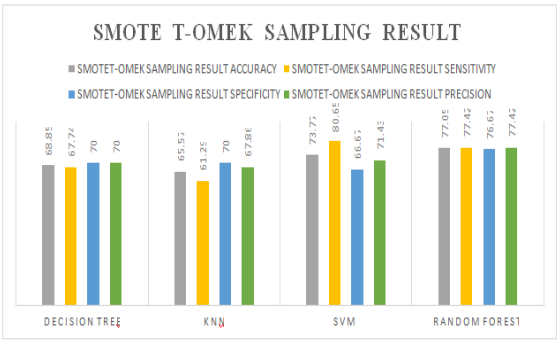



Figure SMOTE T- OMEN sampling techniques result

| K-MEANS SMOTE SAMPLING RESULT | | | | |
|---|---|---|---|---|
| ALGORITHM | ACCURACY | SENSITIVITY | SPECIFICITY | PRECISION |
| DECISION TREE | 100 | 100 | 100 | 100 |
| KNN | 75 | 20 | 100 | 100 |
| SVM | 100 | 100 | 100 | 100 |
| RANDOM FOREST | 100 | 100 | 100 | 100 |

Table K-MEANS SMOTE sampling techniques result

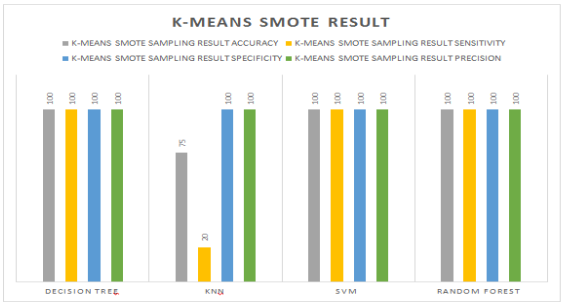



Figure K-MEANS SMOTE sampling techniques result

## CONCLUSION

Imbalanced data poses a difficult task for many classification algorithms Resembling preparing information toward a more adjusted conveyance is an successful way to combat this issue autonomously of the choice of the classifier. To solve the problem of data imbalance, a lot of solutions have proposed in recent years, and the synthetic minority oversampling technique SMOTE has been well developed. There are two strategies to improve the oversampling technique. One is to select the key minority class samples and then to generate samples with the key samples. The selection of the key samples usually has the problem of selecting incorrect samples or missing crucial samples, which could make the problem more complex. And the performance of the classification cannot be effectively improved. The other is to synthesize minority course tests firstly, and after that erase the tests which are effortlessly misclassified by utilizing the information cleaning strategy. Implementation and compare of SMOTE, ADASYN: Adaptive Synthetic Sampling Approach, SMOTE + Tomek Links, K- means SMOTE method belongs to the second strategy. The compared method achieves these properties by clustering the data using k- means, allowing to focus data generation on crucial areas of the input space. A tall proportion of minority perceptions is utilized as a marker that a cluster could be a secure region. Oversampling as it were secure clusters empowers k-means Destroyed to dodge clamor era. Besides, the normal remove among a cluster's minority tests is utilized to find inadequate ranges. Inadequate minority clusters are doled out more manufactured tests, which lighten within-class lopsidedness. Finally, over fitting is discouraged by generating genuinely new observations using SMOTE rather than replicating existing ones. This paper is implement by SMOTE, ADASYN: Adaptive Synthetic Sampling Approach, SMOTE + Tomek Links, K- means SMOTE and tested with without using data sampling techniques. After data sampling, data undergone feature selection techniques such as the correlation method then final train and test with different machine learning algorithms such as e Decision tree, Random Forest, SVM and KNN. According to our result DT, SVM and RF performance extremely good for accuracy not in terms of Specificity value. Because all these algorithms gave 100% value. It's not good for good performance metrics. According to criteria of selecting good Classification algorithms along with good data sampling techniques, This is found by KNN for good algorithm and SMOTE data sampling techniques very good for CKD dataset.

## FUTUREENHANCEMENT

May subsequently center on applying k-means Destroyed to different other real-world issues. Also, finding ideal values of k and other hyper parameters is however to be guided by rules of thumb, which might be deducted from advance examinations of the relationship between ideal hyper parameters for a given dataset and the dataset's properties.

## REFERENCES

1.GunarathneW.H.S.D,PereraK.D.M,KahandawaarachchiK .A.D.C.P,"Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)", 2017 IEEE 17th International Conference on Bio informatics and Bioengineering.

2.S.Ramya,Dr.N.Radha," Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms, "Proc. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January2016.

3. S. Dilli Arasuand Dr. R. Thirumalaiselvi, "Review of Chronic Kidney Disease based on Data Mining Techniques", International Journal of Applied Engineering Research ISSN0973 -4562 Volume12, Number23 (2017)pp.13498-13505

4. L.Rubini, "Early stage of chronic kidney disease UCI machine learning repository," 2015. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/ChronicKidneyDiseas.

5. S. A. Shinde and P. R. Rajeswari, “Intelligent health risk prediction system susing machine learning: a review,”IJET,vol.7,no.3,pp.1019–1023,2018.

6. Himanshu Sharma,M A Rizvi,”Prediction of Heart Disease using Machine Learning Algorithms: A Survey ” , International Journal on Recent and Innovation Trendsin Computing and Communication ISSN:2321-8169, Volume: 5 Issue: 8

7. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease andSelecting Important Predictive Attributes," Proc. IEEE International Conference on Health care Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.