A proportional test study on Machine Learning Algorithms

Dr. Rajesh S. Walse[1], Gajanan N. Narnaware[2]

[1] Assistant Professor (Computer Science) & Head, Department of Dairy Business Management, College of Dairy Technology, Warud (Pusad), M.A.F.S.U, Nagpur, Maharashtra State, India (Corresponding Author)

[2] Assistant Professor &, Department of Dairy Business Management, College of Dairy Technology, Warud (Pusad), M.A.F.S.U, Nagpur, Maharashtra State, India

Abstract

The researcher employed a classification approach to conduct a comparative analysis of data from chronic kidney patients. Two methods were used: Supervised with SMO-SVM and Unsupervised Hierarchical Clustering. Our proposed model was developed by leveraging chronic kidney disease data consisting of 25 attributes and 400 instances, including a class label. Initially, the Supervised-SMO-SVM approach was employed to classify a specific attribute, "htn Vs Class," using Cross-validation ranging from 4 to 15 folds with a 70% data split. Logistic calibration and Polynomial kernel were utilized.

The results indicated that the ROC area value for the CV fold of 4 was 0.794, with the same weighted average value for both CKD and non-CKD classes. The accuracy of correctly classified instances was 74.25%. The Confusion matrix results were consistent across Cross-Validation Folds from 5 to 15. Moreover, the ROC area value for the CCI remained consistent across CV folds from 5 to 15.

Furthermore, the researcher performed a comparison with the unsupervised hierarchical clustering algorithm using all 24 attributes and 400 instances. This analysis also yielded a highly accurate and optimal result. Notably, the hierarchical algorithm model exhibited the highest accuracy among all predictive models.

In summary, the researcher utilized classification methods to analyze chronic kidney patient data, both through supervised and unsupervised techniques. The proposed model's performance was evaluated and compared, with the hierarchical clustering algorithm showing the most accurate predictions.

Keywords: ML Algorithm**,** WEKA, CKD.

Introduction

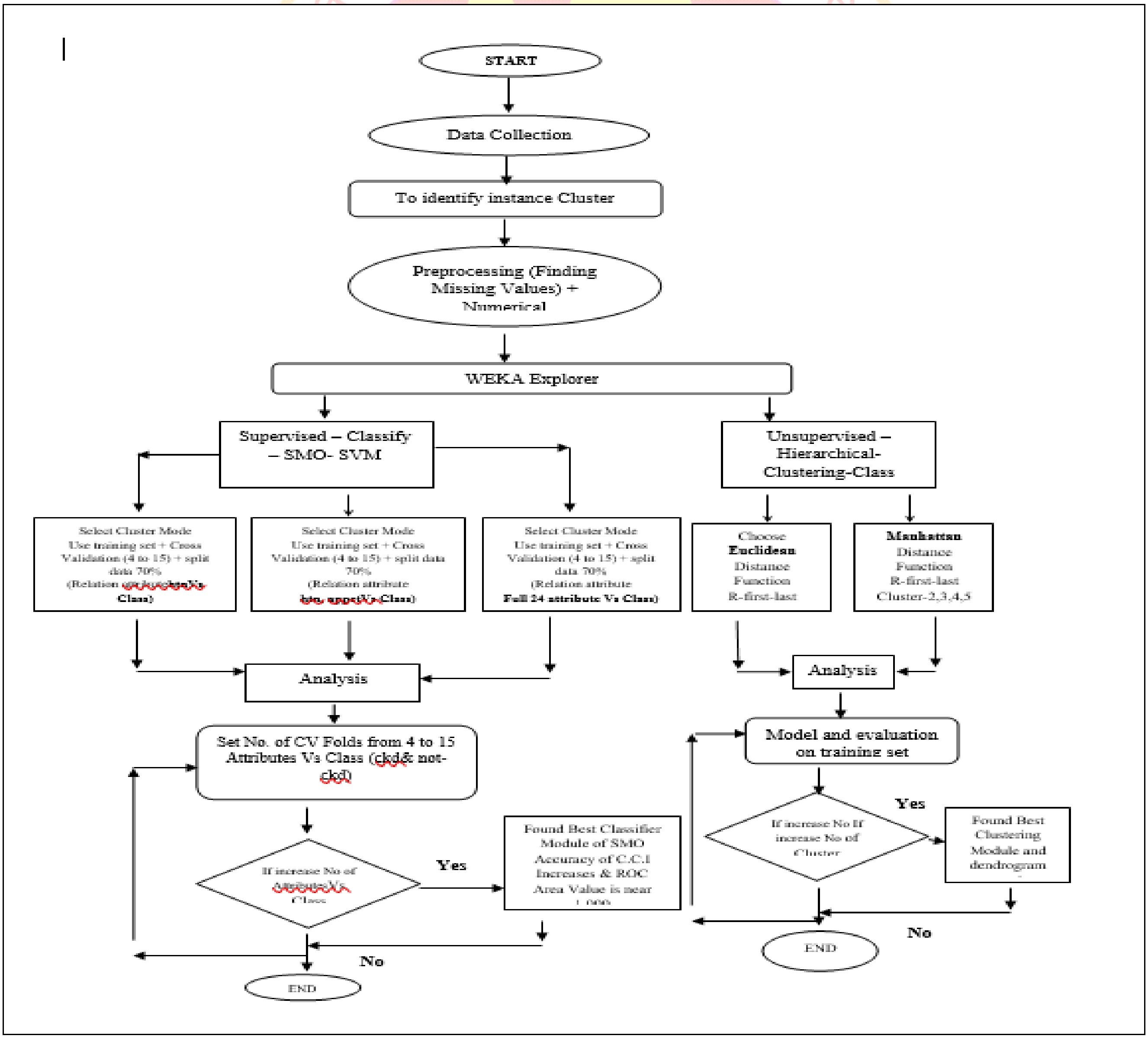
In recent times, the prevalence of chronic kidney disease (CKD) among patients in India has been steadily rising due to dietary habits and various health concerns. Over the past decade, the number of CKD patients has shown a significant increase, as documented by the Indian Journal of Nephrology and others [12]. Consequently, future research endeavors of this nature are poised to offer valuable insights to medical practitioners and the healthcare industry. Such studies could aid in predicting the likelihood of individuals developing CKD or not based on their health parameters. By doing so, these efforts aim to curb the escalating CKD patient figures and mitigate the further deterioration of kidney health.

Literature Review:

In this study, various algorithms were examined, including the NB Classifier, J48, and Random Forest Decision Tree. Data cleaning within the domain of Data Mining was employed to eliminate noise and inconsistent data, integrating multiple types of data for enhanced accuracy. The assessment of data involved the utilization of secondary data sourced from repositories such as the UCI Machine Learning Repository [14] and Jnephrol [13].

The United States has witnessed a substantial 30% rise in the prevalence of CKD over the past decade due to increased life expectancy and the growing incidence of lifestyle-related diseases [13]. In contrast, India lacks comprehensive longitudinal studies and possesses limited data regarding CKD incidence. Presently, the deteriorating quality of life and dietary habits are adversely affecting health, contributing to a surge in kidney-related disorders in India. While kidney health used to be influenced by dietary patterns from four decades ago or older, today, kidney disease transcends factors like diabetes or hypertension; it stems from diverse causes, including chemical-laden cereals, vegetables, and fruits which constitute our daily diet. The cumulative impact of these factors gradually diminishes kidney function, culminating in kidney failure. This trend is highlighted in Jnephrol [13].

Due to the absence of comprehensive longitudinal CKD studies and limited data in India, this study endeavors to address the issue by employing Naïve Bayes, decision tree J48, and random forest algorithms within the realm of machine learning. Our research aims to ascertain whether these algorithms can effectively predict acute kidney disease or its future development.



Methodology:

Supervised – Classifier – SMO Function

Step-1

Step-1 (htnVs Class) – One Attribute

Supervised- classify – SMO- SVM- 2 Attributes – htnVs Class

Applying Cross validation from 4 to 15 + 70% split

Calibrator: Logistic and Kernel: Polynomial

=== Classifier model (full training set) ===, SMO , Kernel used:Linear Kernel: K(x,y) = <x,y>

Classifier for classes: ckd, notckdBinarySMO, Machine linear: showing attribute weights, not support vectors.

2 * (normalized) htn=no, -1 Comparative Study supervised of SMO, calibrator: logistic and kernel: polynomical using cross validation fold 4 to 15 with 70% of split classifier

Table: Detailed accuracy by Class for htnVs class

Cross Validation Fold	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
4	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
5	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
6	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
7	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
8	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
9	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
10	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
11	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
12	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
13	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
14	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
15	CKD	0.588	0.000	1.000	0.588	0.741	0.590	0.794	0.845
	Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
	Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751

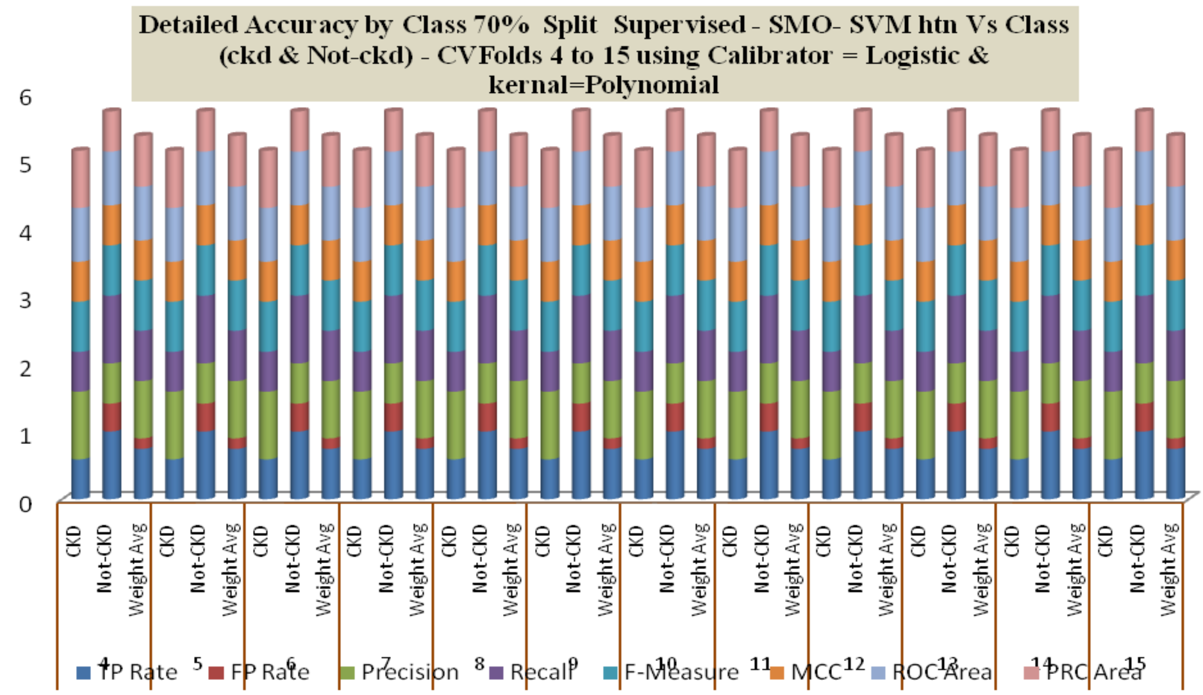


Figure: detailed accuracy by class Supervised SMo-SVM attributehtnVs class with CV folds 4 to 15 using Calibrator and Kernel function.

Table : Summary of Classifier model (full training set) for htnVs Class
1. Test Mode: split 70% , 2. Total Number of Instances=400

Sr. No.	**Particulars**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	Correctly Classified Instances	74.25	74.25	74.25	74.25	74.25	74.25	74.25	74.25	74.25	74.25	74.25	74.25
2	Incorrectly Classified Instances	25.75	25.75	25.75	25.75	25.75	25.75	25.75	25.75	25.75	25.75	25.75	25.75
3	Kappa statistic	0.517	0.517	0.517	0.517	0.517	0.517	0.517	0.517	0.517	0.517	0.517	0.517
4	Mean absolute error	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575	0.2575
5	Root mean squared error	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074	0.5074
6	Relative absolute error	54.90	54.9106	54.911	54.9098	54.91	54.9108	54.9131	54.9105	54.9107	54.9101	54.9114	54.9128
7	Root relative squared error	104.81	104.817	104.8165	104.8123	104.8121	104.8128	104.8172	104.8114	104.8113	104.8098	104.8122	104.8158

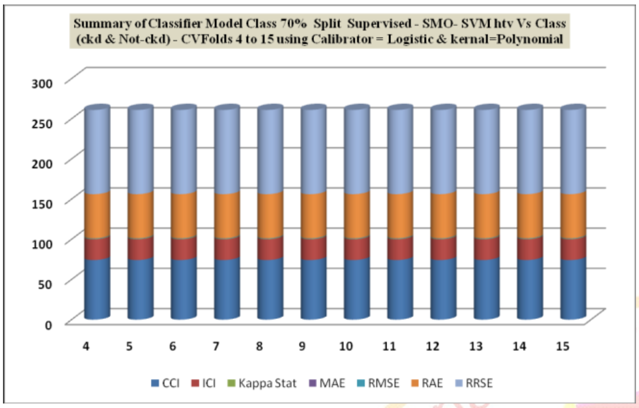


Figure: Summary of classifier model class SMo-SVM attribute htnVs class with CV folds 4 to 15 using Calibrator and Kernel function.

Results:

====== Confusion matrix =======

The above result for all Cross Validation Folds from CVF = 4 to CVF=15 is the same

Table 3:Confusion matrix (full training set) for htnVs Class

Sr. No.	CVF	Predicted (a)	Predicted (b)	< - Classified as
1	4	147	103	| a = ckd
		0	150	| b = not-ckd
2	5	147	103	| a = ckd
		0	150	| b = not-ckd
3	6	147	103	| a = ckd
		0	150	| b = not-ckd
4	7	147	103	| a = ckd
		0	150	| b = not-ckd
5	8	147	103	| a = ckd
		0	150	| b = not-ckd
6	9	147	103	| a = ckd
		0	150	| b = not-ckd
7	10	147	103	| a = ckd
		0	150	| b = not-ckd
8	11	147	103	| a = ckd
		0	150	| b = not-ckd
9	12	147	103	| a = ckd
		0	150	| b = not-ckd
10	13	147	103	| a = ckd
		0	150	| b = not-ckd
11	14	147	103	| a = ckd
		0	150	| b = not-ckd
12	15	147	103	| a = ckd
		0	150	| b = not-ckd

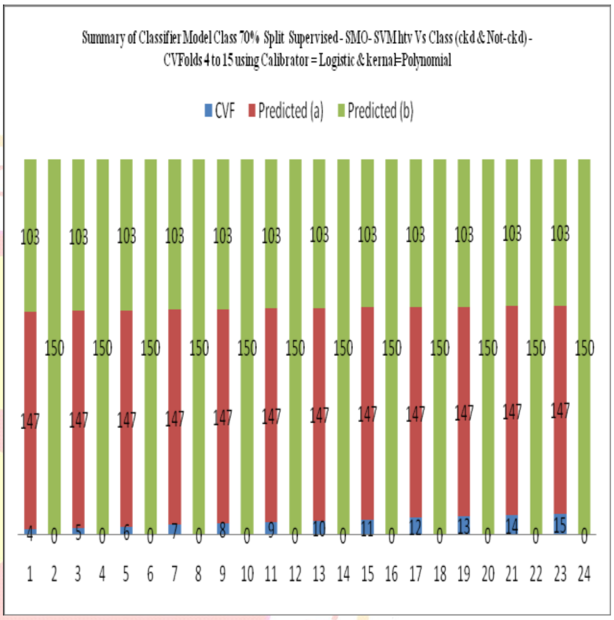


Figure: 9Confusion matrix SMO-SVM attribute htnVs class with CV folds 4 to 15 using Calibrator and Kernel function.

STEP 1:

(Manhattan Distance) = 2 Cluster = Full Attributes

=== Run information ===

Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.ManhattanDistance -R first-last"

Relation: Chronic_Kidney_Disease_(RS Walse)-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 400

Attributes: 25

Time taken to build model (full training data) : 0.72 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 399 (100%)

1 1 (0%)

Class attribute: class

Classes to Clusters:

0 1 <-- assigned to cluster

249 1 | ckd
150 0 | notckd

Cluster 0 <-- ckd
Cluster 1 <-- No class

Incorrectly clustered instances : 151.0 37.75 %

STEP 2:

(Manhattan Distance) = 3 Cluster = Full Attributes

Time taken to build model (full training data) : 0.71 seconds

=== Model and evaluation on training set ===
Clustered Instances

0 398 (100%)
1 1 (0%)
2 1 (0%)

Class attribute: class
Classes to Clusters:
0 1 2 <-- assigned to cluster
248 1 1 | ckd
150 0 0 | notckd

Cluster 0 <-- ckd
Cluster 1 <-- No class
Cluster 2 <-- No class
Incorrectly clustered instances : 152.0 38 %

STEP 3:

(Manhattan Distance) = 4 Cluster = Full Attributes

Time taken to build model (full training data) : 0.73 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 397 (99%)
1 1 (0%)
2 1 (0%)
3 1 (0%)

Class attribute: class
Classes to Clusters:
0 1 2 3 <-- assigned to cluster
247 1 1 1 | ckd
150 0 0 0 | notckd
Cluster 0 <-- ckd
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- No class

STEP 4:

(Manhattan Distance) = 5 Cluster = Full Attributes

=== Model and evaluation on training set ===
Clustered Instances
0 396 (99%)
1 1 (0%)
2 1 (0%)
3 1 (0%)
4 1 (0%)
Class attribute: class
Classes to Clusters:
0 1 2 3 4 <-- assigned to cluster
246 1 1 1 1 | ckd
150 0 0 0 0 | notckd
Cluster 0 <-- ckd
Cluster 1 <-- No class
Cluster 2 <-- No class
Cluster 3 <-- No class
Cluster 4 <-- No class
Incorrectly clustered instances : 154.0 38.5 %

Discussion :

{a} A comparative analysis was conducted using the Supervised SMO-SVM Classifier across different attribute configurations: one attribute, two attributes, and the full set of attributes. The data was split with a 70% partitioning. The analysis incorporated the use of a Logistic Calibrator and a Polynomial Kernel function.

The study involved processing a kidney dataset containing 400 instances and 25 attributes (columns). Each attribute was individually examined to determine its type, whether nominal or not. The analysis also included identifying the presence of missing values and the number of distinct values for each attribute. Attributes categorized as nominal were denoted by the label "Nom-" preceding their name.
For numeric data, descriptive statistics were utilized to provide an overview of the dataset. Additionally, when dealing with qualitative data, it was treated as an attribute class. This information was presented

through label counts and their corresponding weights, represented as true/false or yes/no.

Table 10: Correctness by Class values

	No of Attributes	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
*CVF=4	01 (htnVs Class)	CKD	0.588	0.000	1.000	0.588	0.741	0.590	**0.794**	0.845
		Not-CKD	1.000	0.412	0.593	1.000	0.744	0.590	0.794	0.593
		Weight Avg	0.743	0.155	0.847	0.743	0.742	0.590	0.794	0.751
** CVF=4	02 (htn, appetVs Class)	CKD	0.737	0.000	1.000	0.737	0.848	0.712	**0.868**	0.904
		Not-CKD	1.000	0.263	0.688	1.000	0.815	0.172	0.868	0.688
		Weight Avg	0.833	0.096	0.886	0.833	0.836	0.172	0.868	0.824
*** CVF=5	24 (Full Vs Class)	CKD	0.980	0.000	1.000	0.980	0.990	0.974	**0.990**	0.992
		Not-CKD	1.000	0.020	0.968	1.000	0.984	0.974	0.990	0.968
		Weight Avg	0.988	0.008	0.988	0.988	0.988	0.974	0.990	0.983

* Found max and same accuracy for 5 to 15 CVF
** Found max and same accuracy for 5 to 15 CV
*** 5,6,8,9,10,14,15 found max and same accuracy CVF

The table presents accuracy based on class values for the Supervised SMO-SVM classifier with different attribute configurations: attributes 2, 3, and the full set. Cross-validation was performed with a range of folds from 4 to 15, using a 70% data split. The analysis employed a Logistic Calibrator and a Polynomial Kernel function. The results of this comparative study revealed accurate prediction values.

The research demonstrates that as the number of cross-validation folds increases alongside the corresponding increase in attributes, the accuracy of class value correctness improves. This trend is evident in the ROC Area values: ROC Area = 0.794 for one attribute with CVF=4, ROC Area = 0.868 for two attributes with CVF=4, and ROC Area = 0.990 for five attributes with CVF=5.

Consequently, the study identifies the best and novel predictive module utilizing the SMO-SVM classifier class module, driven by the choice of Calibrator and Kernel function.

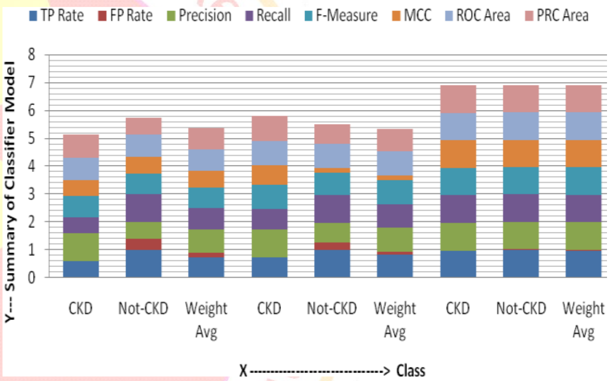


Figure: Comparative study of Correctness by class values of SMO-SVM- Classifier with 1,2 and full attributes with 70% and CVF from 4 to 15.

Table : Summary of Classifier model (Train set data)

Sr. No.	Particulars	01 (htnVs Class)	02 (htn, appetVs Class)	24 (Full Vs Class)
		CV Folds	CV Folds	CV Folds
		4 Found max and same accuracy for 5 to 15 CVF	4 Found max and same accuracy for 5 to 15 CVF	5,6,8,9,10,14,15 found max and same accuracy CVF
1	Test mode: 70% train, 30% test	70.0% train	70.0% train	70.0% train
2	Correctly Classified Instances	**74.25**	**83.33**	**98.75**
3	Incorrectly Classified Instances	25.75	16.66	1.25

4	Kappa statistic	0.517	0.6725	0.9735
5	Mean absolute error	0.2575	0.1667	0.0125
6	Root mean squared error	0.5074	0.4082	0.1118
7	Relative absolute error	54.90	35.6241	2.6656
8	Root relative squared error	104.81	84.6877	23.094
9	Total Number of Instances	400	400	400

(CVF= Cross Validation Folds)

Table: presents a summary of the classifier model using training set data for the SMO-classifier. Different attribute combinations - one, two, and the complete set of 25 attributes - were applied to assess the outcome, yielding valuable and accurate results. These findings contribute to enhancing the applicability of our research, aiding future researchers in predictive endeavors and effectively utilizing the SMO-SVM classifier algorithm.

Moreover, Table 11 showcases the progressive improvement in accuracy based on various parameters, such as Cross Validation Folds, spanning from 4 to 15. Notably, the accuracy surpasses earlier results, with Correctly Classified Instances reaching 74.25%, 83.33%, and 98.75%, respectively. This underscores the discovery of an optimal and innovative summary of the classified module, characterized by the highest accuracy of Correctly Classified

Instances.

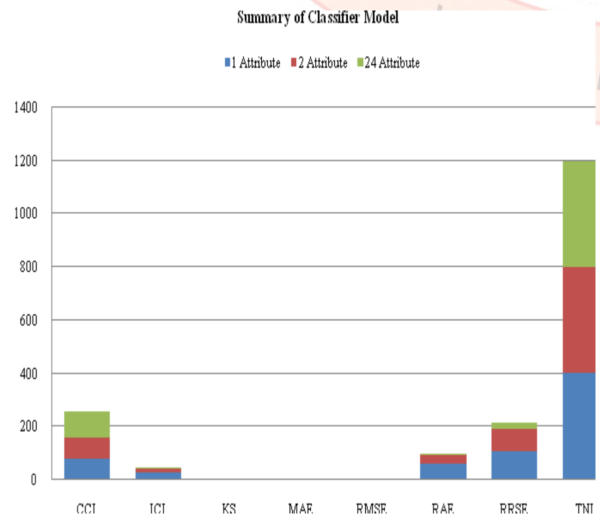


Figure: Comparative study of Summary of Classifier Module of SMO-SVM- Classifier with 1,2 and full attributes with 70% and CVF from 4 to 15.

Table : Comparative statement of Confusion Matrix SMO-SVM- Classifier with one, two & full attributes with 70% split using Calibrator: Logistic and Kernel= Polynomial function

Sr. No.	**No of Attributes**	**CVF**	**Predicted (a)**	**Predicted (b)**	**<-- Classified as**
1	One Attribute htnvs Class	4-15	147	103	| a = ckd
			0	150	| b = not-ckd
2	Two Attributes Htn, appetVs Class	4	147	103	| a = ckd
			0	150	| b = not-ckd
		5-15	172	78	| a = ckd
			0	150	| b = not-ckd
3	Full Attributes 24 Vs Class	4,7,11,12,13	244	6	| a = ckd
			0	150	| b = not-ckd
		5,6,8,9,10,14,15	245	5	| a = ckd
			0	150	| b = not-ckd

(CVF = Cross Validation Folds)

Table 12 displays a Comparative Analysis of the Confusion Matrix for the SMO-SVM Classifier, examining one, two, and the full set of attributes. This assessment was performed with a 70% data split and involved the utilization of a Logistic Calibrator and a Polynomial Kernel function. The table presents threshold Curve and Cost-Benefit Curve for the "ckd" and "Not-ckd" classes, represented as percentages within the confusion matrix and cost matrix. Additionally, it includes Confusion Matrix data for predicted classes "a" and "b," along with corresponding cost matrix values, random values, and gain values.

Notably, the research findings indicate that when employing the full set of attributes, the Confusion Matrix is minimized, resulting in accurate predictions of class values with a reduced number of cluster iterations.

{b} A comparative analysis was conducted on the Unsupervised Hierarchical Clustering algorithm, utilizing both Euclidean and Manhattan distance functions, across different numbers of clusters: 2, 3, 4, and 5.

A comparative assessment of the Hierarchical Clustering algorithm was performed, evaluating the accuracy of results using both Euclidean and Manhattan distance functions.

❖ **Euclidean Distance:**

=== Run information ===

Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Relation:Chronic_Kidney_Disease_(RS Walse)-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

❖ **Manhattan Distance:**

=== Run information ===

Scheme:weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.ManhattanDistance -R first-last"

Relation:Chronic_Kidney_Disease_(RS Walse)-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

=== Clustering model (full training set) ===

Table : Comparative statement of Unsupervised Hierarchical Clustering algorithm using Euclidean and Manhattan Distance Function of full attributes with 70% split. (Cluster 2,3,4& 5)

Sr. No.	No. of Cluster	Euclidean Distance		Manhattan Distance	
		Clustered Instances	Classes to Clusters	Clustered Instances	Classes to Clusters
1	2	0 399(100) % 1 1(0%)	249 1 | ckd 150 0 | notckd	0 399(100) % 1 1(0%)	249 1 | ckd 150 0 | notckd
2	3	0 398 (100%) 1 1 (0%) 2 1 (0%)	248 1 1 | ckd 150 0 0 | notckd	0 398 (100%) 1 1 (0%) 2 1 (0%)	248 1 1 | ckd 150 0 0 | notckd
3	4	0397 (99%) 1 1 (0%) 2 1 (0%) 3 1 (0%)	247 1 1 1 | ckd 150 0 0 0 | notckd	0 397 (99%) 1 1 (0%) 2 1 (0%) 3 1 (0%)	247 1 1 1 | ckd 150 0 0 0 | notckd
4	5	0 396 (99%) 1 1 (0%) 2 1 (0%) 3 1 (0%) 4 1 (0%)	246 1 1 1 1 | ckd 150 0 0 0 0 | notckd	0 396 (99%) 1 1 (0%) 2 1 (0%) 3 1 (0%) 4 1 (0%)	246 1 1 1 1 | ckd 150 0 0 0 0 | notckd

Table: This table illustrates a comparison conducted by the researcher using the Unsupervised Hierarchical Clustering algorithm on Chronic Kidney Disease data. The analysis employed both Euclidean and Manhattan Distance clustering functions, supported by a 70% data split. The properties were configured for the Number of Clusters, ranging from 2, 3, 4, to 5, using the respective Euclidean and Manhattan distance functions.

The findings of this research revealed a noteworthy observation: as the Number of Clusters increased, accuracy improved when utilizing the full set of attributes from the Chronic Kidney Disease data. Moreover, a significant comparative outcome emerged, indicating that the output results for both Euclidean and Manhattan distance functions were consistent across different Number of Clusters settings (2, 3, 4, and 5).

Conclusion:

The CKD dataset was subjected to analysis and prediction using data mining techniques, specifically supervised classifiers such as SMO-SVM, and an unsupervised clustering algorithm, namely Hierarchical clustering. These methodologies were assessed and compared using Weka tools. The final results demonstrated that both the Classification and Clustering algorithms led to the discovery of a novel and optimal module. This module served as a highly accurate classifier, achieving 98.75% Correctly Classified Instances accuracy. This outcome outperformed previous results obtained by employing one or two attributes, and it involved the use of Calibrator: Logistic and Kernel: Polynomial functions. The comparison was conducted using Cross Validation Folds (CVF) ranging from 5 to 15, with consistent results seen in the confusion matrix.

The study also identified a superior predictive model for Supervised SMO in WEKA. This model utilized the three-test approach involving logistic calibration, polynomial kernel, and Cross Validation Folds from 4 to 15, with a 70% data split.

Importantly, it was observed that increasing the number of attributes positively correlated with enhanced accuracy in terms of Correctly Classified Instances (CCI), ROC Area value, and Confusion matrix value.

Similarly, research confirmed the prediction capability of the Unsupervised Hierarchical Clustering algorithm, especially when employing the full set of attributes. Increasing the number of Clusters from 2, 3, 4, to 5, utilizing both Euclidean and Manhattan distance functions, yielded more accurate and valuable clusters.

The chosen methodology illuminated a practical process. Attributes like RBC count, HP, Diabetes Mellitus, CAD, Appetite, Pedal Edema, Anemia, among others, were measured in the research work. Moving forward, this type of study holds potential to aid doctors and the medical industry in predicting CKD and non-CKD patients based on various health parameters. Such predictions aim to curtail the growth rate of CKD cases and mitigate kidney damage. Data mining proves instrumental in foreseeing potential kidney-related health issues.

This study delved into three key algorithms, focusing on chronic kidney disease analysis via the SMO-SVM classifier algorithm with logistic calibration and polynomial kernel. The overarching goal is to provide insight into current kidney health and the likelihood of future kidney disease. By doing so, individuals currently affected by kidney issues can better understand the causes, while those unaffected can ascertain their risk, potentially saving costs on additional tests.

Acknowledgement

The authors express their gratitude to the UCI ML repository for furnishing a comprehensive and essential database. Additionally, special acknowledgment is extended to WEKA for offering a dependable tool that facilitates the extraction and analysis of knowledge from this database.

References

1. Arasu, S. Dilli, and R. Thirumalaiselvi. "A novel imputation method for effective prediction of coronary Kidney disease." In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), pp. 127-136. IEEE, 2017.
2. Ariff, M. H., I. Ismarani, and N. Shamsuddin. "RFID based systematic livestock health management system." In 2014 IEEE Conference on Systems, Process and Control (ICSPC 2014), pp. 111-116. IEEE, 2014.
3. Bharara, Sanyam, A. SaiSabitha, and AbhayBansal. "A review on knowledge extraction for business operations using data mining." In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 512-518. IEEE, 2017.
4. Chuan, Zou, Tang Ying, Bai Li, ZengYuqun, and Lu Fuhua. "Application of clustering analysis to explore syndrome evolution law of peritoneal dialysis patients." In 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 23-26. IEEE, 2013.
5. Due ThanhAnhLuong, Dept of the computer. Sci&Engi, uni, at Buffalo Buffalo, NY, USAA K- Means Approach to Clustering disease Progressions IEEE Keywords, Sep 2017
6. Güllüoğlu, SabriSerkan. "Segmenting customers with data mining techniques." In 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), pp. 154-159. IEEE, 2015.
7. Jinyin, Chen, et al. "A novel cluster center fast determination clustering algorithm." Applied Soft Computing 57 (2017): 539-555.
8. Khanna, Umesh. "The economics of dialysis in India." Indian journal of nephrology 19, no. 1 (2009): 1.
9. Kunwar, Veenita, et al. "Chronic Kidney Disease analysis using data mining classification techniques." 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). IEEE, 2016.
10. Narander Kumar, SabitaKLhatri, Department of Computer, 3rd IEEE International Conference on Computational Intelligence

VOL- X ISSUE- VI JUNE 2023 PEER REVIEW e-JOURNAL IMPACT FACTOR 7.367 ISSN 2349-638x

and Communication Technology (IEEE-CICT 2017) “ Implementing WEKA for medical data classification and early disease prediction 978-1-50, 2017

11. Uboltham, Issariya, NakornthipPrompoon, and Wirichada Pan-Ngum. "AKIHelper: Acute kidney injury diagnostic tool using KDIGO guideline approach." In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1-6. IEEE, 2016.
12. Khanna, Umesh. "The economics of dialysis in India." Indian journal of nephrology 19, no. 1 (2009): 1.
13. Varma, P. P. "Prevalence of chronic kidney disease in India-Where are we heading?." Indian journal of nephrology 25, no. 3 (2015): 133.
14. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease