

COMPARATIVE STUDY ON CKD USING LEARNING ALGORITHMS

Mounika.B¹, M. Nirupama Bhat²

¹PG Student, Department of computer engineering, Vignan's deemed to be university, mounikabingi2@gmail.com

²Professor, Department of computer engineering, Vignan's deemed to be university, nirupamakonda@gmail.com

Abstract— Chronic Kidney Disease (CKD) is one of the common disease that conventionally triggered to the diabetic patients. In every country middle aged people were mostly affected due to stress and hypertension. The objective of this project to find the two feature reduction algorithms to differentiate that which one is a better algorithm than one another. The algorithms that are used to reduce the features are Correlation and LASSO (Least Absolute Shrinkage and Selection Operator). After the features reduced some of the learning algorithms like SVM (Support Vector Machine), Naive Bayes, KNN and Logistic Regression were applied, some of the performance metrics also calculated. This experiment is performed on the diabetic patients records. The LASSO feature reduction algorithm gave better accuracy than a correlation algorithm.

Keywords— SVM, KNN, Naive Bayes, Correlation, LASSO, CKD, Performance metrics.

I. INTRODUCTION

Machine learning in medicine has as of late stood out as truly newsworthy. Google has developed a machine learning algorithm to perceive perilous tumors on mammograms. Stanford is utilizing a profound learning algorithm to distinguish skin disease. An ongoing JAMA article detailed the aftereffects of a profound machine-learning calculation that could analyze diabetic retinopathy in retinal pictures. Plainly machine learning puts another bolt in the bunch of clinical basic leadership. Calculations can furnish quick advantage to disciplines with forms that are reproducible or institutionalized. In India most regular ailment for each one is diabetes this activated because of the pressure and undesirable way of life of a patient. There are 3 kinds of diabetes, they are type-1, type-2 and type-3 happened. Type - 1 is a chronic condition in which the pancreas makes essentially zero insulin, manifestations are high glucose levels and successive urination. Type-2 an unending condition that impacts the way the body shapes glucose. Type-3 is additionally named as Alzheimer's ailment bringing about an insulin obstruction in the mind, the most widely recognized early side effect is trouble in recalling ongoing occasions. United Nations Children's Emergency Fund data show that 28% of children are <2.5 kg at birth. Hypovitaminosis A and other nutritional issues during pregnancy may cause smaller kidney volume at birth and a lower eGFR [1]. Consanguinity and genetic inbreeding increase risk of congenital anomalies of the kidney and urinary tract and obstructive or reflux nephropathy[2].

Chronic kidney disease (CKD) is a kind of kidney infection in which there is a steady loss of kidneys work over a time of months or years. We can't distinguish any reactions opportune. This infirmity can be perceived by the GFR(Glomerular Filtration Rate), Urine tests and Blood tests. Medications and lifestyle changes may help to slow the disease progress if CKD diagnosed early [3]. In India, half of the patients are in the propelled phase of kidney infection i.e., kidney disappointment. In this half, the majority of the patients are beneath the destitution line. The CKD may happen on account of pressure and hypertension, there two conditions are having hazard variables to endure the kidney disappointment. For each patient, the decision to commence that patient is suffering from CKD or not. The choice turns out to be progressively testing in more seasoned patients, not on account of age, it blocks Kidney Failure, but rather on the grounds that contemplations of co morbidity, lessened overall lifespan and the impact on the health itself on quality of life, become more complex with increasing age. In this we performed a comparative study between the feature reduction algorithms, the algorithms that are used in this paper are Correlation and Least Absolute Shrinkage and Selection Operator (LASSO). After the feature reduction applied learning algorithms to compare, the better accuracy to recognize the best fit reduction algorithm on the patient database.

The feature reduction and the comparison of the algorithms can be statistically viewed using the visualization techniques in the R platform. R is a programming language to calculate we have some of the in-built packages to visualize the results statistically Google also launched many applications by using the machine learning algorithms not only in the software sector in the medical sector also machine learning playing a key role for the classification of patient disease in a huge data set Reddy, Manoj and Jhon Cho[4] has told in their research that diabetes is one of the mostly effected and dangerous disease. CKD can be detected in the early stages and can help at-risk patients from a complete kidney failure by simple tests that involve measuring blood pressure, serum creatinine and urine albumin. The person is effected or not effected can be classified by using patients conditions with classification algorithms like SVM(Support Vector Machine), Decision Tree and Ada Boost are used to compare their performance. Some of the unsupervised algorithms are also used like clustering, K-mean algorithm and to reduce the dimensionality PCA (Principal Component Analysis) is used.

The diabetic patients will consult the doctors every month to determine the health conditions like BP, Insulin levels and carbohydrate levels. Bajestani, Narges Shafaei [5] has proved that by using GFR rate, we can identify the stages of the CKD based on the patient's stage we can decrease the number of check ups for a patient and also future occurrence of the disease of a CKD. They used the fuzzy logic to identify the best output for the data set and fuzzy regression model. Fuzzy logic is an algorithm that yields the perfect turnout for a data. Time Series Analysis is the algorithm used to identify the future predictions. Basically, some disease does not have the permanent solution because of this by changing our life style we can postpone the advanced stage of the disease. In this paper they clearly mention that we can delay the sever failure of a kidney.

II. METHODOLOGY

Diabetes Mellitus(DM) is the one of the major illness. There are three types of diseases. Type-1 may happen because of uncalled for the discharge of an insulin. This may come about "high glucose levels" in the body. Type-2 it is because of the relative absence of insulin. It is a long-term complexity which may prompt the heart stroke, diabetes retinopathy (visual impairment) and diabetes nephropathy (kidney disappointment). Type-3 diabetes is a proposed term for Alzheimer's ailment bringing about an insulin opposition in the cerebrum. Of all these, Diabetes Nephropathy is the significant medical issue happening on the planet. The last phase of kidney disease may cause the kidney disappointment, it might prompt the kidney transplantation or passing of a man. As indicated by the World Health Organization(WHO), 61.3% of Indian individuals were experiencing DM. Real level of the general population were affected with Chronic Kidney Disease i.e., Kidney Failure. The capacity of a kidney is to decontaminate the blood and pee made out of the squanders. On account of kidney disappointment, kidney does not perform filtration process. The proposed strategy is to foresee climate the individual is affected is not affected with CKD. It can be ordered utilizing the grouping algorithms.

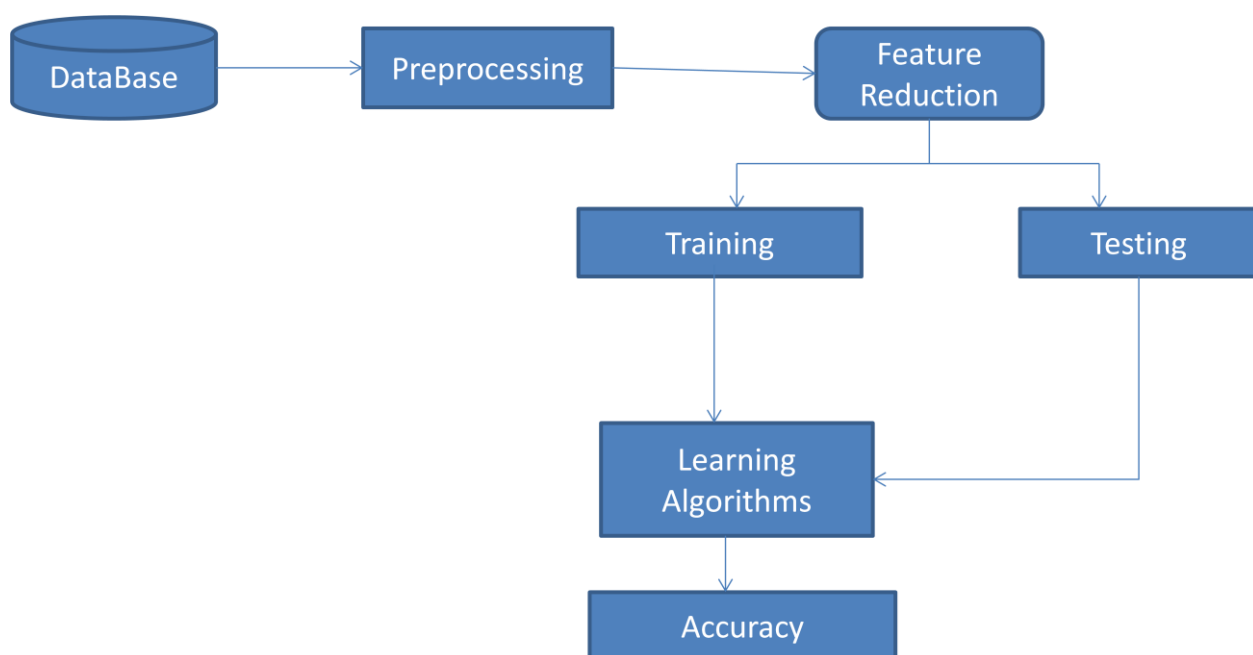


Fig.1.Methodology

A. DATA BASE:

A total 34 features are used in the data set[6].

ID	Diabetes	Height	Fam.PVD
Age	CHF	Poor Vision	Obese
Hyper Tension	Fam.HT	Activity	Anemia
CVD	Stroke	Waist	Insured
SBP	DBP	Education	BMI
Weight	Fam. Diabetes	Racegrp	Female
PVD	Smoker	Fam.CVD	Unmarried
Care Source	HDL	Total. Chol	Dyslipdemia
Income	LDL		

In the informational index 9000 patients records, this is taken based on the patient's conditions. In this, all highlights just 26 factors are the exceptionally co-related factors and non-zero-difference esteems.

B. Pre-Processing :

The patient's informational indexes were gathered from the repository. The information may contain the missing qualities. It can be preprocessed by the preprocessing techniques. By using the relationship calculation very associated traits are taken. The data may contain the missing values these values are filled utilizing the preprocessing techniques. The missing values like exhaust cells, NA'S(not accessible) and zero's values are supplanted with segment mean, median, averages of the information and furthermore some arbitrary qualities that are produced from an algorithm.

C. Feature Reduction:

Feature Selection in the proposed demonstrates is executed utilizing the wrapper techniques. The strategy takes after the way toward choosing the subset of dataset from the dataset and applying the Machine Learning Algorithms to assess the exhibitions. The way toward choosing the subset utilizing calculations is rehashed until the point when the best subset is chosen. The data set used in the proposed model is built based on the analysis made on the CKD occurrence. The risk factors that mainly that occurred the CKD and diabetes are BP, hypertension, Weight, Diabetes, cholesterol this are some of the risk factors, based on the previous family history we can predict the disease to a patient.

C.1. Correlation:

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables[7]. The factors might be two sections of a given informational index of perceptions, frequently called an example or two parts of a multivariate irregular variable with a known

- A correlation coefficient of 1 implies that for each positive increment in one variable, there is a positive increment of a settled extent in the other. For instance, shoe sizes go up in (nearly) idealize relationship with foot length.
- A correlation coefficient of - 1 implies that for each positive increment in one variable, there is a negative diminishing of a settled extent in the other. For instance, the measure of gas in a tank diminishes in (nearly) idealize connection with speed.
- Zero implies that for each expansion, there is certifiably not a positive or negative increment. The two simply aren't connected.

$$r_{xy} = S_{xy} / S_x S_y$$

S_x & S_y = Sample Standard Deviation.

S_{xy} = Sample Covariance

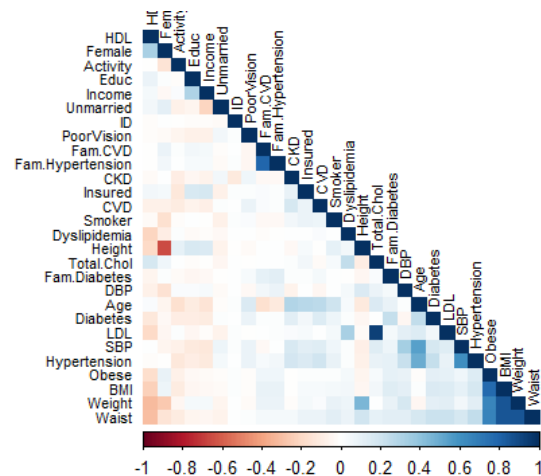


Fig.2:Highly Correlated features

C.2.LASSO

LEAST ABSOLUTE SHRINKAGE and SELECTION OPERATOR is additionally one of the element lessening models. To get the best precision and furthermore locate the better calculation. Tether changes every last coefficient by a consistent segment λ , truncating at zero. Consequently, it is a forward-looking variable choice technique for relapse. It diminishes the remaining entirety of squares subject to the aggregate of the total estimation of the coefficients being not as much of a conspiracy. The rope was initially characterized with regards to slightest squares, yet it can likewise be stretched out to a wide assortment of models. Rope enhances both forecast precision and model interpretability by joining the great characteristics of edge relapse and subset determination. In the event that there is a high relationship in the gathering of indicators, LASSO picks just a single among them and therapists the others to zero. It lessens the changeability of the evaluations by contacting the portion of the coefficients precisely to zero creating effectively interpretable models. Fig 2 demonstrates the variable significance with their probabilities.

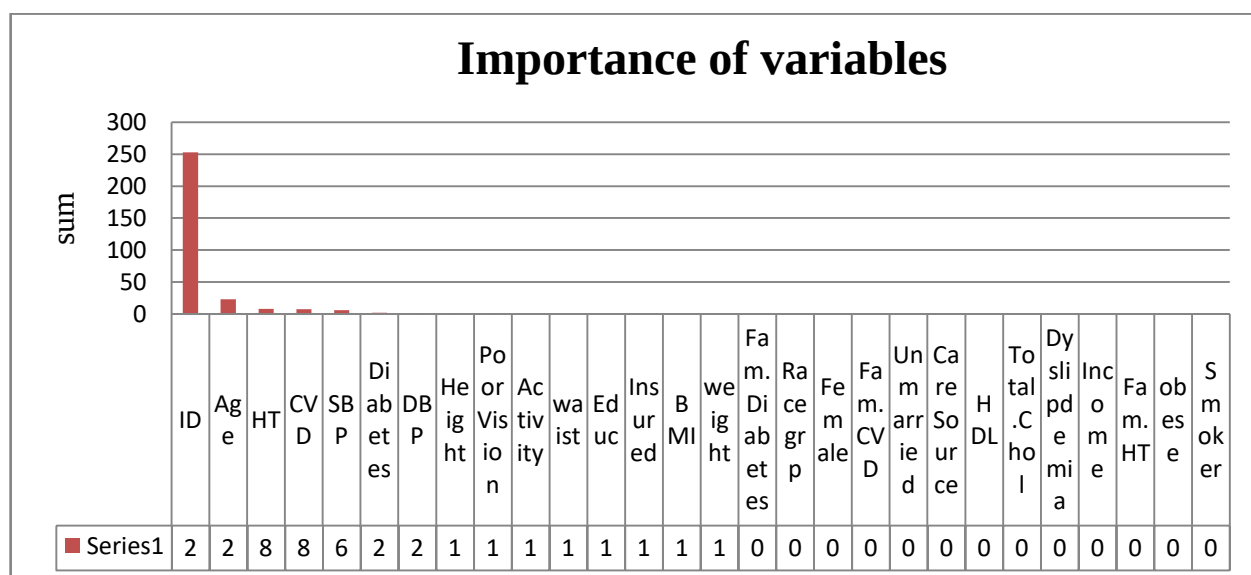


Fig.3.Importance of variables using LASSO

D. Learning Algorithms

The learning algorithms are used to classify the data and predictions, some of the learning algorithms that are used in this paper are SVM, Naive Bayes, Logistic Regression and KNN

SUPPORT VECTOR MACHINE:

In machine learning, Support vector machines (SVMs, additionally support vector networks) are an administered learning model with related learning calculations that investigate information utilized for order and relapse examination. The core idea of SVM is to find the optimal classification hyperplane to meet the classification requirements[8] The commotion focuses close to the edge of the class are frequently blended with legitimate examples in the SVM algorithm. In this way, the arrangement surface is frequently not ideal. The execution of the classifier isn't ideal.

For the order reason, we will utilize a support vector machine that it will express that what number of are accurately recognized. For the detection we will utilize four distinct portions, those are the linear kernel, polynomial kernel, RBF kernel, a signed kernel. Among these four portions, RBF is most productive and viable that it will give better performance when contrasted with others.

Linear SVM:

Support vector machines (SVM) is utilized for grouping, relapse and different assignments. SVM produces hyperplanes for a multi-dimensional zone. The hyperplanes go about as the utilitarian edge which lessens error rate for the classifiers. Spark's ML classifier support binary classification with SVM utilizing linear kernel. The Linear SVC classifier of the spark has outflanked well for the conveyed systems which contain expansive size datasets.

Naïve Bayes:

Naive Bayes is a classification method that uses Bayes theory that is based on probability and statistical knowledge [9]. This method was discovered by Thomas 2018 International Conference on Information and Communications Technology (ICOIACT) 299 Bayes in the 18th century. Decision-making on the Bayes theorem relates to inference probabilities that serve to collect knowledge about previous events by predicting events through the rule base [10]. Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below.

$$P(c/x)=P(x/c)P(c)/p(x)$$

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor

Logistic Regression:

Regression Analysis is an arrangement of factual procedures for assessing the connections among factors. It incorporates numerous methods for demonstrating and breaking down a few factors when the attention is on the connection between a reliant variable and at least one autonomous factors. All the more particularly, regression analysis encourages one to see how the normal estimation of the reliant variable changes when any of the free factors is differed, while the other autonomous factors are held fixed.

$$Ln[P/1-P]=\beta_0+\beta_1X_1+\beta_2X_2+.....+\beta_kX_k$$

KNN:

KNN is an Instance-Based automatic learning algorithm[11]. K-Nearest Neighbor (KNN) is a classification method that classifies data testing based on the distance function between data testing to the nearest training data (Neighbor) which has the highest number [12]. This algorithm will compare data testing with similar training data. if the data is not known then the data will be given the class of training data closest to the pattern space. This method is also called lazy learning method[13].

III. EXPERIMENT AND RESULTS

The data set consists of 9000 patients records with the 36 instances, to improve the performances we utilized the feature reducing algorithms. After the reduction some of the learning algorithms are applied and comparison.

Algorithms	SVM	Naïve Bayes
Correlation	91	86
LASSO	93.9	87.9

Table.2.Confusion Matrix on reduction Algorithms

Table.3. Comparison on Correlation

Performance Metrics	SVM	Naïve Bayes	Logistic Regression	KNN
Accuracy	91	86	90	78.9
Precision	92.2	86.8	92.6	92.5
Recall	94.3	90.5	92.2	79.2
F-Score	93.2	88.6	92.3	85.3

Performance Metrics	SVM	Naïve Bayes
Accuracy	93.9	87.9
Precision	98.1	87.5
Recall	92.6	85.5
F-Score	95.2	86.4

Table.4.Comparison on LASSO

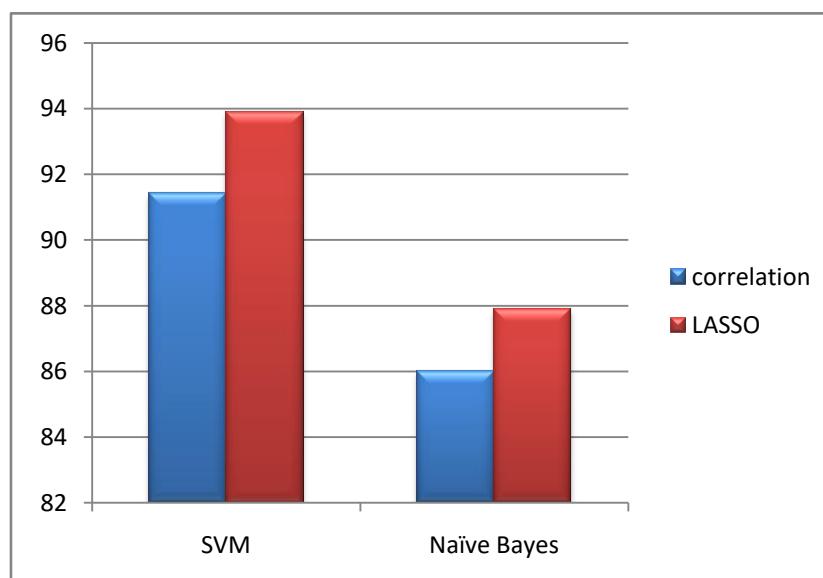


Fig.4. Graph for Confusion matrix of Correlation & LASSO

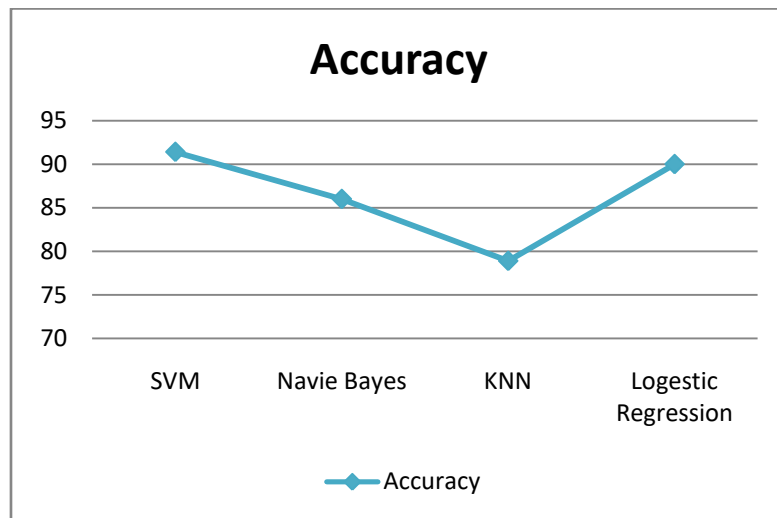


Fig.5.Accuracy Graph on Correlation Algorithm

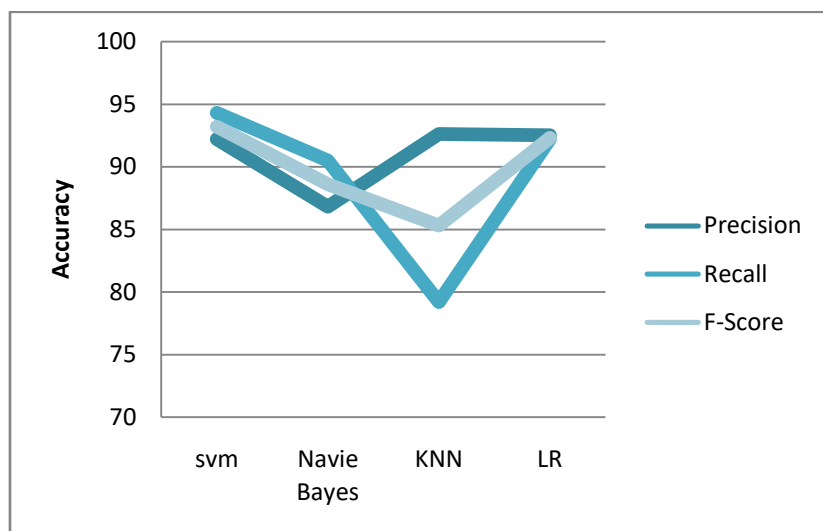


Fig.6.Comparsion On Correlation Algorithm

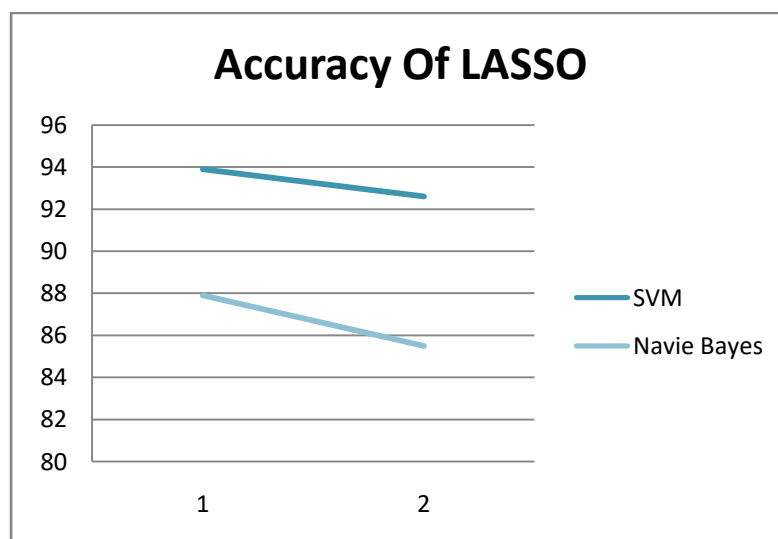


Fig.7.Accuracy graph of LASSO

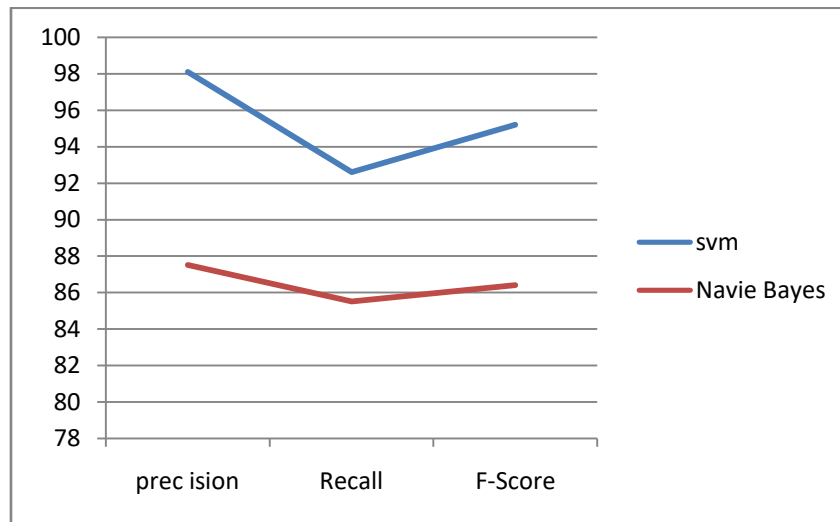


Fig.8.Comparison On LASSO Algorithm

IV. CONCLUSION

The diabetic dataset is utilized to classify the Kidney affected patients. The element decrease calculations like Correlation and LASSO algorithms got a diverse arrangement of highlights. In the wake of playing out this component decrease, the information is prepared. We apply the characterization calculations on this prepared informational collection and furthermore done the examination of this grouping calculations. We computed the execution assessment measurements as far as factual measures are finished for accuracy, review, f-score are figured.

REFERENCES

- [1] **Rajagopalan, Prashanth, et al.** "Population-based estimation of renal function in healthy young Indian adults based on body mass index and sex correlating renal volume, serum creatinine, and cystatin C." *International journal of nephrology and renovascular disease* 9 (2016): 243.
- [2] **Varughese, Santosh, and Georgi Abraham.** "Chronic Kidney Disease in India A Clarion Call for Change." *Clinical Journal of the American Society of Nephrology* 13.5 (2018): 802-804
- [3] "Essential Guide to Kidney Disease", Western Hospital, 2017. [Online]. Available: <http://www.westernhospital.lk/essential-guide-to-kidney-disease>. [Accessed: 22- Aug- 2017].
- [4] **Reddy, Manoj, and John Cho.** "Detecting Chronic Kidney Disease Using Machine Learning." *Qatar Foundation Annual Research Conference Proceedings*. Vol. 2016. No. 1. Qatar: HBKU Press, 2016.
- [5] **Bajestani, Narges Shafaei, et al.** "Nephropathy forecasting in diabetic patients using a GA-based type-2 fuzzy regression model." *Biocybernetics and Biomedical Engineering* 37.2 (2017): 281-289.

- [6] **Levey, Andrew S., et al.** "Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO)." *Kidney international* 67.6 (2005): 2089-2100.
- [7] **correlation coefficient**". NCME.org. National Council on Measurement in Education. Retrieved April 17, 2014. correlation coefficient
- [8] **Dai, Hong.** "Research on SVM improved algorithm for large data classification." *Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on.* IEEE, 2018.
- [9] **M. Mayilvaganan and D. Kalpanadevi,** "Comparison of classification techniques for predicting the performance of students academic environment," *Commun. Netw. Technol. (ICCNT), 2014 Int. Conf. Comput. Intell. Comput. Res.,* pp. 113–118, 2014.
- [10] **P. Suryachandra,** "Comparison of Machine Learning Algorithms," *3rd Int. Conf. Sci. Technol. - Comput. Comp.,* vol. 8, no. 5, pp. 2241–2247, 2017.
- [11] **Chandra, Nidhi, Sunil Kumar Khatri, and Subhranil Som.** "Anti social comment classification based on kNN algorithm." *Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2017 6th International Conference on.* IEEE, 2017.
- [12] **B. Y. Pratama and R. Sarno,** "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," in *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015, 2016,* pp. 170–174.
- [13] **Jupri, M., and Riyanarto Sarno.** "Taxpayer Compliance Classification Using C4. 5, SVM, KNN, Naive Bayes and MLP." (2014).