DATA MINING CLASSIFICATION ALGORITHMS FOR KIDNEY DISEASE PREDICTION

Dr. S. Vijayarani[1], Mr.S.Dhayanand[2],

Assistant Professor[1], M.Phil Research Scholar[2,]
Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India[1, 2].

ABSTRACT

Data mining is a non-trivial process of categorizing valid, novel, potentially useful and ultimately understandable patterns in data. In terms, it accurately state as the extraction of information from a huge database. Data mining is a vital role in several applications such as business organizations, educational institutions, government sectors, health care industry, scientific and engineering. . In the health care industry, the data mining is predominantly used for disease prediction. Enormous data mining techniques are existing for predicting diseases namely classification, clustering, association rules, summarizations, regression and etc. The main objective of this research work is to predict kidney diseases using classification algorithms such as Naïve Bayes and Support Vector Machine. This research work mainly focused on finding the best classification algorithm based on the classification accuracy and execution time performance factors. From the experimental results it is observed that the performance of the SVM is better than the Naive Bayes classifier algorithm.

KEYWORDS

Data mining, Disease prediction, SVM, Naïve Bayes, Glomerular Filtration Rate (GFR)

1. INTRODUCTION

Data mining is an approach which dispense an intermixture of technique to identify a block of data or decision making knowledge in the database and eradicating these data in such a way that they can be put to use in decision support, forecasting and estimation [11]. The data is often voluminous, but it has data that is useful. Two major preferred models that can be created in data mining are predictive and descriptive. Under these two models there are various tasks that are used in the data mining process. On basis of various historical data a predictive model makes estimation about values of data using recognized results found from various data. On the other side, descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model obliges as way to explore the properties of the data observed, not to predict new properties [5]. The algorithms are many in every single task under both the data mining models which are used for various purposes according to the convenient of the use requirements. The various tasks of the predictive and descriptive models are classification, clustering, summarization, prediction, time series analysis, association rules and regression [3].

DOI: 10.5121/ijci.2015.4402

In order to anticipate solution set for various problems data mining technique endeavors distinctive data mining tasks such as classification and clustering. It provides affirmation about the predicted solutions in terms of the stability in prediction and in frequency of legitimate predictions. Based on data mining techniques, many experts develop their research successfully. Some of the technique includes statistics, machine learning, decision trees, hidden markov models, genetic algorithm, Meta learning and so on. Data mining systems depends on database to supply the raw input and this raises problems, such as that database tends to be dynamic, incomplete, noisy and large. Other problems arise as a result of the insufficiency and insignificance of the information stored. The major issues in data mining can be categorized as noise or missing data, Limited information, user interaction, prior knowledge, uncertainty, size, updates and irrelevant fields.

The medical data mining has the elevation potential in medical domain for extracting the hidden patterns in the dataset [9]. These patterns are used for medical diagnosis and prognosis. The medical data are globally scattered, heterogeneous, exaggerate in nature. In order to incur a user oriented approach to novel and hidden patterns of the data, the data should be concerted together [16]. A major problem in health science or bioinformatics exploration is in managing the correct diagnosis of certain important information. Generally multitudinous tests involve the classification or clustering of large scale data for the purpose of esteemed scrutiny.

The test procedures are assumed to be essential in order to reach the ultimate diagnosis. Else,more number tests could obfuscate the main diagnosis process which may result in trouble in gaining the end results, predominantly in the perceptively of finding disease many tests should be performed [12]. This sort of difficulty could be fixed with the support of machine learning which could be used directly to obtain the end result with the assistance of several artificial intelligent algorithms which perform the role as classifiers. Classification is one of the most important techniques in data mining. In order to perform classification process, classifying the data has to be done proceed by coding and then placed into chunk that are submissive by a human. This research work describes classification algorithms and it also analyzes the performance of these algorithms. The performance factors used for analysis are classification accuracy and execution time.

The main objective of this research work is to predict kidney diseases (Acute Nephritic Syndrome, Chronic Kidney disease, Acute Renal Failure, Chronic Glomerulonephritis) using classification algorithms namely SVM and naïve bayes and finding the efficient algorithm.
The remaining portion of the paper is organized as follows. Related works are discussed in Section 2. The proposed methodology is given in Section 3. Section 4 analyzes the experimental results. Section 5 gives conclusion.

2. LITERATURE REVIEW

Giovanni Caocci et.al [7] In order to predict Long Term Kidney Transplantation Outcome, they interpreted discrimination between an Artificial Neural Network and Logistic Regression. Comparison has been done based on the Sensitivity and specificity of Logistic Regression and an Artificial Neural Network in the prediction of Kidney rejection in ten training and validating datasets of kidney transplant recipients. From the experimental results that both the algorithm approaches were complementary and their combined algorithms used to improve the clinical decision-making process and prognosis of kidney transplantation.

Lakshmi.K.R et al [10] analyzed Artificial Neural Networks, Decision tree and Logical Regression supervised machine learning algorithms. These algorithms have been used for Kidney dialysis. For classification process they used a data mining tool named Tanagra. The 10 fold cross validation is used in order to evaluate the classified data proceeded by the comparison of those data. From the experimental result they absorbed that ANN performed better than the Decision tree and Logical Regression algorithms.

Tommaso Di Noia et.al [14] developed a software tool that exploits the power of artificial neural networks to classify patients' health status potentially leading to End Stage of Kidney Disease (ESKD). The classifier influences the results returned by an ensemble of ten networks trained by using data collected in a period of thirty eight years at University of Bari. The tool which has been refined has been made derivable both as an online web application and as an android mobile app. The developed tool is important to clinical usefulness based on the largest cohort worldwide.

Anu Chaudhary et al [2] developed a prediction system using A-priori and k-means algorithm for heart disease and kidney failure prediction. In her survey A-prior and k-mean algorithm algorithms have been used to predict kidney failure patient with 42 attributes. They analyzed the data using machine learning tools such as distribution and attribute statistics, followed by A-prior and k-means algorithms. They evaluated the data using Receiver Operating Characteristic (ROC) plot and calibration plots.

Andrew Kusiak et al [1] have used data preprocessing, data transformations, and a data mining approach to elicit knowledge about the interaction between many of measured parameters and patient survival. Two different data mining algorithms were engaged for extracting knowledge in the form of decision rules. Those rules were used by a decision-making algorithm, which predicts survival of new unseen patients. Important parameters identified by data mining were interpreted for their medical significance. They have introduced a concept in their research work have been applied and tested using collected data at four dialysis sites. The approach presented in their paper reduces the cost and effort of selecting patients for clinical studies. Patients can be chosen based on the prediction results and the most important parameters discovered.

3. METHODOLOGY

3.1 Dataset

The synthetic kidney function test (KFT) dataset have been created for analysis of kidney disease. This dataset contains five hundred and eighty four instances and six attributes are used in this comparative analysis. The attributes in this KFT dataset are Age, Gender, Urea, Creatinine and Glomerular Filtration Rate (GFR). This dataset consists of renal affected diseases.

Blood Urea Nitrogen: Urea is a surplus product that is eliminated by the kidneys. Nitrogen is a derivative product from urea, also eliminated by kidneys. When kidney function reduces, the BUN may be elevated.

Creatinine: this is an excess product of muscles and is normally eliminated by the kidneys. When kidney function reduces, the creatinine may be elevated.

Glomerular Filtration Rate (GFR): This is an essential measure and it is used to calculate the creatinine clearance. Normally this measure is calculated by using the following attributes; they are, age, body, sex of the patient and creatinine. This measure is considered as the best measure for finding the kidney function level and it is represented in percentage (i.e.30%).

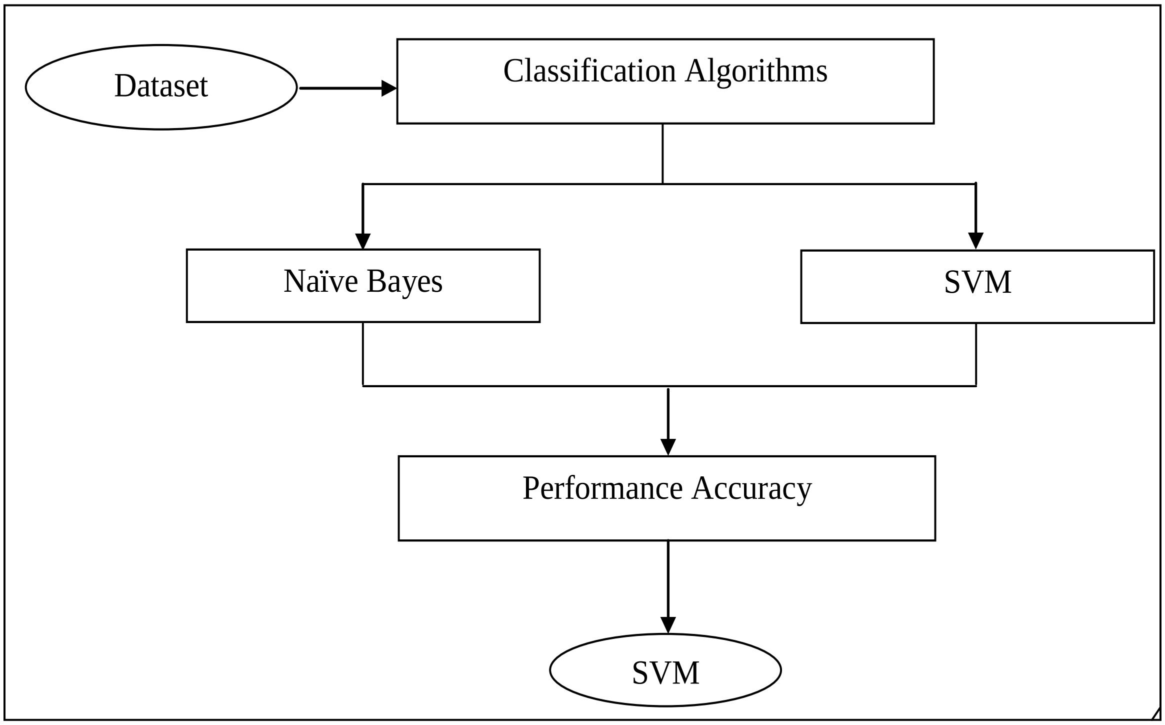

Figure 1. System Architecture

3.2 Classification

Classification – it maps data into predefined groups or classes. In classification the classes are indomitable before examining the data thus it is often mentioned as supervised learning . Classification is the process which classifies the collection of objects,datas or ideas into groups, the members of which have one or more characteristic in common. In this research work Naïve Bayes, SVM, ANN and proposed algorithm namely ANFIS are used to classify different stages of Chronic Kidney Failure disease from the dataset [4].

3.2.1 Naïve Bayes

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". This restricted individuality assumption infrequently clutches true in real world applications, hence the characterization as Naive yet the algorithm inclines to perform well and learn rapidly in various supervised classification problems [6]. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Table 1 represents and explains the Bayes theorm

Table 1. Bayes Theorm

Bayes theorem:

1. P (C|X) = P (X|C) ·P(C) / P(X).
2. P(X) is constant for all classes.
3. P(C) = relative freq of class C samples c such that p is increased=c Such that P (X|C) P(C) is increased
4. Problem: computing P (X|C) is unfeasible! [15] [17].

3.2.2 Support Vector Machine (SVM)

Support vector machine ensures a machine learning technique on the basis of statistical learning theory. It creates a discrete hyperplane in the descriptor space of the training data and compounds are classified based on the side of hyperplane located.

The advantage of the SVM is that, by use of the so-called ‘‘kernel trick’’, the distance between a molecule and the hyperplane can be calculated in a transformed (nonlinear) feature space, lacking of the explicit transformation of the original descriptors. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study. The kernel function is expressed as follows [8]:

$$K(\bar{x},\bar{x}_i)=\exp(-\frac{\|\vec{x}-\vec{x}_i\|^2}{2\alpha^2}) \qquad \text{(a)}$$

In the above equation (a), the kernel width parameters control the amplitude of the Gaussian function reflecting the generalization ability of SVM. The regularization parameter C is censurable for inhibiting transaction between maximizing the margin and minimizing the training error. In00 recent times, particular attention has been dedicated to support vector machines (SVMs) for the classification of diseases. SVMs have frequently been found to provide maximum classification accuracies than other widely used pattern recognition techniques, such as the maximum likelihood and the multilayer perceptron neural network classifiers. Table 2 represents and explains the mathematical formulation of support vector machine.

Table 2 : SVM Mathematical Formulation

Step 1: Let's assume a supervised binary classification problem. Let us consider that the training set consists of N vectors from the -dimensional feature space

$$x_i \in \Re^d \, (i = 1,2,...., N).$$

Step 2: A target $y_i \in \{-1,+1\}$ is associated to each vector x_i.

Step 3: Let us consider that the two classes are linearly separable. This points that it is possible to discovery at least one hyperplane (linear surface) defined by a vector $w \in \Re^d$ (normal to the hyperplane) and a bias $b \in \Re$ that could separate two classes without errors.

Step 4: The membership decision rule can be based on the function sgn [f(x)], where f(x) is the discriminant function associated with the hyperplane and defined as

$$f(x) = w.x + b. \qquad (1)$$

In case to find such a hyperplane, one should estimate w and so that

$$y_i(w.x_i + b) > 0, \qquad \text{with } i = 1,2,..., N. \qquad (2)$$

Step 5: The SVM approach involves in discovering the optimal hyperplane that increases the distance between the neighboring training sample and the splitting hyperplane. It is possible to express this distance as equal to $1/\|w\|$ with a simple rescaling of the hyperplane parameters w and b such that

$$\min_{1=1,2,...,N} y_i(w.x_i + b) \geq 1. \qquad (3)$$

Step 6: Consequently, it changes the optimal hyperplane which can be controlled by the following solution of convex quadratic programming problem

$$\begin{cases} \min imize : \frac{1}{2} \| w \|^2 \\ subject :| \; y_i(w.x_i + b) \geq 1, \end{cases} \qquad i = 1,2,..., N. \qquad (4)$$

Step 7: This traditionally linear constrained optimization problem can be interpreted (using a Lagrangian formulation) into the following dual problem:

$$\begin{cases} \max imize : \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{J=1}^{N} \alpha_i \alpha_j y_i y_j (x_i.x_j) \\ subject.to : \sum_{i=1}^{N} \alpha_i y_i = 0 and \alpha_i \geq 0, \end{cases} \qquad i = 1,2,..., N. \qquad (5)$$

Step 8: The Lagrange formulizers α_i's ($i = 1,2,..., N$) represented in (5) can be assessed using quadratic programming (QP) methods. The discriminant function associated with the optimal hyperplane becomes an equation depending both on the Lagrange multipliers and on the training samples, i.e.,

$$f(x) = \sum_{i \in s} \alpha_i y_i (x_i.x) + b \qquad (6)$$

Where s is the subset of training samples corresponding to the nonzero Lagrange multiplier's. It is worth noting that the Lagrange multipliers effectively weight each training sample according to its importance in determining the discriminant function. The training samples associated to nonzero weights are called support vectors. These lie at a distance exactly equal to $1/\|w\|$ from the optimal separating hyperplane

4. EXPERIMENTAL RESULTS

This work is implemented in Matlab tool. MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. Developed by MathWorks, MATLAB permits matrix manipulations, employment of algorithms, inception of user interfaces, plotting of functions and data and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python. The experimental comparison of classification algorithms are done based on the performance measures of classification accuracy, error rate and execution time.

4.1 Classification Accuracy

Accuracy

Accuracy is defined in the terms of correctly classified instances divided by the total number of instances present in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

TP Rate: It is the ability which is used to find the high true-positive rate. The true-positive rate is also called as sensitivity.

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

Precision

Precision is given the correlation of number of modules correctly classified to the number of entire modules classified fault-prone. It is quantity of units correctly predicted as faulty.

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

F-Measure

F- Measure is the one has the combination of both precision and recall which is used to compute the score. In the field of Information Retrieval the F-measure is habitually used in order to guesstimate the query classification performance.

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

Table 5 represents the performance of classification accuracy measure of the datasets using classification algorithms such as SVM and Naïve Bayes.

Table 5: Accuracy Measure for Classifier Algorithms

Algorithms	**Correctly Classified Instances (%)**	**Incorrectly Classified Instances (%)**	**TP Rate**	**Precision**	**F Measure**	**Recall**
Naïve Bayes	70.96	29.04	0.709	0.809	0.192	0.109
SVM	76.32	23.68	0.763	0.820	0.213	0.173

Figure 2 represents the accuracy measure and figure 3 represents the performance measure for the classification algorithms namely Naive Bayes and SVM. From the experimental result, SVM performs best in classifying process than Naïve Bayes algorithm. This chart represented as given in table 5.

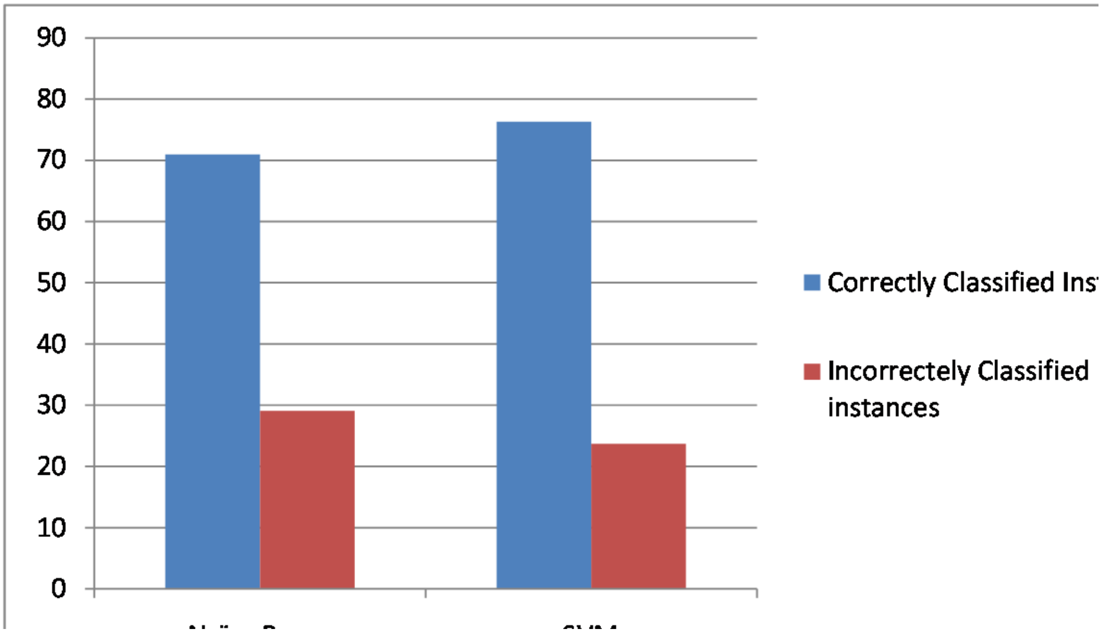

Figure 2: Accuracy measure for Classification Algorithms

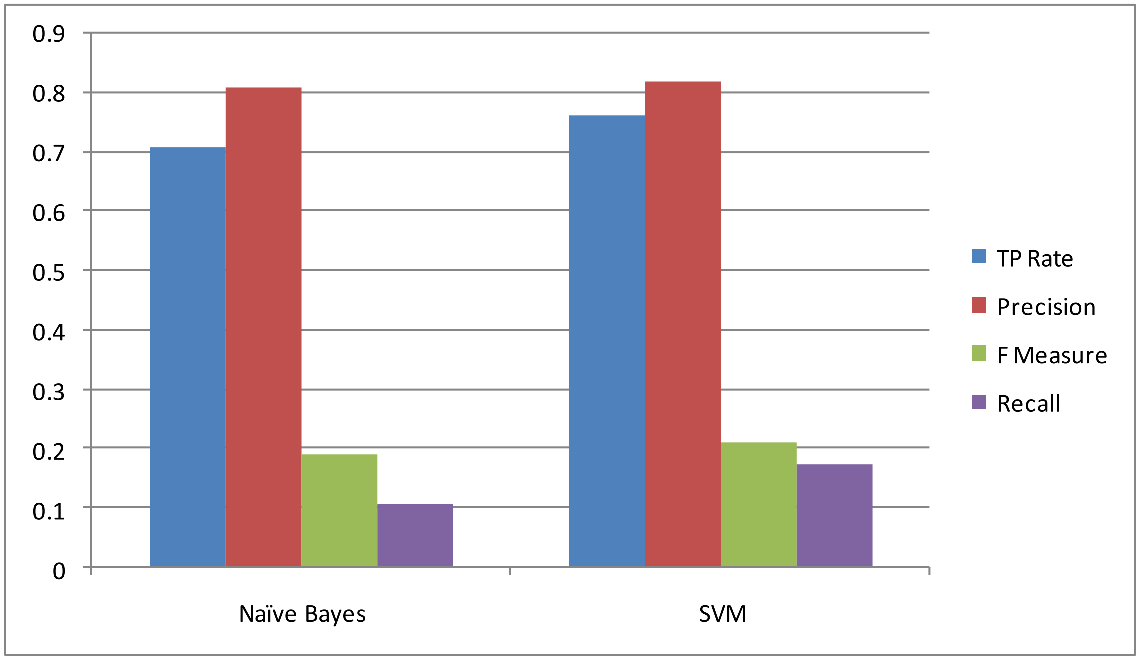

Figure 3: Performance measure for Classification Algorithms

4.2 Execution Time

Table 6 represents the execution time of the classification algorithms

Table 6: Execution time Analysis

Algorithms	**Execution Time in Seconds**
Naïve Bayes	1.29
SVM	3.22

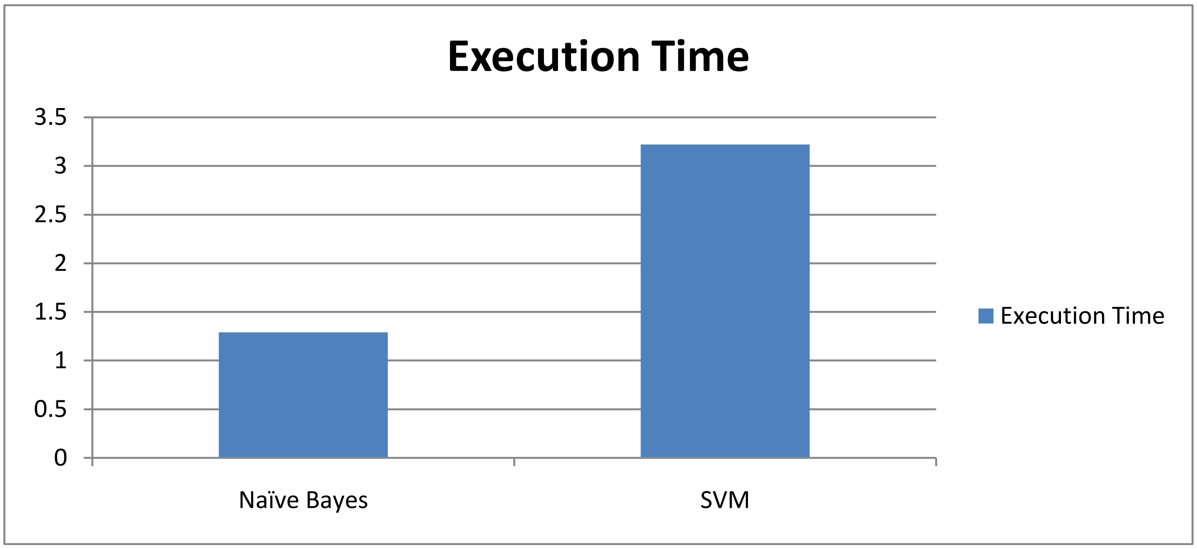

Figure : Execution Time of Classification Algorithms

Figure 4 represents the time taken for execution process. Naïve Bayes performs with minimum period of execution time than the other algorithms. This chart represented as given in table 6.

Table 7 represents and describes the classification of kidney diseases as given below.

Table 7. Classification of Kidney Diseases

Classifiers → **Kidney Disease**	**Naïve Bayes**	**SVM**
Normal	428	435
Acute Nephritic Syndrome	49	45
Chronic Kidney disease	35	42
Acute Renal Failure	19	19
Chronic Glomerulonephritis	52	42

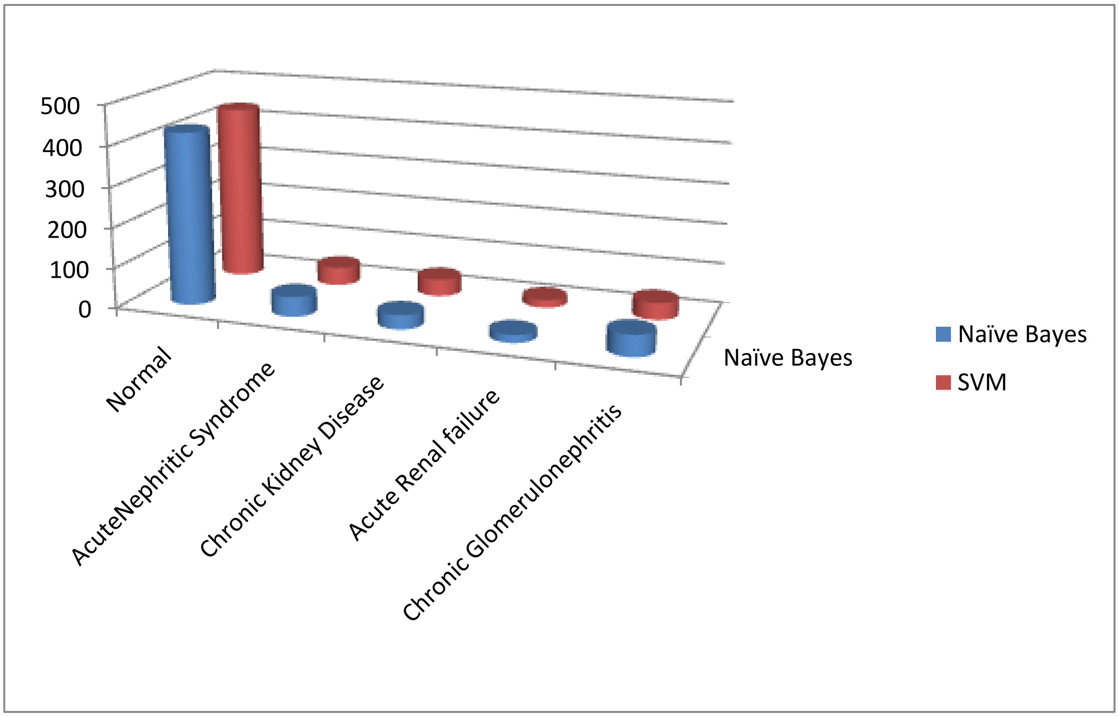

Figure 5: Classification of Kidney Diseases

Figure 5 represents the Kidney diseases classified by different types of classification algorithms, Naïve Bayes and SVM algorithms. Based on chart analysis, SVM gives the overall best classification result than other algorithm.

5. RESULT AND DISCUSSION

The algorithm which has the higher accuracy with the minimum execution time has chosen as the best algorithm. In this classification, each classifier shows different accuracy rate. SVM has the maximum classification accuracy and it is considered as the best classification algorithm. But Naïve Bayes perform as best with minimum execution time.

6. CONCLUSION

In this research work classification process is used to classify four types of kidney diseases. Comparison of Support Vector Machine (SVM) and Naïve Bayes classification algorithms is done based on the performance factors classification accuracy and execution time. From the results, it can be concluded that the SVM achieves increased classification performance, yields results that are accurate, hence it is considered as best classifier when compared with Naïve Bayes classifier algorithm. Perhaps, Naïve Bayes classifier classifies the data with minimum execution time.

REFERENCE

[1] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) *Predicting survival time for kidney dialysis patients: a data mining approach*, Elsevier Publication, Computers in Biology and Medicine 35, page no 311–327

[2] Anu Chaudhary, Puneet Garg,(2014) *Detecting and Diagnosing a Disease by Patient Monitoring System*, International Journal of Mechanical Engineering And Information Technology, Vol. 2 Issue 6 //June //Page No: 493-499.

[3] Approaches, *Knowledge-Oriented Applications in Data Mining*, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1,InTech,http://www.intechopen.com/books/knowledge-oriented-applications-in-datamining/mining-enrollment-data-using-descriptive-and-predictive-approaches

[4] Cristóbal Romero, *Data Mining Algorithms to Classify Students*, http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf

[5] Fadzilah Siraj, Mansour Ali Abdoulha, (2011). *Mining Enrollment Data Using Descriptive and Predictive*

[6] George Dimitoglou, *Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability*

[7] Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orrù, Carlo Carcassi and Giorgio La Nasa, *Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome*, Chapter 5, an open access article distributed under the terms of the Creative Commons Attribution License, http://dx.doi.org/10.5772/53104

[8] Gualtieri. J. A, Chettri. S. R, Cromp. R. F and Johnson.L. F, (1999) *Support vector machine classifiers as applied to AVIRIS data*, in Summaries 8th JPL Airborne Earth Science Workshop, JPL Pub. 99-17, pp. 217–227.

[9] Ian H. Witten and Eibe Frank.(2005) *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition

[10] Lakshmi. K.R, Nagesh. Y and VeeraKrishna. M, (2014) *Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability*, International Journal of Advances in Engineering & Technology, Mar., Vol. 7, Issue 1, pg no. 242-254.

[11] Mahesh Mudhol Purushothama Gowda,(2004) *Data Mining in the Process of Knowledge Discovery in Digital Libraries*, 2nd Convention PLANNER, Manipur Uni., Imphal, 4-5 November, 2004, page no 164-167

[12] Ruben D. Canlas Jr,(2009) *Data Mining In Healthcare: Current Applications And Issues*, August

[13] Tadjudin. S and Landgrebe. D.A, (1999) *Covariance estimation with limited training samples*, IEEE Trans. Geosci. Remote. Sensing, vol. 37, pp. 2113–2118, July

[14] Tommaso Di Noia, Vito Claudio Ostuni, Francesco Pesce, Giulio Binetti, David Naso, Francesco Paolo Schena, Eugenio Di Sciascio,(2013) *An end stage kidney disease predictor based on an artificial neural networks ensemble*, Elsevier Publication, Expert Systems with Applications 40, page no 4438–4445

[15] Uffe B. Kjærulff, Anders L. Madsen, (2005) *Probabilistic Networks — an Introduction to Bayesian Networks and Influence Diagrams*, 10 May

[16] Vijayarani. S, Sudha. S, (2013) *Comparative Analysis of Classification Function Techniques for Heart Disease Prediction*, International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May, page no 735- 741

[17] Zhang H.; Su J, *Naive Bayesian classifiers for ranking*. Paper appeared in ECML2004 15th European Conference on Machine Learning, Pisa, Italy.

AUTHORS

Dr. S. Vijayarani has completed MCA, M.Phil and Ph.D in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Mr. S. Dhayanand has completed MSc, in Software Systems. He is currently pursuing his M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. His fields of research interest are data mining and medical mining. He has presented research papers in international, national conferences and Symposiums.