Critical Disease Prediction using Supervised Classification Model for Healthcare Applications

Veerpal Kaur[1], Simranjit Kaur[2]
Swami Vivekanand Institute of Engineering & Technology, Banur (Chandigarh)

ABSTRACT--*A set of algorithms is required in the balanced combination in order to achieve the goal related to the chronic kidney disease classification for online healthcare databases. The information hiding method has been already implemented using the various forms of variable normalization algorithms for quantitative and qualitative algorithms. The proposed model is also the approach towards the new-age data filtering model by using the layered approach for qualitative data, which involves the labeling and dummy variable transformations. The proposed model utilizes the maximum-minimum scaling method to scale the quantitative variables on 0-1 scale, after handling the missing values with column mean value. The SVM and KNN based classification method are used to predict the patterns for the chronic kidney diseases. The experimental results have proved the proposed model based on SVM classification as the most efficient algorithm for the purpose of kidney disease prediction. The SVM has been recorded with 98.92% (mean) and 99% (median) of accuracy, which is significantly higher than KNN's 97% (both mean and median). Also SVM outperformed KNN on the basis of precision by (98.37% mean) and recall (99.94% mean) against 96.95% (precision mean) and 98.18% (recall mean).*

KEYWORDS—Machine learning, KNN, SVM, CRIDM

I. INTRODUCTION

Information about the mining could be the technique of inspecting & gather info by different perspectives along with outlining the idea in beneficial details. The idea details helpful to improve income, reduces prices, or both equally. It really is used for gather the data by different- different web sites. Information about the mining could be the technique of finding info within big relational listings.The item a good choice for end users to handle the information effortlessly. Process for finding files habits undetectable inside significant files units. In other words Data mining features captivated a great deal of consideration inside the facts market and also inside modern society as a whole nowadays. Data mining is the term for removing or even "mining" understanding by a lot connected with files, typically immediately compiled.

Data mining finds valuable information hidden in large volumes of data. It is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The database system industry has witnessed an evolutionary path in the development of the following functionalities:

- Data collection and database creation
- Data management (including data storage and retrieval, and database transaction processing)
- Advanced data analysis (involving data warehousing and data mining).

Amassing far more amounts of data is referred to as files exploration. In the beginning, with the entire introduction connected with pc in addition to means for mass digital storage space, we started gathering in addition to holding all kinds of files, relying on the facility connected with PCs to help you examine this particular amalgam connected with data. The actual proliferations connected with database management methods in addition have led in order to latest massive get together connected with all kinds of data.

II. LITERATURE REVIEW

Dhutraj et al. proposed the system for hiding sensitive association rules using hybrid algorithm where the dataset is distributed over the network. In this cryptographic technique are used for providing better security when data transfer from each party to trusted third party. Hybrid algorithm used is the combination of ISL and DSR technique and association rule hiding is based on modifying the database transactions so that the confidence of the association rules can be reduced. **Domadiya et al.** proposed that a heuristic based algorithm named MDSRRC (Modified Decrease Support of R.H.S. item

of Rule Clusters) to hide the sensitive association rules with multiple items in consequent (R.H.S) and antecedent (L.H.S). This algorithm overcomes the limitation of existing rule hiding algorithm DSRRC. Proposed algorithm selects the items and transactions based on certain criteria which modify transactions to hide the sensitive information. **Jadav et al.** surveyed methods of hiding sensitive association rules by identifying some open challenges that will be useful to research community in this area. It is found that finding an optimal solution for sanitizing database (to protect privacy of sensitive information) is NP-Hard. Existing approaches provide only the approximate solution to hide sensitive knowledge. There is need of finding exact solution to the privacy problem in database disclosure. **Shah et al.** described association rule hiding approaches and surveyed existing algorithm for association rule hiding. Based on this, comparative analysis of heuristic algorithms described. **Thakur et. al.** discussed the basic of PPDM and its different approaches. Subsequently, association rule hiding approaches and metrics for performance comparison of those approaches are discussed. This paper provided an overview of heuristic approaches.

III. PROBLEM FORMULATION

Chronic kidney disease is a condition characterized by a gradual loss of kidney function over time. If kidney disease gets worse, wastes can build to high levels in your blood and make you feel sick. You may develop complications like high blood pressure, anemia, weak bones, poor nutritional health and nerve damage. Early detection and treatment can often keep chronic kidney disease from getting worse. Review has shown that the use of K- nearest neighbor to detect the problem of chronic kidney disease is ignored by researchers in most of the cases. The use of support vectors and nearest neighbors is ignored for the accurate results from the predicting data. Most of the existing techniques and algorithms are limited to some extent for prediction of chronic kidney diseases therefore utilizing more features may provide more significant results. The existing model is found inefficient in the case of low order or high order feature extraction, which computes the probability, factorization and entity relationships to describe the overall relationships between then entities. Also the use of dimensionality reduction can improve the elapsed time of the proposed model to meet the rising volumes of data every year. In order to overcome these issues, this research work has proposed a technique. The proposed technique will utilize KNN and SVM algorithms in MATLAB by accessing dataset in itself to predict the disease in a human. Also to improve the accuracy rate, further the supervised learning is also used as pre-analysis or processing while classifying chronic kidney disease dataset.

IV. RESEARCH GAP

1. The use of KNN to easily detect how easily and how long chronic kidney diseases exist in a person is ignored by most of the researchers. The use of naive bayes is ignored to check the accuracy depending upon the prediction and detection.
2. Most of the existing techniques are limited to some significant features of predicting CKD therefore utilizing more features may provide more significant results. In the case of existing model, the accuracy of nearly 99% has been achieved in the case of multi-layer preceptron, J48 and Decision table algorithms. The accuracy of data may decrease with the rising number of records. Hence the more descriptive method can be utilized for the feature extraction, which involves the probabilities, factorization, relationship library, etc.
3. The existing model does not utilize any of the dimensionality reduction algorithms, such as independent component analysis, etc., which can improve the elapsed time of the proposed model. In the era of big data, the data mining techniques must be efficient and quick to meet the rising volumes of data, which makes it mandatory to improve the elapsed time.

V. EXPERIMETNAL DESIGN

In this algorithm, the scaling of the data is being processed before undergoing the Euclidean distance in order to avoid the column dominance. The categorical feature handling must possess the steps to ensure the maximum classification accuracy, which includes missing values followed by numeric labeling & dummy variable creation with one column removal.

$$KCR = \log_{10}\left(\frac{N_{keypoints}}{N_{corners}}\right) \overset{?}{\leq} T_1.$$

To fit the supervised classification model over the set of the points in the given dataset for the major five stock markets, which includes the data of technology, political, business and sport news in the text file form encoded “UTF-8” encoding. The standard supervised classification equation stated, y = mx + b, is used for the supervised classification model in this price prediction model, where the m denotes the line’s or curve’s slope, b denotes the y-intercept over the line and x gives the properties of data in the given data. For the best fitting and prediction, we need to utilize the best set of the points around the slope (denoted m) and intercept curve or values (denoted b) with the y-intercept.

The standard error functions are used to compute the error or cost using the supervised classification equations, which takes the input of the data values and return the cost, which defines

the predicted values of the price in our model. The conventional to squared distance is computed to ensure the value as positive for the prediction of the price with flexibility. The normal equation is given as the following:

$$\text{Error}_{(m,b)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$

Where the influential factors (quantity and number of suppliers) have been denoted by m and b respectively, x gives the price data and y denotes the y-intercept to discover the new trend. N represents the total number of the data rows, whereas Error $_{(m,b)}$ gives the result computed by the normal equation.

Gradient Descent is also an error or cost function to predict the future price. The equation of the gradient descent can be given with the following equation:

$$\frac{\partial}{\partial \text{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \text{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

Where the influential factors (quantity and number of suppliers) have been denoted by m and b respectively, x gives the price data and y denotes the y-intercept to discover the new trend, specifically, in y_i, i index gives index of the different price entities. N represents the total number of the data rows, whereas There are two types of gradient descent, m and b,

which is denoted by the (delta / delta * m) and (delta / delta * b) in the above normal equation.

Simple Supervised Classification: The model for Simple Supervised classification is given by $\mathbf{Y} = \beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_n x + \varepsilon$, where

- **Y** is the dependent variable
- X is the independent variable
- $\boldsymbol{\varepsilon}$ is the random error variable
- $\beta_\circ$ is the y-intercept of the line $y = \beta_1 + \beta_0 x$
- β_1 is the slope of the line $y = \beta_1 + \beta_0 x$

In the model above:

Y and X are assumed to be **correlated**, i.e., linearly related, and thus the model function takes the form of a line, $\boldsymbol{Y} = \beta_0 + \beta_1 X$. Although we have discussed the complete algorithm in section 4.6, which elaborates the overall working of the simple supervised classification classifier for the test of validity of this hypothesis deciding the category of the news data. The simple linear classification revolves around the fitting of the equation with all of the independent variable and coefficients as the equation design. The final result is derived from the list of squared distances, which defines the real-time differences from the training data. The match with the lowest distance decides the class or category of the target news data.

VI. METHODOLOGY

The proposed algorithm named as Critical Relevant Information Description Mechanism (CRIDM) algorithm. Critical Relevant Information Description Mechanism (CRIDM) has equally performed or outperformed the existing information discovery algorithms in terms of data classification. The initial design analysis of the algorithm has stated that when the performance parameter of elapsed time would be tested on the latter mentioned four databases it will yield good and acceptable results. An equal or better data representation function values with less elapsed time will be calculated to prove the proposed algorithm better than existing algorithms for large datasets. In future, this algorithm will be tested with an adequate number of datasets and will be compared with the existing information discovery or classification algorithms in the terms of other performance parameters also. Its performance will be also tested and compared with other similar algorithms on the basis of various datasets and more performance parameters. Because the proposed algorithm is proved to be useful for the disease pattern mining, it will be enhanced to perform better than the proposed algorithm by combining it with different algorithms to develop new algorithms using new algorithmic combinations or newly developed algorithms

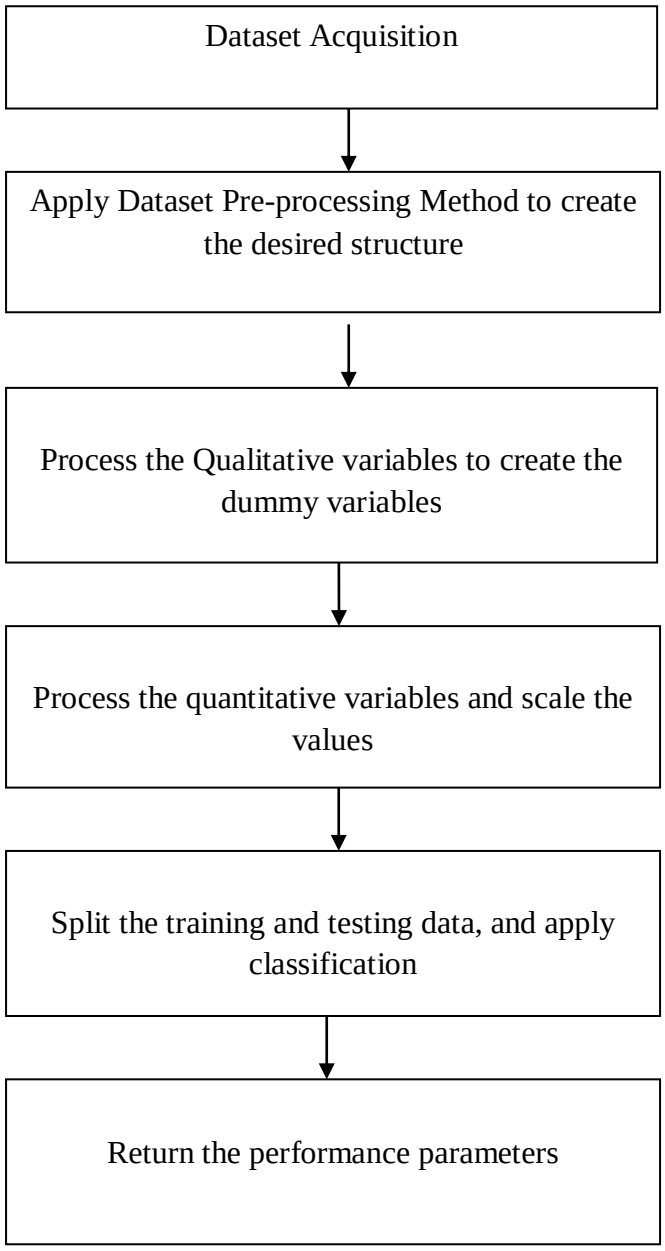

Figure 1: System design of the Proposed Research

VII. RESULT ANALYSIS

The comparison between the classification accuracy, precision, recall, f1 error and statistical errors has been conducted in this section. The comparison includes the average, standard deviation, median, minimum and maximum values. The comparative analysis is supposed to show the clear analysis, and helps to declare the best algorithm. Hence, the averaging factors play the key roles to distinguish the performance. The following table compares both classification, KNN & SVM on the basis of the true type errors from the statistical errors. The SVM is known to produce the more number of true positive on the average than KNN, although the maximum number of is true negative is higher in case of KNN than SVM.

Table 1: Comparative Analysis of KNN and SVM based on type-1 parameters

Parameters	SVM		KNN	
	TP	TN	TP	TN
Mean	60.72	38.2	59.84	37.16
Median	60	39	60	37
Max	67	45	67	46
Min	53	29	52	29

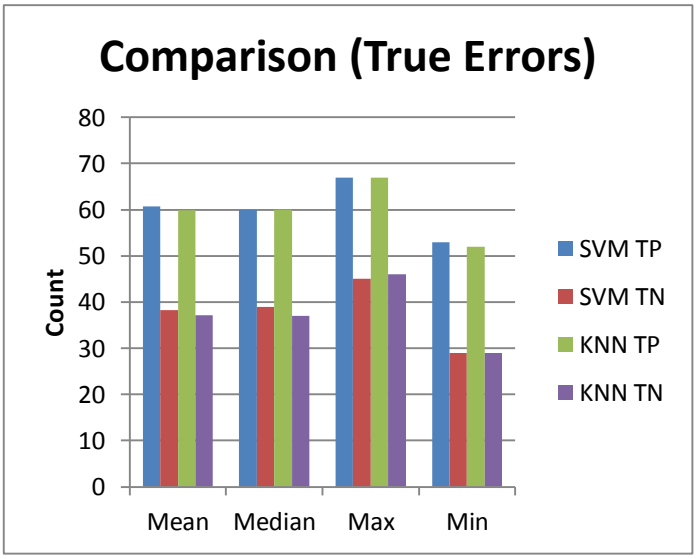

Figure 2: Comparative Analysis of KNN and SVM based on type 1 parameters

The SVM and KNN have been compared on the basis of false type parameters, which include the false positive and false negative parameters. The following table shows the higher performance of SVM on the basis of average false positive cases, which is 1.04 in comparison to 1.88 for KNN. This

pattern is identical with lower difference on the basis of false negative.

Table 2: Comparative Analysis of KNN and SVM based on type 2 parameters

Parameters	SVM		KNN	
	FP	FN	FP	FN
Mean	1.04	0.04	1.88	1.12
Median	1	0	2	1
Max	5	1	5	3
Min	0	0	0	0

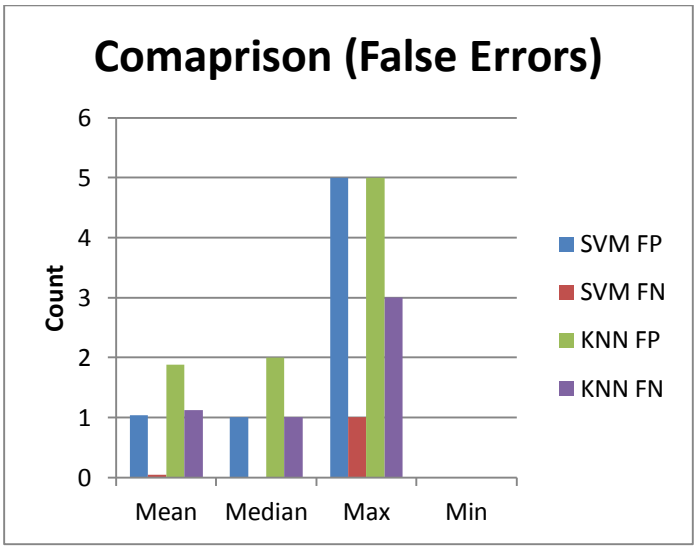

Figure 3: Comparative Analysis of KNN and SVM based on type 2 parameters

The SVM shows the higher performance than KNN on the basis of accuracy, precision and recall based parameters. The following table shows the higher mean for all accuracy, precision and recall than KNN, which clearly signifies the higher performance.

Table 3: Comparative Analysis of KNN and SVM based on Accuracy, Precision and Recall

Parameters	SVM			KNN		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Mean	98.92	98.37481	99.93548	97	96.94921	98.18315
Median	99	98.50746	100	97	97.10145	98.4375
Max	100	100	100	100	100	100
Min	95	92.64706	98.3871	94	91.22807	94.91525

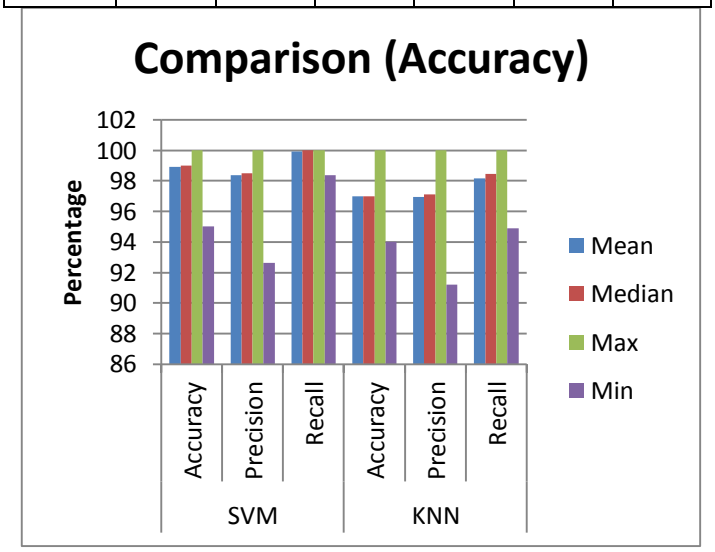

Figure 4: Comparative Analysis of KNN and SVM based on Accuracy, Precision and Recall

VIII. CONCLUSION

The support vector machine (SVM) and k-nearest neighbor (KNN) based models are used to predict the chronic kidney diseases. The dataset includes 35 features, out of which some of them are quantitative (both continuous and discrete) and other all categorical (qualitative). Different modules are developed to handle the categorical and quantitative variables in the dataset to avoid the problems related to the column

dominance, execution errors, etc. The quantitative variables undergo the maximum minimum scaling, which is known to convert the data values to 0-1 scale. In this thesis, the SVM has been found better than KNN on the basis of nearly all of the parameters. Also SVM outperformed KNN on the basis of precision by (98%) and recall (99%) against 97% (precision) and 98% (recall)**.** In the future, the deep learning classification can be utilized to improve the overall classification performance. Also, the optimization algorithms can be used to create the more balanced and advanced feature descriptors to obtain the higher accuracy.

REFERENCES

1. Agrawal R, Srikant R.," Privacy-preserving data mining"Proceedings of the ACM SIGMOD Inter-national Conference on Management of Data; Dallas,Texas; 2000, 439–450.
2. Atallah M., ElmagarmidA.,Ibrahim M. Bertino E., and VerykiosV.,"Disclosure limitation of sensitive rules," in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX'99. Washington, DC, USA: IEEE Computer Society, pp. 45–52, 1999.
3. Bertino E, Lin D, Jiang W.,"A survey of quantification of privacy preserving data mining algorithms, In "Privacy-Preserving Data Mining—Models and Algorithms. Advances in Database Systems.Springer, Berlin,34:183–205,2005
4. Dhutraj N.,Sasane S.,Kshirsagar V., "Hiding Sensitive Association Rule for Privacy Preservation", Institute of Electrical and Electronics Engineers Transactions on knowledge and data engineering,2013 .
5. Domadiya N.,RaoU.,"Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database", Institute of Electrical and Electronics Engineers,pp.1306-1310 ,2013.
6. Jain Y.,Yadav V., Panday G., "An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining", International Journal on Computer Science and Engineering , Vol. 3(7),pp-2792-2798,2011.
7. Jadav K.,Vania J., Patel D. (2013), "A Survey on Association Rule Hiding Methods", International Journal of Computer Applications, Vol. 82 (13), pp-20-25.
8. Kaur C., "Association Rule Mining using Apriori Algorithm: A Survey", International Journal of Advanced Research in Computer Engineering & Technology, Vol. 2(6), pp-893-900 ,2013.
9. Koh J. and Shieh S.(2004)," An Efficient Approach for Maintaining Association Rules Based on Adjusting FP-Tree Structures",in LNCS 2973, pp. 417–424, 2004.
10. Lan Q., Zhang .D,Wu B.,"A New Algorithm For Frequent Itemsets Mining", Institute of Electrical and Electronics Engineers,pp.360-364,2009.
11. Natarajan R.,Sugumar R., Mahendran M.,Anbazhagan K. , "Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1(7), pp.486-492, 2012.
12. Oliveira S., Zaiane O., " Privacy preserving frequent item set mining" in proceedings of the IEEE International Conference on Privacy, Security and Data Mining,Maebashi City, Japan, pp. 43–54, 2002.
13. Patel A., Rao U., PateeD.,"Privacy Preserving Association Rules in Unsecured Distributed Environment Using Cryptography",Institute of Electrical and Electronics Engineers IEEE-20180.
14. Patidar V., Raghuvanshi A., Shrivastava V. , "Literature Survey of Association Rule Based Techniques for Preserving Privacy", COMPUSOFT, An international journal of advanced computer technology, Vol. 2, pp-59-64,2013.
15. Saygin Y., Verykios V.S., Elmagarmid A., "Privacy preserving association rule mining." in RIDEInstitute of Electrical and Electronics Engineers Computer Society, pp.151–158,2002.
16. Saygin Y., VerykiosV., Clifton C., "Using unknowns to prevent discovery of association rules," SIGMOD Rec., Vol. 30(4), pp. 45 54, 2001.
17. Shah K., Thakur A., Ganatra A.,"A Study on Association Rule Hiding Approaches", International Journal of Engineering and Advanced Technology, Vol. 1(3), pp-72-76, 2012.
18. Thakur D. and Gupta H.,"An Exemplary Study of Privacy Preserving Association Rule Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering ,Vol.3(11), pp.2081-2084.
19. Vaidya J., and Clifton C., "Privacy preserving association rule mining in vertically partitioned data," in proceeding international Conference Knowledge Discovery and Data Mining, pp. 639–644, 2002.
20. Wang Y, Wu X.," Approximate inverse frequent item set mining: privacy, complexity, and approximation" in proceedings of the Institute of Electrical and Electronics Engineers, International Conference on Data Mining, 482–489, 2005.