Foundation University Journal of Engineering and Applied Sciences
Vol. 3, Issue 2, 2022.
DOI: 10.33897/fujeas.v3i2.565

Research Article

Heart Diseases Prediction and Diagnosis using Supervised Learning

Wajiha Safat[1], and Ijaz Hussain[2*]

[1]Department of Computer Science, CUI Islamabad, Pakistan.
[2]Department of Computer and Information Sciences, PIEAS Islamabad, Pakistan.
[*]Corresponding Author: Dr. Ijaz Hussain, Email: ijazhussain@pieas.edu.pk

Abstract:

The existing data for clinical diagnosis are often enlarged, but available tools are not efficient enough for decision making. Data mining techniques provide a user-oriented approach for clinical diagnosis and reduce risk factors. To improve clinical diagnosis, particularly for Cardiovascular diseases, nine different data mining techniques have been applied for classification and clustering. We compare all these techniques for better prediction. Despite all recent research efforts, the literature lacks the application of multiple techniques on multiple data sets for Cardiovascular disease prediction, which helps in decision making. In particular, this study is the augmentation of techniques for multiple data analysis by comparing four datasets with 14 attributes and a different number of instances. Another challenge is how to increase the accuracy of the decision-making process. Our research findings predict the better accuracy by using SMO and classification via regression for all data sets which shows the significant difference. Consequently, this research further helps to integrate the clinical decision support, thereby reducing medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient recovery.

Keywords: Data mining; Classification techniques; Cardiovascular disease prediction

1. Introduction

Data is generally gathered from various sectors like e-business, marketing, health and other industries to predict useful information for future assessments (El-Hasnony et al., 2022). Given that, raw data have been usually heterogeneous and thus difficult to understand. The available data is quite enriched and analysis tools are not efficient enough for decision making. Since data mining techniques are used to extract meaningful information for future estimates, therefore, important for research and development process (*EBSCOhost | 124636309 | A Descriptive Study of Predictive Models of MERS-CoV Outbreak.*, n.d.). The process of automatic creation of useful information from large data repository is called data mining. A term KDD (Knowledge discovery in databases) is generally used in data mining for decision making, where raw data is transformed into useful information (Palaniappan & Awang, 2008). The artificial intelligence, machine learning, databases, statistics, and pattern recognition are also the core of data mining (*An Overview of Knowledge Discovery Database and Data... - Google Scholar*, n.d.). Nonetheless, data mining involves multiple methods to accomplish different tasks. Intelligent methods are being utilized as an essential data mining process to extract data patterns and knowledge discovery (*EBSCOhost | 124636309 | A Descriptive Study of Predictive Models of MERS-CoV Outbreak.*, n.d.). All these methods attempt to fit a data into the model using different algorithms. The closer model is determined using these algorithms according to the charachteristics of data that are being examined (P. C. Chen et al., 2010). Medical predictors also use KDD to improve the quality of health services. Data mining techniques provide a user-oriented approach to discover hidden patterns that are further used for clinical diagnosis to reduce risk

factors (*EBSCOhost | 124636309 | A Descriptive Study of Predictive Models of MERS-CoV Outbreak.*, n.d.). Clinical diagnosis is viewed as an essential, but a complex job that should be executed precisely and legitimately. Developing an application to predict the outcome of diseases is the most interesting and challenging task in data mining. There is also a field in medical prognosis called survival analyses where various applications are used to deal with historical data in order to predict the survival of a particular patient suffering from a disease over a particular time period (J. Chen et al., 2009).

Other studies have been conducted for decision making on different diseases, which include Hepatitis, Lung Cancer, Liver Disorder, Breast Cancer, Diabetes and Thyroid disease etc. (*Cluster Analysis - Basic Concepts and Algorithms - Google Scholar*, n.d.). Despite all these researches, particularly for Cardiovascular disease prediction, literature lacks the implementation of multiple techniques which help in decision making (Soni Ujma Ansari Dipesh Sharma & Associate Professor, 2011), (Zhang et al., 2014), (Jindal et al., 2021). However, as far as recent literature is concerned, the available studies are limited in accordance with the comparison of multiple techniques on multiple data sets for better prediction. The major challenge is concerned with the accuracy of the decision-making process. In addition, according to the world health organization seventeen million deaths happen globally due to Cardiovascular diseases. Nevertheless, application of data mining techniques is still needed to be focused for Cardiovascular disease prediction. Therefore, our study, particularly focused on Cardiovascular diseases predictions and compared nine approaches; Decision trees J48, Naïve Bayes, REPTree, Decision table, Bayes net, classification via regression, bagging, Sequential Minimal Optimization (SMO) and K-means clustering using four data sets including Cleveland, Hungarian, VA Long Beach and Switzerland. This study is designed to compare all these methods against performance parameters for better prediction. Research findings predict the better accuracy by using SMO and classification via regression on all data sets.

The rest of the paper is organized as follows: Section 2 presents the proposed methodology along with summary of different techniques from the literature. Section 3 describes the results and discussion while Section 4 concludes the work.

2. Methodology

This section presents the extraction of significant patterns from the Cardiovascular disease data warehouse. The clinical data is the screening of patients affected by different heart problems (*World Health Organization: Death and Disability Due... - Google Scholar*, n.d.). Primarily the data are considered to improve the health standards and services. We took data from the UCI repository to ensure the efficient and explicit processes for mining, which is already preprocessed excluding VA Long Beach dataset. This study comprises three important phases: data understanding, data modeling, and results evaluation.

Data understanding identifies the preliminary insights about attributes and their definitions. In the modeling phase above, mentioned techniques are applied to the data sets to produce optimal values by using WEKA 3.8.1. In evaluation phase, results against performance parameters are compared. Figure 1 shows the proposed methodology framework.

2.1 Dataset

Our dataset contains the statistics of two major cities of Italy: Hungarian and Cleveland from the UCI machine learning repository. The first phase of methodology involves the understanding of preprocessed data which includes the attribute characteristics. Details of datasets are listed in Table 1. We implemented nine different models: eight models for classification and one for clustering. These models are specifically

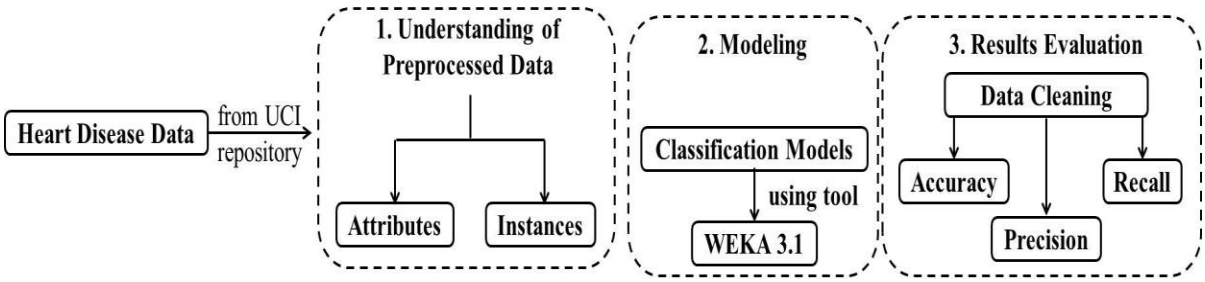


Figure 1: Proposed methodology

Table 1: Dataset details

Datasets	No of Attributes	No of instances
Cleveland	14	219
Hungarian	14	219
VA Long Beach	14	200
Switzerland	14	124

selected for their popularity and they also produce better average performance, according to the comparative studies that have been recently published in the literature (J. Chen et al., 2009). The classifiers, we used in this study evaluates that how good it is to predict the class of instances for which it is trained on. We applied classifiers on the training set by using Weka.

2.2 Summary of Techniques

Following is the detailed description of the mentioned techniques to seek deeper knowledge which is further applied in this study.

2.2.1 Naïve Bayes

Feature selection is a vital preprocessing technology to improve the efficiency, accuracy, and scalability of classifiers specifically in text classification. Usually, domain and algorithm characteristics are considered important for better feature selection. Feature selection is quite simple and efficient in a Naïve Bayes classifier, as it is highly sensitive to the results generated by using this technique for feature selection is highly significant (Ali et al., 2021). It provides fast and easy implementation, so it is used as a baseline for text classification. It suits best specifically when inputs have higher dimensions. The Naïve Bayes model uses the maximum likelihood criteria for parameter inference and performs better in complex real-world situations (El-Hasnony et al., 2022).

As a statistical classifier, the Naïve Bayes does not assume any dependency between attributes. It can work without using any Bayesian methods and can also produce better classification accuracy as compared to other algorithms (*EBSCOhost | 124636309 | A Descriptive Study of Predictive Models of MERS-CoV Outbreak.*, n.d.). The formula used for Naïve Byes (Augusto Gonçalves & Geraldo Pereira Barbosa, 2017) is described below:

$$\boldsymbol{P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}} \qquad \textbf{(1)}$$

Where $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*), $P(c)$ is the prior probability of *class*, $P(x|c)$ is the probability of *predictor* given *class*, $P(x)$ is the prior probability of the *predictor*.

2.2.2 REPTree

The basic principle of the REPTree algorithm is to calculate the information gain with entropy and use variance to reduce the occurrence of an error. REPTree algorithm helps to reduce the complexity of the decision tree model by using "reduced error pruning method" so the occurrence of error will be reduced from the variance. REPTree only sorts numeric attributes and builds a fast decision tree by using information gain (Platt, 1998).

2.2.3 Decision Tree

Decision Tree is the most widely used classifier which is easy to understand and configure as compared to other algorithms (Chaurasia & Pal, 2014). It uses divide and rule approach to form a data structure in the form of a tree. It uses supervised learning to structure a model with the set of divisions where local regions found recursively.

The general equation used for decision tree is given below where the information gain of X is calculated when Y is the conditional entropy:

$$\begin{aligned} &\boldsymbol{(Y) = -\Sigma(Y = yi)\; logP(Y = yi) k\; i = 1} \\ &\boldsymbol{(Y|X) = -\Sigma(X = xi)\;(Y|X = xi) l\; i = 1} \\ &\boldsymbol{(Y; X) = (Y) - (Y|X)} \end{aligned} \qquad \textbf{(2)}$$

There are two types of pruning in decision tree: pre-pruning and post-pruning. In contrast to each other pre-pruning produces faster trees and post-pruning produce more successful tree (Platt, 1998). Decision Tree stills have a problem of redundancy so necessary steps should be taken to resolve the replication and repetition (Mathuria, 2013).

2.2.4 Decision Tree J48

Decision tree J48 is developed by WEKA team and implemented by ID3 (Iterative Dichotomiser 3) algorithm. Derivation of rules, decision tree pruning, and missing values are the additional features of J48. This algorithm can be used for precision in case of overfitting pruning. Usually, the classification algorithm performs pruning until the best possible classification of data is done. The objective of this algorithm is to generate rules for data identification and generalize the decision tree until it meets the accuracy.

The disadvantage of the J48 algorithm is that the size of the tree increases linearly with the number of examples which increases the complexity. Consequently, tree depth is linked with tree size which cannot be greater than the number of attributes. When the depth of tree increases; space complexity, occurs to stores the values in the array and rules get slow down for large and noisy datasets (Alam & Pachauri, 2017).

$$\textbf{Entropy } \boldsymbol{E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i} \qquad \textbf{(3)}$$

$$\textbf{Information gain Gain } \boldsymbol{(T, X) = Entropy(T) - Rntropy(T, X)} \qquad \textbf{(4)}$$

2.2.5 Bayes Network

Probabilistic models for the variables of interest can be encoded graphically by using Bayesian networks. It is more efficient when multiple statistical techniques are merged to model graphically. Bayesian network provides an adequate graphical model even though some data entries are missing and variables have dependencies among each other.

The Bayesian network helps to seek a better understanding of difficult domains and casual relationships. This model gives the best representation of data by combining data and prior knowledge by using probabilistic semantics. Over fitting of data can be avoided efficiently by combining Bayesian networks and Bayesian statistical methods. The formula that is commonly used for the Bayes network is defined below:

$$\boldsymbol{P(C=T|A=T)} = \frac{\boldsymbol{(P(C=T,\ A=T)}}{\boldsymbol{P(A=T)}} \quad \textbf{(5)}$$

2.2.6 Classification via Regression

The Classification via Regression comprises of three major levels that involve encoding, linear regression and decoding. Multivariate adaptive regression splines and kernel tricks are the strategies for conventional adaptive non-parametric regression to apply on the nonlinear extension. To encode class label, the particular scoring scheme is used in literature named as optimal scoring. Average of squared regression residuals can be minimized by optimal scoring. This regression technique is efficient to extract low dimensional features (Chaurasia et al., n.d.).

2.2.7 Bagging

Bagging is basically a meta-algorithm which is designed to improve the accuracy and scalability of machine learning algorithms. It is mostly used in statistical classification and regression which helps to avoid over fitting and reduces variance. It can be used with any type of technique, but mostly applied to decision tree technique. It is also called Bootstrap aggregating (Sankar et.al, 2014).

It is an alternate to cross-validation method which is the mixture of different models. The results are generated in the form of different combinations of training data based on the Bootstrap method after learning weak training data. Bagging is a voting method which helps to generate multiple instances from a single sample. It uses the displacement method to produce multiple instances from the original sample (Platt, 1998). The formula for bagging is formulated as follows:

$$\underbrace{\boldsymbol{E}[[\boldsymbol{h_D}(\boldsymbol{x})-\boldsymbol{y})^2]}_{\textbf{Error}} = \underbrace{\boldsymbol{E}\left[\left(\boldsymbol{h_D}(\boldsymbol{x})-\overline{\boldsymbol{h}}(\boldsymbol{x})\right)^2\right]}_{\textbf{Variance}} + \underbrace{\boldsymbol{E}\left[\left(\overline{\boldsymbol{h}}(\boldsymbol{x})-\overline{\boldsymbol{y}}(\boldsymbol{x})\right)^2\right]}_{\textbf{Bias}} + \underbrace{\boldsymbol{E}[(\overline{\boldsymbol{y}}(\boldsymbol{x})-\boldsymbol{y}(\boldsymbol{x}))^2)]}_{\textbf{Noise}} \quad \textbf{(6)}$$

2.2.8 SMO

Sequential Minimal Optimization (SMO) is an iterative algorithm used to solve problems of Quadratic Programming (QP), where QP problems occurred during the training process of a support vector machine. John Platt proposed this algorithm to resolve constraint optimization problems (Nirve et.al, 2013). The algorithm of SMO is described below:

$$\sum_{i=1}^{n} \boldsymbol{a_i} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{y_i y_j K(x_i, x_j) a_i a_j}\ ,$$

$$\textit{Subject to}$$
$$0 \leq a_i \leq C, for\ i = 1.2, \ldots n,$$
$$\sum_{i=1}^{n} y_i a_i = 0 \qquad (7)$$

Where C is an SVM hyper-parameter and $K(x_i, x_j)$ is the kernel function, both supplied by the user; and the variables are Lagrange multipliers.

2.2.9 K-means Clustering

K-means is an unsupervised learning where available data have no specified groups or categories. This algorithm assigns data point to each K in the group, according to the features decided iteratively. Data points are decided according to the similarity of data which are in the form of clusters. Clustering allows finding and analyzing the groups that are formed spontaneously rather than assuming the groups before observing the data. The resulting groups are defined by the clusters and each cluster has a centroid which shows the collection of feature values. These feature values are used as the weights for centroids which further interpret the quality of cluster represented in the group (Www et al., 2008).

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets. Formally, the objective is to find:

$$argmin_s \sum_{i=1}^{k} c \sum_{x \in S_i} ||x - \mu_i||^2 = argmin_s \sum_{i=1}^{k} |S_i| Var S_i \qquad (8)$$

Where μ_i is the mean of points in S_i.

3. Results and Discussion

Third phase of the study involves the evaluation of results for all cities, after applying mentioned techniques for classifications and clustering. Following are the performances of all data sets with their summaries, evaluation criteria and confusion matrix.

3.1 Performance Study of Cleveland City

Table 2 shows the performance summary of these techniques for Cleveland city.

Table 2: Performance summary of classifiers for Cleveland city

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPTree
Correctly Classified Instances	205	199	219	175	219	141	177	121
Incorrectly Classified Instances	14	20	0	44	0	78	42	98
Kappa Statistics	0.898	0.854	1	0.667	1	0.339	0.665	0
Mean Absolute Error	0.050	0.074	0.24	0.131	0.038	0.241	0.107	0.2541
Root Mean Squared Error	0.140	0.173	0.315	0.2359	0.095	0.329	0.231	0.3565
Relative Absolute Error (%)	19.680	29.11	93.899	51.237	15.228	94.449	42.04	99.424
Root Relative Squared Error (%)	39.313	48.518	88.489	66.169	26.693	92.411	65.03	99.993
Predictive Accuracy (%)	93.607	90.87	100	79.908	100	64.383	80.82	55.251
Time to Build Model (sec)	0	0.01	0.08	0	0.02	0	0.02	0

Table 3 shows the evaluation criteria for Cleveland City after applying all techniques.

Table 3: Comparison of estimates under certain evaluation criteria for Cleveland city

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPTree
TP Rate	0.936	0.909	1	0.799	1	0.644	0.808	0.553
FP Rate	0.050	0.069	1	0.155	1	0.303	0.193	0.553
Precision	0.936	0.908	1	0.797	1	0.565	0.823	0.305
Recall	0.936	0.909	1	0.799	1	0.644	0.808	0.553
F-Measure	0.935	0.907	1	0.793	1	0.575	0.793	0.393
MCC	0.890	0.849	1	0.677	1	0.404	0.698	0.000
ROC Area	0.989	0.979	1	0.955	1	0.810	0.905	0.500
PRC Area	0.983	0.964	1	0.890	1	0.615	0.811	0.365

3.2 Performance Study of Hungarian City

Table 4 shows the performance of different techniques on Hungarian city dataset.

Table 4: Performance summary of classifiers for Hungarian city

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPT ree
Correctly Classified Instances	191	180	219	180	217	149	180	141
Incorrectly Classified Instances	28	39	0	39	2	70	39	78
Kappa Statistics	0.768	0.661	1	0.657	0.983	0.253	0.625	0
Mean Absolute Error	0.070	0.090	0.24	0.115	0.040	0.191	0.110	0.219
Root Mean Squared Error	0.189	0.221	0.315	0.220	0.105	0.304	0.235	0.331
Relative Absolute Error (%)	31.929	40.766	108.06	52.200	18.430	86.414	49.819	98.993
Root Relative Squared error (%)	57.202	66.865	95.252	66.418	31.688	91.897	70.932	99.988
Predictive Accuracy (%)	87.214	82.191	100	82.191	99.086	68.036	82.191	64.383

Table 5 shows the evaluation criteria for Hungarian city.

Table 5: Comparison of estimates under certain evaluation criteria for Hungarian city

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPTree
TP Rate	0.872	0.822	1	0.822	0.991	0.680	0.822	0.644
FP Rate	0.108	0.183	1	0.184	0.001	0.453	0.252	0.644
Precision	0.874	0.820	1	0.823	0.992	0.525	0.826	0.415
Recall	0.872	0.822	1	0.822	0.991	0.680	0.822	0.644
F-Measure	0.872	0.817	1	0.815	0.991	0.590	0.803	0.504
MCC	0.762	0.665	1	0.672	0.986	0.309	0.676	0
ROC Area	0.971	0.954	1	0.964	1	0.734	0.841	0.500
PRC Area	0.959	0.924	1	0.909	0.998	0.594	0.768	0.450

3.3 Performance Study of VA Long Beach City

Table 6 shows the performance of these techniques on VA Long Beach data set.

Table 6: Performance summary of classifiers for VA Long Beach

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPTree
Correctly Classified Instances	162	152	199	165	196	72	59	59
Incorrectly Classified Instances	38	48	1	35	4	128	141	141
Kappa Statistics	0.753	0.686	0.993	0.771	0.974	0.113	0	0
Mean Absolute Error	0.096	0.125	0.222	0.133	0.054	0.251	0.256	0.256
Root Mean Squared Error	0.211	0.241	0.310	0.226	0.121	0.352	0.358	0.358
Relative Absolute Error (%)	37.287	48.86	86.38	51.976	21.018	97.771	99.76	99.762
Root Relative Squared Error (%)	58.908	67.25	86.60	63.285	33.764	98.363	99.99	99.996
Predictive Accuracy (%)	81	76	95.5	82.5	98	36	29.5	29.5
Time to Build Model (sec)	0.03	0.03	0.08	0.13	1.19	0.11	0.01	0

Table 7 shows the evaluation criteria for VA Long Beach after applying all techniques:

Table 7: Comparison of estimates under certain evaluation criteria for VA long beach

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REPTree
TP Rate	0.810	0.760	0.995	0.825	0.980	0.360	0.295	0.295
FP Rate	0.058	0.075	0.002	0.054	0.005	0.248	0.295	0.295
Precision	0.831	0.780	0.995	0.825	0.980	0.197	0.087	0.087
Recall	0.810	0.760	0.995	0.825	0.980	0.360	0.295	0.295
F-Measure	0.815	0.762	0.995	0.824	0.980	0.252	0.134	0.134
MCC	0.761	0.695	0.993	0.773	0.974	0.104	0.000	0.000
ROC Area	0.962	0.936	0.998	0.969	0.999	0.593	0.500	0.500
PRC Area	0.916	0.873	0.998	0.890	0.998	0.286	0.230	0.230

3.4 Performance Study of Switzerland City

Table 8 shows the performance of these techniques on data set of Switzerland.

Table 8: Performance summary of classifiers for Switzerland

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification via Regression	Decision Table	J48	REP Tree
Correctly Classified Instances	105	99	122	82	118	56	49	49
Incorrectly Classified Instances	19	25	2	42	6	68	75	75
Kappa Statistics	0.781	0.710	0.977	0.515	0.931	0.157	0	0
Mean Absolute Error	0.114	0.146	0.240	0.175	0.080	0.283	0.285	0.285
Root Mean Squared Error	0.208	0.241	0.316	0.281	0.159	0.372	0.377	0.377
Relative Absolute Error (%)	39.964	51.026	83.988	61.318	28.178	99.023	99.528	99.53
Root Relative Squared Error (%)	55.314	63.998	83.845	74.593	42.310	98.580	99.990	99.99
Predictive Accuracy (%)	84.677	79.838	98.387	66.129	95.161	45.161	39.516	39.52
Time to Build Model (sec)	0.04	0.07	0.02	0.17	2.04	0.34	0.4	0.13

Table 9 shows the evaluation criteria for Switzerland after applying all the techniques:

Table 9: Comparison of estimates under certain evaluation criteria for VA long beach

Evaluation Criteria	Bayes Network	Naïve Bayes	SMO	Bagging	Classification Via Regression	Decision Table	J48	REPTree
TP Rate	0.810	0.760	0.995	0.825	0.980	0.360	0.295	0.295
FP Rate	0.058	0.075	0.002	0.054	0.005	0.248	0.295	0.295
Precision	0.831	0.780	0.995	0.825	0.980	0.197	0.087	0.087
Recall	0.810	0.760	0.995	0.825	0.980	0.360	0.295	0.295
F-Measure	0.815	0.762	0.995	0.824	0.980	0.252	0.134	0.134
MCC	0.761	0.695	0.993	0.773	0.974	0.104	0.000	0.000
ROC Area	0.962	0.936	0.998	0.969	0.999	0.593	0.500	0.500
PRC Area	0.916	0.873	0.998	0.890	0.998	0.286	0.230	0.230

3.5 K-means Clustering

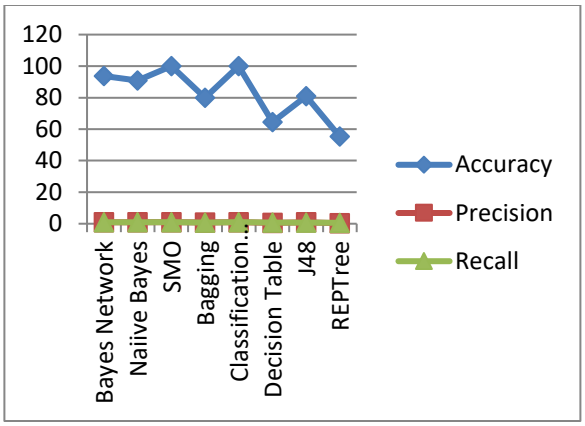
The clustering model describes the statistics assigned to different centroids of the cluster which includes the number and percentage of instances. The characteristics of cluster depend on the values represented by each centroid which includes mean vectors and dimension values (Kapoor et al., 2017). Every new cluster has cluster instances which describes what number of instances are generated in that cluster. Table 10 shows the summary of cluster centroids and cluster instances of all datasets including Cleveland, Hungarian, VA Long Beach, and Switzerland.

Table 10: Final summary of cluster centroids.

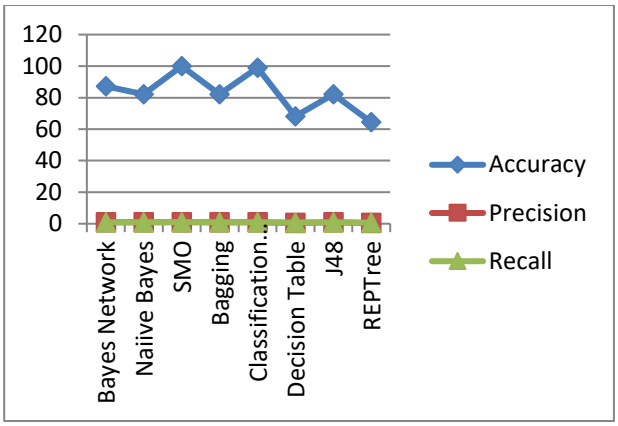
K-mean Clustering	Cleveland	Hungarian	VA Long Beach	Switzerland
Number of Iterations	3	3	7	3
Within Cluster Sum of Squared Error	1651.0	1179.0	1273.0	642.0
Time to Build Model	0.01 sec	0.01 sec	0.22	0.07
Clustered instances at 0	139 (63%)	74 (34%)	118 (59%)	97 (78%)
Clustered Instances at 1	80 (37%)	145 (66%)	82 (41%)	27 (22%)

3.4 Comparison

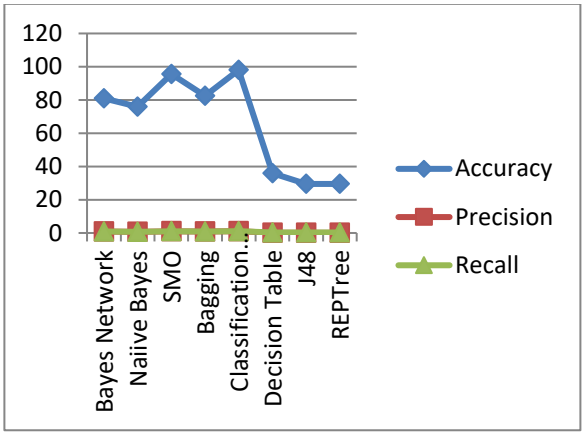
Figure 2 shows the performance of all four cities with respect to parameters including accuracy, precision, and recall after applying classification techniques.



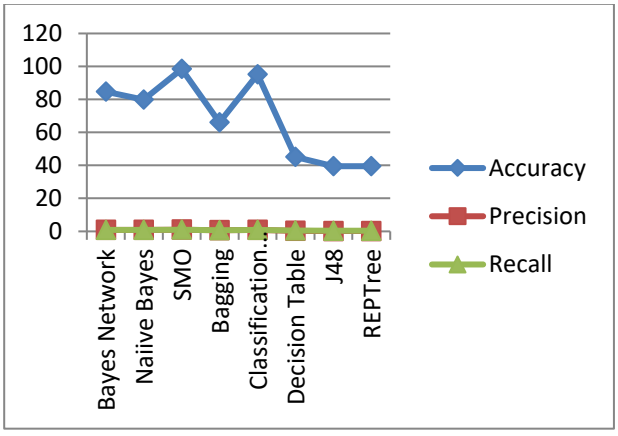
(a) Cleveland



(b) Hungarian



(c) VA Long Beach



(d) Switzerland

Figure 2: Performance visualization of all datasets (Cleveland, Hungarian, VA Long Beach and Switzerland).

4. Conclusion

The objective of this study is to compare all these methods against performance parameters including precision, recall, and accuracy for better prediction of Cardiovascular disease. Existing techniques are limited in accordance with the comparison of multiple techniques on multiple data sets. Our study is the augmentation of techniques for multiple dataset analysis where nine multiple techniques are applied to a greater number of instances and attributes. We use multiple instances on all four datasets--Cleveland, Hungarian, VA Long Beach, and Switzerland to evaluate the precision, recall, and accuracy of these techniques.

Results depict that the accuracy of all cities by using SMO and classification via regression is more than 95%. Accuracy by applying SMO shows the dramatic difference as no other research shows this kind of difference using SMO. While Byes Network and Naïve Bayes are around 80% to 90% in all cities. Whereas accuracy J48 and REPTree are showing approximately the same results for all cities. Bagging is about 80% for all cities, whereas Decision Tree shows poor performance in the whole scenario. Precision and recall are approaching to 1 for SMO, while REPTree shows the lowest for all. Furthermore, a comparison of all

these techniques helps to integrate the clinical decision support that could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient recovery.

In future, this study can be enhanced by adding the automatic prediction of other diseases instead of the heart. Other data mining techniques can be incorporated using the same data set such as time series, fuzzy sets, and rule-based association. In addition, we want to look at how various preprocessing methods affect clustering algorithms. Also, producing datasets with a missing values rate of more than 20% will be taken into consideration to identify the optimal preprocessing methods to use for such datasets.

References

Alam, F., & Pachauri, S. (2017). *Comparative Study of J 48, Naive Bayes and OneR Classification Technique for Credit Card Fraud Detection using WEKA*.

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/J.COMPBIOMED.2021.104672

An overview of knowledge discovery database and data... - Google Scholar. (n.d.). Retrieved March 10, 2023, from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=An+overview+of+knowledge+discovery+database+and+data+mining+techniques&btnG=

Augusto Gonçalves, A., & Geraldo Pereira Barbosa, J. (2017). *The Development of an ICT framework for Business Intelligence at the Brazilian national Cancer Institute: a Case study of organizational learning and Innovation*. *10*, 2017–2551. https://doi.org/10.5902/19834659

Chaurasia, V., and, S. P.-C. J. of S., & 2013, undefined. (n.d.). Early prediction of heart diseases using data mining techniques. *Papers.Ssrn.Com*. Retrieved March 9, 2023, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2991237

Chaurasia, V., & Pal, S. (2014). Data Mining Approach to Detect Heart Diseases. *Http://Ljournal.Ru/Wp-Content/Uploads/2016/08/d-2016-154.Pdf*. https://doi.org/10.18411/D-2016-154

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, *36*(3), 5432–5435. https://doi.org/10.1016/J.ESWA.2008.06.054

Chen, P. C., Lee, K. Y., Lee, T. J., Lee, Y. J., & Huang, S. Y. (2010). Multiclass support vector classification via coding and regression. *Neurocomputing*, *73*(7–9), 1501–1512. https://doi.org/10.1016/J.NEUCOM.2009.11.005

Cluster analysis - Basic concepts and algorithms - Google Scholar. (n.d.). Retrieved March 10, 2023, from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Cluster+analysis+-+Basic+concepts+and+algorithms&btnG=

EBSCOhost | 124636309 | A Descriptive Study of Predictive Models of MERS-CoV Outbreak. (n.d.). Retrieved March 10, 2023, from https://web.p.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09765697&AN=124636309&h=oimqXYKlYiZhFFV4zUa%2f%2fPqDsC3X2utoDiW8olp4RR8iNaaQLns0W4nUecdJXgVmOyN5C38A3sdNFiYNB716eg%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d09765697%26AN%3d124636309

El-Hasnony, I. M., Elzeki, O. M., Alshehri, A., & Salem, H. (2022). Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors (Basel, Switzerland)*, *22*(3). https://doi.org/10.3390/S22031184

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012072. https://doi.org/10.1088/1757-899X/1022/1/012072

Kapoor, P., Arora, D., & Kumar, A. (2017). Effects of mean metric value over CK metrics distribution towards improved software fault predictions. *Advances in Intelligent Systems and Computing*, *553*, 57–71. https://doi.org/10.1007/978-981-10-3770-2_6/COVER

Mathuria, M. (2013). Decision Tree Analysis on J48 Algorithm for Data Mining. *Intrenational Journal of Advanced Research in Computer Science and Software Engineering*. https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*, 108–115. https://doi.org/10.1109/AICCSA.2008.4493524

Platt, J. C. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/

Soni Ujma Ansari Dipesh Sharma, J., & Associate Professor, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction Sunita Soni. *International Journal of Computer Applications*, *17*(8), 975–8887.

World Health Organization: Death and disability due... - Google Scholar. (n.d.). Retrieved March 10, 2023, from https://scholar.google.com/scholar?cluster=14506513145034282120&hl=en&as_sdt=2005&sciodt=0,5

Www, W.:, Sawant, A. A., & Chawan, P. M. (2008). International Journal of Emerging Technology and Advanced Engineering Comparison of Data Mining Techniques used for Financial Data Analysis. *Certified Journal*, *9001*(6). www.ijetae.com

Zhang, W., Montewka, J., & Goerlandt, F. (2014). Semi-qualitative method for ship collision risk assessment. *Safety and Reliability: Methodology and Applications*, 1563–1572. https://doi.org/10.1201/b17399-216/semi-qualitative-method-ship-collision-risk-assessment-zhang-montewka-goerlandt