International Journal of Scientific and Innovative Mathematical Research (IJSIMR)
Volume 7, Issue 6, 2019, PP 30-37
ISSN No. (Print) 2347-307X & ISSN No. (Online) 2347-3142
DOI: http://dx.doi.org/10.20431/2347-3142.0706005
www.arcjournals.org

Analysis of Master Health Checkup Data Using Data Mining Classification and Cross Validation

R. Lakshmi Priya[1], G. Manimannan[2*], N. Manjula Devi[3]

[1]*Assistant Professor, Department of Statistics, Dr. Ambedkar Govt. Arts College, Chennai*

[2]*Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai*

[3]*Post Graduate Student, Department of Bio-Statistics, ICMR-NIE, Chennai*

Corresponding Author:** ***G. Manimannan, *Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai*

Abstract: *This research paper attempts to identify the Blood pressure based BP Systolic and BP Diastolic data and to cross validate the changes of Blood pressure using various machine learning method. The data were collected from secondary source containing 460 patients. The case sheet deals with demographic characteristics, Blood Pressure, Fat, Liver and diabetic parameters. This study concentrates on age, BP Systolic and diastolic only. Machine learning methods such as Logistic Regression, Support Vector Machine and Random Forest Model were used as data mining tools to explore the classification model and to cross validate the present dataset. All the three classification models were applied and extracted. Area under the Curve, Classification Accuracy, F1 Score, Precision and Recall are all closer to unity. The above measures shows three major categories of classification based on Blood pressure parameters. Machine learning methods achieved best model and are labeled as Normal, Elevated and Hypertension. The results of the present study indicate that the machine Learning Data Mining Tools can be used as a feasible tool for the analysis of large set of Blood pressure data. Finally, the three model classification is visualized using Silhouette plot.*

Keywords: *Master Health Checkup, Blood Pressure, Machine Learning Methods and Silhouette plot*

1. Introduction

The Master Health Check-up (MHC) offered by various hospitals and medical research institutes is a programme that attempts to reduce health care costs by prevention and early diagnosis. A variety of chronic diseases afflict us, most of which take their toll after the fifth decade of life. Diabetes, hypertension, heart attacks, stroke and cancer are some of the more common examples. Almost all of these problems first go through a long quiescent phase where they produce no symptoms. This period can be as long as 10 - 20 years. It makes sense, therefore, that a programme which attempts to detect and correct these problems during this silent phase will decrease the ultimate morbidity from all these diseases. In the early days of preventive health check-ups, every conceivable test and technology was ordered in the hope that some would be abnormal and provide an avenue of approach. A handful of items, mostly simple, appear to provide the greatest value. The MHC offered at various hospitals and institutes is a carefully constructed programme that offers a panel of tests that are proven to be valuable. As an incentive to those who have taken the efforts to control their health problems, the programme also includes two or more follow up visits within a year of MHC and change in physician of the check up. Good health is by itself of great value. It enhances market earnings by increasing the number of healthy days an individual has available for work (Grossman 1972) and increases nonmarket productivity, allowing more time for household production (Becker 1976). A health checkup helps to secure and maintain good health.

2. Review of Literature

The Master Health Check (MHC) is a series of tests to screen each functional area closely to detect even the smallest symptom of a major illness. It also helps to identify the reason for minor ailments, which are constant. MHC is considered to be the most comprehensive prevention check. Master Health Check consists of five permanent packages, which are as follows: Master Health Check, Executive Health

Check, Heart Check, Whole Body Check and Well Women Check up [1] B. Krishan Reddy, G.V.R.K. Acharyulu , 2002). In the present context the problem of MHC patients has been studied, without making any assumptions with regard to the number of groups or any other structural patterns in advance, which reflected the classification of patients based on certain medical observations (G. Manimannan, S. Hari and G. Vijaythiraviyam) [2]. The main objective of the present study is to identify the structural data mining model, classification, cross validation and visualization of BP Systolic and BP Diastolic data in Master Health Checkup data using different data mining techniques: (a) To develop data mining model and identify the pattern of BP Systolic and BP Diastolic data in the study period using Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) methods. (b) To identify the final test score, classification and cross validation of BP Systolic and BP Diastolic data using the above methods. (c) Finally, to visualize the BP Systolic and BP Diastolic based on extracted test score using Silhouette plot

3. Methodology and Database

This section brings out the discussion of the database, the MHC (Master Health Checkup) parameters selected and the Data Mining Techniques. The MHC data were collected from secondary source of OPD (Out Patients Department) containing 295 patients in Private Multi Specialty Hospital. Johns Hospital, Bangalore was considered as the database. The data mainly consists of five major categories, such as socio economic and demographic characteristic, Blood Pressure, Fat, Liver and diabetic related parameters. Among the listed patients, number of patients varied over the study period owing to removal of those patients for which the required data are not available or outliers.

3.1 Selection of Variables

In this study, BP Systolic and BP Diastolic medical parameters were chosen among the many that had been used in MHC case sheets. These BP Systolic and BP Diastolic medical observations were chosen to Blood Pressure, Fat, Liver and diabetic. Some of them are given below.

Table1. *MHC Medical observations during the study period*

Parameters	Description
BP_Syst	Blood Pressure Systolic
BP_Dias	Blood Pressure diastolic
Blood_Hb	haemoglobin
Blood_PCV	Packed Cell Volume
Blood_TC	Total Count
Diabetess_Fasting	Diabetes Fasting
Diabetes_Post Pran	Diabetes Post Prandial
Cholesterol	Cholesterol
FAT_VLD	High-density lipoprotein
FAT_LDL	Low-density lipoprotein
Liver_SAP	Alkaline phosphate
Liver_ALT	Alanine transaminase

4. Methodology

4.1. Data Mining

Data Mining Technique is an interdisciplinary field, the confluence of a set of disciplines in the following heads including database systems, Statistics, machine learning, visualization and information science. Although data mining is a new term, the technology is not. Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. In the present context data mining exhibits the patterns and cross validation by applying few techniques namely, logistic regression, random forest method and support vector machine rule. As such KDD is an iterative process, which mainly consist of the following steps: Step 1: Data cleaning; Step 2: Data Integration; Step 3: Data selection and transformation; Step 4: Data Mining and Step 5: Knowledge representation. Of the above iterative process Steps 4 and 5 are most important. If clever techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

In this research paper, the researcher uses orange data mining software. Orange is an open source machine learning and data visualization for learner as well as experts. Interactive data analysis work flows with a large toolbox that is available in this package. The software is developed with python script. Python is an interpreted high-level programming language for general-purpose programming and it was created by Guido van Rossum Guido van Rossum [3].

4.2. Logistic Regression

Regression analysis always requires numeric data, when attributes are categorical, the researcher have to change to numerical values to apply regression analysis [4]. Regression and classification are data mining techniques used to solve similar problems, but they are frequently confused. Both are used in prediction analysis, but regression is used to predict a numeric or continuous data while classification assigns data into discrete data. Classification is a data mining technique that assigns categories to a collection of data in order to aid in more accurate predictions and analysis. Classification is one of the several methods intended to make analysis of very large datasets effective.

The general model for logistic regression is $p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \Rightarrow \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$. The logistic distribution constrains the estimated probabilities to lie between 0 and 1.

4.3. Random Forest Model (RFM)

Random forest is a group learning method used for classification, regression and other tasks. It was first proposed by Tin Kam Ho and further developed by Leo Breiman [5] and Adele Cutler. **Random Forest** builds a set of decision trees. Each tree is developed from a bootstrap sample from the training data. When developing individual trees, an arbitrary subset of attributes is drawn (hence the term “Random”), from which the best attribute for the split is selected. The final model is based on the majority vote from individually developed trees in the forest. **Random Forest** works for both classification and regression tasks.

4.4. Support Vector Machine (SVM)

Support vector machine (SVM) is a machine learning technique that separates the attribute space with a hyper plane, thus maximizing the margin between the instances of different classes or class values. The technique often yields supreme predictive performance results. Orange embeds a popular implementation of SVM from the LIBSVM package. Support Vector Machine (SVM) is primarily a classifier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. For categorical variables a dummy variable is created with case values as either 0 or 1.

4.5. Cross Validation

The data set is split into training and test cases randomly. The centroids of the clusters from the training cases can be used to cluster the test cases. The centroids of the clusters formed by test data are then computed and compared with the training data. Comparable results validate the clustering that has achieved.

Orange Data Mining Classification Algorithm

4.6. Proposed Algorithm and Cross Validation with Various Models Work Flow

Step 1: Initially, the database of Master Health Checkup open through the file widget and connect through the work flow with Test score widget, after that Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) widget connecter with test score.

Step 2: The Test score widget assign the training and testing data sets with 10 folds cross validation. The training data sets are 70 percent of the original database using random sampling method up to target class.

Step 3: The data table widget connected to file widget for checking the original data.

Step 4: Repeat the database to train or test the database to reach best model with the help of step 1, change the training data and cross validation folds. Figure 1 represents the data mining work flow for various data mining Techniques, Confusion Matrix and Visualization of the database.

4.7. Logistic Regression Algorithm

Step 1: The logistic regression widget chooses from various machines learning method and connects to test score widget.

Step 2: Open logistic regression widget and select regularization type (Ridge L2 by default and widely used model).

Step 3: The logistic regression strength must always $C = 1$ is in the middle of the model and the remaining two extreme points of left and right side of the line are labeled as weak and strong strength.

Step 4: Repeat the step 3 with various folds and C value, to get the better model.

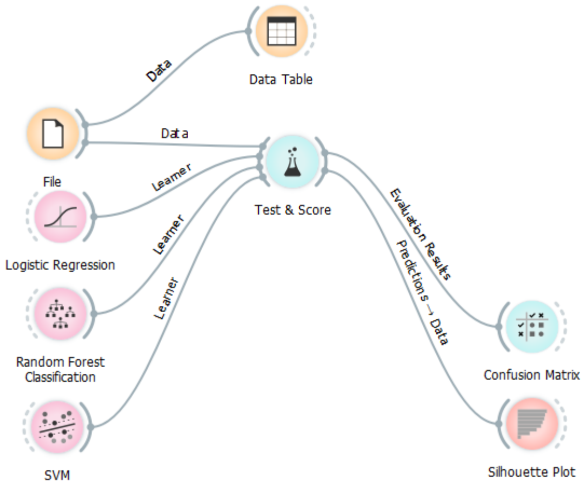


Figure1. *Work Flow diagram Models of classification, Cross Validation and Machine Learning Methods*

4.8. Random Forest Classification

Step 1: The Random Forest widget chooses from various machine learning method and will be connected to test score widget.

Step 2: Open Random Forest widget and select growth control model and number of trees are 10 from basic properties with number of attributes considered at each split at 5.

Step 3: Repeat the step 3 with various size of trees and splits to get a better model. .

4.9. Support Vector Machine

Step 1: The Support Vector Machine widget chooses from various machine learning method and will be connected to test score widget.

Step 2: Open Support Vector Machine (SVM) type and select SVM cost type $C = 1{,}00$, Regression loss epsilon $(\varepsilon) = 0{,}10$ and choose the menu of optimization parameter set to 0.0010 with iteration limit 100

Step 3: Repeat the step 2 with various cost, epsilon, optimization parameter and iteration limit, to get better model of SVM.

5. Result and Discussion

5.1. Test and Scoring Methods

In general, the test and scoring methods are AUC, CA (Classification Accuracy), F1, Precision and Recall measures are displayed in the output window. The definition of each measure is given below.

5.2. Area Under the Curve (AUC)

The Area under the curve is a performance metrics for a binary classification in data mining. By comparing the Receiver Operating Characteristic Curve with the area under the curve, or Area Under the Curve (AUC), it captures the extent to which the curve is up in the Northwest corner. The AUC score less than 0.5 is not a better random estimate. The measure of AUC with 0.9 would be a very good model, but a score of 0.999 would be the best model to be true and indicate correct model (Table 1).

5.3. Classification Accuracy (CA)

The classification accuracy with 1 is the model which is the best and 0 is the model as worst (Table 1), the following formula is used to calculate the CA measure based on Type I and Type II errors of statistics,

$$CA = \frac{true\ \ positive + true\ \ negative}{true\ \ \text{positive} + true\ \ negative + \ false\ positive + \ false\ negative} = \frac{tp + tn}{tp + tn + \ fp + \ fn}$$

5.4. F1 Score

In statistical study of binary classification, the F_1 score, alternatively named as (F-score or F-measure) a measure of a test's precision. The F_1 score considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier. The r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F_1 score is the harmonic mean of the precision and recall, where a F_1 score reaches its best model at 1 (that is perfect Precision and Recall score) and worst model at 0 (Table 1). The general formula for F_1 score is the harmonic average of p and r

$$F_1 = 2 * \frac{1}{\frac{1}{r} + \frac{1}{p}} \quad = 2 * \frac{p * r}{p + r}$$

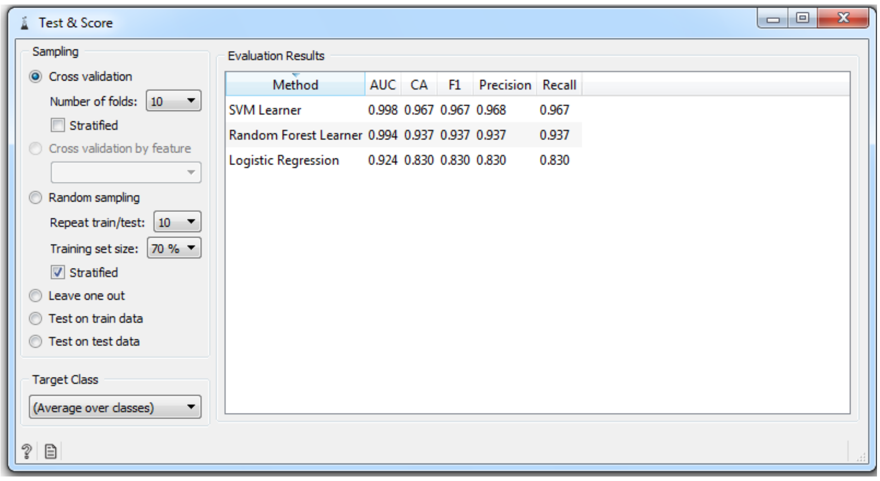
5.5. Precision of Test Score

The precision measure is the ratio of $P = \dfrac{true\ \ positive}{true\ \ positive\ + \ false\ positive} = \dfrac{tp}{tp\ + \ fp}$ where tp is the number of true positives and fp is the number of false positives. The precision is naturally the ability of classifier not to label as positive a sample that is negative. The precision value 1 is the best model and 0 is the worst model (Table 1).

5.6. Recall of test Score

Recall measure is the ratio of $R = \dfrac{true\ \ positive}{true\ \ positive\ + \ false\ \ negative} = \dfrac{tp}{tp\ + \ fn}$ where tp is the number of true positives and fn is the number of false negatives [8]. The precision is naturally the ability of classifier not to label as positive a sample that is negative. The precision value is 1 (Table 1). .

Analysis of Master Health Checkup Data Using Data Mining Classification and Cross Validation



Sampling

● Cross validation

Number of folds: 10

□ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 70 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Target Class

(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM Learner	0.998	0.967	0.967	0.968	0.967
Random Forest Learner	0.994	0.937	0.937	0.937	0.937
Logistic Regression	0.924	0.830	0.830	0.830	0.830

Table2. *Test Score Machine Learning Methods*

The above table shows the test and their scores of various data mining techniques. All the three methods of AUC, CA, F1, Precision and Recall values are closer to unity. The results achieved best model and cross validation of rainfall database. The classification and confusion matrix of three models are classified as 95 percent and above and the remaining five percent are misclassified (*Table 2 to 5*) due to NP Systolic and BP Diastolic and various MHC parameters. Based on this classification accuracy the BP Systolic and BP Diastolic database were classified and labeled as Normal, Elevated and Hypertension categories.

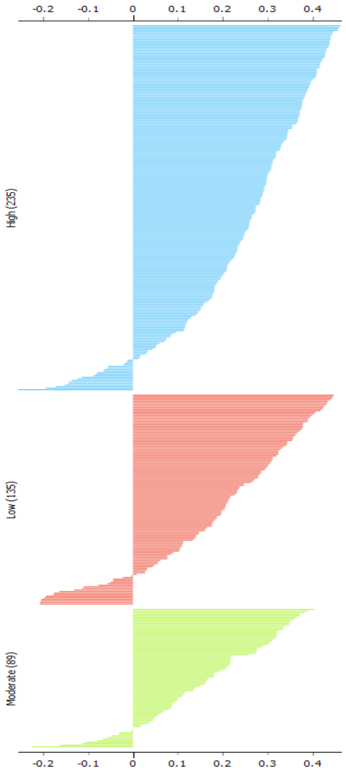


Figure2. *Silhouette Plot for Logistic Regression*

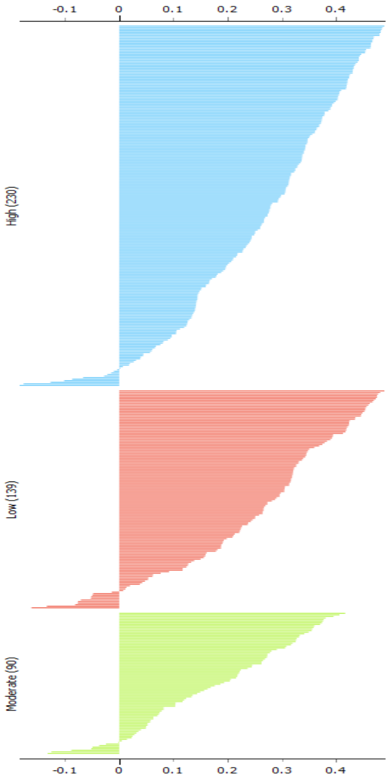


Figure3. *Silhouette Plot for Random Forest Model*

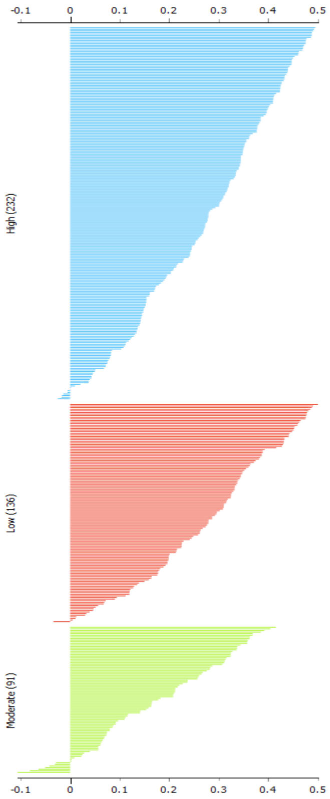
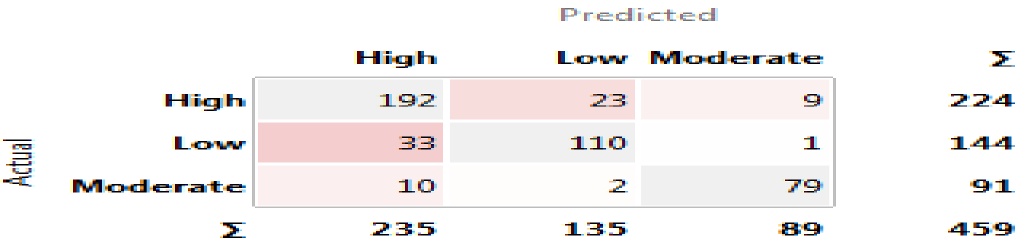


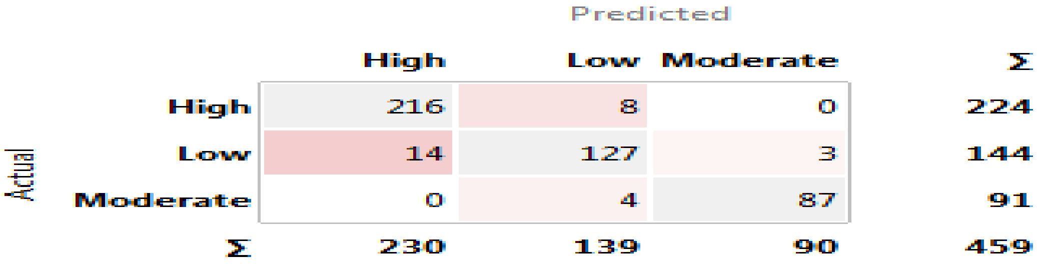
Figure4. *Silhouette Plot for Support Vector Machine*

Table3. *Confusion Matrix for Logistic Regression*



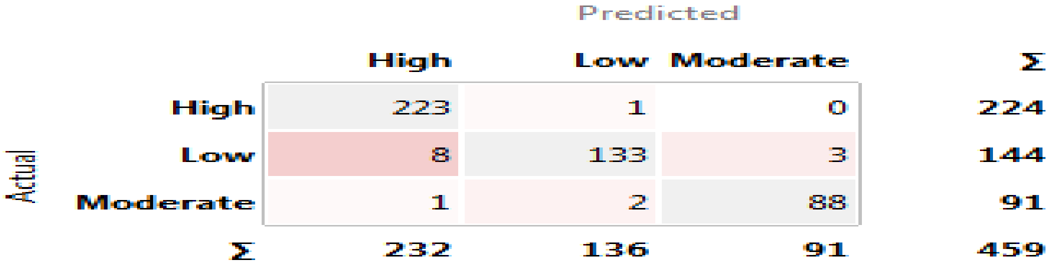
		Predicted			
		High	Low	Moderate	Σ
Actual	High	192	23	9	224
	Low	33	110	1	144
	Moderate	10	2	79	91
	Σ	235	135	89	459

Table4. *Confusion Matrix for Support Vector Machine*



		Predicted			
		High	Low	Moderate	Σ
Actual	High	216	8	0	224
	Low	14	127	3	144
	Moderate	0	4	87	91
	Σ	230	139	90	459

Table5. *Random Forest Model*



		Predicted			
		High	Low	Moderate	Σ
Actual	High	223	1	0	224
	Low	8	133	3	144
	Moderate	1	2	88	91
	Σ	232	136	91	459

The silhouette plot visualize (Figure 2 to 4) five categories of BP Systolic and BP Diastolic data from MHC Database and all the three methods has less than five percent misclassification due to MHC parameters of the study period. After performing data mining techniques of logistic regression, random forest model and support vector machine, the next stage is to assign initial classification based BP Systolic and BP Diastolic. Formations of classification are explored by considering 2- classification, 3- classification 4- classification and so on. Out of all the possible splits, 3- classification exhibited meaningful interpretation than two, four and higher splits. Having decided to consider only 3 classifications, it is possible to group the BP Systolic and BP Diastolic as Normal, Elevated and Hypertension classification depending on whether the MHC parameters belonged to classification 1, classification 2 and classification 3 respectively. In spite of incorporating the results of data mining for the study period, only the summary statistics are reported in Table 2 to 5. All the data mining methods attained best model test and score.

6. Conclusion

This research paper attempts to identify the Blood pressure based BP Systolic and BP Diastolic data and to cross validate the changes of Blood pressure using various machine learning method. The data were collected from secondary source containing 460 patients. The case sheet deals with demographic characteristics, Blood Pressure, Fat, Liver and diabetic parameters. This study concentrates on age, BP Systolic and diastolic only. Machine learning methods such as Logistic Regression, Support Vector Machine and Random Forest Model were used as data mining tools to explore the classification model and to cross validate the present dataset. All the three classification models were applied and extracted. Area under the Curve, Classification Accuracy, F1 Score, Precision and Recall are all closer to unity. The above measures shows three major categories of classification based on Blood pressure parameters. Machine learning methods achieved best model and are labeled as Normal, Elevated and Hypertension. The results of the present study indicate that the machine Learning Data Mining Tools can be used as a feasible tool for the analysis of large set of Blood pressure data. Finally, the three model classification is visualized using Silhouette plot. The scope and future study is to identify the prediction and classification for any type of MHC database with the help of other data mining techniques.

References

[1] B.Krishan Reddy, G.V.R.K. Acharyulu, (2002) Customer Relationship Management (CRM) in Health Care Sector - A Case Study on Master Health Check, *Journal of the Academy of Hospital Administration*, Vol. 14, No. 1 . (2002-01 - 2002-06).

[2] G. Manimannan, S. Hari and G. Vijay Thiraviyam (2012), Data Mining Applications in Master Health Checkup: a Statistical Exploration, Paper presented at *National Conference on Statistics for Twenty First Century-2012 (NCSTC-2012)*, Department of Statistics, University of Kerala, Trivandrum, Kerala, India.

[3] A.B.M. Shawahat Ali and Saleh A. Wasimi (2009), Data Mining: Methods and Techniques, Cenage Learning India Private Limitted, New Delhi.

[4] Leo Breiman (2001), Random Forest, Statistics Department, University of California, Berkeley, CA 9472, USA, 1-33.

[5] David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63.

[6] Guido van Rossum (1991), Interpreted high level programming language, Python Software Foundation

Citation: *G. Manimannan, et.al., (2019). Analysis of Master Health Checkup Data Using Data Mining Classification and Cross Validation. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 7(6), pp.30-37. http://dx.doi.org/10.20431/2347-3142.0706005*