Temporal Data Farming using Iterative Prediction on HDD

Mohd. Shahnawaz, Kanak Saxena

Abstract— Data farming is a process to grow data by applying various statistical predictions methods, machine learning and data mining approach on the available data while in temporal data farming we can generate data from temporal data. Regression is one of the statistical methods to predict, the value of an attribute (dependent variable) and effect of the other attribute (independent variable) on the predicted attribute. A regression equation is developed in parametric regression, which is a function of independent variables. In this paper we propose an algorithm for temporal data farming which is based on one of the parametric regressions known as linear regression to predict the class attribute and iteratively attempt to reduce prediction error tends to zero. Proposed algorithm is used in scenario execution loop of the data farming method and returns farmed data as crops, to merge in the original dataset to improve the dataset adequateness to mining. Case study with the proposed algorithm is made on the real life Heart Disease Data (HDD), in which we are trying to predict the correct dose of the dobutamine, should be given to the patient. Incorrect volume of the dose may affect the body of patient in other ways. This research may be helpful to the doctor's in the deep diagnosis of patient data and its disease. Research for the heart diseases is very much significant for the society.

Index Terms— Data Farming, HDD, Linear Regression, Scenario Building loop, Scenario Execution loop, Statistical Prediction, Temporal data.

1 Introduction

THIS paper presents an algorithm for temporal data farming based on iterative predictions using linear regression. Data farming is a process to generate more and more data [3]. In this paper, HDD (Heart Disease Data) is used as seed data and perform farming. HDD used in this work contain some temporal aspects like date of birth of the patient which is one of the important information direcltly effect the diagnosis the disease. Proposed algorithm can handle such type of temporal data and efficiently farm more data for mining and analysis purposes. In the proposed algorithm, we are generating prediction vector of the class attribute in response to reduce the error vector tends to zero. We developed the proposed work on weka 3.6.2. It is a data mining tool coded in java.

This paper is organized in 5 sections; in section 1 Introduction, section 2 describe the prediction process used in this work, section 3 contain a brief overview on data farming process. Section 4 proposed methodology and section 5 describe the outcomes as result analysis and finally section 6 conclude the entire paper.

2 Prediction Process

Prediction is a statistical method to estimate value of an unknown variable with the help of unknown variable. Linear regression is one of prediction methods [11]. It was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine. Linear regression is an approach to modeling the relationship between a scalar dependent variable y and one or more explanatory independent variables denoted X. The case of one explanatory variable is called Simple linear regression. More than one explanatory variable is multiple regressions [24]. In linear regression, data are modeled using linear functions, and unknown model parameters are estimated from the data. Such models are called linear models. Commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. linear regression could refer to a model in which the median, or some other quintile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis. Linear regression has many practical uses. Most applications of linear regression can be categories as [24].

1. Construction of Predictive Model: for prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of y and X values.

2. Relationship Finding: Given a variable y and a number of variables $x_1, ..., x_p$ that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the x_i, to assess which x_i may have no relationship with y at all, and to identify which subsets of the x_i contain redundant information about y.

3 DATA FARMING

Development of novel data mining applications calls for the definition of appropriate features and data collection at minimal cost [1]. Data farming is a process to get data on minimum cost. Data Farming is an iterative process [20]. Figure 1 presents the Data Farming process as a set of imbedded loops. This process normally requires input and participation by subject matter experts, modelers, analysts, and decision-makers.

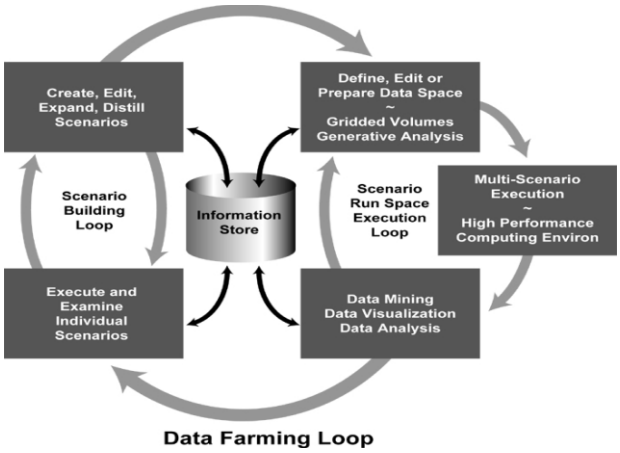

Fig. 1. Data Farming an Iterative Process

Data farming terminology is as follows:

1. **Distillations** - A software system that consists of a specific set of agents units, rules, variables, and environmental data that can be used to establish and execute a distillation or simulation scenario.
2. **Generative analysis** consists of automated methodologies & algorithm to drive parameter and rule variations in the iterative data farming process.
3. **Scenario** - is a specific starting configuration of assumptions, conditions, agents units, rules, variables, and environment data for a model.
4. **Excursion** - is a variation of a scenario with a defined set of altered rules, parameters or other starting conditions.
5. **Parameters** - The set of assumptions, conditions, agents units, rules, variables, starting conditions and environment data that define a scenario and an Excursion. The data model provides an extensible table of possible parameters.
6. **Replicate** - is a single execution of a scenario with a known random seed for stochastic properties.

The "Scenario Creation" loop shown on the left side of the figure involves developing and honing a model that adequately represents the system that addresses the question being asked by the decision-maker [18]. This is an iterative process that often requires honing the question as well. The "Scenario Run Space Execution" loop shown in Figure 1 is entered once the *basecase* of the scenario is complete.

In this loop the algorithm defines a *study* which determines which scenario input parameters should be examined and what processes should be used to vary them. Here the algorithm is exploring the possible variations (or *excursions* of the basecase) in the initial conditions of the scenario. Specifically those parameters that address the question being posed are considered. The defined study is used to guide the execution of many runs of the model in the HPC environment. Each run produces output which is collected by the Data Farming system and provided as output to analysis capabilities [19]. After analysis of the results, the system (or an algorithm) may decide to adjust or produce a new study or adjust the model to more adequately address the question. This process continues until insight related to the decision-maker's question has been gained.

4 PROPOSED METHODOLOGY

Methodology proposed in this research work based on the existing data farming loop or cycle with newly developed **Iterative_Data_farming** algorithm runs in the scenario execution loop. Data farming procedures as shown in figure in 1 contains information store, Scenario building loop and Scenario execution loop.

4.1 Information Store: it is the data storage, which provide the initial data for execution of the data farming algorithm and collect the harvested data after each iteration.

4.2 Scenario Building Loop

Scenario Building loop is responsible for providing data to Scenario Execution loop to prepare the data space (a data set ready to be processed by the data farming algorithm). It is a cyclic process; in each cycle following two sub procces are happened.

1. **Create Scenario, Edit, Expand, Distill Scenarios:** Scenario is the initial dataset sample for the algorithm. It's the initiliaztion of the proposed algorithm in which a particular dataset sample is taken and proceeds with subprocess 2 as under.
2. **Execute and Examine Individual Scenarios:** After selection of the initial dataset sample, it is must to examine individual tuple to be an instance for dataspace in scenario execution loop. Hence, in this step

- *Mohd Shahnawaz is currently pursuing doctorate degree program in computer Application under the faculty of Computer Science & Information Technology in Rajiv Gandhi Technical University, Bhopal - India, PH-09755624180. E-mail: shahnawaznbd@gmail.com*
- *Kanak Saxena is currently working as Professor in the department of Computer Applications at Samrat Ashok Technological Institute, Vidisha - India, PH-09826280706. E-mail: ks.pub.2011@gmail.com*

individual tuple are examined to be or not to be an instance of the data space with the help of the value of error thresold.

4.3 Scenario Execution loop

In this loop a data space is prepared by the input dataset sample recived from the scenario building loop, execution of farming algorithm and finally output of this loop is farmed data (and data mining, data analysis result) which can be used as initial dataset sample for the next iteration of data farming. This loop contains 3 sub processes as under.

1. **Define, Edit or Prepare Data Space**
   Recived the dataset sample (scenario) from scenario building loop, prepare it as data space and forward data space to the farming algorithm for execution.

2. **Scenario Execution**
   The actual data farming algorithm runs here, proposed algorithm **Iterative_Data_farming** given in the subsequent section execute in this step and return the *prediction_vector* as the output farmed data.

3. **Data Mining, Data Visualization, Data analysis:**
   Farmed data can be used for data mining, data visualization or data analysis in any way. Farmed data can be reused for next iteration of the data farming process. So it is collected and store in the information store database permanently.

This section describe proposed algorithm for temporal data farming which based on prediction by linear regression. Proposed algorithm will be run in the scenario execution loop of the data farming. In this algorithm, inner while loop checks the constraint on error threshold and terminate the iteration while errors becomes less then threshold. In outer for loop values of *class_attribute* are reassigned with the *prediction_vector,* finally the *prediction_vector* of the just previous iteration before termination of both the loop are return as the farmed data. Following pseudo code is given for the proposed methodology.

Algorithm: Iterative_Data_farming (data_space, class_attribute, error_thresold)

```
//data_space, it contain data sample in the form a multidimensional array
// class_attribute, Nominate the class attribute (Predicted attribute)
// error_thresold, permissible error in the actual and predicted values of class attribute
// error_vector, it contain the values of error after each iteration
// prediction_vector, it contain the predicted values after each iteration
{
 Initialize error_vector =0;
  for i = 0 to count (data_space)
    {
     while ( error_thresold  >  |error_vector [i] |)
        {
          Apply linear regression
          & evaluate prediction_vector
        }
      error_vector[i] = class_attribute[i] – prediction_vector[i];
      class_attribute[i] = prediction_vector[i];
     }
   return prediction_vector;
}
```

5 Result Analysis

The proposed algorithm is analyzed on weka 3.6.2 open source software for data mining. To perform the experiment we pick dataset of cardiac patient from HDD medical domain having 42 attribute & 558 instances, select a portion i.e. 25 instance and 12 attribute# including one class attribute named as 'Dose'. Sample dataset is given in table 1. We can see the result, decreasing values of errors range from iteration 1 to iteration 2 from 26.222 to 3.572 and after 3rd iteration it becomes zero. As in 3rd iteration errors are less or zero we can harvest 3rd iteration predicted values and accommodate with original dataset. In general case after nth iteration we can found adequate dataset with less errors, it termed as the new 'data farm'.

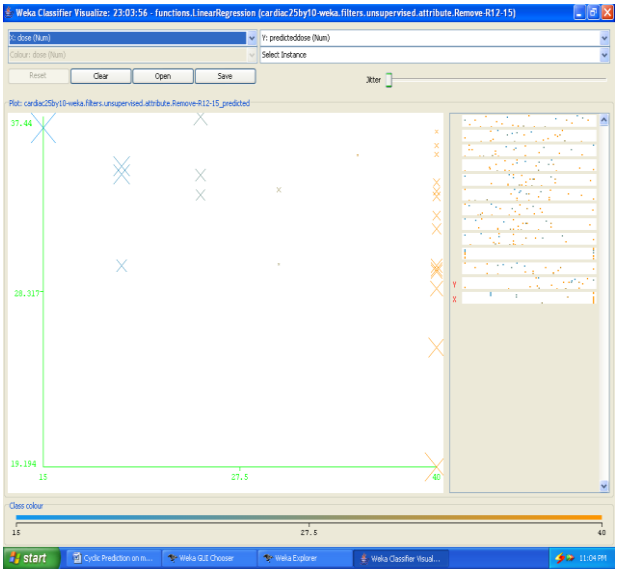

Fig. 2. Error Representation on weka

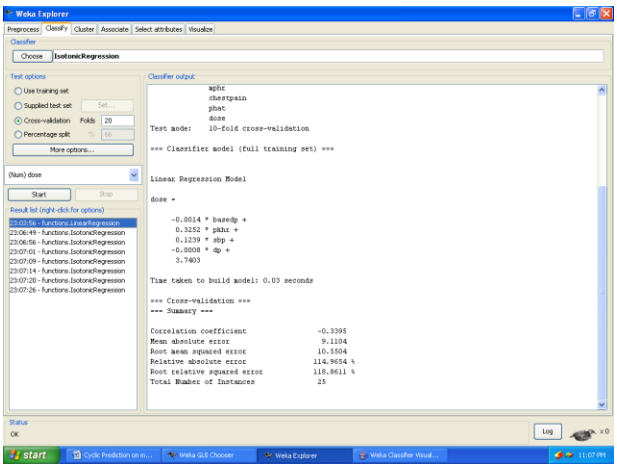

Fig. 3. Running Snapshot Linear Regression on weka

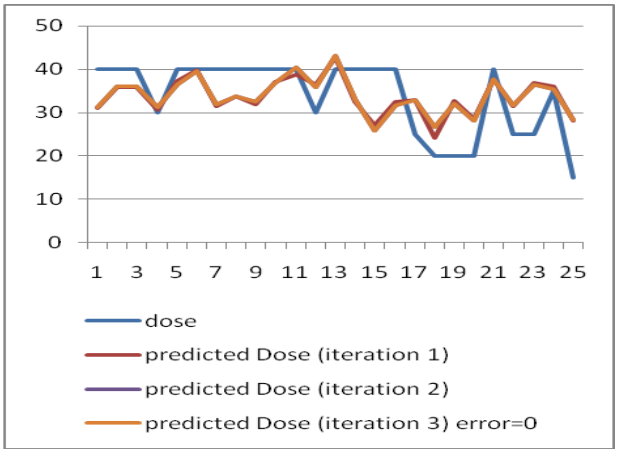

Fig. 5. Iterative Prediction on Dose.

Fig. 4. Iteratively error reduction graph for a sample of 25 instance.

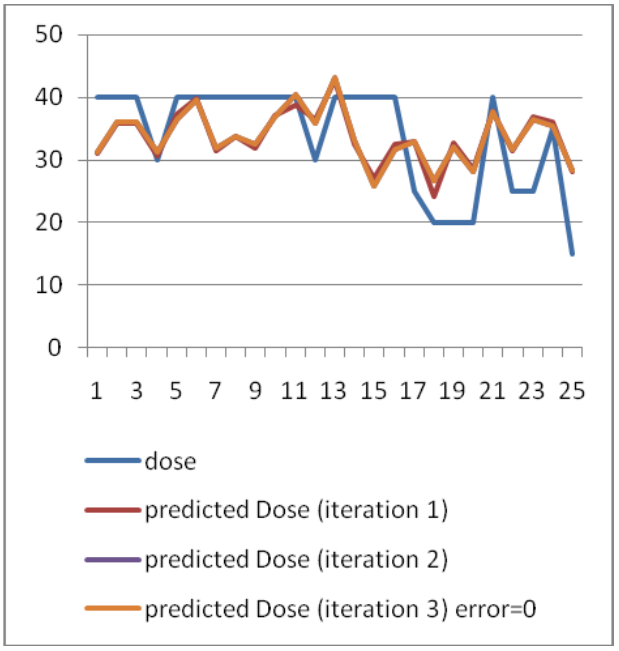

Fig. 6. Iterative Error Smoothing.

Figure 5 shows the graphical view of the class attribute, graph shows that after each iteration values of attribute 'dose' are less fluctuated. In this graph 25 tuple are shown with index 1 to 25 with their actual & predicted values for "dose" in 1st, 2nd & 3rd iteration are represent by the series.

Errors are reduces in each iteration, figure 4 shows a graph in which errors after each iteration represents by series of differ-

ent colors, we can see in 1 iteration error varies from – 12.97 to +13.252 while in 2 iteration - 1.265 to +2.307 and it become zero in 3rd iteration. Hence in iterative cycle data is farmed which can be used in various data mining projects. Proposed algorithm takes linear time in execution.

6 Conclusion

Proposed algorithm improves the adequateness of the available dataset for data mining, by generating prediction vector. Proposed algorithm farms more and more data until unless error threshold is not reached. Iterative prediction process minimizes the error rate. We can see in the result analysis error range from iteration 1 to iteration 2 decreases from 26.222 to 3.572. The proposed algorithm iteratively improve the adequateness of the data for mining process and side by side increase the size of the dataset in the form of additional predicted values. Proposed algorithm can handle the factor of temporality in the seed data, as in case study DOB is the temporal attribute. In this case study proposed algorithm is very useful to predict the correct dose of dobutamine should be given to the patient. High & low dose may cause life loss of the heart patient. In general case after n^{th} iteration we can found adequate dataset with less errors, it termed as the new 'data farm'. Hence the proposed algorithm is effective in temporal data farming process and error reduction, proposed algorithm not only applicable in HDD it can be used in any application like marketing, defence and medical etc.

References

[1] Dr. Alfred G. Brandstein, Dr. Gary E. Horne, Data Farming: A Meta-technique for Research in the 21st Century, Maneuver Warfare Science 1998

[2] Gary E. Horne, Klaus-Peter Schwierz, DATA FARMING AROUND THE WORLD OVERVIEW, Proceedings of the 2008 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

[3] SRIVATSAN LAXMAN and P S SASTRY, A survey of temporal data mining, S¯adhan¯a Vol. 31, Part 2, April 2006, pp. 173–198. © Printed in India

[4] Gary E. Horne, Ted E. Meyer, DATA FARMING: DISCOVERING SURPRISE, Proceedings of the 2004 Winter Simulation Conference, R. G. Ingalls,M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[5] Andrew Kusiak, Data Farming Methods for Temporal Data Mining, Intelligent Systems Laboratory, 2139 Seamans Center, The University of Iowa, Iowa City, Iowa 52242 - 1527

[6] Adam J. Forsyth, Gary E. Horne, Stephen C. Upton, MARINE CORPS APPLICATONS OF DATA FARMING, Proceedings of the 2005 Winter Simulation Conference, M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

[7] Philip Barry, Mathew Koehler, SIMULATION IN CONTEXT: USING DATA FARMING FOR DECISION SUPPORT, Proceedings of the 2004 Winter Simulation Conference,R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[8] M.Fleury, A.C.Downton and A.F.Clark, Scheduling schemes for data farming, IEE Proc.-Comput. & Digit. Tech., Vol. 146, No. 5, September 1999

[9] C.S. Choo,E.C. Ng,Dave Ang, C.L. Chua, DATA FARMING IN SINGAPORE: A BRIEF HISTORY , Proceedings of the 2008 Winter Simulation Conference S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

[10] Cattral, R., F. Oppacher, and D. Deugo (2001), Supervised and unsupervised data mining with an evolutionary algorithm, Proceedings of the 2001 Congress on Evolutionary Computation, IEEE Press, Piscataway, NJ, pp. 767-776.

[11] Han, J. and M. Kamber (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann, San Diego, CA.

[12] Kovacs, T. (2001), What should a classifier system learn, Proceedings of the 2001 Congress on Evolutionary Computation, IEEE Press, Piscataway, NJ, pp. 775-782.

[13] Kusiak, A. (2000), Decomposition in data mining: an industrial case study, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, pp. 345-353.

[14] Kusiak, A. (2001), Feature transformation methods in data mining, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24 (in print).

[15] Kusiak, A., J.A. Kern, K.H. Kernstine, and T.L. Tseng (2000), Autonomous decision-making: A data mining approach, IEEE Transactions on Information Technology in Biomedicine, Vol. 4, No. 4, pp. 274-284.

[16] Brian F. Tivnan, Data Farming Co evolutionary Dynamics in Repast, Proceedings of the 2004 Winter Simulation Conference R. G. Ingalls,M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

[17] Horne, G. And Meyer, T. 2004. Data Farming, Briefing Presented at the Informs National Meeting. Denver, Co.

[18] Simulation Experiments and Efficient Design (SEED) centre for data farming (2009) http:// harvest.nps.edu.

[19] Gary Horn, Stephen Seichter, Karsten Haymann, Data Farming in Support of Military Decision Makers 2010. http:// harvest.nps.edu.

[20] Gary Horne, Ted Meyer Data Farming and Defense Applications, Naval Postgraduate School, gehorne@nps.edu, temeyer@nps.edu

[21] Mohd Shahnawaz, Analysis of Data Farming Methods, 3rd National Conference on Emerging Trends in Software Engineering & Information Technology ETSEIT-2009 on 21-22 March 2009 GEC, Gwalior.

[22] Mohd Shahnawaz & Kanak Saxena, An Overview : Temporal Data Mining (TDM) International Journal of Computer Science & Management Systems. Vol 2 No. 2 Dec 2010.ISSN 0975-5349.pp 39-46

[23] Mohd Shahnawaz & Kanak Saxena, Analysis of Missing Value Estimation Algorithms for data Farming. International Journal of Engineering Sciences Special issue Sep-2011, Vol. 4, ISSN: 2229-6913.pp 496- 504.

[24] Mohd Shahnawaz & Kanak Saxena, A Comparative Study of Various Regression Model for Data Farming, International Journal of Wisdom Based Computing (IJWBC) Vol 2(1), 2012 ISSN 2231-4857.pp 29-34.

TABLE 1
SAMPLE DATA SET

Dob#	Bhr	basebp	basedp	pkhr	sbp	dp	maxhr	mphr	chest pain	phat	dose	predicted Dose (iteration 1)	Error (iteration 1)	predicted Dose (iteration 2)	Error (iteration 2)	predicted Dose (iteration 3)	Error (iteration 3)
31/12/1925	92	103	9476	130	86	9804	100	74	1	0.500272	40	31.057	-8.943	31.302	0.245	31.302	0
3/1/1937	62	139	8618	120	158	18960	120	82	1	0.548361	40	36.008	-3.992	35.995	-0.013	35.995	0
15/06/1938	62	139	8618	120	157	18840	120	82	1	0.548361	40	35.977	-4.023	35.969	-0.008	35.969	0
10/2/1962	93	118	10974	118	105	12390	118	72	1	0.646591	30	30.684	0.684	31.268	0.584	31.268	0
11/11/1976	89	103	9167	129	173	22317	129	69	0	0.522896	40	37.454	-2.546	36.439	-1.015	36.439	0
21/06/1939	58	100	5800	123	140	17220	123	83	1	0.646591	40	39.914	-0.086	39.659	-0.255	39.659	0
17/08/1929	63	120	7560	120	130	12740	98	71	0	0.629887	40	31.628	-8.372	31.825	0.197	31.825	0
31/12/1921	86	161	13846	160	157	22608	144	111	1	0.646591	40	33.791	-6.209	33.828	0.037	33.828	0
1/9/1932	69	143	9867	115	118	13570	113	81	1	0.552864	40	31.905	-8.095	32.411	0.506	32.411	0
22/07/1926	76	105	7980	126	125	15750	126	94	1	0.673639	40	37.218	-2.782	36.921	-0.297	36.921	0
19/11/1950	105	134	14070	171	182	31122	171	108	1	0.689343	40	38.781	-1.219	40.372	1.591	40.372	0
13/04/1947	72	112	8064	127	95	12065	125	80	1	0.689343	30	36.562	6.562	35.883	-0.679	35.883	0
30/06/1924	90	120	10800	169	184	31096	169	126	0	0.774755	40	42.824	2.824	43.198	0.374	43.198	0
19/09/1982	81	110	8910	110	130	14300	110	58	1	0.570676	40	32.497	-7.503	33.193	0.696	33.193	0
1/2/1939	84	176	14784	110	194	21340	110	74	0	0.689343	40	27.03	-12.97	25.765	-1.265	25.765	0
27/10/1947	100	122	12200	134	90	12060	136	87	1	0.570676	40	32.625	-7.375	31.621	-1.004	31.621	0
24/08/1919	69	169	11661	128	138	17664	128	98	1	0.599939	25	33.016	8.016	32.959	-0.057	32.959	0
20/10/1963	79	127	10033	79	133	10507	96	55	0	0.582453	20	24.201	4.201	26.508	2.307	26.508	0
19/08/1945	71	133	9443	114	132	15048	114	74	0	0.628505	20	32.745	12.745	32.147	-0.598	32.147	0
9/7/1929	65	183	11895	111	128	14208	89	64	1	0.611459	20	28.605	8.605	28.109	-0.496	28.109	0
17/12/1960	51	148	7548	123	148	18204	123	82	1	0.699578	40	37.778	-2.222	37.65	-0.128	37.65	0
14/02/1927	77	147	11319	122	115	14030	122	89	1	0.570676	25	31.489	6.489	31.7	0.211	31.7	0
5/9/1919	73	151	11023	141	150	21150	141	109	1	0.570676	25	36.897	11.897	36.572	-0.325	36.572	0
14/11/1951	91	132	12012	144	140	20160	144	89	1	0.704623	35	36.061	1.061	35.397	-0.664	35.397	0
2/10/1926	77	142	10934	95	167	15865	95	70	1	0.628505	15	28.252	13.252	28.309	0.057	28.309	0

#dob – date of Birth, bhr – basal heart rate, basebp – basal blood pressure, basedp – basal double product, pkhr – peak heart rate, sbp – systolic blood pressure, dp – double product, maxhr – maximum heart rate, mphr – % of maximum predicted heart rate achieved by the patient, chest pain – 0 means the patient experienced chest pain, phat – phat level of the patient, dose – dose of dobutamine given to the patient