Early Prediction of Heart Disease Using the Most Significant Features of Diabetes by Machine Learning Techniques

Avijit Kumar Chaudhuri
c.avijit@gmail.com

Dr. Anirban Das
anirban-das@live.com

Dr. Deepankar Sinha
dsinha2000@gmail.com

Dr. Dilip K. Banerjee
dkbanrg@gmail.com

***Abstract*: Medical science is witnessing high levels of specialization with doctors specializing in specific areas, say, heart disease (HD), diabetes, nephrology, and the like. In the process, patients have to make multiple visits for treatment of simultaneous ailments. Studies show that there is overlap in causes of different diseases. One such co-existence is observed in patients with diabetes suffering from HD too. In many cases, one precedes the other. Hence, it is worth diagnosing that a patient having a particular ailment is likely to develop another. Artificial Intelligence and machine learning methods are widely used in healthcare. There are few references to such work using data mining approaches. HD is a primary cause of death worldwide. Studies show that diabetes patients also have HD. This paper aims to identify the association and common risk factors between diabetes and HD - this finding aid in anticipating the HD of a diabetic patient. The authors use proven data mining approaches - logistics regression, decision tree, and random forest to arrive at the most accurate results. The validation is done using unsupervised method: K-means Clustering. The initial investigation demonstrates that body-mass-index (BMI) and age are among the key risk factors for diabetes; and smoking habit, age, gender-male and diabetes (glucose level) lead to HD. 31% of diabetic patients had HD.**

Keywords: Data mining, heart disease, diabetes, decision tree, random forest, logistic regression

I. INTRODUCTION

Artificial Intelligence and Data mining approaches have proved very useful in health care as they provide more significant insights into the cause and effect of health problems. Several data mining applications on health care, bio-medicine, and treatments of diseases have been witnessed since its introduction in 1994 [1-3]. Data mining can be described as investigating veracity of patterns in data in a large dataset with different data-types. These methods enable identification of characteristics, predict out comes and classify events [4].

Thus, data mining techniques are useful in classifying, clustering, and finding associations amongst factors that are causes of a problem. In health care, data mining techniques can be used to prevent suffering and mortality. Data mining can be used in the insurance industry to prevent fraud and check accuracy and consistency in claims [1]. Industries profitably implement and use data mining techniques. It has found its usefulness in various other fields such as business, education, medical, scientific research, and many more.

Medical records show a relationship between HD and diabetes [5] – there are instances of about 68% of diabetic patients of 65 years of age and above dying from HD; this includes 16% fatality from heart stroke. The mortality chance of diabetic patients is two to four-times than non-diabetic adults. Diabetes is also treated as one of the seven most significant HD risk factor by American Heart association.

HD is a major cause of mortality and morbidity in people with diabetes. Data from the 2012 National Heart Association indicates that 65% of people with diabetes are going to die of a type of coronary disease or stroke. HD incorporates stroke, artery disease, and peripheral artery disease. Individuals attacked in diabetes are at expanded risk of HD, and these occasions generally happen at a prior age contrasted with individuals with no diabetes. As the number of individuals with diabetes is anticipated to increase, the viewpoints for HD become seven more alarming. Likewise, the risk of HD increases with age. A systematic review of the literature on diabetes HD has been performed across the world. There were significant contrasts found in the approaches utilized in the investigations on HD in individuals with diabetes in various nations, implying that exact worldwide HD evaluations in individuals with diabetes were not ready to be delivered.

In low and medium-income countries, 80 % of deaths are diagnosed with diabetes and HD. Diabetic patients in these countries lack information and suffer from heart ailments. HD may be avoided or deferred by regulating blood glucose, blood pressure, cholesterol, smoking cessation, a nutritious diet, increasing physical activity, and lead a healthy lifestyle. Besides, there is a need to develop health systems capable of detecting and treating diabetes and HD. Non-communicable diseases will begin to dominate mortality rates faster, and more than 75% of deaths worldwide are expected to account for them by 2030. HD is a significant cause of mortality and morbidity in persons with diabetes and a barrier to sustainable development. Thus, it is crucial to reduce the risk of HD in people with diabetes.

The Diabetic HD (DHD) prediction utilizing data mining techniques is valuable for detecting DHD's early detection. This work aims at foreseeing DHD and its characteristics utilizing decision tree(DT), random forest (RF), and logistic regression (LR) data mining techniques.

The paper has 8 sections. The next section discusses the uniqueness and correlation of diabetes and HD. Section 3 states the relevance of data mining techniques in medical analysis; section 4 describes the data set; section 5 compares different supervised classifiers; section 6 analyses the data and section 7 discusses and validates the results. the paper ends with the conclusion - section 8.

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License

II. DIABETES AND HD

A. General Description

Among diabetic people, HD is the primary source of death [6,7]; matured people with diabetes have a double or higher risk of HD than non-diabetics[8, 9]. There is evidence of increased incidence of hypertension and dyslipidemia associated with diabetes[10]. It is reasonable to assume that excess weight in many people increases the likelihood of diabetes, obesity, and dyslipidemia, which leads to HD [11–13].

Although there are various methodologies for evaluating the HD risk and diabetes [14–16], there is less reference to their association. Nevertheless, one such resource (accessible free on the Internet at http:/www.diabetes.org/diabetesphd) has been widely accepted through multiple conflicting medical studies and combines nearly all known risk factors for HD. Studies show that instruments and other risk assessment algorithms are infrequently utilized in clinical practice to determine diabetes and HD's simultaneous impact.

There is evidence identifying different factors that show the co-existence of diabetes and HD [17-20]. These factors include the individual's glucose level in the blood, circulatory strain, cholesterol level (LDL), overweight, and tobacco use.

III. DATA MINING IN THE ANALYSIS OF DISEASES

Data mining has been recognized as a crucial way to categorize and predict, enabling recursive learning. These features lead to developing an Artificial Intelligence(AI) system and use it for diagnosing diseases and anticipating illness (Aishwarya et al., 2013) [21]. AI aids in automatic learning - resulting in faster diagnosis without involving the patients. This aspect has benefits as it saves patients' time and predicts with higher accuracy. The medical records have numerous variables, and the data size is large. It becomes challenging to draw conclusions and find hidden patterns. These difficulties are overcome with AI and data mining techniques that aid in classification, prediction, and derive association rules.

IV. DATA SETS

A. Dataset- 1

The Pima Indian-Diabetes dataset (www.kaggle.com) has been used in this article. The dataset is used to predict whether a patient with diabetes is also likely to suffer from HD. This dataset has 768 patient records and nine features - includes 268 patient-records with diabetes. Table 1 shows the attributes and their characteristics.

Attribute Description

B. Dataset- 2

Data was compiled using Framingham Heart Research dataset (https://www.kaggle.com/amanjmera1/framingham-heart-study-dataset) shown in table 2. Data associated with 4240 hospital patients was put to use for the analysis.

TABLE I. DESCRIPTION OF DIABETES DATASET

	Attributes	**Description**	**Values**
1	Pregnancies	Number of times pregnant	Continuous
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	BloodPressure	Diastolic blood pressure	mm Hg
4	SkinThikness	Triceps skin fold thikness	mm
5	Insulin	2-Hour serum insulin	mu U/ml
6	BMI	Body mass Index	weight in kg/(height in m) ^2
7	DiabetesPedigreeFunction	Diabetes Pedigree Function	
8	Age	Age	Years
9	Outcome	Class variable	0=no diabetes 1=has diabetes

TABLE II. DESCRIPTION OF HEART DISEASE DATASET

Sl. No	**Attributes**	**Description**	**Range of Values**	**Mean**	**Standard Deviation**
1	age	Age at exam time in years	Continuous	49.5801	8.5729
2	male	Male orFemale	0 = Female; 1 = Male		
3	education	Education of the patient	1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college		
4	currentSmoker	At present smoker or	Value 0 for no smoking; value 1 for		

		not	smoking		
5	cigsPerDay	Smoking habits - Average no. of cigarettes/day	Continuous	9.0059	11.9225
6	BPMeds	Blood Pressure medications	Value 0 for not taking any Blood Pressure medications; value 1 for already in Blood Pressure medications		
7	prevalentStroke	Fasting blood sugar>120 mg/dl	0 = false; 1 = true		
8	prevalentHyp				
9	diabetes	Diabetes present or not	0 = No; 1 = Yes		
10	totChol	Total amount of cholesterol present in blood	mg/dL	236.6995	44.5913
11	sysBP	Systolic blood pressure	mmHg	132.3546	22.0333
12	diaBP	Diastoloc blood pressure	mmHg	82.8978	11.9104
13	BMI	Body Mass Index	Weight/Height($kg/m^{2)}$	25.8008	4.0798
14	heartRate	Beats/Min (Ventricular)	Continuous	75	12.0254
15	glucose		mg/dL	81.9637	23.95433
16	TenYearCHD	HD present or not	0 = No; 1 = Yes		

V. CHOICE OF MODELS

This paper proposes using machine learning techniques—namely, LR, RF, and DT for disease cause analysis and disease factor prediction.

A. Decision trees(DT)

Authors propose the DT technique among the other types of Data mining techniques because of the following criteria:

- DT filtration techniques are easy to implement and easily understandable.
- It is a systematic and widely used data mining method.
- In data mining, DT demonstrates a very large-scale achievement in comparison with other techniques.
- Tree-like models are used to make decisions in this decision support system.
- DT techniques can be treated as the proven mechanism in knowledge discovery fields.
- DT classifiers are the most used in supervised classification extracting patterns and insights from set of independent inputs affecting thje output variable, say disease or no disease, (Shouman, Turner & Stocker, 2012)[22].
- This method can handle varied data – ordinal, nominal, ratio and text data.
- The processing of datasets with missing values is also possible.

This method is not affected by linearity in data, missing values, types of data, and outliers. The results are easy to understand and interpret.

B. Random Forest(RF)

RF is an ensemble approach - the dataset is split into small sub-sets, and the decision tree algorithm is run on these sets [23]. The subsets are sampled-with-replacement. The importance of features is obtained by voting. This concept is referred to as the bagging approach or bootstrapping, which reduces the variance without affecting the complete ensemble's bias. The results are validated using out-of-bag scores. This method reduces overfitting and high variance in outcomes [23].

C. Logistic- Regression (LR)

LR enables predicting a disease outcome, a dependent variable, using independent variables of multiple types. This method is suitable where the outcome is a categorical variable, including health care science (Dwivedi, A. K., 2018)[24]. Thus, LR is a useful technique for determining the extent of individual features' impact on diseases' occurrence or non-occurrence.

VI. DATA MINING AND ANALYSIS

The authors used the DT and RF methods to ascertain the significance of features. They applied these techniques to the two data sets associated with the diseases. The dataset with important variables were put in LR to predict the presence of diabetes and HD. The results were compared with LR analysis using original dataset. Thus, this study made the comparisons shown in table 3.

TABLE III. ANALYSIS MATRIX

Prediction	LR	DT	RF
LR	Original Data Set	Revised Data Set comprising significance variables revealed from DT	Revised Data Set comprising significance variables revealed from RF

The association rules were drawn for the both the datasets. The results were validated using K-means clustering, an un-supervised method to draw associations.

A. RF analysis For Diabetes disease dataset

RF analysis of the Diabetes Disease Dataset data reveals that comparatively significant variables are Glucose, BMI, Age, Diabetes PedigreeFunction, Births, SkinThickness, and Insulin.

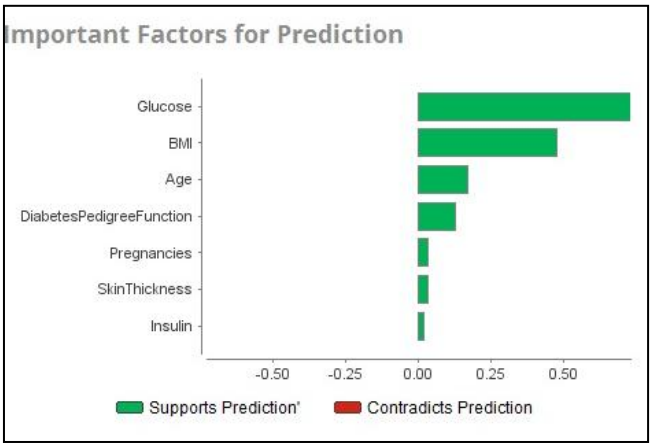

Fig. 1. RF classification to predict features on diabetes dataset

Similarly, the RF analysis of the HD dataset (Fig 2 shown below) shows that the most prominent variables for HD are glucose, sysBP, age, male, diaBP, currentSmoker and prevalentHyp.

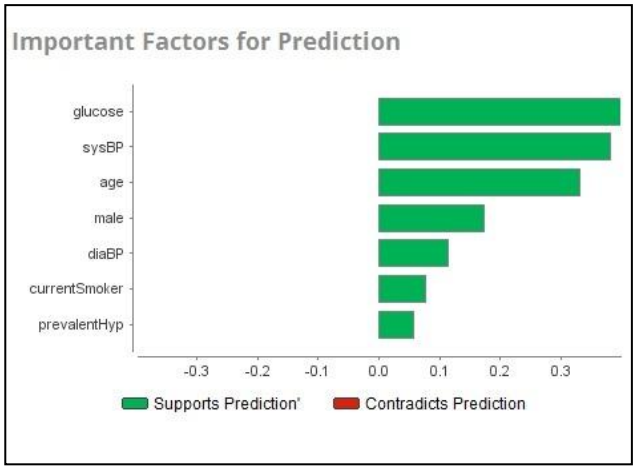

Fig. 2. RF classification to predict features on HD dataset

The patient dataset contained data on blood-pressure, a crucial marker for HD; however, the association rule and the significant factors did not include the same. This indicates that BP is not the cause for diabetes. The reverse can be tested with HD dataset which contains information on diabetes.

B. Decision Tree

Application of DT in Diabetes disease dataset

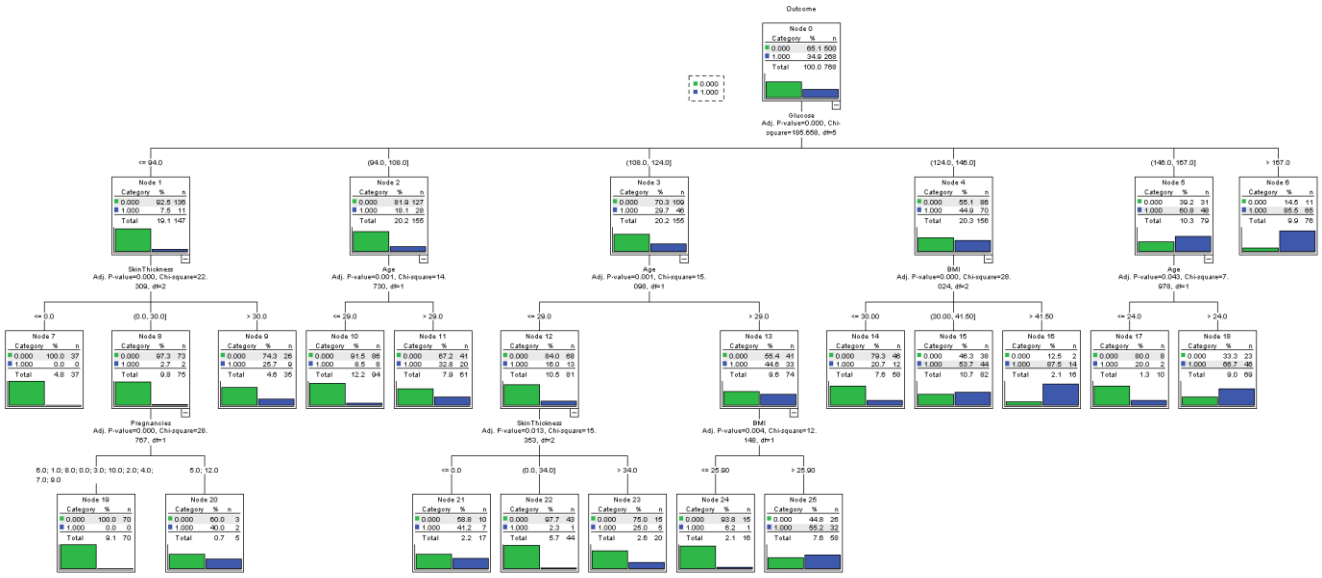

Fig. 3. DT analysis of the data on diabetes dataset

The association rules derived from the above DT analysis are shown below

If glucose < 94.0 then the next best predictor is SkinThickness
If SkinThickness <= 0.0 then probability of diabetes is 0.0%.
This is known as a terminal node because there are no child nodes below it.
If SkinThickness > 0.0 and <= 30.0 then next best predictor is Pregnancies
If Pregnancies = 6.0 or 1.0 or 8.0 or 0.0 or 3.0 or 10.0 or 2.0 or 4.0 or 7.0 or 9.0 then probability of diabetes is 0.0%.
If Pregnancies = 5.0 or 12.0 then probability of diabetes is 40.0%.
This is known as a terminal node because there are no child nodes below it.
If SkinThickness > 30.0 then probability of diabetes is 25.7%.
This is known as a terminal node because there are no child nodes below it.
If 94.0 < glucose <= 108.0 then the next best predictor is age
If age <= 29.0 then probability of diabetes is 8.5%.
If age > 29.0 then probability of diabetes is 32.8%.
This is known as a terminal node because there are no child nodes below it.
If 108.0 < glucose <= 124.0 then the next best predictor is age
If(age <= 29.0) then the next best predictor is SkinThickness
If SkinThickness <= 0.0 then probability of diabetes is 41.2%.
This is known as a terminal node because there are no child nodes below it.
If SkinThickness > 0.0 and <= 34.0 then probability of diabetes is 2.3%.
This is known as a terminal node because there are no child nodes below it.
If SkinThickness > 34.0 then probability of diabetes is 25.0%.
This is known as a terminal node because there are no child nodes below it.
If age > 29.0 then next best predictor is BMI
If BMI <= 25.0 then probability of diabetes is 6.2%.
This is known as a terminal node because there are no child nodes below it.
If BMI > 25.0 then probability of diabetes is 55.2%.
This is known as a terminal node because there are no child nodes below it.
If 124.0 < glucose <= 148.0 then the next best predictor is BMI
If BMI <= 30.0 then probability of diabetes is 20.7%.
This is known as a terminal node because there are no child nodes below it.
If BMI > 30.0 and <= 41.5 then probability of diabetes is 53.7%.
This is known as a terminal node because there are no child nodes below it.
If BMI > 41.5 then probability of diabetes is 87.5%.
This is known as a terminal node because there are no child nodes below it.
If 148.0 < glucose <= 187.0 then the next best predictor is age
If age <= 24.0 then probability of diabetes is 20.0%.
This is known as a terminal node because there are no child nodes below it.
If age > 24.0 then probability of diabetes is 66.7%.
This is known as a terminal node because there are no child nodes below it.
If 187.0 < glucose then probability of diabetes is 85.5%.

TABLE IV. ACCURACY LEVEL OF DT ANALYSIS ON DIABETES DATASET

Classification			
Observed	Predicted		
	0	1	Percent Correct
0	400	100	80.00%
1	67	201	75.00%
Overall Percentage	60.80%	39.20%	78.30%
Growing Method: CHAID			
Dependent Variable: Outcome			

TABLE V. ACCURACY LEVEL OF DT ANALYSIS ON HD DATASET

Classification			
Observed	Predicted		
	0	1	Percent Correct
0	3583	13	99.60%
1	618	26	4.00%
Overall Percentage	99.10%	0.90%	85.10%
Growing Method: CHAID			
Dependent Variable: TenYearCHD			

Table 4 shows that the DT analysis of all factors predicts diabetes with 78.30 % accuracy. The variables considered to be relatively important in the DT analysis are - Glucose accompanied by SkinThickness, Pregnancies, Age and BMI; this outcome does not differ from the RF analysis.

The patient dataset contained data on blood-pressure, a crucial marker for HD; however, the association rule and the significant factors did not include the same. This indicates that BP is not the cause for diabetes. The reverse can be tested with HD dataset which contains information on diabetes.

Application of DT in HD Dataset

Fig 4 displays the effect of the DT analysis conducted on all features in the HD dataset.

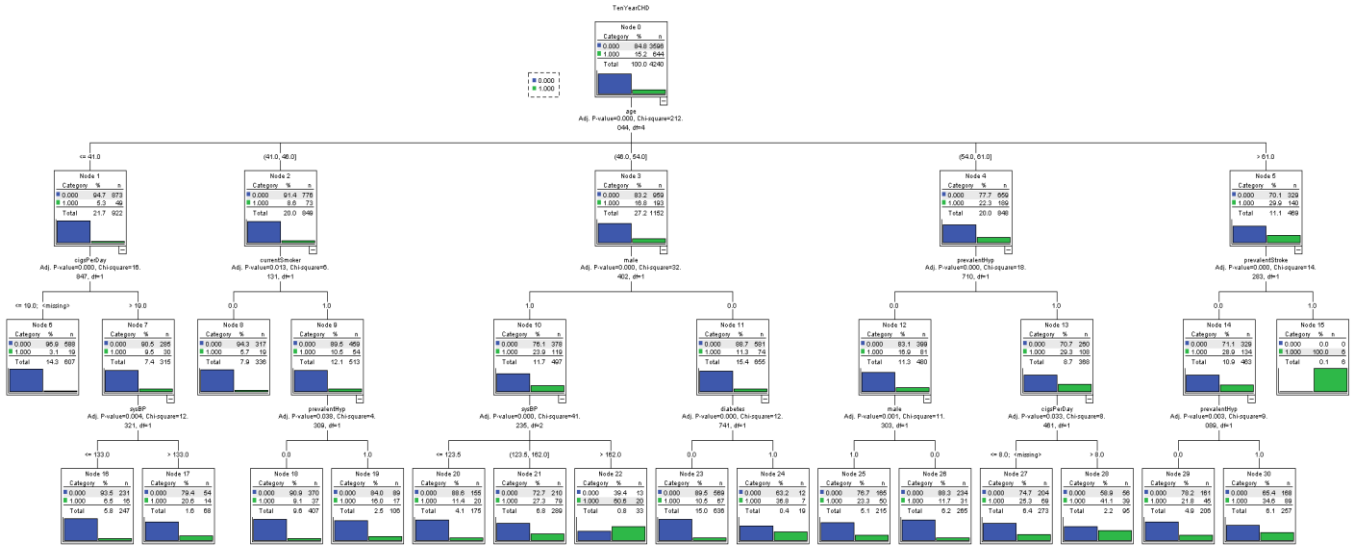

Fig. 4. DT analysis of HD dataset

The analysis estimates the association rules as follows:

age factor is the best factor for the detection of HD
For age <= 41, next best factor is cigsPerDay
If cigsPerDay <= 19.0 or missing then 3.1% patients have HD
If cigsPerDay > 19.0 then sysBP is the next best predictor
This is known as a terminal node because there are no child nodes below it.
If sysBP <= 133.0 then 6.5% patients have HD
If sysBP > 133.0 then 20.6% patients have HD
For age > 41 and age <= 46, next best factor is currentSmoker
If currentSmoker = 0.0 then 5.7% patients have HD
If currentSmoker = 1.0 then prevalentHyp is the next best predictor
This is known as a terminal node because there are no child nodes below it.
If prevalentHyp = 0.0 then 9.1% patients have HD
If prevalentHyp = 1.0 then 16.0% patients have HD
For age > 46 and age <= 54, next best factor is male(gender)
If male = 1.0 then sysBP is the next best predictor
This is known as a terminal node because there are no child nodes below it.
If sysBP <= 123.5 then 11.4% patients have HD
If sysBP > 123.5 and sysBP <= 162.0 then 27.3% patients have HD
If sysBP > 162.0 then 60.6% patients have HD
If male = 0.0 then diabetes is the next best predictor
This is known as a terminal node because there are no child nodes below it.
If diabetes = 0.0 then 10.5% patients have HD
If diabetes = 1.0 then 30.8% patients have HD
For age > 54 and age <= 61, next best factor is prevalentHyp
If prevalentHyp = 0.0 then next best factor is male
This is known as a terminal node because there are no child nodes below it.
If male = 1.0 then 23.3% patients have HD
If male = 0.0 then 11.7% patients have HD
If prevalentHyp = 1.0 then next best factor is cigsPerDay
This is known as a terminal node because there are no child nodes below it.
If cigsPerDay <= 8.0 then 25.3% patients have HD
If cigsPerDay > 8.0 then 41.1% patients have HD
For age > 61, next best factor is prevalentStroke
If prevalentStroke = 0.0 then next best factor is prevalentHyp
This is known as a terminal node because there are no child nodes below it.
If prevalentHyp == 0.0 then 21.8% patients have HD
If prevalentHyp == 1.0 then 34.6% patients have HD
If prevalentStroke = 1.0 then 100% patients have HD

Table 5 demonstrates that the DT study of all variables predicts HD with an accuracy of 85.1%.

DT and RF analysis with both showed that the variables - age followed by cigsPerDay, currentSmoker, male, prevalentHyp, prevalentStroke, sysBP, and diabetes are significant for HD. It includes diabetes as a significant variable. Moreover, the association rule obtained from DT analysis indicated that diabetes caused HD in 31% of the patients.

C. LR classification of Diabetes dataset considering all features

*1) The LR analysis of **all features** reveals that the variables Pregnancies, Glucose, BloodPressure, BMI and DiabetesPedigreeFunction are important together with the constant that predicts the existence of diabetes.*

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -8.405 + 0.123*\text{Pregnancies} + 0.035*\text{Glucose} - 0.013*\text{BloodPressure} + 0.090*\text{BMI} + 0.945*\text{DiabetesPedigreeFunction} \quad (1)$$

Table 6 shows that LR analysis on diabetes data set with all variables showed exactly accuracy similar to DT .

TABLE VI. ACCURACY LEVEL OF LR ANALYSIS ON DIABETES DATASET

Classification Table[a]

Observed			Predicted		
			Outcome		Percentage Correct
			0	1	
Step 1	Outcome	0	445	55	89
		1	112	156	58.2
	Overall Percentage				78.3

a. The cut value is .500

TABLE VII. ACCURACY LEVEL OF LR ANALYSIS WITH THE VARIABLES SELECTED BY DT ON DIABETES DATASET

Classification Table[a]

Observed			Predicted		
			Outcome		Percentage Correct
			0	1	
Step 1	Outcome	0	439	61	87.8
		1	117	151	56.3
	Overall Percentage				76.8

a. The cut value is .500

2) LR classification of Diabetes dataset considering relatively significant features from DT classification

The LR on characteristics considered to be important using DT shows that, along with the constant, Glucose, Pregnancies and BMI are statistically relevant and predict the existence of diabetes. With these factors, Equation 2 helps the estimation of the disease.

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -8.411 + 0.033 * \text{Glucose} + 0.113 * \text{Pregnancies} + 0.086 * \text{BMI} \quad (2)$$

The results showed an overall accuracy of 76.8% (as shown in Table 7) which is less than analysis with DT.

Therefore, in predicting the existence of diabetes, LR classification on the dataset with the features chosen by DT demonstrated lower accuracy.

3) LR classification of Diabetes dataset considering relatively significant features from RF classification

Glucose, BMI, DiabetesPedigreeFunction and Pregnancies, along with the constant, are statistically important and estimate the presence of Diabetes in the LR classification of the variable considered significant using RF. With these factors, Equation 3 helps the estimation of the disease.

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -8.837 + 0.034 * \text{Glucose} + 0.084 * \text{BMI} + 0.957 * \text{DiabetesPedigreeFunction} + 0.118 * \text{Pregnancies} \quad (3)$$

The findings revealed an average classification accuracy of 77.6% (Table 8), which is lower than the DT

TABLE VIII. CLASSIFICATION ACCURACY OF LR ON DIABETES DATASET CONSIDERING RELATIVELY IMPORTANT FEATURES DETERMINED FROM RF

Classification Table[a]

Observed			Predicted		
			Outcome		Percentage Correct
			0	1	
Step 1	Outcome	0	446	54	89.2
		1	118	150	56
	Overall Percentage				77.6

a. The cut value is .500

TABLE IX. CLASSIFICATION ACCURACY OF LR OF HD DATASET CONSIDERING ALL FEATURES

Classification Table[a]

Observed			Predicted		
			TenYearCHD		Percentage Correct
			0	1	
Step 1	TenYear CHD	0	3082	19	99.4
		1	506	51	9.2
	Overall Percentage				85.6

a. The cut value is .500

Thus, in predicting the existence of diabetes, LR on the dataset with the features chosen by RF demonstrated lower accuracy.

a) LR Classification of HD Dataset considering all features

The LR classification for all variables indicates that age, cigsPerDay, male, prevalent Stroke, sysBP, and diabetes are statistically relevant and predicted HD along with the constant. Equation 4 shows the assessment of the disease for these features.

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -8.328 + 0.555 * \text{male} + 0.064 * \text{age} + 0.018 * \text{cigsPerDay} + 0.002 * \text{totChol} + 0.015 * \text{sysBP} + 0.001 * \text{glucose} \quad (4)$$

The accuracy level using this method was 85.6%; higher compared to overall accuracy depicted by DT including the presence of HD.

LR analysis thus showed better accuracy in predicting the presence of HD on the dataset with the variables selected by DT analysis.

TABLE X. CLASSIFICATION ACCURACY OF LR ON HD DATASET CONSIDERING RELATIVELY IMPORTANT FEATURES DETERMINED FROM DT

Classification Table[a]					
			Predicted		
			TenYearCHD		Percentage
Observed			0	1	Correct
Step 1	TenYearCHD	0	3542	27	99.2
		1	599	43	6.7
	Overall Percentage				85.1
a. The cut value is .500					

TABLE XI. CLASSIFICATION ACCURACY OF LR ON HD DATASET CONSIDERING RELATIVELY IMPORTANT FEATURES DETERMINED FROM RF

Classification Table[a]					
			Predicted		
			TenYearCHD		Percentage
Observed			0	1	Correct
Step 1	TenYearCHD	0	3232	26	99.2
		1	551	43	7.2
	Overall Percentage				85
a. The cut value is .500					

b) LR classification of HD dataset considering relatively significant features from DT classification

The LR classification technique for features deemed to be important by DT classification technique indicates that age, prevalent Stroke, cigsPerDay, male, diabetes and sysBP and the constant are significant, predicting incidence of HD. The relationship is shown in equation 5.

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -7.540 + 0.065 * age + 0.020 * cigsPerDay + 0.476 * male + 1.021 * prevalentStroke + 0.014 * sysBP + 0.794 * diabetes \quad (5)$$

LR showed similar results compared to DT in terms of overall accuracy (of 85.1%). But the accuracy of the presence of HD in this method is greater than the DT analysis of all variables (4.0 %). This outcome showed that use of important variables predicted by DT, in LR, has improved the findings.

c) LR classification of HD dataset considering relatively significant features from RF classification

The LR classification of features considered to be important using the RF method shows that, along with the constant, glucose, sysBP, age, male and currentSmoker are statistically significant and predict the presence of HD. For these features, Equation 6 allows disease prediction.

LR Equation

$$\log\left(\frac{p}{1-p}\right) = -8.061 + 0.007 * glucose + 0.016 * sysBP + 0.64 * age + 0.594 * male + 0.382 * currentSmoker \quad (6)$$

Table 11 shows an overall accuracy of 85.0%

Therefore, in predicting the existence of HD, LR on the dataset with all variables demonstrated greater accuracy.

VII. RESULTS AND DISCUSSIONS

No single method showed better performance in all cases – diabetes and HD dataset; original and revised dataset; DT, RF and LR methods. The finding were validated using unsupervised clustering technique – K means method.

A. K means method

K means clustering method was applied to the diabetes dataset for identification of cluster characteristics where the outcome is one, i.e., diabetic patients. Table 12 lists the final clusters of diabetes dataset. It shows that blood sugar level of 130 and systolic blood pressure with value greater than 72 and above could suffer from diabetes.

TABLE XII. FINAL CLUSTER CENTERS

	Cluster									
	1	2	3	4	5	6	7	8	9	10
Pregnancies	4	4	5	1	4	4	2	4	3	2
Glucose	100	153	155	189	131	134	97	117	160	181
BloodPressure	72	72	79	60	73	72	68	1	73	80
SkinThickness	14	30	11	23	30	31	26	2	35	36
Insulin	1	309	0	846	127	196	70	1	502	712
BMI	30.5	35.1	33.4	30.1	33.5	34.8	30.5	25.8	35.6	44.5
DiabetesPedigreeFunction	.398	.546	.473	.398	.476	.604	.490	.393	.570	1.378
Age	33	34	41	59	33	34	27	30	33	27
Outcome	0	0	1	1	0	1	0	0	1	0

The cluster (3) with highest number of patients show that glucose level of 155 and BP greater that equal to 155 suffer from diabetes (shown in table 13).

TABLE XIII. CLUSTER-WISE NUMBER OF CASES

Cluster	1	228
	2	37
	3	121
	4	1
	5	110
	6	81
	7	136
	8	36
	9	16
	10	2
Valid		768
Missing		0

Similarly, the HD dataset was subject to K-means clustering to identify the cluster characteristics where the outcome is one, i.e., HD patients. Table 14 lists the final clusters of HD dataset. It shows that even with low value of diastolic blood pressure (85) diabetic patient suffered from HD – cluster 2. Whereas patients with diastolic blood pressure greater than 90 did not suffer from HD as they were non-diabetic.

TABLE XIV. FINAL CLUSTER CENTERS OF HD DATASET

	Cluster									
	1	2	3	4	5	6	7	8	9	10
male	0	1	1	0	0	0	0	0	1	0
age	44	57	54	50	55	54	46	51	52	53
education	2.06	1.92	1.72	2.03	1.83	1.96	2.08	1.94	1.99	1.81
currentSmoker	1	0	0	1	0	0	1	0	1	0
cigsPerDay	9	8	5	10	7	8	10	9	5	8
BPMeds	.01	.17	.05	.01	.12	.06	.00	.02	.00	.06
prevalentStroke	0	0	0	0	0	0	0	0	0	0
prevalentHyp	0	1	1	0	1	1	0	0	1	1
diabetes	0	1	1	0	0	0	0	0	1	0
totChol	172.6	250.5	241.6	243.6	257.6	336.8	210.4	283.5	648.0	203.8
sysBP	118.9	154.9	144.7	125.6	169.2	144.3	117.8	128.6	158.3	154.1
diaBP	76.2	85.3	86.6	80.0	99.2	87.3	76.1	81.9	90.5	93.7
BMI	24.27	25.62	28.57	25.66	27.90	26.59	24.58	26.04	26.36	27.22
heartRate	74	81	80	75	81	80	73	76	87	77
glucose	78.96	343.33	206.13	79.85	82.01	81.46	78.10	79.84	112.00	83.17
Outcome	0	1	0	0	0	0	0	0	1	0

Thus, the results from supervised classification were validated using the un-supervised method.

VIII. CONCLUSION

The incidence of diabetes is rapidly increasing worldwide and influencing heart-disease, causing a fatality. This paper does four critical analysis:

i. identifies significant features affecting diabetes and HD;

ii. finds linkage between diabetes and heart disease;

iii. compares the performance of well-proven supervised data mining techniques - decision tree, random forest, and logistics regression;

iv. validates the findings using an unsupervised method, i.e., K-means clustering.

The diabetic patients were primarily affected by t body-mass-index (BMI) followed by pedigree function and then age - shown in table 15.

TABLE XV. COMPARISON OF THE DEGREE OF ACCURACY AND DETECTION OF SIGNIFICANT FEATURES (DIABETES DATASET)

Comparison of Accuracy Levels and Identification of significant Variables for Diabetes Dataset										
Approaches	**Pregnancies**	**Glucose**	**Blood Pressure**	**Skin Thikness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Accuracy Predicting 1**	**Accuracy of Predicting 0**
RF	1	1		1	1	1	1	1	40.9	
DT - I	1	1		1		1		1	75	80
LR - I	1	1	1			1	1		58.2	89
LR - II	1	1				1	1		56	89.2
LR - III	1	1				1			56.3	87.8
TOTAL	5	5	1	2	1	5	3	2		

The outcome is the difference in skin thickness and glucose levels. The probability of this disease is high amongst pregnant females. Factors, namely smoking habits followed by diabetes (and glucose levels) and gender - male lead to HD. The outcome is high blood pressure and stroke.

DT and the ensemble method such as RF showed similar performance in overall accuracy. DT performed better in the diabetes dataset (75% correct prediction with outcome equal to 1) than HD dataset (4% correct prediction with outcome equal to 1). LR demonstrated better results for diabetes (58% correct prediction with outcome equal to 1) than HD (9% correct prediction with outcome equal to 1).

DT predicted better than any other method, i.e., 13.7% accurate in predicting HD, while RF predicted better than any other method, i.e., 13.7% accurate in predicting diabetes (table 16). Thus, no single method can be recommended in general - a comparative performance always leads to prediction (table 15).

TABLE XVI. COMPARISON OF THE DEGREE OF ACCURACY AND DETECTION OF SIGNIFICANT FEATURES (HD DATASET)

Comparison of Accuracy Levels and Identification of significant Variables for HD Dataset													
Approaches	**age**	**male**	**currentSmoker**	**cigsPerDay**	**prevalentStroke**	**prevalentHyp**	**diabetes**	**totChol**	**sysBP**	**diaBP**	**glucose**	**Accuracy Predicting 1**	**Accuracy of Predicting 0**
RF	1	1	1			1			1	1	1	13.7	
DT - I	1	1	1	1	1	1	1		1			4	99.6
LR - I	1	1		1				1	1		1	9.2	99.4
LR - II	1	1		1	1		1		1			6.7	99.2
LR - III	1	1	1						1		1	7.2	99.2
TOTAL	5	5	3	3	2	2	2	1	5	1	3		

Further analysis using improved ensemble techniques such as the extra-tree method can shed more light on diabetes and HD's co-existence. Besides, a single dataset indicating both HD and diabetes can lead to additional knowledge on the subject.

REFERENCES:

[1] Raghupathi, W., &Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential, Health information science and systems, 2(1), 3.

[2] Tomar D., Agarwal S., 2013. A survey on Data Mining approaches for Healthcare, Int. J. Bio-Sci. Bio-Technol , 5:241–266.

[3] Yoo I., Alafaireet P., Marinov M., Pena-Hernandez K., Gopidi R., Chang J.-F., Hua L., 2012. Data mining in healthcare and biomedicine: A survey of the literature, J. Med. Syst., 36:2431–2448.

[4] Crockett, D. &Eliason B., 2017. What is Data Mining in Healthcare, Health Catalyst.

[5] http://www.heart.org/HEARTORG/Conditions/More/Diabetes/WhyDiabetesMatters/%20 Cardiovascular - Disease-Diabetes_UCM_313865_Article.jsp#.XRddylQzbIU

[6] Engelgau MM., Geiss LS., Saaddine JB., Boyle JP., Benjamin SM., Gregg EW., Tierney EF., Rios-Burrows N., Mokdad AH., Ford ES., Imperatore G., Narayan KM., 2004. The evolving diabetes burden in the United States, Ann Intern Med 140:945–950.

[7] Haffner SM., Lehto S., Ronnemaa T., Pyorala K., Laakso M., 1998. Mortality from coronary Heart disease in subjects with type 2 diabetes and in non diabetic subjects with and without prior myocardial infarction, N Engl J Med 339:229–234.

[8] Hu FB., Stampfer MJ., Solomon CG., Liu S., Willett WC., Speizer FE., Nathan DM., Manson JE., 2001. The impact of diabetes mellitus on mortality from all causes and coronary Heart disease in women: 20 years of follow-up, Arch Intern Med 161:1717–1723.

[9] Fox CS., Coady S., Sorlie PD., Levy D., Meigs JB., D'Agostino RB Sr., Wilson, PW., Savage PJ., 2004. Trends in cardiovascular complications of diabetes, JAMA 292:2495–2499.

[10] Mokdad AH., Ford ES., Bowman BA., Dietz WH., Vinicor F., Bales VS., Marks JS. , 2003. Prevalence of obesity, diabetes, and obesity related health risk factors, JAMA 289:76 –79.

[11] Jonsson S., Hedblad B., Engstrom G., Nisson P., Berglund G., Janzon L., 2002. Influence of obesity on cardiovascular risk: twenty three-year follow-up of 22,025 men from an urban Swedish population, Int J ObesRelatMetabDisord 8:1046 –1053.

[12] Schulte H., Cullen P., Assmann G., 1999. Obesity, mortality and cardiovascular disease in the Munster Heart Study (PROCAM), Atherosclerosis 144:199–209.

[13] Thomas F., Bean K., Pannier B., Oppert JM., Guize L., Benetos A., 2005. Cardiovascular mortality in overweight subjects: the key role of associated risk factors, Hypertension 46: 654–659.

[14] Wilson PW., D'Agostino RB., Levy D., Belanger AM., Silbershatz H., Kannel WB., 1998 Prediction of coronary Heart disease using risk factor categories, Circulation 97: 1837–1847.

[15] Stevens RJ., Kothari V., Adler AI., Stratton IM., 2001. The United Kingdom Prospective Diabetes Study (UKPDS) Group: The UKPDS risk engine: a model for the risk of coronary Heart disease in type II diabetes, (UKPDS 56) ClinSci (Lond) 101:671– 679.

[16] Assmann G., Cullen P., Schulte H., 2002. Simple scoring scheme for calculating the risk of acute coronary events based on the 10- year follow-up of the prospective cardiovascular Munster (PROCAM) study, Circulation 105:310–315.

[17] Brunner EJ., Shipley MJ., Witte DR., Fuller JH., Marmot MG., 2006 Relation between blood glucose and coronary mortality over 33 years in the Whitehall Study, Diabetes Care 29:26 –31.

[18] Yusuf S., Hawken S., Ounpuu S., Bautista L., Franzosi MG., Commerford P., Lang CC., Rumboldt Z., Onen CL., Lisheng L., Tanomsup S., Wangai P. Jr., Razak F., Sharma AM., Anand SS., 2005. The INTERHEART Study Investigators: Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study, Lancet 366:1640–1649.

[19] Eberly LE., Prineas R., Cohen JD., Vazquez G., Zhi X., Neaton JD., Kuller LH., 2006. The Multiple Risk Factor Intervention Trial Research Group: Metabolic syndrome: risk factor distribution and 18-year mortality in the Multiple Risk Factor Intervention Trial, Diabetes Care 29:123–130.

[20] Wilson PW., D'Agostino RB., Parise H., Sullivan L., Meigs JB., 2005. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus, Circulation 112: 3066–3072.

[21] Aishwarya R., Gayathri P., N. Jaisankar., 2013. A Method for Classification Using Machine Learning Technique for Diabetes, International Journal of Engineering and Technology (IJET) ,Vol 5 No 3.

[22] Shouman M., Turner T. & Stocker R., 2012. Using data mining techniques in heart disease diagnosis and treatment, pp, 173–177.

[23] Breiman L., 2001. Random forests", Mach Learn, 45(1):5–32.

[24] Dwivedi, A. K., 2018. Performance evaluation of different machine learning techniques for prediction of heart disease, Neural Computing and Applications, 29(10), 685-693.