HPPS: HEART PROBLEM PREDICTION SYSTEM USING MACHINE LEARNING

Nimai Chand Das Adhikari[1], Arpana Alka[1] and Rajat Garg[2]

[1]Department of Mathematics, Indian Institute of Space Science and Technology, Thiruvananthapuram, India
[2]Department of Biotechnology, National Institute of Technology, Jalandhar, India

ABSTRACT

Heart is the most important organ of a human body. It circulates oxygen and other vital nutrients through blood to different parts of the body and helps in the metabolic activities. Apart from this it also helps in removal of the metabolic wastes. Thus, even minor problems in heart can affect the whole organism. Researchers are diverting a lot of data analysis work for assisting the doctors to predict the heart problem. So, an analysis of the data related to different health problems and its functioning can help in predicting with a certain probability for the wellness of this organ. In this paper we have analysed the different prescribed data of 1094 patients from different parts of India. Using this data, we have built a model which gets trained using this data and tries to predict whether a new out-of-sample data has a probability of having any heart attack or not. This model can help in decision making along with the doctor to treat the patient well and creating a transparency between the doctor and the patient. In the validation set of the data, it's not only the accuracy that the model has to take care, rather the True Positive Rate and False-Negative Rate along with the AUC-ROC helps in building/fixing the algorithm inside the model.

KEYWORDS

Heart Attack, Computation, Machine Learning, Data Analysis, Recommendation Systems, Neural Networks, Data Mining, Visualization, Artificial Intelligence

1. INTRODUCTION

The mortality rate in India and abroad is mainly due to heart attack. This calls for a vital check of the organ periodically for the wellness of all human beings. From the below figure of the heart, any major heart problem occurs when there is a blockage in the major arteries that carries the oxygenated blood [1]. The blockage causes huge pressure on the organ to pump the required amount of pure blood to the other parts of the body. The health care industry has huge amount of data that can be utilized to find the different patterns related to the heart problems with a probabilistic score. Here, we have collected the data from a survey of around 1000 patients from different parts of India and found out the correlation among the different risk factors that we have gathered.

The risk factors that has been taken as an input. in this survey are Family History, Smoking, Hypertension, Dyslipidemia, Fasting Glucose, Obesity, Life Style, CABG and High Serum in blood. Apart from the mentioned risk-factors, we have the demographic details as well. The most

Dhinaharan Nagamalai et al. (Eds) : AI, CSTY, SIGI - 2017
pp. 23– 37, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71803

important thing that each diagnosis should prevent is the exposure to a normal human body to the CT Scan radioactive rays [2][3]. The CCTA (Coronary computed tomography angiography) is an imaging test for the heart to find out the places for the plaques build up in the blood vessels. This has an increased prone to the cancer for the human body exposed to high radiation [4]. Plaque is majorly built up due to the circulating substances in blood like fat, cholesterol and calcium, whose deposit in the inner side of blood vessel can effect the normal blood flow and can result in excessive pressure on the heart pump.

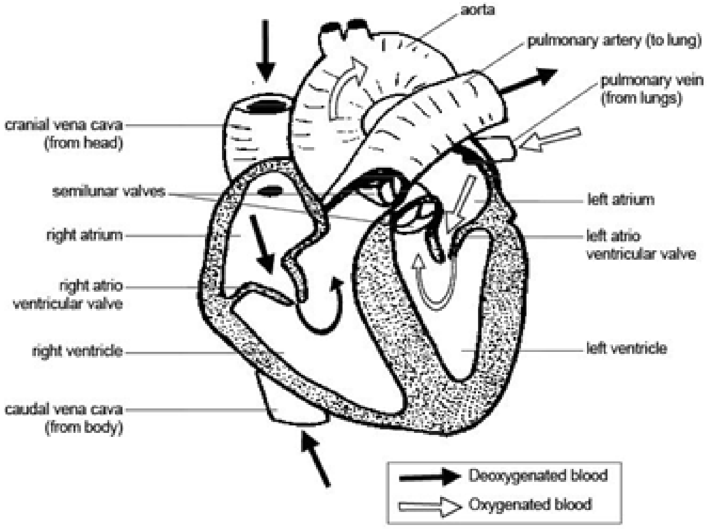

aorta
pulmonary artery (to lung)
pulmonary vein (from lungs)
cranial vena cava (from head)
semilunar valves
left atrium
right atrium
left atrio ventricular valve
right atrio ventricular valve
right ventricle
left ventricle
caudal vena cava (from body)
Deoxygenated blood
Oxygenated blood

Figure1: Diagram of Human Heart

So, the main intension of this paper is to help in the decision making of a doctor for detecting the possibility or identifying the patient’s suffering or going to suffer from heart problems. Apart from the above mentioned, this method should also help in diminishing the False Negative Rate of the prediction. It is the number of the actual positives which is negative through the prediction to the total negatives. In statistical hypothesis testing, this ratio is represented by the letter β. In the following sections we will discuss the different terminologies and factors related to this project and the methodology of HPPS, which can be a partner of the doctor in the decision making of whether the patient is going to suffer from any heart attack or not. In the next section we will discuss about the factors that we have taken for the survey and their correlations with the predictor output, followed by the proposed model and scenarios and lastly with the results for the selection of the algorithms.

2. Dataset Description and Analysis

The survey contains the data of 1094 patients from 5 different cities of India Delhi, Chennai, Bangalore, Kolkata and Hyderabad. The attributes that define as the features for the model are the different demographic details of the patients like Age and Sex with the different Risk Factors which we have defined previously. Here the predictor variable is Heart Problem or Not. Thus, there are many terminologies that define this. Some of them are:

1. Heart disease due to atherosclerosis [5]: In this case the walls of the arteries become stiff or hard due to the fatty deposits which in medical term known as plaques.

2. Cerebrovascular disease [6]: This is mainly due to the blockage in the blood flow through the blood vessels to the brain.

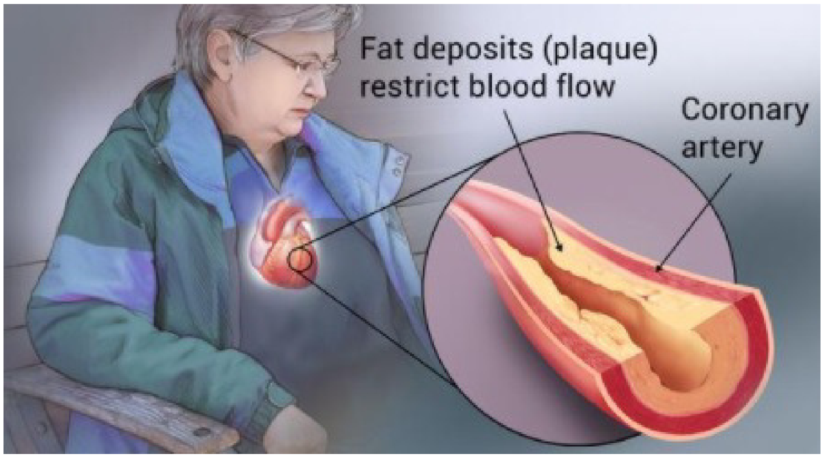


Figure 2: Heart Blockage

3. Ischemic heart disease [7]: This is mainly due to the deposit of the cholesterol on the walls of the arteries. Figure 2 shows how the deposit looks like in this similar case.

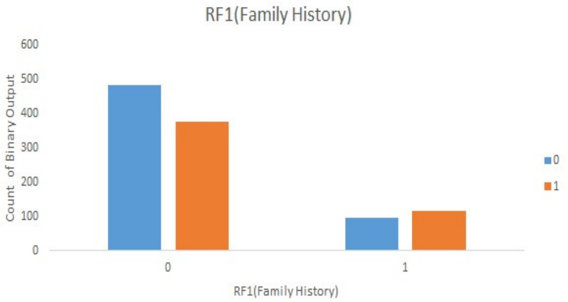
4. Hypertensive heart disease [8]: This happens mostly due to high blood pressure.

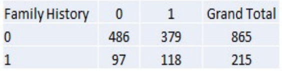
The above is some of the types of heart problems that we have discussed. There are many apart from the ones described before as the heart is one of the vital organs that help in the transportation of the oxygenated blood and nutrients and removal of wastes from the body. In the predicted value, we have given the value as 1 for the heart related problems and 0 as no problem in the heart.

Below is the analysis of the different risk-factors for the heart problem detection.

2.1. RISK FACTOR 1: FAMILY HISTORY

This is one of the important risk-factor as it depends on the hereditary behaviour of the heart [9]. Here, we have the values of 1080 patients and the rest are NA or No values. For those missing values we have assigned the value as 0 or the maximum of the value that appears in this risk factor. Which we will discuss in the results section.



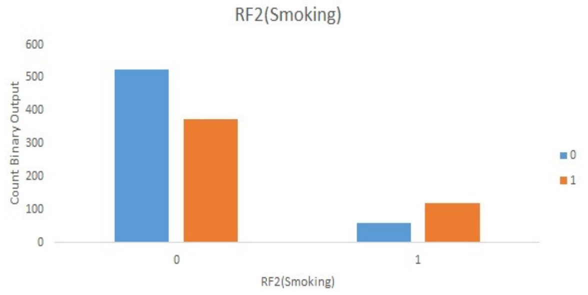


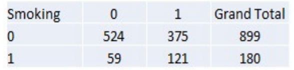
Family History	0	1	Grand Total
0	486	379	865
1	97	118	215

In the analysis we found that, when Family History is 1, then 118 out of 215 patient suffer from heart problem i.e 55%.

2.2. Risk Factor 2: Smoking

It leads to the developing of the cardiovascular diseases, which includes heart attack and stroke. It leads to damaging the lining of the arteries which ultimately leads to atheroma. Below is the analysis of the data for the smoking that we have established.



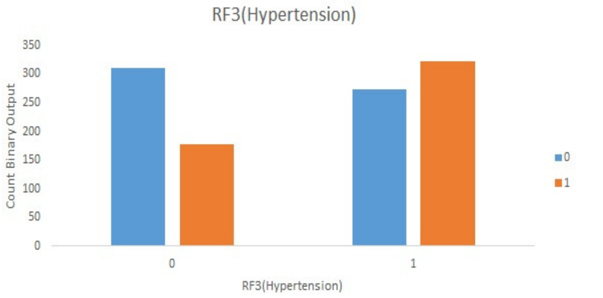


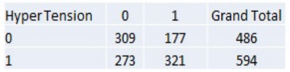
Smoking	0	1	Grand Total
0	524	375	899
1	59	121	180

The above curves show that if the patient has smoking as a characteristic, then 67.22% chances is, he/she will suffer from the heart related problems [10] [11].

2.3. Risk Factor 3: Hypertension

This leads to the heart diseases that occur due to high blood pressure over a long period of time [12][13]. Due to blood pressure, the heart has to do pump more against this pressure, adding extra pressure to heart resulting into the thickening of the heart muscle.





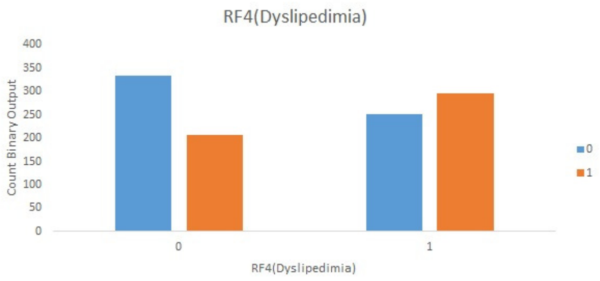
HyperTension	0	1	Grand Total
0	309	177	486
1	273	321	594

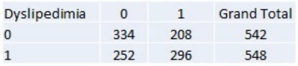
Figure 3: Hypertension

In the analysis done and represented in the figure 3, we can find that 54% chances is there for a hypertensive patient to suffer from any heart related problem.

2.4. RISK FACTOR 4: DYSLIPEDIMIA

This is a high level of lipids like cholesterol, triglycerides carried through the lipo-proteins present in the blood. The risk of *Atherosclerosis* increases due to the increase in the above-mentioned lipids in the blood leading to excessive pressure on the blood flow [14].



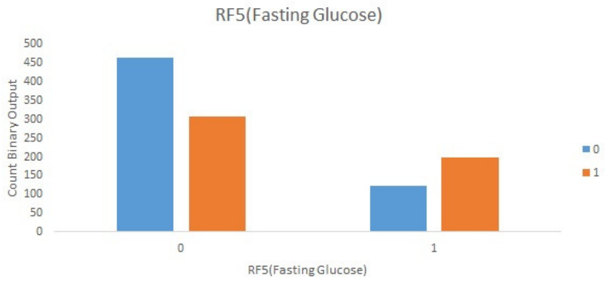


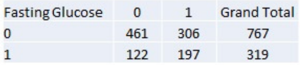
Dyslipedimia	0	1	Grand Total
0	334	208	542
1	252	296	548

In this analysis, we found that out of 1090 patients having the details of suffering from dyslipedimia which has been captured by the doctor, 548 suffer from the same. Out of 548, 296 patients suffered from heart related problems, which is a little over 54%.

2.5. Risk Factor 5: Fasting Glucose

Fasting Glucose greater than a certain value leads to type 2 diabetes and it is proved that type 2 diabetes increase marks the risk of Cardiovascular Disease(CVD) and ischemic heart disease(IHD) [15] [16] [17].



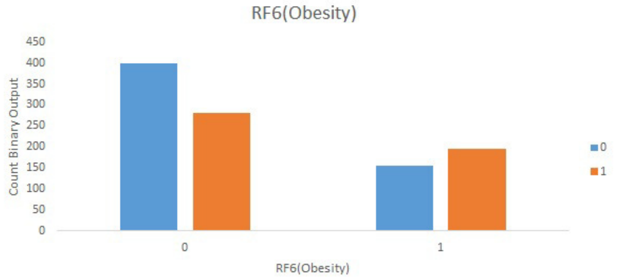


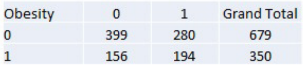
Fasting Glucose	0	1	Grand Total
0	461	306	767
1	122	197	319

According to our analysis, we found that 1066 data of the patients had this risk factor captured. Out of this, 319 had Fasting Glucose as marked 1. About 62% of those having 1 in this risk-factor suffered from the heart attack, which proved the analysis with that of the proven results.

2.6. Risk Factor 6: Obesity

The role of diet in the prevention of CVD is very crucial as it is a very key risk factor for CVD. Thus, obesity leads to the development of hypertension, diabetes, musculoskeletal disorder, thus putting in a high risk of CVD [18].



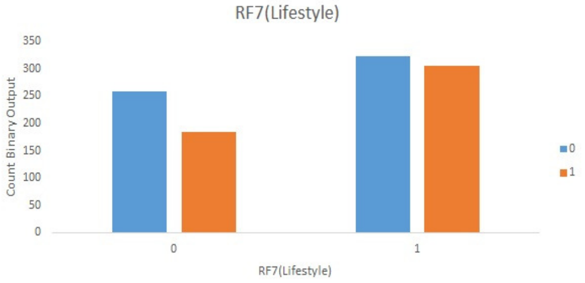


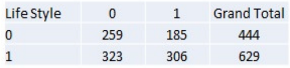
Obesity	0	1	Grand Total
0	399	280	679
1	156	194	350

According to the analysis, we found that 194 patients having Obesity suffered from heart related problems which accounts to 56%.

2.7. RISK FACTOR 7: LIFE STYLE

It is one of the most important factors in controlling the heart related problems. Some of the major lifestyle effects that can control in the prevention and keeping the heart in a good shape are Stop Smoking, Choosing Good Nutrition, High Blood Cholesterol, Lowering High Blood Pressure, Being Physically Active, Aiming for a healthy weight, Managing Diabetes, Reducing Stress and drinking alcohol etc. [19][20].



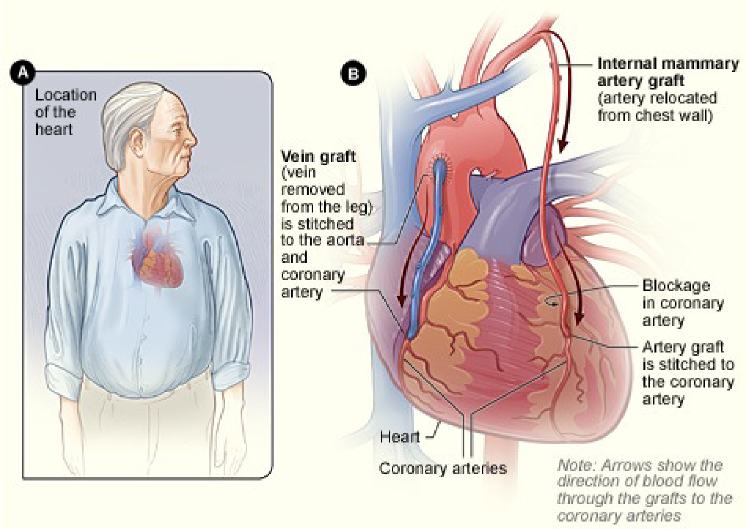


Life Style	0	1	Grand Total
0	259	185	444
1	323	306	629

In the analysis above, we find that 306 cases out of 629 marked as 1, suffered from heart related disease. Thus, this is around 49% of the cases. But if we see the two bar plots above we can find that the conversion of the heart problem is in a greater percentage in case of the bad life style. Thus, marking this risk factor to be one of the most important factors in determining the CVD.

2.8. RISK FACTOR 8: CABG

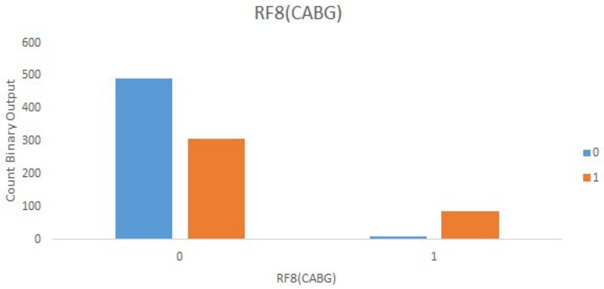
Coronary Artery Bypass Grafting is a kind of surgery done for those patients who have suffered from severe CHD.

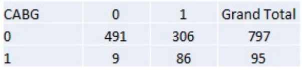

A
Location of the heart
B
Internal mammary artery graft (artery relocated from chest wall)
Vein graft (vein removed from the leg) is stitched to the aorta and coronary artery
Blockage in coronary artery
Artery graft is stitched to the coronary artery
Heart
Coronary arteries

Note: Arrows show the direction of blood flow through the grafts to the coronary arteries

Figure 4: Bypass Grafts in heart

This is mostly whether a patient suffered from the serious heart attack and has a graft anywhere in the heart. Thus, this will be having a very high correlation for the heart being regularly checked up.



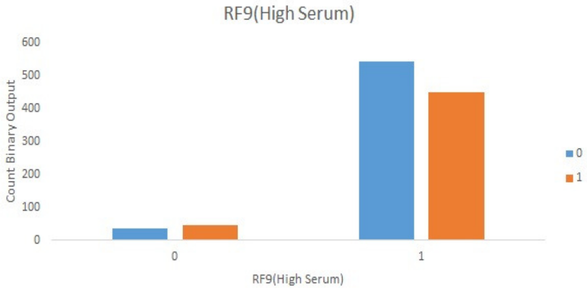


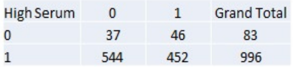
CABG	0	1	Grand Total
0	491	306	797
1	9	86	95

From our analysis done, we found that 95 of the cases of the patients had grafts present or this risk factor being high according to the doctor. Among the data, 86 do have severe heart problem and being asked for re-check-up.

2.9. Risk Factor 9: High Serum

A Serum test is a measure of the amount of iron which is present in the left over liquid after the red blood cells and the clotting factors being removed from the blood. Hence having too much iron content in the blood can cause serious health problem. This has a direct correlation with the heart related problems [21].

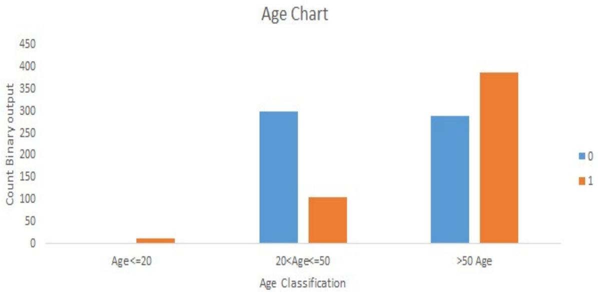


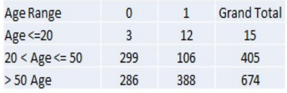


High Serum	0	1	Grand Total
0	37	46	83
1	544	452	996

In this analysis, we found that 452 cases having suffered from heart problems out of 996 having High Serum.

Apart from the above Risk Factors we have different other attributes like Age, Sex, Location and Vascular Pattern. The analysis of the Age feature is shown below in binned form.

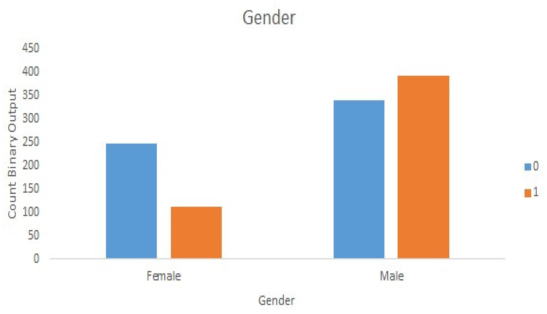


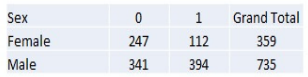


Age Range	0	1	Grand Total
Age<=20	3	12	15
20 < Age <= 50	299	106	405
> 50 Age	286	388	674

We have divided the age continuous values into three groups ’age < 20years’, ’age between 20 years and 50 years’ and ’age > than 50 years’. We can find from the analysis that most of the cases the age group more than 50 years have suffered from heart related problems which is not the case in case of the middle age group. Thus, the heart problem is skewed towards the more than 50 age group.

In the below graph we are showing the analysis for the heart problem with that of the gender or sex category of a patient.

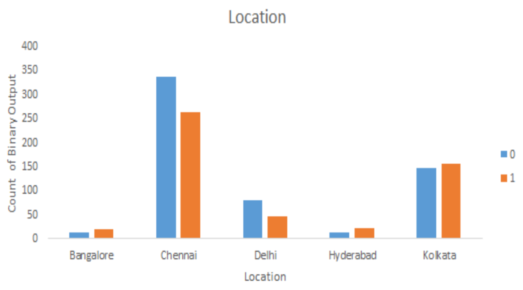


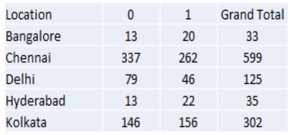


Sex	0	1	Grand Total
Female	247	112	359
Male	341	394	735

Males are more prone to heart related problems than female as can be seen from the analysis.

In another analysis which represent the variation of the patients with the location of 5 different cities all over the India i.e Chennai, Delhi, Kolkata, Bangalore and Hyderabad, we find that the data is skewed towards Chennai as more data is available from that region and when the patient suffering from heart related problem is seen, Chennai, Delhi and Kolkata are having patients details more than 100 and out of them Kolkata region has more patients suffering from heart problem and is recorded as 51.66% which can be seen from the location graph below.





Location	0	1	Grand Total
Bangalore	13	20	33
Chennai	337	262	599
Delhi	79	46	125
Hyderabad	13	22	35
Kolkata	146	156	302

3. EXISTING PROCEDURE AND LITERATURE SURVEY

As talked earlier in this paper heart disease remains one of the main causes for deaths worldwide. About 7.4 million people died due to coronary heart disease, and 6.7 million were only due to stroke (WHO, 2015). In order to investigate the misfortune of heart attack, certain factors that are associated with different risks need to be addressed. Therefore, people with heart disease due to the presence of chest pain, resting blood pressure, cholesterol, fasting blood sugar resting electro cardiographic and maximum heart rate need early detection and prediction for better counseling and appropriate medicine. According to Anooj(2012) and Hedeshi and Abadeh(2014), the decision to make for the presence of any problem in heart sorely depends on the physicians intuition, experience and experience. This is a very challenging task and needs to take care of a number of factors. Mostly the work related to the prediction and figuring out the heart problem, many data driven techniques has been used in past and the work inclines towards the classification problem. This is a process used to tune a model and then predict the class for whether the patient is suffering from any heart related problem or not. To talk about the intelligent methods in the medical sector, a vast number of related works has been performed (Muthukaruppan & Er, 2012; Sikchi et al., 2012; Kumar, 2013; Sikchi et al., 2013). The practitioners make use of these computerized intelligent methods for assist in the diagnosis to give suggestions with certain probability. In 2012 Opeyemi and Justice suggested one of the best and effcient technique to deal with the uncertainty by incorporating fuzzy logic and neural network. There are many diverse studies that tend to the ANFIS methodologies (Palaniappan & Awang, 2008; Patil & Kumaraswamy, 2009; Abdullah et al., 2011; Zhu et al., 2012; Kar & Ghosh, 2014; Mayilvaganan & Rajeswari, 2014; Yang et al., 2014). This research involves in the developing a framework that includes hybrid learning algorithms to find the least square estimates with gradient descent and Levenberg-Marquardt algorithms for training Statlog-Cleveland Heart Disease Dataset [24]. Some of the recent work on the heart problem prediction has been done using naive bayes [25] [26]. In [27], many classification algorithms like Naive Bayes, Decision Tree, K-NN and Neural Network is used for Prediction of Heart Disease and the result proves that Naive Bayes technique outperformed other used techniques. Similar to this [28] tree based algorithms J48, Bayes Net, Simple Cart, and REPTREE along with and Naive Bayes algorithm is used to classify and develop a model which diagnose heart attacks in the patient data. Three popular data mining algorithms (support vector machine, artificial neural network and decision tree) were employed to develop a prediction model using 502 cases for better prediction of heart problems [29]. SVM became the best prediction model followed by artificial neural network. In [30] a new concept of Weighted Associative Classifier was used where it was used to predict the probability of patients receiving heart attacks. In this Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. Coming forward a new approach different from above in [31] based on adaptive neuro-fuzzy models are presented was proposed. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work effciency in performing analysis for coronary heart disease occurrences [32].

4. PROPOSED SYSTEM

In the proposed model, we want to give a brief idea about how our system looks like and behaves. Below is the flow chart of our model:

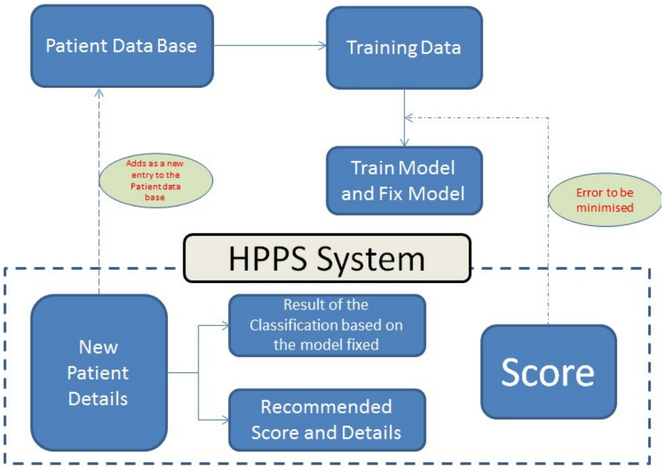

Patient Data Base
Training Data
Train Model and Fix Model
Adds as a new entry to the Patient data base
Error to be minimised
HPPS System
New Patient Details
Result of the Classification based on the model fixed
Recommended Score and Details
Score

Figure 5: Flow Chart of HPPS

Dataset: There is a common database for the patient from where the data will be taken by the model to finalize the algorithm. The Database will be for a particular hospital where HPPS is being installed or it can be an on-line stored data where all the details of the patient for those hospitals who use HPPS can access it. This will be helpful for both Classification Score and Recommendation [23].

Algorithms: We have used a wide range of algorithms and in the validation set, the algorithm that gives a better Selection Value i.e.

$$\text{Selection}_{\text{value}} = 0.6 * (1 - FNR) + 0.4 * Accuracy$$

Here, we have assigned 0.6 to the term having the FNR, as we wanted to diminish the False Negative Rate more than the Accuracy. Using the above metric, the algorithm which gives the maximum score in the validation set, is selected.

Recommender System: When a new patient detail is input to HPPS, using the risk factor combinations, all those similar patient details is made a clustered display using the cosine-similarity.

$$\text{score}_{\text{similarity}} = \frac{< patient_{new} \, , patient_{old} >}{|patient_{new}|_2^2 * |patient_{old}|_2^2}$$

Recommender Score: The Voted Output of the recommended patient details will be shown in the dash board [figure 6].

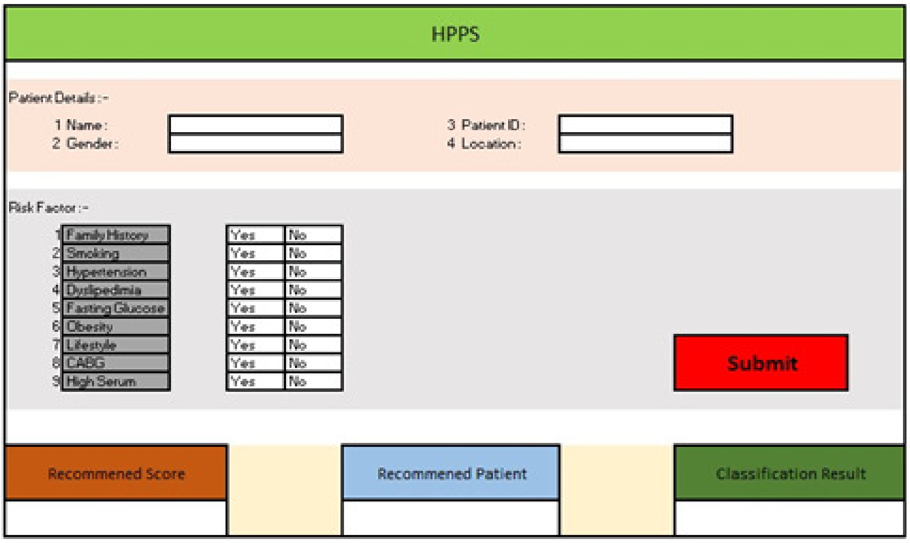


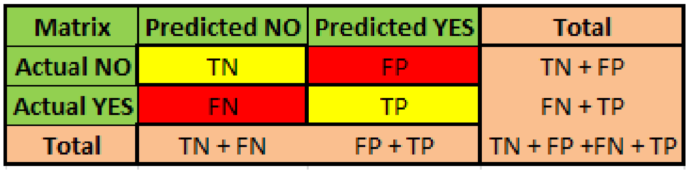
Figure 6: Dashboard of HPPS

Using the above information, the doctor will have multiple scenarios and also help him in aiding to his decision. This will also help to create a transparency among the doctor and the patient. So, here we want to showcase a system which can create a confidence in the patients mind for he/she is going to have any heart problem in the future or not, so as to take better care.

In the dashboard shown in figure 6, there are three sections. In the first section, the demographic details of the patient will be recorded and the Patient ID will be automatically filled. This will mostly depend upon the hospital id and the patient number. The second section is the risk-factors details section. Here the values that will be input will be mostly Boolean i.e. Yes or No. In the right side of this section one red button labelled Submit is present. Once the button is pressed, Section 3 will generate it's all relevant values.

5. RESULTS

In this section we have presented the results comparison for different algorithms that we used in the model. Here, we have analysed the details of 1094 patients having label as 1 or 0. Here 1 is represented for the patient's suffering from any kinds of heart disease and vice versa. Also, for small plaques, the label is given as 0. For training and validation to check how the algorithm is performing, we have used the holdout technique with 70:30 ratios. There are many others cross validation techniques but we have fixed our model to start the testing phase with the 70:30 percent holdout technique.

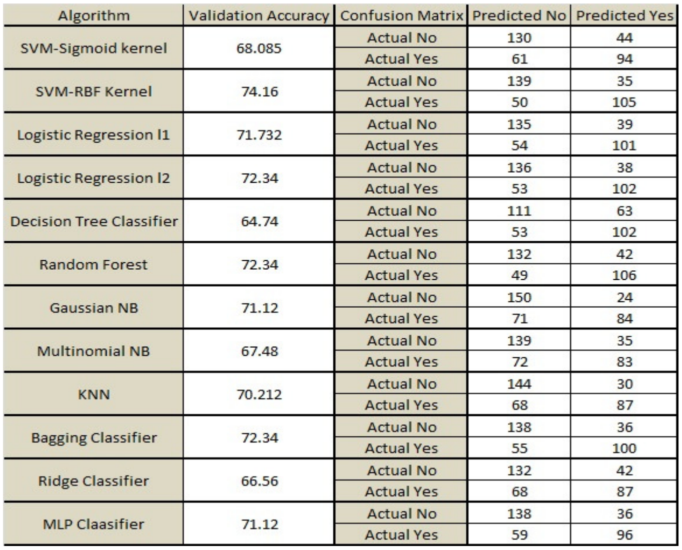


Matrix	Predicted NO	Predicted YES	Total
Actual NO	TN	FP	TN + FP
Actual YES	FN	TP	FN + TP
Total	TN + FN	FP + TP	TN + FP +FN + TP

Figure 7: Confusion Matrix

In figure 7, we present an example of a confusion matrix and the interpretation of it. The metric Accuracy is the ratio of the sum of TN and TP to the sum of TN, TP, FN and FP. Apart from the accuracy, we believe that we have to diminish the False Negative Rate which is the ratio between the FN and sum of TN and FN. Using these two metrics we define our own metric which we use to select the best algorithm i.e. $Selection_{value}$.

We want to penalize the model for predicting wrong for a patient having the chance for heart attack or heart problem but predictive No for that case. This we have taken into the consideration because the patients who have the chance of suffering from any heart problem cannot be predicted wrong. Using the above metric as Selection Value, we have found that particular algorithm in both the cases which gives that particular algorithm as a trade-off. Below are the results for the verification of different algorithms which are present in the model. All the accuracy that we present it here are the validation accuracy. It is how correctly the algorithm has predicted the validation set. 329 samples out of the total dataset is used for the validation set. The algorithm that we have used in our model are SVM-rbf, SVM-sigmoid, Logistic Regression, Decision Tree Classifier, Random Forest, Naive Bayes.

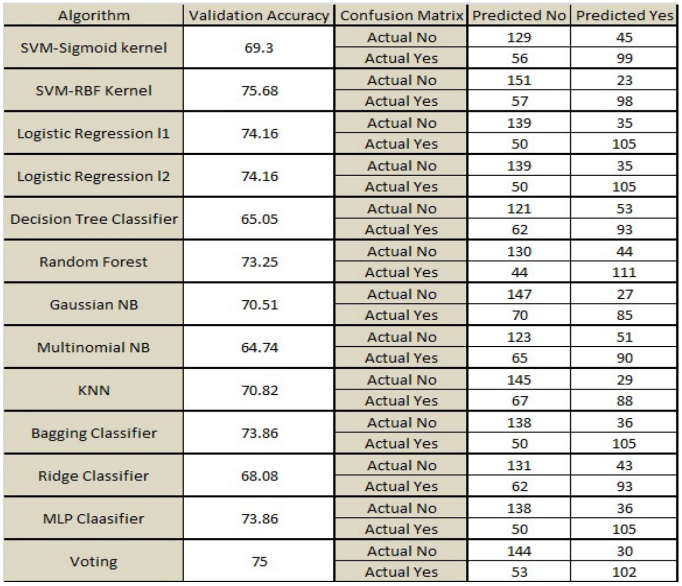


Algorithm	Validation Accuracy	Confusion Matrix	Predicted No	Predicted Yes
SVM-Sigmoid kernel	68.085	Actual No	130	44
		Actual Yes	61	94
SVM-RBF Kernel	74.16	Actual No	139	35
		Actual Yes	50	105
Logistic Regression l1	71.732	Actual No	135	39
		Actual Yes	54	101
Logistic Regression l2	72.34	Actual No	136	38
		Actual Yes	53	102
Decision Tree Classifier	64.74	Actual No	111	63
		Actual Yes	53	102
Random Forest	72.34	Actual No	132	42
		Actual Yes	49	106
Gaussian NB	71.12	Actual No	150	24
		Actual Yes	71	84
Multinomial NB	67.48	Actual No	139	35
		Actual Yes	72	83
KNN	70.212	Actual No	144	30
		Actual Yes	68	87
Bagging Classifier	72.34	Actual No	138	36
		Actual Yes	55	100
Ridge Classifier	66.56	Actual No	132	42
		Actual Yes	68	87
MLP Claasifier	71.12	Actual No	138	36
		Actual Yes	59	96

Figure 8: Results 1

In the figure 8, the results for the various algorithm is analyzed with the 0 as the imputation for the missing values. In this if we check the accuracy alone, SVM with rbf kernel gives a better result with 74.16 % accuracy followed by 72.34% with Random Forest, Bagging and Logistic Regression l2 norm. Apart from the above accuracy measure, we want to minimize the False Negative Rate i.e Actual is 1 but predicted is 0. The algorithm that best performed is SVMrbf with 29.118%.

In the figure 9, we have imputed the missing values if present in the data, with the maximum frequency present and we see increased values or accuracies for all the algorithms and SVM-rbf performed better with 75.68% accuracy. But if we check the False-Negative Rate, Random Forest performed better in this category. Even in the previous scenario, Random forest had this actual number lesser but the rate was higher. When checked with the Selection Value, Random Forest is the better algorithm with selection probability of 0.741 in comparison to 0.738 of SNM-rbf. These results will pop up in the section 3 of the Dash board and will take a decision making in case of the prediction.



Algorithm	Validation Accuracy	Confusion Matrix	Predicted No	Predicted Yes
SVM-Sigmoid kernel	69.3	Actual No	129	45
		Actual Yes	56	99
SVM-RBF Kernel	75.68	Actual No	151	23
		Actual Yes	57	98
Logistic Regression l1	74.16	Actual No	139	35
		Actual Yes	50	105
Logistic Regression l2	74.16	Actual No	139	35
		Actual Yes	50	105
Decision Tree Classifier	65.05	Actual No	121	53
		Actual Yes	62	93
Random Forest	73.25	Actual No	130	44
		Actual Yes	44	111
Gaussian NB	70.51	Actual No	147	27
		Actual Yes	70	85
Multinomial NB	64.74	Actual No	123	51
		Actual Yes	65	90
KNN	70.82	Actual No	145	29
		Actual Yes	67	88
Bagging Classifier	73.86	Actual No	138	36
		Actual Yes	50	105
Ridge Classifier	68.08	Actual No	131	43
		Actual Yes	62	93
MLP Claasifier	73.86	Actual No	138	36
		Actual Yes	50	105
Voting	75	Actual No	144	30
		Actual Yes	53	102

Figure 9: Results 2

Thus, with the view of the above results, we have used the type-2 case for the data processing and as from the validation score from the Selection Value, Random Forest as the brain behind the model. The algorithm can vary whenever a new patient details is fed into the system.

6. CONCLUSION

In the above procedure, we not only want to maximize the accuracy of the algorithm that we select to help the doctor take a decision rather, we want to decrease and penalize the model for having a bad prediction for the cases where the patient has a high probability for the heart attack but the model predicting for no heart problem. We hence stated one new metric called Selection Value which takes care of these scenarios and selects that algorithm which gives maximum S.V. We do not want to bias the doctor with the results of the classification rather as discussed in the proposed scenario section; we try to give the doctor with the better option with the history similar data results. Using these data, the doctor can have a transparency with the patient and the patient won't feel cheated at the end. With the more amounts of data being fed into the data base, the system will be very intelligent.

REFERENCES

[1] Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms, Sanjay Kumar Sen

[2] Peylan-Ramu, Nili, et al. ”Abnormal CT scans of the brain in asymptomatic children with acute lymphocytic leukemia after prophylactic treatment of the central nervous system with radiation and intrathecal chemotherapy.” New England Journal of Medicine 298.15 (1978): 815-818.

[3] Decramer, Isabel, et al. ”Effects of sublingual nitroglycerin on coronary lumen diameter and number of visualized septal branches on 64-MDCT angiography.” American Journal of Roentgenology 190.1 (2008): 219-225.

[4] Alkhorayef M, Babikir E, Alrushoud A, Al-Mohammed H, Sulieman A. Patient radiation biological risk in computed tomography angiography procedure. Saudi Journal of Biological Sciences. 2017;24(2):235-240. doi:10.1016/j.sjbs.2016.01.011.

[5] Diaz, Marco N., et al. ”Antioxidants and atherosclerotic heart disease.” New England Journal of Medicine 337.6 (1997): 408-416.

[6] Rodgers, Anthony, et al. ”Blood pressure and risk of stroke in patients with cerebrovascular disease.” Bmj 313.7050 (1996): 147.

[7] Gertler, Menard M., et al. ”Ischemic heart disease.” Circulation46.1 (1972): 103-111.

[8] Diamond, Joseph A., and Robert A. Phillips. ”Hypertensive heart disease.” Hypertension research 28.3 (2005): 191-202.

[9] Leander, Karin, et al. ”Family history of coronary heart disease, a strong risk factor for myocardial infarction interacting with other cardiovascular risk factors: results from the Stockholm Heart Epidemiology Program (SHEEP).” Epidemiology 12.2 (2001): 215-221.

[10] US Department of Health and Human Services. ”The health consequences of smoking: a report of the Surgeon General.” (2004): 62.

[11] Hjermann, I., et al. ”Effect of diet and smoking intervention on the incidence of coronary heart disease: report from the Oslo Study Group of a randomised trial in healthy men.” The Lancet318.8259 (1981): 1303-1310.

[12] Collins, Rory, et al. ”Blood pressure, stroke, and coronary heart disease: part 2, short-term reductions in blood pressure: overview of randomised drug trials in their epidemiological context.” The Lancet 335.8693 (1990): 827-838.

[13] Wolf, Philip A., Robert D. Abbott, and William B. Kannel. ”Atrial fibrillation as an independent risk factor for stroke: the Framingham Study.” Stroke 22.8 (1991): 983-988.

[14] Miller, M. ”Dyslipidemia and cardiovascular risk: the importance of early prevention.” QJM: An International Journal of Medicine 102.9 (2009): 657-667.

[15] Haffner, Steven M., et al. ”Reduced coronary events in simvastatin-treated patients with coronary heart disease and diabetes or impaired fasting glucose levels: subgroup analyses in the Scandinavian Simvastatin Survival Study.” Archives of Internal Medicine 159.22 (1999): 2661-2667.

[16] Emerging Risk Factors Collaboration. ”Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies.” The Lancet 375.9733 (2010): 2215-2222.

[17] Jee, Sun Ha, et al. ”A coronary heart disease prediction model: the Korean Heart Study.” BMJ open 4.5 (2014): e005025.

[18] Poirier, Paul, et al. ”Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss.” Circulation 113.6 (2006): 898-918.

[19] Ornish, Dean, et al. ”Can lifestyle changes reverse coronary heart disease?: The Lifestyle Heart Trial.” The Lancet336.8708 (1990): 129-133.

[20] Villareal, Dennis T., et al. ”Effect of lifestyle intervention on metabolic coronary heart disease risk factors in obese older adults.” The American journal of clinical nutrition 84.6 (2006): 1317-1323.

[21] Killip, Thomas, and Mary Ann Payne. ”High serum transaminase activity in heart disease.” Circulation 21.5 (1960): 646-660.

[22] Sowjanya, K., Ayush Singhal, and Chaitali Choudhary. ”MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices.” Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015.

[23] Pazzani, Michael J., and Daniel Billsus. ”Content-based recommendation systems.” The adaptive web. Springer, Berlin, Heidelberg, 2007. 325-341.

[24] Sagir, Abdu Masanawa, and Saratha Sathasivam. ”A Novel Adaptive Neuro Fuzzy Inference System Based Classification Model for Heart Disease Prediction.” Pertanika Journal of Science & Technology 25.1 (2017).

[25] Pattekari, Shadab Adam, and Asma Parveen. ”Prediction system for heart disease using Nave Bayes.” International Journal of Advanced Computer and Mathematical Sciences3.3 (2012): 290-294.

[26] Medhekar, Dhanashree S., Mayur P. Bote, and Shruti D. Deshmukh. ”Heart disease prediction system using naive Bayes.” Int. J. Enhanced Res. Sci. Technol. Eng 2.3 (2013).

[27] Peter, T. John, and K. Somasundaram. ”An empirical study on prediction of heart disease using classification data mining techniques.” Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on. IEEE, 2012.

[28] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. ”Prediction of heart disease using classification algorithms.” Proceedings of the world Congress on Engineering and computer Science. Vol. 2. 2014.

[29] Xing, Yanwei, Jie Wang, and Zhihong Zhao. ”Combination data mining methods with new medical data to predicting outcome of coronary heart disease.” Convergence Information Technology, 2007. International Conference on. IEEE, 2007.

[30] Ratnaparkhi, Devendra, Tushar Mahajan, and Vishal Jadhav. ”Heart Disease Prediction System Using Data Mining Technique.” International Research Journal of Engineering and Technology (IRJET) 2.08 (2015): 2395-0056.

[31] Sagir, Abdu Masanawa, and Saratha Sathasivam. ”A Novel Adaptive Neuro Fuzzy Inference System Based Classification Model for Heart Disease Prediction.” Pertanika Journal of Science & Technology 25.1 (2017).

[32] Sen, Ashish Kumar, Shamsher Bahadur Patel, and D. P. Shukla. ”A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level.” International Journal of Engineering and Computer Science 2.9 (2013): 1663-1671

AUTHORS

Nimai Chand Das Adhikari received his Master's in Machine Learning and Computing from Indian Institute of Space Science and Technology, Thiruvananthapuram in the year 2016 and did his Bachelor's in Electrical Engineering from College of Engineering and Technology in the year 2011. He is currently working as a Data Scientist for Philips Lighting (SS Supply Chain Solutions Pvt. Ltd.). He is a vivid researcher and his research interest areas include computer vision, health care and deep learning.

Arpana Alka received her Master's in Machine Learning and Computing from Indian Institute of Space Science and Technology, Thiruvananthapuram in the year 2017 and did her Bachelor's in Computer Science Engineering from National Institute of Technology, Surat in the year 2014. She is currently working as a Data Science Engineer for Busigence Technologies. Her interest areas include deep learning, video analytics, medical application and NLP.

Rajat Garg received his Bachelor's in Biotechnology Engineering from National Institute of Technology, Jalandhar and is currently working as Data a Scientist in Philips Lighting (SS Supply Chain Solutions Pvt. Ltd.). His interest areas include Machine Learning, Computer Vision and Data Analysis.