Online at: https://researchtrendsjournal.com ISSN No: 2584-282X

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 2; Issue 1; 2024; Page No. 175-182

Received: 02-10-2023
Accepted: 06-11-2023

Analysis of decision rules-based prediction of cardiac risk factors in disease prognosis

[1]Prathima Y and [2]Dr. Manish Saxena

[1]Research Scholar, Department of Computer Science, Himalayan University, Arunachal Pradesh, India
[2]Assistant Professor, Department of Computer Science, Himalayan University, Arunachal Pradesh, India

DOI: https://doi.org/10.5281/zenodo.12823682

Corresponding Author: Prathima Y

Abstract

Numerous serious illnesses can be cured by advancements in the pharmaceutical industry and the healthcare system. However, the main obstacle to preventing fatalities is determining the right timing for detection and the precision of the detection technique. A major illness that can be fatal to a person is coronary disease, which is characterized by a heart condition that tends to malfunction. Men are more likely than women to get chronic diseases, especially heart disease. Heart disease can be broadly classified as coronary artery disease, arrhythmias, and congenital heart illnesses, though it is primarily caused by a variety of factors. Heart disease has no symptoms at first, and people are more vulnerable if a chronic heart disease event takes place. According to Yang and Garibaldi, age, gender, high blood pressure, cholesterol, smoking, diabetes, physical inactivity, and obesity are common risk factors for heart disease. These risk variables can be divided into categories that are manageable and unmanageable. According to Hajar *et al.* (2017), non-manageable risk variables include age and family history, while manageable risk factors include smoking, physical activity, diet, and obesity. The suggested framework presents a novel method based on the blockages of the heart's major blood channels for determining the severity of heart illnesses utilizing a multilayer perceptron approach.

Keywords: Chronic heart disease, pharmaceutical, coronary, characterized, smoking, physical

Introduction

Smokers are particularly vulnerable, and high cholesterol increases the risk of atherosclerosis, a disorder in which fat deposits obstruct blood vessels and cause them to narrow. Damage to the heart muscle and cardiac failure can result from blood artery blockage. Globally, the leading cause of death is Chronic Heart Disease (CHD). Men who are 40 years of age or older may develop CHD. Inactivity raises the risk of heart disease and contributes to fat storage. Each person has different risk factors based on their lifestyle. It is difficult for doctors to rule out illness risk in its early stages. To rule out the presence of the disease, investigations, symptoms, and physical examinations are required. Patients must simultaneously invest a significant amount of money in clinical research. Years of experience are needed for an accurate forecast and diagnosis of heart disease, and not all doctors can gain this experience. Because the data provided by patients is largely redundant and cannot be connected to that of other patients. Additionally, because a single symptom may be associated with numerous diseases, patients tend to dismiss multiple symptoms. Physicians cannot rely on ambiguities that may arise during the majority of the clinical decision-making process (Hsu *et al.* 2018) [9].

Physicians require data on which to base their diagnoses, and the volume of data gathered by medical records is sufficient to screen through a manual procedure. Additionally, doctors must intervene quickly in cases with CHD because failure to do so could result in the patient's death. Physicians lack the funds and resources necessary to put up facilities that can handle CHD incidents, according to the case study. They frequently deal with patients who report chest pain in the case of CHD; however, in order to differentiate between chest pain that is related to the heart and that is not, additional factors including age, gender, and cardiovascular illnesses must be taken into consideration

(Ilayaraja & Meyyappan 2015) [10]. A minimal number of risk indicators are necessary to identify heart disease since chest discomfort, by itself, cannot be utilized to assess the risk of coronary heart disease (CHD). Determining the correlation between risk factors is crucial in these situations as well (Taslimitehrani *et al*. 2016) [11].

Application of decision rules

The main purpose of decision support systems is to assist physicians in making decisions. To make decisions, a decision support system may use the patient's medical history, clinical testing, and physical examination. The fact that decision support systems focus on facts rather than subjective judgments makes them crucial to use (Reilly & Evans 2006) [12]. To determine the relationships between the risk factors, physicians employ decision rules. A hospital and an emergency department may have different contexts for the same illness; there are always differences between the two (Buntinx *et al*. 2001) [13]. Systems for supporting decisions help to lower the cost and limits associated with analysis. Decision support rules have the potential to improve the efficacy of treatments and therapies while effectively addressing personnel shortages, healthcare system constraints, and healthcare intervention costs (Van *et al*. 2012) [3].

Determining decision rules is comparable to making predictions; the model's performance and accuracy have an impact on the caliber of the generated rules. In addition to generating conditions, decision rules also reflect actions.

Proposed methodology

The current study suggested a novel approach utilizing a probability function in particular to extract the risk elements from the rules produced in order to get beyond the shortcomings of the current system. The primary objectives of the suggested model are to produce guidelines and identify the risk variables that affect an illness's prognosis. Figure-1 shows the suggested model's schematic form. By calculating the likelihood that each feature level of a leaf node will occur, the probability function seeks to minimize the size of the tree with the best cut points while also enhancing the prediction probability of the classes and raising the likelihood of creating high-quality decision rules. Additionally, by comparing the leaf nodes on various splits, the best cut point is selected to split or reduce the tree based on the feature levels.

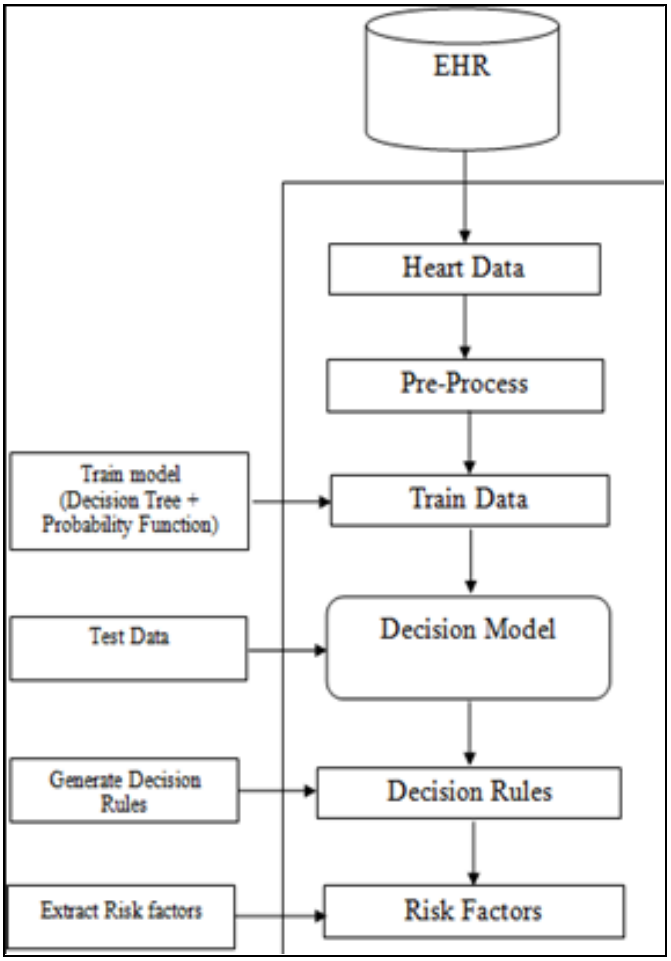

Fig 1: Flow diagram for building decision rules

A decision tree consists of three nodes: the root, branch, and leaf nodes. Each node in the tree defines a feature; the leaf node shows the result of the class labels; and the branch node defines the condition. Figure-2 shows an example of a decision tree with features, conditions, and an outcome. Beginning as a single node with the greatest amount of information, the root node branches into branch nodes, which then branch into leaf nodes. A decision rule is the path that a node takes from its root to its branch and the leaf node that corresponds to it.

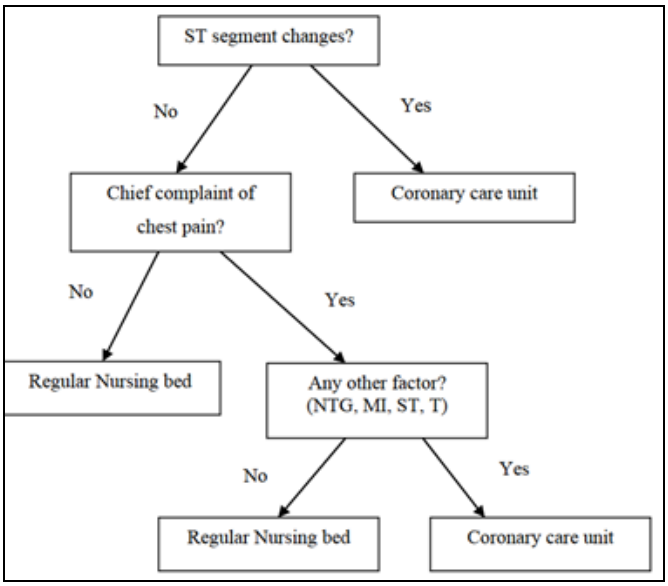

Fig 2: Example for decision tree

The feature with the greatest information content is designated as the root using entropy values, and the tree is divided based on the levels of that selected feature. The definition of entropy is

$$E(T) = -\sum_{d=1}^{|k|} p(c) log_2 p(c)$$

$$Gain\ (T,a) = E(T) - \sum_{v=1}^{AV} \frac{T^{Td}}{T}, E(T^{AV})$$

Where p(c) is the ratio of the items in c to the proportion of elements in T, T is the dataset, and c is the class inside dataset T. The features of the dataset T are represented as f={f1,f2,f3...fn}, where fn denotes the number of features in the dataset as {1,2,3,...}, and p(c) is the percentage of a feature tuple in T that falls into a specific class. Ci, Td is the total number of elements in the training data, and Tav is the various subsets of levels in a feature. An attribute can have several levels of values, and each level is represented by the notation AV={a1,a2,...av}, where av is the total number of levels or values for the attribute.
Algorithm 1

Input: dataset T Output: a decision tree
- Create node
- If sample comes from the same class. Then
- The node is as leaf node and named as Ci; Return 4 End if
- If Tav = 0 or there are no sub values in T, Then
- The node is labelled as leaf node; Return 7 Else
- info gain is used to select the best sublevels of fi from f using info gain;
- For every Tav in fn do
- Create a branch for node; Av is the sample subset from Tav in T
- If Tav is zero. Then
- The node is labelled as a leaf node; Return
- Else
- Define the branch as a branch node, return Step 8. 15
- End if
- End For
- End if
- End for

The likelihood degree of each level Tav of a feature on the branch nodes is determined using the probability function. T is represented as T={(x1,y1),(x2,y2),...(xn,yn)}, and the features are represented as f={f1,f2,f3...fn}; fn denotes the number of features as {1,2,3,...}, and the labels are represented as Ci={C1,C2...Cn}. Equation is used to determine the probability based on the presence of each level in a feature.

$$P(T_i^{av} \rightarrow T) = |T_{fi}^{av} \cup Ci| \, |T_{fi}^{av}|$$

Rule is given by

$$R(T_{fi}^{av}) = |T_{fi}^{av} \cup Ci| \, / | \, TT_{fi}^{av} \, |$$

Where |T av| is the number of elements contained in a specific sublevel of a feature, T is the input data, Fi is the attribute levels, Ci is the labels in the data, and T av is the subset for each attribute level. Algorithm 2 is used to build the suggested decision tree using the probability function equation.

Algorithm 2
Input: dataset T Output: a decision tree 1 Create node
If sample comes from the same class. Then
The node is as leaf node and named as Ci; Return 4 End if
If Tav = 0 or there are no sub values in T, Then
The node is labelled as leaf node; Return 7 Else
info gain is used to select the best sublevels of fi from f using info gain;
For every value Tav in f_n do
Create a branch for node; Av is the sample subset from Tav in T
If Tav is zero. Then
The node is labelled as a leaf node; Return
Else
define the branch as a branch node, return
Step 8. 15 For each branch of Tav do
Calculate split level by using probability function
replace condition attributes based on probability value 17
define the branch as a branch node, return Step 8.
End if
End for
End if
End For 22
End if
The primary benefit of the suggested model is in its capacity to prevent overfitting, as the probability of each level is determined by counting its occurrences, and the sublevels used to divide a node are determined by the items found in each feature's sublevels. Since the process of growing a tree is recursive, the tree tends to converge further toward generalization after pruning. Pruning the tree minimizes prediction errors by removing low- discriminating leaf and branch nodes. In order to maintain the number of splits (N), pruning proceeds from leaf nodes to root nodes, with the

error improvement being computed. The number of nodes on the branches is kept if the error rate decreases; if not, the branch node is substituted with a leaf node.

Experiment and Analysis

Dataset

The UCI repository's heart (SA), thyroid, and diabetes datasets are used to assess the performance of the suggested model. Details about the dataset are provided in Table-1. Heart (SA), Diabetes, and Thyroid are included in the dataset. There are two classes and 768 instances of nine features in the Diabetes dataset. The Thyroid dataset has 22 attributes, 7200 instances, and three classes: class 1 represents normal thyroid status, class 2 represents hypothyroidism, and class 3 represents hyperthyroidism. Class 1 represents the presence of diabetes illness, while class 2 represents its absence. The Heart (SA) dataset includes two classes, 462 records, and 10 characteristics. In the cardiac dataset, class 1 denotes the absence of heart disease, while class 2 indicates the presence of heart illness.

Table 1: Dataset used in the study

S. No	Dataset	Instances	No of Features
1	Thyroid	7200	22
2	Diabetes	768	9
3	Heart (SA)	462	10

Model Building

40% of the dataset is designated as a testing set and 60% as a training set in order to test the suggested risk estimate model. The training set is used to train the suggested model, while the testing set is used to assess it. As covered in section 3.4, evaluation metrics are used to assess the model's performance.

Results and Discussion

Building decision tree models on the training set and evaluating the model on the testing set allows researchers to study the prediction of heart disease risk variables using decision trees. Decision tree models are built using Algorithms 1 and 2, and the results for the two models are contrasted. Figure-3 displays the decision tree created for the diabetic dataset, Figure-4 displays the decision tree created for the heart dataset, and Figure-5 displays the decision tree created for the thyroid dataset. For the heart (SA) dataset, the decision tree model's accuracy is 82.7%; for the diabetes dataset, it is 87.4%; and for the thyroid dataset, it is 98.3%. Using the probability function increased the model's accuracy to 89% on the heart dataset, 92% on the diabetes dataset, and 99% on the thyroid dataset. Table-2 provides the accuracy details for both the suggested and the current approaches. Using the decision rules from each model, the risk factors are obtained and assessed. In the next sections, the number of rules produced by each model and their importance as risk markers for heart disease are discussed in detail.

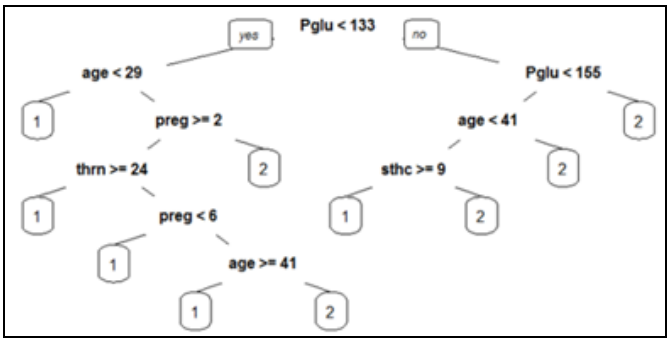

Fig 3: Decision tree for diabetes dataset

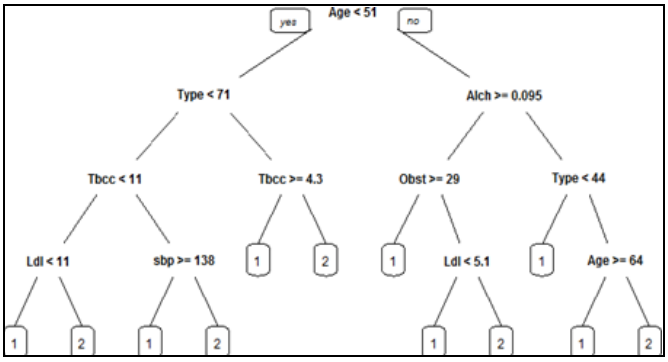

Fig 4: Decision tree for heart dataset

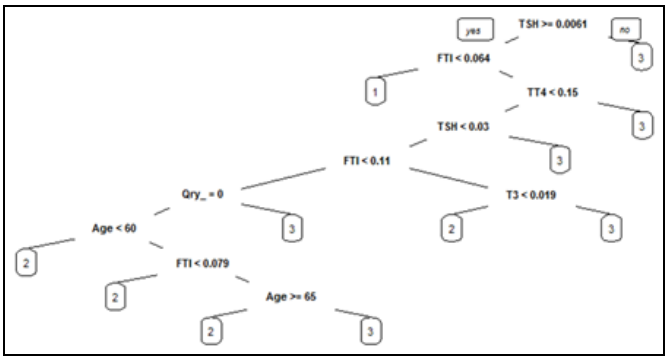

Fig 5: Decision tree for thyroid dataset

Table 2: Accuracy of the models

	Heart	Diabetes	Thyroid
DTPR	83%	86%	97%
DT	82.7%	87.4%	98.3%
Proposed Method	89.20%	92.20%	99.20%

With the help of a probability function, the suggested strategy produced 16 rules for the heart dataset, 17 rules for diabetes, and 11 rules for the thyroid dataset. Table-3 displays the eight rules that the decision tree produced for the heart dataset, ten rules for diabetes, and eight rules for the thyroid dataset. Table-4 contains the decision rules developed for the diabetes dataset, Table-5 contains the decision rules generated for the thyroid dataset, and Table-6 contains the decision rules generated for the heart dataset.

Table 3: Number of rules generated

No of Rules	Heart	Diabetes	Thyroid
DT	8	10	8
Proposed Method	16	17	13

Table 4: Rules generated for diabetes dataset using decision tree

No	Rules	Class
1	Pglu< 137.5 & age< 28.5	1
2	Pglu< 137.5 & age>=28.5 & pedigree< 0.202	1
3	Pglu<137.5 & age>=28.5 & pedigree>=0.202 &sbp>=77 &sthick< 30.5	1
4	Pglu<137.5 & age>=28.5 & pedigree>=0.202 &sbp>=77 &sthick>=30.5	2
5	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp<77 &Pglu< 120.5 & pedigree>=0.379 &	1
6	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp<77 &Pglu< 120.5 & pedigree< 0.379	2
7	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp<77 &Pglu>=120.5	2
8	Pglu>=137.5 &Pglu< 154.5 & age< 40	1
9	Pglu>=137.5 &Pglu< 154.5 & age>=40	2
10	Pglu>=137.5 &Pglu>=154.5	2

Table 5: Rules generated for thyroid dataset using decision tree

No	Rules	Class
1	TSH>=0.00605 & FTI< 0.06446	1
2	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435&T4U< 0.0935 &	2
3	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435& T4U>=0.0935& T3>=0.0214	2
4	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 & T4U>=0.0935 & T3< 0.0214 & T3< 0.02005 & T3>=0.0165	2
5	TSH>=0.00605& FTI>=0.06446 & TT4< 0.1435 T4U>=0.0935 & T3< 0.0214 & T3< 0.02005 & T3< 0.0165 & TSH>=0.0081	2
6	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 & T4U>=0.0935 & T3< 0.0214 & T3< 0.02005 & T3< 0.0165 & TSH< 0.0081	3
7	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 & T4U>=0.0935 & T3< 0.0214 & T3>=0.02005	3
8	TSH>=0.00605 & FTI>=0.06446 & TT4>=0.1435	3

Table 6: Rules generated for heart dataset using decision tree

No	Rules	Class
1	Age>=29.5 &Typea< 66.5 & Tobacco< 0.46	1
2	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp>=145.5	1
3	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp<145.5 &Typea>=45 &Typea< 50.5	1
4	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp<145.5 &Typea>=45 &Typea>=50.5&Tobacco>=5.25	1
5	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp<145.5 &Typea>=45 &Typea>=50.5&Tobacco< 5.25 &Age>=48	1
6	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp<145.5 &Typea>=45 &Typea>=50.5 &Tobacco< 5.25 &Age< 48	2
7	Age>=29.5 &Typea< 66.5 &Tobacco>=0.46 &sbp<145.5 &Typea< 45 &	2
8	Age>=29.5 &Typea>=66.5	2

Table-7 contains the decision rules developed for the thyroid dataset; Table-8 contains the decision rules generated for the diabetic dataset; and Table-9 contains the decision rules generated for the heart dataset. The suggested approach and the decision rules produced by DT diverge greatly. The probability calculation of each feature's sublevel accounts for the discrepancy. Regarding heart disease, the suggested DT rule indicates that Age>=29.5 &Typea< 66.5 & Tobacco>=0.46 &sbp>=145.5 relates to no heart disease. If you are under 29.5 years old and have not yet had heart disease, your score is less than 29.5 years old and your sbp is less than 132. This indicates that you have heart disease. The suggested approach effectively captures the event that corresponds to the lower of sbp as a risk factor for developing heart disease, and the prediction levels of sbp are further refined by the subset of features in the probability function. While the suggested DT indicates that sbp<132 has the danger of developing heart disease and sbp>=132 has no heart disease, the rules from DT demonstrate that sbp>=145.5. Age >= 29.5 and tobacco >= 2 remain constant for both models in the two rules.

DT created the rule for the likelihood of not developing heart disease. The suggested technique produced the following rule: Age>=29.5 &Typea< 68.5 & Tobacco>=7.94 & Tobacco<12.17 & Obesity< 29.88 & Adiposity>=17.27 as the risk of developing cardiac diseases: sbp< 145.5 &Typea>=45 &Typea>=50.5 & Tobacco< 5.25 & Age>=48. While the suggested technique predicted the risk of developing heart disease using age, typea, tobacco, obesity, and adiposity, the DT developed the rule utilizing features like age, tobacco, and sbp to predict no heart disease. The risk variables that have been found may serve as indicators of heart disease.

The decision rule changed as a result of taking into account the likelihood of each feature's sublevels while creating rules. By taking into account the nodes and utilizing the maximum information feature, DT produced the decision rules. The nodes are based on subsets of the features since the probability function defines the information by calculating the likelihood that a feature will contain sublevels. The variation in the number of node levels in both models can be explained by the trees' complexity parameter and validation error.

Table 7: Rules generated for thyroid dataset (proposed)

No	Rules	Class
1	TSH>=0.00605 & FTI< 0.064461	1
2	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 & Thyroid_ surgery< 0.5 & T4U< 0.0935 & TT4>=0.0585	2
3	TSH>=0.00605 & FTI>=006446 & TT4< 0.1435 & Thyroid_surgery< 0.5 & T4U< 0.0935 TT4< 0.0585	3
4	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI<0.1035 & T3>=0.015	2
5	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI< 0.1035 & T3< 0.015 & TSH>=0.007995 & TSH< 0.0275	2
6	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI< 0.1035 & T3< 0.015 & TSH>=0.007995 & TSH>=0.02753	3
7	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI< 0.1035 & T3< 0.015 & TSH< 0.007995	3
8	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI>=0.1035 & TT4>=0.1141 & TSH< 0.018	2
9	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI>=0.1035 & TT4>=0.1141 & TSH>=0.0183	3
10	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery< 0.5 & T4U>=0.0935 & FTI>=0.1035 & TT4< 0.1141	3
11	TSH>=0.00605 & FTI>=0.06446 & TT4< 0.1435 &Thyroid_surgery>=0.5	3
12	TSH>=0.00605 & FTI>=0.06446 & TT4>=0.1435	3
13	TSH< 0.00605	3

Table 8: Rules generated for diabetes dataset (proposed)

No	Rules	Class
1	Pglu<137.5 & age< 28.5 &sbp< 103	1
2	Pglu<137.5& age< 28.5 &sbp>=103	2
3	Pglu<137.5 & age>=28.5 & pedigree< 0.202	1
4	Pglu<137.5 & age>=28.5 & pedigree>=0.202 &sbp>=77 &sthick< 30.5	1
5	Pglu<137.5 & age>=28.5 & pedigree>=0.202 &sbp>=77 &sthick>=30.5 &sbp>=86	1
6	Pglu<137.5 & age>=28.5 & pedigree>=0.202 &sbp>=77 &sthick>=30.5 &sbp< 86	2
7	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp< 77 &Pglu< 120.5 &Pglu>=116.5	1
8	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp< 77 &Pglu< 120.5 &Pglu< 116.5	2
9	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp< 77 &Pglu>=120.5 & mass< 27.2	1
10	Pglu< 137.5 & age>=28.5 & pedigree>=0.202 &sbp< 77 &Pglu>=120.5 & mass>=27.2	2
11	Pglu>=137.5 &Pglu< 154.5 & age< 40 &sbp>=57 & mass< 42.2	1
12	Pglu>=137.5 &Pglu< 154.5 & age< 40 &sbp>=57 & mass>=42.2	2
13	Pglu>=137.5 &Pglu< 154.5 & age< 40 &sbp< 57	2
14	Pglu>=137.5 &Pglu< 154.5 & age>=40 &preg< 2	1
15	Pglu>=137.5 &Pglu< 154.5 & age>=40 &preg>=2	2
16	Pglu>=137.5 &Pglu>=154.5 & age>=62.5	1
17	Pglu>=137.5 &Pglu>=154.5 & age< 62.5	2

Table 9: Rules generated for heart dataset (proposed)

No	Rules	Class
1	Age< 29.5 &Tobacco< 2.025	1
2	Age< 29.5 &Tobacco>=2.025 &sbp>=132	1
3	Age< 29.5 &Tobacco>=2.025 &sbp< 132	2
4	Age>=29.5 &Typea< 68.5 &Tobacco< 7.941&Tobacco>=5.2 &Typea< 63	1
5	Age>=29.5 &Typea< 68.5 & Tobacco< 7.94&Tobacco>=5.2 &Typea>=63 &sbp>=130	1
6	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94&Tobacco>=5.2 &Typea>=63 &sbp< 130	2
7	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94 &Tobacco<5.2 &Tobacco< 0.46 &&Alcohol< 20.68	1
8	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94 &Tobacco<5.2 &Tobacco< 0.46 &Alcohol>=20.68	2
9	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94 &Tobacco<5.2 &Tobacco>=0.46 &Alcohol>=30.81	1
10	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94 &Tobacco<5.2 &Tobacco>=0.46 &Alcohol< 30.81 &Typea< 50.5	1
11	Age>=29.5 &Typea< 68.5 &Tobacco< 7.94 &Tobacco<5.2 &Tobacco>=0.46 &Alcohol< 30.81 &Typea>=50.5	2
12	Age>=29.5 &Typea< 68.5 &Tobacco>=7.94&Tobacco>=12.17	1
13	Age>=29.5 &Typea< 68.5 &Tobacco>=7.94&Tobacco< 12.17 &Obesity>=29.88	1
14	Age>=29.5 &Typea< 68.5 &Tobacco>=7.94&Tobacco< 12.17 &Obesity< 29.88 &Adiposity< 17.27	1
15	Age>=29.5 &Typea< 68.5 &Tobacco>=7.94&Tobacco< 12.17 &Obesity< 29.88 &Adiposity>=17.27	2
16	Age>=29.5 &Typea>=68.5	2

Better rules are produced by maximizing the margins in a decision tree, which is sized according to the complexity parameter. Furthermore, the validation error influences the maximizing. Only in cases where there is an increase in

cross validation error is maximizing advantageous. The tree cannot split enough data if there is no increase in cross validation. Table-10 complexity parameter for the suggested model demonstrates an enhanced split level from the DT model. While the intended DT contains six nodes, the DT model only has three. The error rate decreases to 0.233 for node split 3 and 0.0232 for node split 16 when node split is maximized from 3 to 14.

Table 10: Complexity parameter table for heart dataset using decision tree

CP	N split	rel error	X error	X std
0.069767	0	1	1	0.12673
0.046512	3	0.76744	1.3953	0.1358
0.01	4	0.72093	1.5116	0.1368

Table 11: Complexity parameter table for heart dataset (proposed)

	CP	N split	Rel error	X error	X std
1	0.077519	0	1	1	0.12673
2	0.069767	3	0.76744	1.2558	0.13364
3	0.046512	8	0.4186	1.2326	0.13317
4	0.023256	10	0.32558	1.1163	0.13037
5	0.011628	14	0.23256	1.186	0.13214
6	0.01	16	0.2093	1.2093	0.13267

The diabetes dataset shows that the splits are maximized at node 11, where there is a decrease in error rate and an improvement in model performance. Table-12 illustrates the error rate gain to around 0.0525 for 8 nodes and 0 for 11 nodes. According to Table-13, the thyroid dataset's maximum split of 14 can be attained with an error rate gain of 0.0129.

Table 12: Complexity parameter table for diabetes dataset (proposed)

	CP	N split	Rel error	X error	X std
1	0.273684	0	1	1	0.078723
2	0.189474	1	0.72632	1.01053	0.078845
3	0.052632	2	0.53684	0.71579	0.072915
4	0.036842	3	0.48421	0.70526	0.072599
5	0.031579	5	0.41053	0.72632	0.073224
6	0.021053	8	0.31579	0.67368	0.071601
7	0.015789	11	0.25263	0.67368	0.071601
8	0.010526	13	0.22105	0.65263	0.070894

Table 13: Complexity parameter table for thyroid dataset (proposed)

	CP	N split	Rel error	X error	X std
1	0.3051948	0	1	1	0.077657
2	0.2857143	1	0.69481	0.9026	0.074053
3	0.0454545	2	0.40909	0.42208	0.051559
4	0.0324675	3	0.36364	0.44805	0.053071
5	0.025974	5	0.2987	0.43506	0.052321
6	0.0194805	6	0.27273	0.4026	0.050391
7	0.0162338	8	0.23377	0.4026	0.050391
8	0.012987	12	0.16883	0.4026	0.050391
9	0.0064935	14	0.14286	0.38961	0.049595

Table-14 displays the accuracy of the decision tree (DT) model for the heart (SA) dataset at 82.7%, the diabetes dataset at 87.4%, and the thyroid dataset at 98.3%. The decision tree-prism model (DTPR) that is now in use has an accuracy of 83% for heart disease, 86% for diabetes, and 97% for thyroid illness. Using the probability function, the suggested model's accuracy was 99.20% on the thyroid dataset, 92.20% on the diabetes dataset, and 89.20% on the heart dataset. The suggested model obtained 94% accuracy, DT achieved 89% accuracy, and DTPR achieved 89% accuracy on average.

Table 14: Accuracy of the models vs datasets

Models	Heart	Diabetes	Thyroid	Average
DTPR	83%	86%	97%	89%
DT	82.7%	87.4%	98.3%	89%
Proposed Method	89.20%	92.20%	99.20%	94%

However, the accuracy of the proposed model is lower than that of association rule-PSO-SVM (98.9%), the Cluster-SVM+ The suggested approach produced better outcomes, and it is advised for the identification of heart disease risk factors. The likelihood that a subset will occur in a feature enhanced the model's performance and deepened the tree. Consequently, a more accurate construction of the decision rules is made using the risk factors and their sublevels. The deeper a tree gets, the more rules it generates as well. The number of rules grew from 8 to 16 for the heart dataset, from 10 to 17 for the diabetes dataset, and from 8 to 11 for the thyroid dataset. On three datasets, the probability-based function works & enhances the creation of rules & trees without degrading the performance of the model.

Conclusion

This paper goes into great length regarding the application of the suggested decision tree model for heart disease in the prediction of heart disease risk factors in illness prognosis. A probability function is used by the decision tree model to enhance the depth and tree building. An increasing number of tree splits occur as tree depth increases. The subsets of values contained in a feature are correlated with the gain in tree splits. Feature levels are used in the performance analysis, and decision rules are created. The decision rules filter out the risk elements. The heart, thyroid, and diabetes datasets are used to compare the performance of the suggested model, and its benefits are examined. The outcomes and conclusions reveal that the suggested model has a superior prediction capacity with more rules across the board for all datasets.

References

1. Voruganti S. Effective IoT Techniques to Monitor the Levels of Garbage in Smart Dustbins. International Research Journal of Engineering and Technology. 2020;7(6):6549-6554.
2. Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iran J Basic Med Sci. 2016;19(5):476-482.
3. Camm AJ, Lip GY, De Caterina R, Savelieva I, Atar D, Hohnloser SH, *et al*. Focused update of the ESC guidelines for the management of atrial fibrillation: an update of the 2010 ESC guidelines for the management of atrial fibrillation developed with the special contribution of the European Heart Rhythm Association. European heart journal. 2012;33(21):2719-

2747.
4. Sairam U. Multi-Functional Blind Stick for Visually Impaired People. IEEE Explore; c2020. p. 895-899. ISBN:978-1-7281-5371-1.
5. Sairam U, Bhanu Prakash MV. DL And ML Approaches Along with Blockchain Towards IoT Security. Int J Adv Sci Technol. 2020;29(4s):826-832.
6. Guru N, Dahiya A, Rajpal N. Decision Support System for Heart Disease Diagnosis Using Neural Network. Delhi Bus Rev. 2007;8(1):99-101.
7. Voruganti S. Local Security Enhancement and Intrusion Prevention in Android Devices. Int Res J Eng Technol. 2020;7(1):205-211.
8. Kunta V, Tuniki C, Sairam U. Multi-Functional Blind Stick for Visually Impaired People. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES); c2020. p. 895-899.
9. Hsu CT, Matsuo T, Liu JY. Impact of assimilating the FORMOSAT-3/COSMIC and FORMOSAT-7/COSMIC-2 RO data on the Midlatitude and low-latitude ionospheric specification. Earth and Space Science. 2018;5(12):875-890.
10. Ilayaraja M, Meyyappan T. Efficient data mining method to predict the risk of heart diseases through frequent itemsets. Procedia Computer Science. 2015;70:586-592.
11. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. Journal of biomedical informatics. 2016;60:260-269.
12. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Annals of internal medicine. 2006;144(3):201-209.
13. Buntinx F, Knockaert D, Bruyninckx R, De Blaey N, Aerts M, Knottnerus JA, *et al*. Chest pain in general practice or in the hospital emergency department: is it the same?. Family Practice. 2001;18(6):586-589.