Rule Generation of Cataract Patient Data Using Random Forest Algorithm

Mamta Santosh Nair[1](✉) and Umesh Kumar Pandey[2](✉)

[1] MATS University, Raipur, India
mamtanair@yahoo.com
[2] MSIT, MATS University, Raipur, India
Umesh6326@gmail.com

Abstract. Cataract is one of the common problems among the humans. Cataract is the condition caused due to clouding of lens in the eye which eventually may lead to blindness. In last few years, data mining has been widely used to build the predictive model in various fields. In this paper, historical data of cataract patient has been used to build the predictive model. Random forest algorithm is one of the decision tree algorithms for predictive modeling. Random forest algorithm incorporates advantages of classification and regression. Present study uses random forest method to create a model for prediction of cataract. The random forest algorithm is also tested for Out of Bag estimation error.

Keywords: Data mining · Classification · Random forest · Rule generation · Out of bag error · Decision support system

1 Introduction

Decision making is a tough job. One decision relies on many factors. In data mining algorithms, decision tree algorithms are one of the most widely used algorithms for predictions. Random forest algorithm is one of such robust and simplistic algorithm that works on ensemble learning method.

Data mining is applied in medical science, astronomy and other field to extract information from the data set. This data set has large number of attribute and complexity of inferring information.

One of the major causes of blindness in the world is cataract. Cataract is the preventable blindness if the patient is operated in time. Several organizations worldwide are working towards spreading the word about cataract and also about surgeries performed for cataract. According to World Health Organization, Cataract is responsible for 51% of world blindness [15]. World Health Organization defines this condition as "Cataract is clouding of the lens of the eye which impedes the passage of light. Although most cases of cataract are related to the aging process, occasionally children can be born with the condition, or a cataract may develop after eye injuries, inflammation, and some other eye diseases" [15]. The statistics collected from many agencies and previous literatures are serious enough to take a giant step towards preserving the vison. Data of cataract

U. S. Tiwary and S. Chaudhury (Eds.): IHCI 2019, LNCS 11886, pp. 175–188, 2020.
https://doi.org/10.1007/978-3-030-44689-5_16

patients' needs to be studied and analyzed so as to reveal the hidden trends which can further be used to create awareness among general population.

In this paper, we have collected the data of patients with eye problems among which many are suffering from cataract. And the collection also includes other details of patients like dietary habits, addictions, living environment etc. which may help to predict the chances of getting cataract. Data mining algorithms carry out this assessment to assist in the decision making process.

2 Literature Review

Data mining can help see us what is not directly visible but is underlying the obvious. It finds out the pearls of patterns and trends from the oceans of data. Data mining performs analysis of information to find possible outputs [3]. The methods where the hidden trends of data are identified, analyzed and then categorized into useful knowledge is known as Data Mining [4]. It finds patterns or trends, which are interesting and useful too. It helps to see beyond all the knowledge. And finally, it allows one to decide upon facts and predict the classes. Data Mining can play a significant role in arranging the data into different classes [6].

Decision tree algorithm breaks the dataset multiple times from top to bottom approach and then later horizontally at the same level till all the data items belonging to a class are identified [5]. A decision tree structure is made of root, internal and leaf nodes. Most decision tree classifiers perform grouping or classification in two steps: firstly, a tree is grown fully and then shortening or trimming of trees are done. The tree is grown from the top first then it is divided further into branches till all class labels are identified. While trimming process is carried on, a tree is cut wherever required to improve the accuracy. The trimming begins from lowermost node [10].

A decision tree is like a flowchart in structure and layout where every inner node represents a condition on an attribute and each branch represents a yes/no result of the condition and class label is represented by each leaf node (or terminal node). The leaf node is the last node. Classification rules are generated going from the top node to the terminal node of the decision tree [2].

Classification algorithm learns in supervised environment. It finds out and allocates class labels to data items by applying the already acquired knowledge of class which the data records belong [1]. Classification technique can be solving several problems in different fields like medicine, industry, business, and science. Basically it involves finding rules that categorize the data into disjoint groups [14].

The objective of the classification is to build a model based on some example cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the model is used to predict the group attributes of new cases from the domain based on the values of other attributes [12].

Classification is the step wise process of finding a set of models which describe and performs allocation of data classes. The derived model is based on the analysis of a set of training data (i.e. data objects whose class label is known) [13].

Random Forest algorithm is a classifier model consisting of collection of trees or jungle like appearance where independent random vectors are distributed identically and every tree ends terminally for the accurate class [8]. At each step new random vector is

generated which is independent of the previous random vectors with same distribution and then forms a tree using the training set [9]. Random Forest uses decision Trees as base classifier. This ensemble learning method is used for classification and regression of data. An ensemble consists of number of trained models whose predictors are combined to classify new variables.

Random forests are an effective tool in prediction. Because of the Law of Large Numbers, they do not overfit. It inserts just the right amount of randomness and we get good and accurate classifiers and regressors [7]. The random selection of dimensions to choose the splitting variable can be done as well as the choice of coefficients for random combinations of features [11].

Nayer [18] did his research work on diabetes mellitus detection using machine learning. Stacking ensemble method used in this research work built upon linear discriminant analysis, recursive tree and KNN.

Beaulac and Rosenthal [19] studied undergraduate students of Canada university in past 10 years using random forest. Using random forest, they identified most important variable useful to the classifier that reveals information for the university administration.

Sugandhi, Yasodha, Kannan [20] used five classification algorithms for prediction of cataract. The algorithms used by them were Naïve Bayes, SMO, J48, REP Tree and Random Tree. Authors also found mean absolute error and correctly classified instance generated by all the algorithms. They found random forest algorithm to be most accurate classifier with prediction accuracy at 84.87%.

Niya [21] developed automatic cataract detection methodology. The methodology involved pre-processing, feature extraction and classification. SVM classifier was used for prediction of cataract and regression method used for grading of cataract.

3 Data Collection and Research Instrument

The research work uses cataract patient data for the study. Dataset used in this research work is primary data collected through questionnaire. Questionnaire has been designed in consultation with Ophthalmologists. The questionnaire has also been designed considering the factors responsible for cataract as per specified by World Health Organization website. World Health organization mentions smoking, diabetes mellitus, exposure to ultra violet rays and high body mass index to be some of the cataract causing parameters [15]. Keeping in view of all factors total 43 different parameters selected for the data collection. These parameters included personal details, food habits, medical and birth history and addictions etc. The target location of the data collection is Raigad District of Maharashtra, India. Questionnaire was prepared in English and Marathi language. This questionnaire distributed among the cataract patient of approximately 700. Because of low education, most of the respondents are not familiar with the questionnaire system, thus assistance provided for the form filling. The data include people of both genders of different age groups. The data also had good mix of rural including tribal as well as urban population. Total approx. 500 forms received and filled at the camps and outpatient department (OPD) of doctors. Only 297 forms found complete and were selected for analysis. Certain parameters in questionnaire have received no answers or very less amount of entries. Thus, those attributes were removed from the dataset and only 17 attributes were considered for the study.

From the dataset attribute 'cataract' is used as a class name and other 16 variables are predictor variable. The dataset is studied in R software for performing random forest algorithm. R has inbuilt packages for random forest. Packages used in this study are "randomForest", "dplyr", "readxl" and "reprtree". Table 1 represents Attribute name and symbolic name used in the code development and to increase the visibility of the tree.

Table 1. Attribute names and abbreviations

Symbolic name	Attribute name	Symbolic name	Attribute name	Symbolic name	Attribute name
A	Age	G	Addiction	M	Occupation history in years
B	Gender	H	Hypertension duration	N	Sun exposure in hours
C	Occupation	I	Diabetes duration	O	History of trauma (yes/no)
D	Height	J	Cholesterol duration	P	Spectacle use duration
E	Weight	K	Surgical history (yes/no)	Q	Cataract (yes/no)
F	Diet	L	Type of surgery		

4 Importance of Attributes

One of the most robust characteristics provided by Random Forest is the importance factor of attributes. Table 2 gives the list of attributes with their importance in Class 1 and Class 2. Table 2 also gives the Mean Decrease Accuracy and Mean Decrease Gini. Mean Decrease Accuracy is where values of variables are randomly permuted and it is also known as permutation importance.

Mean decrease Gini is also known as Gini importance. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a particular variable is used to split a node, the Gini coefficient for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). Attributes with a large mean decrease in accuracy are more important for classification of the data. In the given Table 2 attribute age is highest important with the Meandecrease accuracy at 29.2196387 followed by attribute type of surgery at 17.944127 and so on. Similarly, Meandecreasegini is highest for attribute age at 37.377705 followed by attribute weight and so on.

Table 2. Importance of attributes

Symbolic name	1	2	Mean Decrease Accuracy	Mean Decrease Gini
a	32.11137067	10.9070787	29.2196387	37.377705
b	0.04448175	1.3997418	1.1273503	1.243247
c	2.54318416	6.4052249	6.9375734	12.35707
d	0.33578291	1.7408604	1.4615954	13.086229
e	−0.83469508	6.668167	4.5231379	20.628917
f	1.16039169	3.4639696	3.1889527	2.61541
g	4.55669683	3.0163488	4.8897378	2.234993
h	7.7225398	0.4988208	6.067578	8.993176
i	7.57758079	−1.6094264	4.6023365	5.432224
j	0.78475064	−0.3456008	0.1164765	1.35444
k	7.65858699	2.7526541	7.989449	2.167219
l	20.67840801	2.376934	17.944127	6.086469
m	7.02724257	−0.3946136	4.6610465	8.43427
n	0.10091254	8.399354	5.9438672	9.882421
o	−0.51346518	6.1557188	3.4134775	2.085709
p	0.84577193	−0.3872764	0.2524296	13.915559

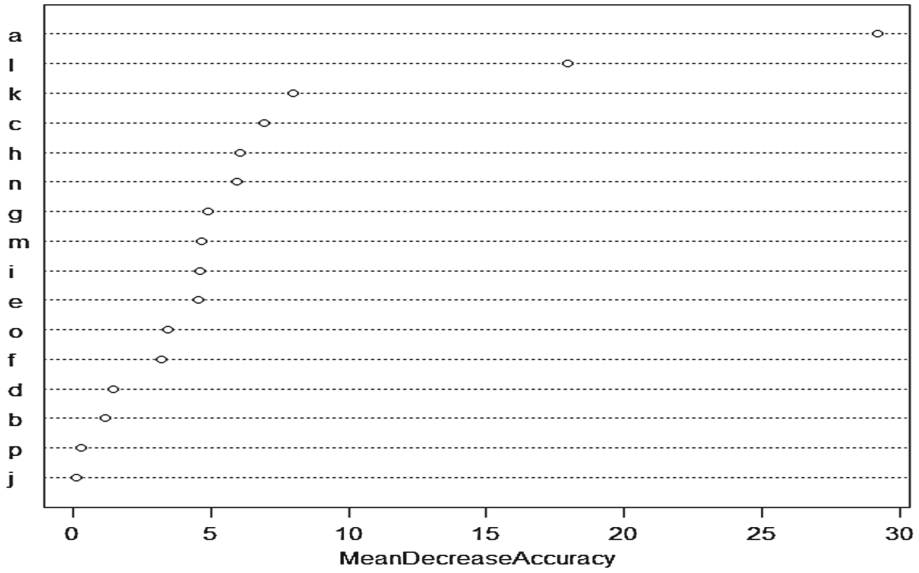

Fig. 1. Graph depiction of Mean Decrease Accuracy

The graph plotted for the variable importance is shown in Figs. 1 and 2. The plot shows each variable on the y-axis, and their importance on the x-axis. Attributes are ordered top-to-bottom as most- to least-important. Therefore, the most important attributes are at the top and an estimate of their importance is given by the position of the dot on the x-axis. Three least important variables were removed but OOB estimation error increased after removing it. Random forest algorithm used all 16 variables for rule generation.

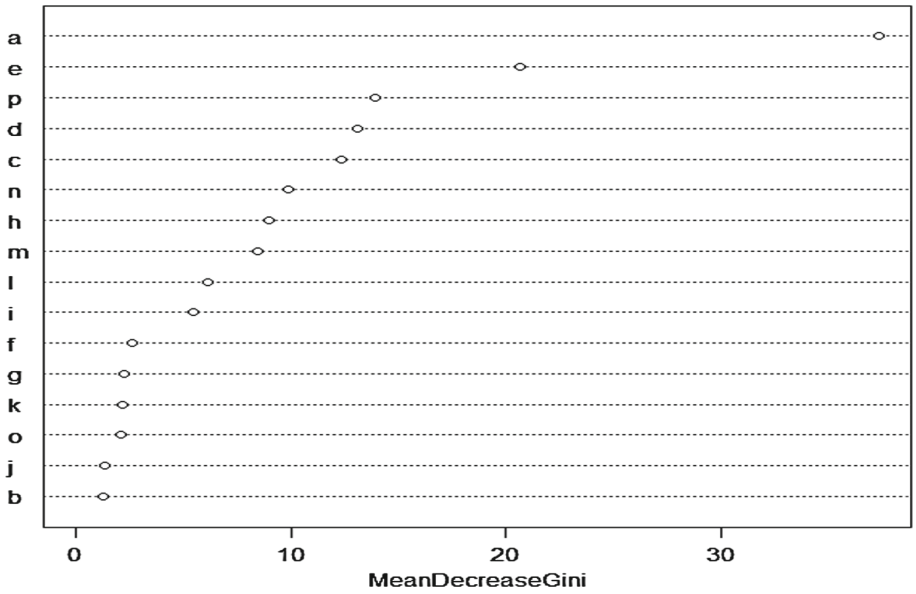

Fig. 2. Graph depiction of Mean Decrease Gini

5 OOB Estimation Error

The out-of-bag error is an error estimation technique often used to evaluate the accuracy of a random forest and to select appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split, referred to as mtry [16]. For each observation zi = (xi, yi), construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which zi did not appear [17]. The out of bag estimate chooses all the samples which were left during the tree creation and error is estimated for that sample.

6 Analysis and Discussions

Random forest generates an OOB error estimation depending upon the seed value and mtry. Table 3 shows OOB error estimation at different seed value and mtry. Random forest code run in R, for different values of mtry ranging from 3 to 13 and the value of set.seed was from 1 to 10. Set.seed is used for staring point of random number generation and mtry tuning parameters. Thus total 130, OOB estimation recorded as shown in Table 3.

Among the obtained OOB error estimation values, lowest value (30.98) is obtained at set.seed value 3 and mtry value 12, which is used for further study and generating rule. The R Code generates the 500 trees and selecting 12 variables at each spilt. Total 297 records considered for developing the model.

Table 3. OOB error estimation at set.seed and mtry

	mtry											
		3	4	5	6	7	8	9	10	11	12	13
Set.seed	1	31.31	33	32.66	34.01	34.34	34.01	34.34	32.32	33.33	34.01	33.67
	2	33.33	34.01	33.33	32.32	32.66	33.67	34.68	34.01	34.34	34.01	32.66
	3	33.67	35.02	32.32	35.02	33.67	32.32	35.35	34.01	34.68	**30.98**	34.01
	4	34.68	34.01	33	32.66	34.34	34.34	35.02	34.68	32.32	33.67	35.69
	5	31.65	32.32	35.02	33.67	34.68	34.01	31.99	33	33	31.99	33.67
	6	32.32	32.66	33.67	33.67	34.34	33	33.33	35.69	35.02	35.02	33.33
	7	33.33	34.01	31.65	32.66	35.35	33.33	35.35	36.03	35.35	35.35	34.68
	8	33	34.34	31.99	34.01	32.66	33.67	34.01	33.33	34.68	33.67	34.68
	9	31.99	33	33.33	34.01	35.35	33.33	34.34	34.68	35.69	34.34	35.02
	10	32.66	34.01	33.67	33.67	34.01	33	34.34	33	31.99	34.01	34.01

Table 4 shows confusion matrix. Confusion matrix shows that total 92 records are correctly classified into class 1 whereas 54 records are wrongly classified. Similarly, 113 records classify correctly into the class 2 whereas 38 records are wrongly classified. Classification error for the class 1 is 0.3698 and for the class 2 classification error is 0.2516556.

Table 4. Confusion matrix

N = 297	1	2	Total	Class. error
1	92	54	146	0.3698630
2	38	113	151	0.2516556
Total	130	167		

Accuracy related various parameters calculated from the confusion matrix are given in the Table 5 as diagnostic testing of accuracy. In Table 5 accuracy of classification of the tree reported 69.0326%. Another important point to identify the accuracy is precision and recall. Precision and recall both collectively represent detailed picture of accuracy. At one side precision represents relevancy whereas recall represent correctness of the model. Precision value of the decision tree is 67.6646% whereas recall is 74.8344%. Precision value explains 67.6646% of positive identification actually correct and recall explains that 74.8344% actual positives identified correctly. Misclassification rate of the decision tree is 28.9562%. Epidemiologist and other use prevalence which is in contrast incidence measure new cases in the population. Point prevalence reported in the Table 5 is 50.8417%. Prevalence explains that reported percentage of people are having condition of cataract at the time of collection of data. False positive rate i.e. 36.986% condition

Table 5. Parameters obtained from the values of confusion matrix

Parameter	Value	Parameter	Value	Parameter	Value
Accuracy of classifier	0.6902356	Precision	0.676646	Recall	0.748344
Misclassification rate	0.289562	Prevalence	0.508417	False positive rate	0.369863
F-score	0.710691	True negative rate	0.63013		

improperly exist. True negative rate is 63.013 reported in the Table 5 explains that actual nonexistence of condition is correctly classified. F score indicator represents harmonic mean between precision and recall. F score value is reported in Table 5 is 71.0691% represents similarity between the groups.

Table 6 shows the database of tree generation. Sr. no. shows the node number. Left daughter column indicates the node number which is associated with the left part of the splitting node. Right daughter indicates that which node number is associated with right part of the splitting node. Split var indicates the name of variable which is used for the splitting. Status column indicates that whether node is terminal or non-terminal node. If the status is 1 it means it is non-terminal and −1 status indicates that it is terminal node and indicates class name. Predication column shows the name of the class. <NA> in prediction column indicates that node is not leaf node and has further left or right or both sub-trees.

Table 6 enlists the rules generated by random forest. First column of the table represents node of the tree. Second column highlights left child of the current node. Third column represents right child of the current node. Fourth column represents code name of the splitting variable. Splitting code is defined in the Table 2. Fifth column is the split point that represents threshold value. Continuous values less than goes to the left side of the tree and greater than and equal into right side of the tree; in case of categorical variable respective values are mentioned in the column. Column 6 is status represents whether the current node belongs to leaf node. Status 1 represents non-leaf node whereas

Table 6. Rules generated by Random Forest

Node	Left daughter	Right daughter	Split var	Split point	Status	Prediction
1	2	3	A	54.5	1	<NA>
2	4	5	L	2	1	<NA>
3	6	7	G	1	1	<NA>
4	8	9	E	49	1	<NA>
5	10	11	C	167	1	<NA>
6	12	13	P	21	1	<NA>

(*continued*)

Table 6. (*continued*)

Node	Left daughter	Right daughter	Split var	Split point	Status	Prediction
7	14	15	H	6.5	1	<NA>
8	0	0	<NA>	0	−1	1
9	16	17	F	1	1	<NA>
10	18	19	H	0.5	1	<NA>
11	20	21	D	5.4	1	<NA>
12	22	23	C	239	1	<NA>
13	24	25	A	58.5	1	<NA>
14	26	27	A	58.5	1	<NA>
15	28	29	P	5	1	<NA>
16	30	31	P	8.5	1	<NA>
17	0	0	<NA>	0	−1	2
18	32	33	P	18	1	<NA>
19	34	35	P	2.5	1	<NA>
20	36	37	N	3	1	<NA>
21	0	0	<NA>	0	−1	1
22	38	39	B	1	1	<NA>
23	0	0	<NA>	0	−1	1
24	0	0	<NA>	0	−1	2
25	0	0	<NA>	0	−1	1
26	40	41	H	1	1	<NA>
27	0	0	<NA>	0	−1	2
28	0	0	<NA>	0	−1	1
29	0	0	<NA>	0	−1	2
30	0	0	<NA>	0	−1	1
31	0	0	<NA>	0	−1	2
32	42	43	K	1	1	<NA>
33	44	45	D	5.4	1	<NA>
34	0	0	<NA>	0	−1	2
35	46	47	E	74	1	<NA>
36	0	0	<NA>	0	−1	2
37	48	49	P	7.5	1	<NA>
38	50	51	I	1.5	1	<NA>

(*continued*)

Table 6. (*continued*)

Node	Left daughter	Right daughter	Split var	Split point	Status	Prediction
39	52	53	L	2	1	<NA>
40	54	55	N	1.5	1	<NA>
41	0	0	<NA>	0	−1	2
42	56	57	N	7	1	<NA>
43	58	59	P	9.5	1	<NA>
44	0	0	<NA>	0	−1	1
45	0	0	<NA>	0	−1	2
46	0	0	<NA>	0	−1	1
47	60	61	I	5	1	<NA>
48	0	0	<NA>	0	−1	1
49	0	0	<NA>	0	−1	2
50	62	63	D	5.25	1	<NA>
51	0	0	<NA>	0	−1	1
52	0	0	<NA>	0	−1	1
53	64	65	E	76	1	<NA>
54	0	0	<NA>	0	−1	1
55	0	0	<NA>	0	−1	2
56	0	0	<NA>	0	−1	1
57	66	67	E	76.5	1	<NA>
58	0	0	<NA>	0	−1	1
59	0	0	<NA>	0	−1	2
60	0	0	<NA>	0	−1	2
61	0	0	<NA>	0	−1	1
62	68	69	D	4.5	1	<NA>
63	70	71	A	56.5	1	<NA>
64	72	73	N	0.5	1	<NA>
65	74	75	A	67	1	<NA>
66	0	0	<NA>	0	−1	2
67	0	0	<NA>	0	−1	1
68	76	77	C	1	1	<NA>
69	78	79	K	1	1	<NA>
70	0	0	<NA>	0	−1	1

(*continued*)

Table 6. (*continued*)

Node	Left daughter	Right daughter	Split var	Split point	Status	Prediction
71	80	81	A	79	1	<NA>
72	0	0	<NA>	0	−1	1
73	0	0	<NA>	0	−1	2
74	0	0	<NA>	0	−1	2
75	0	0	<NA>	0	−1	1
76	0	0	<NA>	0	−1	1
77	0	0	<NA>	0	−1	2
78	82	83	O	1	1	<NA>
79	84	85	D	5.05	1	<NA>
80	86	87	N	0.5	1	<NA>
81	0	0	<NA>	0	−1	1
82	0	0	<NA>	0	−1	2
83	88	89	E	50	1	<NA>
84	0	0	<NA>	0	−1	2
85	90	91	P	0.5	1	<NA>
86	0	0	<NA>	0	−1	1
87	92	93	E	71	1	<NA>
88	0	0	<NA>	0	−1	2
89	0	0	<NA>	0	−1	1
90	0	0	<NA>	0	−1	2
91	94	95	H	5	1	<NA>
92	96	97	E	63.5	1	<NA>
93	0	0	<NA>	0	−1	2
94	0	0	<NA>	0	−1	1
95	0	0	<NA>	0	−1	2
96	98	99	E	58.5	1	<NA>
97	0	0	<NA>	0	−1	1
98	100	101	N	1.5	1	<NA>
99	0	0	<NA>	0	−1	2
100	102	103	E	53.5	1	<NA>
101	0	0	<NA>	0	−1	1
102	0	0	<NA>	0	−1	2
103	0	0	<NA>	0	−1	1

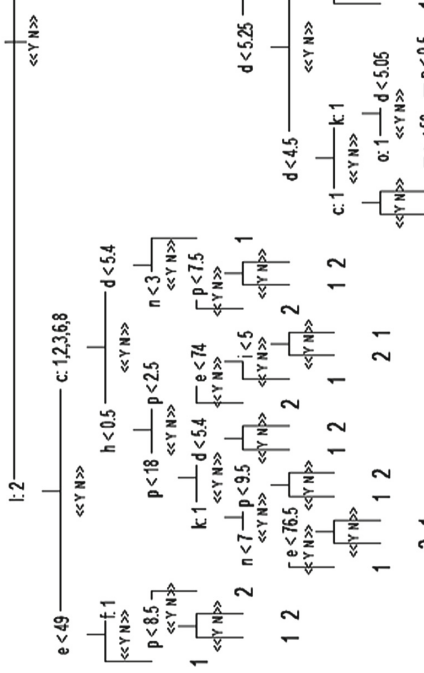

Fig. 3. Random forest tree

−1 represents leaf node. Last column is prediction that shows class label. If the node is non leaf node then this column contains value <NA> which means class identification is not required. Figure 3 is the tree representation of random forest generated rule.

7 Conclusion

Cataract condition develops in the lens of eye. Ophthalmologist consider numerous factors like living habits, age, gender as well medical conditions like diabetes, cholesterol level etc. for cause of cataract. In consultation with ophthalmologist, primary data was collected and studied using random forest algorithm. Random forest algorithm has given lowest OOB error estimation when value of set.seed was set to 3 and value of mtry set to 12. From Table 1, it is concluded that most important attribute for predicting the cataract is age and other factors in the order of importance has been shown in the Figs. 1 and 2. To predict the presence of cataract in the patient have been shown as rules in Table 6 and also visualized in Fig. 3. From confusion matrix shown in Table 3, it is concluded that for cataract yes the classification is more accurate (error is 0.2516556) and less accurate (0.3698630) for no cataract. Variable importance and tree path are useful to predict possibility of the cataract in individuals.

References

1. Anyanwu, M.N., Shiva, S.G.: Comparative analysis of serial decision tree classification algorithms. Int. J. Comput. Sci. Secur. (IJCSS) **3**(3), 230–240 (2009)
2. Sharma, H., Kumar, S.: A survey on decision tree algorithms of classification in data mining. Int. J. Sci. Res. (IJSR) **5**, 2094–2097 (2016)
3. Gupta, B., et al.: Analysis of various decision tree algorithms for classification in data mining. Int. J. Comput. Appl. **163**(8), 15–19 (2017). (0975–8887)
4. Kesavaraj, G., Sukumaran, S.: A study on classification techniques in data mining. IEEE-31661, 4–6 July 2013
5. Black, P.E.: Greedy Algorithm, in Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Technology, February 2005. NIST-greedy algorithm
6. Alaoui, S.S., Labsiv, Y., Aksasse, B.: Classification algorithms in data mining. Int. J. Tomogr. Simul **31**, 34–44 (2018)
7. Breiman, L. Random Forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
8. Goel, E., et al.: Int. J. Adv. Res. Comput. Sci. Softw. Eng. **7**(1), 251–257 (2017)
9. Brieman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
10. Uwe, K., Dunemann, S.O.: SQL database primitives for decision tree classifiers. In: CIKM 2001. ACM, Atlanta (2001)
11. Denil, M., et al.: Narrowing the gap: random forests in theory and in practice. In: Proceedings of the 31st International Conference on Machine Learning, Beijing, China, vol. 32. JMLR: W&CP (2014)
12. Kaur, S., Grewal, A.K.: A review paper on data mining classification techniques for detection of lung cancer. Int. Res. J. Eng. Technol. (IRJET) **03**(11), 1334–1338 (2016). e-ISSN: 2395-0056, p-ISSN: 2395-0072
13. Parashar, H.J., Vijendra, S., Vasudeva, N.: An efficient classification approach for data mining. Int. J. Mach. Learn. Comput. **2**(4), 466 (2012)

14. Kumar, S.V.K., Kiruthika, P.: An overview of classification algorithm in data mining. Int. J. Adv. Res. Comput. Commun. Eng. **4**(12), 255–257 (2015). ISSN (Online) 2278-1021, ISSN (Print) 2319 5940
15. https://www.who.int/blindness/causes/priority/en/index1.html. Accessed 29 Sept 2019
16. Janitza, S., Hornung, R.: On the overestimation of random forest's out-of-bag error. PLoS ONE **13**(8), e0201904 (2018). https://doi.org/10.1371/journal.pone.0201904
17. Hastie, T., et al.: The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
18. Nayer, A.M.: Detecting diabetes mellitus using machine learning. Int. J. Comput. Syst. **03**(12), 670–677 (2016). ISSN 2394-1065
19. Beaulac, C., Rosenthal, J.S.: Predicting university students academic success and major using random forest (2019). https://doi.org/10.1007/s11162-019-09546-y
20. Sugandhi, C., Yasodha, P., Kannan, M.: Analysis of a population of cataract patients databases in weka tool. Int. J. Sci. Eng. Res. **2**(10), 1 (2011). ISSN 2229-5518
21. Niya, C.P.: Automatic cataract detection and classification systems: a survey. TechS Vidya e-J. Res. **3**(2014–15), 28–36 (2015). ISSN 2322-0791