Performance Analysis of Different Machine Learning Algorithms and to Predict Trachoma Using R Language

Akbar khan, Dr.Faizullah Khan, Dr.Surat Khan, Ishtiaq Marwat

Abstract— Machine learning has remained widely and effectively utilized as a part of anticipating diverse sicknesses, including eye related. Trachoma is an eye sickness. Trachoma is the most widely recognized eye sicknesses in worldwide and brought on by visual impairment. The primary point of this paper is to foresee Trachoma among the understudies of diverse schools here in BALOCHISTAN of distinctive ages, and about 1000 patients were analyzed with the help of an eye expert and his team. And to find out the variable that can cause the trachoma most and prior and compared all the results.

Index Terms— Machine Learning, Logistic Regression, Support Vector Machine, Random forest, SMOTE, WEKA

1 Introduction

Machine learning is a sub field of Computer science that made from determined learning hypothesis in medical Trachoma is the most widely recognized eye illnesses in worldwide and caused by blindness. Infection with trachoma is most usually found in youngsters and grown-ups just about. It is transmitted through the release from tainted kids eyes and went on by hands (fingers), on garments or by flies that terrains on the eyes of no inflected children.it happens in poor individual group where the climatic condition is dry ,hot, repulsive and dusty climatic and influence the most marginalized and denied individuals from the group.

By World Health Organization (WHO) trachoma is right now in charge of more than 3% of world's visual deficiency. [1]

In Africa, 27.8 million instances of dynamic trachoma (68.5% of all) and 3.8 million instances of trachealis (46.6%) of all are observed and it is accepted to be endemic in 33 of the 56 nations in Africa. The peak occurrence of dynamic trachoma and trachealis stays in the Sahel region of West Africa and Savannah zones of East and Central Africa. A great extent of TF pervasiveness in 1–9-year-olds in South Sudan (83%), Ethiopia (64%), Guinea (half), Uganda (37%), Chad (38%), Central Africa Republic (38%), and Tanzania (32%) .Studies in Gambia, Cameroon, and Nigeria additionally demonstrated that the general commonness of dynamic trachoma in youngsters matured 1–9 years old were 3.8%, 12%, and 37.7%, separately .[2]

- *Akbar Khan is currently serving as aLecturer in the Department of Computer Engineering, BUITEMS, PAKISTAN.*
- *Akbar.khan@buitms.edu.pk*
- *Dr. Faizullah Khan is working as Chair Person Telecommunication Engineering Department, Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Quetta, Pakistan,*
- *E-mail: faizullah.khan@buitms.edu.pk*

Different danger variables are seen to be there connected with element trachoma, small monetary status, swarming, nonattendance of cleanliness and social mentalities are identified not basic factors of component trachoma .the locale of flies stays to an extraordinary degree enormous for trachoma event and a tarnished face can be brought on trachoma [3]

Data mining techniques are the strategies wanted to observe and measure data for the point if comprehension and determining essential results and illustrating frameworks considering proposition result. [4]

2 Materials and Methods

2.1 Data source and Data collection

The primary data was collected using a questioners based which includes a questions related to several personal, socio economic, psychological and behavioral factors. The key looking at unit was finished Cluster/Villages. The amount of gatherings per Evaluation Unit (EU) for each association has been determined. A multi-stage pack examining system - to make sense of which gatherings and family units will be investigated in the midst of the study. With the help of an eye expert and their team has examined each and every man and woman individually and concluded their case result on the basis of affected condition

2.2 Data structure

The general structure for all variables is a table like structure where information is put away in instances and attributes. Each instance comprises of information identified with each other, separated up in various segments where every section

can have its own information sort.

The classification of target class contains two means binary class.

- Patient has trachoma
- Patient has no trachoma means normal

This generates a problem that is known as binary class.

Table.1
List of Variables

S. No	Variables	Description	Type
1	Page	1-5, 6-10, 11-15, 16-20, 21-40, 41-60 and 60+	Nominal
2	District	KSF,Kech,Loralai,Quetta	Nominal
3	DFW	Daily face washing	Nominal
4	Sex	Male and Female	Nominal
5	UoL	Use of Latrine	Nominal
6	T	Trachoma	Nominal
7	WFW	Water for face washing	Nominal
8	SWCD	Solid waste collected disposed	Nominal
9	SWCS	Solid waste collection system	Nominal
10	TDW	Type of drinking water	Nominal
11	SDW	Source of drinking water	Nominal
12	T&D	Time and distance	Nominal
13	HC	House connected	Nominal
14	LoH	Latrine of house hold	Nominal
15	ToL	Type of latrine	Nominal

2.3 Data preprocessing steps

Data is basically evaluated with all variables involved. The first step is to analyze data is to select 100 instances randomly from the entire data set contain .once this experiment has been done and as a result there are some data set that contain no Trachoma at this point the said experiment is little bit weak and we will increase the number of instances per data set so it may make the required result clear mean the data set will contain the class Trachoma selected randomly.

Following are the concerned code.

Fig .1 Codes

```
Function rndnmb() As Integer
rndnmb = Int((1000 - 1 + 1) * Rnd + 1)
End Function
```

```
Sub Macro1()
Application.ScreenUpdating = False
Sheets("Datanew").Activate
no = InputBox("Enter Name of the Sheet")
noreps = 120
Dim nums(120)
For i = 1 To noreps

    nums(i) = rndnmb()
    For j = i - 1 To 1 Step -1
        If (nums(j) = nums(i)) Then
            nums(i) = rndnmb()
            j = i - 1
        End If
    Next j
Next i

    Sheets("Datanew").Activate
    Sheets("Datanew").Range("A1").Select
    ActiveCell.EntireRow.Copy

    Sheets(no).Activate
    Sheets(no).Range("A1").Select
    Sheets(no).Paste

For i = 1 To noreps

    Sheets("Datanew").Activate
    Sheets("Datanew").Range("A" & nums(i) + 1).Select
    ActiveCell.EntireRow.Copy

    Sheets(no).Activate


    Sheets(no).Range("A" & i + 1).Select
    Sheets(no).Paste
Next i

Sheets(no).Activate
Sheets(no).Range("L2").Activate
tracs = 0
ntracs = 0

For i = 2 To noreps + 1
    If (ActiveCell.Value = "T") Then
        tracs = tracs + 1
        ActiveCell.Offset(1, 0).Activate
    Else
        If (ActiveCell.Value = "N") Then
            ntracs = ntracs + 1
            ActiveCell.Offset(1, 0).Activate
        End If
    End If
Next i

Sheets(no).Range("Y1").Activate
ActiveCell.Value = "Trachoma"
ActiveCell.Offset(0, 1).Value = "Non-Trach"
Sheets(no).Range("Y2").Activate
ActiveCell.Value = tracs
ActiveCell.Offset(0, 1).Value = ntracs

    Range("Y1:Z2").Select
    ActiveSheet.Shapes.AddChart.Select
    ActiveChart.ChartType = xlPie
    ActiveChart.SetSourceData Source:=Range("$Y$1:$Z$2")
    ActiveChart.SeriesCollection(1).Select
    ActiveChart.SeriesCollection(1).ApplyDataLabels
    ActiveChart.Legend.Select
    ActiveChart.SeriesCollection(1).Select
```

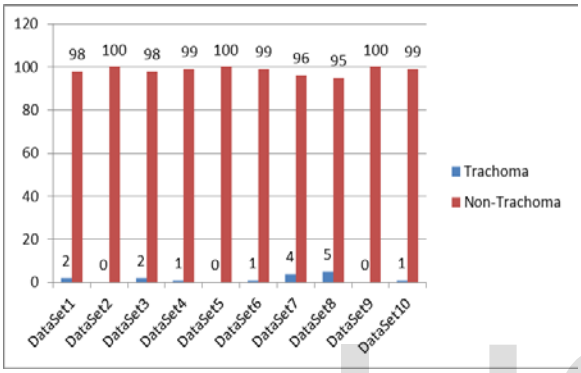
```
        ActiveChart.SeriesCollection(1).DataLabels.Select
        Selection.ShowCategoryName = True
    End Sub
```

2.4 Following are the test sets randamoly selected.

Test .1

- Data Set with No Trachoma 3
- Data Sets 10
- Sample Size / Data Set 100

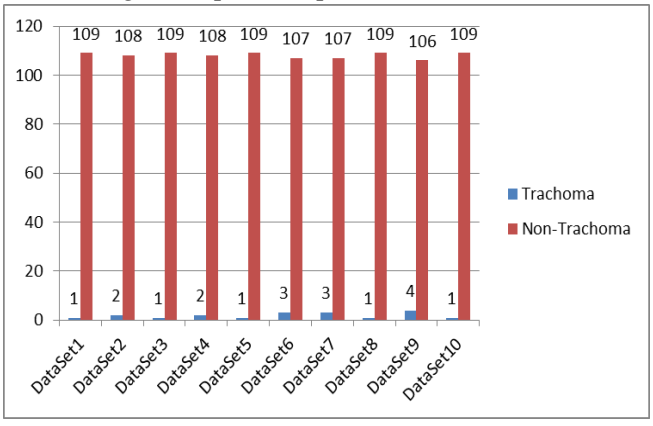
Fig .2 Graphical representations for test 1



Test .2

- Data Set with No Trachoma 0
- Data Sets 10
- Sample Size / Data Set 110

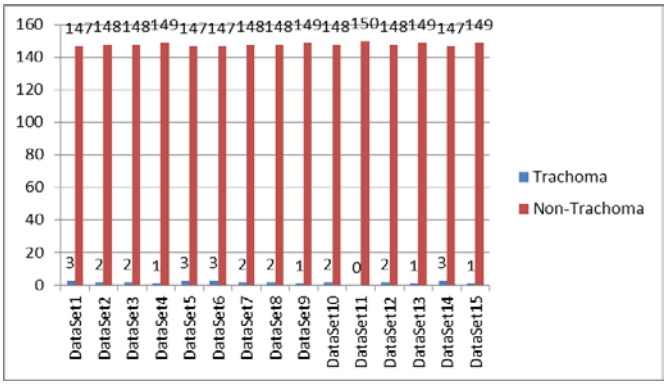
Fig .3 Graphical representations for test .2



Test .3

- Data Set with No Trachoma 1
- Data Sets 15
- Sample Size / Data Set 150
-

Fig.4 Graphical representation for test .3



3. SMOTE

A technique which is useful for the class imbalance through SMOTE the said class can be balanced while using oversampling and under sampling. Oversampling technique is used to increase the minority class and undersampling is used to decrease the majority class. [5]

4. Results

The Data set was gathered from the concerned authority in the field related. The information set is imbalanced with 12 out of 1000 implies that there are 12 positive and rest of the 988 are negative. When we have imbalanced information, there are predominantly three things we need to embrace in the exploration.

1) Ignoring the issue,

2) Under examining the larger part class,

3) Over inspecting the minority class.

A standout amongst the most widely recognized being the SMOTE system in any case; something to remember is that while over testing utilizing SMOTE improves the choice limits. On the off chance that we utilize the same information for training and validation, results will be better lastly comes about while doing proper cross validation with oversample utilizing SMOTE strategy.

Inside the cross-validation loop, get a specimen out and don't utilize it for anything identified with components determination, oversampling or model building.

Oversample your minority class, without the example you as of now excluded.

Utilize the avoided test for validation, and the oversampled minority class + the larger part class, to make the model.

Do again n times, where n is your number of tests (if doing forget one member cross-approval).

While implementing SMOTE in R language. The above given snippet shows that how the data is loaded. The data is loaded from a fixed path i.e. "E: /only4/". After setting the working directory features are loaded in to R environment in the above mentioned variable name "tpehgdb_ features"Variables are then factorized

Here the part of the program mentions the variables to be considered in variable "features" for cross validation we will leave one participant out and then for result storage from different algorithms variables used are shown above.

4.1 By selecting the following features

features <-
c("page","district","dfw","sex","uol","wfw","swcd","swcs","tdw","sdw","hc","loh","tol")

Fig .5 Graphical views for sensitivity

Fig .6 Graphical views for specificity

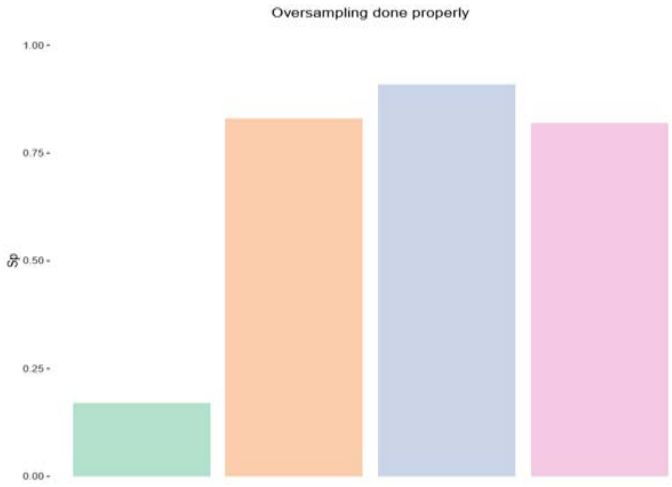


Table 2.2

Table 2.2 shows complete results of the experiments

	se	sp	ppv	npv	auc	classifier
1	0.75	0.17	0.01	0.98	0.50	logistic_regression
2	0.33	0.82	0.02	0.99	0.51	tree
3	0.08	0.91	0.01	0.99	0.50	svm
4	0.25	0.83	0.02	0.99	0.50	random_forests

5. Conclusion

The primary collected data was too much imbalance with 12 positive and 988 patients were negative .To handle and to learn from imbalance data a particular and successful technique SMOTE which is widely used in machine learning. In SMOTE I have applied oversampling technique to increase the minority class as level of requirement .the classification of target class is contained in binary pattern i.e. patients having trachoma and having not.

Data was essentially evaluated with all variables it contain the first step was to select 100 instances randomly from the original data set. I have repeated the said experiment multiple times to increase the number of instances to find out the target class in VB.

After that I did cross validation in R language then oversample the minority class using SMOTE technique and leave one participant out cross validation and applied all the four

algorithms namely Logistic regression, Random forest, Support vector machine and Decision tree to obtain the required result the said results shown that LR is the best algorithm as compared to others algorithms.

REFERENCES

[1] J. Morris, "Beyond clinical documentation: using the EMR as a quality tool," *Health Manag. Technol.*, vol. 25, no. 11, pp. 20–24, 2004.

[2] M. Alemayehu, D. N. Koye, A. Tariku, and K. Yimam, "Prevalence of active trachoma and its associated factors among rural and urban children in dera woreda, northwest ethiopia: A comparative cross-sectional study," *Biomed Res. Int.*, vol. 2015, 2015.

[3] M. A. Mowafy, N. E. Saad, H. M. El-Mofty, M. G. ElAnany, and M. S. Mohamed, "The prevalence of chlamydia trachomatis among patients with acute conjunctivitis in kasr alainy ophthalmology clinic," *Pan Afr. Med. J.*, vol. 17, 2014.

[4] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[5] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. data*, 2009.