ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

INTERNATIONAL JOURNAL OF RESEARCH – GRANTHAALAYAH
A knowledge Repository

Science

PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHM ON DIABETES HEALTHCARE DATASET

Subhankar Manna [*1], Malathi G. [2]
[*1] MCA, VIT University, Chennai Campus, India
[2] Associate Professor, School of Computing Science & Engineering, VIT University, Chennai Campus, India

Abstract

Healthcare industry collects huge amount of unclassified data every day. For an effective diagnosis and decision making, we need to discover hidden data patterns. An instance of such dataset is associated with a group of metabolic diseases that vary greatly in their range of attributes. The objective of this paper is to classify the diabetic dataset using classification techniques like Naive Bayes, ID3 and k means classification. The secondary objective is to study the performance of various classification algorithms used in this work. We propose to implement the classification algorithm using R package. This work used the dataset that is imported from the UCI Machine Learning Repository, Diabetes 130-US hospitals for years 1999-2008 Data Set.
Motivation/Background: Naïve Bayes is a probabilistic classifier based on Bayes theorem. It provides useful perception for understanding many algorithms. In this paper when Bayesian algorithm applied on diabetes dataset, it shows high accuracy. Is assumes variables are independent of each other.
In this paper, we construct a decision tree from diabetes dataset in which it selects attributes at each other node of the tree like graph and model, each branch represents an outcome of the test, and each node hold a class attribute. This technique separates observation into branches to construct tree. In this technique tree is split in a recursive way called recursive partitioning. Decision tree is widely used in various areas because it is good enough for dataset distribution. For example, by using ID3 (Decision tree) algorithm we get a result like they are belong to diabetes or not.
Method: We will use Naïve Bayes for probabilistic classification and ID3 for decision tree.
Results: The dataset is related to Diabetes dataset. There are 18 columns like – Races, Gender, Take_metformin, Take_repaglinide, Insulin, Body_mass_index, Self_reported_health etc. and 623 rows. Naive Bayes Classifier algorithm will be used for getting the probability of having diabetes or not.
Here Diabetes is the class for Diabetes data set. There are two conditions “Yes” and “No” and have some personal information about the patient like - Races, Gender, Take_metformin, Take_repaglinide, Insulin, Body_mass_index, Self_reported_health etc. We will see the probability that for “Yes” what unit of probability and for “No” what unit of probability which is given bellow. For Example: Gender – Female have 0.4964 for “No” and 0.5581 for “Yes” and for Male 0.5035 is for “No” and 0.4418 for “Yes”.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

Conclusions: In this paper two algorithms had been implemented Naive Bayes Classifier algorithm and ID3 algorithm. From Naive Bayes Classifier algorithm, the probability of having diabetes has been predicted and from ID3 algorithm a decision tree has been generated.

Keywords: Classification; Probabilistic Classification; Naïve Bayes Methodology; ID3 Methodology.

Cite This Article: Subhankar Manna, and Malathi G.. (2017). "PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHM ON DIABETES HEALTHCARE DATASET." *International Journal of Research - Granthaalayah*, 5(8), 260-266. https://doi.org/10.5281/zenodo.890581.

1. Introduction

The leading reason of death among is diabetes. The health industry is more in need of data mining today. When data mining algorithm used, at the end get meaningful information from large dataset and that can help to medical industry to take a good decision and improve health service. In datamining, a few arguments that can support the use of data mining in health industry for diabetes like classification. R is one of best tool for contains supervised learning as well classification of the dataset. By R we can do classification, clustering, association mining, selection etc. The main reason to using R is help research like implementation of classification algorithm and compare data mining technique very easily on algorithm. R also good for developing new technique. R is an open source software.

Diabetes is a disease in which the body's ability to produce or respond to the hormone insulin is impaired, resulting in abnormal metabolism of carbohydrates and elevated levels of glucose in the blood. Age, weight, medicine information history is some such factor being considered for diabetes.

Controlling blood sugar levels is the main treatment for diabetes, in order to prevent complications of the disease. Type one diabetes is managed with insulin as well as dietary changes and exercise. Type two diabetes may be managed with non-insulin medications, insulin and weight reduction.

2. Materials and Methods

Naïve Bayes is a probabilistic classifier based on Bayes theorem. It provides useful perception for understanding many algorithms. When Bayesian algorithm applied on large dataset, it shows high accuracy. Is assumes variables are independent of each other.

Bayes theorem provides a way to calculate posterior probability P(h | x) from P(h). P(x) and P(x | h).

$$P(h \mid x) = P(x \mid h)\, P(h) / P(x)$$

P(h | x) = Posterior Probability,
P(x | h) = Likelihood,

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

P(h) = Class Prior Probability,
P(x) = Predictor Prior Probability.

It constructs a decision tree from dataset in which it selects attributes at each other node of the tree like graph and model, each branch represents an outcome of the test, and each node hold a class attribute. This technique separates observation into branches to construct tree. In this technique tree is split in a recursive way called recursive partitioning. Decision tree is widely used in various areas because it is good enough for dataset distribution. For example, by using ID3 (Decision tree) algorithm we get a result like they are belong to diabetes or not.

$$\text{Info}(D) = -\sum p_i \log_2(p_i)$$

Dataset needed to classified a tuple in D.

$$\text{infoA}(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * I(Dj)$$

Dataset needed (after using A to spill D into V position) to classification.
Dataset gained by branching an attribute A.

$\text{Gain}(A) = \text{info}(D) - \text{info}_A(D)$.

3. Results and Discussions

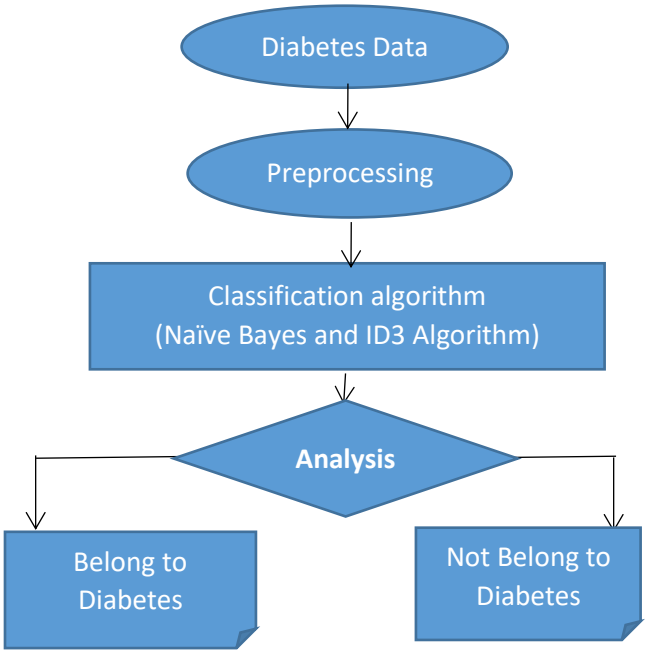

Figure 1:

Fig: 1 is the architecture for how to implement the dataset using Naive Bayes Classifier algorithm and ID3 (Decision Tree) algorithm.

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

The dataset is related to Diabetes dataset. There are 18 columns like – Races, Gender, Take_metformin, Take_repaglinide, Insulin, Body_mass_index, Self_reported_health etc. and 623 rows. Naive Bayes Classifier algorithm will be used for getting the probability of having diabetes or not.

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
       No       Yes
0.4218009 0.5781991

Conditional probabilities:
     Races
Y     AfricanAmerican     Alaska   American California  Caucasian     Chines
  No       0.10112360 0.05056180 0.12359551 0.12359551 0.09550562 0.04494382
  Yes      0.11065574 0.06967213 0.14754098 0.04918033 0.10245902 0.03688525
     Races
Y       Colorado     Indian     Navada    NewYork      Other   Pakistan
  No  0.03370787 0.15168539 0.13483146 0.03370787 0.05617978 0.00000000
  Yes 0.03278689 0.15573770 0.10245902 0.07786885 0.06147541 0.01639344
     Races
Y     SouthCarolina
  No     0.05056180
  Yes    0.03688525

     Gender
Y         Female      Male
  No  0.5112360 0.4887640
  Yes 0.5614754 0.4385246

     Take_metformin
Y           No      Yes
  No  0.258427 0.741573
  Yes 0.295082 0.704918

     Take_repaglinide
Y            No       Yes
  No  0.2134831 0.7865169
  Yes 0.2172131 0.7827869

     Insulin
Y           Down         No     Steady         Up
  No  0.17977528 0.32584270 0.40449438 0.08988764
  Yes 0.15163934 0.42622951 0.34426230 0.07786885

     Body_mass_index
Y         Normal     Obese OverWeight
  No  0.4550562 0.3426966  0.2022472
  Yes 0.4672131 0.2295082  0.3032787

     Self_reported_health
Y          Good      Poor
  No  0.6123596 0.3876404
  Yes 0.6680328 0.3319672
```

Here Diabetes is the class for Diabetes data set. There are two conditions “Yes” and “No” and have some personal information about the patient like - Races, Gender, Take_metformin,

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

Take_repaglinide, Insulin, Body_mass_index, Self_reported_health etc. If we see the above result we can see the probability that for "Yes" what unit of probability and for "No" what unit of probability. For Example: Gender – Female have 0.4964 for "No" and 0.5581 for "Yes" and for Male 0.5035 is for "No" and 0.4418 for "Yes".

ID3 (Decision Tree) algorithm perform on the diabetes dataset for making the decision tree.

```
library(rpart.plot)
> rpart.plot(dtm)
> rpart.plot(dtm,type = 4, extra = 101)
> plot(dtm)
> text(dtm)
```

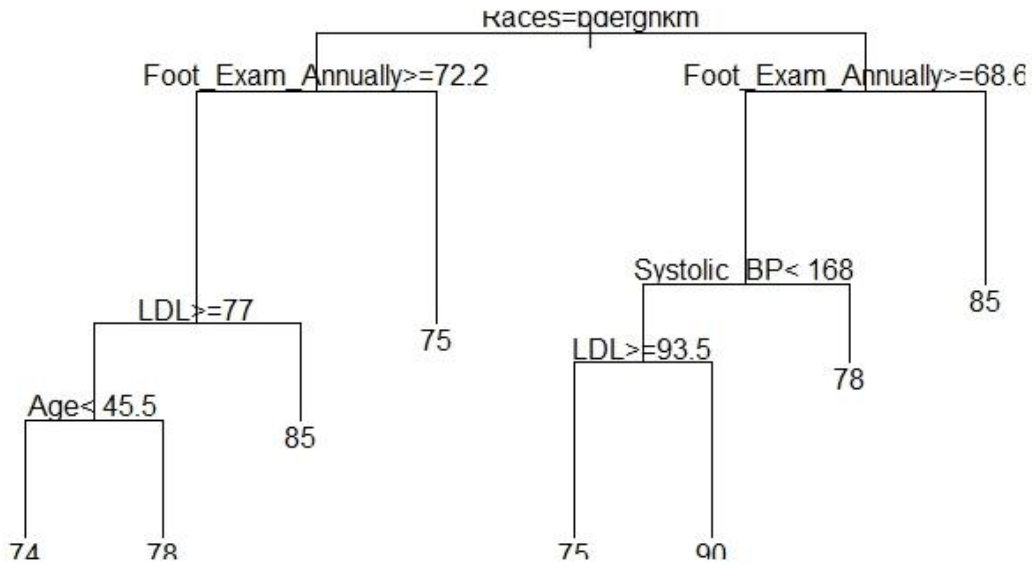

Figure 2:

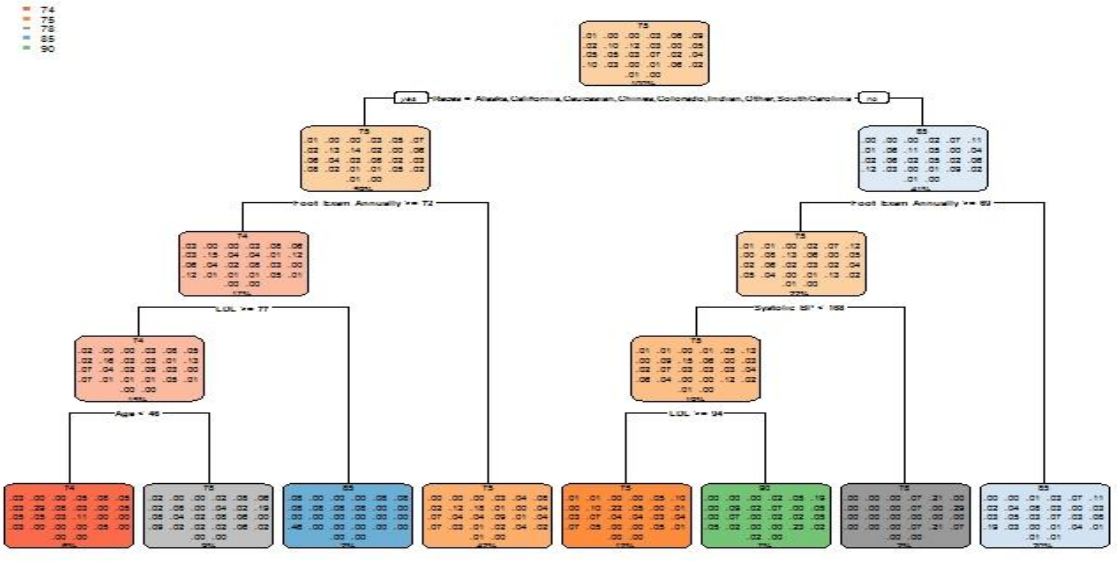

Figure 3:

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

For Fig: 2 and Fig: 3 we get a decision to clear consolation. All the numeric data comes into a table with its attribute.

4. Conclusions & Recommendations

In this paper two algorithms had been implemented Naive Bayes Classifier algorithm and ID3 algorithm. From Naive Bayes Classifier algorithm, the probability of having diabetes has been predicted and from ID3 algorithm a decision tree has been generated.

5. Appendices

- Preprocess the diabetes dataset,
- Implement the Bayesian algorithm with those datasets using R,
- Implement ID3 algorithm with those datasets using r,
- Get a probability that having diabetes or not to taking a class of those diabetes dataset,
- Make a decision tree with those datasets using R.

Health care industries are providing several benefits like fraud detection in health insurance, availability of medical facilities to patients at inexpensive process, improve patient care and hospital infection control.

Data mining is the process of extraction hidden information from massive dataset using classification technique. The technique used for classification: Naïve Bayes, ID3, K-means.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Races	Gender	Take_met	Take_repe	Insulin	Body_mas	Self_repo	Age	LDL	Systolic_B	Diastolic_	Weight	A1c_Chec	Foot_Exar	Dilated_E	Ever_Atte	Daily_Self	Diabetic
2	AfricanAn	Female	Yes	No	No	Normal	Poor	52	98	120	80	96	72.5	58	59.3	58.6	65.2	Yes
3	Caucasian	Female	No	Yes	Up	Normal	Good	42	88	130	80	82	70.5	70.2	63.2	65.2	74.2	Yes
4	Indian	Female	Yes	Yes	No	OverWeig	Poor	63	87	128	77	70	80.6	79.3	60.5	60.2	65.2	Yes
5	AfricanAn	Male	No	No	Up	Normal	Poor	58	99	115	80	95	56.9	70.6	71.2	63.1	65.1	No
6	AfricanAn	Male	Yes	No	Steady	Normal	Poor	45	67	130	80	91	72.1	58.6	72.2	57.5	74	No
7	Indian	Male	Yes	Yes	Steady	Normal	Good	32	100	120	75	78	58	69.5	56.3	72.6	75.3	Yes
8	Other	Male	Yes	Yes	Steady	Normal	Good	36	76	120	12	72	70.2	80	54.2	70.2	65.5	No
9	Pakistan	Male	Yes	Yes	No	OverWeig	Good	56	111	140	85	80	79.3	63.2	56.3	63.5	60.2	Yes
10	Chines	Female	Yes	Yes	Steady	OverWeig	Good	85	102	130	80	79	70.6	85	65.2	69.8	69.3	No
11	American	Female	Yes	Yes	Steady	Normal	Good	63	45	145	72	82	58.6	69.6	65.2	63.2	65.2	Yes
12	Alaska	Female	Yes	Yes	Steady	OverWeig	Good	36	112	180	70	72	69.5	80.2	70.1	65.2	69.5	Yes
13	California	Male	No	Yes	Steady	Obese	Good	34	125	145	75	75	80	75.5	63	60.3	74.5	Yes
14	Indian	Female	Yes	Yes	Down	Obese	Good	31	145	120	80	71	63.2	74.3	65	60.5	70.2	No
15	Indian	Male	No	No	Steady	Normal	Poor	56	150	110	95	102	85	69.3	65.4	71.2	75	No
16	Indian	Female	No	Yes	Steady	Normal	Good	52	170	145	75	96	69.6	52.6	78.3	72.2	78.5	Yes
17	California	Male	Yes	Yes	Up	Normal	Good	32	180	140	78	82	80.2	74.3	69.3	56.3	79	No
18	California	Male	Yes	Yes	Steady	Normal	Good	32	185	145	79	81	75.5	56.9	56.9	54.2	75.2	Yes
19	Alaska	Female	Yes	No	No	Normal	Good	32	190	125	82	79	74.3	72.1	72.1	56.3	74.2	Yes
20	AfricanAn	Male	Yes	Yes	Steady	Normal	Poor	33	85	163	83	98	69.3	58	58	65.2	75.6	Yes
21	Colorado	Male	No	Yes	Steady	Obese	Poor	34	102	130	80	85	52.6	70.2	70.2	65.2	78.6	Yes
22	Navada	Female	No	Yes	Down	Normal	Good	34	145	160	75	81	74.3	79.3	79.3	70.1	63.2	Yes
23	Colorado	Male	Yes	Yes	Steady	Normal	Good	34	164	173	12	86	56.9	70.6	70.6	57.5	65.2	No

References

[1] MR. CHINTAN SHAH from Information Technology, SHANKERSINH VAGHELA BAPU from Institute of Technology Gandhinagar, India. (2013). “COMPARISON OF DATA MINING CLASSIFICATION ALGORITHM FOR BREAST CANCER PREDICTION.”

[2] ZEINAB SEDIGHI, HOSSEIN EBRAHIMPOUR-KOMLEH, SEYED JALALEDDIN MOUSAVIRAD, Department of Computer Engineering, Faculty of Computer and Electrical

ISSN- 2350-0530(O), ISSN- 2394-3629(P)
DOI: 10.5281/zenodo.890581

Engineering, University of Kashan, Kashan, I.R.Iran, November, (2015), "FEATUE SELECTION EFFECTS ON KIDNEY DESEASE ANALYSIS."

[3] VEENIT KUNWAR, KHUSBOO CHANDEL, A. SAI SABITHA, Amity University Uttar Pradesh, July, (2013). "CHRONIC KIDNEY DISEASE ANALYSIS USING DATA MINING CLASSIFICATION TECHNIQUES."

[4] HAMIDAH JANTAN, ABDUL RAZAK HAMDAN AND ZULAIHA ALI OTHMAN, University Teknologi MARA (UiTM) Terengganu, 23000 Dungun, Terengganu, "MALAYSIA.POTENTIAL DATA MINING CLASSIFICATION TECHNIQUES FOR ACADEMIC TALENT FORECASTING."

[5] D. RAJESWARA RAO, VIDYALLATA PELLAKUN, SATHISH TALLAM, RAMYA HARIKA, K L University, Guntur, Andhra Pradesh. (2015) "PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS USING HEALTHCARE DATASET."

[6] KETAN SANJAY DESALE, CHANDRAKANT KUMATHEKAR, ARJUN PRAMOD CHAVAN from DYPSOEA, Pune, Maharahtra. (2015), "EFFICIENT INTRUSION DETECTION SYSTEM USING STREAM DATA MINING CLASSIFICATION TECHNIQUE."

[7] GRIGORIOS CHYSOS, PANAGIOTIS DAGRIZIKOS, IOANNIS PAPAEFSTATHIOU, APOSTOLOS DOLLAS, Microprocessor & Hardware Laboratory, Dept of Electronic and Computer Engineering, Chaina, Greece. (2012), "NOVEL AND HIGHLY EFFICIENT RECONFIGURABLE IMPLEMENTATION OF DATA MINING CLASSIFICATION TREE."

[8] C. M. VELU, K. R. KASHWAN, Dept of Computer Science and Engineering Dattakala Group of Institution, Swami Chincholi, Pune. (2013), "VISUAL DATA MINING TECHNIQUES FOR CLASSIFICATION OF DIABETIC PATIENTS."

[9] A. SWARUPA RANI, S. JYOTHI, Dept of Computer Science Sri Padmavathi Visvavidyalayam, Tirupati, AP. (2016), "PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHM UNDER DIABETIC DATASET."

*Corresponding author.
E-mail address: subhankar.manna2016@ vitstudent.ac.in