Understanding the Lifestyle of People to Identify the Reasons of Diabetes Using Data Mining

Gavin Pinto, Sunil Jangid, Radhika Desai

Post- Graduate Student, Post- Graduate Student, Project Guide

Information Technology,

Thakur College of Science & Commerce, Mumbai-400101, India

Abstract: Data mining techniques explore critical information in various domains (for example in CRM (customer relationship management), HR (Human Resource), GIS (Geographic Information System) etc.) but most importantly in medical domain. In medical domain, data mining can assist on minimizing the risk of developing some of the stereotyped diseases such as cancer, heart diseases, diabetes etc. Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated or unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves all this critical problem. The motive of this study is to design a website which can give a probability of having diabetes depending upon the information user is entering and giving a prediction of if the person is prone to having diabetes in the near future. Therefore, two machine learning classification algorithms namely SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on a dataset collected by a google form survey. The performances of all the algorithms are evaluated on various measures like Precision and Accuracy. Accuracy is measured over correctly and incorrectly on classified instances. Results obtained show Naive Bayes outperforms with the highest accuracy of 60.30% comparatively other algorithms.

Keywords- Data Mining, Blood Pressure, Naïve Bayes, SVM

I. INTRODUCTION

Data mining is the process of extracting hidden knowledge from a large volumes of data. It is the analytical process designed to explore data in search of consistent patterns and find systematic relationships between variables. The application areas of data mining are in field of education system, market basket analysis, customer relationship management, banking application, sports and in Health care system. In recent years medical data mining has become prominent, since there is enormous amount of medical data available which can be used for discovering useful patterns. The data mining techniques such as classification, clustering, association, outlier analysis help in finding useful patters from the huge amount of medical data. Data mining has great potential for the healthcare industry since it helps health systems to use medical data for analysis and to offer improved healthcare at reduced cost. The data mining techniques when applied to health care play a significant role in prediction and diagnosis of various health problems like heart disease, diabetes, cancer, skin disease and many more.

Classification

Data mining includes classification as one of the most fundamental task. Classification is used to predict the group membership of data instances. Classification is applied in manyareas such as weather prediction, medical diagnosing, scientific experiments etc. The classification technique is mainly used in medical data mining. The classification techniques generally used are Decision trees, Bayesian classifier, Random Forest, Random tree, classification by back propagation and rule based classifiers. Classification is performed in two steps: Model construction: In this step the prediction model is built using appropriate algorithm. In this step the prediction model is applied to actual data and prediction is done accordingly.

II. LITERATURE REVIEW

A Research Paper given by Mukesh kumari1, Dr. RajanVohra 2,Anshul arora3 1,3Student of M.Tech (C.E) 2Head of Department Department of computer science & engineering P.D.M College of Engineering helps in predicting diabetes by applying data mining technique. The discovery of knowledge from medical datasets is an important aspect in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. Diabetes mellitus is a chronic disease and it is a major public health challenge worldwide. Using data mining methods in order to aid people to predict diabetes has gain major popularity. In this paper, Bayesian Network classifier was used to predict the persons whether they are diabetic or not. We used python for the experiment and analysis. Classification algorithm is applied on the dataset of persons collected from a Google form survey. Results have been obtained[1]

In a Research paper presented by Yang Guo , GuohuaBai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, The discovery of knowledge from medical databases is important in order to make a effective medical diagnosis. The dataset that was used was the Pima Indian diabetes dataset. Preprocessing was used to improve the quality of data. The classifier applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3%. [2]

A Research Paper given by SudajaiLowanichchai, SaisuneeJabjone, TidanutPuthasimma Assistant Professor, Informatic Program Faculty of Science and Technology NakhonRatchsimaRajabhat University it proposed the application Information

technology of knowledge-based DSS for the analysis of diabetes of elder people using decision tree. The result showed that the Random Tree model has the highest accuracy in the classification is 99.60 percent when compared with the medical diagnosis that the error MAE is 0.004 and RMSE is 0.0447. The NBTree model has lowest accuracy in the classification is 70.60 percent when compared with the medical diagnosis that the error MAE is 0.3327 and RMSE is 0.454 [3]

.

III .METHODOLOGY

CONCEPTUAL FRAMEWORK

Data mining is the analysis of the hidden patterns of data according to the different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, mining algorithms, facilitating business decision making other information requirements to ultimately cut costs and increase revenue. Data mining is also referred as data discovery and knowledge discovery.

The first step in data mining is gathering the relevant data critical for business. Company data is either transactional, non-operational or metadata. Transactional data deals with the day-to-day operations like sales, inventory and cost etc. Non-operational data is basically forecast, while metadata is concerned with logical database design. Patterns and relationships among data elements render the relevant information, which may increase the organizational revenue.

The second step in data mining is the selection of a suitable algorithm - a mechanism producing a data mining model. The general working of the algorithm mainly involves identifying trends in a set of data and using the output for parameter definition. The most popular algorithms used for data mining are classification algorithms and regression algorithms, which are used to identify relationships among data elements. Major database vendors like Oracle and SQL incorporate data mining algorithms, such as clustering and regression tress, to meet the demand for data mining.[4]

Naive Bayes classifiers

Naive Bayes has been studied extensively since the year 1960s. It was introduced (though not under that name) into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging the documents as the belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing of data, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, with a requirement of a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood

Training can be done by the evaluation of a closed-form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.[5]

SVM (Support Vector Machine)

Support vector machine is another algorithm that every machine learning expert should have in his/her arsenal. Support vector machine is a highly preferred algorithm by many as it produces significant accuracy with less computation power. Support Vector Machine(SVM) can be used for both regression and classification tasks. But, it is widely used in classification objectives.

Python

Python is a high-level programming language that is designed to be easy to read and simple to implement. It is aopen source, which means it is free to use, even for commercial applications. Python can mainly run on Mac, Windows, and Unix systems and has also been ported to Java and .NET virtual machines.

Python is considered as a scripting language, like Ruby or Perl and is often used for the creation of Web applications and dynamic Web content. It is also supported by a number of 2D and 3D imaging programs, enabling users to create the custom plug-ins and extensions with Python.

IV. PROBLEM STATEMENT

We will collect the dataset through a google form survey.
The dataset will be gathered from various sources and the entry will be maintained into the excel sheet.
The excel file will contain peoples name, weight, age, gender, are they exercising, their eating habits etc.
Based on people's lifestyle, status of their health will be measured.
All the entries from dataset are maintained in the excel sheet.
Based on the patients information on the dataset, a graph will be generated by the system, which will be categorized as age, gender, lifestyle, eating habits
The dataset will be gathered from various sources and the entry will be maintained into the excel sheet.
The excel file will contain peoples name, weight, age, gender, are they exercising, their eating habits etc.
Based on people's lifestyle, status of their health will be measured.
All the entries from dataset are maintained in the excel sheet.
Based on the patients information on the dataset, a graph will be generated by the system, which will be categorized as age, gender, lifestyle, eating habits

V. RESULTS & DISCUSSION

. The dataset that is taken for this research work contains 535 records and 10 attributes for the purpose of predicting whether a person is diabetic or non diabetic based on the symptoms. This dataset is designed in MS excel format.

Preparation of Dataset :- This is the sample of dataset used for prediction. The dataset used contains 535instances . all instances have 10 input attributes(2 to 10) and one output attribute(1).table shows the attribute of this dataset.

S.No	Name	Description
1	Gender	Male or Female
2	Age	Age(in years)
3	Exercise	Times you exercise
4	Meals	Meals per day
5	Sleep	Hours of Sleep
6	Weight	Weight(in kg)
7	Blood Pressure	Normal, Low, High
8	Smoking	Yes or No
9	Alcohol	Preserve alcohol
10	Fast-food	Junk food in a week

Attributes of dataset

Sample Dataset[]

	A	B	C	D	E	F	G	H	I	J	K
1	SRNO	Gender	Age	How many ti	Meals per	Hours of s	Weight	Blood Pre	Smoking	Do you pr	Fast-food/Ju
2	1	1	22	0	3	8	57	2	0	0	3
3	2	0	21	3	3	8	48	2	0	1	3
4	3	1	19	5	3	7	54	2	0	1	4
5	4	1	22	3	3	6	95	2	0	0	3
6	5	1	19	3	2	7	51	2	0	1	2
7	6	1	25	5	4	7	67	2	1	1	2
8	7	0	22	0	2	7	45	2	0	0	3
9	8	1	16	5	3	7	66	2	0	0	3
10	9	1	25	0	3	8	60	2	0	0	2
11	10	0	22	0	2	6	56	1	0	0	4
12	11	1	28	0	3	7	55	2	0	0	2
13	12	1	19	0	4	8	75	2	0	0	2
14	13	0	24	0	4	7	53	2	0	0	2
15	14	1	23	3	3	6	70	2	0	0	2
16	15	0	22	4	2	6	75	2	0	0	3
17	16	0	20	3	3	8	49	2	0	0	2
18	17	0	21	0	2	6	62	2	0	1	4
19	18	1	24	4	3	7	50	2	0	1	2
20	19	1	22	4	4	7	72	2	0	1	4
21	20	0	19	3	3	8	56	2	0	0	4
22	21	1	21	0	3	8	85	2	0	0	4
23	22	1	18	5	2	6	57	2	0	0	4
24	23	0	21	0	3	7	60	2	0	0	3
25	24	0	21	0	3	6	61	2	0	0	4
26	25	0	21	0	2	7	37	2	0	0	4

FIGURE 1: Sample Dataset

Name	Type	Size	
X	int64	(534, 8)	[[0 3 8 ... 0 0 3] [3 3 8 ... 0 1 3]]
X_test	int64	(134, 8)	[[5 4 7 ... 0 1 2] [3 2 8 ... 0 1 3]]
X_train	int64	(400, 8)	[[3 3 7 ... 0 0 3] [0 4 8 ... 0 0 4]]
Y	int64	(534,)	[1 0 1 ... 1 1 1]
Y_pred	int64	(134,)	[1 1 0 ... 1 1 1]
Y_pred1	int64	(134,)	[1 0 0 ... 1 1 1]
Y_test	int64	(134,)	[1 1 0 ... 1 1 0]
Y_train	int64	(400,)	[0 0 1 ... 0 1 1]
dataset	DataFrame	(534, 11)	Column names: SRNO, Gender,
features	int	1	10
length	int	1	534

FIGURE 2: Variable Explorer

Fig 2 contains all the variables and the prediction values that are obtained from the desired dataset.

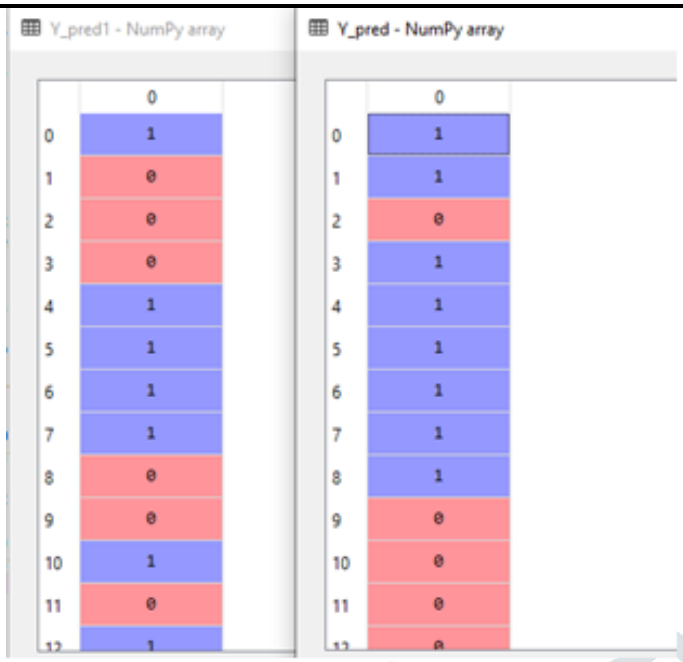

FIGURE 3: Result

Here Y_pred is the result obtained for the Naïve Bayes algorithm and Y_pred1 is the result that is obtained for SVM(Support Vector Machine).

Algorithms	Accuracy
Naïve Bayes	60.44
SVM	64.92

Table : Accuracy

VI. CONCLUSION

The disease identification data using data mining techniques are already available but always there is a need for improved accuracy for identification, prediction and diagnosis. It is also very useful in disease sub type prediction and categorization. To evaluate medical data, Various machine learning methods and mining algorithms are available. The main objective of this paper is to identify the best classifier for prediction of diabetes with a algorithm of high accuracy rate. For that the classifiers which are highly used in medical diagnosis are compared in this work. The selected classifiers and performance measures are implemented with data set. Based on the evaluation and result Naïve Bayes classifier achieved higher accuracy rate than SVM for Prediction of Diabetes dataset. The best performance of the classifier is obtained in terms of precision and error rate.

VII. FUTURE SCOPE

The survey helps us to identify the techniques that data mining uses to predict diabetes from the dataset applied thus application can be created to check whether the person can have a probability of getting diabetes based on the survey that is obtained from our research work. This will help a lot of patients as well as other users who are willing to check their health and if they are prone to having diabetes in the near future based on the data entered by the user on the application which will help users to take the necessary precautions quickly.

VIII. ACKNOWLEDGEMENT

Authors are very thankful to their guides Ms. Radhika Desai and Mr. Omkar Singh of IT Department for the guidance and help they provide for the conception of the idea of the research work and support during the research.

IX . REFERENCES

[1] Mukesh kumari1, Dr. RajanVohra 2,Anshul arora3 “Prediction of Diabetes Using Bayesian Network” (IJCSIT) International Journal of Computer Science and Information Technologies 2014

[2] SudajaiLowanichchai, SaisuneeJabjone, TidanutPuthasimma, ”Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree”

[3] SudajaiLowanichchai, SaisuneeJabjone, TidanutPuthasimma, ”Knowledge-based DSS for an Analysis Diabetes of Elder using Decision Tree”

[4] https://www.techopedia.com/definition/1181/data-mining

[5] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] Abdullah, A. A., Zakaria, Z., & Mohamad, N. F. (2011). Design and development of fuzzy expert system for diagnosis of hypertension.Proceedings of IEEE International Conference on Intelligent Systems, Modelling and Simulation, 2011, 113-17.

[7] Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.