VOL 3 (2019) NO 2-2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Predicting Diabetes by adopting Classification Approach in Data Mining

Rapinder Kaur [#]

[#] *Department of Computer Science Engineering, Chandigarh University, Punjab, India.*
E-mail: rapindersaini94@gmail.com

***Abstract*— As the world is growing fast, the metamorphosing of things, lifestyle, perceptions of people and resources is taking place. But the elevation in technology has become a challenge now as the ideas, innovations are amplifying. One of the biggest things the advancement and elevations in technology has given birth is "Big Data". In this data massive amount of information is hidden. In order to refine or process this data and to find out and unmask the insights, many techniques and algorithms have been evolved, one of which is the data mining. The data mining is the approach or procedure which helps in detaching or extracting profitable and fruitful knowledge, reports and facts from the rough or impure data. The prediction analysis is approach comprehended from data mining to forecast and figure out the future making using classification technique. This research work is based on the diabetes prediction by making use of classification approach. In the existing approach SVM classifier is applied for the prediction analysis. To increase accuracy approach of KNN classifier is applied for the prediction analysis. Both the proposed and existing methods are implemented in Python. The simulation results show that accuracy of KNN is increased and execution time is reduced.**

***Keywords*— Diabetes, SVM, KNN.**

I. INTRODUCTION

Advancement in the world has resulted in large number of specializations and domains which has given light to large data and this data is being dominating every day in every single second. Data mining can be used to discover and acquire the advantageous and fruitful information from the abundant and immeasurable datasets. Data mining is being exercised in miscellaneous domains like counterfeit or bluffer revelation, communal analysis and risk supervision, market supervision and analysis, traffic management, sports, transportation, education, healthcare, insurance, banking, medications, exploration of science and many more.

Data mining is the mechanism or procedure of separation or derivation of hidden insights from immeasurable volumes of raw datasets [1]. Data mining can be used for exploring and supervising or analysing large volumes of

Data mining is defined as the process in which useful information is derived or extracted from the raw data. In order to amass and gather the imperative, fundamental and necessitous knowledge it is a necessity to extract large amount of data. This process of extraction is also known as misnomer. Currently in every domain, there is large amount of data present and analysing or refining the whole amount of data is a way difficult and furthermore it wears out a lot of time. This present data is in rough or raw form which is of no use until or unless a proper data mining technique to find out insights is applied in order to extract fruitful knowledge. The process of extracting raw material is characterized as mining. This is a world where having access to or acquiring a lot of information procure to power and success and this is only attainable because of sophisticated technologies such as satellites, computers. With the advent in the technology in the mass digital storage and computers it becomes easier to grasp large amount of information by which different types of data is stored [3]. In the cluster analysis, image processing, market research, data analysis and pattern recognition are some major application of this technique. In the clustering technique, customer categorized group and purchasing patterns has been done in order to discover their customer's interest by the marketers. It is also utilized in biology as it derives the plant and animal taxonomies and also categorizes genes with similar functionality. In geology this technique is used to identify the similar houses and lands areas. Information clustering can be used to discover new theories that classify all documents available on Web. The group of metabolic diseases in which a person has high blood sugar is commonly referred as diabetes and in the scientific term as Diabetes mellitus. There are two reasons for the presence of high blood sugar in the body:

(1) Enough insulin is not produced by the pancreas,

(2) No response by cells to the produced insulin [4]. Hence, it is the infirmity that occurs in the human body due to omission

or absence of appropriate insulin. There are various types of diabetes exists such as diabetes insipidus. For the knowledge acquisition, Medical data mining has been utilized as it condemns all the information from research reports, medical reports, flowcharts, evidence tables. All this information is fruitful for judgements and decision making whether patient is affected by or suffering from diabetes or not. In India, Diabetes is one of the foremost and a major health dispute. There are various consequences and repercussion of this disease on human body such as risk of kidney deterioration or rupture and eye issues. In order to get of all these entanglement, early detection of the disease and proper care management is required [5]. The main aspiration of inventing a diabetes data system helps the diabetic patients during the disease. It is necessary for the diabetic patient to have daily glucoses rate and insulin dosages that can be possible by diabetes data system as it care the daily dosages a person inhale. This system is not only for the diabetic patient but also for those who suspect if they are diabetic. Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced. Type II - Diabetes is a chronic disease that is also known as Non - Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. The adequate insulin is produced by the patient, which cannot be utilized by body due to lack of sensitivity to insulin by the cells of the body. At the age of 40, type II disorder occurs mostly in human beings [6]. Devastating results are provided by the diabetic foot among the chronic diabetic as it produce various complications. Loss of sensation is also experienced by the Diabetes patients in their feet even a small injury can cause infection that is very difficult to cure. Foot ulcers problem also occur in the 15% of patients suffering with diabetes due to nerve damage and reduced blood flow. Diabetes in the person minimizes the vision to see and also cause common blindness and cataracts the diabetic person. Every year more than 50,000 leg amputations take place in India due to diabetes [7]. In the medical imaging intelligence the classification of imbalanced data is now common working. In order to tackle the computational problems the synthetic minority oversampling technique (SMOTE) has been utilized that is a powerful approach as it oversamples the positive class or the minority class. The local space between any two positive instances cab be positive or belong to lower class hence, it is fully based on assumption. This assumption is not true all the time but can be true if training data is not linearly separable [8]. Therefore, in order to use the SMOTE algorithm it is necessary to map the training data into a more linearly separable space.

II. Literature Review

Bayu Adhi Tama, et.al (2016) presented in this paper a chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes [9]. This range and data will increase in nearby future as there is no appropriate method till date that minimize the effects and prevent it completely. Type 2 diabetes (TTD) is the most common type of diabetes. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and help in knowledge discovery from data. In the data mining process, support vector machine (SVM) was utilized that acquire all the information extract all the data of patients from previous records. The early detection of TTD provides the support to take effective decision.

Yu-Xuan Wang, et.al, (2017) analyzed various applications that provide significance of the data mining and machine learning in different fields. Different data mining and machine learning techniques has been utilized to analyze the huge amount of data. It creates more commercial values in high end enterprise systems. It becomes easy with the advancement in technology to use data mining and machine learning on personal computers or embedded systems that are type low end systems [10]. For the validation of the proposed method cache design was utilized that automatically control the replacement of cached contents to make decisions. All the collected data from the system was analyzed when reply is obtained from a data miner. As per performed experiments, it is concluded that proposed method provides effective results.

Zhiqiang Ge, et.al, (2017) presented a review on existing data mining and analytics applications by the author which is used in industry for various applications. For the data mining and analytics eight unsupervised and ten supervised learning algorithms were considered for the investigation purpose [11]. To the semi-supervised learning algorithms an application status was given in this paper. In the recent years, the semi-supervised machine learning has been introduced. Therefore, it is demonstrated that an essential role is played by the data mining and analytics in the process of industry as it leads to develop new machine learning technique.

Jahin Majumdar, et.al, (2016) presented the most popular research areas in computer science that is data mining and machine learning is utilized in order to provide essential data or information. Huge amount of data is present in today's world hence it is difficult to collect only useful data as the size of data is getting increased. Therefore, it is necessary to invent a method that extracts useful information from data that will be helpful in industry and markets [12]. In order to improve the data classification and pattern recognition in Data Mining mainly feature selection various existing approaches were focused and experimented. As per performed experiments, it is concluded that comparison between the existing techniques was done in order to find out the best method. The theoretical limitations of existing algorithms were overcome by proposed method.

M. Sharma, et.al, (2017) presented data mining techniques that are utilized to investigate the known and unknown available patterns in medical databases, effect of preprocessing and performance of different data mining techniques for the used dataset [13]. The main objective of author was to provide various models in the medical science that utilized the data mining techniques. In order to mine data various efforts was made in the field of Cardiology and Diabetes. On the basis of survey it is concluded that number of papers published in cardio, diabetes, digestive, dentistry and ophthalmology disease diagnosis using data mining are 42%, 26%, 18%, 10% and 4% respectively. For the ophthalmology, dentistry and digestive disorders type's diseases author provided various models.

Han Wu, et.al (2018) proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus

(T2DM). The main objective of this paper is to improve the accuracy of the prediction model and to more than one dataset model is made adaptive in nature. Proposed model comprised of two parts based on a series of pre-processing procedures [14]. These two parts are improved K-means algorithm and the logistic regression algorithm. As per performed experiments, it is concluded that proposed model show netter accuracy as compared to other methods and also provide the sufficient dataset quality. In order to evaluate the performance of the model it is applied to other diabetes dataset, in which good performance is shown by both the methods.

III. Research Methodology

The classification techniques can be applied for the prediction analysis. These research works is based on the diabetes prediction and below steps are applied to do so:

1. Pre-Processing:- In this phase, the data is given as input and furthermore data is cleaned and the missing values, redundant values are evacuated. The data set is described in terms of standard deviation, mean etc. values are calculated.

2. Prediction Phase: - In the phase, the input dataset is divided into two parts i.e. training and test part. The 60 percent of the whole data is considered at the training part and rest 40 percent is the test part. The KNN classifier is applied for the prediction analysis which takes input test and training data and output in the form of predicted data.

One of the simplest algorithms amongst all the learning machine algorithms is the K-Nearest Neighbor (KNN) algorithm. Since there are no assumptions made on the underlying data distribution, KNN is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors, the unlabelled question point is doled out during the classification process.

Through majority share cote, on the basis of labels of its k nearest neighbours, the object is characterized. The object is classified essentially as the class of the object that is nearest to it in the event when k=1. k is known to be an odd integer in case when there are only two classes present. During the performance of multiclass categorization, there can be tie in case when k is an odd whole number. The classification of samples on the basis of majority class of its nearest neighbor is the major task of KNN algorithms.

$$Class = arg_v max \sum_{(X_i,y_i)\in D_z} I(v = y_i) \qquad \ldots (1)$$

Here, the class label is represented by v. The class label for i^{th} nearest neighbors is denoted by y_i. An indicator function is denoted by I, in which if the argument is true, the value of 1 is returned and otherwise, 0 value is returned. Thus, within the class of its K nearest neighbors, the samples are assigned. A set of labeled objects, a distance or similarity metric that calculates the distance amongst objects and the number of nearest neighbors that is the value of k, are the three important elements within the KNN approach. In order to make the recognition task successful, the selection of an appropriate similarity function as well as value for parameter k is important. For understanding as well as implementation of classification techniques, KNN classification is known to be simple and easy.

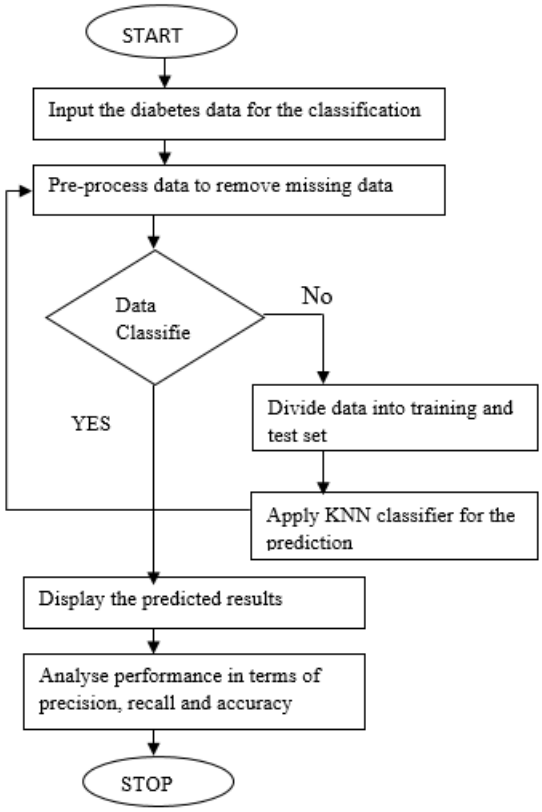

Fig 1: Proposed Methodology

IV. Result and Discussion

The proposed algorithm is implemented in Anaconda in which python programming language is used for the result analysis. The data set is collected from the UCI repository which has 9 attributes.

1. **For SVM:** The diabetes dataset has been predicted using the SVM classifier by implementing it using Python and the final results has been analysed in terms of precision, recall and accuracy. The existing system is based on the SVM classifier

```
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
The accuracy is 80.00%
```

Permissions: RW | End-of-lines: CRLF | Encoding: UTF-8

Fig 2: SVM Results

2. **For KNN:** The dataset of the diabetes has been then implemented using the KNN classifier in the python and the final results has been analysed in terms of precision, recall and accuracy. Due to the multiclass classification the KNN is more accurate.

```
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
Accuracy of K-NN classifier on training set: 0.79
Accuracy of K-NN classifier on test set: 0.78
The accuracy is
83.1597222222
```

Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Li

Fig 3: KNN Results

Comparison Results of SVM and KNN: After analysing the predicted results the performance of both the classifiers is analysed on the basis of accuracy and execution time.

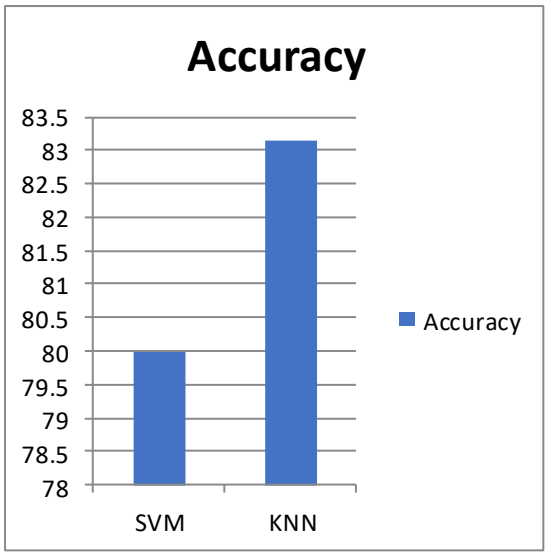

Fig 4: Accuracy Comparison

As shown in figure 4, the accuracy of SVM and KNN is compared for the diabetes prediction. The KNN has the multiple hyper planes due to which classification accuracy is increased at steady rate.

Fig 5: Execution Time

As shown in figure 5, the execution time of KNN classifier is compared with SVM classifier. Due to multiclass classification execution time of KNN is less as compared to SVM classifier

V. Conclusion

In this work, it is concluded that diabetes prediction is applied using the approach of classifications. To implement prediction analysis, whole data is divided into training and test sets. In the existing system, SVM classifier is applied for the prediction analysis. In the proposed system, KNN classifier is implemented for the prediction analysis. The performance of both classifiers is compared in terms of accuracy and execution time. The results show that KNN is more accurate as compare to SVM and moreover the time taken by KNN is comparatively less than SVM.

References

[1] Ashish kumar Dogra and Tanu Walia, A Review Paper on Data Mining Techniques and Algorithms, May 2015, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5.

[2] Nirmal Kaur and Gurpinder Singh, A Review Paper On Data Mining And Big Data, May 2017 (Special Issue), International Journal of Advanced Research in Computer Science, Volume 8, No. 4.

[3] Yanhui Sun, Liying Fang and Pu Wang, Improved k-means clustering based on Efros distance for longitudinal data, 2016 Chinese Control and Decision Conference (CCDC), Vol. 11, issue 3, pp. 12-23, 2016.

[4] Shunye Wang, Improved K-means clustering algorithm based on the optimized initial centroids, 2013 3rd International Conference on Computer Science and Network Technology (ICCSNT), Vol. 11, issue 3, pp. 12-23, 2013.

[5] Phattharat Songthung and Kunwadee Sripanidkulchai, Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification, 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Vol. 11, issue 3, pp. 12-23, 2016.

[6] Jyoti, Neha Kaushik, Rekha, "Review paper on Clustering and Validation Techniques", International Journal for Research in Applied Science and Engineering Technology", vol. 2, pp. 182-186, 2014.

[7] Dr. Sankar Rajagopal, "Customer data clustering using data mining technique", International Journal of Database Management Systems (IJDMS), vol. 3, pp. 21- 32, 2011.

[8] Shai Shalev-Shwartz, Shai Ben-David, "Understanding Machine Learning: From Theory to Algorithms", vol. 8, issue 4, pp. 1-499, 2014.

[9] Bayu Adhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.

[10] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, "Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization", Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.

[11] Zhiqiang Ge, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning", 2017 IEEE. Translations and content mining are permitted for academic research only, vol. 5, pp. 20590-20616, 2017.

[12] Jahin Majumdar, Anwesha Mal, Shruti Gupta, "Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.

[13] M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202-222, 2017.

[14] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.