Prediction of Diabetes Using Data Mining Techniques and its Applications

PROF.KALAM NARREN
PROF.V.VINAY KRISHNA

[1]M.Tech Scholar, Department of Computer Science and Engineering, JNTUH College of Engineering, Jagityal.

[2]Professor of CSI, ISIE and Computer Science and Engineering, JNTUH College of Engineering, Jagityal.

***Abstract*—** Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is help to make predictions on medical data. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. In this paper, we explore the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The experiment result proves that ANN (Artificial Neural Network) provides the highest accuracy than other techniques.

• INTRODUCTION

HUMAN body needs energy for activation. The carbohydrates are broken down to glucose, which is the important energy source for human body cells. Insulin is needed to transport the glucose into body cells. The blood glucose is supplied with insulin and glucagon hormones produced by pancreas. Insulin hormones produced by the beta cells of the islets of Langerhans and glucagon hormones are produced by the alpha cells of the islets of Langerhans in the pancreas. When the blood glucose increases, beta cells are stimulated and insulin is given to the blood. Insulin enables blood glucose to get into the cells and this glucose is used for energy. So blood glucose is kept in a narrow range. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors. Diabetes affects human organs such as kidney, eye, heart, nerves, foot etc.

• RELATED WORK

[2] Survey of Machine Learning Algorithms for Disease Diagnostic

In medical imaging, Computer Aided Diagnosis (CAD) is a rapidly growing dynamic area of research. In recent years, significant attempts are made for the enhancement of computer aided diagnosis applications because errors in medical diagnostic systems can result in seriously misleading medical treatments. Machine learning is important in Computer Aided Diagnosis. After using an easy equation, objects such as organs may not be indicated accurately. So, pattern recognition fundamentally involves learning from examples. In the field of bio-medical, pattern recognition and machine learning promise the improved accuracy of perception and diagnosis of disease. They also promote the objectivity of decision-making process. For the analysis of high-dimensional and multimodal bio-medical data, machine learning offers a worthy approach for making classy and automatic algorithms. This survey paper provides the comparative analysis of different machine learning algorithms for diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue disease and hepatitis disease. It brings attention towards the suite of machine learning algorithms and tools that are used for the analysis of diseases and decision-making process accordingly.

Statistical models for estimation that are not capable to produce good performance results have flooded the assessment area. Statistical models are unsuccessful to hold categorical data, deal with missing values and large data points. All these reasons arise the importance of MLT. ML plays a vital role in many applications, e.g. image detection, data mining, natural language processing, and disease diagnostics. In all these domains, ML offers possible solutions. This paper provides the survey of different machine learning techniques for diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue and hepatitis disease. Many algorithms have shown good results because they identify the attribute accurately. From previous study, it is observed that for the detection of heart disease, SVM provides improved accuracy of 94.60%. Diabetes disease is accurately diagnosed by Naive Bayes. It offers the highest classification accuracy of 95%. FT provides 97.10% of correctness for the liver disease diagnosis. For dengue disease detection, 100% accuracy is achieved by RS theory. The feed forward neural network correctly classifies hepatitis disease as it provides 98% accuracy. Survey highlights the advantages and disadvantages of these algorithms. Improvement graphs of machine learning algorithms for prediction of diseases are presented in detail. From analysis, it can be clearly observed that these algorithms provide enhanced accuracy on different diseases. This survey paper also provides a suite of tools that are developed in community of AI. These tools are very useful for the analysis of such problems and also provide opportunity for the improved decision making process.

[3] Extreme learning machine: Theory and applications

It is clear that the learning speed of feedforward neural networks is in general far slower than required and it has been a major bottleneck in their applications for past decades. Two key reasons behind may be: (1) the slow gradient-based learning algorithms are extensively used to train neural networks, and (2) all the parameters of the networks are tuned iteratively by using such learning algorithms. Unlike these conventional implementations, this paper proposes a new learning algorithm called extreme learning machine (ELM) for single-hidden layer feedforward neural networks (SLFNs) which randomly chooses hidden nodes and analytically determines the output weights of SLFNs. In theory, this algorithm tends to provide good generalization performance at extremely fast learning speed. The experimental results based on a few artificial and real benchmark function approximation and classification problems including very large complex applications show that the new algorithm can produce good generalization performance in most cases and can learn thousands of times faster than conventional popular learning algorithms for feedforward neural networks.

This paper has demonstrated that ELM can be used efficiently in many applications, however, two more interesting aspects are still open: the universal approximation capability of ELM and the performance of ELM in sparse high-dimensional applications, which are currently under our investigation.

• FRAMEWORK

Neural Networks are essentially mathematical models to solve an optimization problem. They are made of neurons, the basic computation unit of neural networks. A neuron takes an input (say x), do some computation on it (say: multiply it with a variable w and adds another variable b) to produce a value (say; z= wx + b). This value is passed to a non-linear function called activation function (f) to produce the final output (activation) of a neuron. There are many kinds of activation functions. One of the popular activation function is Sigmoid. The neuron which uses sigmoid function as an activation function will be called sigmoid neuron.

Depending on the activation functions, neurons are named and there are many kinds of them like RELU, TanH.

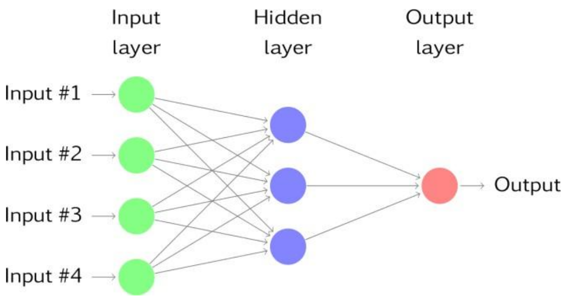


Fig.1. System Design

To predict disease class multiple layers operate on each other to get best match layer and this process continues till no more improvement left.

SVM Algorithm:

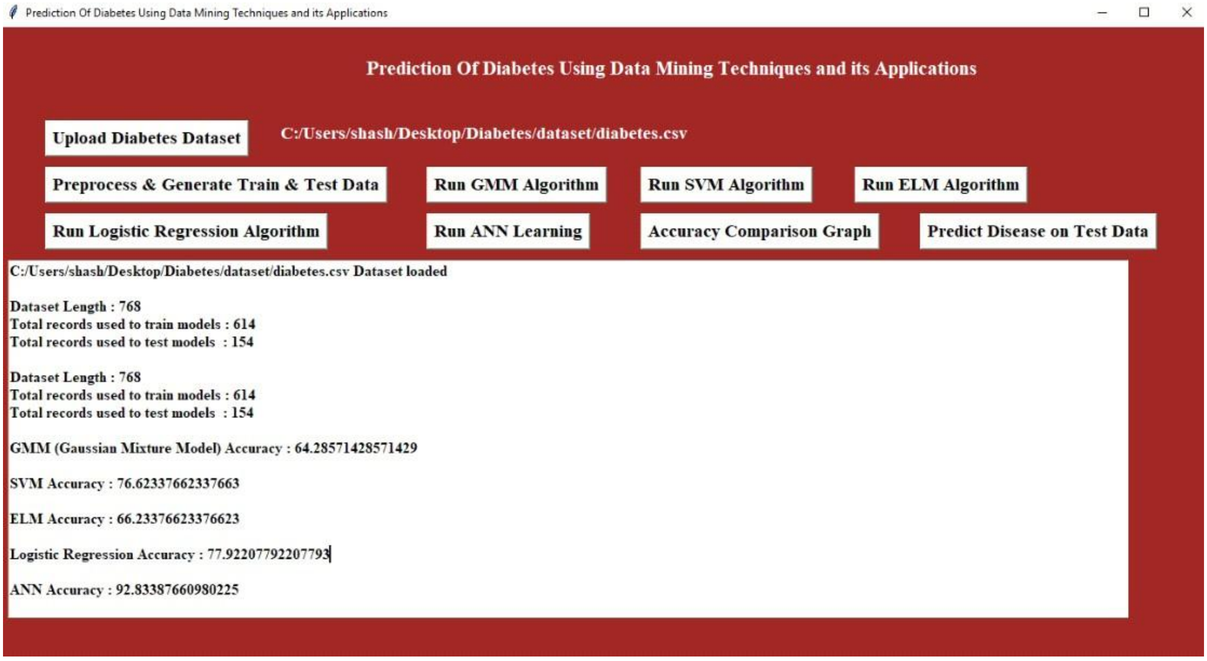
Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

ANN Algorithm:

To demonstrate how to build an ANN neural network based image classifier, we shall build a 6 layer neural network that will identify and separate one image from other. This network that we shall build is a very small network that we can run on a CPU as well. Traditional neural networks that are very good at doing image classification have many more parameters and take a lot of time if trained on normal CPU. However, our objective is to show how to build a real-world convolutional neural network using TENSORFLOW.

• EXPERIMENTAL RESULTS

In this paper author is evaluating performance of various data mining algorithm such as SVM, Logistic Regression, Gaussian Mixture Model (GMM), Artificial Neural Network (ANN) and EML (Extreme Machine Learning) to predict diabetes disease. Among all algorithms ANN is giving better prediction Accuracy.


Prediction Of Diabetes Using Data Mining Techniques and its Applications
Prediction Of Diabetes Using Data Mining Techniques and its Applications
Upload Diabetes Dataset
C:/Users/shash/Desktop/Diabetes/dataset/diabetes.csv
Preprocess & Generate Train & Test Data
Run GMM Algorithm
Run SVM Algorithm
Run ELM Algorithm
Run Logistic Regression Algorithm
Run ANN Learning
Accuracy Comparison Graph
Predict Disease on Test Data

```
C:/Users/shash/Desktop/Diabetes/dataset/diabetes.csv Dataset loaded

Dataset Length : 768
Total records used to train models : 614
Total records used to test models  : 154

Dataset Length : 768
Total records used to train models : 614
Total records used to test models  : 154

GMM (Gaussian Mixture Model) Accuracy : 64.28571428571429

SVM Accuracy : 76.62337662337663

ELM Accuracy : 66.23376623376623

Logistic Regression Accuracy : 77.92207792207793

ANN Accuracy : 92.83387660980225
```

Fig.2. In above screen with ANN we got 92% accuracy and below is the ANN model details in console.

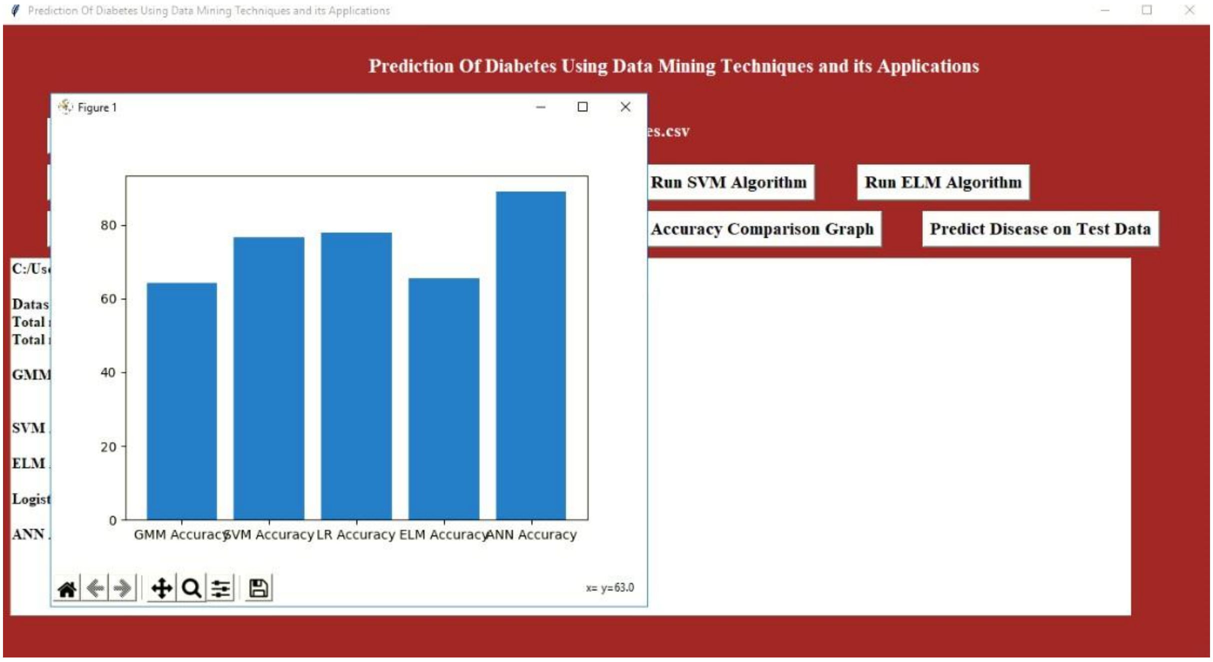
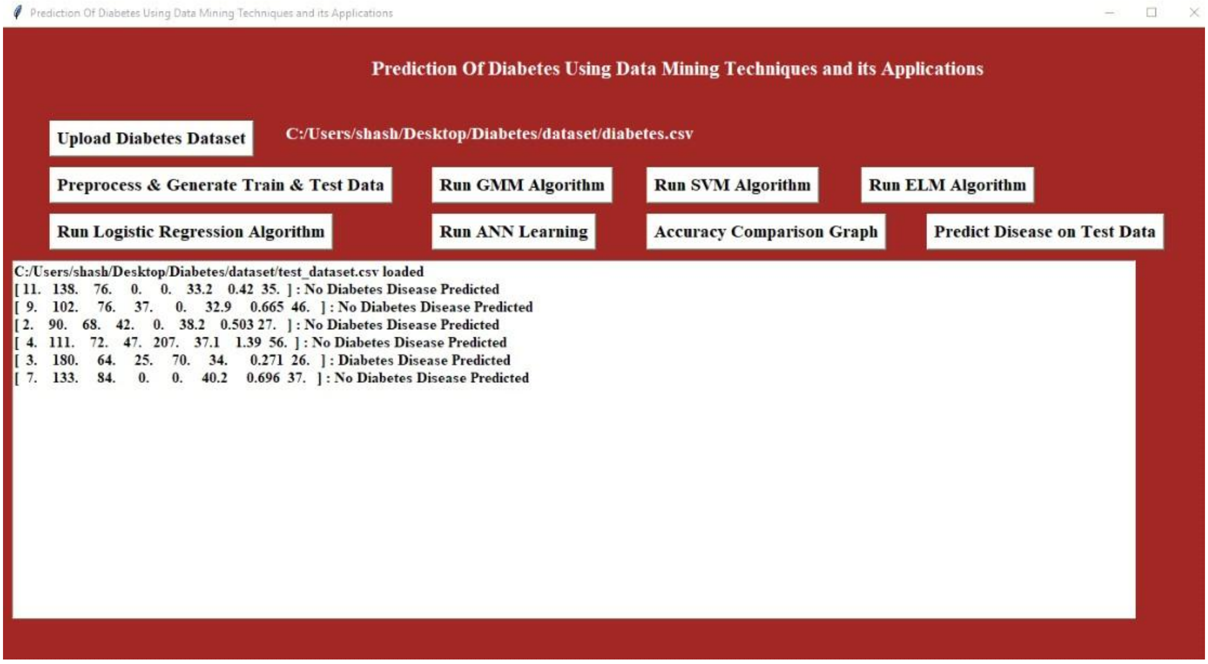


Fig.3. In above screen x-axis represents algorithm names and y-axis represents accuracy of those algorithms and from above graph we can conclude that ANN is better in performance.


Prediction Of Diabetes Using Data Mining Techniques and its Applications
Prediction Of Diabetes Using Data Mining Techniques and its Applications
Upload Diabetes Dataset
C:/Users/shash/Desktop/Diabetes/dataset/diabetes.csv
Preprocess & Generate Train & Test Data
Run GMM Algorithm
Run SVM Algorithm
Run ELM Algorithm
Run Logistic Regression Algorithm
Run ANN Learning
Accuracy Comparison Graph
Predict Disease on Test Data

```
C:/Users/shash/Desktop/Diabetes/dataset/test_dataset.csv loaded
[ 11.  138.   76.   0.   0.   33.2   0.42  35.  ] : No Diabetes Disease Predicted
[ 9.   102.   76.   37.   0.   32.9   0.665  46.  ] : No Diabetes Disease Predicted
[ 2.   90.   68.   42.   0.   38.2   0.503 27.  ] : No Diabetes Disease Predicted
[ 4.   111.   72.   47.   207.   37.1   1.39  56.  ] : No Diabetes Disease Predicted
[ 3.   180.   64.   25.   70.   34.   0.271  26.  ] : Diabetes Disease Predicted
[ 7.   133.   84.   0.   0.   40.2   0.696  37.  ] : No Diabetes Disease Predicted
```

Fig.4 In above screen first we are displaying test values and beside it display predicted result as diabetes detected or not.

• CONCLUSION

The diabetes prediction system is developed using five data mining classification modeling techniques. These models are trained and validated against a test dataset. All five models are able to extract patterns in response to the predictable states. The most effective model to predict patient with diabetes appear to be ANN followed by ELM and GMM. Although not the most effective model, the Logistic regression result is easier to read and interpret, what is more, the training over Logistic regression is very efficient. Although the ANN do outperform other data mining methods, the relationship between attributes and the [mal result is more difficult to understand. Although we achieved a fair accuracy over the prediction of diabetes, our study still has several limitations. The primary limitation of this study is its small sample size, which made it very difficult for any of the endpoints to achieve statistical significance. The second limitation was that we did not directly measure medication adherences. Finally, our data was mainly based on patient information. However, this study only illustrates a potential use of the data mining method. In the medical field accuracy in prediction of the diseases is the most important factor. In the analysis of data mining 1009 techniques, ANN classifier gives 89% of highest accuracy.

REFERENCES

[1] Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1 (2016): 727-730.

[2] Berry, Michael 1., and Gordon Linoff. Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc., 1997

[3] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[4] Emoto, Takuo, et al. "Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease." Heart and vessels 32.1 (2017): 39-46.

[5] Giri, Donna, et al. "Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform." Knowledge-Based Systems 37 (2013): 274-282.

[6] Fatima, Meherwar, and Maruf Pasha. "Survey of Machine Learning Algorithms for Disease Diagnostic." Journal of Intelligent Learning Systems and Applications 9.01 (2017): 1.

[7] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[8] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70.1 (2006): 489-501.

[9] Tiwari, Mukesh, Jan Adamowski, and KazimierzAdamowski. "Water demand forecasting using extreme learning machines." Journal of Water and Land Development 28.1 (2016): 37-52.

[10] U-;;ar, Ay�egUI, Yakup Demir, and CUneytGUzeli�. "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine

initialized with spherical clustering." Neural Computing and Applications 27.1 (2016): 131-142