PREDICTION OF DIABETES MELLITUS USING RBF NEURAL MODEL AND GENETIC ALGORITHM

K Rashmi¹, N V Krishna Rao², Dr. M Madhu Bala³, M Lahari⁴, Neha Fathima⁵, V Prudhvi⁶
Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, Hyderabad-500043

ABSTRACT

Diabetes Mellitus is a common disease of the human body caused by a group of metabolic disorders where the sugar levels over a prolonged period are very high. It affects various organs in the human body, causing damage to a large number of body systems, especially the blood vessels and nerves. Early detection and treatment of such diseases will save lives. To achieve the target, this study uses machine learning techniques to investigate various risk factors associated with this disease. Machine learning techniques provide an efficient result to extract knowledge by constructing predicting models from diagnostic medical data sets collected from diabetic patients. It may be possible to predict diabetic patients by extracting information from such data. In this work, we employ four popular machine learning algorithms, namely Support Vector Machine, Naive Bayes, K-Nearest Neighbor, C4.5 Decision Tree including and optimization of Radial Basis Neural Network using Cluster Validity Index and Genetic Algorithm on adult population data to predict Diabetes Mellitus.

Keywords: Diabetes Mellitus, RBF Neural Network, Genetic Algorithm, Machine Learning

I. INTRODUCTION

Diabetes is a metabolic and hereditary disease caused due to deficiency of insulin hormone in the human body. Insulin plays a key role in the conversion of food into energy. Lack of sufficient insulin causes the presence of excess sugar levels in the blood. As a consequence, diabetic patients' glucose levels are higher than average. Diabetes is referred to as Diabetes Mellitus (DM). It has symptoms like frequent urination, increased hunger, increased thirst, and high blood sugar. Diabetes is the fastest rising long-term illness condition that impacts lots of people globally. Excess blood sugar in the blood vessels can cause damage to the blood vessels, which can lead to complications such as cardiovascular damage, kidney damage, nerve damage, eye damage, and stroke. World Health Organization (WHO) statistics show diabetes contributed a major share in Non-Communal Disease (NCD) deaths across the worldwide population. More commonly referred to as “diabetes” -- a chronic disease associated with abnormally high levels of sugar glucose in the blood. Diabetes is due to one of two mechanisms:

- Inadequate production of insulin (which is made by the pancreas and lowers blood glucose), or
- Inadequate sensitivity of cells to the action of insulin.

Insulin-dependent (type1) and non-insulin-dependent (type2) diabetes are the two primary forms of diabetes that lead to these two mechanisms. In type 1 diabetes, there is little or insufficient insulin. There is usually enough insulin in type 2 diabetes, but the cells on which it should function are not necessarily responsive to it. Increased urine production, reduced appetite, and fatigue are both signs and symptoms of both forms of diabetes. Diabetes is diagnosed by blood glucose testing, the glucose tolerance test, and testing of the level of glycosylated hemoglobin (hemoglobin A1C). The mode of treatment depends on the type of diabetes. Diabetes has many major complications, including dangerously high blood sugar, abnormally low blood sugar caused by diabetes medications, and blood vessel disease, which can damage the eyes, kidneys, nerves, and heart. This study focuses on predicting diabetes using various machine learning methods while taking into account various risk factors associated with the disease.
Software Details:
- Programming Language: Python3, SQL
- Packages Used: NLP, python libraries (like Scipy, Sklearn)

Hardware Details:
- Device (Mobile/Laptop)
- Monitor to view the output
- Web browser

Fig.1: Block diagram for Predicting Diabetes Mellitus

The Existing solutions: Many researchers have conducted studies in the area of diabetes by using machine learning techniques to extract knowledge from available medical data. The recent study developed a predictive analysis model using KNN, J48, SVM, and Random Forest algorithm, where the J48 machine learning algorithm provides better performance and accuracy than others before preprocessing technique. The cross-validation method did not improve the classification algorithms' efficiency. To predict and control diabetes, data mining techniques like IB1, Naive Bayes, and C4.5 have been used on a dataset gathered from Ulster Community and Hospitals Trust (UCHT). By applying the feature selection technique, the performance of IB1 and Naive Bayes provided a better result. ANN, Logistic regression, and J48 have been used to predict diabetic diseases using real-world data sets by collecting information by the distributed questioner. Finally, it was concluded as J48 machine learning techniques provided efficient and better accuracy than others.

II. RELATED WORK

Performance Analysis of Classifier Models to Predict Diabetes Mellitus-2015:

In this approach, a comparison of four prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. One is before preprocessing the dataset. Here the studies conclude that the decision tree J48 classifier achieves a higher accuracy of 73.82% than the other three classifiers. In other cases that are after pre-processing the dataset, we have more accurate results when compared to the first study. In this case, both KNN (k=1) and Random Forest perform much better than the other three classifiers and they provide 100% accuracy. In this approach, the model did not perform well when it had noisy data.

Performance Analysis of Classification Algorithms on Diabetes Dataset-2016:

The output of an algorithm and the time it takes to create a model are both influenced by the type of data we use as input, and the data mining methods we use are also influenced by the dataset's design. Various classification techniques have been analyzed and it is inferred that the classification techniques best suit for the prediction of
results. Though various techniques have been used, the diagnosis of disease suffers false alarm, and the detection rate is low.

Machine Learning-based Framework to Identify Type-2 Diabetes through Electronic Health Record-2017: This work proposed an accurate and efficient framework as a pilot study to identify subjects with and without T2DM from EHR data. The framework leverages machine learning to automatically extract patterns of T2DM and further boost its predictive power by overcoming the wide separation range of cases and controls in expert algorithms. The experimental results show that this framework can identify subjects with and without T2DM at an average AUC of around 0.98, significantly outperforming the state-of-the-art at an AUC of 0.71.

III. PROCEDURE AND METHODOLOGY

Data Acquisition

We have input data from the Pima Indians diabetes database for the prediction of diabetes. The data includes 768 rows and 9 columns of features like glucose, BMI, age, pregnancies, insulin, skin thickness, diabetes pedigree function, blood pressure.

Data Preprocessing

Data preprocessing is a data mining technique that entails converting raw data into an understandable format that our system can understand and process more efficiently. Real-world data is often incomplete, unreliable, and/or deficient in specific habits or patterns, as well as containing numerous errors. Data preprocessing is a proven method of resolving such issues. Preprocessing usually includes removing/replacing null values, encoding values, to capture the essential features, and remove the unnecessary features among others that are highly correlated or redundant data. Applying standard scalar before training the model as StandardScalar performs the task of Standardization. In most cases, a dataset includes variables of varying scales. For e.g. an Employee dataset will contain an AGE column with values on a scale of 20-70 and a SALARY column with values on a scale of 10000-80000. Since the scales of these two columns vary, they are standardized to create a similar scale while creating a machine learning model.

Training the Data

Training data is used to train an algorithm. Generally, training data is a certain percentage of an overall dataset along with a testing set. The data is divided into training and testing sets. The training data is used to ensure that the machine recognizes patterns in the data, ensuring that the algorithm used to train the machine is more accurate and effective. As a rule, the better the training data, the better the algorithm or classifier performs. To generate an output, the neuron process the input signal through a function called the activation function of the neuron. In RBF activation function of a hidden neuron is \( \phi(X) \) i.e. for an input vector \( X \) output produced by the hidden neuron will be \( \phi(X) \). Gaussian neural activation function with center (mean) for 1-D input \( x \). The basic concept behind this model is that Gaussian neural nodes partition the entire feature vector space, with each node generating a signal corresponding to an input vector, and the intensity of each neuron's signal being determined by the distance between its center and the input vector. The output signals produced must also be similar for inputs that are closer in Euclidian vector space. Here, \( \mu \) is the center of the neuron and \( \phi(X) \) is the response of the neuron corresponding to \( X \).

Gaussian neural nodes and the boundary of circles represents the range of the corresponding nodes also known as the receptive field of neurons. Here 2-D vector space is partitioned by 12 Gaussian nodes. Every input vector activates the collective system of neurons to some extent and the combination of these activations enables RBF to decide how to respond. The above configuration of neurons will generate similar output signals for input vectors A and B whereas for C output generated will be quite different.
In RBF architecture, weights connecting input vector to hidden neurons represents the center of the corresponding neuron. These weights are predetermined in such a way that the entire space is covered by the receptive field of these neurons, whereas values of weights connecting hidden neurons to output neurons are determined to train the network.

K Means Clustering algorithm:

- Choose the number of cluster centers “K”.
- randomly choose K points from the dataset and set them as K centroids of the data.
- for all the points in the dataset, determine the centroid closest to it.
- for all centroids, calculate the average of all the points lying closest to the same centroid.
- Change the value of all the centroids to corresponding averages calculated in (4).
- Go to (3) until convergence.

The range of receptive fields is chosen such that the entire domain of the input vector is covered by the receptive field of the neurons. So, the value of sigma is chosen according to the maximum distance “d” between two hidden neurons. Where d is the maximum distance between two hidden neurons and M is the number of hidden neurons.

$$\sigma = \frac{d}{\sqrt{2M}}$$

Applying Trained Data on Various Machine Learning Classifiers

Various machine learning techniques can predict diabetes mellitus. However, it is very difficult to choose the best technique to predict based on such attributes. So we are applying trained data on SVM, Naive Bayes, decision tree, k-means and is made available for testing. But along with it, we are using optimized radial basis neural
function as it does not pertain to the model we are implementing the entire logic, preprocessed data is followed by making clusters of data using k-means clustering logic, then applying Gaussian function as kernel function to optimize the clusters formed, and to name clusters class by class we are using Dunn, Davies Bouldin, silhouette validity indexes and computing inverse-weight matrix to train the model. We are applying a genetic algorithm to know the accuracy of the model when only the best features selected from the data are given to the output.

Algorithm

- Define the number of hidden neurons “K”.
- set the positions of RBF centers using the K-means clustering algorithm.
- Calculate \( \sigma \) using equation (2)
- Calculate actions of RBF node using equation (1)
- Train the output using equation (3)

Testing the Trained Data

The test set is a set of observations used to evaluate the performance of the model using some performance metrics here we used accuracy.

IV. IMPLEMENTATION AND RESULTS

Dataset Preprocessing and Training the data

Data preprocessing is the essential section before applying a model involves the steps like importing the libraries, importing the Dataset, Handling of Missing Data, Handling of Categorical Data, Splitting the dataset into training and testing datasets, and finally Feature Scaling. The following picture depicts the output of Data Preprocessing and Training the dataset for prediction of Diabetes.

Fig. 3: Description of The Diabetes Dataset

Applying Clustering Techniques
Fig 4: Indexing To Clusters

Clustering which is an unsupervised machine learning task is used here for labeling the clusters for further use in the RBFN model. Here K-Means Clustering is used for assigning examples to clusters to minimize the variance within each cluster. Along with K-Means, the Dunn Index (DI) metric is used for judging a clustering algorithm. The Davies–Bouldin index is used as a metric for evaluating clustering algorithms. This is an internal assessment scheme that uses the dataset's inherent quantities and features to validate how well the clustering was performed.

Constructing the RBFN and Predicting Its Accuracy

In this section, we would be predicting the RBFN accuracy by applying the K-means clustering technique implemented above.

Computing Various Machine Learning Techniques

Various machine learning accuracies like gradient Boosting, Naive Bayes, SVM, Decision Tree, and Logistic Regression are calculated.

Predicting Genetic Algorithm Accuracy Using SVM Classifier as RBF kernel Function
The following output predicts the Genetic Algorithm's accuracy using SVM Classifier as RBF kernel Function with gamma and cost as hyper-parameters by selecting the key features trained in the dataset.

**Performance Analysis of Computed Models as a Bar Plot**

![Bar Plot of Performance Analysis](image)

**Comparative Difference of All Computed Models**

On the same dataset, different models were run, and each model produced different performance metrics. The accuracy achieved by the models is summarized in the table below.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radial Basis Neural Network Function</td>
<td>72%</td>
</tr>
<tr>
<td>Genetic Algorithm</td>
<td>67%</td>
</tr>
<tr>
<td>DUNN Index</td>
<td>5%</td>
</tr>
<tr>
<td>Davies Bouldin Index</td>
<td>71%</td>
</tr>
<tr>
<td>Silhouette Index</td>
<td>56.8%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>78.5%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>80.5%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>69%</td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td>75.9%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>79%</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>78.5%</td>
</tr>
</tbody>
</table>

### V. CONCLUSION AND FUTURE SCOPE

This project proposed a new classification model for classifying diabetes patients by taking into account various factors related to this disease. The proposed model integrates cluster validity index with k-means clustering algorithm in Radial Basis Neural Network. To predict diabetes mellitus, we have also done our experiments using machine learning algorithms, namely Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and decision tree, on adult population data to predict diabetes mellitus. Though the Naïve Bayes algorithm yields the highest accuracy, i.e. 79.2% but Naïve Bayes assumes class conditional independence and therefore loses inaccuracy. To have an accurate model, we have used a genetic algorithm with an SVM classifier as kernel function and RBFNN neural model producing accuracies of 67% and 72%. Our model was used to classify diabetes patients into one of two classes (positive/negative). Cluster validity index integrated with the k-
means algorithm to guarantee the optimal cluster locations. Optimizing cluster centers minimized the classification time by reducing network complexity. This project, therefore, is useful to determine the prediction of Diabetes using the Genetic Algorithm model to have an accurate diabetes prediction reducing the false alarm. As future work, hybrid particle swarm optimization can be used for determining weights. Also, we can apply other kernel functions like polyharmonic spline, inverse quadratic, etc., in the classification phase.

REFERENCES

10. Abdullah A. Aljumah et al., Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, July 2013, Pages 127-136