**DIABETES PREDICTION**

**INTRODUCTION**

Data mining is the process of extracting hidden knowledge from large volumes of data. The goal of the data mining methodology is to think data from a data set and change it into a reasonable structure for further use. Data mining plays an efficient role in prediction of diseases in health care industry.

Prediction of diseases by analysis of voluminous historical data is one of the most significant applications of data mining. Medical data mining is the process of finding useful patterns that would be helpful in medical diagnosis. The predictability of diabetes will be more effective since earlier detection of disease will be helpful for the patients to take care of themselves.

**PURPOSE**

Our examination concentrates on this part of Medical conclusion learning design through the gathered data of **diabetes** and to create smart therapeutic choice emotionally supportive network to help the physicians.

In this system, we propose the use of algorithms like PSO to apply on diabetes patient's database and analyse them by taking various attributes of diabetes for prediction of diabetes disease.

The involvement of intelligent optimization techniques has been found effective to enhance the complex, real time, and costly data mining process. One of such optimization technique is a **Particle Swarm Optimization (PSO)** based clustering technique.

**PROBLEM STATEMENT**

Diabetes is a common disease that affects a vast majority of the people in many parts of the world. Diabetes affects people usually after the age of 20. According to WHO statistics, the global prevalence of diabetes among adults above 18 years of age has risen to 8.5% in 2014. It becomes a cause for other illnesses also like blindness, kidney failure, cholesterol and heart diseases. The deaths due to diabetes and high blood glucose are on the rise.

Prediction of diabetes at an early stage would help the patients to maintain the sugar level under control. As data mining techniques prove to be good in predictive analyses, a data mining approach is used to predict the risk of diabetes in the proposed approach. The performance of the algorithm is also measured and improved using feature selection.

**METHODOLOGY**

The sample dataset is selected and separated into feature (labels to be considered for prediction) and target (label to predict). Feature selection is an important problem in knowledge discovery. The main aim is to find a feature subset that produces higher classification accuracy. After selection of features, the classification algorithm is applied to build the classification model. Then the model is applied to the data set for predicting the diabetes risk. The performance metrics are measured and evaluated. Feature selection is a data pre-processing step. This will select the subset of features from whole feature set based on statistical score and will remove redundant features that do not contribute to performance.

**PROPOSED SYSTEM**

In the proposed system data mining has been done in different phases data set has to be separated into feature (labels to be considered for prediction) and target (label to predict).

**Pre-processing**: In this phase dataset has been pre-processed for the removal of empty sets and anomalies available in the dataset. Data cleaning removes the observations containing noise and those with missing data.

**Feature Extraction**: Data mining approaches have been implemented in the data set to develop the rule that can be implemented on the dataset to identify the hidden pattern from the dataset that can be utilized on the time of classification. Data mining techniques are operated on large volumes of data to explore hidden patterns and relationships that are further helpful in decision making.  While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

**Parameter Evaluation**: After extraction of the features from the dataset different results have been evaluated by using such labels in it. On the basis of these samples various parameters have to be evaluated for performance evaluation and validation of the proposed work.

**CONCLUSION**

The model has been trained to classify the diabetes patients from non-diabetes persons and it is used to predict the risk of diabetes on dataset. The performance of the model has been evaluated using the performance measures such as accuracy, sensitivity and specificity. The performance of the algorithm has been improved by feature subset selection and by varying the size of the dataset. The experiment result proves that **ANN (Artificial Neural Network)** provides the highest accuracy than other techniques.