Healthcare Recommendation System Using Ann Algorithm

[1] Mrs. N. Vishnu Devi, [2] S. Bala Subiramanian, [3] R. Navaneethan, [4] S. Bogar

[1]Assistant Professor, [2,3,4] Students
Information Technology,
K.L.N. College of Engineering, Madurai, India

***Abstract*—Diabetes, also known as chronic illness, is a group of metabolic diseases due to a high level of sugar in the blood over a long period. The risk factor and severity of diabetes can be reduced significantly if the precise early prediction is possible. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting diabetes with ML techniques. The prediction model is introduced with different combinations of features and several known classification techniques. The system is developed based on classification algorithms includes Random Forest, Logistic Regression, Gradient Boosting and Artificial Neural Network algorithms have been used. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms.**

I. INTRODUCTION

It is difficult to identify diabetes because of several contributory risk factors such as high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of diabetes among humans. The severity of the disease is classified based on various methods like Random Forest and Logistic Regression. The nature of diabetes is complex and hence, the disease must be handled carefully. We have also seen hybrid ml model is used in predicting the accuracy of events related to diabetes. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of diabetes. Diagnosis of diabetes is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of diabetes. Moreover, it is expensive and computationally difficult to analyse. Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the diabetes as a result, the ratio of death decreases. The main objective of this research is to improve the performance accuracy of diabetes prediction. We proposed a machine learning based diagnosis method for the identification of diabetes in this research work.

2. EXISTING SYSTEM

Diabetes is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Diabetes is very dangerous if not immediately treated on time. The existing system doesn't effectively classify and predict the disease in human body. Practical use oh healthcare database systems and knowledge discovery is difficult in diabetes diagnosis.

Disadvantages

- Doesn't Efficient for handling large volume of data.
- Theoretical Limits
- Incorrect Classification Results.
- Less Prediction Accuracy.

2.2 PROPOSED SYSTEM

The proposed model is introduced to overcome all the disadvantages that arises in the existing system. This system will increase the accuracy of the Supervised classification results by classifying the data based on the diabetic prediction and others using Random Forest classification algorithm. It enhances the performance of the overall classification results. Apply hybrid data mining techniques to the dataset to investigate if ML & DL techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.

Advantages

- High performance.
- Provide accurate prediction results.

It avoids sparsity problems.

3. IMPLEMENTATIONS

DATA SELECTION AND LOADING

- Data selection is the process of determining the appropriate data type and source, as well as suitable instruments to collect data.
- Data selection precedes the actual practice of data collection and it is the process where data relevant to the analysis is decided and retrieved from the data collection.
- Data loading refers to the "load" component.
- After data is retrieved and combined from multiple sources, cleaned and formatted, it is then loaded into a storage system, such as a cloud data warehouse.
- In this project, the diabetes dataset is used for detecting disease.

DATA PREPROCESSING

- The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.
- **Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  - ✓ Ignore the tuples:
    This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  - ✓ Fill the Missing values:
    There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- **Regression:**
  Data can be made smooth by fitting it to a regression function. The regression used may be linear or multiple.

SPLITTING DATASET INTO TRAIN AND TEST DATA

- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.
- Separating data into training and testing sets is an important part of evaluating data mining models.
- Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.
- To train any machine learning model irrespective what type of dataset is being used you have to split the dataset into training data and testing data.

CLASSIFICATION

Classification is the problem of identifying to which of a set of categories, a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on**.**

Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Logistic Regression Logistic Regression is a Machine Learning algorithm which is used for the **classification problems**, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

Artificial Neural Network Tutorial provides basic and advanced concepts of ANNs. Our Artificial Neural Network tutorial is developed for beginners as well as professions. The term "Artificial neural network" refers to a biologically inspired sub-field of artificial intelligence modeled after the brain. An Artificial neural network is usually a computational network based on biological neural networks that construct the structure of the human brain. Similar to a human brain has neurons interconnected to each other, artificial neural networks also have neurons that are linked to each other in various layers of the networks. These neurons are known as nodes. Artificial neural network tutorial covers all the aspects related to the artificial neural network

PREDICTION

Predictive analytics algorithms try to achieve the lowest error possible by either using “boosting” or “bagging”.

Accuracy − Accuracy of classifier refers to the ability of classifier. It predict the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Speed − Refers to the computational cost in generating and using the classifier or predictor.

Robustness − It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

Scalability − Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

Interpretability − It refers to what extent the classifier or predictor understands.

RESULT GENERATION

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

- Accuracy

  Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

$$AC= \frac{TP+TN}{TP+TN+FP+FN}$$

- Precision

  Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

$$Precision=\frac{TP}{TP+FP}$$

- Recall

  Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

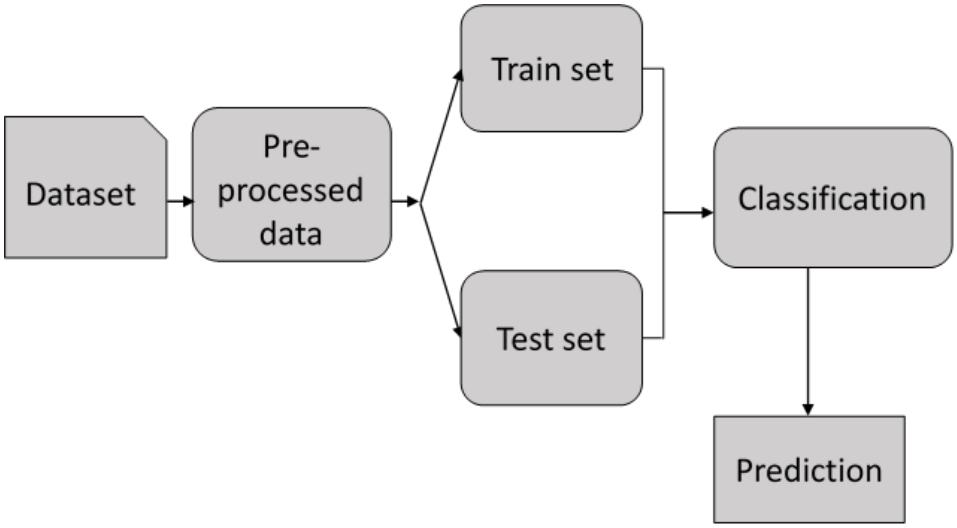
$$Recall=\frac{TP}{TP+FN}$$

- F-Measure

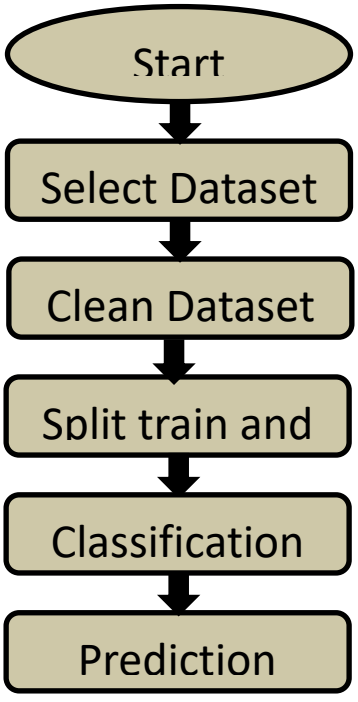
  F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

$$F\text{-}measure=\frac{2TP}{2TP+FP+FN}$$

ARCHITECTURE DIAGRAM



FLOW DIAGRAM



4.METHODOLOGY

The methodology of the project is to predict the diabetic disease. The predicted feature has a value of 1, then will be a diseased person; if the value is 0, then it will be Normal. The dataset is divided into two parts: 70% of the data is reserved for training and 30% of the data is reserved for testing. The machine learning and deep learning techniques are used to detect the diabetes disease. The proposed deep learning algorithm is Artificial Neural Network is used to detect the diabetes disease by applying the K-fold cross validation method to increases the performance of prediction. Finally, it will generates the metrics in terms of accuracy, precision, recall and f1-score.

ANN

Artificial neural networks (ANNs) use learning algorithms that can independently adjust – or learn, in a sense – as they receive new input. This makes them a very effective tool for non-linear statistical data modeling.

Deep learning ANNs play an important role in machine learning (ML) and support the broader field of artificial intelligence (AI) technology.

An artificial neural network has three or more layers that are interconnected. The first layer consists of input neurons. Those neurons send data on to the deeper layers, which in turn will send the final output data to the last output layer.

All the inner layers are hidden and are formed by units which adaptively change the information received from layer to layer through a series of transformations. Each layer acts both as an input and output layer that allows the ANN to understand more complex objects. Collectively, these inner layers are called the neural layer.

The units in the neural layer try to learn about the information gathered by weighing it according to the ANN's internal system. These guidelines allow units to generate a transformed result, which is then provided as an output to the next layer.

Hence, the error is used to recalibrate the weight of the ANN's unit connections to take into account the difference between the desired outcome and the actual one. In due time, the ANN will "learn" how to minimize the chance for errors and unwanted results.

5.CONCLUSION

In this process, we present the hybrid predictive models by using machine learning method Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF) and Artificial Neural Network (ANN) to predict diabetes disease. By method to Improve Expected Output of Semi-structured Sequential Data. It will enhance the performance of the predicted result. Finally generate the result based on accuracy, precision, recall and f1-score.

REFERENCES:

1. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. RezaAlbarrán, and K. L. Ramaiya, "Diabetes in developing countries," Journal of Diabetes, vol. 11, no. 7, pp. 522-539, Mar. 2019.
2. R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in Proc. International Conference on Computing Networking and Informatics, Oct. 2017, pp. 1-5.
3. Emerging Risk Factors Collaboration and other, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies," The Lancet, vol. 375, no. 9733, pp. 2215-2222, Jul. 2010.
4. N. H. Choac, J. E. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. W. Ohlrogge, and B. Malandaa, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Research and Clinical Practice, vol. 138, pp. 271-281, Apr. 2018.
5. P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karurangaa, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, R. Williams, and IDF Diabetes Atlas Committee, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation," Diabetes Research and Clinical Practice, vol. 157, pp. 107843, Nov. 2019.
6. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Annual Symposium on Computer Application in Medical Care, Nov. 1988, pp. 261-265.
7. M. Maniruzzaman, M. J. Rahman, M. A. M. Hasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," Journal of Medical Systems, vol. 42, no. 5, pp. 92, May 2018.
8. G. J. McLachlan, "Discriminant analysis and statistical pattern recognition," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 168, no. 3, pp. 635-636, Jun. 2005.
9. T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," IEEE Transactions on Electronic Computers, vol. 14, no. 3, pp. 326-334, Jun. 1965.
10. G. I. Webb, J. R. Boughton, and Zhihai Wang, "Not So Naive Bayes: Aggregating one-dependence estimators," Machine learning, vol. 58, no. 1, pp. 5-24, Jan. 2005.