Analysis and Future Risk Prediction of Diabetes Mellitus

Project ID:21D211230

Review -I

Group Members

RA1711003030211 Praffulla Kumar Dubey

RA1711003030230 Udbhav Naryani

Supervised By:

Ms. Medhavi Malik

Assistant Professor

Department of Computer Science & Engineering

Faculty of Engineering & Technology

SRM Institute of Science & Technology

March 18, 2021

Table of Contents

Abstract

The number of patients suffering from diabetes in India are increasing rapidly. One of the major reasons behind this can be that most of the population does not take regular check-ups for diabetes. Hence they are unable to take proper precautions and have a risk of getting diabetes in the future. The current methodologies followed by the doctors are manual and can be time consuming, hence we have come up with a solution for this problem, by providing a model in which the user can enter their health report details and know their future risk of having diabetes and if required they can consult a doctor for the same. In this project, we propose to analyse the Pima Indian diabetes dataset and develop an intelligent system for predicting future risk of diabetes using supervised machine/deep learning algorithms. We would be coding in python using Jupyter Notebook for analysis, training and testing of our model. We would be deploying our prediction model on localhost using Spyder IDE.

Objective

- Analyse and gain insights from the collected data.

- Develop a diabetes prediction model using supervised machine/deep learning algorithms.

- Predict future risk of diabetes.

- Develop a user friendly GUI (Graphical User Interface).

Literature Survey I

- S. Barik et al.,[2] in their paper "Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques"
  - Used XGBoost and Random Forest Algorithm and they did not perform data cleaning on the dataset.
  - They achieved an accuracy of 74.1% with XGBoost algorithm.
- H. Balaji et al.,[1] in their paper "Optimal predictive analytics of pima diabetics using deep learning"
  - Used only Recurrent Neural Network for training and testing their models.
- A. Mohebbi, et al.,[7] in their paper "A deep learning approach to adherence detection for type 2 diabetics."
  - Used Convolution Neural Network and the data-set used for training the model was small.

Literature Survey II

- Y. N. and SriPreethaa[8] in their paper "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster"
  - Used only Random Forest Algorithm on Hadoop Cluster.
  - They discussed information gain methods used for feature selection but pre-processing steps were not mentioned.
- F. Mercaldo et al.,[6] in their paper "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques"
  - Used HoeffdingTree, JRip, BayesNet, Random Forest and the highest accuracy achieved was 76% by using HoeffdingTree.
  - They used WEKA tool for analysis and prediction but they did not mention the pre-processing steps.

Literature Survey III

- S. Lekha and Suchetha M.[5] in their paper "Real-time non-invasive detection and classification of diabetes using modified convolution neural network"
  - Used one dimensional Modified Convolution Neural Network on data-set having eleven healthy, nine diabetic patient of type 2 and 5 diabetic patients of type 1.
  - The data collected was very small and no pre-processing of data-set was performed.
- J Pradeep and Saminathan[4] in their paper "Performance analysis of classifier models to predict diabetes mellitus"
  - Used J48, KNN and Random Forest Algorithm for prediction and got 73.8% accuracy with J48 algorithm.
  - They did not mention the data pre-processing techniques and methodology followed.

Literature Survey IV

- A. Iyer et al.,[3] in their paper "Diagnosis of diabetes using classification mining techniques"
  - Used only Naive Bayes and Decision Tree Algorithms and they got an accuracy of 79.5% from Naive Bayes classifier.
  - Pre-processing and data-set transformation was done using WEKA tool and we propose to use Python.
- Q. Zou et al.,[10] in their paper "Predicting diabetes mellitus with machine learning techniques."
  - Used Neural Networks, Random Forest and Decision Tree algorithms.
  - They achieved the highest accuracy of 76% with Neural Networks.
  - MATLAB tool was used for model creation by the authors and we propose to use python.
- S Nanda et al.,[9] in their paper "Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks"
  - Used only Logistic Regression for prediction and the accuracy of their model was not mentioned.

Architectural Design for Proposed System I

The architectural design for the proposed system is divided into 3 layers:

- **Data Layer:** Pre-process the data and gain knowledge from it.

- **Application Layer:** Create, train and test the diabetes prediction model.

- **Presentation Layer:** Create a GUI (Graphical User Interface) for the user to use the diabetes prediction model.
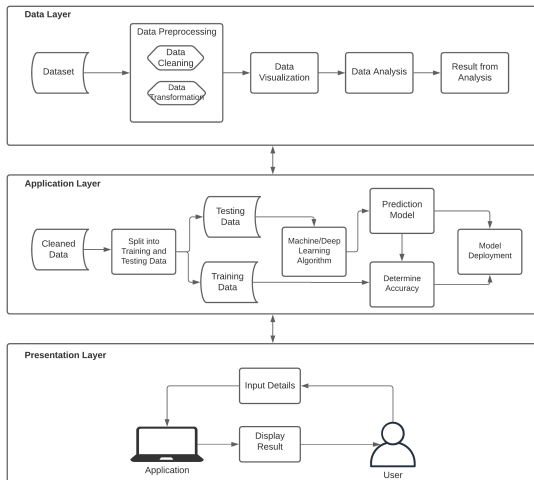
Figure: Architectural Design for Proposed System

Algorithms/Techniques to be used I

- **Tools/IDE to be used:**

  - **Jupyter Notebook:** Used for analysis of data-set and creation of prediction model using Python.

  - **Spyder IDE:** Used for developing GUI using Python.

- **Data Cleaning:** Cleaning the data-set. The data-set collected has a lot of null values and they have to be handled properly. Columns which are not relevant to analysis and prediction should be dropped from the data-set.

Algorithms/Techniques to be used II

- **Data Visualisation:** Visualise the data by plotting various graphs. This technique makes analysis of the data present in the data-set easy and can also be used for finding co-relation between different attributes of the data-set.

- **Data Analysis:** Analyse the various results of visualisations and gain knowledge about different attributes of the data-set which will help in selection of attributes for machine learning algorithm.

- **Feature Scaling:** Using Standard Scaler to scale all the values in data-set as the values in the data-set are of different units. This step might help in improving the accuracy of our model.

Algorithms/Techniques to be used III

- **Select Machine Learning Algorithm:** Select the best Machine/Deep Learning Algorithm (K-Nearest Neighbour, Support Vector Machine, Logistic Regression, Naive Bayes, Random Forest, Decision Tree, Perceptron and Multi-layer Perceptron) for creation of our prediction model.

- **Prediction Model:** Create the Prediction Model for predicting diabetes.

- **Design User Interface:** Develop a Graphical User Interface (GUI) design for the prediction model to be easily used by the user.

📄 H. Balaji, N. Iyengar, and R. D. Caytiles.
Optimal predictive analytics of pima diabetics using deep learning.
International journal of database theory and application, 10:47–62,
2017.

📄 S. Barik, S. Mohanty, S. Mohanty, and D. Singh.
*Analysis of Prediction Accuracy of Diabetes Using Classifier and
Hybrid Machine Learning Techniques*, pages 399–409.
01 2021.

📄 A. Iyer, s. Jeyalatha, and R. Sumbaly.
Diagnosis of diabetes using classification mining techniques.
*International Journal of Data Mining Knowledge Management
Process*, 5:1–14, 02 2015.

References II

📄 J. p. Kandhasamy and S. Balamurali.
Performance analysis of classifier models to predict diabetes mellitus.
Procedia Computer Science, 47:45–51, 12 2015.

📄 S. Lekha and S. M.
Real-time non-invasive detection and classification of diabetes using modified convolution neural network.
IEEE Journal of Biomedical and Health Informatics, 22(5):1630–1636, 2018.

📄 F. Mercaldo, V. Nardone, and A. Santone.
Diabetes mellitus affected patients classification and diagnosis through machine learning techniques.
Procedia Computer Science, 112:2519–2528, 12 2017.

📄 A. Mohebbi, T. Aradóttir, A. Johansen, H. Bengtsson, M. Fraccaro, and M. Mørup.
A deep learning approach to adherence detection for type 2 diabetics.
In *Proceedings of 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society*, 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (embc), pages 2896–9, United States, 2017. IEEE. 2017 39th Annual International Conference of the Ieee Engineering in Medicine and Biology Society, EMBC 2017 ; Conference date: 11-07-2017 Through 15-07-2017.

📄 Y. N. and K. SriPreethaa.
Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster.
Cluster Computing, 22, 01 2019.

S. Nanda, M. Savvidou, A. Syngelaki, R. Akolekar, and K. Nicolaides.
Prediction of gestational diabetes mellitus by maternal factors and
biomarkers at 11 to 13 weeks.
Prenatal Diagnosis, 31(2):135 – 141, Feb. 2011.

Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang.
Predicting diabetes mellitus with machine learning techniques.
Frontiers in Genetics, 9:515, 2018.