

Analysis and Future Risk Prediction of Diabetes Mellitus

Project ID:21D211230

Review - II

Group Members

RA1711003030211 Prafulla 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

April 5, 2021

Table of Contents

- 1 Abstract
- 2 Objective
- 3 Literature Survey
- 4 Architectural Design for Proposed System
- 5 Data-set Specification
- 6 Algorithms/Techniques to be used
- 7 Partial Implementation
- 8 Expected Outcome
- 9 References

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.

- 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

| Ref | Author | Year | Algorithm | Findings |
|------|--------------------------|------|---------------------|--|
| [2] | S. Barik et al. | 2021 | XGBoost: 74.1% | The data collected was very small and no pre-processing of data-set was performed. |
| [8] | Y. N. and SriPreethaa | 2019 | RF: NA | Used only Random Forest Algorithm on Hadoop Cluster. |
| [5] | S. Lekha and Suchetha M. | 2018 | CNN: NA | The data collected was very small and no pre-processing of data-set was performed. |
| [10] | Q. Zou et al. | 2018 | NN: 76% | MATLAB tool was used for model creation by the authors and we propose to use python. |
| [6] | F. Mer-caldo et al. | 2017 | Hoeffding Tree: 76% | They used WEKA tool for analysis and prediction but they did not mention the pre-processing steps. |

Table: Literature Survey I

Literature Survey II

| Ref | Author | Year | Algorithm | Findings |
|-----|--------------------------|------|------------|---|
| [1] | H. Balaji et al. | 2017 | RNN: NA | Used only Recurrent Neural Network for training and testing their models. |
| [7] | A. Mohebbi, et al. | 2017 | CNN: NA | Used Convolution Neural Network and the data-set used for training the model was small. |
| [4] | J Pradeep and Saminathan | 2015 | J48: 73.8% | They did not mention the data pre-processing techniques and methodology followed. |
| [3] | A. Iyer et al. | 2015 | NB: 79.5% | Pre-processing and data-set transformation was done using WEKA tool and we propose to use Python. |
| [9] | S Nanda et al. | 2011 | LR: NA | Used only Logistic Regression for prediction and the accuracy of their model was not mentioned. |

Table: Literature Survey II

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.

Architectural Design for Proposed System II

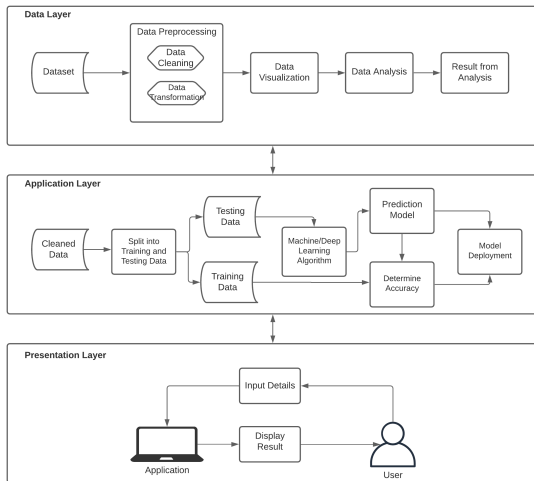


Figure: Architectural Design for Proposed System

Data-set Specification

| S.no | Attribute | Description |
|------|---------------------------|---|
| 1. | Pregnancies | Number of times pregnant. |
| 2. | Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test. |
| 3. | BloodPressure | Pressure of circulating blood against the walls of blood vessels (mm Hg). |
| 4. | SkinThickness | Triceps skin fold thickness (mm). |
| 5. | Insulin | Hormone that lowers the level of glucose in blood (μ U/ml). |
| 6. | BMI | Body mass index ($\text{weight in kg}/(\text{height in m})^2$). |
| 7. | DiabetesPedigree Function | A function which scores likelihood of diabetes based on family history. |
| 8. | Age | Age (in years). |
| 9. | Outcome | Class variable (0 or 1), 0 = Non Diabetic 1 = Diabetic |
| 10. | Doctor | Name of Doctor in-charge. |
| 11. | Hospital | Hospital Name. |

Table: Data-set Specification

- **Techniques:**

- **Data Cleaning:** Preparing the data-set for analysis and prediction.
 - We dropped columns "**doctor**" and "**hospital**" as they were irrelevant for analysis and prediction.
 - The data-set collected has a lot of null values and they have to be handled properly.
 - We replaced all the null values of a particular attribute by the class median.
- **Data Visualisation:** Represent information and data graphically.
 - We used "**seaborn**" library to plot "**heatmap**" and "**pairplot**" for the data-set.
- **Data Analysis:** Extracting useful information from data and making the decision based upon the data analyzed.
 - As glucose levels increase the chances of being diabetic increases.
 - As insulin decreases chances of being diabetic increases.

- **Algorithms:**

- **K-Nearest Algorithm:**

- It identifies data points that are separated into several classes to predict the classification of a sample point.
- It uses Euclidean distance in order to find nearest neighbors. As shown in formula below: $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- It classifies new points based on similarity measures.

- **Support Vector Machine:**

- It is a linear model for classification as well as regression problems.
- It is used to solve both linear as well as non-linear problems.
- Basically, SVM creates a hyper-plane or a line which separates the data into classes.
- SVM are one of the robust prediction algorithms.
- It is useful in outlier detection.

● Logistic Regression:

- It is used when target variable is categorical for example: to predict tumor (if 1 then true else if 0 then false). It is used for predictive analysis.
- It can be used for classification algorithms.
- Logistic Regression uses logistic function to model a binary dependent variable as defined below: $f(x) = L/(1 + e^{-k(x-x_0)})$ where,
 - x_0 is the x value of sigmoid midpoint
 - L is curves maximum learning value.
 - k is the steepness of the curve or the logistic growth rate.

● Decision Tree:

- It is a classifier in the structure of tree.
- Internal nodes of the decision tree represent features of data-set, branch of tree represents decision rules and leaf nodes of the tree represents the outcome.
- It is a graphical representation of all possible solution to a problem based in the conditions given.

● Naive Bayes:

- It is mainly used for solving classification problems.
- It predicts the outcome on the basis of the probability of the data and hence it is called as **Probabilistic Classifier**. Naive stands for strong assumption and it uses Bayes theorem which is given as:

$P(A|B) = P(B|A)P(A)/P(B)$ where,

- $P(A|B)$ is **Posterior Probability**: Probability of hypothesis A on the observed event B.
- $P(B|A)$ is **Likelihood Probability**: Probability of evidence given that probability of hypothesis is true.
- $P(A)$ is **Prior Probability**: Probability of hypothesis before observing the evidence.
- $P(B)$ is **Marginal Probability**: Probability of evidence.

● Random Forest:

- Random forest is a classifier that contains number of decision trees on various subset of given data-set and takes average to improve the accuracy of prediction. The greater number of tree the greater is the accuracy and it prevents overfitting.

● Perceptron:

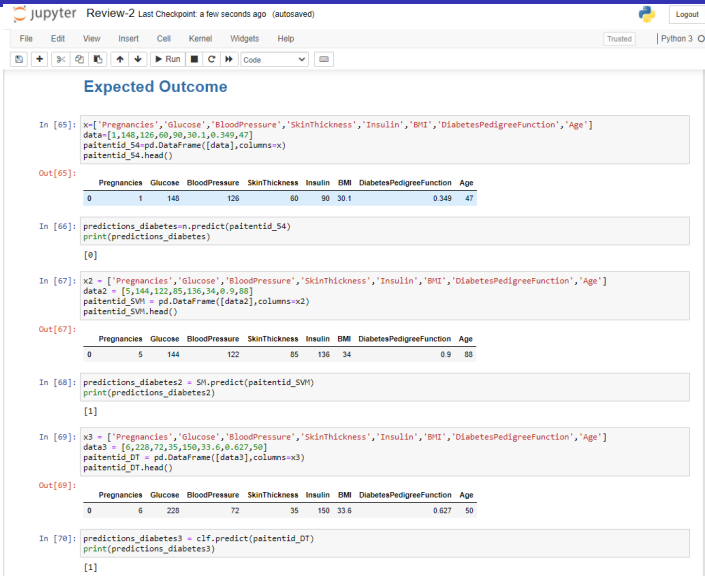
- It is an algorithm used for binary classifiers in Supervised Learning.
- Perceptron is a single layered neural network.
- They consist of four main parts:
 - input values
 - weights
 - bias
 - activation function (It is the function which outputs a small value for small inputs, and a larger value if its inputs exceed a threshold)
 - net sum

● Multi-layer Perceptron:

- It classifies data-sets that are not linearly separable.
- It has the same input and output layers of Perceptron but in addition to that it has multiple hidden layers.
- MLP form the basis for all neural networks and they have greatly improved the results of classification as well as regression problems.

The partial implementation is showed through the Jupyter Notebook IDE.

Expected Outcome



References I



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 III



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.

References IV

-  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.