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

February 13, 2021

Table of Contents

Objective

To analyse and gain insights from the collected data in order to develop a diabetes prediction model using supervised machine learning algorithms which can help the user to predict future risk of diabetes using a user friendly GUI (Graphical User Interface).

Literature Survey I

- Aiswarya, Jeyalatha and Ronak in their paper used Naive Bayes and Decision Tree Algorithms. They got an accuracy of 79.5% from Naive Bayes classifier. They did not use any other Supervised Learning Algorithms like SVM, KNN and Random Forest. Pre-processing and Data-set Transformation was done using WEKA tool and we propose to use Python.[2]
- J Pradeep and Saminathan used J48, KNN and Random Forest Algorithm for prediction and got 73.8% accuracy with J48 algorithm. The authors did not mention the Data pre-processing techniques and methodology followed by them.[3]
- S Nanda, M Savvidou, A Syngelaki, R Akolekar, KH Nicolaides used only Logistic Regression for prediction. They did not mention the accuracy of their machine learning prediction model. [6]

Literature Survey II

- Shekharesh, Sambit, Surajit and Debabrata in their paper used XGBoost and Random Forest Algorithm. The authors did not perform Data Cleaning on the dataset. It is mentioned in their paper that they achieved an accuracy of 74.1% with XGBoost prediction algorithm. [1]
- Yuvaraj and SriPreethaa in their paper used Random Forest Algorithm on Hadoop Cluster. The authors discussed information gain methods used for feature selection but failed to mention the pre-processing steps. [5]
- Francesco, Vittoria and Antonella used HoeffdingTree, JRip, BayesNet, Random Forest. The highest accuracy achieved was 76% by using HoeffdingTree. The authors used WEKA tool for analysis and prediction and did not mention the pre-processing steps. [4]

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.

Architectural Design for Proposed System II

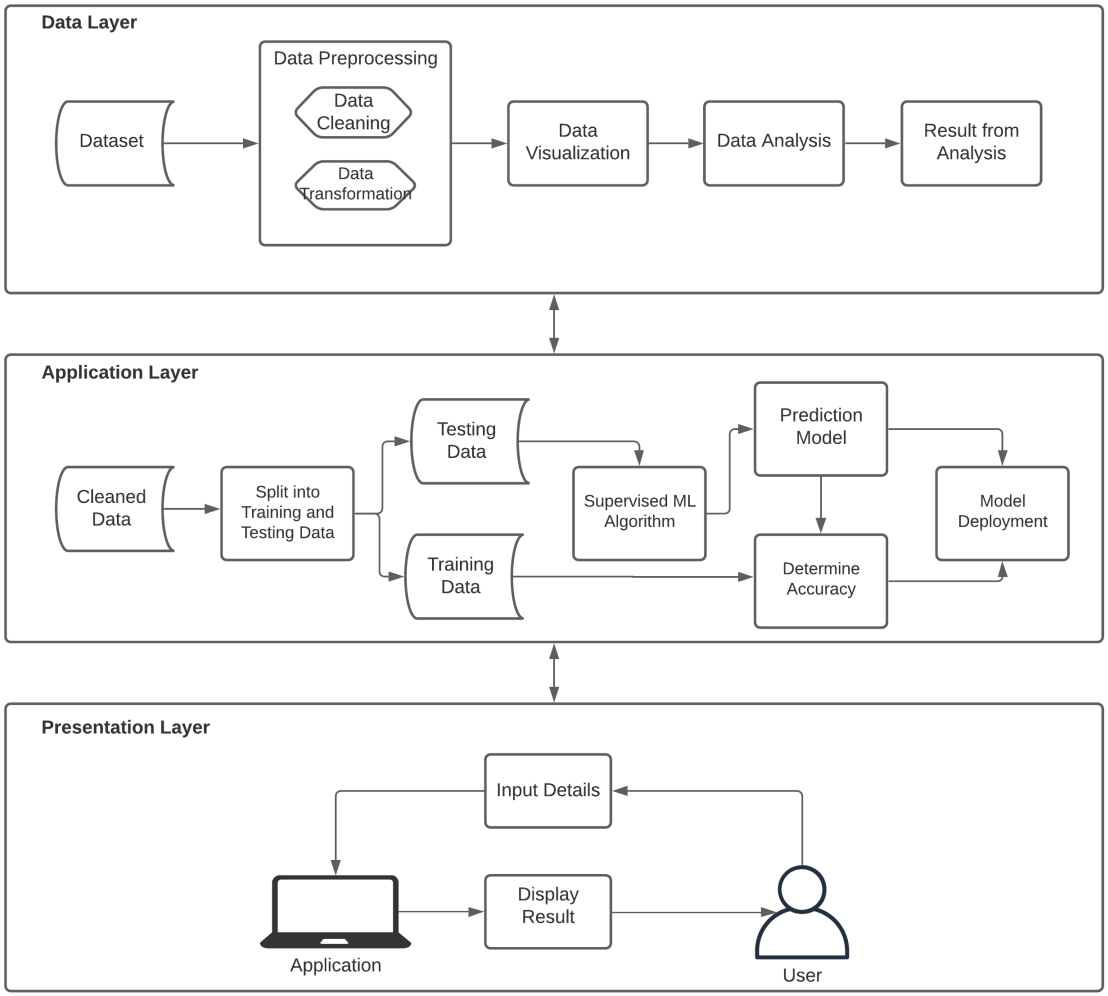

Figure: Architectural Design for Proposed System

Techniques to be used I

- **Data Collection:** Collecting the data-set to be used for data analysis and prediction. We have downloaded the Pima Indians Diabetes Database from data.world provided by National Institute of Diabetes for Digestive and Kidney Diseases.
- **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.
- **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.

Techniques to be used II

- **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.
- **Select Machine Learning Algorithm:** Select the best Supervised Learning Algorithm (K-Nearest Neighbour, Support Vector Machine, Logistic Regression, Naive Bayes, Random Forest and Decision Tree) 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.

References I


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.


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

References II


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.


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.