Diabetes disease prediction using a web application with the help of a machine learning model.

Introduction.
Diabetes is a disease that occurs when your blood glucose, also called blood sugar is too high in the body. Blood glucose is the main source of energy and comes from the food we eat. Insulin made by pancreas helps glucose get into our cells to be used for energy. However when blood glucose is high, people get very sick. Early signs of diabetes include Hunger and fatigue, blurred vision, dry mouth and itchy skin and also peeing more often and being thirstier. If not detected early can lead to death. This is where machine learning in the healthcare comes into help. With the advancement of technology, the better computing power and availability of data-sets on open source repositories have further increased the use of machine learning. Machine learning is used in healthcare in vast areas. The healthcare produces sector produces large amounts of data in terms of images, patient data and so on that helps to identify patterns and make predictions.
Thus, making a machine learning model, training it on the data-set and entering individual patient details can help in prediction. The prediction result will be according to the data entered and hence specific to that individual.

Problem Statement.
The problem this project is trying to address is late diagnosis or detection of Diabetes which tends to lead to high death rates in people and also manual detection and classification of Diabetes disease by health workers.

Main objective.
To develop a web application using a machine learning model to aid medical doctors to detect, predict and classify diabetes disease in human beings.

Specific objectives.
- To collect requirements needed to develop the proposed system.
- To build, test and evaluate the machine learning model which will be used to make predictions.
- To implement a web application which will be used in diabetes disease prediction.
- To study the current existing system used in diabetes disease prediction.

Methodology.
To address the problem identified, the project needs to meet the objectives stated above.
This part describes how I intend to achieve the objectives of the project including the technologies used.

1. The data that will be used.

Source: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Data description: The Pima Indians Diabetes Dataset consists of several medical parameters and one dependent parameter (outcome) of binary values. The data-set is mainly for Female gender of at least 21 years old of Pima Indian heritage and the description of the data-set is as follows;
* 9 columns with 8 independent parameters, 1 outcome parameter with uniquely identified 768 observations having 268 positive for diabetes (1) and 500 negative for diabetes (0).
The 9 columns are the following.
Pregnancies: Number of times pregnant.
Glucose: Oral glucose tolerant test result.
Blood pressure: Diastolic blood pressure values in mmHg
skin thickness: triceps skin fold thickness in mm

Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index.
DiabetesPedigreeFunction : Diabetes Pedigree function.
Age: Age in years.
Outcome: Class 1 indicates person with diabetes, 0 indicates other.

Early Stage Diabetes Prediction.

About Data

This dataset contains the sign and symptpom data of newly diabetic or would be diabetic patient.This has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor.

Features of the dataset: Diabetes_data_upload.csv

The dataset consist of total 15 features and one target variable named class.

1. Age: Age in years ranging from (20years to 65 years)
2. Gender: Male / Female
3. Polyuria: Yes / No
4. Polydipsia: Yes/ No
5. Sudden weight loss: Yes/ No
6. Weakness: Yes/ No
7. Polyphagia: Yes/ No
8. Genital Thrush: Yes/ No
9. Visual blurring: Yes/ No
10. Itching: Yes/ No
11. Irritability: Yes/No
12. Delayed healing: Yes/ No
13. Partial Paresis: Yes/ No
14. Muscle stiffness: yes/ No
15. Alopecia: Yes/ No
16. Obesity: Yes/ No

Class: Positive / Negative

Relevant Papers:

Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques [Web Link] Authors and affiliations M. M. Faniqul IslamEmail Rahatara Ferdousi Sadikur Rahman Humayra Yasmin Bushra

Citation Request:

Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125

Method of collection: Downloading the data-set from the link
https://www.kaggle.com/uciml/pima-indians-diabetes-database

Analysis techniques.
- Taking a peek at the raw data eg reviewing the first 20 rows of data using the head() function.
- Reviewing the dimensions of the data-set eg using shape attribute.
- Reviewing the data types of the attributes in the data eg using dtypes.
- Summarizing the distribution of instances across classes in the data-set eg using the describe() function on the pandas data-frame.
- Summarizing data using descriptive statistics eg separating diabetes patients from non diabetes patients.
- Understanding the relationships in the data using correlations eg using corr() function.
- Reviewing the skew of the distributions of each attribute using the skew() function.
- Data visualizations with the use of Histograms, density plots, box and whisker plots, correlation matrix plot, scatter plot matrix.

2. The Algorithms for analyzing the data and justification for each.

Logistic regression: can model binary classification problems.
Support vector machine (SVM): seeks a line that best separates two classes.
K-nearest neighbors: takes the mean outcome of the neighbors as the prediction.
Naive Bayes: calculates the probability and conditional probability of each class given each input value.
Random forests Classifier.
XGB classifier
Decision tree classifier.

3. The evaluation methods for the algorithms and justification ie why they are applicable.

Confusion matrix.
It gives a matrix as the output and describes the complete performance of the model.

Logarithmic loss.
It works well for multi class classification.

Area under ROC curve.
Its mainly for binary classification problems.

F1 score.
Its used to measure a test's accuracy.
It tries to find a balance between precision and recall. It does well if the positive predicted are actively positives (precision) and does not miss out on positives and predicts them negative (recall).

Mean squared error (MSE).
It provides a good idea of the magnitude of the error.
Its easier to compute the gradient which is not the case for Mean Absolute error since it requires complicated linear programming tools to compute the gradient.

Mean Absolute error (MAE).
It gives an idea of how wrong the predictions were.

Precision.

4. The technologies that will be used to build the system.

Client side.
Front-end: Flask framework.
Back-end: Python programming language.

Server side.
Back-end: Python3.
Working environment: Jupyter notebook
Important libraries: tensorflow, numpy, pandas, seaborn, pickle, matplotlib.