Final Project - Technical Report

Diabetes Disease Prediction Web Application

Application URL:

I have hosted application on **Heroku: Cloud Application Platform**
URL: https://diabetes-disease-predictions.herokuapp.com/

Full GitHub URL:

I have uploaded the application source code on github.
URL: https://github.com/AAbubaker2020/diabetes-predicts

Project Summary

The aim of the project to predicate weather the user is diabetes, pre-diabetes or no-diabetes and help the user to be more aware about their diabetes disease health status. The data used to develop multinomial logistic regression machine learning model was founded on Kaggle and the original source was the CDC Centers for Disease Control and Prevention diabetes _ 012_health_indicators_BRFSS2015.csv is a clean dataset of 253,680 survey responses to CDC's BRFSS22015.

Project Objectives and Usefulness:

The data on the prevalence and incidence of diabetes and prediabetes, risk factors for complications, acute and long-term complications, deaths, and costs of diabetes disease.

These data can help focus efforts to prevent and control diabetes across the United States. By utilizing this data and build a machine learning model to classify the user's information provided to be able help give a predicted status of diabetes disease which it would help make a precaution decisions and necessary steps.

Through this project, I planned to success the following objectives:

- Give the user an accessibility to a source of prediction to one of the most chronic diseases in our live time and aiming it could lead to necessary steps if needed.
- Provide user friendly application to predict the user diabetes disease status.
- Make the user able to get the prediction free of any mistake by applying constrains, range of entry, entry type and mandatory entry for all the fields to match the machine learning model requirements of the data.
- Achieve the highest accuracy of prediction based on provided data.

Technical Description:

Data:

The diabetes _ 012_health_indicators_BRFSS2015.csv dataset(The response to CDC survey) is downloaded from Kaggle and it is originated by the Centers for Disease Control and Prevention (CDC) . This datasetis distributed under the Open Data Commons Public Domain Dedication and License. The main purpose of the data is "The National Diabetes Statistics Report" provides up-to-date information on the prevalence and incidence of diabetes and prediabetes, risk factors for complications, acute and long-term complications, deaths, and costs.

I loaded the csv data file into Pandas data frame and did the following data preprocessing:

1- Making sure no missing data.
2- Checking the features type.
3- Checking the data correlation between the data features.
4- Checking the consistency of the data.
5- Perform features selection to use the best 11 features out of the 22 original features
6- Normalize the data.
7- Split the data to train and test data sets.

Tools:

I have implemented the application with VSC (**V**isual **S**tudio **C**ode) architecturepattern and Python language for server-side programming.

Model:
The model is developed using Flask & Python on source-code editor visual studio code "VSC".

View:
The views are developed Flask render templated with HTML and CSS.

Controller:
The controllers are developed using Flask Web Application routers

Deployment Platform:
Heroku cloud-based PaaS platform which supports build and deploy web application.
Tools Used:
Front End: Flask - HTML – CSS
Back End: Python
Tools: Visual Studio Code, Python, Jupyter Notebook, Pandas , Numpy ,sklearn, seaborn and matplotlib.

User Functionalities:

Application supports following functionalities to predict the diabetes diseases status for the users.

- Give detailed instruction for the web application best practices.
    - What type of entry and what the range of entry.
    - Perform user entry validations.
    - Restrict the type of value of fields
    - Make sure all the fields are mandatory to be entered because of the ML model requirements.
- Model prediction.
    - Give the user the prediction based on the information been provided.
- View visualization about the data.
    - Heatmap representing the data correlations.
    - Histogram plot for each dataset feature been used on the model

Evaluation:

It was an individual project which is give me the opportunity to through every step of the machine learning web application life cycle started working with data and make sure that data is representing the objective on hand, going through the data consistency, integrity and no missing or outlier data, splitting the data to train & test datasets and lastly normalized which is the steps of data processing part of the project and then going the model selection to be able to classify the user information using multinomial logistic regression machine and tuning the parameters to achieve the highest accuracy possible.
And finally going to the web application developing working with Flask, HTML and CSS to deploy the ML model through Heroku.
It was a great challenge and experience for me in the whole cycle of the machine learning web application project.

This is an individual project by:

Abubaker Ahmed
M.S. in Data Science
Subject: Applied data science