ML Application In Healthy Living Partner Project Proposal

Machine Learning Engineering Project

Francis Odo

STATUS: First DRAFT

Background

Costs of healthcare continue to rise at a significantly high rate in the United States with no limit in sight. More troubling is the cost of diagnosing chronic diseases. Healthcare providers, patients and insurance service providers are faced with this very challenging problem.

Healthcare Service Providers have expressed interests in finding solutions that are based on preventative approaches to help patients through increased and consistent physical activities that will achieve healthy habits and establish a long-term healthy lifestyle. The payback can be described two folds, improved patient health condition and reduced cost of diagnostics and treatment. Potentially, it could be a win-win situation for patients, healthcare service providers and pharmaceutical industries and the insurance carriers.

Overall, Healthcare Service Providers want to decrease spending on chronic diseases such as type 2 diabetes and related conditions.

Problem Statement

Chronic disease such as Type-2 Diabetes is affecting quality of lives significantly across age categories. Research shows that this disease often affects major arteries in humans. It causes damage to large and small blood vessels. Most often, problems extend to deterioration of the kidney and vision performance.

Due to the nature of the disease and how widely spread in the society, the cost of diagnosis and treatment is increasingly high. Healthcare and pharmaceutical industries are faced with uncontrollable costs, projecting 26% increase over a five year period.

The most viable and cost effective preventative approach to reducing and possibly eliminating this staggering cost problem is through controlled healthy habits.

This can be achieved with a Machine Learning binary classifier algorithm such as the Logistic Regression model and Neural Network (TensorFlow/Keras) model. The algorithm works as the decision support core component of the application.

There are a number of research and articles in this area of ML/AI in Healthcare particularly for identifying chronic illness.

Healthcare providers can save significantly on costs to diagnose and treat less severe cases.

Datasets and Inputs

Ideally, users of the app will consistently generate data over a period of time if committed to using the app and following the guidelines and recommendation offered. The data is then used as input to train a supervised Machine Learning model, which can be deployed and to make predictions with respect to health status classification.

Since this proposal doesn't have its own generated data, I am using a sample data obtained from KAGGLE for this application. The data is CSV formatted (diabetes.csv). The difference will be that the data generated by the user reflects personal activity and performance within the app itself, compared to a general sample with similar set of features.

Data Exploration

The nature and characteristics of the dataset consists of 768 data points and 9 features. We are only interested in features that are contributing factors to the development of Diabetes disease that the Machine Learning Algorithms can train on. Among them are:

- Glucose - Measures the type of sugar you get from the food you eat
- BloodPressure - The pressure of the blood within the Arteries
- Insulin - Scores the regulation of glucose level
- DiabetesPedigreeFunction - Measures likelihood of diabetes based on family history
- BMI - Blood Mass Index
- Outcome - Indicates Diabetic with (1) and Non-Diabetic with (0)

The source of data is the user and user activities. As the user participates, data is created and generated daily across the app and warehoused. The goal is to operate with each of the features at a reasonable level that will enhance healthy living.

This dataset serves as input to the Machine Learning component of the solution. It is important to note that this data is taken from pregnant women only. The dataset does not represent the general population. It is meant to demonstrate the application concept in Healthy Living Partner.

Solution Statement

This project is mainly focusing on the application of Machine Learning techniques and methodologies using multiple algorithms in the proposed product solution. The objective is not to develop the entire product.

Healthy Living Partner is a web-based mobile application that assists users to form a habit and live a healthy life through monitored and controlled practice, thereby preventing chronic diseases.

The system collects user data with similar features as shown below. With the application of Machine Learning algorithms the Logistic Regression model classifies the user data as diabetic or non-diabetic. The supervised Machine Learning model trains and predicts based on the trend of the user data collected over a period of time. It is the decision-based component of the solution.

Neural Network model applies Tensorflow/Keras algorithm for its accuracy and layered optimization techniques with adjustable model hyper-parameters.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Imbalance in dataset

In order to address the imbalance in the dataset, the function value_count() is used to obtain the class distribution(0:500,1:268). Oversampling/Undersampling and feature selection techniques could be used to mitigate the imbalance in the dataset classes, along with the evaluation metrics(Confusion Matrix, Precision, Recall, F1-Score).

Product Data Collection Features

1. Health Status Check (Health Status)

   Provides analytical report of user health condition
2. Glucose/Sugar-Level Monitor (Glucose Monitor)

   Provides the user with measures of Blood Glucose level
3. Performance Monitor (Performance Monitor)

   Provides the user with an analysis and percentage rating of completion of tasks in a

routine.

4. Application Programming Interface (API)

Benchmark Model

- Comparison of different classifier performance in Training/Testing
- Accuracy of prediction 70% or higher
- Loss

Evaluation Metrics

Evaluation of Machine Learning performance will depend on:

1. Confusion Matrix
2. Precision
3. F1-Score
4. Recall

Project Design

As the system collects more user data, it trains itself and continues to track user activities and reveal trends and decisions using Machine Learning classification algorithms.

Data Pre-Processing

Depending on the quality of the available data, preparation of data requires some critical analytical steps. Data pre-processing includes:

a) Inspecting data features, filling nulls, dropping NaNs,Skin Thickness and the Target column.
b) Normalize if we need to change values to numeric(only if necessary)
c) Get the data points and standardize with standardscaler() i.e remove the mean and scale according to standard deviation
d) Obtain dataset class distribution and balance the dataset
d) Create visualization to see the distribution's graphical representation
d) Split into training and testing

The Machine Learning process involves:

1. Define the model
2. Train the model
3. Evaluate the model

Algorithms
Logistic Regression and Neural Network
Hyperparameter Tuning - Number of Hidden Layers, Size of Layer and choice of Activation Function

Ensemble and XGBoost classifiers.

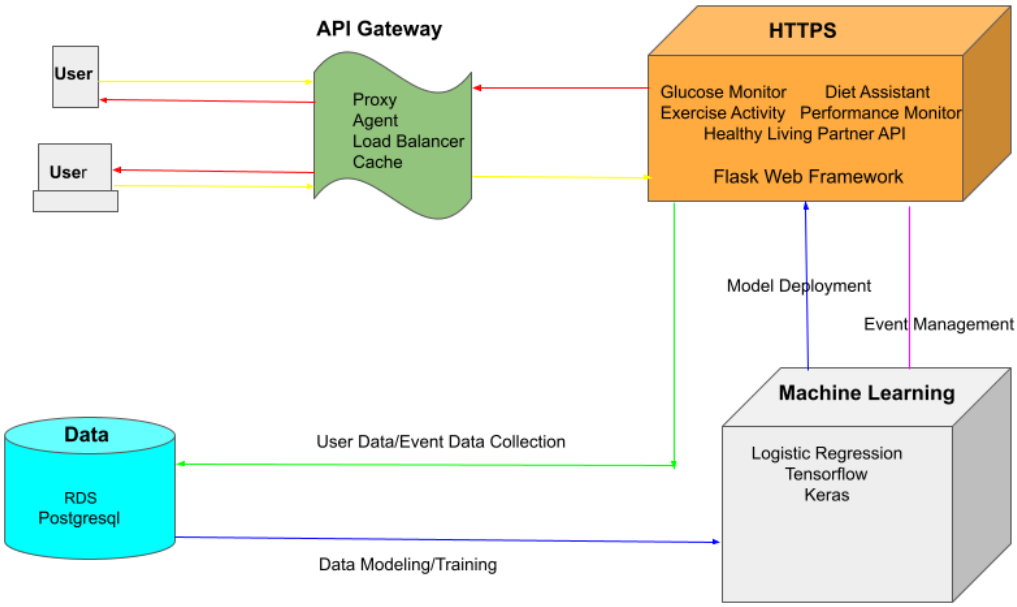

Figure - Diagram showing Healthy Living Partner System with Machine Learning Component

References

1. Accuracy
https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b
2. Pima Indians Diabetic Data
https://www.kaggle.com/uciml/pima-indians-diabetes-database
3. Economic Costs of Diabetes in the US 2017
https://care.diabetesjournals.org/content/early/2018/03/20/dci18-0007
4. Healthcare statistics for 2021
https://policyadvice.net/insurance/insights/healthcare-statistics
5. Complications of Type1 Diabetes
https://jdrf.org.uk/information-support/about-type-1-diabetes/complications/
6. Predictive Models for Diabetes Mellitus Using machine Learning
https://doi.org/10.1186/s12902-019-0436-6
7. Balanced and Imbalanced Dataset
https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5#:~:text=Balance%20Dataset,positive%20values%20and%20negative%20values.