Prediction of health disease based on day-to-day life activity using Machine Learning Approach

Gouse Baig Mohammad
Vardhaman College of Engineering
Srihari K.S (srihariksresearch@gmail.com)
SNSCT: SNS College of Technology
S. Shitharth
Vardhaman College of Engineering
Vijaya Kumar Reddy Radha
Pradad V Potluri Siddhartha Institute Of Technology
Buraga Srinivasa Rao
Lakireddy Balireddy College of Engineering
Puranam Revanth Kumar
Icfaitech (faculty of science and technology)

Research Article

Keywords: Healthcare data analysis, Machine learning in healthcare, Data analytics, Health status estimation

Posted Date: April 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-391346/v1

Abstract

In the current generation, it is very important to monitor our health. With the busy lives of people nowadays, many are experiencing health-related issues at an early age. Many of these issues arise because of our daily life activities. People are interested in many activities, but they hardly know the consequences of those activities. Hence it is very important to detect daily life activities that affect the health of a person and predict the diseases that may come in the future. However, there are existing methods for predicting a particular kind of disease like diabetes, tuberculosis, etc., based on electronic health records. The proposed system predicts the overall health status of a person using machine learning techniques. The overall health status includes how well a person is sleeping, eating, doing physical activity, etc. Also, the prosed system monitors the health of persons and alerts when they are deviating from a normal state. In this chapter, we will discuss the data collection approach, architecture of the system, overall health estimation models, implementation details, and the analysis of the result.

Introduction

1.1 Health Status of an Individual

The overall health status of a person is assessed by comparing the level of wellness with the level of illness. The health status can be estimated through many parameters. Some of the parameters are (i) Sleep status: the health level of a person is depending on his/her sleep timings, (ii) Screen status: the health level of a person is depending on the amount of time spent on screen, (iii) Drink status: the health level of a person is depending on his/her drinking habits, (iv) Smoke status: the health level of a person is depending on his/her smoking activities (v) Calories status: the health level of a person is depending on the calories consumed and physical activities.

1.2 Activities and Measures of an Individual

The things that an individual does daily can be referred to as activities. Some of the activities include sleeping, watching television, consuming alcohol, smoking cigarette, listening to the radio, reading books, etc. Measures of an Individual include physical measures like height, weight, and some other measures like age, gender, etc. Basically, many of the measures are permanent they will not change frequently, whereas the activities might change frequently.

1.3 Traditional Approach to predict health status

In general, health status can be predicted by consultancy experts. If an individual wants to know about their sleep status (i.e. whether their sleep pattern is good? And whether they are taking the adequate amount of sleep?), they can consult an expert at sleep centers. If an individual wants to know about their calorie status (i.e. How much calories they need to consume to maintain/ increase/ decrease weight? How much exercise they need to do to maintain the calories in balance?), they can consult physicians.

But what the experts do, they give some suggestions by considering the measures and activities mentioned previously. For this, the experts use some rules and conditions on the measures and activities. For example, ‘A boy of age 21, height 176cm, weight 63kg with a less physical activity needs to consume 1950 calories per day to maintain weight’. But the limitation of this approach is not considered some of the important parameters like (i) Different Health Parameters (Sleep Status, Calorie Status, etc.) have different consultants. (ii) They may not be very accurate in predicting them manually without any calculations. (iii) Consulting experts might be costly for a low-middle and middle-class family.

Thus, it demands the need for designing a model that can predict their health status from daily life activities.

Related Study

It is important for everyone to understand their health status, it helps to avoid future diseases. As mentioned previously some of the parameters of the health status are sleep status, smoke status, drink status, disease status, etc. Directly or indirectly they depend on the individual’s daily life activities and physical measures. In healthcare data management, a huge amount of structured or unstructured data related to the patient is generated from the diagnostic reports, doctor's prescription, and the wearable devices. In recent years the healthcare data analysis and estimating the future health status are the major focused domains in healthcare. Disease Prediction has a major impact on healthcare analytics as it predicts outbreaks of epidemics to avoidable diseases and improves the quality of life. Some of the recent works proposed a verity of models to predict health status a person with the help of various factors. Researchers in (P. K. Sahoo, S. K. Mohapatra, & S.-L. Wu, 2016), the authors proposed a cloud-based probabilistic data acquisition method and also, designed an approach to predict the impending health state of a person based on the current health status. A work by Hirshkowitz M, et al.(2015), proposed a method to evaluate and recommended sleep duration for individuals based on their age categories. Researchers in (Min Chen, YixueHao, Kai Hwang, Lu Wang, & Lin Wang, 2017), proposed a new approach for the disease risk prediction, in that they also proposed the Convolutional Neural Network (CNN) based on unimodal disease risk prediction and CNN-based Multimodal Disease Risk Prediction. Authors in (C.-H. Weng, T. C.-K. Huang, & R.-P. Han, 2016), discussed different types of artificial neural network (ANN) techniques for disease prediction and evaluated all the methods based on statistical tests. Researchers in (L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, & T. S. Chua, 2015), the authors proposed a system to collect health data through some questionnaires and analyzed using deep learning architectures.

A work by Tayeb S et al. (2017), proposed a method based on the popular machine learning algorithm KNN to predict heart disease and chronic kidney failure. Researchers in(Hung Chen-Ying, Chen Wei-Chen,Lai Po-Tsun, Lin Ching-Heng& Lee Chi-Chun, 2017), Authors proposed an automated system for the prediction of stroke based on Electronic Medical Claims (EMCs), and they compared the Deep Neural Network (DNN) with the gradient boosting decision tree (GBDT), logistic regression (LR) and support vector machine (SVM) approach. Researchers in(M. Chen, Y. Ma, J. Song, C. Lai & B. Hu, 2016) authors

proposed the cloud base smart clothing system for sustainable monitoring of human health. They also discussed the technologies and the implementation of methodologies. Authors in (Schmidt S. C. E., Tittlbach S., Bös K., &Woll A, 2017) analyzed varieties of physical activity, fitness and health, they considered 18 years duration for study and identified interesting insights. In a recent work on Analyzing University Fitness Center data (Yunshu Du, Assefaw H. Gebremedhin, 2019 & Matthew E. Taylor, 2019)), the user's fitness activity data is collected to predict the crowd at the fitness center. But the fitness activity data can be used to predict more than that.

A lot of research was done on measuring health parameters numerically. Also, there are many works on calculating some health parameters from other parameters. A work by Harris-Benedict (Harris JA & Benedict FG, 1918) calculates Basal Metabolic Rate from an individual's physical measures. It is used to estimate the number of calories needed for an individual to maintain good health. Our work incorporated the effect of daily life activities on health status. But that data can be used to personalize health predictions and suggestions. This motivated to design a model that predicts health status from the daily life activities of individuals.

Problem Statement

Let A_t be the set of daily life activities done by an individual t day's back. Thus, A_0 is the set of activities done by an individual today, A_1 be the set of activities done by an individual yesterday, and so on. A is the collection of the activities of an individual for many days. M be the set of physical measures of an individual. H be the health status matrix.

Definition 2.1: Health Status Matrix: A health status matrix M describes the outcome of various parameters of health status. Each row of the matrix is considered as a vector of possible outcomes of the respective parameter of the health status. Examples of health status parameters are sleep status, smoke status, drink status, etc.

Given a set of daily life activities and physical measures of users over a few days and their health status. The health status of a set of users already defined, known as labeled users U_L. Whereas the health status of other sets of users is not defined, known as unlabeled users U_V. The aim of the proposed model is to learn a function that uses the information of the labeled users' U_L and find the health status of the unlabeled users U_V.

Given a series of activities from last t days, the objective is to learn a function F,

$H = F(M, A_0, A_1, A_2, \ldots. A_t)$,

where$M$ is the set of physical measures of a user A_t is the set of activities of the user t days back. H is a health status matrix

Proposed Architecture

The figure 2.1 describes the architecture of the proposed model. The set of daily life activities and physical measures of an individual is taken from the users and fed into a pre-processor phase, which processes the input by reducing the number of features and does the required data pre-processing operations.

4.1 Pre-processing:

The daily life activities of an individual that are mainly considered are screen time, sleep time, physical activity, number of cigarettes smoked, units of alcohol consumed. The measures that are mainly considered are age, gender, height, weight, calorie intake. Thus, there are ten features that are collected from an individual. Then, in the pre-processing step, the number of features is reduced by removing the activities and measures that do not have any direct effect on health status. This is achieved by using the Harris-Benedict Equation (Harris JA & Benedict FG,1918).

The Harris-Benedict Equation (Harris JA & Benedict FG, 1918) is a method used to estimate an individual's basal metabolic rate (BMR). It says

For Men	BMR =(10 × Weight in kg) + (6.25 × Height in cm) - (5 × Age in years) + 5
For Women	BMR = (10 × Weight in kg) + (6.25 × Height in cm) - (5 × Age in years) - 161

As per the Harris-Benedict Equation (Harris JA & Benedict FG, 1918), the calories to be consumed is depending on the BMR value and the physical activity.

Calories to be consumed = BMR * Physical Activity

Calorie Difference = (Calories Consumed) - (Calories to be consumed)

In the proposed method the number of features is reduced to seven. They are age, gender, sleep time, screen time, number of cigarettes, units of alcohol consumed, and calorie intake.

4.2 Phase-I:

The Phase-I of the model, process the data received from both the data sources and the user. In this phase, a decision tree classifier is used to estimate the health parameter of the user. Initially, the model Is trained with the dataset received from the data sources. The Phase-I of the model estimates the health status of an individual for a particular day. But an individual’s health status can’t be accurate just by considering one day's output. In Phase-I the decision tree classifier is used, it takes the activities of an individual as input and produces the status of the health parameters for one day. Thus, the output of Phase-I is collected over a week and feeds it to Phase-II.

4.2 Phase-II:

The Phase-II of the model, process the data received from the data sources and the output of the Phase-I. In this phase, the decision tree classifier is used to estimate the health parameter of the user. Initially, the model is trained with the dataset received from the data sources. The Phase-II of the model estimates the health status of an individual for a week. The output of Phase-II estimates the health status and generates the alerts and suggestions that are to be notified to the individual. In Phase-II the decision tree classifier is used, it takes the daily status of the health parameters over a week as input (i.e. the output of Phase-I) and outputs alerts & predictions of that health parameter.

4.3 Dataset Generation

Below sub-section provides the details of the rules collection and the dataset generation. The generated dataset is used for training the model proposed in the previous section.

4.3.1 Rules Collection

For preparing the datasets a proper set of rules is required on how the daily life activities of an individual affect his health status. From different trusted sources (Hirshkowitz M, et al., 2015, Bjartveit K, &Tverdal A, 2005) the rules are collected. Based on the activities and measures of an individual, these rules give the overall health status of an individual. For example, the recommended sleep time for the person aged between 6 to 13 years is 9 to 11 hours. if the sleep time is between 7 to 8, it is a little less than normal. if the sleep time is between 11 to 12, it is a little more than normal. if the sleep time is more than 12 or less than 8, then it affects health.

4.3.2 Feature Selection

Selecting the features from the rules that are collected and these rules depend on some activities and measures of an individual. For example, alcohol consumption rules for females are different from males. Similarly, the calorie value recommended for a person of 100kg is different than that of a person of 50kg(Bjartveit K, &Tverdal A, 2005). In these examples, gender and weight are the features that are selected. In a similar fashion all the features like age, gender, height, weight, calorie intake, units smoked, units drunk, physical activity, screen time and sleep time were collected

4.3.3 Feature Reduction

Although the features were collected, some of them might not affect the health status of a person directly. Thus, the collected features need to be transformed into the actual features which affect the health status. Here, the Harris-Benedict equation is used to reduce the features. The Harris-Benedict equation (Harris JA & Benedict FG, 1918) is a method used to estimate an individual's basal metabolic rate (BMR). It says that the calories to be consumed depends on the BMR value and physical activity.

For example, If the physical activity is sedentary or a little active, then the calories to be consumed is 1.2*BMR. If the physical activity is lightly active, then the calories to be consumed is 1.375*BMR. If physical activity is moderate, then the calories to be consumed is 1.55*BMR. If physical activity is an

intense exercise, then the calories to be consumed is 1.725*BMR. If physical activity is an extra hard exercise, then the calories to be consumed is 1.9*BMR.

Calorie Difference = (Calories Consumed) - (Calories to be consumed) (1.1)

Thus, the total number of inputs is reduced to seven. They are Age, Gender, Number of units smoked, Units of Alcohol Consumed, Screen Time, Sleep Time, Calorie Difference.

4.3.4 Dataset Generation from Rules

Based on the rules discussed in section 2.4.3.1, all the required features are extracted. The features include daily life activities and physical measures of an individual. From the features extracted, the number of features is reduced using some standard techniques as discussed (Harris JA & Benedict FG, 1918).

There are two phases in the proposed system. Thus, the Phase-I needs one dataset and the Phase-II needs a different dataset with class labels. The example dataset is described in Table 2.1

Table 2.1: Sample Dataset for Phase-I

Class	Condition	Class label	Description
sleep			
0	for age less than 2 sleep value between 11 and 14 For age between 3-5 sleep value between 10 and 13 For age between 6-13 sleep value between 9 and 11 For age between 14-17 sleep value between 8 and 10 For age between 18-25 sleep value between 7 and 9 For age between 26-64 sleep value between 7 and 9 For age greater than 65 sleep value between 7 and 8	normal	It tells the optimal sleep value for different age groups
1	for age less than 2 sleep value between 9 and 10 For age between 3-5 sleep value between 8 and 9 For age between 6-13 sleep value between 7 and 8 For age between 14-17 sleep value between 7 and 8 For age between 18-25 sleep value between 6 and 7 For age between 26-64 sleep value between 6 and 7 For age greater than 65 sleep value between 5 and 6	less sleep	It tells the sleep value is less than the optimal value for different age groups
2	for age less than 2 sleep value between 15 and 16 For age between 3-5 sleep value between 13 and 14 For age between 6-13 sleep value between 11 and 12 For age between 14-17 sleep value between 10 and 11	more sleep	It tells the sleep value is more than the optimal value for different age groups

	For age between 18-25 sleep value between 9 and 10 For age between 26-64 sleep value between 9 and 10 For age greater than 65 sleep value between 8 and 9	
Smoke		
0	if the number of cigars smoked is 0	good smoke status
1	if the number of cigars smoked is between 1 and 4	smoking status is reasonable
2	if the number of cigars smoked is between 5 and 15	bad smoking status
3	if the number of cigars smoked is more than 15	dangerous smoking status
Drink		
0	if the number of units consumed is 0	drinking status is good
1	if gender is male and the number of units consumed is less than 2 If gender is female and the number of units consumed is less than 1	drinking status is reasonable
2	if gender is male and the number of units consumed is between 3 and 4 If gender is female and the number of units consumed is less than 2 and 3	drinking status is bad

4.3.5 Example

Let the individual's activities and measures for a day are:

Input = (Age=21) ∩ (Gender=Male) ∩ (No. of cigars smoked=0) ∩ (Units of Alcohol Consumed=2) ∩ (Screen Time=6) ∩ (Sleep Time=8) ∩ (Height=176) ∩ (Weight=63) ∩ (Calorie Intake=1800) ∩ (Physical Activity=Lightly Active)

5.1 Pre-processing:

BMR = (10 × Weight in kg) + (6.25 × Height in cm) - (5 × Age in years) + 5 -------(from Harris JA & Benedict FG,1918) BMR = 10 × 63 + 6.25 × 176 – 5 × 21 + 5 = 1630

Calories needs to be consumed = BMR × Physical Activity = 1630 × 1.375 = 2241.25

Calorie Difference = Calories consumed – Calories needs to be consumed= 1800 – 2241.25 = -441.25

Thus, inputs after pre-processing are:

Input1 = (Age=21) ∩ (Gender=Male) ∩ (No. of cigars smoked=0) ∩ (Units of Alcohol Consumed=2) ∩ (Screen Time=6) ∩ (Sleep Time=8) ∩ (Calorie Difference=-441.25)

Experimental Results

We have developed two models in this chapter based on the two popular machine learning algorithms are Decision tree and Random and tested both the models based on the synthetic dataset. We have developed a web-based application to demonstrate the models proposed in this chapter, below are a few screenshots of the application.

5.1 Performance Metrics

To analyze the effectiveness and the performance of the model proposed in this chapter, we used the standard performance metrics (Tom. M. Mitchell, 1997 &EthemAlpaydın, 2010) accuracy, precision, recall, and F1-score.

5.1.1 Accuracy: The accuracy of the model is calculated using the equation given below.

Table 2.2 shows the accuracy of the model for the decision tree proposed in this chapter.

Table 2.2: Accuracy of the model

Health Status	Model 1	Model 2		
Accuracy: Phase-I	Accuracy: Phase-II	Accuracy: Phase-I	Accuracy: Phase-II	
Sleep	90.54	93.64	91.54	94.64
Smoke	92.21	94.01	94.21	96.01
Drink	94.63	95.99	96.63	97.99
Screen	93.11	94.76	94.11	95.76
Calories	94.00	97.83	95.00	98.83

Figure 2.3 shows the accuracy comparison between the two models which are proposed in this chapter and it is observed the model-II gives more accuracy than the model-I.

5.2.2 Precision: The precision of the model is calculated using the equation given below.

See formula 1 in the supplementary files.

Table 2.3: Precision of the model

Health Status	Model 1	Model 2		
Precision: Phase-I	Precision: Phase-II	Precision: Phase-I	Precision: Phase-II	
Sleep	95.55	97.82	95.60	97.87
Smoke	95.69	95.83	95.83	97.89
Drink	95.55	97.87	97.89	98.94
Screen	96.70	97.84	97.82	96.84
Calories	97.34	97.93	97.84	98.96

Figure 2.4 shows the precision comparison between the two models which are proposed in this chapter and it is observed the model-II gives more accuracy than the model-I.

5.2.3 Recall: The recall of the model is calculated using the equation given below.

See formula 2 in the supplementary files.

Table 2.4: Recall of the model

Health Status	Model 1	Model 2		
Recall: Phase-I	Recall: Phase-II	Recall: Phase-I	Recall: Phase-II	
Sleep	93.47	94.73	94.56	95.83
Smoke	95.69	97.87	97.87	97.89
Drink	93.47	96.84	97.89	98.94
Screen	95.65	95.78	95.74	97.87
Calories	95.78	98.95	96.80	98.96

Figure 2.5 shows the Recall comparison between the two models which are proposed in this chapter and it is observed the model-II gives more accuracy than the model-I.

5.2.4 F1-Score: The F1-score is the harmonic mean of precision and recall. Below equation used to calculate the F1-score.

See formula 3 in the supplementary files.

Table 2.5: F1-score of the model

Health Status	Model 1	Model 2		
F1-score: Phase-I	F1-score: Phase-II	F1-score: Phase-I	F1-score: Phase-II	
Sleep	94.50	96.25	95.08	96.84
Smoke	95.69	96.84	96.84	97.89
Drink	94.50	97.35	97.89	98.94
Screen	96.17	96.80	96.77	97.35
Calories	96.80	98.44	97.32	98.96

Figure 2.6 shows the F1-score comparison between the two models which are proposed in this chapter and it is observed the model-II gives more accuracy than the model-I.

Conclusion

In this chapter, we have proposed an architecture based on machine learning algorithms. Basically, we focus on a challenging problem of predicting the overallhealth status of an individual based on their daily life activities and measures. The proposed system predicts the overall health status of a person and future diseases using machine learning techniques. To demonstrate the proposed model, we have created a web-basedapplication.The proposed model helps the user to understand their health status by submitting their details. For training and testing we used the synthetic data, in the future we need to test the proposed model using the real data by collecting from the users. In this work, we attempted a general healthcare problem and a lot more has to be done in the future.The future work is to predict the diseases based on the overall health status estimation using the models proposed in this chapter.

Declarations

Conflict of interest:

There is no conflict of interest.

Funding information:

There is no funding information.

Availability of data and material:

There is no availability of data and material.

Code availability:

There is no code availability.

Author's contribution:

There is no author's contribution.

References

Bjartveit K,&Tverdal A. (2005). Health consequences of smoking 1-4 cigarettes per day. Tobacco Control, 14(5), 315–320.

C.-H. Weng, T. C.-K. Huang,&R.-P. Han. (2016). Disease prediction with different types of neural network classifiers. Telematics and Informatics, 33(2), 277–292.

EthemAlpaydın. (2010). IntroductiontoMachineLearning", 2nd edition.

Harris JA & Benedict FG. (1918). A Biometric Study of Human Basal Metabolism. Proceedings of the National Academy of Sciences of the United States of America, 4(12), 370–373.

Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L … Hillard P J A. (2015). National Sleep Foundation's sleep time duration recommendations: methodology and results summary, Sleep Health, 1(1), 40-43.

Hung Chen-Ying, Chen Wei-Chen,Lai Po-Tsun, Lin Ching-Heng& Lee Chi-Chun. (2017) Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database, 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 3110-3113.

Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, & T. S. Chua. (2015). Disease inference from health-related questions via sparse deep learning, IEEE Transactions on Knowledge and Data Engineering, 27(8), 2107–2119.

Chen, Y. Ma, J. Song, C. Lai & B. Hu. (2016). Smart clothing: Connecting human with clouds and big data for sustainable health monitoring, ACM/SpringerMobile Networks and Applications Mobile, 21(5), 825-845.

Min Chen, YixueHao, Kai Hwang, Lu Wang, & Lin Wang. (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities", IEEE Access, 5, 8869-8879.

K. Sahoo, S. K. Mohapatra, & S.-L. Wu (2016). Analyzing healthcare big data with prediction for future health condition, IEEE Access, 4, 9786-9799.

Schmidt S. C. E., Tittlbach S., Bös K., &Woll A. (2017). Different Types of Physical Activity and Fitness and Health in Adults: An 18-Year Longitudinal Study. BioMed Research International.

Tayeb S, MatinPirouz, Johann Sun, Kaylee Hall, Andrew Chang, Jessica Li, … ShahramLatifi. (2017). Toward Predicting Medical Conditions Using kNearest Neighbors, IEEE International Conference on Big Data, 3897-3903.

Tom. M. Mitchell. (1997). Machine Learning, McGraw Hill International Edition.

Yunshu Du, Assefaw H. Gebremedhin, & Matthew E. Taylor (2019). Analysis of university fitness center data uncovers interesting patterns, Enables prediction. IEEE transactions on knowledge and data engineering, 31(8), 1478 – 1490.

Figures

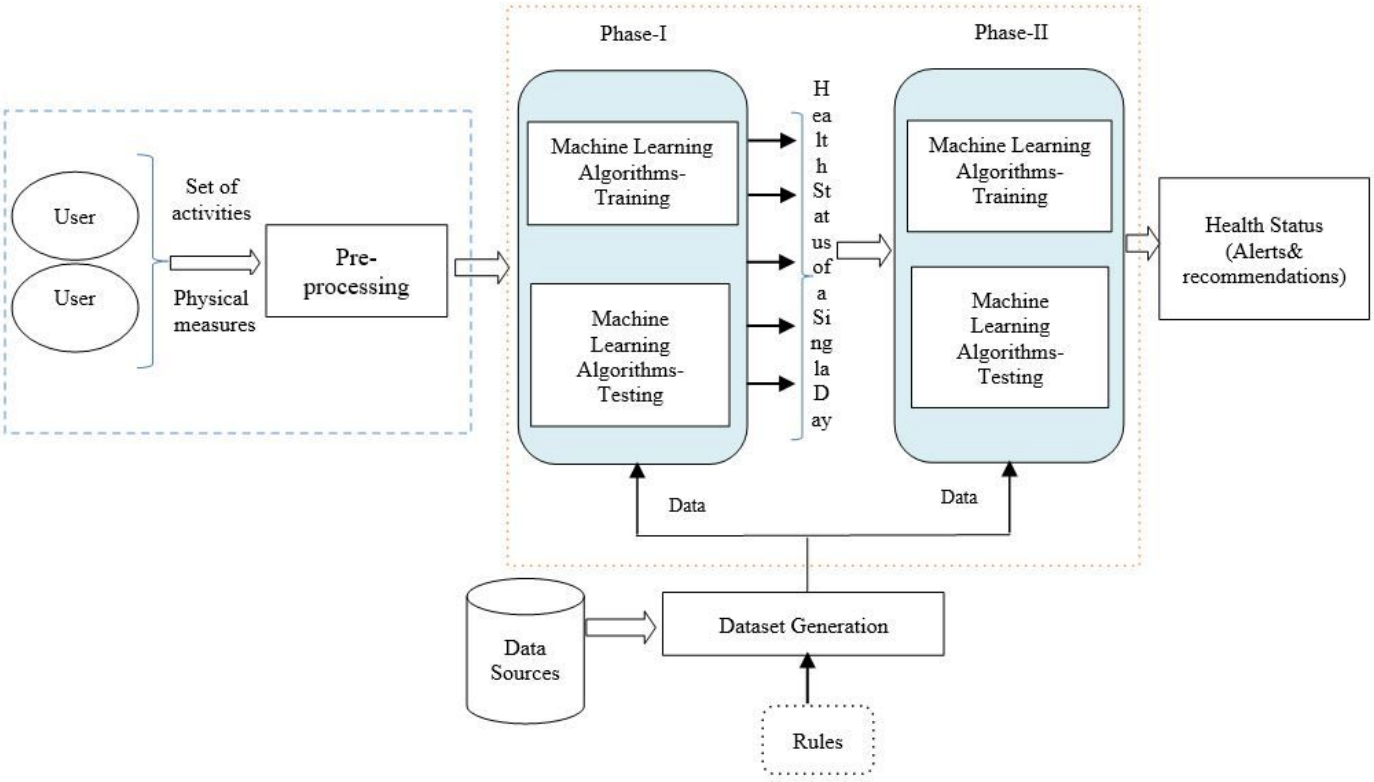

Figure 1

Architecture of the Model

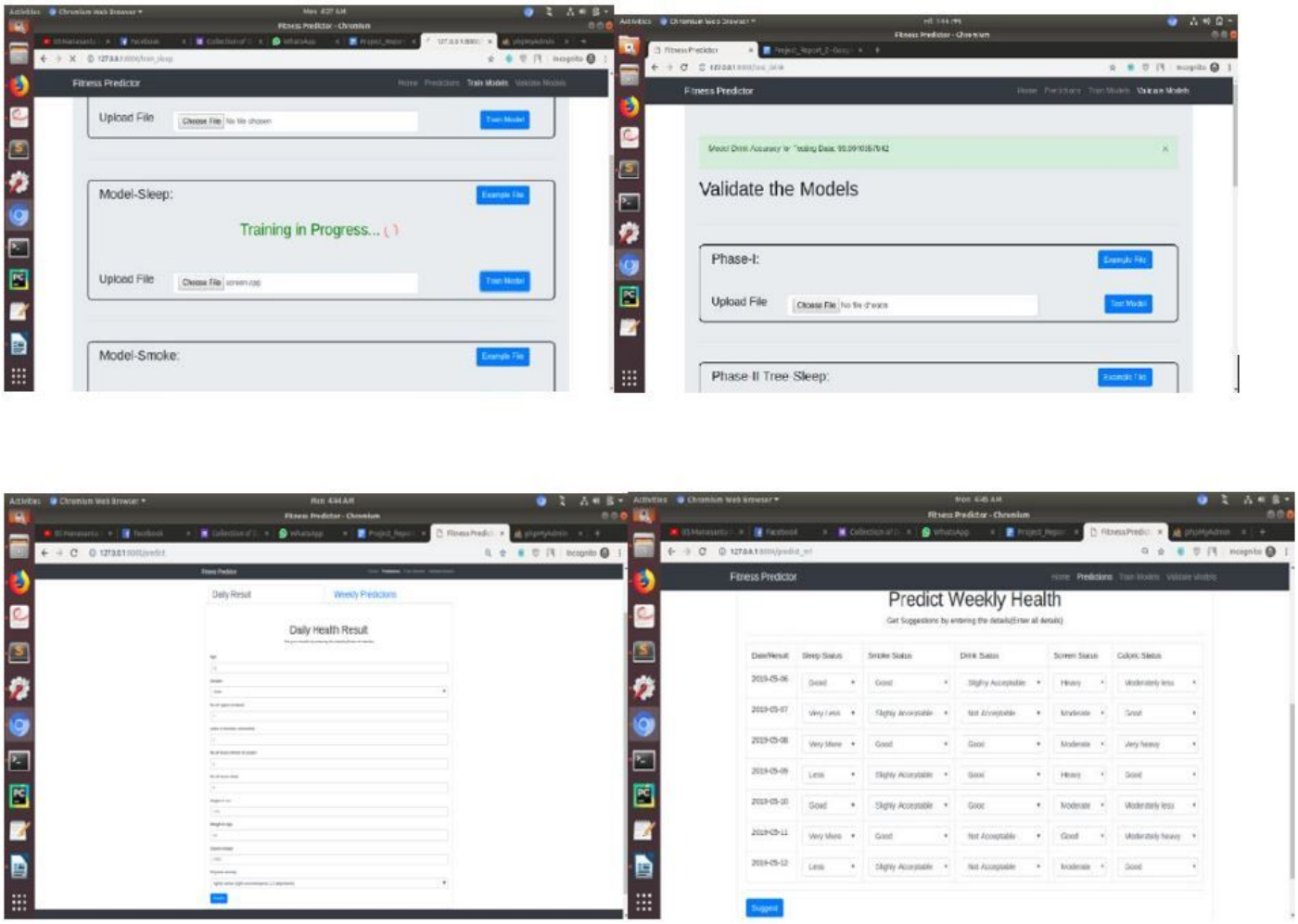

Figure 2

Screenshots of the web application

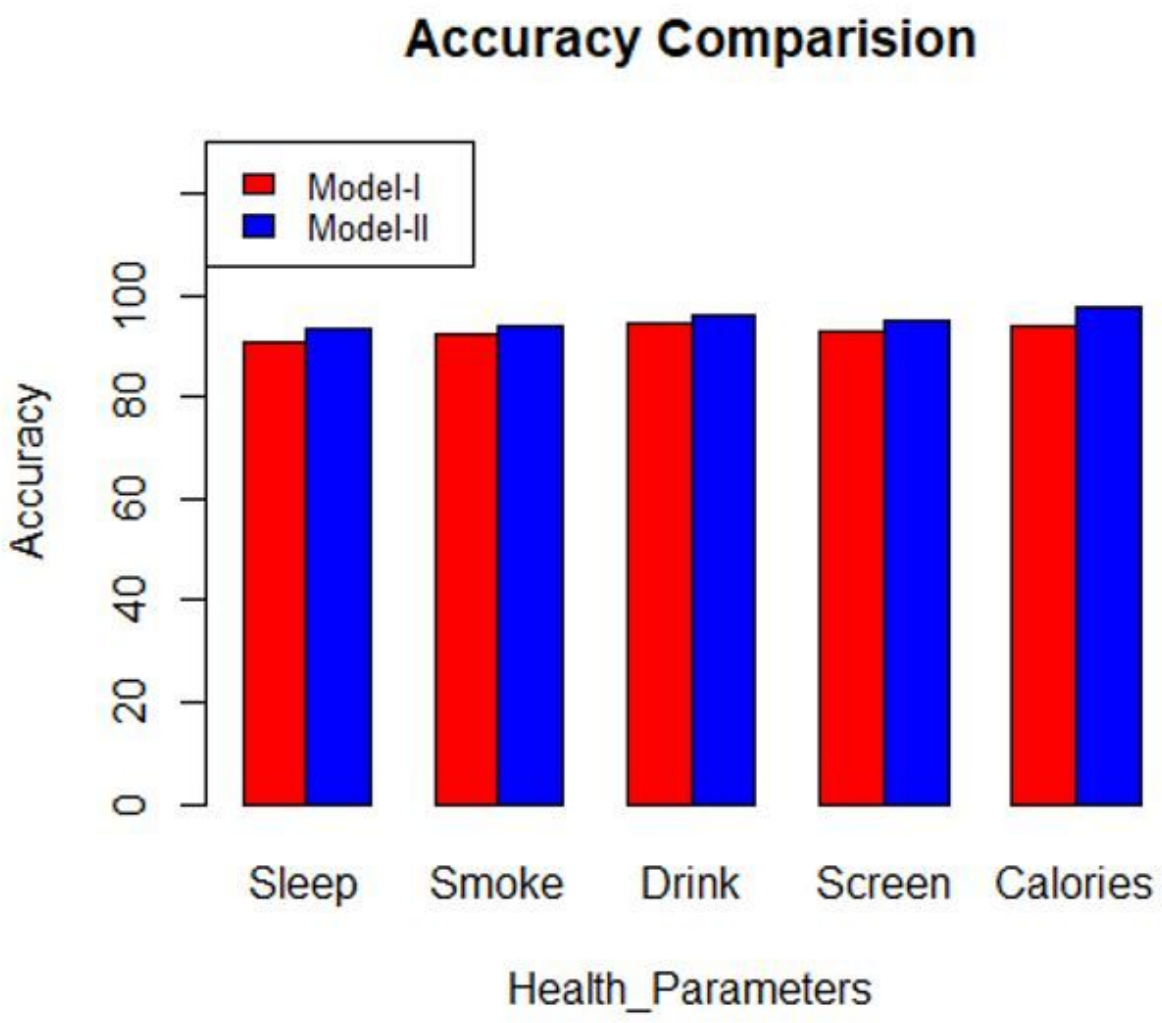

Figure 3

Accuracy: Model-I vs Model-II

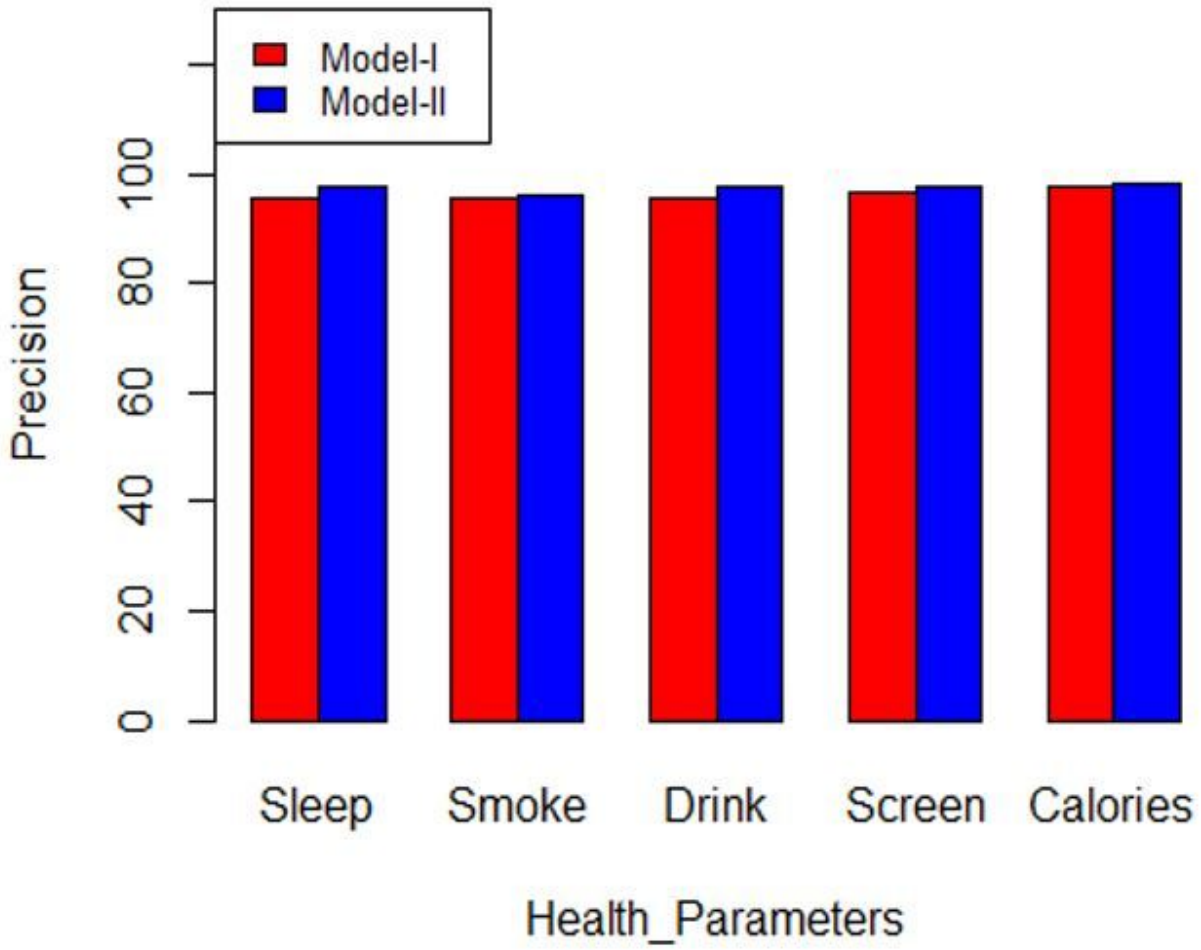

Figure 4

Precision: Model-I vs Model-II

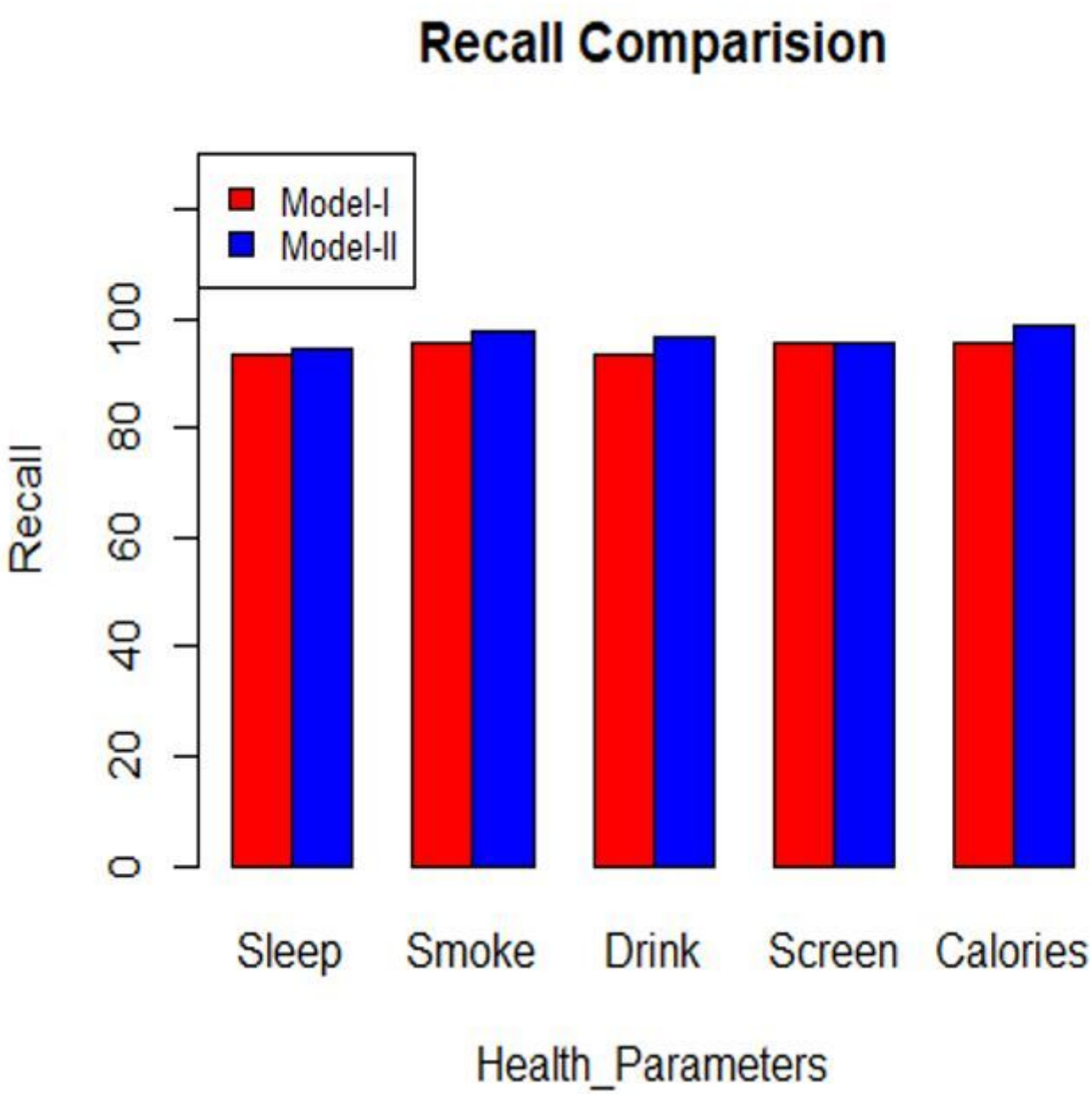

Figure 5

Recall: Model-I vs Model-II

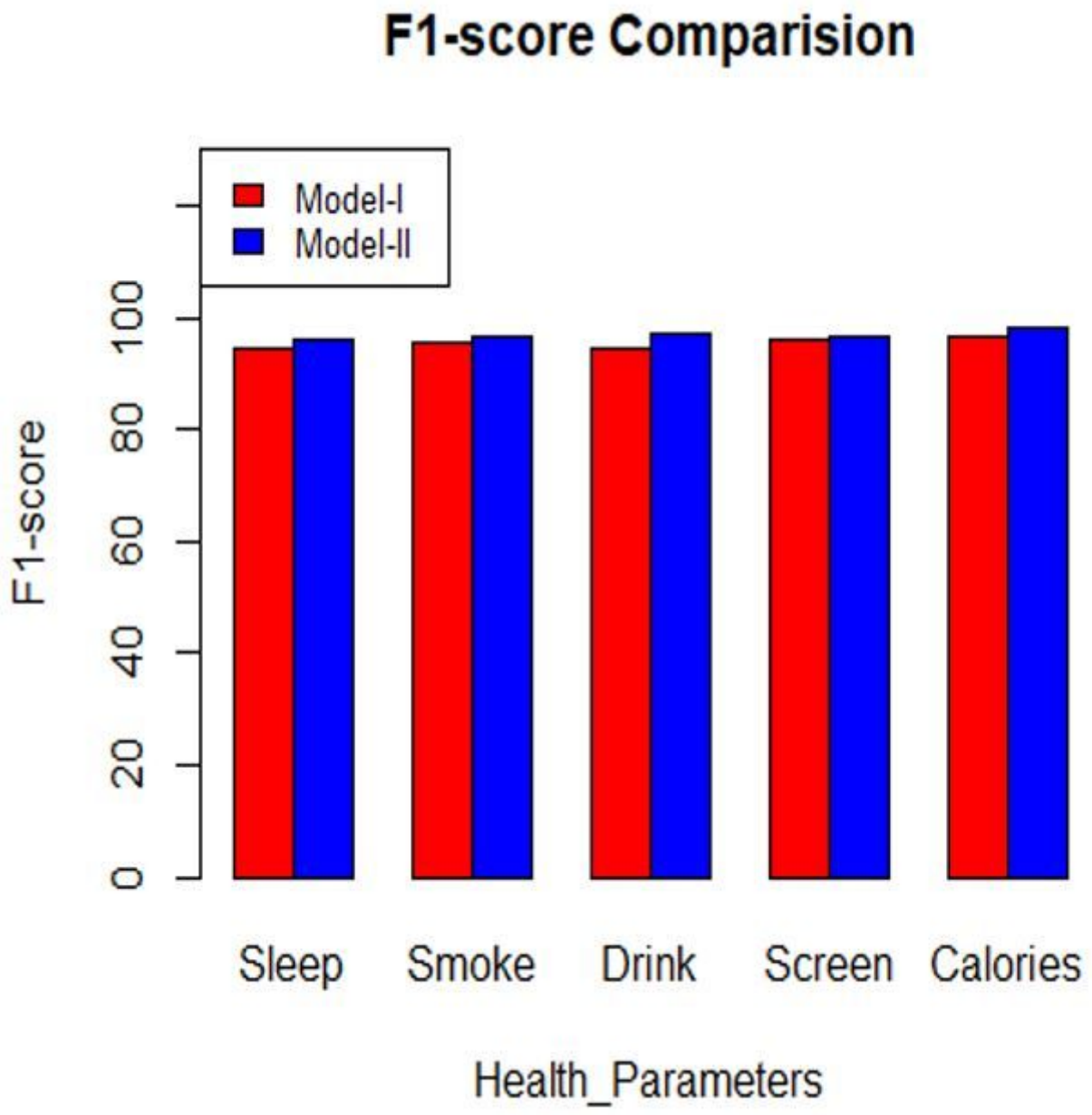

Figure 6

Recall: Model-I vs Model-II

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- formula.docx