SMART HEALTH PREDICTION SYSTEM

A system to provide efficient diagnosis based on existing diagnosis reports

- AARON THOMAS
- POOJA GHATGE
- KARTIK JADHAV
- PIYUSH MANTRI

The goal is to turn data into information, and information into insight.

—Carly Fiorina

Abstract

The aim of this project is to develop a system that uses machine learning algorithms and give the user an estimate of disease risk.The system shall focus on the major diseases that are dependent on an average person’s lifestyle and a few factors rather than all the diseases.

The system will be provided data relevant from his/her medical history/reports. The system will sent this data to the model it created from the datasets of disease records. The purpose of such an application is to help user to understand different medical threats that he/she might be prone to.

Motivation

Since the user may want to get a second opinion for his/her diagnosis, we want to provide the end user with a user-friendly and easily accessible portal where he/she can get to know his/her ailments simply by inputting the symptom values pertaining to a specific disease.Our motivation is to provide accurate prediction of diseases , that too free of cost.

FUNCTIONAL REQUIREMENTS

- Creating an accurate model to find the factors influencing the risk of contracting the disease.
- Incorporating new data updates:The central repository will be updated from time to time with the latest discoveries in medicine by incorporating relevant data from the web into the central repository.
- Devising a healthcare prediction system which would incorporate a wide range of lifestyle diseases in future ie. scalability to incorporate additional disease identification.

NON-FUNCTIONAL REQUIREMENTS

- The data should be properly pre-processed,all anomalies must either be removed or suitably modified.
- The system should be user friendly and user specific.
- The system should store the data sets as well as user data securely.
- The calculation of disease probability should be as precise and accurate.

DATASET

We would be operating on datasets provided by the following ML repository.
http://archive.ics.uci.edu/ml/

The datasets within this repository contains genuine records of patient history, with the following characteristics:

- Source information(Doctor,hospitals,etc.)
- Relevant symptom labels pertaining to the disease and the corresponding range of values.
- The actual condition of the patient,yes/positive if he indeed suffered from the disease,and no/negative if he wasn't suffering.
- Other characteristics of the dataset(classes,missing data,etc.)

SAMPLE DATASET FOR HEART DISEASE

	age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_rate	exercice_angina	disease
2	43	asympt	140	f	normal	135	yes	positive
3	39	atyp_angina	120	f	normal	160	yes	negative
4	39	non_anginal	160	t	normal	160	no	negative
5	42	non_anginal	160	f	normal	146	no	negative
6	49	asympt	140	f	normal	130	no	negative
7	50	asympt	140	f	normal	135	no	negative
8	59	asympt	140	t	left_vent_hyper	119	yes	positive
9	54	asympt	200	f	normal	142	yes	positive
10	59	asympt	130	f	normal	125	no	positive
11	56	asympt	170	f	st_t_wave_abnormality	122	yes	positive
12	52	non_anginal	140	f	st_t_wave_abnormality	170	no	negative
13	60	asympt	100	f	normal	125	no	positive
14	55	atyp_angina	160	t	normal	143	yes	positive
15	57	atyp_angina	140	t	normal	140	no	negative
16	38	asympt	110	f	normal	166	no	positive
17	60	non_anginal	120	f	left_vent_hyper	135	no	negative
18	55	atyp_angina	140	f	normal	150	no	negative
19	50	asympt	140	f	st_t_wave_abnormality	140	yes	positive
20	48	asympt	106	t	normal	110	no	positive
21	39	atyp_angina	190	f	normal	106	no	negative
22	66	asympt	140	f	normal	94	yes	positive
23	56	asympt	155	t	normal	150	yes	positive
24	44	asympt	135	f	normal	135	no	positive
25	43	asympt	120	f	normal	120	yes	positive
26	54	asympt	140	f	normal	118	yes	positive
27	52	atyp_angina	140	f	normal	138	yes	negative
28	48	asympt	120	f	normal	115	no	positive
29	51	non_anginal	135	f	normal	150	no	positive
30	59	non_anginal	180	f	normal	100	no	negative
31	58	atyp_angina	130	f	normal	110	no	negative
32	46	asympt	118	f	normal	124	no	positive
33	54	asympt	130	f	normal	91	yes	positive
34	48	asympt	160	f	normal	92	yes	positive
35	38	asympt	110	f	normal	150	yes	positive
36	39	atyp_angina	130	f	normal	120	no	negative
37	46	asympt	120	f	normal	115	yes	positive
38	33	non_anginal	120	f	normal	185	no	negative
39	50	asympt	145	f	normal	150	no	positive
40	41	atyp_angina	125	f	normal	144	no	negative
41	49	asympt	140	f	normal	140	yes	positive

Detailed comparison of algorithms

1. Naive Bayes

The initial proposal after a short survey of popular ML algorithms led us to this algorithm,which is simple and easy to implement.It is one of the most efficient algorithm for ML,especially for classification.

Problems with Naive Bayes

- Assumes that there is no dependency between the attributes of a dataset.This means that each column contributes independently to the final outcome.This may be detrimental in certain cases and hampers the overall accuracy of the system.
- Not all situations and systems can be determined/solved by probability conditions and calculations.

For these mentioned reasons,we have decided to not go with this algorithm.

2. Logistic regression

In order to create an initial prototype of the system, we have selected logistic regression.The reason for picking this algorithm is that it has a simple logic and is easy to implement.The reason for selecting an algorithm upfront and why we proceeded to create a prototype was to get a better understanding of how the frontend of the system would communicate with the backend.

Working

The algorithm tries to fit a straight line to the data.Consider the following equation:

$$comp=\Theta_0+\Theta_1*x_1+\Theta_2*x_2+\Theta_3*x_3+.....+\Theta_n*x_n$$

Where ,

Y ->output

X’s->the input from the user,in our case the user symptoms

Θ’s ->the coefficients,which our algorithm tries to optimize.

Steps for Logistic regression

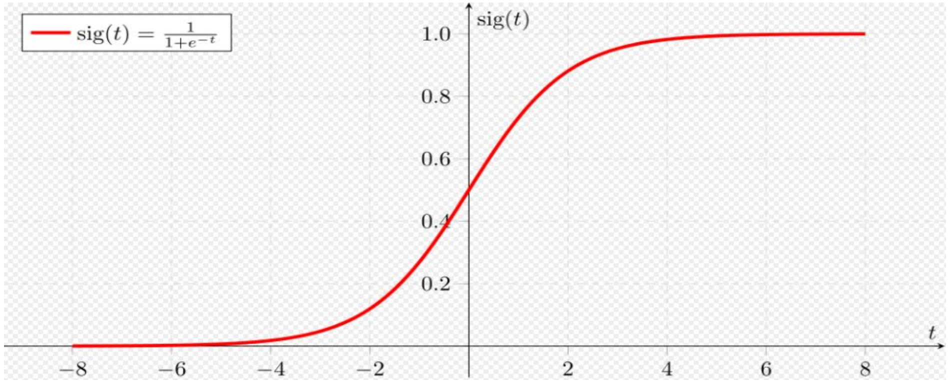
1. Computing the hypothesis.
2. Calculating error using Cost function J(Θ).
3. Calculating derivative(slope) for the gradient curve
4. Fitting the user data to the linear equation.

1. Computing the hypothesis

We randomly initialize the theta(coefficient) values and multiply it with the corresponding x values from the dataset to get the value of comp.This value is given to a sigmoid function $h_\Theta(x)$ in order to compress the value within [0-1].The sigmoid value is the hypothesis i.e the guess of our algorithm.

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Sigmoid function



Curve obtained by Sigmoid function

2. Calculating error using Cost function J(Θ)

We use a logarithmic convex curve cost function to converge to the global minimum of J(Θ).

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: $y = 0$ or 1 always

- So, in summary, our cost function for the θ parameters can be defined as

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

3. Calculating derivative(slope) for the gradient curve

We need to take a step in the gradient curve to improve our Θ values. We do this by taking the derivative of the cost and updating the Θ values using learning rate α. The steps are taken till we reach the global minimum.When J(Θ) reaches the global minimum,it means the error is least.This means that our algorithm has identified the best coefficient values for the data

Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

}

4. Fitting user data to the linear equation.

Once the algorithm computes the Θ values,we can take the user input(symptoms) i.e. the x’s.

Then we will apply the following equation to get the comp value.

$$comp=\Theta_0+\Theta_1*x_1+\Theta_2*x_2+\Theta_3*x_3+.....+\Theta_n*x_n$$

Then we apply the sigmoid function $h_\Theta(x)$ to get the probability value of the disease risk.We output this to the user.

Performance on the test set

The test set has 104 rows.The algorithm gives 79% accuracy in the prediction results.

	Predicted 0’s	Predicted 1’s	Total
Actual 0’s	50	9	59
Actual 1’s	13	32	45
Total	63	41	104

Problems with Logistic Regression

- Very difficult to model the dependency between the attributes of a dataset.As the number of columns keep on increasing,the number of parameters which can be clubbed together increases,also the power of the input attributes is difficult to ascertain.
- The algorithm assumes that the training set can be linearly fitted(logistic regression can be nonlinear,but it is very difficult to deduce the powers of the variables. We assume linearity for the sake of simplicity). While this is certainly possible,it may give poor outcomes in a non-linear datasets.

User Interface

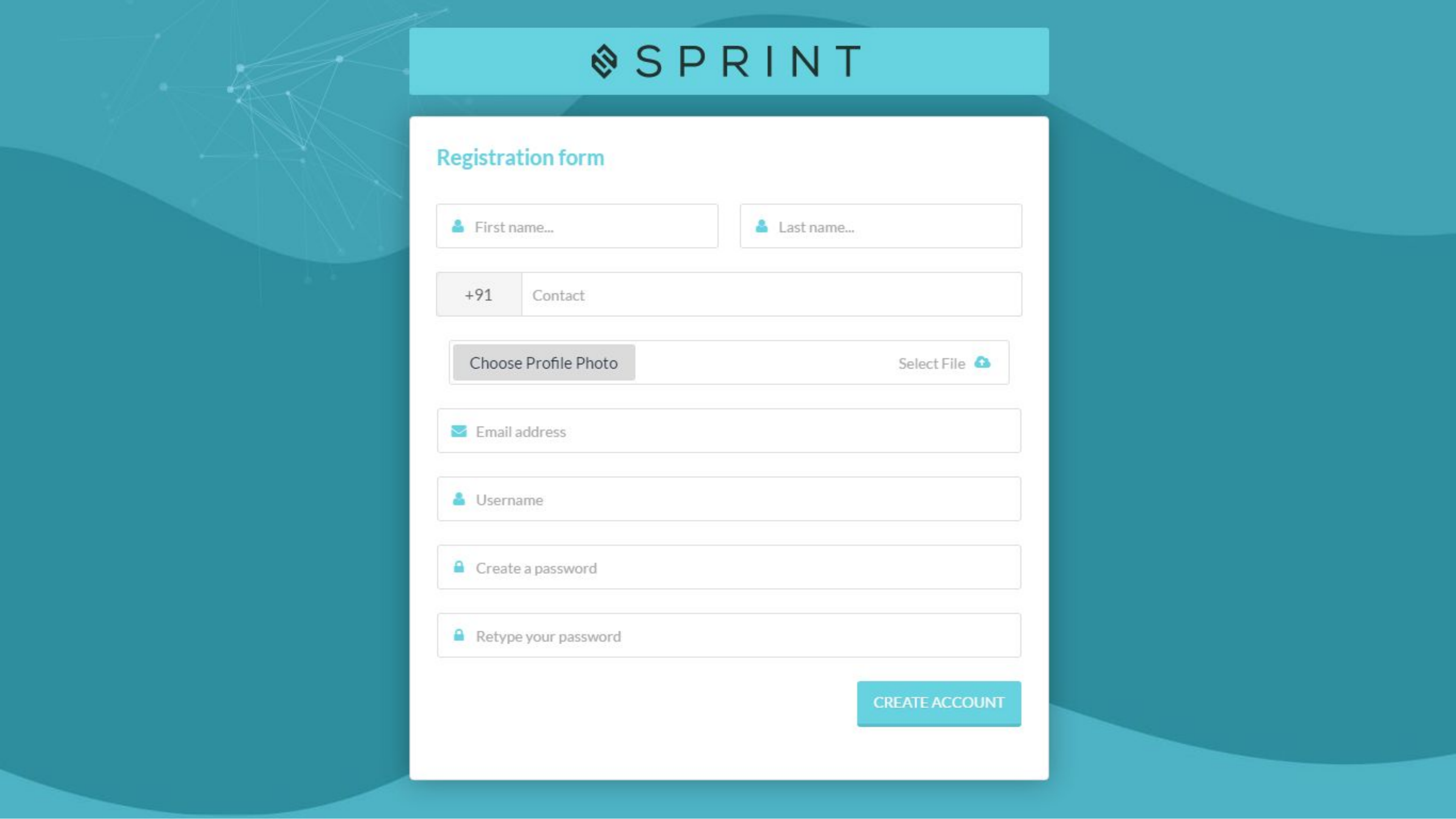
PRINT

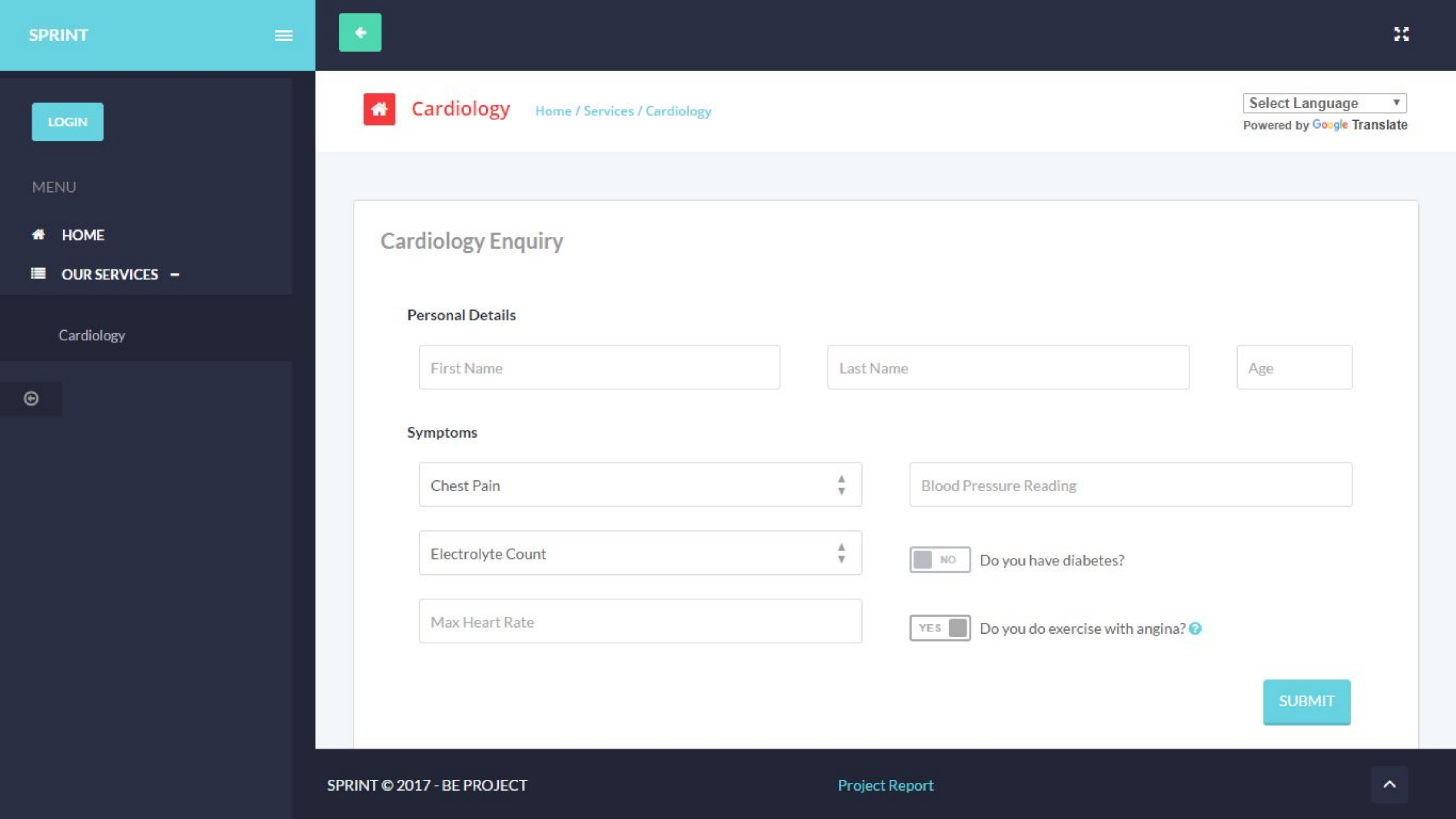
Username

Password

Register

LOGIN


SPRINT
Registration form
First name...
Last name...
+91
Contact
Choose Profile Photo
Select File
Email address
Username
Create a password
Retype your password
CREATE ACCOUNT


SPRINT
LOGIN
MENU
HOME
OUR SERVICES
Cardiology
Cardiology
Home / Services / Cardiology
Select Language
Powered by Google Translate
Cardiology Enquiry
Personal Details
First Name
Last Name
Age
Symptoms
Chest Pain
Blood Pressure Reading
Electrolyte Count
NO
Do you have diabetes?
Max Heart Rate
YES
Do you do exercise with angina?
SUBMIT
SPRINT © 2017 - BE PROJECT
Project Report

SPRINT

LOGIN

MENU

HOME

OUR SERVICES –

Cardiology

Select Language

Powered by Google Translate

Angina & Physical Activity

Prescription for Exercise

Angina

& Physical Activity

The heart, like any other muscle, needs physical activity to keep it in good condition. In coronary heart disease there is narrowing of the arteries that supply blood to the heart. Angina is pain that comes from the heart. This can be severe and very limiting for some and only very mild in others. In an unhealthy heart, any extra blood supply cannot get past the narrowed coronary arteries, which causes pain.

Physical activity reduces your risk of having further problems. Conditioning the heart reduces symptoms of angina and prevents it from getting worse. It can have a positive effect on other risk factors including: high blood pressure, high cholesterol levels (by raising the amount of 'good' cholesterol – HDL), diabetes (by gaining better control of blood sugar), having a family history of heart disease, smoking and increased body fat (in particular having lots of fat around the middle).

Regular physical activity also gives you more energy, builds confidence and can help you to sleep more soundly at night. You can combine your activity time with family and friends or use it as an opportunity to reflect on things and listen to your favourite music.

Physical Activity Recommendations for currently inactive adults with **Angina**

Aim to do the following three types of activity:

- **Aerobic activity** at relative moderate intensity for at least 150 minutes (2 hours and 30 minutes) a week – one way to approach this is to do 30 minutes on at least five days each week.

Plan your lifestyle change

Keep it simple: Don't make drastic commitments. Choose activities that are easy, simple and enjoyable to maintain.

Set a goal and monitor: Set weekly targets that are achievable and keep a record of what you do. If you fail, create barriers to the things that stop you from reaching them.

Go Public: Discuss your goals and activities with others to keep you motivated for longer.

Ca

Age

ding

ve diabetes?

exercise with angina?

SUBMIT

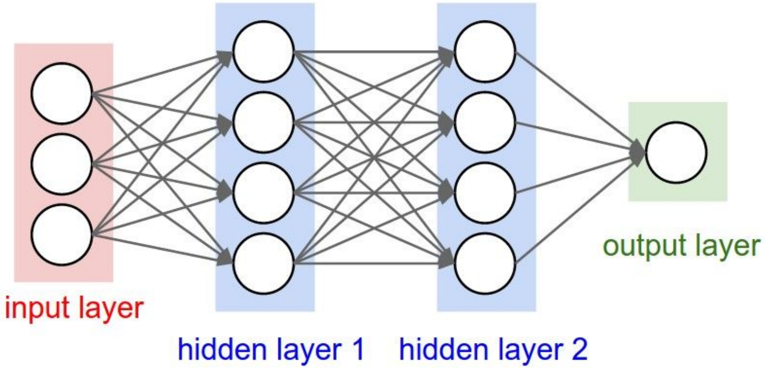
Future Work

We would like to improve the efficiency of the existing system and at the same time address the drawbacks we had with naive bayes and logistic regression algorithm. Focus is on creating/shortlisting an algorithm which:

- Provides a better accuracy.
- Which is nonlinear in nature.
- Develop a relationship between the attributes/columns of the dataset.
- Avoiding/Reducing the complexity of the final equation.

Why Neural Networks?

Neural networks , also called bayesian belief network, is an advanced nonlinear algorithm used to calculate the final outcome by copying the architecture of the neurons in the human body. By the use of layers and nodes within these layers, we can form an interlinking structure which captures the interdependency of attributes. It also avoids over complicated equations of high degree.



APPLICATIONS

Machine learning can be used in the health care to get innovative outcomes in the following areas.

- Personalized healthcare - Predictive data analysis systems can provide early detection of a disease before a patient actually develops disease symptoms.
- Population health - Analytics solutions can mine web based media data to predict future trends.
- Evidence based medicine- Evidence-based medicine involves the use of quantified research and statistical studies by doctors to form diagnosis. This enables doctors to make better decisions not only based on their own judgement and perceptions but also from the best available evidences. It also provides a means of validating and verifying scientific hypotheses with statistical health models .

Thank You