The final report

Introduction

Rationales

Heart disease refers to various types of conditions that can affect heart function. These types include coronary artery disease, valvular heart disease, cardiomyopathies, heart rhythm disturbances, and congenital heart diseases. Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. It difficult to identify high risk patients because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol et cetera. Because of the above problems, scientists have chosen machine learning. due to its superiority in pattern detection and classification. Proven to be effective in aiding decision-making and risk assessment based on large amounts of data.

Objectives

The Objectives of machine learning of this piece to help increase physicians' experience and ability to make informed and accurate decisions about patients with heart disease This is a disease that is difficult to detect and there are many people who are at risk of developing it. We therefore offer an effective diagnostic method for heart disease. using a neural network model

Literature review

1. PREDICTING THE TEN YEAR RISK OF DEVELOPING HEART DISEASE USING MACHINE LEARNING

This article is a development of a screening tool. to predict patients at 10-year risk of coronary heart disease by relying on machine learning on the Framingham dataset

The Framingham dataset is from a cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors. Attributes in dataset:

- Demographic: Sex and Age

- Behavioral: Current Smoker and Cigs Per Day

- Information on medical history: Blood pressure medication, Prevalent Stroke, Prevalent hypertensive and Diabetes

-Information on current medical condition: Total cholesterol level, Systolic blood pressure, Diastolic blood pressure, Body Mass Index , Heart Rate and Glucose level.

Target variable to predict: 10 year risk of coronary heart disease (CHD)

Take the dataset through Data cleaning, pre-processing, Exploratory Data Analysis and Feature Selection use the Boruta algorithm. age and systolic blood pressures are selected as the most important features for predicting. But this project will use the six most important features to build our models that is Age, Total cholesterol, Systolic blood pressure, Diastolic blood pressure, BMI, Heart rate and Blood glucose.

Models and predictions because the dataset is imbalanced. The classifier may have high accuracy but poor precision and recall. To address this they will balance the dataset using the Synthetic Minority Oversampling Technique (SMOTE).

The four algorithms that used are: 1. Logistic Regression 2. k-Nearest Neighbours
3. Decision Trees 4. Support Vector Machine

Conclusion

```
my_data = my_data[top_features]
my_data
```

	age	totChol	sysBP	diaBP	BMI	heartRate	glucose
1	23.0	205.0	138.0	91.0	21.43	85.0	77.0

```
prediction = svm_clf.predict(my_data)
```

```
print('You are not at risk') if prediction[0] == 0 else print('You are at risk')
```

```
You are not at risk
```

The model created can assist in screening. by filling in the information: age, BMI, systolic and diastolic blood pressures, heart rate, and blood glucose levels. The model then runs and makes predictions.

Witchakon Panpai 6205063

Reference: Amayo Mordecai (2020, Apr 29). Heart Attack Risk Prediction Using Machine Learning. from https://towardsdatascience.com/heart-disease-risk-assessment-using-machine-learning-83335d077dad

2. HEALTH FACTORS AND RISK OF ALL-CAUSE, CARDIOVASCULAR, AND CORONARY HEART DISEASE MORTALITY: FINDINGS FROM THE MONICA AND HAPIEE STUDIES IN LITHUANIA

This article discusses the factors that cause CVD(Cardiovascular Disease) and CHD(Coronary Heart Disease) based on data collection in Lithuania. These data were randomly drawn from a population of 9,209 people aged 45-64 (7,648 were free from CHD and stroke at baseline) from 1983 to 2008 which risk factors include smoking, BMI, blood pressure, level of total serum cholesterol, physical activity and level of fasting glucose. After collecting the data, participants were followed up with information on the cause of death whether it was caused by CVD, CHD or other causes.

In this article, statistically analyzed data using P**<0.05** values were taken as statistically significant and 95% confidence intervals (CI). By visualization the data is presented as a table divided into male and female columns, with rows representing 6 risk factors (smoking, BMI, blood pressure, level of total serum cholesterol, physical activity and level of fasting glucose) that each risk factor affects mortality. There is also another table showing how when participants had more than one risk factor affect the cause of death for CVD and CHD. It was concluded that in the prospective analyzes, ideal or intermediate levels of most cardiovascular health factors were associated with significantly lower all-cause. Of course, the data in this paper comes from a population sample of Lithuania, unlike the main project data coming from Cleveland, Hungary, Switzerland, and Long Beach, including methods of collecting data on the cause of death (in our project to predict the disease which means including the living people) which can cause inaccuracies in the analysis.

Pitchat Thanintorn 6205253

Reference: Abdonas T. et al.(2014, December 5) Health factors and risk of all-cause, cardiovascular, and coronary heart disease mortality. Retrieved September 25, 2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114283

3. EFFECTIVE DIAGNOSIS OF HEART DISEASE THROUGH NEURAL NETWORKS ENSEMBLES

Introduction

Recent advances in artificial intelligence have led to the emergence of expert systems for medical applications. The computational tool is designed to improve the physician's experience and ability to make informed decisions about patients. in this study We therefore offer an effective method for diagnosing heart disease. The proposed system uses a neural network model.

Database description

The heart disease database was taken from UCI machine learning repository . The Cleveland heart disease data was obtained from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation from Dr. Robert Detrano. The database contains 303 samples of which 297 are complete samples and six are samples with missing attributes.

Proposed methodology and implementation

A multi-layer feedforward neural network typically has an input layer, an output layer and one or more hidden layers. In multi-layer feedforward networks, neurons are arranged in layers and there is a connection among the neurons of other layers. The inputs are applied to the input layer the output layer contributes to the output directly. Other layers between input and output layers are called hidden layers. Inputs are propagated in gradually modified form in the forward direction, finally reaching the output layer. The backpropagation learning algorithm has been used in the feedforward, single hidden layer neural network. We used 14 neurons in the hidden layer. The initial weights were chosen randomly. Ensemble component

was used to create new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.

Experimental results and discussion

In this study, there were two diagnosis classes: healthy and a patient who is subject to possible heart disease. As it was noted earlier in the background section, several researchers proposed various methods for diagnosing the heart disease. The reported accuracies vary between 50% and 87%. The database contains 303 samples . While 70% of the heart disease database was used for training the neural networks ensemble model, the rest of the heart disease database (30%) was used for validation of the proposed system.

Author (year)	Method	Accuracy (%)
ToolDiag	IB1-4	50.00
WEKA, RA	InductH	58.50
ToolDiag, RA	RBF	60.00
WEKA, RA	FOIL	64.00
ToolDiag, RA	MLP + BP	65.60
WEKA, RA	T2	68.10
WEKA, RA	1R	71.40
WEKA, RA	IB1c	74.00
WEKA, RA	K*	76.70
Robert Detrano	Logistic regression	77.00
Newton Cheung (2001)	C4.5	81.11
Newton Cheung (2001)	Naive Bayes	81.48
Newton Cheung (2001)	BNND	81.11
Newton Cheung (2001)	BNNF	80.96
WEKA, RA	Naive Bayes	83.60
Polat et al. (2005)	AIRS	84.50
Polat et al. (2006)	Fuzzy–AIRS–Knn based system	87.00
Our proposal – SAS base (2008)	Neural networks ensemble	89.01

Conclusions

In this study, SAS enterprise miner 5.2 was used to construct a neural networks ensemble based methodology for diagnosing of the heart disease. Experiments were conducted on the heart disease dataset to diagnose heart disease in a fully automatic manner. Three independent neural networks models were used to construct the ensemble model. The number of neural networks node in the ensemble model was also increased but no performance improvement was obtained. The experimental results gained 89.01% classification accuracy, 80.95% sensitivity and 95.91% specificity values for heart disease diagnosis.

Apisit amnatwong 6205199

Reference: (19 September 2008.) Department of Informatics, Firat University, 23119 Elazig, Turkey Effective diagnosis of heart disease through neural networks ensembles
https://www.sciencedirect.com/science/article/pii/S095741740800657X#fig1

4. ON DEEP NEURAL NETWORKS FOR DETECTING HEART DISEASE

Introduction:

Heart disease is the leading cause of death worldwide, killing twenty million people per year. In many cases there have no clear biomarkers. From the foregoing, most doctor use the American heart association (AHA),which test eight widely recognized risk factor such as hypertension, cholesterol, smoking and diabetes. However, the American heart association (AHA) still has weaknesses, which is this risk assessment model is flawed since it based on assumed linear relationship between risk factor and heart disease outcome, but relationship are complex and with non- linear interactions

Machine learning(ML) techniques can alleviate the possibility of human error and human expertise. Deep neural network (DNN) was designed and adjusted until it can predict the heart disease. The Accuracy is up to 99 percent. The result were evaluated and validated using k-way ,cross-validation

Dataset:

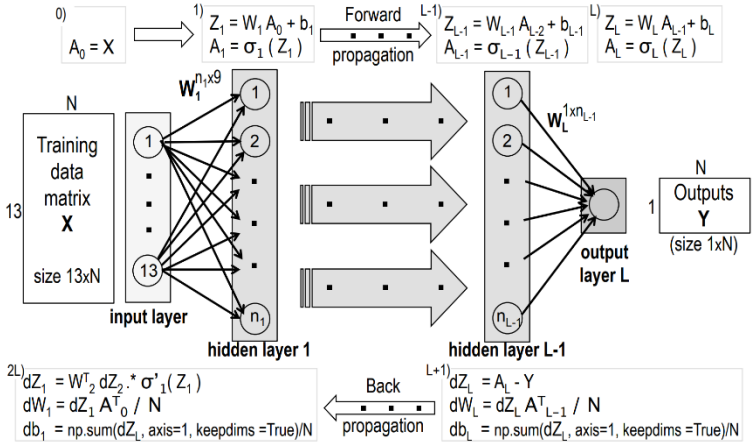
The dataset consist 13attribute:

1) age, 2) sex, 3) chest pain type, 4) resting blood pressure, 5) cholesterol, 6) fasting blood sugar, 7) resting electrocardiographic results, 8) maximum heart rate achieved, 9) exercise-induced angina, 10) ST depression, 11) slope of the peak exercise ST segment, 12) major vessels colored by fluoroscopy, and 13) thallium heart scan results.

DNN:

The neural network is organized into L fully-connected 'layer'(i=1,….L) with in node(artificial neurons) per layer that work together to make a prediction. The connections between layer i-1 and I are represented by numerical weights, stored in matrix w_i of size n_ixn_{i-1},and vector bi of length n_i.

Thus, If the input value for layer I, given by the value at the n_i-1 nodes of layer $i-1$,are represented as vector ai-1 of size ni-1, The output of layer I will be a vector of size n_i ,given by the matrix-vector product$w_i a_i - 1 b_i$ as training will be in pararelle for a batch of nb vector, the input a_{i-1}will be matrices $A_{i-1}$of size n_{i-1} x nb and the output will be given by the matrix-matrix products $z_i = w_i + A_{i-1} + b_i$,



The Forward propagation process

Step0,…,,L

Represent a non-linear hypothesis/prediction function HW,b(X) ≡ AL for given inputs X and fixed weights W, b. The weights must be modified so that the predictions HW,b(X) become close to given/known outcomes stored in Y. The modification of the weights is defined as a minimization problem on a convex cost function J

$$\min_{W,b} J(W,b), \text{ where } J(W,b) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log H_{W,b}(x_i) + (1-y_i)\log(1-H_{W,b}(x_i)).$$

the backward propagation process

steps L+1, ..., 2L,

modify their respective weights Wi , bi during the iterative training process for each layer i as

$$W_i = W_i - \lambda dW_i, \quad b_i = b_i - \lambda db_i,$$

λ is a hyperparameter referred to as learning rate.

Hidden layer

Each of layer uses a different activation function. This paper coded activation function choices for ReLU, sigmoid, tanh, and leaky ReLU.

Conclusion:

This work showed that potential of using DNN-based for detecting heart disease . From routine clinical data, The result show DNN data analysis techniques can yield very high accuracy by 99% accuracy and 0.98MCC .which significantly outperforms currently published research in the area. At present it is still being developed continuously to become a doctor consultation and collect more diagnostic data

Thanakorn chaichiratikul 6205113

Reference: Nathalie-Sofia Tomov (Wed, 22 Aug 2018). On Deep Neural Networks for Detecting Heart Disease.researchgate.RetrievedSeptember28,2021 ,

From https://arxiv.org/pdf/1808.07168.pdf

Methodology

Machine learning algorithm

Neural Network is a mathematical model or computer model for processing information with a connected computation. A neuron network is made up of different neurons according to layers .Input Layer is responsible for receiving data into the neural network by the Input Layer is only one layer and there is a page to send data to the hidden layer. Hidden Layers get data from the previous layer. Notice that the hidden layer basically, if we need more precision we can increase the number of hidden layers and the number of neurons. Output layer receives values from the hidden layer and produces the output.

The obvious point is feature extraction that the neuron network will do manually, but in machine learning, we will have to do it manually before we can integrate it into our model. Deep learning will automatically try to identify the strengths of the input. The output is a class of heart disease and non-cardiovascular disease.

Data source

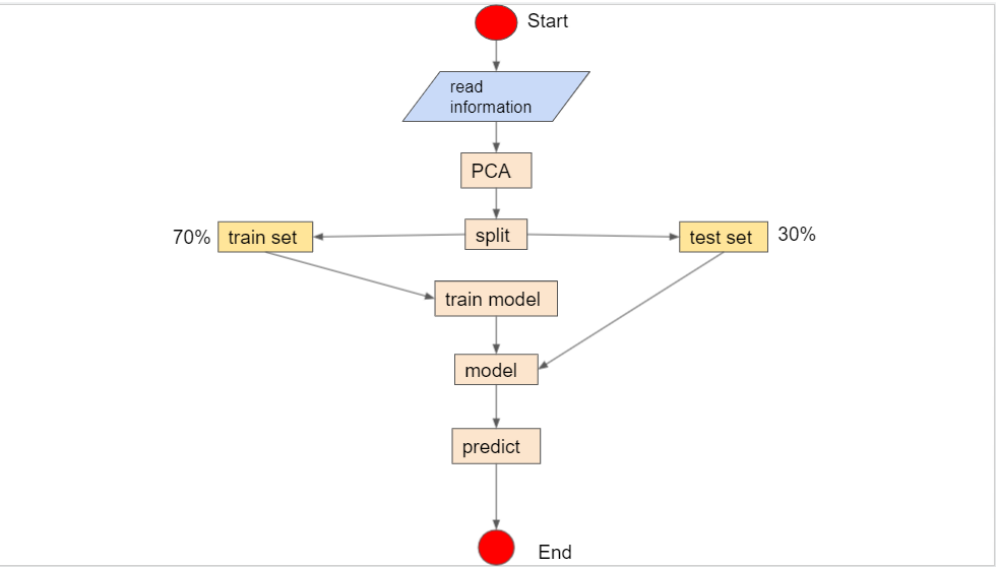
This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

System overall

Input: 9 risk factors affecting CVD and CHD by comparing the main data of the project through the Principal Component Analysis (PCA) process

Output: Heart disease prediction

Process:



1. Read data from a data set that collects information from people about heart disease.

2. Send data to do Principal component analysis to reduce the request dimension to 9 attributes.

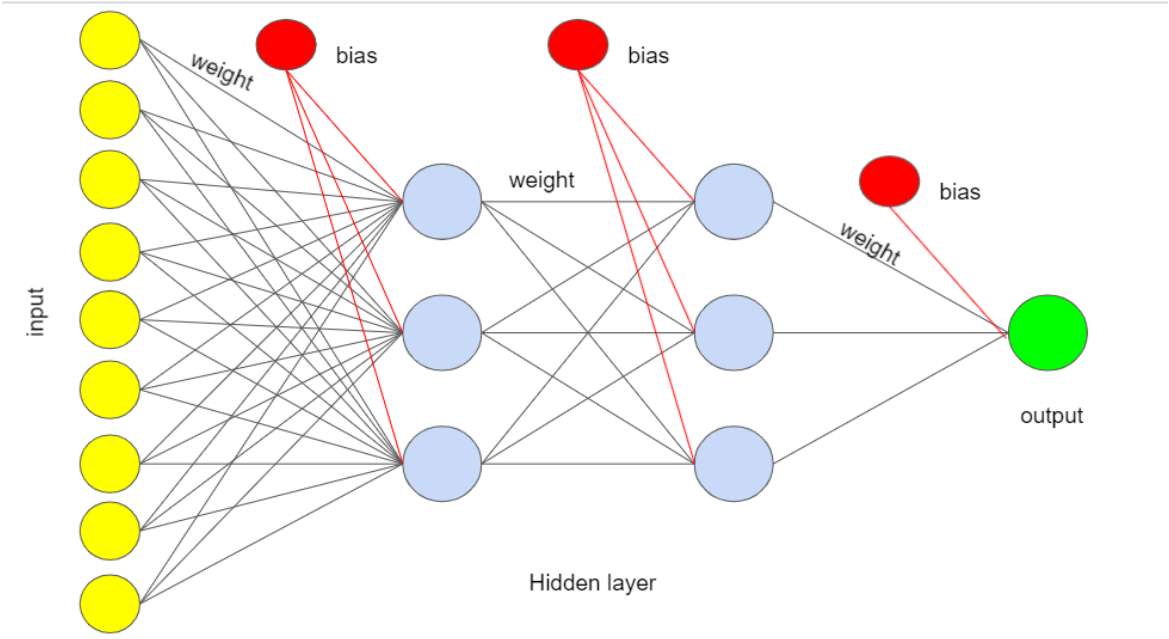
3. Split data into 30% test data, 70% learned data.

4. Take 70% of the data to be learned in the neural network model.

5. get a model

6. Take the data, 30% test data, and test it in the neural network model.

7. Take the value obtained from the test model to find the accuracy value. Neural network model The first layer is called “input” and the last one is the “output”.



One or more layer(s) in between are called “hidden layers”.

1. Forward propagation

1.1 Send 9 inputs (attributes) to calculate weight and bias.

1.2 Calculation of Activation functions In this task, use the Sigmoid function to store the values in the 1st hidden layers for all 3 nodes.

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

1.3 Send 3 new input values to calculate the weight and bias of the next layers.

1.4 Compute Activation functions and store the values in the 2nd hidden layers.

1.5 Pass the value from the 2nd hidden layers, compute the same and pass the value to the output.

2. The back propagation

2.1 Take the output value to calculate Gradient descent

- In output layer $\delta_j^L = (a_j^L - y_j)\sigma(z_j^L)\left(1 - \sigma(z_j^L)\right)$
- In other layers $\delta_j^l = \sum_k \delta_k^{l+1} w_{kj}^{l+1} \sigma'(z_j^l)$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1}\delta_j^l$$

2.2 Subtract the Gradient descent value from the weight and bias of the previous layers.

$$w_{jk}^{l*} = w_{jk}^l - \eta \frac{\partial C}{\partial w_{jk}^l}$$

2.3 Keep calculating

Training and testing process

After that we transfer the data PCA. The PCA data shows which columns are important to the data by matching between sorted eigenvalues and the previous order of eigenvalues. And we got the new_x for training the model.

with the following columns:

```
import numpy as np
import pandas as pd
from sklearn.decomposition import PCA

X = heart2[["age" ,"sex","cp" ,"trestbps" ,"chol" ,"fbs"  ,"restecg"  ,"thalach","exang", "oldpeak" ,"slope","ca","thal"]]
D = X.values
X = D - D.mean(axis=0, keepdims=True)
X = X/D.std(axis=0,keepdims=True)

pca = PCA(n_components=9)
new_heart2 = pd.DataFrame(pca.fit_transform(X))

new_heart2
```

	0	1	2	3	4	5	6	7	8
0	-0.520765	-1.115588	0.958239	-1.147563	-0.607720	-1.483691	0.083895	0.053069	0.872829
1	2.590875	-0.523070	1.464292	1.535439	1.402281	1.491756	1.455630	0.592370	-0.137894
2	3.044483	-1.326406	-0.427875	1.565692	0.260563	-0.737883	0.383635	-1.399363	-0.834595
3	-0.491855	-0.280027	0.802094	-0.981783	-0.535596	-1.422797	0.394991	-1.566299	0.086887
4	2.185672	1.948291	-0.382287	0.298020	-2.408085	-0.478247	1.029869	1.682547	0.445119

1.Age(age in years)

2.Sex(1 = male; 0 = female)

3.Chol(serum cholestoral in mg/dl)

4.FBS(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

5.Thalach(maximum heart rate achieved)

6.Exang(exercise induced angina) (1 = yes; 0 = no)

7.Oldpeak(ST depression induced by exercise relative to rest)

8.Slope(the slope of the peak exercise ST segment)

9.CA(number of major vessels (0-3) colored by fluoroscopy)

Code for random splitting train set and test set

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3,
random_state = 0)
x = X_train
y = y_train
```

We choose a 70/30 ratio of train set and test set because our dataset is small(n<10,000).

Result

```
test= heart1.loc[13:17]
test
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
13	51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
14	52	1	0	128	204	1	1	156	1	1.0	1	0	0	0
15	34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
16	51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
17	54	1	0	124	266	0	0	109	1	2.2	1	1	3	0

```
test= new_heart3.loc[13:17]
test = np.array(test)
for k in range(len(test)) :
  h1 = predict(test[k])
  print(h1)
  print("====================")
```

```
0.08012546528430166
====================
0.0801297742127323
====================
0.8902988963722434
====================
0.8540533954027565
====================
0.08012546528430296
====================
```

Values less than 0.5 are interpreted as integers. 0 = No disease and values greater than 0.5 are assumed to be integer 1 = Diseased. Based on data from all 5 patients, conclusions 1,2. ,5 has no disease, 3,4 has disease. The actual value is "target".

```
#print("จำนวนที่ทายถูก",ac,"จากทั้งหมด",len(y_test))
au = (ac/len(y_test)*100)
print("accuracy",au,"%")
```

```
accuracy 90.25974025974025 %
```

```
[35] #confusion_matrix
from sklearn.metrics import confusion_matrix
y_true = y_test
y_pre = kk
confusion_matrix(y_true,y_pre)
```

```
array([[131,  14],
       [ 16, 147]])
```

```
[36] # f1
from sklearn.metrics import confusion_matrix, classification_report
print ( classification_report(kk, y_test))
```

```
              precision    recall  f1-score   support

           0       0.90      0.89      0.90       147
           1       0.90      0.91      0.91       161

    accuracy                           0.90       308
   macro avg       0.90      0.90      0.90       308
weighted avg       0.90      0.90      0.90       308
```

The f1 score was used to predict whether a patient with heart disease f1 score would perform well for each class of validation methods. In order to measure the effectiveness of both classes, in this work, heart disease is a complex disease, to predict the accuracy of both classes to be both high. If any class is not predictive well, we can see and improve it.

Discussion

Unexpected result

1. In pca if not standardization This will result in the value in the first attribute being too proportional to weight or taking precedence over the other attributes.
   Idea to improve: So we do standardization. To keep the data neutral and to focus on other attributes, PHE pipes the precision of the resistant model.
2. Normally, the predicted neural networks approach 0 and 1, but in this work the predicted values range from 0.85 - 1 and 0 - 0.1. This is due to our model having the number of nodes and the number of rayers small size
   Idea to improve: Add nodes and rayers to make the model more complex.
3. Normally this neural network can be up to 99% accurate, but this model can only predict 90% due to 1000 learning cycles and the complexity of the model.
   Idea to improve: Add nodes, rayers, and learning cycles. to make the model more complicated.

Performance

From the values of table f1, it can be concluded that the model is able to distinguish the two classes equally, with class 0 (without disease) predicting 90% and class 1 (with disease) predicting 91%. is 90%. It can be concluded that this model is good at predicting both classes. at 90% accuracy

Conclusion

In this task, the dataset is brought to pca first as a feature selection in order to reduce the attributes and bring only the important attributes. Then take the new data from the pca to divide the ratio of train 70% data and test 30% data. Bring data train to learn in the model neural network, the purpose of training so that the weight of each node in each neural network ray is adjusted to a value that is suitable for the range of data. which after doing the training Bring data_test to test. In the evaluation model, f1-score was used to see the performance of both classes. In conclusion, the disease-free class performed 90% and the disease-free class 91%. From both outcomes, it was concluded that the model predicted both classes equally well. together at 90%

Pitchat Thanintorn 6205253

Witchakon Panpai 6205063

Apisit amnatwong 6205199

thanakorn chaichiratikul 6205113