*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 11 Issue 2 February 2023

Proposing Heart Disease Alarm System using Arduino based on analysis with Linear Regression and Decision Tree

Author: Seung-Bin Hwang

Affiliation: Portsmouth Abbey School

E-mail: hsbhcy0508@gmail.com

DOI: 10.26821/IJSHRE.11.2.2023.110205

ABSTRACT

Heart disease has a higher average mortality rate than cancer and is the leading cause of death in the United States. The high incidence is a problem, but the difficulty in predicting heart disease is a greater issue. Because heart disease lacks apparent symptoms, it is difficult to forecast and test for, and testing also results in property loss. To make up for this, a decision tree and linear regression system were used to construct an audio basis system that could predict and notify the presence of heart disease. The numerical data were included to the linear regression in order to discover the most highly correlated component. These parameters were age, cholesterol, and restoring blood pressure, and the maximum heart rate was found to be the predicted value. With the projected maximum heart rate, categorical data variables were uploaded to the decision tree to categorize the incidence of heart disease as 0 or 1. The linear regression and decision tree findings are included in the evaluation section. In addition, surveys of the utilized factor data and heart disease are presented. The study concludes by discussing the utility of the proposed system and future RNN-based work.

Keywords: Classification, Data mining, Decision tree, Heart disease, Linear regression, Prediction

1. INTRODUCTION

COVID-19, a novel virus that arose in the 21st century, caused a significant number of fatalities not only in the United States but also worldwide. This was comparable to an unprotected disaster because there was no prior knowledge of the virus. However, there is a disease with a higher fatality rate than this frightening virus, which is heart disease, the leading cause of death in the United States. According to the Centers for Disease Control and Prevention (CDC), heart disease will account for 698,880 of the 3,358,800 deaths in the United States in 2020 [1]. This is almost twice the number of fatalities caused by COVID-19, indicating that 20% of all deaths were caused by cardiovascular disease. Heart disease is not an emerging disease like COVID-19. However, I was compelled to investigate the cause and prevention of why one person dies every 50 seconds from the disease. Heart disease is a terrible condition that results in death within an hour of onset without any deadly cause or warning symptoms, such as sudden cardiac death. In addition, the treatment of heart disease is quite expensive. Each year, around 805,000 Americans suffer from heart attacks. Only around 200,000 of these occur in individuals with a history of heart disease who have never suffered a heart attack. Approximately $363.4 billion was spent on medical services, medications, and productivity losses in 2016 and 2017. For instance, a heart attack of $12.1 billion and Coronary Heart Disease (CHD) of $9.0 billion were included in two of the ten most costly diseases treated in U.S. hospitals in 2013[1][3]. One-fifth of all heart attacks that occur each year are asymptomatic, meaning that the affected individual was unaware that a heart attack was imminent. Pre-testing is a method of preventing heart disease. However, this has limits as well. First, people's lack of knowledge about heart disease is a concern. To confirm cardiac disease, periodic testing is required, but the

majority of uninformed individuals do not know when or how often to test, thus it is finally late [1]. When patients experience a heart attack or become aware of the signs of heart disease, it is too late. Second, hospital visits necessitated by pre-testing might incur expenses on their own. When it comes to things that are unsure, you shouldn't only pay the cost of the test; you need also invest your time and effort. The research began with the question: Is there any way for people to justify going to the hospital not too late while assessing the risk of heart disease in a friendlier and more comfortable manner? This study presents a system that informs individuals with potential heart disease based on ECG measurements and user factors. In the paper, I present an Arduino-based system prototype. The alert is activated based on the analysis of the patient's heart disease data (condition data of patients with heart disease) using linear regression and a decision tree. First, analyze the database of heart disease patients using linear regression and a decision tree. It is possible to forecast the incidence of heart disease based on patient conditions using linear regression and decision trees. Linear regression is a basic and widely employed data mining technique. Regression is utilized to explain the relationship between a dependent variable and one or more independent variables and to forecast the value of a variable. For instance, if you lay out the incidence of heart disease and resting blood pressure data and draw the line that best describes both, you may forecast the incidence of heart disease based on the resting heart rate data.

2. Background

A heart disease is caused by a shortage of blood flow to a portion of the heart muscle due to constriction of the blood vessels (coronary arteries) that supply the heart with blood. Ischemic heart disease comprises angina, myocardial infarction, and sudden death as a result of inadequate oxygen and nutrition delivery to the heart due to inadequate blood flow. Additionally, heart disease is a set of diseases and conditions that cause cardiovascular issues. Each form of cardiac disease is brought on by a cause that is specific to that condition [2]. Surprisingly, high blood pressure and low blood pressure are also forms of cardiac disease. The official terms for them are hypertension and hypotension, respectively. Hypotension is subdivided into essential hypotension and symptomatic hypotension based on the likelihood of disease development [5]. Coronary Artery Disease (CAD) is the most prevalent kind of heart disease in the United States. This occurs when plaque builds up on the arterial blood channel walls, which can restrict blood flow to the heart and lead to a heart attack. Other common heart disorders include ischemic heart disease, angina pectoris, myocardial infarction, arteriosclerosis, and cardiac failure. Ischemic heart disease, for instance, is caused by a shortage of blood flow to a portion of the heart muscle due to the constriction of the blood vessels (tubular arteries) that supply blood to the heart. Necrosis of myocardial cells may develop if the blood supply declines or the time of disconnection is extended; depending on the region and severity, there is a significant risk of myocardial infarction, and sudden death may occur. Angina pectoris and myocardial infarction are referred to combine as coronary artery disease or ischemic heart disease. Angina pectoris is caused by cholesterol and other foreign material collecting in the coronary arteries and narrowing blood vessels. The coronary artery feeds the heart muscle with the energy it needs to continue pumping continuously, especially in the left ventricle. With oxygen and nutrients from the coronary arteries, the myocardium functions as a source of energy. If the blood supply of oxygen and nutrients is diminished for whatever reason, the muscles will experience pain. This discomfort is known as angina, this condition is known as angina attack, and the disease is known as angina pectoris. There are four typical forms of ischemic heart disease, including stable angina, unstable angina, variant angina, and acute myocardial infarction. Stable angina is caused by constriction of the coronary arteries, and chest pain only arises when the amount of oxygen that the heart muscle demands increase. Unstable angina is brought on by an abrupt constriction of the coronary arteries and a reduction in oxygen flow. Variant angina is not caused by anatomical obstruction of the coronary artery, but by coronary artery spasm. Acute myocardial infarction is a disease in which the coronary arteries become abruptly blocked and the blood-supplying heart muscle dies [4]. Tobacco use, diabetes, hyperlipidemia, and high blood pressure are the four main causes of cardiovascular disease. Smoking causes blood arteries to constrict, limiting blood circulation and leading to the formation of blood clots that can obstruct heart

vessels, resulting in myocardial infarction. In addition, when the body is obese, cholesterol builds in the blood vessels. Cholesterol is an important lipid for maintaining the body's regular activities, but excessive cholesterol in the blood due to overeating can create deposits on the artery wall and lead to cardiovascular disease [7]. Heart disease is caused when the coronary arteries that provide blood to the heart are clogged and blood flow is disrupted. In the blood artery, waste such as cholesterol builds up and causes constriction. Included among these are Angina Pectoris and myocardial infarction. People with certain chronic disorders are six times more likely to get an acute myocardial infarction than those without these conditions. Additionally, family history impacts diabetes, hypertension, and hyperlipidemia, hence raising the chance of sudden death. It is known that the risk of sudden death is three to four times greater in families with a history of cardiovascular disease deaths [6]. In addition, if a person has had high blood pressure for an extended period of time, the artery wall is subject to high pressure, which can damage the smooth arterial wall, causing cholesterol and fat to accumulate along the arterial wall. Consequently, blood arteries become constricted and the heart's blood circulation drops. Blood arteries across the entire body also contract. This can lead to heart disease due to the increased workload placed on the heart. Particularly, LDL cholesterol causes arteriosclerosis and has a direct effect on the incidence of heart disease, which is increased by smoking and cholesterol consumption. If you have coronary artery disease, you must thus pay extra attention to managing your LDL cholesterol levels. Reduce your intake of meat, fats, and oils and increase your consumption of blue-backed fish, fruits, and vegetables. In addition, regular exercise and stress reduction aid in the prevention of heart disease, as much as lack of exercise and stress lead to cardiovascular issues [7].

3. Materials and Methods

3.1 Linear regression

Linear regression is a fundamental and widely employed data mining technique. Regression is utilized to explain the relationship between a dependent variable and one or more independent variables (factors) and to forecast the value of a dependent variable. Predictions based on linear regression are only achievable when historical data is available. The history should be numerically stated. There should be a single database including both the relevant parameters and responses. A good example of history would look like Table 1 below [8].

Table 1. Price of house per size

size	price
73	980
60	850
62	830
56	720
34	610
50	580
41	560
16	550
47	500
20	400
10	300
3	150
1	95

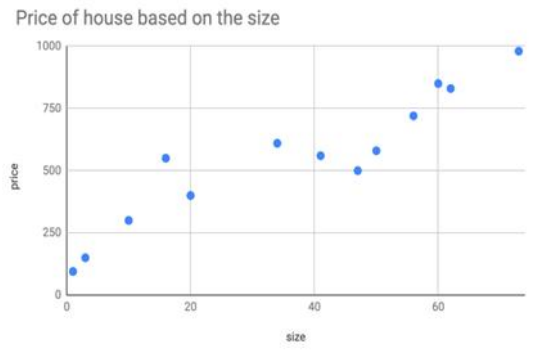
3.1.1 Simple linear regression

Formula 1 defines the simplest form of the regression equation with one dependent and one independent variable:

$$y' = c + b*x, \quad (1)$$

where y' = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Figure 1. Price of house on the size



The estimated dependent variable represents the trend of the history. The history data can be shown as a scatter graph. The Figure 1 is the scatter graph of the history in Table 1.

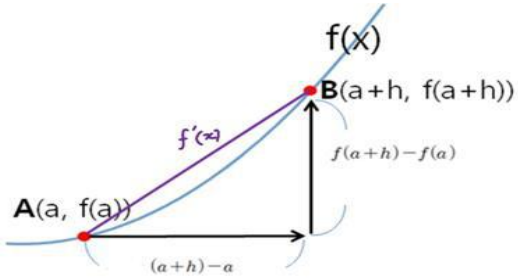
Example 1. This is an illustration of the provided history. The x-axis in this data reflects the size of the house,

while the y-axis represents the price. If you examine the recorded price, it resembles a linear graph, and if you represent this point graph as a linear graph, the y-axis may be predicted along the x-axis. This graph allows us to anticipate the price by inputting the size.

3.1.2 Finding the variable

Deviation refers to a point's slope. Previously, it was established that range errors may be represented by a linearized graph. In order to reduce the error value, it is necessary to square the equation. In this instance, differentiating the related square equation yields the Exeter point, which is the a and b values with the smallest error value.

Figure 2. Way finding Predicted value



$$y''_i = \text{predicted variable } y \text{ based on } a \text{ and } b$$
$$y'_i = \text{mean of } y$$
$$y''_i = ax_i + b$$
$$\Sigma e_i^2 = \Sigma(y_i - ax_i - b)^2$$

(2)

Partial differentiation for b:

$$d\Sigma e_i^{\ 2}/db = 2\Sigma(-1)(y_i - ax_i i - b)$$
$$= -2\{\Sigma y_i - a\Sigma x_i - \Sigma b\}$$
$$= -2\{\Sigma y_i - a\Sigma x_i - n * b\}$$
$$0 = \Sigma y_i - a\Sigma x_i - \Sigma b\}$$
$$n * b = \Sigma y - a\Sigma x_i$$
$$b = (\Sigma y - a\Sigma x)/n$$
$$b = y'/a * x'$$

(3)

Σe_i^2 is differentiated to b, by using the chain rule, $2\Sigma(-b')x(y_i - ax_i - b)$, therefore, $2\Sigma(-1)x(y_i - ax_i i - b)$ comes out as a result. In this case, if y_i, ax_i and $-b$ are all multiplied by Σ, it becomes $-2\{\Sigma y\text{-}a\Sigma x\text{-}\Sigma b\}$. Because Σb is multiplied by b until the nth term, it becomes bx$n$. Therefore, $-2\{\Sigma y\text{-a}\Sigma\text{x-}n\text{*b}\}$ comes out. The differential is the slope of the tangent line, and the pole to be obtained is the interval where the slope of the tangent line is 0, so the value at differentiation, which is $-2\{\Sigma y - a\Sigma x - n * b\}$ must be zero.

$$0 = -2\{\Sigma y - a\Sigma x - nxb\}$$
$$0 = \Sigma y - a\Sigma x - nb$$

(4)

Then, move nxb to the left term to solve b.

$$n * b = \Sigma y - a\Sigma x_i$$
$$b = (\Sigma y - a\Sigma x)/$$
$$yei^2/ya = (-1) * 2\Sigma(y_i - ax_i - b)(x_i)$$
$$0 = -2\{\Sigma y_i * x_i - a\Sigma(x_i)^2 - b\Sigma x_i\}$$
$$a\Sigma(x_i)^2 = \Sigma y_i * x_i - b\Sigma x_i\}$$
$$a\{\Sigma(x_i)^2 - x' * \Sigma x_i\} = \Sigma y_i * x_i - y'\Sigma x_i\}$$
$$a = \Sigma y_i * x_i - y'\Sigma x_i\ /\Sigma(x_i)^2 - x'\Sigma x_i$$

(5)

Σe_i^2 is differentiated to a, by using the chain rule, $2\Sigma(-b')x(y_i - ax_i - b)$, therefore, $(-1)2x(y_i - ax_i i - b)(x_i)$ comes out as a result. ‘a’ is the value which differentiates ei^x to a. Differential calculus is the slope of the tangent line and the extremum point is the point where the slope of the tangent line is zero. So, the value of the derivative, which is $-2\{\Sigma y_i * x_i - a\Sigma(x_i)^2 - b\Sigma x_i\}$ should be 0. Because ‘a’ has to be found, move ‘a’ to the left and solve the equation. Then, $\Sigma y_i * x_i - y\Sigma x_i\ /\Sigma(x_i)^2 - x\Sigma x_i\}$ comes out as the result for ‘a’.

Now, as $b = (\Sigma y - a\Sigma x)/n$, and $a = \Sigma y_i * x_i - y\Sigma x_i\ /\Sigma(x_i)^2 - x\Sigma x_i$

$$b = (\Sigma y - (\Sigma y_i * x_i - y\Sigma x_i\ /\Sigma(x_i)^2 - x\Sigma x_i)\ \Sigma x)/n$$

(6)

The expression b can only be expressed using the variables x and y. Then we can get predicted data(y) with the measured data(x).

Example 2. The table above shows the history of the price of a house based on the size. Variable x is the size and variable y is the price of the house. By one history of the house, we can make a prediction graph. For example,

if the size is 50 then the price is 580. Now, to get prediction data with the least error, find out 'a' and 'b' by using one of the historical data 50 and 580.

$$y''_i = ax_i + b$$
$$b = (\Sigma y - a\Sigma x)/n$$
$$a = \Sigma y_i * x_i - y\Sigma x_i / \Sigma(x_i)^2 - x\Sigma x_i$$
$$b = (7125 - a * 473)/13$$
$$a = (330865 - 548 * 473)/(24221 - 36.38 * 473) =$$
$$10.21793004$$

$$b = (7125 - 10.21793004 * 473)/13 = 176.3014685$$

For example, if the size of the house is 30, the predicted cost would be,

$$y = 10.21592205 * 30 + 176.3014685$$
$$y = 482.77913$$

482.77913 dollars is the predicted cost when the size of the house would be 30.

The goodness of fit test is necessary for the process of predicting data. Two ways of the goodness-of-fit test are 'Standard Error of Estimate' and 'Coefficient of Determination.'

$$SSR = Sum\ of\ Squares: \Sigma(y''_i - y')^2 \quad (7)$$

$$SSE = Sum\ of\ Squares\ errors: \Sigma(y - y_i'')^2 \quad (8)$$

$$SST = Sum\ of\ Squares\ Regression: SSR + SSE$$

$$= \Sigma(y_i - y')^2 = \Sigma(y''_i - y')^2 + \Sigma(y - y_i'')^2 \quad (9)$$

R^2 =Explained variation/total variation= SSR/SST = 1 – SSE/SST

$$= (a\Sigma y_i + b\Sigma x_i y_i - ny'^2) / (\Sigma y_i^2 - ny'^2) \quad (10)$$

The Coefficient of Determination should be between 0 to 1. It is more suitable when SSE is close to 0, and R^2 is close to 1. Considering Figure 1, R^2 be:

$$= 1 - SSE/(SSR + SSE)$$
$$= 1 - (109565.4632/841564.6216)$$
$$= 1 - 0.1301925727$$
$$= 0.8698074273$$

The residual is the difference between predicted values of y and observed values of y.

$$ei = y_i - y'' \quad (11)$$

3.1.3 Evaluation

MSE is the mean squared error

$$MSE = \Sigma(y_i - y'')^2 / n \quad (12)$$

RMSE is the root mean of square error

$$RMSE = \sqrt{(y_i - y)^2/n} \quad (13)$$

RMSE should be used instead of MAE because when the case demands occasional large mistakes in the prediction, the data causing large errors should be avoided. (The value becomes smaller when it has the root on).

Example 3. MSE when the size is 50 then the price is 580 is

$$MSE = \Sigma(y_i - y)^2 / n$$
$$= 109565.4632 / 13 = 8428.112554$$
$$RMSE = \sqrt{8428.112554} = 91.8047523498$$

3.2 Decision Tree

A decision Tree is a categorization technique for both category and numeric values(factors). However, as the number of components increases, so does the depth of the tree. This decreases not just the time efficiency but also the accuracy. The Local database contains both numeric and categorical data, such as maximum heart rate and sex. Impurity refers to the number of persons in certain groups. It refers to the extent to which different classifications are combined. In contrast, purity refers to the number of classes that comprise a single class. For instance, when there are three jars as seen below, jars 1 and 3 may be considered to be 100 percent pure, whereas jar 2 can be said to be highly impure.

Figure 3. Jar explaining purity

Jar 1 **Jar 2** **Jar 3**

Entropy and the Gini Index are indicators that measure this impurity; the algorithm calculated by entropy is ID3, while the algorithm produced by the Gini coefficient is the CART algorithm. Entropy is a measure of the purity of the data distribution; the purer the data, the lower the entropy; conversely, if the purity is poor, the entropy number increases.

3.2.1 Entropy and the Information

$$Entropy(S) = \sum_{i=1}^{c} \quad p_i * I(x_i)$$

(14)

S=a collection of events that have already occurred
c= number of events
I(x) = the amount of information
P(x) = p(x) is the probability that the event x occurs.
The amount of information means the amount of information an event has.

$$I(S) = log_2(1Ip(x))$$ (15)

There is more information if the event is infrequent. As the chance of occurrence of event x grows, the quantity of information reaches zero. In other words, the fewer knowledge people possess, the more common and frequent the events. This implies that when p(x) is small, I(x) is large.

When the impurity is the highest, entropy might also be the highest. The lesser the impurity, the purer the feasible classification. Since entropy and impurity are proportional, the lower the entropy, the lower the impurity; thus, the lower entropy should be the more accurate classifier in the decision tree. The data with the lowest impurity should thus be placed on the first block.

3.2.2 Decision Tree

Decision Tree is a technique that facilitates the classification of both categories and quantitative variables (factors). However, as the number of components increases, so does the depth of the tree. This not only decreases time efficiency but also accuracy.

Example 4. Consider the table below.

Date	weather	temperature		humidity	wind	participation
D1	S	H		H	W	x
D2	S	H		H	S	x
D3	B	H		H	W	o
D4	R	G		H	W	o
D5	R	C		N	W	o
D6	R	C		N	S	x
D7	B	C		N	S	o
D8	S	G		H	W	x
D9	S	C		N	W	o
D10	R	G		N	W	o
D11	S	G		N	S	o
D12	B	G		H	S	o
D13	B	H		N	W	o
D14	R	G		H	S	x

To classify this table as a description tree, the initial placement of classification criteria must be defined. Determine the least entropy standard depending on participation. Nonetheless, the first entropy computation must be performed prior to categorization. The objective is to decide whether or not to play the game.

$$E(participation) = (5/14) * log_2(5/14) + (9/14) * log_2(9/14) \approx 0.94$$

Next, divide each attribute and the perform the entropy calculation

$$IG(S, A) = E(S), -E(S)$$

(16)

a. weather (Sunny, Blurry, Rainy)

$$E(participation|weather) = P(Sunny) * E(participation|Sunny) + P(Blurry) * E(participation|Blurry) + P(Rainy) * E(participation|Rainy)$$

$$= (5/14) * ((2/5) * log_2(2/5)^{-1} + (3/5) log_2(3/5)^{-1}) + (4/14) * (1 * log_2(1)^{-1} + 0 * log_2(0)^{-1})$$

$$(5/14) * ((3/5) * log_2(3/5)^{-1} + (2/5) * log_2(2/5)^{-1} \approx 0.7$$

b. Temperature (H, G, C)

$$E(participation|Temperature) = P(H) * E(participation|H) + P(G) * E(participation|G) + P(C) * E(participation|C)$$

(17)

$$= (4/14) * ((2/4) * log_2(2/4)^{-1} + (2/4) log_2(2/4)^{-1}) + (6/14) * ((4/6) * \log_2(4/6)^{-1} + (2/6 * log_2(2/6)^{-1}) + (4/14) * ((3/4) * log_2(3/4)^{-1} + (1/4) * log_2(1/4)^{-1} \approx 0.91$$

c. Humidity (High, Normal)

$$E(participation|Humidity) = P(High) * E(participation|High) + P(Normal) * E(participation|Normal)$$

$$= (7/14) * ((4/7) * log_2(4/7)^{-1} + (3/7) log_2(3/7)^{-1}) + (7/14) * ((1/7) * log_2(1/7)^{-1} + (6/7) * log_2(6/7)^{-1} \approx 0.74$$

c. Wind (High, Normal)

$$E(participation|wind) = P(weak) * E(participation|weak) + P(strong) * E(participation|strong)$$
$$= (7/14) * ((4/7) * log_2(4/7)^{-1} + (3/7) log_2(3/7)^{-1}) + (7/14) * ((1/7) * log_2(1/7)^{-1} + (6/7) * log_2(6/7)^{-1} \approx 0.782$$

d. Wind (High, Normal)

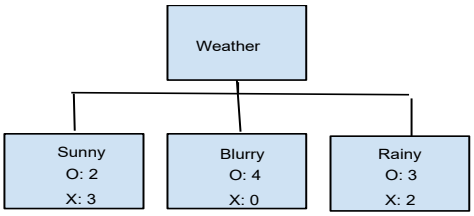
$$E(participation|wind) = P(weak) * E(participation|weak) + P(strong) * E(participation|strong)$$

(18)

$$= (7/14) * ((4/7) * log_2(4/7)^{-1} + (3/7) log_2(3/7)^{-1}) + (7/14) * ((1/7) * log_2(1/7)^{-1} +$$
$$+ (6/7) * log_2(6/7)^{-1} \approx 0.782$$

Therefore, because the entropy of the weather is the lowest, the weather should be the first factor used for classification.

Figure 4. Decision tree example of weather



Now, find each criterion based on the weather and participation. The overcast weather already has zero entropy; hence it is unnecessary to complete the decision tree. The criteria for classifying sunny and overcast weather may differ.

a. Temperature (H, G, C)

The weather should be sunny

$$E(participation|Temperature) = P(H|Sunny) * E(participation|H) + P(G|Sunny) * E(participation|G) + P(C|Sunny) * E(participation|C)$$
$$= (2/5) * (0 * log_2(1/0)^{-1} + 1 * log_2(1)^{-1} + (2/5) * ((1/2) * \log_2(1/2)^{-1} + (1/2 * log_2(1/2)^{-1}) + (1/5) * (1 * log_2(1)^{-1} + 0 * log_2(0)^{-1} \approx 0.4$$

b. Humidity (High, Normal)

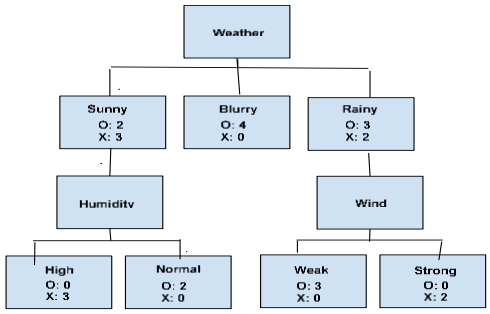
$$E(participation|Humidity) = P(High|Sunny) * E(participation|High) + P(Normal|Sunny) * E(participation|Normal)$$
$$= (3/5) * (0 * log_2(1/0)^{-1} + 1 * log_2(1)^{-1}) + (2/5) * (1 * log_2(1)^{-1} + 0 * log_2(0)^{-1}) = 0$$

c. wind (weak, strong)

$$E(participation|Humidity) = P(weak|Sunny) * E(participation|weak) + P(strong|Sunny) * E(participation|strong)$$
$$= (3/5) * ((1/3) * log_2(1/3)^{-1} + (2/3) * log_2(2/3)^{-1}) + (2/5) * ((1/2) * log_2(1/2)^{-1} + (1/2) * log_2(1/2)^{-1}) \approx 0.951$$

Therefore, humidity should be the next criterion for the sunny branch. Because the Sunny weather branch based on humidity already has zero entropy on both sides, it is unnecessary to continue the decision tree.

Figure 5. Decision tree example of weather



3.2.3 Gini Index

The Gini index or Gini coefficient evaluates the distribution of income throughout a population. Indicators of economic inequality, the Gini index measures the distribution of income or, less frequently, wealth within a population. A Gini coefficient of zero indicates perfect equality, in which all values are identical.

$$G(s) = 1 - \sum_{i=1}^{c} \quad p(i)^2$$

(19)

Example 5. Consider the graph below:

The horizontal line is divided into A and the vertical line is divided into B. Based on the x-axis, the left part is A_1 and the right part is A_2. Based on the y-axis, the upper part is B_1 and the lower part is B_2. The higher the Gini index, the higher the impurity. Therefore, the classification criteria with a low Gini coefficient should be a priority.

1) Gini index when divided based on A

$$G(A_1) = 1 - (8/16)^2 - (8/16)^2 = 0.5$$

$$G(A_2) = 1 - (4/7)^2 - (3/7)^2 = 0.49$$
$$G(A) = (16/23) * 0.5 + (7/23) * 0.49$$
$$= 0.49$$

2) Gini index when divided based on B

$$G(B_1) = 1 - (11/14)^2 - (3/14)^2 = 0.34$$
$$G(B_2) = 1 - (1/9)^2 - (2/9)^2 = 0.2$$
$$G(B) = (14/23) * 0.34 + (9/23) * 0.2 = 0.28$$

Because the Gini index divided based on B is lower than the Gini index based on A, the prior classification criteria should be G (B).

4. Evaluation and Result

4.1 Data Analysis

First, "age" refers to the length of time someone or something has lived or existed. In the sex factor, M signifies male and F represents female. "Chest Pain Type" refers to causes of chest pain connected to the heart. ATA, NAP, ASY, and TA are chest pain types. 'ATA chest pain' is cardiac chest pain resulting from an imbalance between the blood flow to the heart and the oxygen requirements of the heart muscle. "Resting BP" is "resting blood pressure." meaning normal blood pressure that people need to have. Normal resting blood pressure is less than 80 millimeters of mercury. The force exerted by blood on artery walls is known as blood pressure. Normal blood pressure is essential for the circulation of blood from the heart to the organs and tissues of the body. For instance, a man at the top of the list with a resting BP of 140 is risky. Blood cholesterol is a waxy substance. The body needs cholesterol to produce healthy cells, but excessive cholesterol levels can raise the risk of cardiovascular disease. Normal cholesterol levels are typically below 130 mg/dl. If the LDL cholesterol level is below 100 mg/dl, it is regarded as normal, and if it is between 100 and 129 mg/dl, it is considered almost normal. 130 to 159 mg/dl requires caution, 160 to 189 mg/dl is high and greater than 190 mg/dl is extremely high and dangerous. A man at the top of the list, for instance, has a cholesterol level of 289%, which is quite high. The resting ECG is a simple, painless, and rapid test. Certain heart disorders, such as hypertrophy of the heart, ischemia, myocardial infarction, sequelae of myocardial infarction, and cardiac arrhythmias, can be detected by the resting ECG. Electrocardiogram refers to a straightforward test that records the electrical activity of the heart at rest. ECGs can detect a variety of cardiac diseases, such as a rapid, slow, or abnormal heart rhythm, a heart defect, coronary artery disease, heart valve disease, or an enlarged heart. Normal indicates that the heart is beating at a rate between 60 and 100 beats per minute. The 'ST' segment is the flat, isoelectric portion of the electrocardiogram (ECG) between the end of the S wave (the J point) and the start of the T wave. The ST Segment indicates the period between depolarization and repolarization of the ventricle. The term 'Max HR' refers to the maximum number of heartbeats per minute. To get the maximal age-related heart rate, subtract the individual's age from 220. For instance, the predicted maximum age-related heart rate for a 50-year-old male would be computed as 220 - 50 = 170 beats per minute (bpm). For example, a man at the top of the list should have a maximum heart rate of 180 bpm (220-40). Due to his 170 Max HR, he is unaffected by the maximum heart rate. "Exercise angina" is chest discomfort brought on by exercise, stress, or other stimuli that cause the heart to work harder. It is a typical sign of coronary artery disease, which is caused by coronary arteries blocked with cholesterol. 'N' signifies 'no' and 'Y' means 'yes,' which specifies whether or not he or she is experiencing chest pain. "Heart disease" refers to a spectrum of disorders affecting the heart. Problems of the blood vessels, such as coronary artery disease, are included under heart diseases. The four most common types of heart disease are coronary artery disease, arrhythmia, heart valve disease, and heart failure. 0 signifies negative (a person without heart disease), and 1 represents positive (a person with heart disease). Because he has a score of 0 on the chart, he does not suffer from heart disease. Fasting BS- Fasting BS is glucose, often known as "fasting blood sugar." Blood sugar, or glucose, is the most prevalent sugar in the blood. It is derived from food and is the body's primary source of energy. Diabetes is caused by elevated blood sugar levels. A person without diabetes has a normal fasting blood sugar level of less than 100 mg/dl. 0 indicates a normal glucose level, whereas 1 indicates a high glucose level. A man at the top of the list, for instance, has a glucose level of 0, which indicates that his blood sugar level is normal.

Example 6.

Age	The amount of time during which someone or something has lived or existed Heart disease usually developed frequently among people in their 40s and 60s.
Sex	In the sex factor, 'M' is male and 'F' is female.
Chest Pain type	'Chest Pain Type' is the heart-related cause of chest pain. ATA, NAP, ASY, and TA are types of chest pain.
Resting BP	'Resting BP' is 'resting blood pressure.' In other words, it is normal blood pressure that people need to have. A normal resting blood pressure level is less than 80 mmHg.
Cholesterol	'Cholesterol' is a waxy substance in the blood. The body needs cholesterol to build healthy cells, but high levels of cholesterol can increase your risk of heart disease. Usually, normal cholesterol levels are below 130mg/dl.
Fasting BS	Fasting BS is a 'fasting blood sugar' which is known as glucose. Blood sugar, or glucose, is the main sugar found in the blood. A normal fasting blood sugar level for someone without diabetes is less than 100 mg/dl.
Resting ECG	N' means normal, and it shows that the heart is beating at an even rate of 60 to 100 beats per minute.
Max HR	Max Heart Rate, which means the greatest number of beats per minute (bpm).
Exercise Angina	Exercise angina' is a pain in the chest that comes with exercise, stress, or other things that make the heart work harder. 'N' means 'no' and 'Y' means 'yes',
Old peak	'Old peak' is the pointed summit of a mountain
ST Slope	'ST slope' is the segment shift relative to exercise-induced increments in heart rate.
Heart Disease	Describes a range of conditions that affect the heart.

Example 7. Consider the patients below

40	M	ATA	140	289	0	N'	172	N	0	Up	0
49	F	NAP	160	180	0	N'	156	N	1	Flat	1

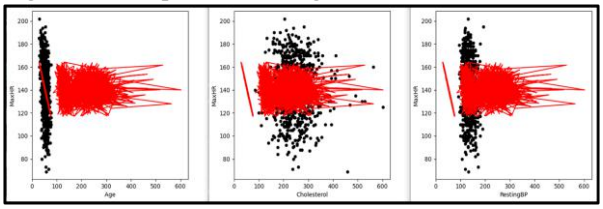
For instance, a 40-year-old man with ATA type chest discomfort, 140 resting blood, and 289 cholesterols has 0 fasting BS. In addition, he has a normal resting ECG, 172 Max HR, no exercise-induced angina, 0 old peak, and a positive ST Slope. Additionally, he has no heart disease. According to the data analysis, a man has both a high resting blood pressure and a high cholesterol level. His heartbeat, however, is normal.

4.2 Result

4.2.1 Linear regression

By entering data such as age, cholesterol, and blood pressure at rest, linear regression may estimate maximal heart rate. Age, cholesterol, resting blood pressure, and maximum heart rate were chosen since these are the only variables that can be measured numerically. Since the elements of the graph might exist in a variety of ways, I have conducted several tests to determine the extent to which data are related.

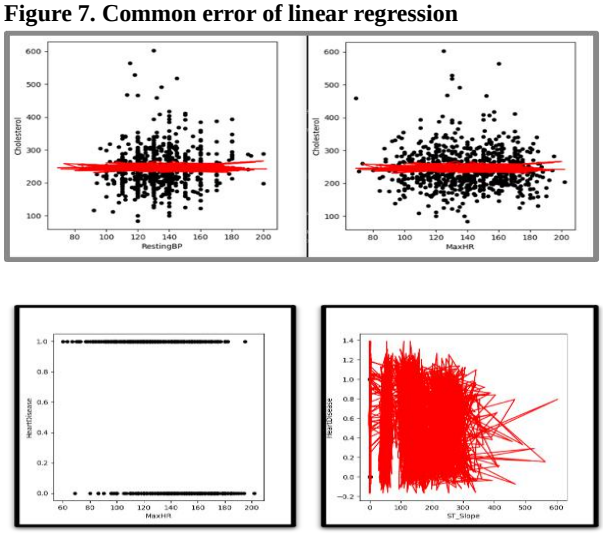
Figure 6. Example of linear regression



This number was determined through linear regression in the project. Maximum Heart Rate (Max HR) is the predicted value, and the factors are {Age, Cholesterol, and Resting Blood Pressure}. This combination is

determined by the R-square value, which represents the correlation score between many value sets, which represent factors. R2 was greatest when the variables were Age, Cholesterol, and Resting Blood Pressure while the value was Maximum heart rate.

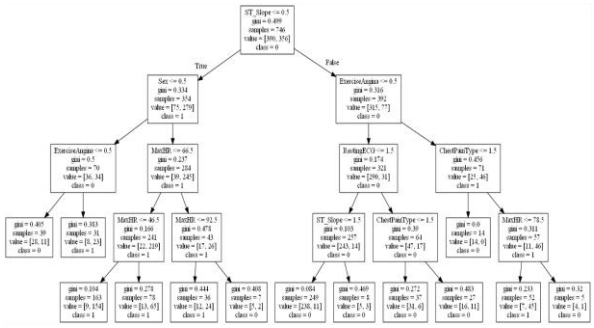
Figure 7. Common error of linear regression



This figure means a linear regression between Age, Cholesterol, Resting Blood Pressure, and Maximum heart rate. The R square value is 0.0093 if the components are Resting Blood Pressure and Maximum Heart Rate and the value is Cholesterol. This is significantly lower than the selected combination. This value, n, is a linear regression error. When data is categorical rather than numerical, an error occurs.

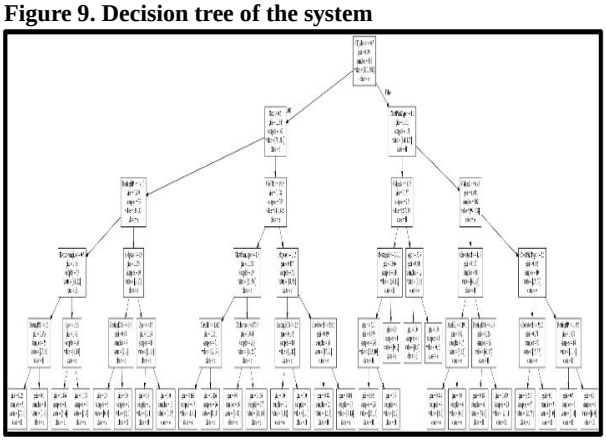
4.2.2 Decision Tree

Figure 8. Result of Decision tree



This figure represents the result of the project's decision tree. The decision tree will predict whether cardiac disease exists. The anticipated value is Heart Disease, which is 1 if it occurs and 0 otherwise. Age, ST slop, Max HR, Sex, Chest pain type, fasting BP, resting ECG, Exercise Angina, and Old peak are the variables. Moreover, Max HR may be predicted using Linear Regression. A decision tree is a method for categorizing both categorical and quantitative elements. Since the decision tree classifier in the Python library only parses numeric values, I had to utilize One Hot Encoding to transform vectors of categorical data to numerical data. This transforms categorical data such as sex ('M' or 'F') into one-hot-vector data ([1,0] or [0,1]). The decision tree must contain more than 40 splits and fewer than 4 levels of depth. As the number of factors increases, the depth of the tree becomes greater than 4.

Figure 9. Decision tree of the system



Similar to figure, which depicts an erroneous decision tree constructed with an excessive number of components, this results in a decline in time efficiency and data accuracy. Therefore, I initially employed linear regression to generate some decision tree variables. Due to the fact that Maximum Heart Rate is predicted by age, cholesterol, and resting blood pressure using linear regression, data cannot be reused in the decision tree.

5. Conclusion and Future works

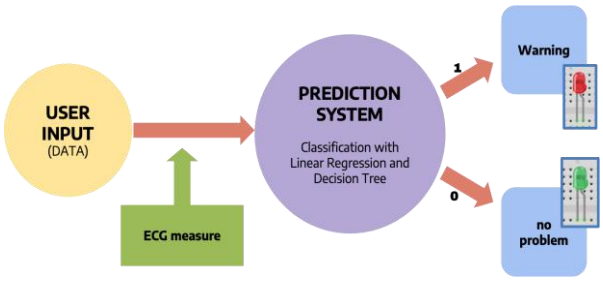
As stated in the introduction, heart disease is a dreadfully serious condition that produces a cascade of health and financial complications. Even if a person has not yet been diagnosed with the condition, pre-testing is prohibitively expensive in nations where government help is limited, including the United States. I offered a solution to this issue by employing data-mining techniques – decision tree and linear regression. The data comprises a number of variables that can affect I integrated an Arduino-implemented ECG sensor into the data collection procedure in order to lower the cost of pre-testing. In the evaluation stage, the suggested system seems reasonable. There is nonetheless room for improvement. This research utilized just numeric and categorical text data. In the actual world, heart disease is diagnosed not only by statistics but also by CT data – which consists of images. For future projects, the system will be augmented with a deep learning model that

has been trained using labeled data (photo: disease). Currently, RNN (recurrent neural network) is the most promising model. RNN is a sequence model that handles sequential inputs and outputs. Since a structure may accept inputs and outputs regardless of the length of the sequence, it is possible to construct it in a number of different ways. One-to-any relations may be utilized for the photo description. On the other hand, we may simply classify the image using the appropriate training data and a many-to-one relation.

6. Proposed System

6.1 Overview

Figure 10. Steps of heart disease alarming system

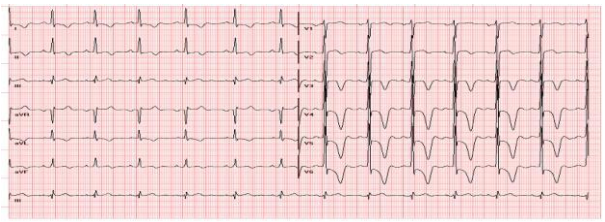


I suggested a system that, using user characteristics and ECG measurements, notifies people with that condition if suspected heart illness occurs. In this project, I propose an Arduino-based logical prototype of the system. The alert is activated based on the analysis of the patient's heart disease data (condition data of patients with heart disease) using linear regression and a decision tree. User input, the initial component of the system, is represented by the yellow portion of the system. The system's users enter information such as their age and gender into the device. Next, the green section is titled "ECG measure" After the ECG has been measured by the Arduino system, further data will be added. The purple circle represents the "prediction system," which displays the predicted outcome. The classification technique is comprised of linear regression and a decision tree. Linear regression and the decision tree model constitute the system. We include ECG measurements in the model's inputs, by entering data such as age, cholesterol, and blood pressure at rest, and the linear regression may estimate maximal heart rate. Also, linear regression first generates decision tree elements. Since Maximum Heart Rate is predicted by age, cholesterol, and blood pressure at rest using linear regression, we do not need to re-enter these values in the decision tree. The decision tree will then forecast the presence of cardiovascular illness. When people have heart disease, the outcome will be 1, and if they do not have heart disease, the result will be 0. The output of the system consists of the numbers '0' and '1'. '0' signifies false, whereas '1' indicates true. If the result is 1, a red LED will indicate a warning for heart disease. If the result is 0, a green light will indicate that the user is healthy.

6.2 ECG monitoring

Certain heart conditions, such as hypertrophy of the heart, ischemia, myocardial infarction, sequelae of myocardial infarction, and cardiac arrhythmias, can be detected by the resting ECG. Electrocardiogram refers to a straightforward test that records the electrical activity of the heart at rest. ECGs can detect a variety of cardiac diseases, such as a rapid, slow, or aberrant heart rhythm, a heart defect, coronary artery disease, heart valve disease, or an enlarged heart.

Figure 11. Image of heart beat



7. ACKNOWLEDGMENTS

My thanks to the people who have supported and given feedback on the paper.

8. REFERENCES

[1]. DJP Barker, “Fatal origins of coronary heart disease.", BMJ, volume 311, 1995, pp. 171-174

[2]. Gerhard Whitworth, "Everything You Need to Know About Heart Disease", healthline, 2020

[3]. McCullough, Peter A., "Coronary Artery Disease." Clinical Journal of the American Society of Nephrology, volume 2(3), 2007, pp.611-616

[4]. Hugh Trowell, "Ischemic heart disease and dietary fiber", The American Journal of

Clinical Nutrition, Volume 25, Issue 9, 1972, pp.926–932

[5]. T Joseph Reeves, Albert Oberman, William B Jones, and Thomas Sheffield, "Natural history of angina pectoris", The American journal of cardiology volume 33 (3), 1974, pp. 423-430

[6]. K.S.Lee, "Warning to the heart diseases", chosen news, 2020

[7]. K.P.Hong, "What causes the heart diseases?",health digest, 2016

[8]. Xiang Li, Christophe Claramunt, "A Spatial Entropy-Based Decision Tree for Classification of Geographical Information" , Transactions in GISVolume 10, Issue 3, 2006, pp. 451-467