**VISUALIZATION, FEEDBACK & DISSEMINATION**

**Heart disease prediction in Tableau**

**Final Assignment**

**Sathwik Nallamada**

**Exclusive summary**

The problem that I going to address is specifically focused on measuring the heartbeat and beats per minute on a real-time basis by using the health trackers connected via the Internet of things(IoT).The specific problem statement I am going to address is “*How accurate is the prediction of human body complications, by measuring the real-time heart beat and beats per minute data collected from the health trackers “ (or) “Can the diseases be predicted by analyzing the real-time heartbeat, beats per minute collected from the health data trackers connected via IoT*”. The dataset I sourced from Kaggle this dataset contains the parameters and the description are listed below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **SNo** | **Name** | **Description** | **measurement type** | **Role** | | --- | --- | --- | --- | --- | | 1 | Age | Age of the participant (in years) | Numeric | Predictor | | 2 | Sex | Gender of the participant | Categorical | Predictor | | 3 | Blood pressure | The value of the resting blood pressure. | Numeric | Predictor | | 4 | Cholesterol | The serum cholesterol level in the body. Measured by mg/dl | Numeric | Predictor | | 5 | Chest pain type | 1-typical angina  2-Atypical angina  3-non-anginal pain  4-asymptotic | Categorical | Outcome | | 6 | FBS over 120 | Blood sugar level | numeric | Outcome | | 7 | ST depression | The level of internal depression of the participant | Numeric | Outcome | | 8 | Number of vessels | The number of vessels in the heart narrowing down | Numeric | Outcome | | 9 | The allium | The allium heart related disease categorized by  3-normal  4-fined effect  7-reversible efffect | Numeric | Outcome | | 10 | Max.HR | The maximum human heart rate | Numeric | Predictor | | 11 | Heart disease | The presence or absence of heart disease | Categorical | Outcome | |  |  |  |  |  | |

**Define:**

I would be considering the continuous variables among the dataset to compute the p-value (Probability value) and R-squared value of the occurrence of heart disease with the three different predictor variables of my choice. The first predictor variable I would like to consider is MAX.HR, this predictor variable explains the maximum heart rate in the specific period. Cholesterol, this variable explains the amount of cholesterol present in the body. BP, the abbreviation of BP is blood pressure this variable calculates the amount of blood pressure in the body.

I collected the dataset from the Kaggle website which it an open-source dataset and the types of each predictor variable are listed below.

| **SNo** | **Name** | **Type** |
| --- | --- | --- |
| 1 | Age | Numerical |
| 2 | Sex | Categorical |
| 3 | Blood pressure | Numerical |
| 4 | Cholesterol | Numerical |
| 5 | Chest pain type | Categorical |
| 6 | FBS over 120 | Categorical |
| 7 | ST depression | Numerical |
| 8 | Number of vessels | Categorical |
| 9 | The allium | Integer |
| 10 | Max.HR | Numerical |
| 11 | Heart disease | Categorical |

**Analysis**

As mentioned in the above problem statement I would be considering three different continuous variables to analyze heart disease. Following that, I would create a linear regression model as a MAX.HR is the outcome variable and Cholesterol level and Blood pressure level as the predictor variables.

In the next step, I would be examining the individual parameters with the linear regression model to find the R-squared value and the P-value of the prediction of heart disease.

To perform the above analysis, I would be creating four different sheets where in which the first will have a range of the values of all the predictor variables.

In the first sheet, I would be creating a linear regression model with the combination mentioned above. Following that, I would be considering the Cholesterol and the linear regression model as the predictor variables and the heart disease prediction as the outcome variable and find the P value and R squared value.

The reason I will be creating a linear regression model is to fulfill my problem statement. As I mentioned in my problem statement, I would be using heart rate, as the main predictor variable to predict heart disease. I had now considered a linear regression model to find the relationship between the MAX.HR and BP, cholesterol levels to predict heart disease.

In the second sheet, I will be analyzing the MAX.HR with the linear regression model to predict the heartbeat.

In the third sheet, I will be analyzing the blood pressure with the regression model to calculate the P value and the R squared value of the presence and absence of heart disease.

Coming to the fourth sheet I would be plotting all three predictor variables on the X-axis and the regression model on the Y-axis. Following that, I will be calculating the R squared value and the P value of all three predictor variables.

**The full model**

Chart

Description automatically generated

I above graphical representation explains the range of values (Note: I did not consider the categorical variables). By glancing at the above graph cholesterol, MAX.HR and BP has the longer range of values with respect to others. So, following through the analysis I would be considering these three variables to predict the heart disease.

**Cholesterol level:**

Chart, bar chart, histogram

Description automatically generated

The blue color denotes the presence of the heart disease, and the orange color denotes the absence of the heart disease. The P value of the cholesterol level is with respect to the presence of heart disease is 0.8426 and the r squared value is 0.004. By this we can conclude that cholesterol is not the significant predictor of the presence of heart disease. It explains only 0.4% of the total prediction of the presence of heart disease.

Coming to absence of heart disease the P value stood around 0.255 and R squared value is 0.013. By this we can conclude cholesterol level alone cannot interpret the absence of heart disease.

**Blood pressure (BP):**

Chart, line chart, histogram

Description automatically generated

The R squared value and the P value stood for the absence of the heart beat stood at 0.0006 and 0.87.The R squared value above explains that BP predictor variable can only predict the 0.6% of the total in absence of heart diseases .While coming to the presence of heart disease the R squared value and the P value are 0.05 and 0.18.By this we can conclude that BP is not the significant predictor for presence or absence of the heart disease.

**MAX.HR**

Graphical user interface, chart

Description automatically generated

The R-Squared value and the P values for the maximum heart rate parameter for the absence of the heart disease stood at 0.051 and 0.064 respectively. Since the P-Value is greater than 0.05 the MAX.HR is not a significant predictor of the heart disease. While the P values and R squared values for in the presence of heart disease is the 0.499 and 0.007. By this we can conclude that MAX.HR is also not a significant predictor in both cases.

**The combined model:**

Graphical user interface, application, table, Excel

Description automatically generated

The combined model is the combination of all three parameters. All three parameters are plotted on the X-axis, while the regression model is plotted on the Y -axis.

In the first graph cholesterol level is examined with the regression model concerning MAX.HR and BP level. The R-squared value and the P values in the presence of heart disease are 0.9238 and <0. 001. By this, we can conclude that Cholesterol level is the significant predictor of heart disease in the presence of MAX.HR and BP.

The R-squared value and P values in the absence of heart disease are 0.366 and <0. 001. This determines that Cholesterol level is the significant predictor of heart disease in the presence of MAX.HR and BP.

Coming to the MAX.HR values the R squared, and p values in the presence and absence of heart disease are (0.00022&0.60) and (0.0017&0.61). This leaves us with that MAX.HR is not a significant predictor even when combined with BP and cholesterol levels.

While on the other hand the values R squared and p values in the presence and absence of heart disease of the BP levels are (0.239 &>0.001) and (0.727 & >0.001). This concludes that BP is the significant predictor when combined with MAX.HR and cholesterol levels, in the presence and absence of heart disease.

**Deployment:**

By looking at the graphs and analysis above, we can assume that no-one predictor is solely responsible for the prediction of heart disease. By glancing at the R squared values and P values generated in the above analysis, the individual parameters can only predict 5% of the actual model, and there is no statistical significance with the above predictor variables since P -values are less than 0.05 in all the above cases.

The analysis in the above scenarios developed when the individual parameters were combined with the rest of the predictor variables. The individual parameters showed greater statistical significance when combined with the rest of the predictor variables in most of the cases, in the combined model. The cholesterol level and BP level showed greater statistical significance while MAX.HR failed to generate statistical significance.

**Conclusion:**

By skimming through the above analysis, we can conclude that as the predictor variables increase, the greater accuracy of the model. The combined model is more accurate than the other three models. The greater the predictor variables we consider in evaluating the models the greater the accuracy of the prediction.

In the future, I would like to consider the categorical variables that are present in the dataset for my analysis.

**Additional analysis:**

I had done some additional analysis with the by creating a linear regression model of the parameters I am interested and analyzed the presence and absence of heart disease with that model. And the results are as shown below

Chart, bar chart

Description automatically generated

References:

<https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction>

Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications*, *68*(16).

Indrakumari, R., Shukla, P., & Sehgal, A. (2021). Heart Disease Prediction Using Tableau. In *Exploratory Data Analytics for Healthcare* (pp. 125-141). CRC Press.

Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart disease prediction using exploratory data analysis. *Procedia Computer Science*, *173*, 130-139.