Detection of Heart Disease

Kanaan Sullivan
Computer Science 5300

April 24th, 2024

Contents

Abstract

All project reports in my courses should be prepared using Overleaf. Through this short tutorial I hope to help you prepare your Overleaf report. You can learn more about "Overleaf" here. You are also welcome to use any templates you want; here is an example.

1 Introduction

Dataset and Problem: The dataset used for this report is the "Logistic Regression To Predict Heart Disease" Dataset found on Kaggle. The problem attempting to be solved is with the inputs given in this dataset can we create a Neural Network to correctly predict heart disease within the next ten (10) years.
Motivations: As someone who has one and a family with a history of heart disease this problem is fairly close to the heart. As such it feels right to select this dataset keeping in mind personal struggles with the subject.

2 Dataset

The "Logistic Regression To Predict Heart Disease" is sourced from Kaggle Data Science[1]. It is a set of data from an ongoing cardiovascular study being performed on the people of Farmingham, Massachusetts. The base dataset contains 15 input fields and over 4000 records. The output field is binary in being a 0 or 1 as for the chance if the record has a chance of heart disease over the next ten years.

2.1 Input Fields

Below are the input fields of the dataset and a short definition.

1. Sex: Is the patient male or female (Binary, 0 for Female, 1 for Male).
2. Age: How old is the patient (in years), continuous.
3. Education: How much education the patient has received (Rated 1-4 for years).
4. currentSmoker: Is the patient a smoker (Binary, 0 for non-smoker, 1 for smoker).
5. cigsPerDay: If the patient is a smoker, how many cigarettes do they smoke per day (Flat number of how many cigarettes are smoked).
6. BPMeds: Is the patient on blood pressure medication (Binary, 0 for not on medication, 1 for being on medication).
7. prevalentStroke: Does the patient have a history of a stroke (Binary, 0 for no history, 1 for past stroke history).
8. prevalentHyp: Does the patient have a history of being hypertensive (Binary, 0 for no history, 1 for past hypertensive history).
9. diabetes: Is the patient diabetic (Binary, 0 for non-diabetic, 1 for diabetic).
10. totalChol: Total cholesterol level of the patient.
11. sysBP: The systolic blood pressure of the patient.
12. diaBP: The diastolic blood pressure of the patient.
13. BMI: The Body Mass Index (BMI) of the patient.
14. heartRate: The heart rate of the patient, continuous.
15. glucose: The glucose level of the patient, continuous.

2.2 Visualization of Input Data Distribution

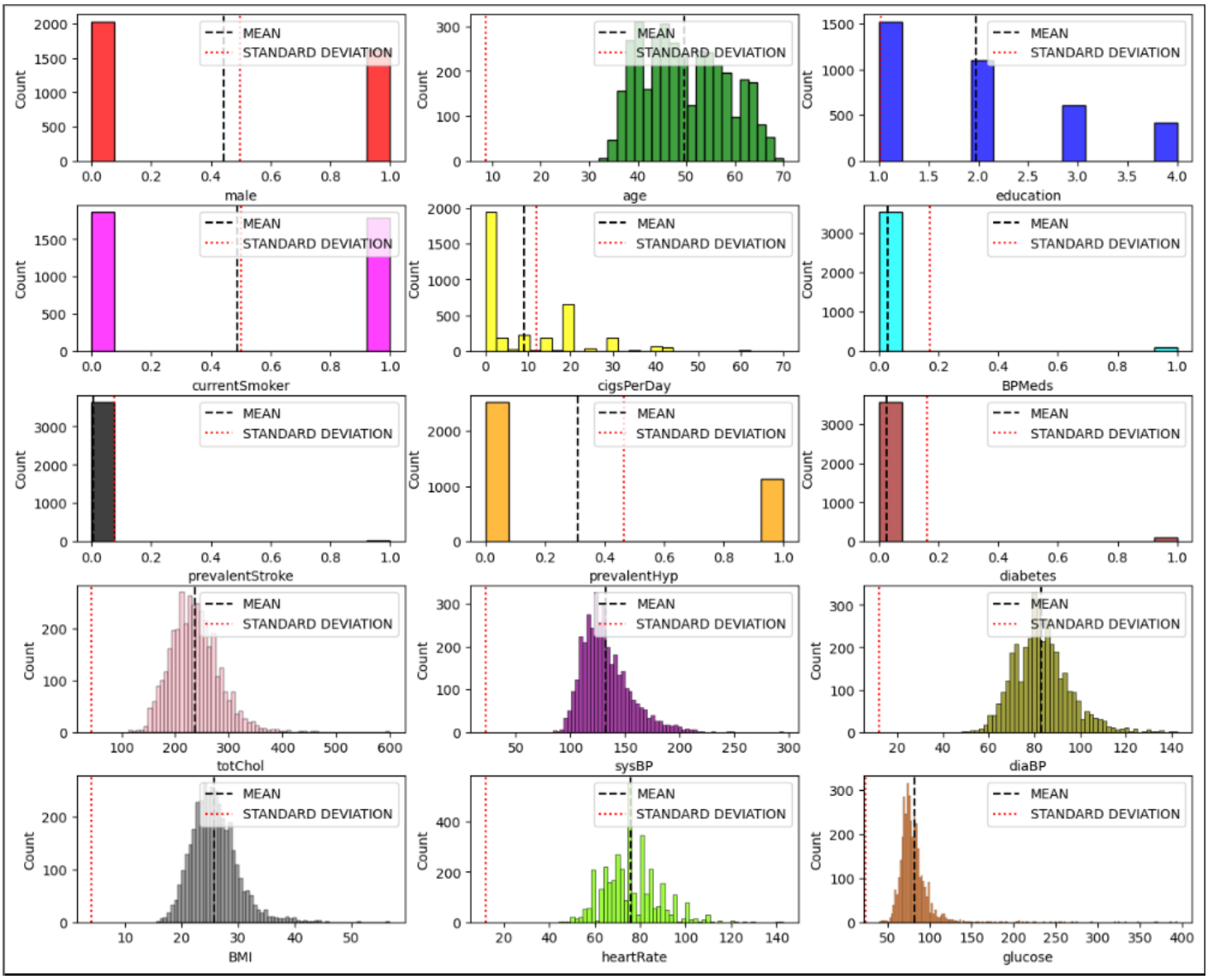

Figure 1: Input Data Histograms

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
count	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000
mean	0.443654	49.557440	1.979759	0.489059	9.022155	0.030361	0.005744	0.311543	0.027079	236.873085	132.368025	82.912062	25.784185	75.730580	81.856127	0.152352
std	0.496883	8.561133	1.022657	0.499949	11.918869	0.171602	0.075581	0.463187	0.162335	44.096223	22.092444	11.974825	4.065913	11.982952	23.910128	0.359411
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	113.000000	83.500000	48.000000	15.540000	44.000000	40.000000	0.000000
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	206.000000	117.000000	75.000000	23.080000	68.000000	71.000000	0.000000
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	234.000000	128.000000	82.000000	25.380000	75.000000	78.000000	0.000000
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.000000	1.000000	0.000000	263.250000	144.000000	90.000000	28.040000	82.000000	87.000000	0.000000
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.000000	1.000000	1.000000	600.000000	295.000000	142.500000	56.800000	143.000000	394.000000	1.000000

Figure 2: Input Features Statistics

2.3 Distribution of Output Data

The output data included in the dataset is a binary class with 0 being for no detected Chance of Heart Disease and a 1 for if this chance was detected within the next 10 years (10YearCHD).

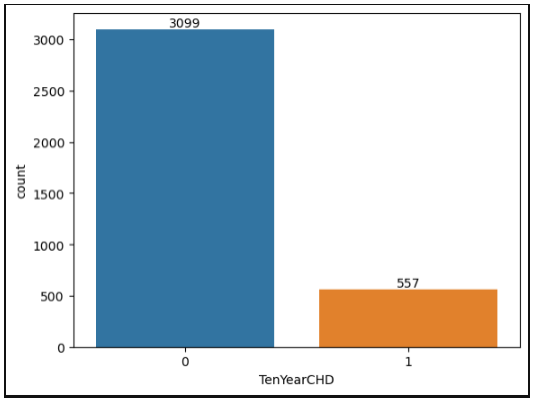

Figure 3: Output data statistics. Showing a 85% chance of no heart disease and a 15% of Heart Disease

3 Data Processing

3.1 Data Normalization

Data Normalization allows a tighter set of constraints to be me made on data. It removes the impact of scale and puts all input fields into the same scale. This "Normalizes" the data and allows for faster processing due to the smaller scalar. After normalization all values of the dataset are between 0 and 1. The process used for this normalization was the Min Normalization function. Which can be defined from the equation below.

$$Xnormal = \frac{X - Xmin}{Xmax - Xmin}$$

3.2 Visualization of Normalized Data

Below are the visualizations of the normalized input fields

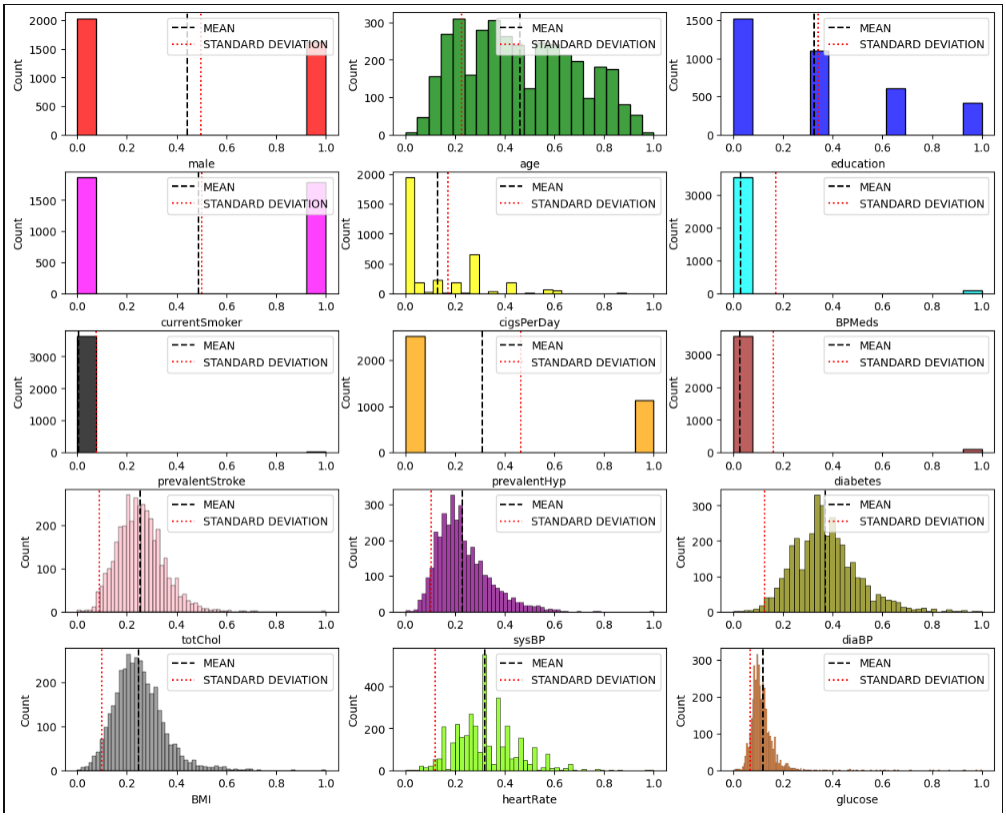

Figure 4: Normalized Input Field Histograms

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
count	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000	3656.000000
mean	0.443654	0.462038	0.326586	0.489059	0.128888	0.030361	0.005744	0.311543	0.027079	0.254360	0.231054	0.369440	0.248284	0.320511	0.118238	0.152352
std	0.496883	0.225293	0.340886	0.499949	0.170270	0.171602	0.075581	0.463187	0.162335	0.090547	0.104456	0.126718	0.098544	0.121040	0.067543	0.359411
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.263158	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.190965	0.158392	0.285714	0.182744	0.242424	0.087571	0.000000
50%	0.000000	0.447368	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.248460	0.210402	0.359788	0.238488	0.313131	0.107345	0.000000
75%	1.000000	0.631579	0.666667	1.000000	0.285714	0.000000	0.000000	1.000000	0.000000	0.308522	0.286052	0.444444	0.302957	0.383838	0.132768	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 5: Normalized Input Features Statistics

4 Initial Modelling

The problem attempting to be solved, the prediction of whether a patient will need to worry about heart disease within the next ten years is a Binary Classification Problem. To fit this data a Logistic Regression model works best for initial testing.

4.1 Overfitting Initial Logistic Regression Model

The first step was to over-fit the model using the entire dataset to properly hone in on what would be the correct size for the Neural Network later on in the next phase. Initial over-fitting was not successful getting stuck usually around 82-87% accuracy. Over-fitting would not be satisfied until near 100% accuracy was reach. Nodes were added to the model in increments of two initially ramping up to five when the former did not change much. Eventually the over-fitting stabilized around 99% with ending statistics of
Accuracy: 99.97%
Precision: 99.82%
Recall: 100.00%
F1-score: 1.00

References

[1] Dileep & Naveen. Logistic regression to predict heart disease. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression.