Sentinus Research Academy
SANO Centre for Computational Personalised Medicine

Heart Disease Predicition using Logistic Regression

Dominik Szablonski

Heart Disease Predicition using Logistic Regression

Dominik Szablonski[1]

[1]Sentinus and SANO

September 14, 2022

Abstract

In this report, I outline my approach to analysing data related to coronary heart disease (CHD), and factors which affect its development. I was able to streamline the data in order to apply logistic regression in order to predict the development of CHD within a subject, while scrutinising the statistical model used for biases and inaccuracies.

Introduction

The World Health Organization (WHO) estimates that in 2019, 17.9 million people died from cardiovascular diseases (CVD), representing 32% of all global deaths, while of 17 million pre-mature deaths due to non-communicable diseases in 2019, 38% were caused by CVDs (1). Further, a WHO report states that the most effective method of preventing, in particular, CHD is through addressing behavioural risk factors, such as diet and smoking (2). This research then aims to be able to predict the development of CHD within a subject through analysing various factors, including behavioural, medical, and demographic, in a logistic regression model.

The following libraries were used during this research:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as st
import statsmodels.api as sm
import sklearn
```

Variables

The dataset provided contains variables of two types: continuous and discrete. Continuous variable types are those who's value may be attained through measurement. There are then two categories of continuous variable data, interval data and ratio data. In our dataset, the continuous variables are represented through interval data where the data is ranked and categorised through precise and continuous intervals, i.e. temperature in degrees Celsius, cholesterol levels, etc. Our discrete variables then also fall into two categories, nominal data and ordinal data. Our dataset is only concerned with nominal data, in which data is fitted into a group with no scale, i.e. colour or other binary values. Much of our nominal data has binary values, which is important within our logistic regression model, particularly with our predicting variable. This is explained further under the section "Logistic Regression".

Below are then listed the variables included within our dataset and the category within which it falls into.

Demographic

- **male:** The sex of the subject. These are binary values, 0 representing female and 1 representing male (Discrete, nominal, binary)
- **age:** The age of the subject, rounded to a whole number. (Continuous, interval)

Behavioural

- **currentSmoker:** If the subject is currently a smoker. (Discrete, nominal, binary)
- **cigsPerDay:** How many cigarettes does the subject consume per day. (Continuous, interval)

Medical (History)

- **BPMeds:** Whether the subject was on blood pressure medication. (Discrete, nominal, binary)
- **prevalentStroke:** Whether or not the subject has experienced stroke before. (Discrete, nominal, binary)
- **prevalentHyp:** Whether or not the subject was previously hypertensive. (Discrete, nominal, binary)

Medical (Current)

- **totChol:** The subject's total cholesterol level. (Continuous, interval)
- **sysBP:** The subject's systolic blood pressure. (Continuous, interval)
- **diaBP:** The subject's diastolic blood pressure. (Continuous, interval)
- **BMI:** The subject's Body Mass index. (Continuous, interval)
- **heartRate:** The subject's heart rate. (Continuous, interval)
- **glucose:** The subject's glucose level. (Continuous, interval)

Predicting Variable

- **TenYearCHD:** 10 year risk of coronary heart disease. This will be the dependant variable within our model. (Discrete, nominal, binary)

Cleaning the Data

Missing Values

Opening the data within the CSV file, we may check for null values as so:

```
data = pd.read_csv("framingham.csv")
print(data.isnull.sum())
```

This leads to the following output:

male	0
age	0
currentSmoker	0
cigsPerDay	29
BPMeds	53
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

These null values only account for a small amount of the data (12%), so we may columns which contain such values through adding the following line

```
data.dropna(axis=0, inplace=True)
```

Checking for Outliers

The next step was to check for outliers with the use of the sklearn library with the code below:

```
fig, ax = plt.subplots(figsize=(10,10), nrows=3, ncols=4)
ax = ax.flatten()

i = 0!
for k,v in data.items():!
    sns.boxplot(y=v, ax=ax[i])
    i+=1!
    if i==12:
        break
plt.tight_layout(pad=1.25, h_pad=0.8, w_pad=0.8)
```

From the boxplots in figure 1, we may identify that outliers are present within the variables totChol, sysBP, diaBP, and BMI, and must be removed.

Initial Analysis

Initial analysis shows us the trend with the subjects within the data. It appears that our data contains more female subjects than male ones, and that most of the subjects are relatively healthy, as seen in Figure 2.

We may then attempt to identify correlation between certain variables with the use of a pair plot in order to identify relations between variables (see Figure 3), as well as by generating a heatmap (see Figure 4).

From both of these figures, we can identify that diastolic blood pressure ad systolic blood pressure are heavily correlated, along with cigarettes per day and current smoker. These variables are highly correlated and are not necessary within our model, so we may remove cigsPerDay and diaBP from our data.

Logistic Regression

Logistic regression is a form of regression analysis, which is a predictive modelling technique which may be used to identify a relationship between

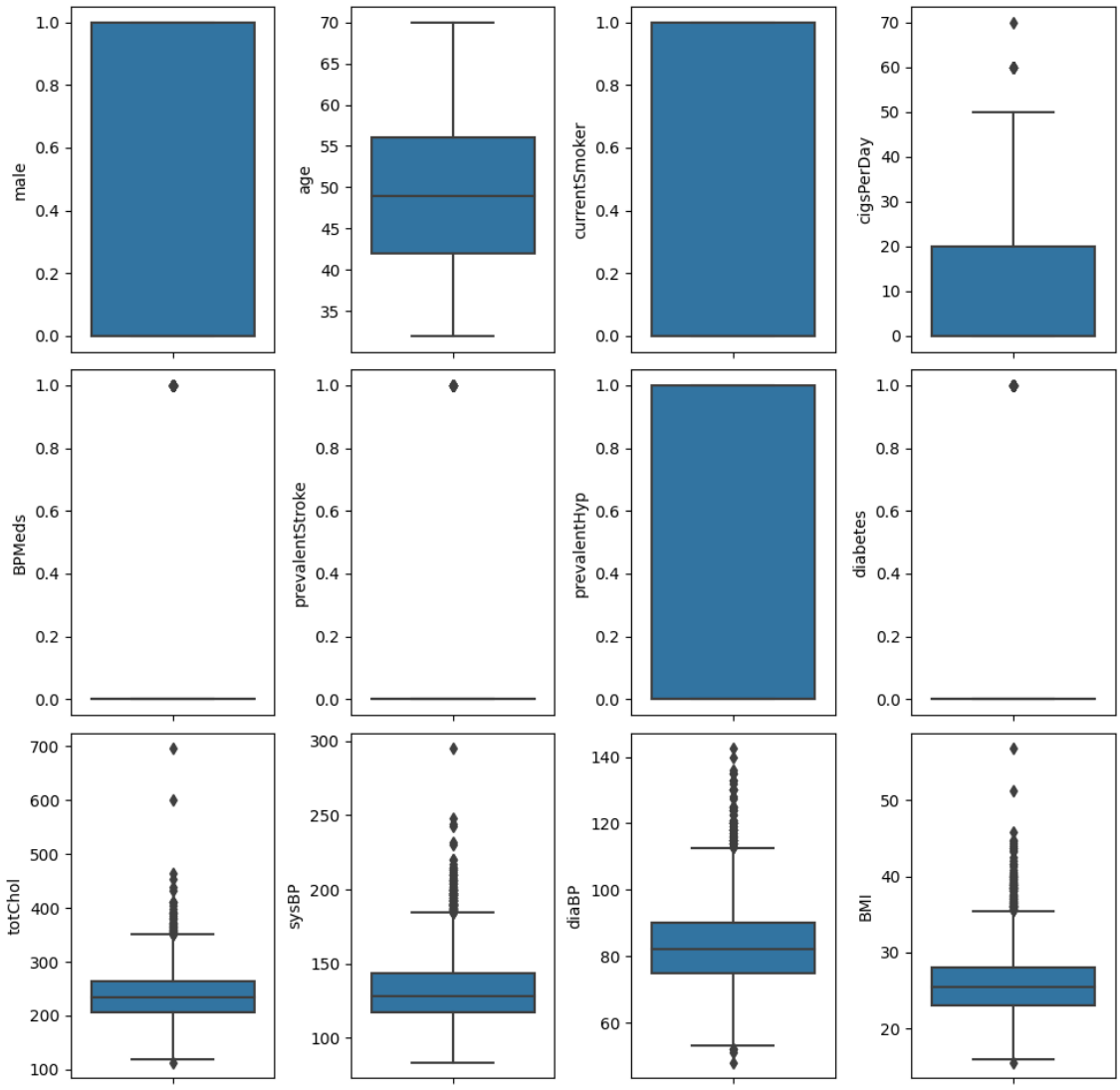

Figure 1: *Set of boxplots to help us identify outliers.*

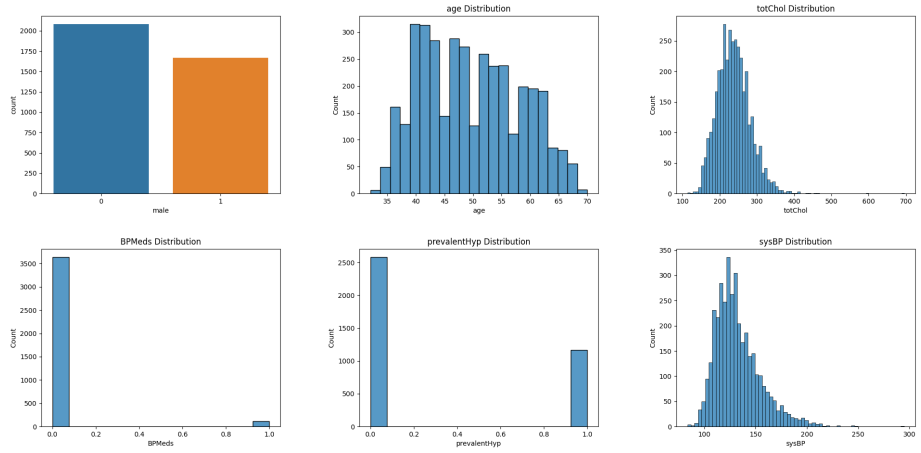

Figure 2: *Histograms of variables within the data. (Left to right; Top: Sex, Age Distribution, Total Cholesterol; Bottom: Blood Pressure Medication, Prevalent Hypertension, Systolic Blood Pressure*

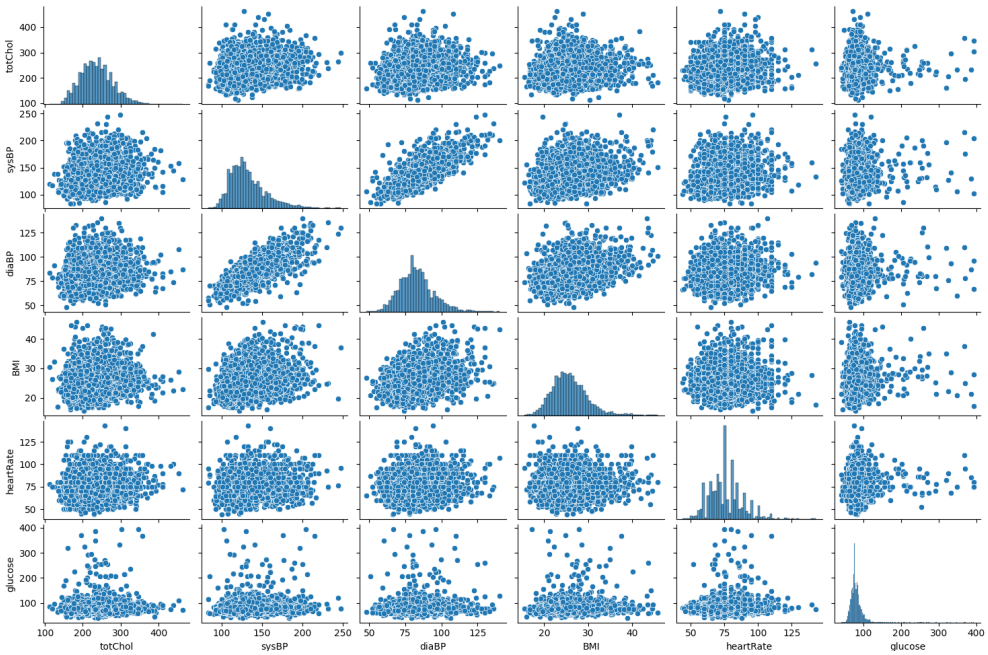

Figure 3: *Pair plot to help identify relations between continuous variables.*

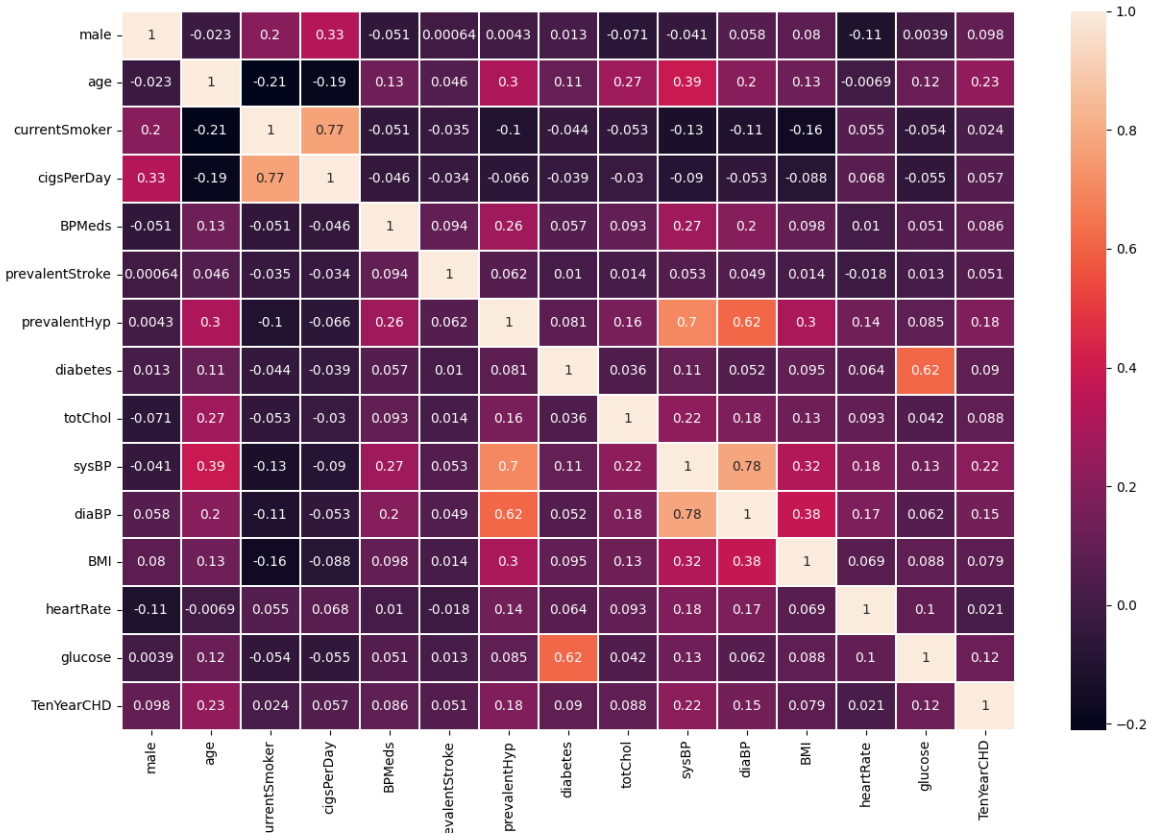

Figure 4: *A heatmap graph used to further identify relations between variables.*

two variables. In the case of logistic regression, this is the prediction of a binary event occurring (in our case, the possibility of developing coronary heart disease within 10 years). The independent variables by which the model is trained on may be either continuous or discrete. Logistic regression follows a rather complex mathematical modal, shown below.

$$p(X) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p}/(1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p})$$

This equation follows where X_j is the j^{th} predictor variable (i.e., cigsPerDay, etc.) and β_j is the estimate for the j^{th} coefficient of the predictor variable. $p(X)$ is then our probability, that within whatever threshold will result in either 1 or 0.

Before continuing with applying our logistic regression, we must first prepare our data further by generating a constant within our data and adjusting the P-values of the data to its ideal value (5%). It will also generate a value for β_j for each variable.

```
from statsmodels.tools import add_constant as add_constant
data_constant = add_constant(data)
st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
cols=data_constant.columns[:-1]
model=sm.Logit(data.TenYearCHD,data_constant[cols])
result=model.fit()
print(result.summary())
```

The result of the code above can be see in Figure 5, many of the P-values above the desired 5%. With variables such as these show a relationship of low statistical significance between a certain variable and the development of heart disease. We must then remove values with a P-value of more than 0.05 by, first, removing the highest P-values, followed by running the regression repeatedly until all P-values are below 0.05. This is demonstrated in the following code.

```
Optimization terminated successfully.
         Current function value: 0.376665
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 3743
Model:                          Logit   Df Residuals:                     3730
Method:                           MLE   Df Model:                           12
Date:                Thu, 08 Sep 2022   Pseudo R-squ.:                  0.1162
Time:                        23:24:23   Log-Likelihood:                -1409.9
converged:                       True   LL-Null:                       -1595.2
Covariance Type:            nonrobust   LLR p-value:                5.806e-72
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             -8.7033      0.657    -13.251      0.000      -9.991      -7.416
male               0.5718      0.107      5.350      0.000       0.362       0.781
age                0.0648      0.006     10.140      0.000       0.052       0.077
cigsPerDay         0.0206      0.004      4.949      0.000       0.012       0.029
BPMeds             0.1804      0.232      0.778      0.437      -0.274       0.635
prevalentStroke    0.8576      0.495      1.733      0.083      -0.112       1.827
prevalentHyp       0.2040      0.136      1.500      0.134      -0.063       0.471
diabetes          -0.0422      0.316     -0.134      0.894      -0.662       0.578
totChol            0.0022      0.001      1.967      0.049    8.39e-06       0.004
sysBP              0.0137      0.003      4.622      0.000       0.008       0.020
BMI                0.0066      0.012      0.525      0.599      -0.018       0.031
heartRate         -0.0022      0.004     -0.538      0.591      -0.010       0.006
glucose            0.0078      0.002      3.496      0.000       0.003       0.012
==============================================================================
```

Figure 5: *The P-values in this output are largely above the desired 5% which must be rectified.*

```
def back_feature_elem (data_frame, dep_var, col_list):
    while len(col_list) > 0 :
        model = sm.Logit(dep_var,data_frame[col_list])
        result = model.fit(disp=0)
        largest_pvalue = round(result.pvalues,3).nlargest(1)
        if largest_pvalue[0] < (0.05):
            return result
            break
        else:
            col_list=col_list.drop(largest_pvalue.index)

result=back_feature_elem(data_constant,data.TenYearCHD,cols)
print(result.summary())
```

The results which follow this are a lot more satisfactory. (See Figure 6.)

Interpreting the Final Results

Final Results For Interpretation:

	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000046	0.000602	0.000166	0.000
male	1.436623	2.184140	1.771379	0.000
age	1.053654	1.080377	1.066932	0.000
cigsPerDay	1.012501	1.029132	1.020783	0.000
BPMeds	0.760072	1.887256	1.197686	0.437
prevalentStroke	0.893858	6.218234	2.357587	0.083
prevalentHyp	0.939413	1.600951	1.226358	0.134
diabetes	0.515800	1.781726	0.958653	0.894
totChol	1.000008	1.004482	1.002242	0.049
sysBP	1.007936	1.019738	1.013820	0.000
BMI	0.982229	1.031556	1.006590	0.599
heartRate	0.989672	1.005929	0.997767	0.591
glucose	1.003438	1.012261	1.007840	0.000

We can draw a lot of conclusions from these results. Males seem to be 77.1% more likely to be diagnosed with coronary heart disease than females, while those with prevalent stroke have a 136% higher chance to be diagnosed with CHD. Surprisingly, we see that glucose and BMI play hardly any role

Optimization terminated successfully.
Current function value: 0.376665
Iterations 7

Logit Regression Results

Dep. Variable:	TenYearCHD	No. Observations:	3743
Model:	Logit	Df Residuals:	3736
Method:	MLE	Df Model:	6
Date:	Thu, 08 Sep 2022	Pseudo R-squ.:	0.1140
Time:	23:36:55	Log-Likelihood:	-1413.4
converged:	True	LL-Null:	-1595.2
Covariance Type:	nonrobust	LLR p-value:	1.854e-75

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.1129	0.474	-19.222	0.000	-10.042	-8.184
male	0.5872	0.106	5.562	0.000	0.380	0.794
age	0.0654	0.006	10.312	0.000	0.053	0.078
cigsPerDay	0.0197	0.004	4.806	0.000	0.012	0.028
totChol	0.0023	0.001	1.996	0.046	4.06e-05	0.004
sysBP	0.0173	0.002	8.007	0.000	0.013	0.022
glucose	0.0075	0.002	4.528	0.000	0.004	0.011

Figure 6: *The features here are a lot more appropriate for our model.*

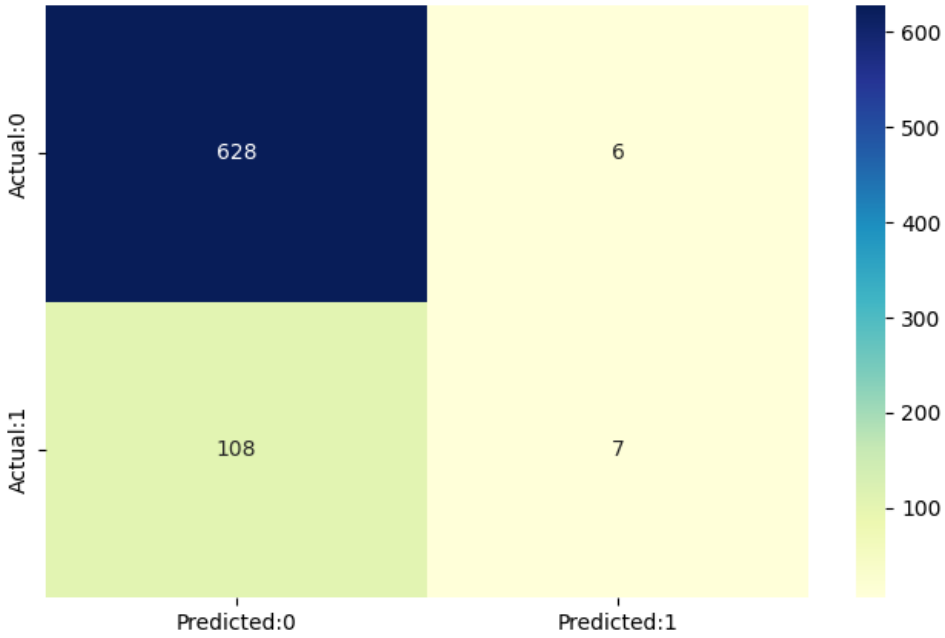

Figure 7: *The model made 634 correct predictions and 115 incorrect ones*

in increasing the chances of potential diagnoses of CHD.

Model Evaluation

However, we cannot assume that our model is 100% correct. Using the `sklearn` library, we may evaluate our model. Through generating a confusion matrix, we can visualise this. (See Figure 7.) Running a metric using `sklearn` has shown us that our model is 84.8% accurate, and in the tests that were run, we received 6 true positives, 628 true negatives, 108 false positives and 7 false negatives. So then, it is clear that our model is flawed in some way, perhaps in the way it identifies positive results. This may be due to our model being highly specific, rather than sensitive, and perhaps also due to the threshold for positive results being too high.

Improving the Model

A method of vastly improving our model is by implementing an AI system which SANO is currently developing. (See Figure 8 and Figure 9). This is a two-system approach based on Daniel Kahneman's theory of "Thinking, Fast and Slow". This proposes two decision making systems, System 1 and System 2. System 1 is the automated, unconscious system of decision making, and System 2 is the cognitive and conscious decision making process. In the case of AI, System 1 takes its role as the seed model, providing a schema for the more advanced System 2 model. System 1 will process data quickly and is able to frame data by presenting the System 2 model with a context of reality. System 2 will then be based on a Statistically Augmented Model. Through being provided a context, it will be able to process data more rigorously and will be able to create knowledge through a dynamic memory system.

We may apply this to our model by potentially applying a model which will first perform the simple, initial regression, then applying a more complex model to identify ways to adjust our threshold values, adjusting the specificity of our model. It could also perhaps indicate to the model whether the P-Values require recalculation, and find algorithms in order to optimise them.

Conclusion

The model outlined in this report has shown the potential for how machine learning and statistical models may allow us to combat non-communicable diseases such as CHD. Our model allowed us to narrow down variables which played significant roles in the development of CHD through analysing their P-values, as well as showing the links between CHD and sex, with males being more likely to be diagnosed with CHD than females, perhaps revealing a bias within our model. It has also led us to a surprising conclusion of how insignificant of a role cholesterol and glucose play in the development of CHD, 0.224% and 0.00784%.

However, our model is not perfect, reaching only 84.8% accuracy in our metric test with many false positives. We may rectify this perhaps by tweaking certain parameters, as mentioned above, or implementing a more complex model. Perhaps something similar to the model which SANO is developing, based on Daniel Kahneman's concept of "Thinking, Fast and Slow".

Whatever the next steps may be, I am hopeful of a future where SANO's

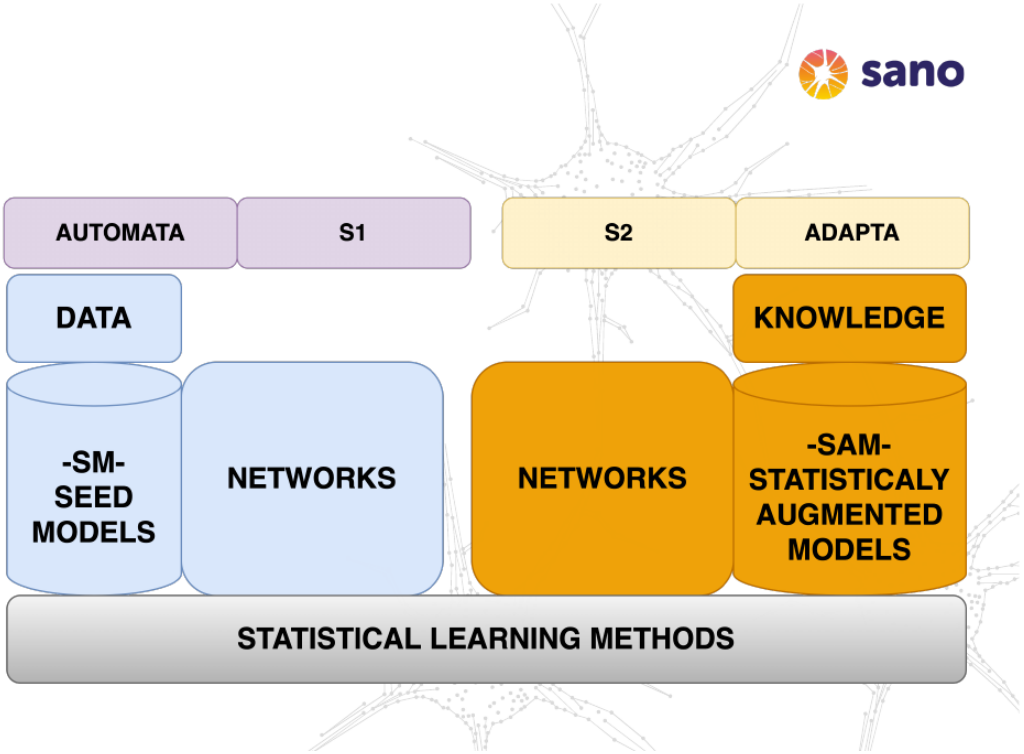

Figure 8: Simple overview of SANO's learning model.

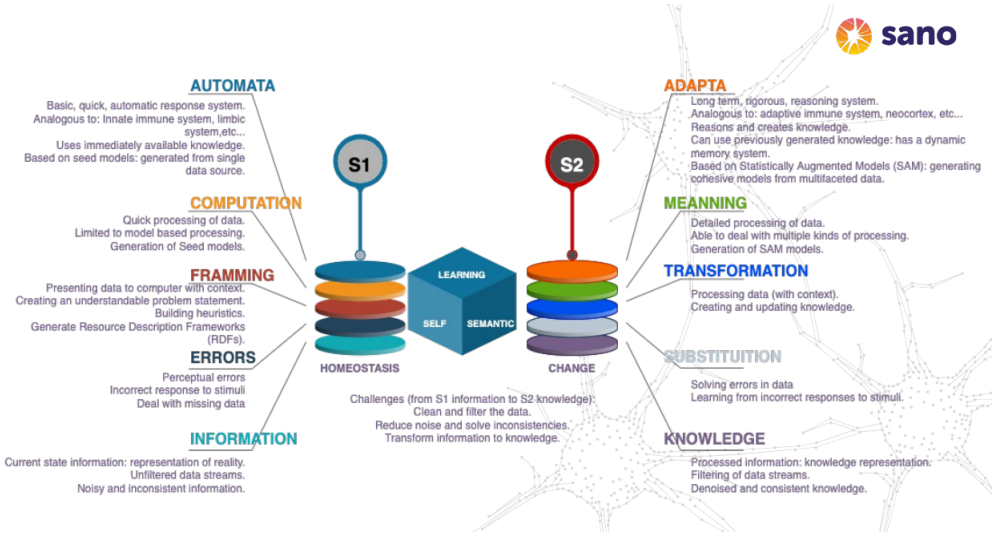

Figure 9: Detailed overview of SANO's learning model.

vision of "Citizen before patient" is realised through new developments within the field of machine learning.

References

[1] World Health Organization, Cardiovascular diseases (CVDs), https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] WHO Expert Committee on Prevention of Coronary Heart Disease & World Health Organization. (1982). Prevention of coronary heart disease : report of a WHO expert committee [meeting held in Geneva from 30 November to 8 December 1981]. World Health Organization. https://apps.who.int/iris/handle/10665/39293