Danmarks
Tekniske
Universitet

Report 2 - Regression and Classification

GROUP 131

Elysia Gao - s222445
Marton Bertalan Limpek - s221835
Katja Naasunnguaq Jagd - s185395

November 15, 2022

Contents

1 Mandatory section

1.1 Contribution

Students	Section covered
Elysia	Regression part A
Katja	Regression part B
Marton (70%), Katja (30%)	Classification
Katja (70%), Elysia (30%)	Discussion and conclusion

1.2 MC questions

Q2 - C

Assuming that each class is evenly distributed at the root

$$x_7 = 2$$

$$N(r) = 135, N(v_1) = 1, N(v_2) = 134$$

$$\text{ClassError}(v) = 1 - \max_c p(c \mid v)$$

$$I_r = 1 - \frac{1}{4} = \frac{3}{4}, I(v_1) = 1 - \frac{1}{1} = 0, I(v_2) = 1 - \frac{\left(\frac{1}{4} \cdot 135\right)}{134}$$

$$\Delta = I_r - \frac{N(v_1)}{N(r)} \cdot I(v_1) - \frac{N(v_2)}{N(r)} \cdot I(v_2) = 0.0074$$

Q3 - A

Each node in the hidden layer receives a vector with the dimension corresponding to the number of attributes, and applies a weight and a bias. Each note in the hidden layer feeds a value to the output layer. The output layer outputs a vector with the dimension corresponding to the number of classes, and a bias is applied to every element of that vector.

$$n_{\text{parameters,hidden}} = 7 \cdot 10 + 1 \cdot 10 = 80$$

$$n_{\text{parameters,output}} = 10 \cdot 4 + 1 \cdot 4 = 44$$

$$n_{\text{parameters,total}} = 80 + 44 = 124$$

Q4 - D

A separates class 3 and 4 from 2: $b_1 \geq -0.76$
C separates class 1 and 3 from 4: $b_1 \geq -0.16$
B separates class 1 from 2: $b_2 \geq 0.03$
D separates class 3 from 1: $b_2 \geq 0.01$

Q6 - B

$$\mathbf{W} = \begin{bmatrix} 1.2 & -2.1 & 3.2 \\ 1.2 & -1.7 & 2.9 \\ 1.3 & -1.1 & 2.2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ -0.6 \\ -16 \end{bmatrix}$$

$$\hat{\mathbf{y}} = \mathbf{W} \cdot \mathbf{b} = \begin{bmatrix} -2.66 \\ -2.42 \\ -1.56 \end{bmatrix}$$

$$P(y = 4 \mid \hat{\mathbf{y}}) = \frac{1}{1 + \mathrm{e}^{-2.66} + \mathrm{e}^{-2.42} + \mathrm{e}^{-1.56}} = 0.73$$

2 Introduction

In this project we will continue to work with the data from "Replication Data for: South African Heart Disease" [1]. The regression problems aims to predict systolic blood pressure from the attributes obesity, tobacco and alcohol, and the classification problem aims to predict the occurrence of heart disease from all 9 attributes.

3 Regression

3.1 Part A

In this section, we would like to use regression to the predict systolic blood pressure (sbp) attribute as a function of the obesity, alcohol, and tobacco continuous variable attributes in order to determine if lifestyle choices correlate well with blood pressure.

We define the variables of our regression problem accordingly:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \text{y} = \text{systolic blood pressure}$$

Consider the simple linear regression model:

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i, \boldsymbol{w}) = \tilde{\boldsymbol{x}}_i^T \boldsymbol{w},$$

The regularized linear regression model can be trained by minimizing the sum-of-squares error term as shown below. λ is defined as the regularization constant in front of the regularization term which allows us to make different models (by adding different degrees of regularization). The most appropriate choice of regularization is then made using cross-validation for model selection.

$$E_\lambda(\boldsymbol{w}, w_0) = \left\| \boldsymbol{y} - w_0 \mathbf{1} - \hat{\boldsymbol{X}} \boldsymbol{w} \right\|^2 + \lambda \|\boldsymbol{w}\|^2, \quad \lambda \geq 0.$$

Before regularization in our regression model, we standardize our data using a feature transformation method which includes subtracting the mean and dividing by the standard deviation. We apply the following standardization to each observation of input features as well as the predicted values (y) :

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_j}{\hat{s}_j}, \quad \mu_j = \frac{1}{N} \sum_{i=1}^{N} X_{kj}, \quad \hat{s}_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \mu_j)^2}$$

We computed the generalization error using 10-fold cross-validation for a lambda range of 10^{-1} to 10^3 for 100 points in order to determine an optimal value for lambda. Figure 1 (left) depicts the mean values of coefficients as a function of the log of the regularization factor (λ). A new data observation for sbp would be predicted according to the linear model with the optimal lambda. Obesity depicts higher mean coefficient values/weights followed by tobacco and alcohol attributes, thus, has a larger affect on sbp prediction (logically also makes sense as obesity should have a larger effect on high blood pressure than tobacco and alcohol consumption). Figure 1 (right) shows the log of the generalization error (squared error of cross validation) as a function of the log of the regularization factor. The optimal lambda is 65.97. The training error curve is also lower than the validation error curve as expected. .

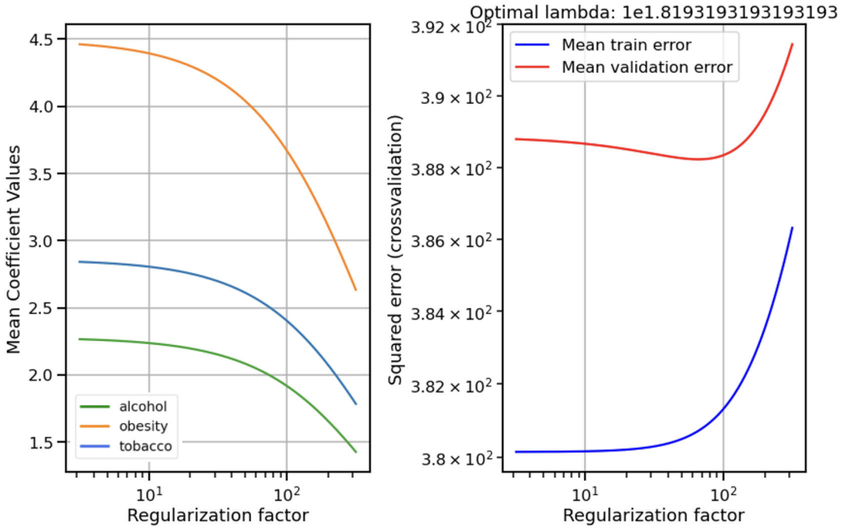

Figure 1: The plots shows results from the regularized linear regression model. In the left plot the mean coefficient values are plotted against the regularization factor λ. In the right plot the squared error is plotted against the regularization factor λ, for both the training and validation data.

3.2 Part B

In this section, three models will be compared. A baseline model, a regularized linear regression model, and an artificial neural network (ANN). The baseline model is simply a linear regression model with no features and thereby predicts y from the mean of the y-values (sbp) of the training data. Two-level cross-validation will be used to answer which of the models are the better choice. The inner loop serves to estimate optimal parameters for the models, and the outer loop serves to estimate the performance of the optimal models which is reflected in the computed generalization errors.

As mentioned, the data has been standardized, and this transformation has been kept throughout this section. This is done because we are applying the same regularization strength to the features, and therefore it is necessary to standardize the data first.

Range of Values for Regularization Parameter λ and Hidden Units h

In the previous section, we found an optimum for the regularization parameter λ that minimized the mean validation error. The range focused on will be the same for the two-level cross validation. The complexity-controlling parameter for the ANN is the number of hidden units h. To select a reasonable range for h, a few test-runs of the training and validation set were performed. Besides running the ANN with different number of hidden units, the number of maximum iterations was also tested. The results for the test runs are shown in Figure 2.

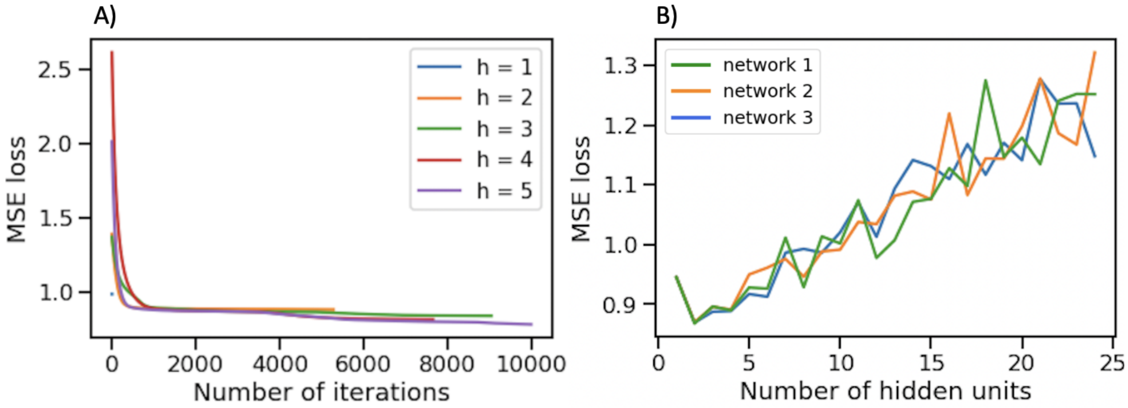

Figure 2: Test-runs with varying parameters h and maximum number of iterations. **A)** Number of iterations plotted against the MSE loss for the training data for $h = 1 - 5$. **B)** h plotted against the MSE for the validation data with number of replicates = 3.

Looking at Figure 2A, the number or iterations is plotted against the mean squared error for the training data for $h = 1 - 5$. All the functions seems to be well converged around 4000 iterations. Some of the functions show a cutoff at different values. This is due to the tolerance criterion for minimum relative change in loss set to 10^{-6}. For $h = 1$ the cutoff already happens after 4 iterations. In Figure 2B, h is plotted against the mean squared error for the validation data. This was done three times, and the three networks seem to follow a similar trend. It can be argued that including more than 5 hidden units for this dataset will only result in overfitting. Taking into account the runtime and chance of overfitting, the selected range of h was chosen to be $h = 1 - 5$, and the max number of iterations = 4000.

Results

Running the two-level cross-validation yielded optimal parameters for the models for the regularized linear regression model and the ANN. The generalization errors E_i^{test} of these models were then generated by evaluating them on the test-set D_i^{test} for each outer fold . The results of the analysis is presented in Table 1. The optimal λ found in section 3.1 $\lambda =$ 65.97 seems to fall in optimal range presented in the table, $\lambda =$ 18.57 - 81.92. By comparing the E_i^{test} of the three models it is hard to tell which one might be performing better than the other. To evaluate the possible difference in performance, a paired t-test was performed for each combination of the three models.

Table 1: Two-level cross-validation results presented as the optimal parameters h and λ, and the generalization error E_i^{test} for the three models. The error measure used is the squared loss per observation.

Outer fold	ANN		Linear regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	1	0.72	47.12	1.67	1.70
2	2	0.73	79.69	0.73	0.83
3	2	0.84	18.57	0.97	0.87
4	1	1.66	51.67	0.75	0.86
5	1	0.84	81.92	0.93	0.88
6	3	1.14	74.02	0.87	0.97
7	4	0.84	66.88	0.98	0.90
8	5	0.92	65.66	0.88	0.93
9	3	0.98	42.38	0.65	0.83
10	1	0.96	44.58	0.79	0.97

Paired t-test

For the paired t-test, the parameters for the regularized linear regression model and the ANN were fixed as $\lambda = 65.97$, $h = 2$ and maximum iterations = 4000. The results from the three models are presented in Table 2.

Table 2: Paired t-test for the three models: Regularized linear regression, ANN and baseline model. Quantities shown for the 3 tests include p-value, confidence-interval CI, width of the CI, and the estimated difference in generalization error $\hat{z}$.

	Linear regression vs. ANN	Linear regression vs. baseline	ANN vs. baseline
P-value	0.25	$6.25 \cdot 10^{-5}$	$8.18 \cdot 10^{-5}$
CI	[-0.011, 0.062]	[-0.148, -0.048]	[-0.197, -0.063]
Width CI	0.073	0.1	0.05
$\hat{\mathbf{z}}$	0.025	-0.098	-0.13

The null hypothesis states that there are no difference between the models, and the test between the regularized linear regression model and the ANN shows that we can not reject the null hypothesis. This is reflected in the confidence interval which includes 0. Since the p-value for this result is higher than the significance level of 0.05, it further supports that we cannot reject the null hypothesis.

When looking at the cases of the two models against the baseline model, both models seem to outperform the baseline model, indicated by the negative $\hat{z}$ value. In both cases the p-value is below the significance level and the confidence interval does not include 0, which indicates that we can reject the null hypothesis.

4 Classification

In this section, we want to predict the binary classification problem of whether a person has heart disease or not. For the predictions, we used all 9 attributes directly without PCA/component reduction. Since heart disease is one of the leading causes of death, the machine learning algorithms for binary classification are extremely useful for early detection of this high-risk disease.

The aim of this exercise is to compare three different classification models (the baseline model, a logistic regression model, and an artificial neural network) with different parameters and to evaluate the best model. To achieve this we used two-level cross-validation algorithm. The errors for each model was calculated in the following way:

$$Error = \frac{Number\ of\ misclassified\ observations}{Number\ of\ test\ examples}$$

4.1 Baseline model

As a baseline model, we used the 'DummyClassifier(strategy="most-frequent")' method from the scikit-learn package with "most-frequent" strategy parameter (always returns the most frequent class label in the observed 'y'). DummyClassifier makes predictions that ignore the input features. This classifier serves as a simple baseline to compare against other more complex classifiers. [2]

4.2 Logistic regression model

We evaluated a logistic regression model using lambda to control the complexity as seen in the previous sections. We also used two-fold cross validation with 10 folds in each outer and inner loop. Here we chose to focus on the range $\lambda = 10^{-1} - 10^{2.2}$ in the logarithmic plot. We also evaluated parameter C, the inverse of the regularization strength of λ. For this logistic regression model, the *saga* solver and *l2* regularization penalty was used. The *l2* regularization penalty means, that $r(w) = \frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^T w$. The C regularization parameter was changed to find an optimal value for C. We had to modify at this model for the C parameter if we wanted to change the strength of the regularization since this was the only option available to change for the method used.

$$\min_{w} C \sum_{i=1}^{n} (-y_i \log(\hat{p}(X_i)) - (1 - y_i) \log(1 - \hat{p}(X_i))) + r(w).$$

Figure 3 is somewhat similar to Figure 1 however this time we are focusing on the classification problem and all of the 9 attributes. Figure 3 (left) shows the mean coefficient values for the attributes, and here, it is clear that a main factor for heart disease is age. The 6 attributes, family history, low density lipoprotein cholesterol, tobacco, type A behavior, sbp and adiposity, follow age and accumulate around the same mean coefficient values . Lastly, comes alcohol and obesity. For Figure 3 (right), we see the same trend as in the previous section where we see a slight increase of the error in the training data around the time a drop in the validation error occurs.

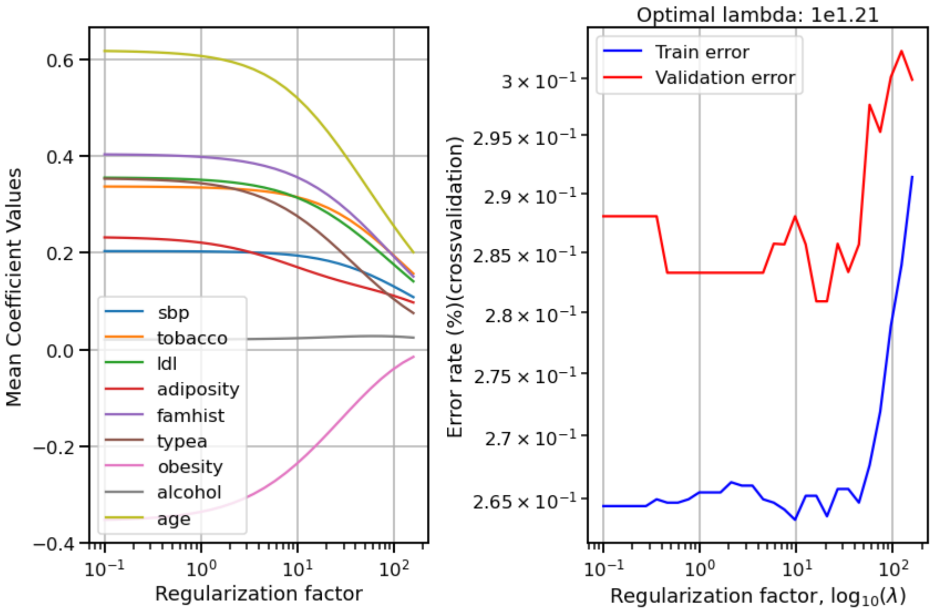

Figure 3: In the left plot the mean coefficient values are plotted against the regularization factor λ. In the right plot the Error rate is plotted against the regularization factor λ, for both the training and validation data.

In Figure 4 we see the results of the mean coefficient values plotted against C instead of λ, for all the attributes except systolic blood pressure sbp, which is a lower in the case with the parameter C.

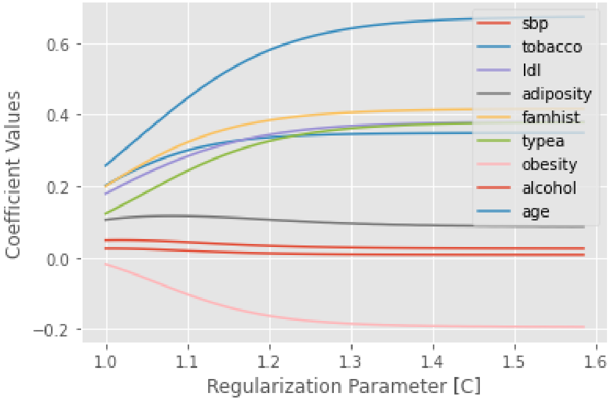

Figure 4: The plots shows results from the logistic regression model. The figure shows the regularization parameter C against the mean coefficient value. C is inverse of the regularization strength λ. Smaller values specify stronger regularization.

4.3 Artificial Neural Network model

To model the ANN in this section, we used the 'MLPClassifier(solver="lbfgs", activation="relu", random-state=1, max-iter=1000, early-stopping=True, hidden-layer-sizes=(complexity-parameter))' method from the scikit-learn package. The ANN has several parameters to adjust, and the chosen values for the parameters can be seen on the parameter list in the function. The ANN in this section will investigate the same complexity parameter as in the previous exercise h, which is the number of the hidden units in the one hidden layer.

4.4 Results

The results of the analysis between the three classification models are presented in table 3. The optimal lambda found in Figure 3, 16.1 lies within the range of λ's presented in the table $\lambda = 7.1 - 33.6$. When running the neural network, we investigated up to $h = 25$, however in the table the optimal h's found only included 2 and 4.

Table 3: Two-level cross-validation results presented as the optimal parameters h, C and λ, and the generalization error E_i^{test} for three models.

Outer fold	ANN		Logistic regression				baseline
i	h_i^*	E_i^{test}	C_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	2	0.32	0.06	0.30	22.8	0.26	0.51
2	2	0.30	0.06	0.28	49.5	0.26	0.30
3	4	0.24	0.04	0.22	7.1	0.26	0.30
4	4	0.24	0.04	0.26	22.8	0.24	0.35
5	4	0.28	0.04	0.37	33.6	0.28	0.35
6	4	0.30	0.04	0.33	10.5	0.28	0.41
7	4	0.22	0.04	0.22	33.6	0.30	0.30
8	4	0.26	0.04	0.22	15.5	0.35	0.33
9	4	0.28	0.04	0.24	22.8	0.30	0.39
10	4	0.30	0.04	0.22	33.6	0.20	0.22

Looking at the generalization errors of the three models, it is difficult to tell which model is better at classifying. However in all cases, ANN and logistic regression performs better than the baseline model which is to be expected. To get a statistical result of the performance between the three classifiers, three correlated t-tests for cross-validation were performed.

Correlated t-test for cross-validation

To compare the accuracy of the three classifiers statistically, three correlated t-tests were performed. Based on the results from table 3, we fixed the parameters as follows: $h = 4$, $\lambda = 16.1$ and maximum iterations = 1000. The results from the three tests are presented in Table 4.

Table 4: Correlated t-test for three models: Logistic regression, ANN and baseline model. Quantities shown for the 3 tests include p-value, confidence-interval CI and $\hat{t}$.

	Logistic regression vs. ANN	Logistic regression vs. baseline	ANN vs. baseline
P-value	0.550	0.089	0.246
$\hat{t}$	0.621	1.904	1.241

When looking at all the p-value, we can see that all of them exceed the significance level of $\alpha = 0.05$. Therefore we cannot reject the null hypothesis in any of the cases, which means that we cannot say there is a statistically significant difference in the performance of the three models.

5 Discussion and conclusion

The aim of the regression problem was to predict the level of systolic blood pressure from a few continuous features, and the aim for the classification problem was to predict the occurrence of heart disease from all 9 features.

For the regression section we chose to focus on the 3 features: alcohol, tobacco and obesity to determine whether these lifestyle attributes have a strong correlation/prediction of sbp. With these small amount of features and also a fairly limited amount of test subjects (462), the complexity of the problem might be limited. By including the baseline model in the cross-validation section and the statistical analysis, the possible simplicity of the regression problem was evaluated. The full analysis of the regression problem was completed by further including a regularized linear regression model and an ANN with tuned parameters.

Before completing the two-level cross-validation, a 10-fold cross validation was done to determine an optimal λ for the regularized linear regression model. Figure 1 shows that a too high of a λ will lead to an overfit model. However, for a certain value of λ, the model will improve as the training error only increases slightly while the validation error drops. To evaluate the appropriate parameter range for the ANN, a few test runs were performed. In figure 2A, we concluded that the functions seemed to be well converged around 4000 iterations. Looking at the curves for $h = 3 - 5$, one could argue that a higher number of iterations should be considered for a higher number of h. The learning curves are computed form the training data, so with the risk of over fitting the validation data was evaluated for a larger range of h, Figure 2B. It is clear that the data only fits a low number of h without generating a high error.

With the chosen range for the parameters, all the models were run through the two-level cross-validation. This was followed by paired t-tests that revealed the performance of the regularized regression model and that the ANN were significantly better than the baseline model. This was somewhat expected as the baseline model merely predicted the mean values of systolic blood pressure measurements.
The regularized regression model and the ANN performance were not found to be significantly different. As ANN's are more computationally demanding, the regularized linear regression model should be considered when trying to predict the systolic blood pressure from these few features.

The parameter found to be most relevant in the linear regression section was obesity followed by tobacco and finally alcohol. In the classification section, the most relevant parameter found for predicting heart disease was age. Comparing the results of the mean coefficients from the linear regression model and from the logistic regression model is somewhat difficult as we investigated two different outcomes. In the regression problem the systolic blood pressure was predicted from 3 chosen attributes. And in the classification problem the prediction of heart disease from all 9 attributes were investigated.
From the statistical analysis in the classification problem, we surprisingly found that we could not reject the null hypothesis in any of the cases. We could have assumed that the logistic regression model and the ANN should outperform the baseline model. But again the results showed no significant difference when classifying. Since the baseline model classifies all cases as the most frequent class, and we know that only 160 out of the 426 are diagnosed with heart disease, it means that the baseline model actually has a rather high accuracy. This provides a possible explanation of how it might perform equally well as the

other models.

Lastly, we compare our results to other machine learning studies done on the same dataset. A previous study done by Khdair and Dasari created a comparative analysis of four different machine learning techniques (Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Multi-layer Perception) [3] in order to predict coronary heart disease in the South African Heart Disease data. We will compare the results of our logistic regression and ANN vs. the logistic regression and multi-layer perception performed by the study. Note that the results for Khdair and Dasari's paper yielded better results as they performed oversampling of the data due to the class imbalance distribution that existed in the original dataset. In the original dataset, the error rate is high since the sample size of 462 is small since there exists a class imbalance distribution. Thus, the previous study included K-means SMOTE to produce minority class samples in safe areas of the input space and to handle the class imbalance distribution. Note that the previous study used a multi-layer neural network (3 layer) which will also provide better results than our 1 layer ANN.

References

[1] "Replication Data for: South African Heart Disease." `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/76SIQD`. Accessed: 2022-09-15.

[2] "sklearn.dummy.DummyClassifier." =https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html. Accessed: 2022-13-11.

[3] H. Khdair, "Exploring machine learning techniques for coronary heart disease prediction," *International Journal of Advanced Computer Science and Applications*, 2021.