Project 2

02450 Introduction to Machine Learning and Data Mining

Danmarks Tekniske Universitet

Group 4

	Section contribution percentage			
Name	Section 1	Section 2	Section 3	Section 4
Astrid Mantzius Jepsen (s214569)	10%	10%	80%	40%
Hoang Anh Do (s250220)	60%	30%	10%	30%
Bálint Kostyál (s250222)	30%	60%	10%	30%

April 8, 2025

Contents

1 Regression

1.1 Regression part A

1.1.1 Regression variables

For our regression analysis, our aim was to predict Type A behavior in men based on various attributes, including blood pressure, cholesterol levels, obesity, smoking habits, and age. Regarding feature transformation, K-coding is not required, since none of our attributes are nominal categories that represent different groups. Although we do have nominal attributes, they are binary indicators that represent the presence or absence. Prior to the utilization of the data matrix X, we standardized the data.

1.1.2 Generalization error calculation

We then introduced a regularization parameter λ and estimated the generalization error for different values of λ using $K = 10$ fold cross-validation. In Figure 1, the left side illustrates how increasing λ causes the coefficients to shrink toward zero, demonstrating the effect of regularization in reducing model complexity. On the right side, we observe that as λ increases, initially, the validation error decreases, reaching an optimal point before rising again due to an incorrect fit. The optimal λ is approximated to be $\lambda \approx 10^{2.24489796}$.

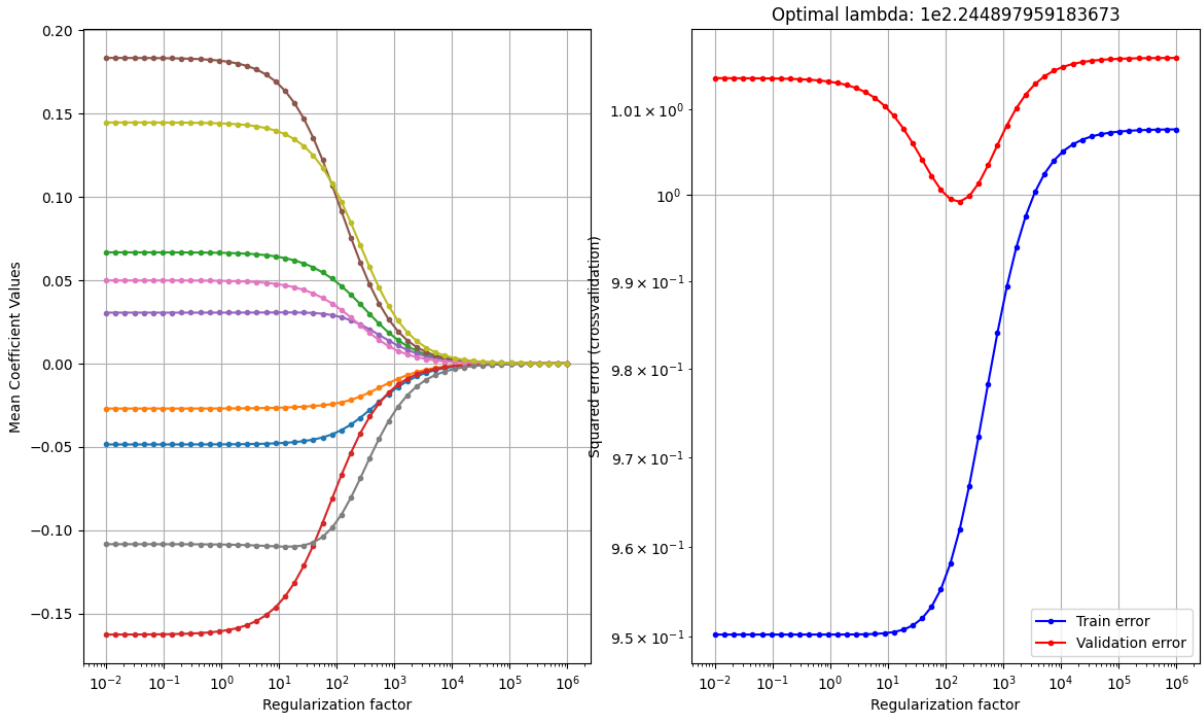

Figure 1: Regularization and mean coefficient values on the left and regularization factor and squared error on the left

1.1.3 Interpreting Output and Attribute Influence in a Linear Model

After training the linear model, the weights and the offset are presented in the Table 1.

	Offset	Sbp	Tobacco	Ldl	Adiposity	Famhist	Obesity	Alcohol	Age	Chd
Weight:	53.44	-0.28	0.03	0.27	-0.71	0.48	0.97	0.59	-1.05	1.23

Table 1: Table of weights of attributes and offset

Thus, Type-A behavior score can calculated as follows:

$$\begin{aligned} y = 53.44 &+ (-0.28 \cdot \text{Sbp}) + (0.03 \cdot \text{Tobacco}) \\ &+ (0.27 \cdot \text{Ldl}) + (-0.71 \cdot \text{Adiposity}) \\ &+ (0.48 \cdot \text{Famhist}) + (0.97 \cdot \text{Obesity}) \\ &+ (0.59 \cdot \text{Alcohol}) + (-1.05 \cdot \text{Age}) \\ &+ (1.23 \cdot \text{Chd}) \end{aligned}$$

Chd (1.23) has the strongest positive effect, indicating that individuals with coronary heart disease are more likely to exhibit Type-A behavior. Age (-1.05) shows a strong negative effect, suggesting that younger individuals are more likely to exhibit Type-A behavior. Obesity (0.97) and alcohol (0.59) positively contribute to Type A behavior, which is consistent with research on stress-related behaviors [1]. Adiposity (-0.71) and Sbp (-0.28) have negative effects, implying that lower body fat and blood pressure are more closely associated with Type-A behavior. These relationships align with existing research, where younger individuals and those with coronary heart disease (CHD) are often found to exhibit more Type-A characteristics [2].

1.2 Regression part B

1.2.1 Two-Level Cross-Validation Implementation

We fitted an ANN to the dataset and used the number of hidden units h as the complexity-controlling parameter. After a few test runs, we found that the optimal number of hidden units was 26 using 10-fold cross-validation Figure 2. Note that changing the number of hidden units does not affect the learning capabilities of the network too much. This could be either because the learning objective is achievable with one neuron or because the network does not converge. We could rule out the latter as we can see on Figure 4 that the network is capable of converging with 10000 learning steps. Based on these conclusions, we chose a reasonable range of values for h from 1 to 30, as the mean squared error showed no significant improvement beyond 30 hidden units.

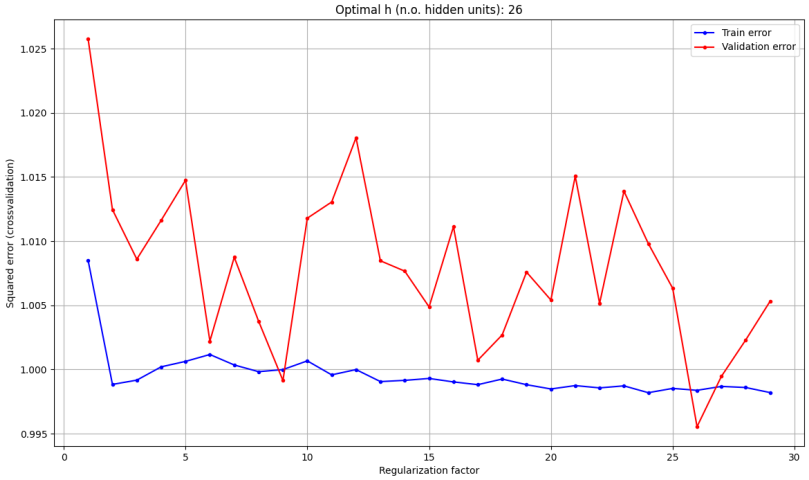

Figure 2: Optimal number of hidden units to use for the ANN

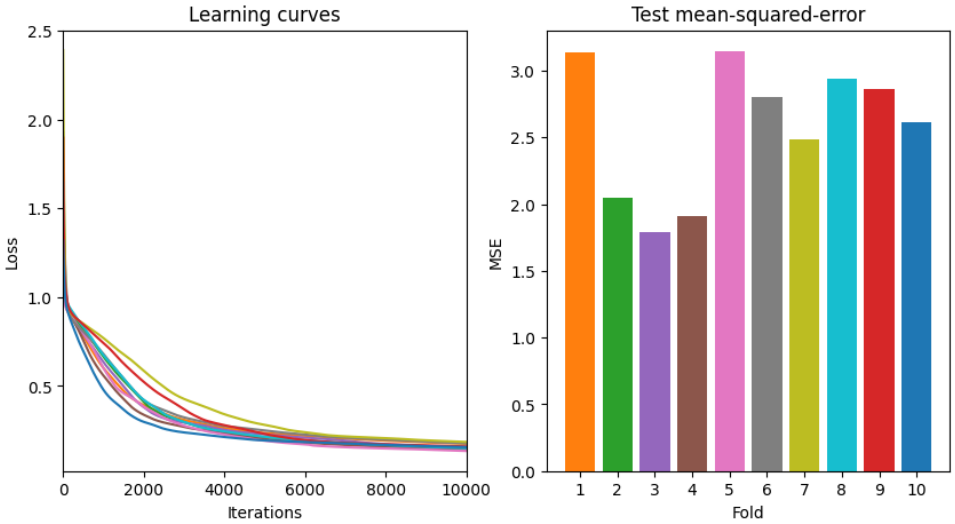

Figure 3: Learning curves of the best performing ANN models for each fold run

For the linear regression model, we chose the regularization parameter λ to range between 10^0 and 10^4, as previous testing on a broader range Figure 1 indicated that the minimum error lies within this interval.

We also implemented a baseline model, which is a linear regression model without any features—that is, it simply predicts the mean of y from the training data.

We then created Table 2 using two-level cross-validation with $K_1 = K_2 = 10$.

Outer fold	**ANN**		**Linear regression**		**baseline**
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	9.0	0.982	78.48	1.095	1.112
2	20.0	0.997	78.48	1.015	0.998
3	5.0	1.024	127.43	0.739	0.748
4	6.0	1.000	29.76	0.989	0.948
5	17.0	0.991	78.48	0.991	1.058
6	23.0	0.988	127.43	1.027	1.054
7	13.0	1.001	335.98	0.890	0.917
8	10.0	0.978	127.43	1.140	1.186
9	14.0	0.982	78.48	1.100	1.173
10	11.0	1.013	78.48	0.791	0.838

Table 2: Table with the values of optimal λ and h for outer-fold, and the result prediction errors of each of the three models.

The table shows, for each of the $K_1 = 10$ outer folds i, the optimal number of hidden units and regularization strength (h_i^* and λ_i^*, respectively) as determined from the inner loop. It also reports the estimated generalization errors E_i^{test}, computed by evaluating on $\mathcal{D}_i^{\text{test}}$, as well as the baseline test error on the same test set. We calculate the error as the following:

$$E_i^{\text{test}} = \frac{1}{N^{test}} \sum_{i=1}^{N^{test}} (y_i - \hat{y}_i)^2 \tag{1}$$

Overall, the three models performed quite similarly, each achieving an error of around 1. The number of hidden units h varied between 10 and 24. Among the models, the ANN performed the best. The error remained consistent across the folds, which could indicate that it performs similarly with different numbers of hidden units—consistent with our earlier observation.

For linear regression, the error decreased from 1.5 to 0.8, but rose to 1.3 with the 10^{th} iteration. This is likely due to overfitting that occurs in the later folds.

1.2.2 Comparison of models

We then statistically evaluated whether there is a significant performance difference between the fitted ANN, the linear regression model, and the baseline. For this, we used Setup I and applied the paired t-test with a confidence level of $\alpha = 0.05$.

	Comparison	Mean Difference	Confidence Interval Min	Confidence Interval Max	p-value
0	ANN vs LR	0.011330	-0.005091	0.027750	0.175793
1	ANN vs Baseline	0.003649	-0.000059	0.007357	0.053741
2	LR vs Baseline	-0.007681	-0.023887	0.008526	0.352174

Table 3: T-testing results for the three models

The mean differences were very close to zero, suggesting that the models perform similarly. This is consistent with the small differences in validation errors shown in Table 2. Since the p-value is greater

than 0.05 for all comparisons, we can conclude that there are no statistically significant differences between the models—though there may be a potential difference between the ANN and the baseline, as the p-value, in that case, is close to the 0.05 threshold. Additionally, because the confidence intervals include zero, we can again conclude that there may be no real difference in performance between the models. The ANN might be slightly better than the other, but we do not have strong enough evidence to confirm that. In Figure 4, the p-values with their corresponding confidence intervals are illustrated:

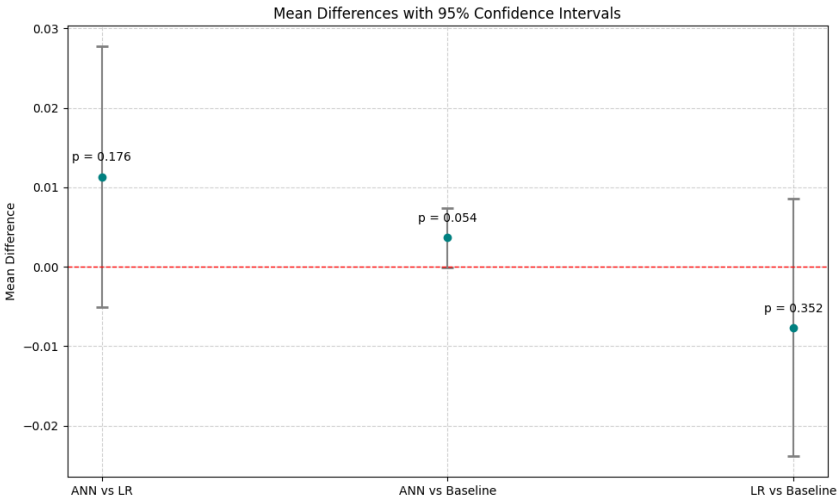

Figure 4: The mean differences between models, with error bars showing the 95% confidence intervals. The dashed red line at 0 indicates "no difference."

2 Classification

The chosen classification problem focusing on detecting males with coronary heart disease (CHD). This is a binary classification task, where males with CHD are labeled as 1, and others as 0. Out of a total of 462 observations in this dataset, 160 males are diagnosed with CHD.

Classification is performed using three different models: logistic regression, K-nearest neighbors (KNN), and a baseline model. For logistic regression, a regularization parameter is applied to control the complexity of the model. The inner-fold of the cross-validation is used to determine the optimal regularization parameter for each training set. This optimal value is then used in the outer-fold to train the model and evaluate performance on the test data. The optimal value of λ is searched within the range $\lambda = [1, 50]$, using 30 steps. This range was determined based on a 10-fold cross-validation trial run.

For the K-nearest neighbors model, cosine distance is chosen as the distance measure, as it is good for high-dimensional data and correlation between the many features. The optimal number of neighbors K is also determined in the same inner-fold, with values searched in the range $K = [10, 50]$.

The baseline model predicts all test samples as belonging to the majority class in the corresponding training fold of each cross-validation split.

The training and testing of the models is performed via a two-level cross-validation. Prior to using the dataset, all features are standardized in the same manner as done for the regression model in subsubsection 1.1.1.

2.1 Two-Level Cross-validation

For each outer-fold (i), the error rate of the predicted test data is found for each of the three models as:

$$E_i^{\text{test}} = \frac{\{\text{Number of misclassified observations}\}}{N_i^{\text{test}}} \tag{2}$$

The determined values of λ_i and K_i for each outer-fold, and the belonging error rates of the three models are listed in Table 4.

Outer fold	KNN		Logistic Regression		Baseline
i	K_i	E_i^{test}	λ_i	E_i^{test}	E_i^{test}
1	35	38.3	13	29.8	51.1
2	49	31.9	23	29.8	29.8
3	31	30.4	11	23.9	30.4
4	38	26.1	3	28.3	34.8
5	36	30.4	16	37.0	34.8
6	42	30.4	8	30.4	41.3
7	40	19.6	8	21.7	30.4
8	28	21.7	25	21.7	32.6
9	24	30.4	28	23.9	39.1
10	39	23.9	20	21.7	21.7

Table 4: Table with the determined values of optimal K_i and λ_i for outer-fold, and the result prediction errors of each of the three models.

Overall, the determined error rates are quite high for all three models. The K-nearest neighbor model gave error rates in the range of approximately 20% to 38%. The optimal K-value for each outer fold is found to be quite high, ranging from $K = [24, 49]$. Thus, a large number of neighbors are necessary to determine whether a male is diagnosed with coronary heart disease (CHD) or not.

The error rates for the logistic regression model are approximately the same, ranging from around 21% to 37% at the highest. The range of the optimal λ_i for each outer fold is from 3 to 28, with most values found in the upper part of the range. This indicates that the data observations may have high noisy variance, which can be reduced by introducing a large λ, resulting in a lower testing error.

For the baseline model, most error rates are between 30% and 40%, however, there is also an error rate of 21.7% in the last outer fold and 51% in the first.

The relatively high error rates for the three models may be due to the fact that the dataset is highly unbalanced, with a significant overweight of males who are not diagnosed with CHD ($y = 0$). This could lead to an overrepresentation of this class in each iteration, which especially affects the predictions made by the baseline and K-nearest neighbor models. Another reason for the high error rates could be the relatively small number of observations (462). When the dataset is split into training and test sets, the amount of data used to train the models is quite low.

A potential solution could be oversampling of the dataset to reduce the imbalance. Another approach would be to use a model more suitable for unbalanced datasets, such as random forest. This method divides the dataset into different subsets with different available features. Then, the model makes predictions using Hunt's algorithm for each subset and combines the predictions to select the final outcome.

2.1.1 Comparison of models

To determine which of the models is best, a statistical comparison of the models is performed using McNemar's test. In Python, the function `mcnemar` from `dtuimldmtools` is applied. The Null-hypothesis used for the statistical analysis is that the predictions made by the two classification models are the same[1].

$$H_0 = \text{Model A has same performance as model B}$$

The comparison is done with a confidence interval of 95% ($\alpha = 0.05$), and the resulting intervals and p-values are presented in Table 5, along with the determined values of θ for each comparison. The McNemar

	p	Confidence Interval	θ
KNN vs. Logistic Regression	0.382	[-0.0442, 0.0139]	-0.0151
KNN vs. Baseline	0.0113	[0.0161, 0.109]	0.0618
Baseline vs. Logistic Regression	0.00107	[-0.123, -0.0328]	-0.0779

Table 5: McNemars test for comparison of the three models.

test between the K-nearest neighbor and the logistic regression shows a value of $p = 0.382$, which is above the confidence level at $\alpha = 0.05$. Additionally, with a confidence interval including 0, it is not possible to reject the formulated null hypothesis and conclude whether the KNN model performs differently from the logistic regression model. However, the determined $\theta = -0.0151$ is negative, indicating that the logistic regression model might perform slightly better than the KNN model.

The comparison between the KNN and the baseline model gives $p = 0.0113$, which shows that the null hypothesis can be rejected, implying that there is a significant difference between the models. This is

[1] From the book: Introduction to Machine Learning and Data Mining

also supported by the confidence interval $CI = [0.0161, 0.109]$, which excludes 0. Together with the determined $\theta = \theta_{KNN} - \theta_{baseline} = 0.0618 > 0$, it confirms that the KNN model performs better than the baseline model.

The p-value found for the comparison between the predictions made by the baseline model and the logistic regression model is even lower, $p = 0.00107$. This strongly implies that the baseline model does not have the same performance as the logistic regression model. The confidence interval is negative, $CI = [-0.123, -0.0328]$, indicating that the logistic regression model performs better. The determined $\theta = -0.0779 < 0$ is also negative, which confirms this conclusion.

Another way to statistically compare the models could be by calculating the precision and recall. These error rates determine the number of predicted positive CHD cases and how many of them are correctly identified. Thus, they are more applicable for datasets that are highly unbalanced.

2.2 Logistic Regression

A logistic regression model is fitted to a training dataset via the "hold-out" method of 1/3 being the test dataset. The model is fitted with a regularization parameter of $\lambda = 20$, which is chosen from the performances in Table 4. Prior to the fitting, a column of ones is added to determine the intercept of the regression model. The obtained weights of each feature are found as: The weight $w_0 = -6.7 \cdot 10^{-5}$ is

	Offset	Sbp	Tobacco	Ldl	Adiposity	Famhist	typea	Obesity	Alcohol	Age
Weight:	$-6.7 \cdot 10^{-5}$	0.086	0.25	0.25	0.10	0.23	0.17	-0.088	-0.079	0.45

Table 6: Table of weights of attributes and offset for classification of coronary heart disease.

very close to zero, indicating that the logistic regression model is likely to classify any male as not having coronary heart disease (CHD) by default. This is due to the heavy class imbalance. Both the obesity and alcohol features have negative weights, meaning that higher values for these features contribute to a lower model output, thereby making it more likely that the male is classified as not having CHD. However, these negative weights are relatively small compared to the positive weight of age. The blood pressure feature (Sbp) also has a very low positive weight, suggesting that alcohol, obesity, and Sbp contribute little to the prediction of whether a male has CHD.

As mentioned, age has a high positive weight, meaning that the older the male is, the more likely he is to be diagnosed with CHD. Other features such as tobacco consumption, cholesterol level (ldl), and family history (famhist) also have positive weights, implying that higher values for these features increase the likelihood of a CHD diagnosis.

If these weights are compared to those derived in Table 1 from the regression part a, it is seen that they differ. For instance, Sbp, has a negative influence on Type A behaviour, whereas it has a positive weight in the classification of coronary heart disease. Similar differences in sign are observed for adiposity, obesity, alcohol consumption, and age. The only similar weight is seen for cholesterol, indicating that a high LDL level is associated with a higher Type A score and an increased likelihood of being diagnosed with CHD. The weight of the Type A score in the classification is only 0.17, meaning that a Type A score will increase the risk of being diagnosed with CHD. Oppositely, the weight of CHD in the regression section is 1.23. This indicates that having coronary heart disease increases the Type A score by 1.23.

3 Discussion

For the regression task, the linear regression model, the ANN, and the baseline model all performed with relatively low errors. Since the dataset was not explicitly focused on Type A behavior, there was no imbalance in the representation of Type A individuals during data collection. In addition, the data set included attributes closely related to Type A behavior [2] [1], which made it easier to predict it.

Using the t-test, we found no significant differences between the three models. The ANN model may perform slightly better than the others, but the difference is minimal based on our comparison.

In the classification task, the KNN model, the logistic regression model, and the baseline model all performed relatively poorly. But overall, the McNemars test showed that both the KNN and logistic regression performed better than the baseline, however it was not clearly possible to conclude wether of these two models performed best despite form looking at the θ, which was negative in Table 5, thus indicating that the logistic regression might perform a bit better.

From the logistic regression performed in the classification part, it was seen that the features of age, tobacco consumption, and cholesterol level had a high positive influence on the probability of being diagnosed with CHD. Whereas, the age had a highly negative influence on the Type A-score, derived in the regression part a. Except for the weight of LDL, no weights were similar for these two types of regressions.

The observed high error rates in the classification may result from the dataset being highly imbalanced, with a significant overrepresentation of males not diagnosed with CHD. For this reason, oversampling may be relevant to create a more balanced dataset and to provide additional data for training, as the dataset contains only 462 observations.

Additionally, it may be more appropriate to evaluate the performance of the models using other metrics such as precision, recall, and accuracy. It could also be beneficial to apply a different classification model that is better suited for imbalanced datasets and capable of yielding a lower error rate. One possible model is the random forest, which separates the dataset by features and then combines the predictions of individual trees to generate a final prediction.

Given these conclusions, this dataset may be more suitable for regression problems than for classification tasks, depending on the chosen models.

3.1 Comparison to another study

In 2021, Khdair and Dasari examined the same dataset in [1]. They applied it to the same classification problem, where the objective was to determine whether males were diagnosed with coronary heart disease (CHD). Four models were used: Support Vector Machine (SVM), K-nearest neighbors (KNN), logistic regression, and MLP neural networks. For the KNN method, they used the "Minkowski" distance measure with $K = 17$, and for the logistic regression, they used $C = 0.25 \rightarrow \lambda = 4$ as the regularization parameter.

Prior to training the models, they applied different methods such as the ANOVA method and Permutation Feature Importance using random forest as the model to evaluate the importance of the features to the coronary heart disease. They found that features like age, tobacco consumption, cholesterol level (ldl), blood pressure (Sbp), adiposity, and family history (famhist) were most important. Comparing these findings to the weights found in the logistic regression in Table 6, it is also shown from the magnitude of the weights that the same features, except for Sbp, contributed the most to the logistic regression. But, unlike the study, our results indicate that type A behavior contributes more to the diagnosis than blood

pressure.

Additionally, they used Pearson's correlation method to assess feature correlations and discovered that obesity was highly correlated with adiposity. As a result, the obesity feature was removed from the dataset.

Following that, the models were trained and fitted, and their performances were evaluated using accuracy, F1 score, precision, recall, and specificity. They used 10-fold cross-validation and found that KNN performed best in terms of precision, with 0.7 wrong predictions. Logistic regression had a precision of 0.633. When evaluating recall, logistic regression performed best but with only 0.506, while KNN had a recall of 0.394. For comparison, the precision and recall for KNN and logistic regression predictions can be found in subsection 2.1 as[2]: These results show that the predictions in this project are slightly

	Precision	Recall
KNN	0.62	0.47
Logistic Regression	0.66	0.47

Table 7: Precision and Recall for predictions made for classification in subsection 2.1.

less precise compared to the study in [1]. This difference might be due to the fact that they used a fixed value of $\lambda = 4$, which is much lower than the one chosen in the two-level cross-validation. The same can be said about the K values we chose, which are at least 24 but mostly above 30.

Overall, Khdair and Dasari attributed the high error rate to the highly imbalanced dataset. To solve this, they tried to improve the number of positive CHD cases by oversampling. They then trained and tested the same models on the oversampled dataset, finding that the overall prediction error was reduced. KNN now showed a precision of 0.8, and logistic regression had a precision of 0.755. The recall values also increased significantly, reaching 0.737 for KNN and 0.795 for logistic regression.

Thus, in this project, reducing the imbalance between the CHD classes should help lower the error rates identified in subsection 2.1. Additionally, evaluating the correlations between the features before fitting and validating the models might have been a good choice to reduce the complexity of the models.

[2]Expression for precision and recall is found in the book: Introduction to Machine Learning and Data Mining

References

[1] Naga M. Khdair Hisham; Dasari. "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction". English. In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 12.5 (2021). Also available as ISSN: 2156-5570, pp. 28–36. ISSN: 2158-107X. DOI: 10.14569/IJACSA.2021.0120505.

[2] R. H. Rosenman and M. A. Chesney. "The relationship of type A behavior pattern to coronary heart disease". In: *Acta Nervosa Superior (Praha)* 22.1 (Mar. 1980), pp. 1–45.