Predictive Analytics
(INFS 5100)

Assignment 2

Mahmoud Ghazi

Table of Contents

Introduction and Recap

Our team conducted a literature review and data analysis regarding the heart disease dataset to predict target values. Initially, related studies from the past were reviewed, and brief summaries were given, which aimed to provide more domain knowledge in diagnosing heart disease and steer the direction of research in this paper, which mostly was done by my teammate. Next, I describe the pre-processing to conduct data exploration and understand the dataset, such as what types of data there is and the pattern of the data through visualizing them in charts and graphs as well as performing statistical analysis to handle missing data, removing outlier and noise and discretization of the age attribute.

I explained the features I decided to extract from the dataset consisting of 13 independent variables and provided descriptive statistics, including value distributions, skewness, histograms, bar plots, and box plots. I decided to choose features that have a significant difference in their distribution concerning the target variable as attributes that can contribute more to the model. Besides the univariate and the correlation between variables analysis, I considered feature selection according to the score obtained from the mlr3 library. According to all investigations, I concluded to drop two features (fbs, trestbps) and keep the rest of the 11 variables for the training decision tree model.

To evaluate the decision tree model behaviour, I divide the dataset into train and test. I assumed considering 20% for a test could be a good ratio for representing most of the variance in the dataset, and the remaining 80% is a good chunk for the training mode. I fit and explained three models by applying various parameters such as maximum depth (maxdepth), the minimum number of samples a node must have before it can split (minsplit), and complexity parameter (cp). I compared models by performance metrics which the decision tree model trained with default parameters of cp, maxdepth and minsplit shows higher accuracy, recall, F1-score, and kappa compared with other models.

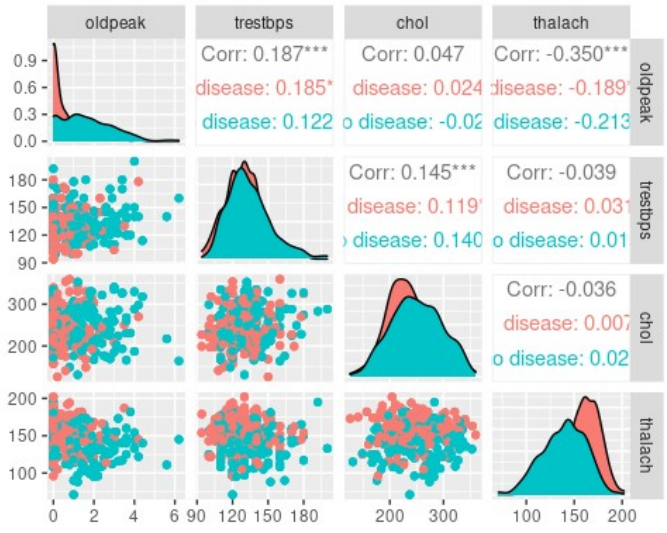

Figure 1: correlation between numerical variables

Data exploration and Feature Selection

This graphic in figure 2 shows the Pearson correlation between numerical variables and the distribution of each variable regarding the categorical target variable. As can be seen, there is no strong relationship between numerical variables, which infers dropping one variable similar to the other. The plots imply a significant difference in the distribution in oldpeak, thalach and major_vessele regarding the target variable (disease, no-disease). Chol shows a slight difference in distribution, and Trestbps distribution difference between disease and no-disease are almost the same. Furthermore, a boxplot is an excellent approach to depict the difference in mean, median and interquartile of a numerical attribute concerning with categorical attribute.

Figure 2 implies that all the numerical variables have outliers as some points fall far away from whiskers. Also, outliers can be identified by values adopted with a z-score of more than three, which implies it deviated significantly from the majority instance. Furthermore, as figure 2 depict, there is a different range of value for numerical variables that a normalization method has been applied to decrease the effect of irregularity on the model.

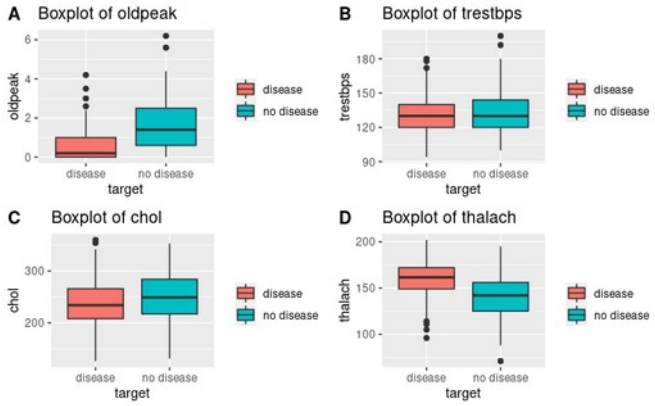

Figure 2: numerical variables

I selected attributes for the classification model by considering three stages:

1- Choose attributes that have a significant distribution difference concerning the categorical target variable. Chol, Trestbps from numerical variable and fbs from categorical variables were the candidate to drop as their distribution regarding the target variable were not significantly different.

2- Along with the univariate, the correlation between variables was considered and selected one out of two variables had a strong correlation (I could not find any proper pair variable with a strong correlation to drop at this stage).

3- I considered feature selection according to the score obtained from the mlr3 library. It filters features from all types, including numerical and categorical variables, based on the implemented filter (information_gain was considered). Trestbps and fbs obtained the minimum score among all variables.

According to all analyses, I decided to drop two candidate features of fbs, trestbps as two steps of feature selection analysis voted to remove. I trained the classification model based on the 11 variables.

Building Classification Models

For this study, four machine learning classifiers, including Support Vector Machine (SVM), Neural Networks, K Nearest Neighbours and Decision tree, have been employed to classify the heart disease data. Each classification has been examined with various parameters and the model with highest performance metrics has been chosen as a candidate.

Support Vector Machine (SVM)

The objective of SVM is to find an optimal hyperplane to divide classes with the maximum margin. Different kernel, which SVM uses in the case of non-linearly separable data points, transform data into higher dimensions. In this study, Linear, Polynomial, RBF kernels with various optimization options have been examined. The setting for the three models come as follow:

- Having specified the kernel as linear, I determine the strength of the regularization parameter 'Cost'. A range of values including 0.001, 0.01, 0.1, 1, 5, 10 has been set for cost parameter.
- A polynomial kernel needs parameters such as degree, which is the degree of the polynomial kernel function, and Coef0 as an independent term in 'poly' and 'sigmoid' kernel function to be tuned. A range of values including 3,4,5 for degree and 0.001, 0.01, 0.1, 1, 5, 10 for coef0 has been considered.
- The RBF (radial) kernel needs degree and gamma as the parameter. Gamma is a Kernel coefficient for 'rbf', 'poly' and 'sigmoid'. For this model values including 0.001, 0.1, 0.5, 1, 5, 10 has been considered for gamma.

According to the performance metrics, among the various kernels, a polynomial kernel is selected. The order of degree and coefficient have been considered to tune the parameters in the model, and a degree 5 polynomial kernel with coefficient 5 is the best parameter for the classification model to predict disease and no disease in the test data set. Table 1 shows the performance metrics and figure 3 shows the AUC plot for the selected model.

Accuracy	0.98
Kappa	0.97
Sensitivity	0.97
Specificity	0.99
AUC	0.98

Table 1: performance metrics

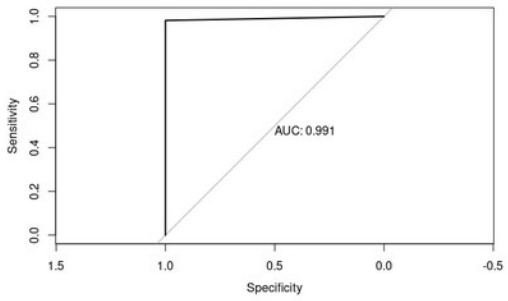

Figure 3: AUC plot

Neural Networks

Neural networks use complex patterns in datasets using multiple hidden layers and non-linear activation functions to predict unseen data. In neural networks, it is a good idea to normalize/scale data as it uses the activation function between -1 and 1. Furthermore, it helps faster approaching the global minimum (minimizing the loss function).
I apply min-max normalization for numerical variables and a dummy method for categorical variables as neural network input. Several models with different hyper-parameters have been examined, which come as follows.

- Hidden layer and neurons transform the input as non-linear into the network after several trials and errors. The more hidden layers and neurons we apply, the more complex function for classifying data we have. Five neurons for first layer and 3 neurons for the second layer is the best setting for the best model. Figure 4 shows the layers and neurons.

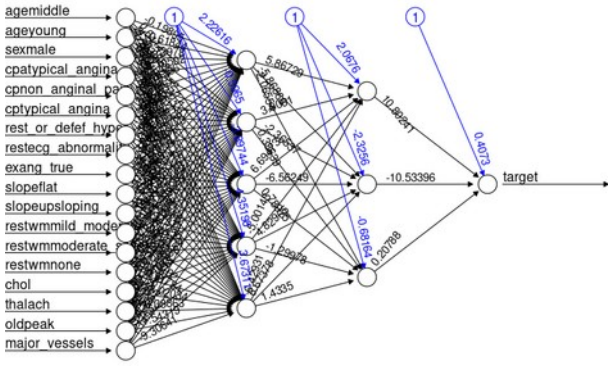

Figure 4: Neural Network for selected model

- Threshold: The threshold is set to 0.01, which means that no further optimization will be carried out if the change in error during an iteration is less than 1%.

- The learning rate(alpha) for the back propagation algorithm needs to be set small enough to avoid running into an error. The order of values has been examined for the alpha, and 0.01 is considered as the best one, which leads to convergence and represents suitable performance measures.

The performance for the best neural networks model can be seen in table 2 and figure 5.

Accuracy	0.95
Kappa	0.90
Sensitivity	0.96
Specificity	0.94
AUC	0.95

Table 2: performance metrics

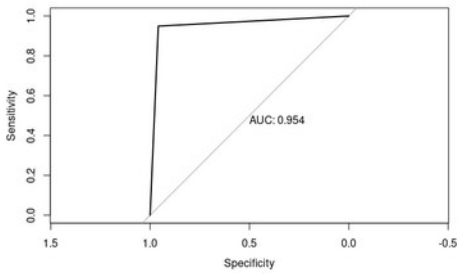

Figure 5: AUC plot for selected model

K Nearest Neighbours

KNN is a supervised machine learning algorithm that is mainly based on feature similarity. It classifies the data by checking how similar a data point is to its neighbours. KNN is a distance-based algorithm and can be affected by the scale of features. In this study, variables are get scaled to be used in the KNN model. The range of values has been examined to find the optimal value for K.

Figure 6 shows the trend of accuracy by increasing the number of neighbours. It implies that for a smaller value of K, the model shows higher accuracy. The value of 4 is the optimal value for the number of neighbours.

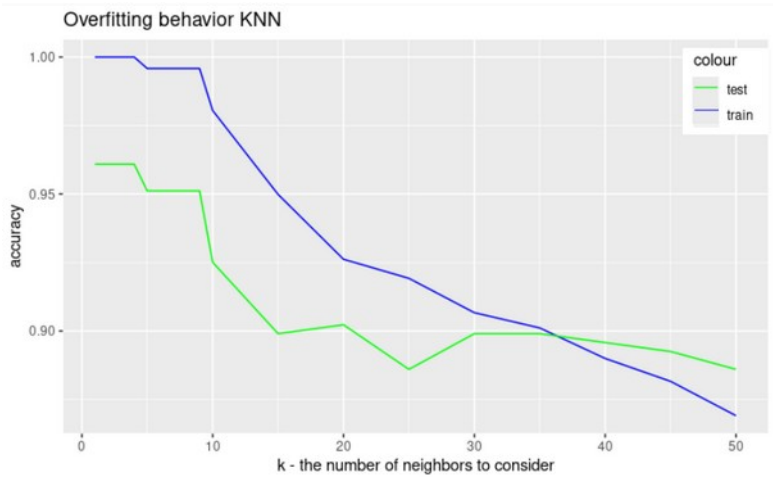

Figure 6: Accuracy for train and test

The performance for the best model can be seen in table 3 and figure 7.

Accuracy	0.99
Kappa	0.98
Sensitivity	0.98
Specificity	1
AUC	0.99

Table 3: Performance metrics

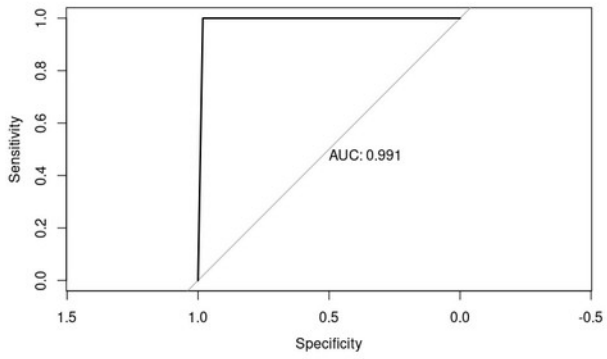

Figure 7: AUC plot

Decision tree

Decision tree is a decision make process that can be optimized by parameters like complexity, Maxdepth and minsplit.

We can terminate the splitting of the decision tree based on some thresholds. For instance, if the decision tree splits the data into two groups of 45 and 5 nodes, five is too small and may cause overfitting. Minsplit parameter determines the minimum number of observations in the parent node. The default is 20, which means if we have less than 20 observations in a parent node, this label will be considered a terminal node.

Maxdepth is another parameter that prevents the tree from growing to exceed a certain height. The default is 30, and the root node is counted as depth 0.

The complexity parameter is a stopping parameter. It aims to speed up the research and save computing time by identifying splits that prune is not worthwhile. The default value for CP is 0.01. The performance metrics for models have been calculated with the caret library. Figure 8 shows the features for selected model.

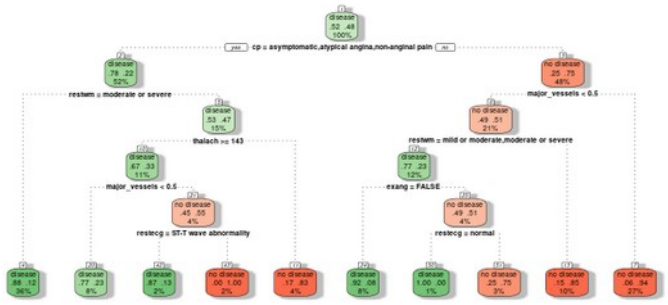

Figure 8: Decision tree map for selected model

The performance for the best model can be seen in table 4 and figure 9.

Accuracy	0.88
Kappa	0.76
Sensitivity	0.93
Specificity	0.95
AUC	0.94

Table 4: performance metrics

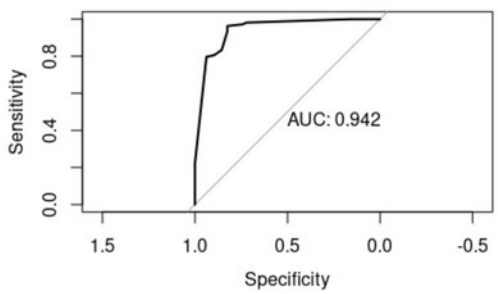

Figure 9: AUC plot

Model Comparison

By comparing the four models performance in table 5, we are noticed:

- Accuracy, which is the percentage of correct predictions, is highest in the KNN model.
- The precision that is a ratio of correctly predicted positive in true predicted observations is highest in a in the KNN classifier.
- Recall or sensitivity, which is the ratio of correctly predicted positive in all positive populations, is highest in the KNN model.
- Kappa, which is a metric for comparing how our result comes with chance, in in the KNN is the highest.
- AUC, which implies the model's ability to discriminate between positive and negative classes, in the KNN is the highest.

	SVM	Neural Networks	KNN	Decision tree
Accuracy	**0.98**	**0.95**	**0.99**	**0.88**
Kappa	**0.97**	**0.90**	**0.98**	**0.76**
Sensitivity	**0.97**	**0.96**	**0.98**	**0.93**
Specificity	**0.99**	**0.94**	**1**	**0.95**
AUC	**0.98**	**0.95**	**0.99**	**0.94**

Table 5: All selected models from different ML models

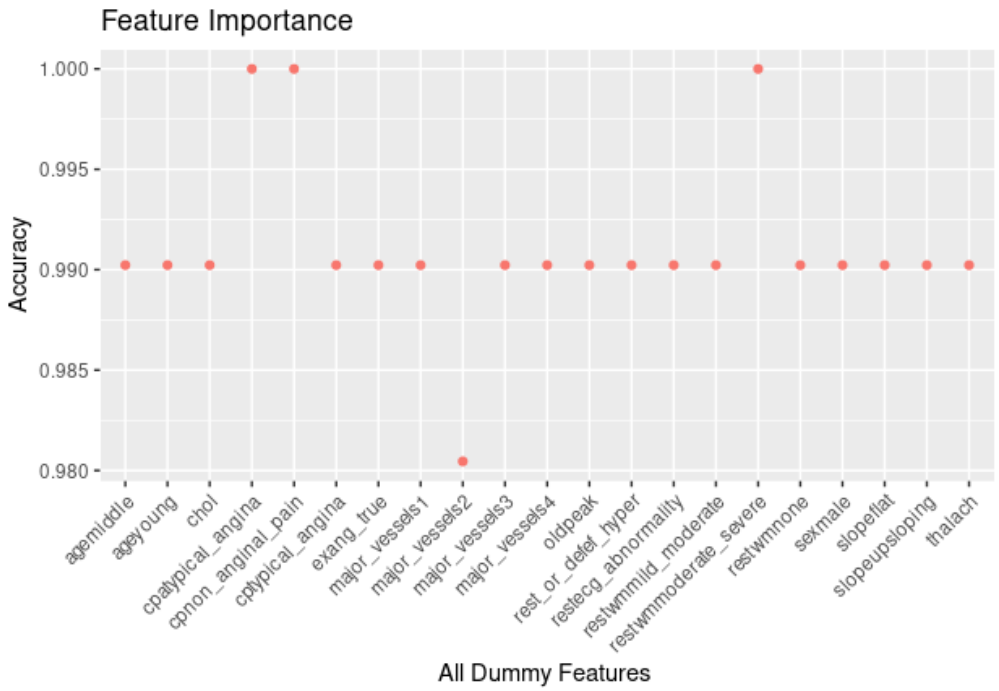

Figure 10: Accuracy for a model without specific feature

For feature importance analysis, I applied a loop on all features to drop one by one and calculate the KNN and its accuracy without that feature. As figure 33 shows, there is 21 independent feature after the dummy, and by dropping some variable, the accuracy changes. For example, the model's accuracy improves by removing two features related to the CP variable and a feature related to restwm in the original dataset. On the other hand, All the features whose removal from the list of predictive features for KNN model did not increase the model's accuracy could account for the most predictive features. For example, dropping a feature related to the major vessel in the original data leads to a drop in accuracy, implying a most predictive feature.
