DSA Assignment Sem 2 AY 24/25

Part 0: Introduction

Heart disease remains one of the most significant public health challenges globally, accounting for a substantial proportion of mortality each year. This report aims to identify the most suitable model for predicting one's heart disease status based on the dataset "heart-disease-dsa.csv", which contains medical conditions of the health of 300 individuals, including both quantitative and categorical variables like blood pressure, type of chest pain, cholesterol level, heart rate etc. We will use the dataset and determine the most suitable variables to develop models using K-Nearest Neighbours (KNN), Decision Tree (DT), and Logistic Regression (LR). The best version of each method will be chosen, and the performance of each model will then subsequently be assessed for its respective goodness of fit based on metrics like True Positive Rate (TPR), precision, Receiver Operating Characteristic (ROC) curve, and Area Under Curve (AUC). The pros and cons of the 3 models will be compared so that the best model can be determined.

Part I: EDA – Exploring the variables & association

(A) Summary + Strength of Association between Quantitative Input Variables & Response Variable

Histograms were used to visualise the distribution of each quantitative input variable.

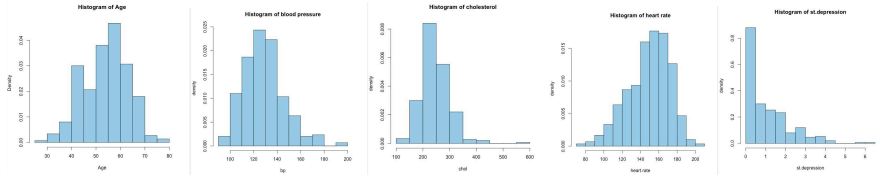

Fig. 1: Histograms of quantitative input variables

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     40      50      60      60      70      80
> summary(bp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   94.0   120.0   130.0   131.7   140.0   200.0
> summary(chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    126     211     240     246     274     564
> summary(heart.rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   71.0   133.0   152.0   149.5   166.0   202.0
> summary(st.depression)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00    0.00    0.80    1.05    1.60    6.20
```

Fig. 2: Summary of quantitative input variables

Box plots were used to visualise and identify the strength of association between each quantitative input variable and the response variable – disease status, so as to determine which inputs to add to the 3 models developed.

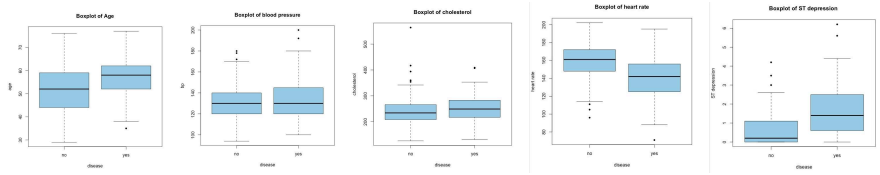

Fig. 3: Boxplots of quantitative input features – age, bp, chol, heart.rate, st.depression (from left to right) against response variable disease

age: Represents the age of patients in years. As people age, their risk of heart disease increases. Fig. 1 shows a mostly symmetrical, bimodal distribution. Fig. 2 indicates most ages fall between 40–70, with a few outliers. Fig. 3 suggests individuals with heart disease have a higher median age, indicating older individuals have higher disease prevalence.
bp: Represents resting blood pressure (mmHg) upon admission. High blood pressure increases the risk of heart disease. Fig. 1 shows a unimodal, slightly right-skewed distribution. Fig. 2 shows most bp values range from 120–140, with some outliers. Fig. 3 shows little difference between individuals with and without heart disease, suggesting a weak association.
chol: Represents serum cholesterol level (mg/dl). High cholesterol increases heart disease risk. Fig. 1 shows a unimodal, right-skewed distribution. Fig. 2 shows most values between 211–274, but outliers exist (max 564). Fig. 3 shows no significant difference in chol distribution between individuals with and without heart disease, suggesting a weak association.
heart.rate: Represents the highest heart rate achieved during exercise. Persistently high heart rate increases heart disease risk. Fig. 1 shows a unimodal, left-skewed distribution. Fig. 2 shows most values between 133–166, with some outliers. Fig. 3 shows individuals with heart disease have a lower median heart rate, suggesting a potential association.
st.depression: Represents ST depression in ECG during exercise relative to rest (mm). High ST depression increases heart disease risk. Fig. 1 shows a unimodal, extremely right-skewed distribution. Fig. 2 shows most values between 0–1.60, with outliers (max 6.20). Fig. 3 shows individuals without heart disease have significantly lower median ST depression levels, suggesting a potential association.

(B) Summary + Strength of Association between Categorical Input Variables & Response Variable

Tables were used to visualise the distribution of each categorical input variable.

```
> table(sex)
sex
  0   1
 94 206
> table(chest.pain)
chest.pain
  0   1   2   3
142  50  86  22
> table(fbs)
fbs
  0   1
255  45
> table(rest.ecg)
rest.ecg
  0   1   2
145 151   4
> table(vessels)
vessels
  0   1   2   3
173  64  38  20
> table(angina)
angina
  0   1
201  99
> table(blood.disorder)
blood.disorder
  0   1   2   3
  2  18 163 117
```

Fig. 4: Summary of categorical input variables

Contingency tables of joint proportions, which determined the conditional probability of heart disease given each categorical input variable, as well as barplots were used to visualise and identify the strength of association between the variables.

```
        disease
sex           no       yes
  0   0.7446809 0.2553191
  1   0.4466019 0.5533981
        disease
chest.pain        no       yes
         0 0.2676056 0.7323944
         1 0.8200000 0.1800000
         2 0.7906977 0.2093023
         3 0.6818182 0.3181818
        disease
fbs           no       yes
  0   0.5450980 0.4549020
  1   0.5111111 0.4888889
        disease
rest.ecg      no       yes
       0 0.4551724 0.5448276
       1 0.6291391 0.3708609
       2 0.2500000 0.7500000
        disease
vessels       no       yes
       0 0.7398844 0.2601156
       1 0.3125000 0.6875000
       2 0.1842105 0.8157895
       3 0.1500000 0.8500000
       4 0.8000000 0.2000000
        disease
angina        no       yes
  0   0.6915423 0.3084577
  1   0.2323232 0.7676768
        disease
blood.disorder        no       yes
             0
             1 0.3333333 0.6666667
             2 0.7757576 0.2242424
             3 0.2393162 0.7606838
```

Fig. 5: Contingency tables of proportion by categorical input variables

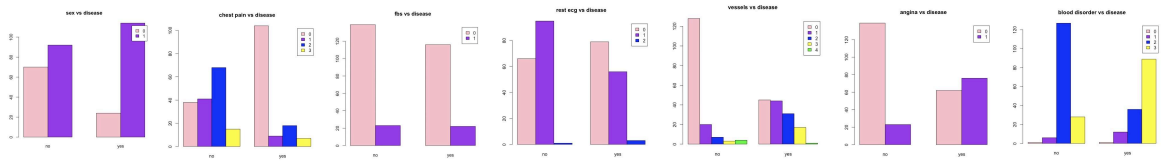

Fig. 6: Bar Plots between categorical input variables & response variable

sex: Indicates patient gender, with “1” for males and “0” for females. Men are at higher risk of heart disease at a younger age, while women are more prone to osteoporosis and autoimmune disorders. Fig. 5 & 6 show that males (55%) have a higher probability of heart disease than females (26%), suggesting a potential association.
chest.pain: Indicates chest pain type: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic. Chest pain can signal serious conditions like heart attacks. Fig. 5 & 6 show disease prevalence differs across types, with type 0 (73%) strongly associated with heart disease compared to types 1 (18%), 2 (21%), and 3 (32%).
fbs: Indicates if fasting blood sugar >120 mg/dl (“1” = True, “0” = False). High fasting blood sugar increases risk for diabetes, heart disease, and kidney damage. Fig. 5 & 6 show little difference: 46% disease presence for fbs = 0 and 49% for fbs = 1, suggesting weak association.
rest.ecg: Indicates resting electrocardiogram result: 0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy. Abnormal ECGs can signal heart issues. Fig. 5 & 6 show probabilities of disease: type 0 = 54%, type 1 = 37%, and type 2 = 75%, suggesting moderate to little association.
angina: Indicates whether exercise-induced angina occurred (“1” = yes, “0” = no). Fig. 5 & 6 show individuals with angina (77%) have a much higher probability of heart disease than those without (31%), suggesting a strong association.
vessels: Indicates number of major vessels visible by fluoroscopy (1–4). More affected vessels increase heart disease risk. Fig. 5 & 6 show probabilities: 0 vessels = 26%, 1 = 69%, 2 = 82%, 3 = 85% (highest), and 4 = 20%, suggesting vessel abnormality is indicative of heart disease.
blood.disorder: Represents blood disorder status: 1 = normal, 2 = fixed defect, 3 = reversible defect, 0 = missing. Blood disorders can increase heart disease risk. Fig. 5 & 6 show disease probabilities: type 1 = 67%, type 2 = 22%, and type 3 = 76% (highest), suggesting certain blood disorders are associated with heart disease.

(C) Summary of Response Variable
disease is a categorical response variable, which represents the heart disease status of patients, with “yes” for patients with heart disease and “no” for those patients without.

```
> table(disease)
disease
 no yes
162 138
```

Fig. 7: Summary of Categorical response Variable

Part II: Methods – Building KNN, DT & LR Classifiers
For all my models, I have decided to drop the variables *bp*, *chol*, *fbs and rest.ecg* because they were found to have little to no association to the response variable. For the variable *blood.disorder,* there were 2 missing values classified as “0”, which I then proceeded to classify them to the modal category which is *blood.disorder* = 2. For reproducibility, set.seed(1101) was used.In order to find the best arguments to use for the models, I will be utilising 5-fold cross-validation (CV), dividing the dataset into 5 subsets to evaluate the models based on True Positive Rate (TPR). Keeping to the recommended 8:2 ratio for splitting the data into training and testing respectively for

each fold, the training dataset consists of 240 observations, while the testing dataset consists of 60 observations, so as to ensure every data point will be used in the evaluation, providing a more comprehensive assessment. I will then derive the goodness of fit of each model by finding the values of TPR, Precision, AUC and plotting ROC curves on the full data set of 300 observations. TPR shows the proportion of positive observations the classifier correctly identified, Precision shows the percentage of observations that are actually positive among the marked positives, while AUC measures the area under the ROC Curve – the larger the magnitude of AUC, the better the prediction of the final class of whether one has heart disease or not.

Model 1: K-Nearest Neighbours (KNN) Classifier

KNN is a simple, instance-based algorithm that classifies a new data point based on the majority label of its k closest neighbors in the feature space. It calculates Euclidean distances to find the nearest neighbors and assigns the most common label among them to the new point. Since the model only works with numeric variables, I transformed the categorical variables that are ordinal in nature into numeric variables (*chest.pain, vessels, blood.disorder*). The rest of the categorical variables were dropped. To prepare the data for KNN, all features were then standardised using the scale function, ensuring each feature contributes proportionally to the distance calculation when making a prediction of a certain test point. To choose the best K value for the model, I used the 5-fold CV to evaluate between a pre-chosen set of K values – c(1, 3, 5, 7, 9, 11, 13, 15, 17), which was chosen based on the conditions that only odd values can be used and that K is capped at the square root of n (size of the training data). Through evaluating based on True Positive Rate (TPR), the best K value of 7 was chosen. The performance of Model 1 when K = 7 is shown below.

Indicator	Value (4sf)
TPR	0.8261
Precision	0.8702
AUC	0.5535

Table 1: Indicators to assess goodness of fit of KNN model (Model 1)

Model 2: Decision Trees (DT) Classifier

A decision tree is a classification method that splits data into branches based on feature values (both categorical & quantitative) to make predictions. At each node, it chooses the feature that best separates the data (information gain/ Gini), creating a structure of decision rules. The process continues until it reaches a stopping criteria, such as a minimum number of samples per node (minsplit). To choose the best minsplit value for the model, I used the 5-fold CV to evaluate between a pre-chosen set of minsplit value – c(1:100), which gives flexibility in finding the optimal tree. A low minsplit value could lead to risks of over-fitting while a high minsplit value could result in the model missing out on detailed patterns. Through evaluating based on True Positive Rate (TPR), the best minsplit value of 9 was chosen. The performance of Model 2 when minsplit = 9 is shown below.

Indicator	Value (4sf)

TPR	0.8333
Precision	0.8984
AUC	0.9148

Table 2: Indicators to assess goodness of fit of DT model (Model 2)

Model 3: Linear Regression (LR) Classifier

Logistic regression is a binary classification model that predicts the probability of a class using a linear combination of features (categorical & quantitative) passed through a logistic function. The probability output will be between 0 and 1, which will be used as a threshold to assign class labels. To determine the most appropriate threshold values, I plotted a contingency table to determine the proportion of the sample that has heart disease, which was . The threshold of 0.46 was hence used instead of the usual value of 0.5 as the model has been adjusted to be more sensitive to the actual distribution, potentially improving classification accuracy. The performance of Model 3 when the threshold value of 0.46 is used is shown below.

Indicator	Value (4sf)
TPR	0.8406
Precision	0.8529
AUC	0.9312

Table 3: Indicators to assess goodness of fit of LR model (Model 3)

Pros & Cons of Each Classifier

Classifier	Pros	Cons
Model 1 (KNN)	KNN can perform well on small datasets, where computation time is manageable, providing good results without heavy model tuning.	Only works for numeric variables, so need to convert categorical variables to numeric inputs, but this is only possible for ordinal categorical variables → excludes some important non-ordinal variables → reduces model's accuracy. ROC may not be a reliable metric as it is based on probabilities rather than binary values.
Model 2 (DT)	DT can take both categorical and quantitative variables. Visualisation of tree plot helps to identify which variables are most	Unsuitable minsplit value leads to over-fitting of DT as they can include insignificant details within the data unnecessarily → over

	crucial in making predictions.	complex tree plots created → leads to biased predictions → reduce accuracy.
Model 3 (LR)	DT can take both categorical and quantitative variables. Coefficients of LR give direct insight into how each predictor affects the outcome, allowing for easy interpretation.	LR only works well if the relationship between the predictor and response variables is linear. Although it can incorporate categorical variables by using encoding techniques, it may struggle to accurately model complex, non-linear interactions between features, potentially affecting its overall predictive accuracy.

Table 4: Pros & Cons of Each Classifier

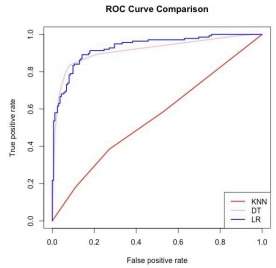

Fig.8: Combined ROC curve of all 3 classifiers

Part III: Conclusion – The Best Model/ Classifier

To determine the best model, we will be evaluating the metric values for each classifier. Based on a quantitative comparison using the results from our analysis, Logistic Regression achieved the highest AUC value of 0.9312, indicating the best overall ability to distinguish between classes. It also recorded the highest TPR at 0.8406, suggesting it correctly identifies more positive cases compared to the other models. Although Decision Trees showed the highest Precision at 0.8984, and performed reasonably well across all metrics, Logistic Regression still maintains an advantage due to its superior balance between TPR, Precision, and AUC. K-Nearest Neighbors (KNN), while demonstrating decent TPR and Precision, had a significantly lower AUC of 0.5535, suggesting poor class separation, and is therefore less suited for this dataset. Given its strong performance across the most important metrics, particularly AUC and TPR — Logistic Regression is considered the most reliable model for this task.