Binary Classification of Heart Disease Using Random Forests and k-Nearest Neighbour

Audrey Karabayinga

2022-11-22

Contents

1 Introduction

1.1 Cardiovascular and Heart Disease

Cardiovascular disease (CVD)[1] is a term for all types of diseases affecting the heart or blood vessels, including coronary heart disease, stroke, congenital heart defects and peripheral artery disease (NHLBI, n.d.). All heart diseases (HDs) are CVDs but not all CVDs are HDs. According to the World Health Organisation (WHO), CVDs are the leading cause of morbidity and mortality worldwide (WHO 2021). In 2019, approximately 17.9 million people (32% of worldwide deaths) died from CVDs. Over 75% of these deaths occur in low- and middle-income countries, where there is inadequate access to effective and affordable primary health care that prevents early and accurate detection, diagnosis and timely intervention (WHO 2021).

Preventable risk factors for HDs and stroke include poor diet, lack of physical activity, ineffective use of alcohol and tobacco (WHO 2021; Fuchs and Whelton 2020; Benjamin et al. 2018; Peters, Muntner, and Woodward 2019). These manifest in the body as elevated blood pressure (BP), blood sugar (BS) and blood lipids, and overweight or obesity (WHO 2021).

1.1.1 Machine Learning and Diagnosis of CVD

Machine learning (ML), a scientific discipline which uses computer algorithms that learn and adapt, can be exploited to diagnose a number of medical disorders (Sajda 2006; Deo 2015). ML approaches include *supervised learning*, which entails using labeled data to train algorithms to classify data or predict outcomes, and *unsupervised learning*, which aims to identify naturally occurring patterns and trends or clusters data into groups using unlabelled data (Deo 2015).

In the context of disease diagnosis, supervised learning entails training algorithms using data for which the disease status of patients is known and applying the trained algorithm to make predictions regarding the disease status of patients whose disease status is unknown (Akella and Akella 2021).

Supervised learning models are categorised as:

- *Classification models* - predict specific discrete or categorical values; or
- *Regression models* - predict continuous numerical values.

1.2 Data set Description

The heart failure prediction data set (fedesoriano 2021) was created by combining the following 5 data sets from the Heart Disease Data set directory of the UCI Machine Learning Repository:

[1] According to the National Heart, Lung and Blood Institute (NHLBI), the terms 'cardiovascular disease (CVD)', 'heart disease (HD)' and ' coronary heart disease' are often used interchangeably (NHLBI, n.d.). The NHLBI, defines them as follows:

- CVD: A term for all types of diseases affecting the heart or blood vessels, including coronary heart disease, stroke, congenital heart defects and peripheral artery disease.
- HD: A catch-all phrase for a variety of conditions affecting the heart structure and function. The most common type of HD is coronary heart disease.
- Coronary heart disease: Build up of plaque in the arteries (atherosclerosis). It is another term for coronary artery disease and is one type of HD but not the only one.

- Cleveland Clinic Foundation - 303 observations
- Hungarian Institute of Cardiology, Budapest - 294 observations
- V.A. Medical Center, Long Beach, CA - 200 observations
- University Hospital, Zurich, Switzerland - 123 observations
- Statlog (Heart) data - 270 observations

Combined, these data sets resulted in a total of 1190 observations. The creator of the *heart* data set found and removed 272 duplicates, resulting in a data set of 918 observations and 12 variables. The codebook accompanying the data set defined the variables as follows:

- Outcome variable:

  i. *HeartDisease*: output class (1: heart disease, 0: normal)

- Predictor variables:

  i. *Age*: age of patient (years)
  ii. *Sex*: sex of the patient
     - M: Male
     - F: Female
  iii. *ChestPainType*: chest pain type
     - TA: Typical Angina
     - ATA: Atypical Angina
     - NAP: Non-Anginal Pain
     - ASY: Asymptomatic
  iv. *RestingBP*: resting blood pressure [mm Hg on admission to the hospital]
  v. *Cholesterol*: serum cholesterol [mg/dL]
  vi. *FastingBS*: fasting blood sugar
     - 1: if FastingBS > 120 mg/dl
     - 0: otherwise
  vii. *RestingECG*: resting electrocardiogram results
     - Normal: Normal
     - ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
     - LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
  viii. *MaxHR*: maximum heart rate achieved [Numeric value between 60 and 202]
  ix. *ExerciseAngina*: exercise-induced angina
     - Y: Yes
     - N: No
  x. *Oldpeak*: oldpeak = ST depression induced by exercise relative to rest [Numeric value measured in depression]
  xi. *ST_Slope*: the slope of the peak exercise ST segment
     - Up: upsloping
     - Flat: flat
     - Down: downsloping

1.3 Capstone Project

This report was created for the 'Choose Your Own' Project, the final of two Capstone projects required to fulfill the HarvardX Data Science Professional Certificate series. The stated aim of the project was to solve a problem of the student's choice using a publicly available data set and 2 ML algorithm techniques, with at least one algorithm that is more advanced than standard linear regression.

This project entailed a binary classification prediction project, which used the heart failure prediction data set, *heart*, to predict *presence* of HD in patients. For this project, the *heart* data set was first imported from kaggle and formatted. It was then divided into: i) a *model* data set, which contained 90% of *heart* data and was used for exploratory data analysis (EDA) and algorithm development; and ii) a *validation* data set, which contained 10% of *heart* data and was used as the final hold-out set to evaluate the final algorithm's performance. Data exploration and visualisation were conducted only on the *model* data set. Prior to algorithm development, the *model* data set was further divided into: i) a *train* data set, which contained 90% of *model* data and was used for model training and hyper-parameter tuning during algorithm development; and ii) a *test* data set, which was used for initial model evaluation during algorithm development.

Random forest (RF) and k-nearest neighbour (kNN) were the 2 ML algorithms selected for this project. Data pre-processing entailed kNN-imputation of improbable values for both the RF and kNN models and creation of dummy variables for categorical variables, and normalisation and skewness removal for continuous variables for the kNN model. Both models also utilized 10-fold cross-validation (of the *train* set) for hyper-parameter tuning. The best hyper-parameters for each model were chosen based on the performance metric, accuracy, and used to create the best RF and kNN models. The effectiveness of the 2 best models was then assessed by fitting them on the entire *train* data set and comparing their accuracy at predicting HD on the *test* data set. The model with the highest accuracy, the kNN model, achieved an accuracy of 0.78 and was chosen as the final algorithm. The RF model achieved an accuracy of 0.76. The final algorithm's performance on new data was finally assessed by fitting it on the entire *model* data set and evaluating it on the *validation* data set, resulting in an accuracy of 0.86.

The documents submitted for assessment of this project are: i) a report in R markdown format; ii) a report in PDF file format (knit from the Rmd file); iii) a script in R format; iv) the *heart* data set (heart.csv), and; v) file containing the list of citations used (citations-heart.bib).

1.3.1 Acknowledgments

The creators of these data sets are:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

2 Methods

2.1 Data Import, Format and Overview

The heart failure prediction data set was imported into R and saved as *heart*. Table 1 shows an overview of the *heart* data set. It contained 918 observations and 12 variables—11 predictor variables and 1 outcome variable, *heartdisease*. The outcome variable indicated whether HD was *present* or *absent*. Table 1 also shows that *fastingbs* and *heartdisease* were classified as ‘numeric’ but should be of class ‘factor’. The data set was re-formatted and all character variables were converted to class ‘factor’.

Table 1: Overview of heart data set

age	sex	chestpaintype	restingbp	cholesterol	fastingbs	restingecg	maxhr	exerciseangina	oldpeak	st_slope	heartdisease
Class											
numeric	character	character	numeric	numeric	numeric	character	numeric	character	numeric	character	numeric
First 10 observations											
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0

Table 2 shows the structure of the *heart* data set after formatting, including all variable labels, names, classes, and levels, where available. Prior to data partitioning, the *heart* data set was checked for missing values and found to have none.

Table 2: Structure of heart data set

Variable label	Description	Class	Variable levels
Outcome variable			
heartdisease	Heart disease	Factor	0: Absent; 1: Present
Predictor variables			
age	Age (in years)	Numeric - integer	
restingbp	Resting blood pressure (in mmHg on admission to the hospital)	Numeric - integer	
cholesterol	Serum cholesterol (in mg/dL)	Numeric - integer	
maxhr	Maximum heart rate achieved	Numeric - integer	
oldpeak	ST depression induced by exercise relative to rest	Numeric	
sex	Patient’s gender	Factor	M: Male; F: Female
chestpaintype	Type of chest pain	Factor	TA: Typical angina; ATA: Atypical angina; NAP: Non-anginal pain; ASY: asymptomatic
fastingbs	Fasting blood sugar	Factor	0: </= 120 mg/dL; 1: > 120 mg/dL
restingecg	Resting electrocardiographic results	Factor	Normal: normal; ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); LVH: showing probable or definite left ventricular hypertrophy by Estes’ criteria
exerciseangina	Exercise-induced angina	Factor	Y: Yes; N: No
st_slope	Slope of the peak exercise ST segment	Factor	Up: Upsloping; Flat: Flat; Down: Downsloping

2.1.1 Data Partitioning I: Creating *model* and *validation* Sets from *heart* Set

Before performing a more in-depth EDA, the *heart* data set was divided into 2 (Figure 1):

i) a *model* data set, which contained 90% of *heart* data and was used for EDA and algorithm development; and
ii) the *validation* data set, which contained 10% of *heart* data and was used as the final hold-out set to evaluate the final algorithm’s performance.

The splitting was stratified using the outcome variable, *heartdisease,* to ensure that the the frequency distribution of the outcome in the new data sets was representative of the *heart* data set. Table 3 shows the proportion of patients with and without HD in each data set. A second data partitioning, where the *model* data set was divided into the *train* and *test* data sets as shown in Figure 1, was done in the Modeling Approach section.

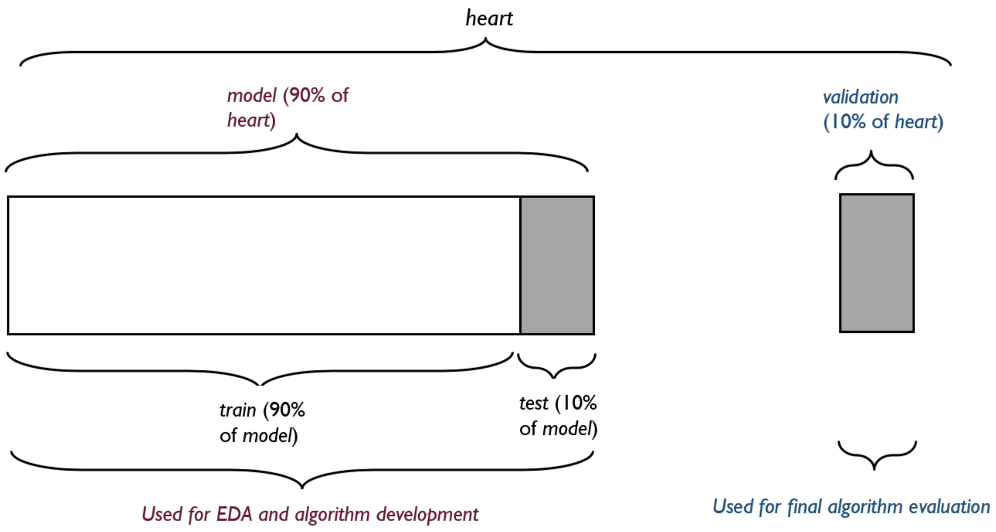

heart
model (90% of heart)
validation (10% of heart)
train (90% of model)
test (10% of model)
Used for EDA and algorithm development
Used for final algorithm evaluation

Figure 1: Data set partitions

Table 3: Data partitioning I - Overview of data sets observations and proportions with and without heart disease

Data set	Number of observations	Proportion of heart data set	Proportion with HD	Proportion without HD
Heart	918	100%	55.3%	44.7%
Model	826	90%	55.3%	44.7%
Validation	92	10%	55.4%	44.6%

2.2 Data Cleaning, Exploration and Visualisation

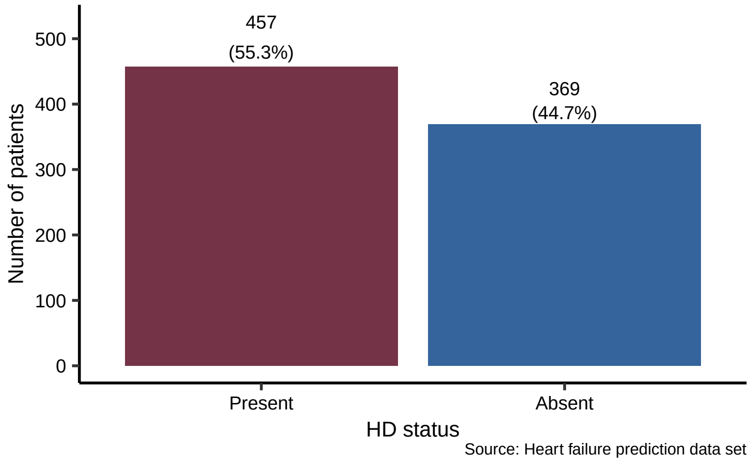
Data exploration and visualisation was performed using only the *model* set.

2.2.1 Outcome Variable

2.2.1.1 Heart Disease The *heartdisease* column contained the outcome variable, which indicated presence or absence of HD. Out of a total of 826 patients, HD was present in 457 (55.3%) and absent in 369 (44.7%) (Figure 2). This shows that the data set was balanced.

Table 4 shows the descriptive statistics of all predictor variables stratified by HD status. HD appears to be associated with :

- male gender,
- asymptomatic chest pain,
- fasting BS > 120 mg/dL,
- resting ECG results with ST-T wave abnormality and LVH,
- presence of exercise-induced angina,
- downsloping and flat slow of peak exercise ST segment (st_slope)
- older age,
- slightly higher resting BP,
- lower serum cholesterol and maximum HR achieved, and



Source: Heart failure prediction data set

Figure 2: Distribution of HD status

- higher oldpeak (ST depression induced by exercise) values.

2.2.1.2 Data Cleaning Table 4 revealed improbable values for some variables. Out of the 826 patients in the *model* data set, 157 (19%) had cholesterol levels of 0, 1 (0.1%) had resting BP levels of 0, and 11(1.3%) had oldpeak values < 0. kNN imputation was carried out on the *model* data set and the improbable values were imputed using the default number of neighbours (i.e., k = 5). Figure 3 shows plots before and after imputation and Table 5 shows the descriptive statistics after imputation.

The kNN-imputed *model* data set was used for the rest of the exploration and visualisation section.

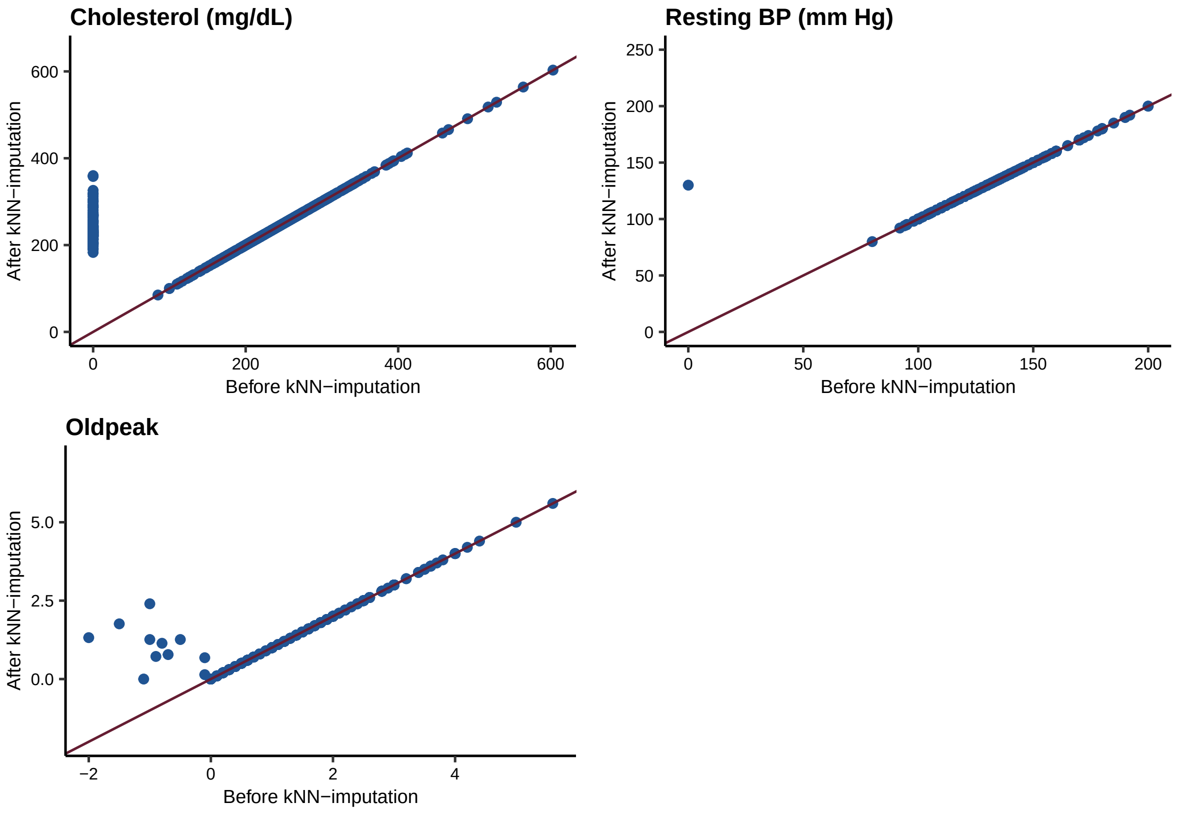
2.2.2 Predictor Variables - Categorical

2.2.2.1 Sex The *sex* column contained the patient's gender. Studies show that CVD tends to primarily affect men, with largely higher rates being found in men compared to women in most age groups (Peters, Muntner, and Woodward 2019; Benjamin et al. 2018). In our data, HD was observed to be more prevalent among males compared to females. Figure 4A shows that out of a total of 658 males, 415 (63.1%) were found to have HD compared to 243 (36.9%) who did not. Out of a total of 168 females, 42 (25%) were found to have HD compared to 126 (75%) who did not.

2.2.2.2 Chest Pain Type The phrase 'chest pain,' defined as "pain, pressure, tightness, or discomfort in the chest, shoulders, arms, neck, back, upper abdomen, or jaw, as well as shortness of breath and fatigue," is a common and recognizable symptom across various types of CVDs (Members et al. 2021). The *chestpaintype* column contained the patients' type of chest pain (also known as *angina pectoris*), which was categorised into 4 classes defined as follows (Detrano et al. 1984):

- Typical angina: Pain that occurs in the anterior thorax, neck, shoulders, jaw, or arms is precipitated by exertion and relieved within 20 min by rest;

- Atypical angina: Pain in none of the above locations and either not precipitated by exertion or not relieved by rest within 20 min;



Source: Heart failure prediction data set

Figure 3: Scatter plots of cholesterol, resting BP and oldpeak values before and after kNN-imputation of the model data set

Table 4: Descriptive statistics table for the model data set

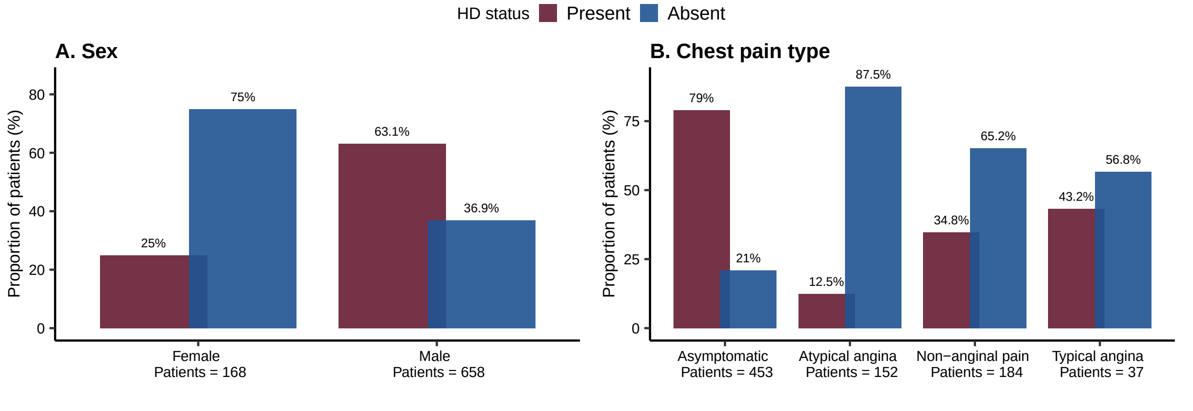
	All patients	**HD - Present**	**HD - Absent**
	(N=826)	**(N=457)**	**(N=369)**
Categorical variables			
Sex			
Female	168 (20.3%)	42 (9.2%)	126 (34.1%)
Male	658 (79.7%)	415 (90.8%)	243 (65.9%)
Chest pain type			
Typical angina	37 (4.5%)	16 (3.5%)	21 (5.7%)
Atypical angina	152 (18.4%)	19 (4.2%)	133 (36.0%)
Non-anginal pain	184 (22.3%)	64 (14.0%)	120 (32.5%)
Asymptomatic	453 (54.8%)	358 (78.3%)	95 (25.7%)
Fasting blood sugar			
</= 120 mg/dL	631 (76.4%)	302 (66.1%)	329 (89.2%)
> 120 mg/dL	195 (23.6%)	155 (33.9%)	40 (10.8%)
Resting ECG			
ST-T wave abnormality	162 (19.6%)	106 (23.2%)	56 (15.2%)
Left ventricular hypertrophy	167 (20.2%)	94 (20.6%)	73 (19.8%)
Normal	497 (60.2%)	257 (56.2%)	240 (65.0%)
Exercise-induced angina			
Yes	337 (40.8%)	287 (62.8%)	50 (13.6%)
No	489 (59.2%)	170 (37.2%)	319 (86.4%)
Slope of peak exercise ST segment (st_slope)			
Downsloping	57 (6.9%)	45 (9.8%)	12 (3.3%)
Flat	414 (50.1%)	340 (74.4%)	74 (20.1%)
Upsloping	355 (43.0%)	72 (15.8%)	283 (76.7%)
Continuous variables			
Age (years)			
Mean (SD)	53.4 (9.37)	55.8 (8.68)	50.4 (9.35)
Median [Min, Max]	54.0 [28.0, 77.0]	57.0 [31.0, 77.0]	51.0 [28.0, 76.0]
Resting BP - Systolic (mmHg)			
Mean (SD)	132 (18.1)	134 (19.3)	130 (16.2)
Median [Min, Max]	130 [0, 200]	131 [0, 200]	130 [80.0, 190]
Serum cholesterol (mg/dL)			
Mean (SD)	197 (109)	173 (127)	227 (70.7)
Median [Min, Max]	221 [0, 603]	216 [0, 603]	226 [0, 564]
Maximum HR achieved			
Mean (SD)	137 (25.5)	127 (23.3)	148 (23.4)
Median [Min, Max]	138 [60.0, 202]	125 [60.0, 195]	150 [69.0, 202]
ST depression induced by exercise (oldpeak)			
Mean (SD)	0.877 (1.05)	1.26 (1.13)	0.400 (0.674)
Median [Min, Max]	0.500 [-2.00, 5.60]	1.20 [-2.00, 5.60]	0 [-1.10, 3.50]

Table 5: Descriptive statistics table for the model data set - kNN-imputed variables

	All patients	**HD - Present**	**HD - Absent**
	(N=826)	**(N=457)**	**(N=369)**
Resting BP - Systolic (mmHg)			
Mean (SD)	132 (17.5)	134 (18.3)	130 (16.2)
Median [Min, Max]	130 [80.0, 200]	131 [92.0, 200]	130 [80.0, 190]
Serum cholesterol (mg/dL)			
Mean (SD)	243 (54.8)	249 (55.6)	237 (53.2)
Median [Min, Max]	235 [85.0, 603]	239 [100, 603]	230 [85.0, 564]
ST depression induced by exercise (oldpeak)			
Mean (SD)	0.903 (1.03)	1.30 (1.09)	0.405 (0.669)
Median [Min, Max]	0.600 [0, 5.60]	1.20 [0, 5.60]	0 [0, 3.50]

- Non-anginal pain: Pain not located in any of the above locations, or if so located not related to exertion, and lasting less than 10 sec or longer than 30 min;
- Asymptomatic: No pain.

In our data set, HD was observed to be more prevalent among patients with asymptomatic chest pain compared to those with typical angina, non-anginal pain, and atypical angina. Figure 4B shows that out of a total of 453 patients with asymptomatic chest pain type, 358 (79%) had HD compared to 95 (21%) without HD. Among 37 patients with typical angina, 16 (43.2%) had HD compared to 21 (56.8%) who did not. Among 152 patients with atypical angina, 19 (12.5%) had HD compared to 133 (87.5%) who did not. Among 184 patients with non-anginal pain, 64 (34.8%) had HD compared to 120 (65.2%) who did not.



Source: Heart failure prediction data set

Figure 4: Sex and chest pain type by HD status

2.2.2.3 Fasting Blood Sugar

Diabetes Mellitus (DM) is a significant risk factor for CVD (Benjamin et al. 2018). Fasting blood sugar (BS), or fasting blood glucose, levels are used to classify patients with DM. The American Heart Association (AHA) defines DM diagnosis as shown in Table 6 (AHA, n.d.a). The *fastingbs* column in the *heart* data contained patients' fasting BS levels as a categorical variable, with patients classified as having fasting BS levels $\leq$ 120 mg/dL or > 120 mg/dL.

Table 6: Categories of FBS in adults

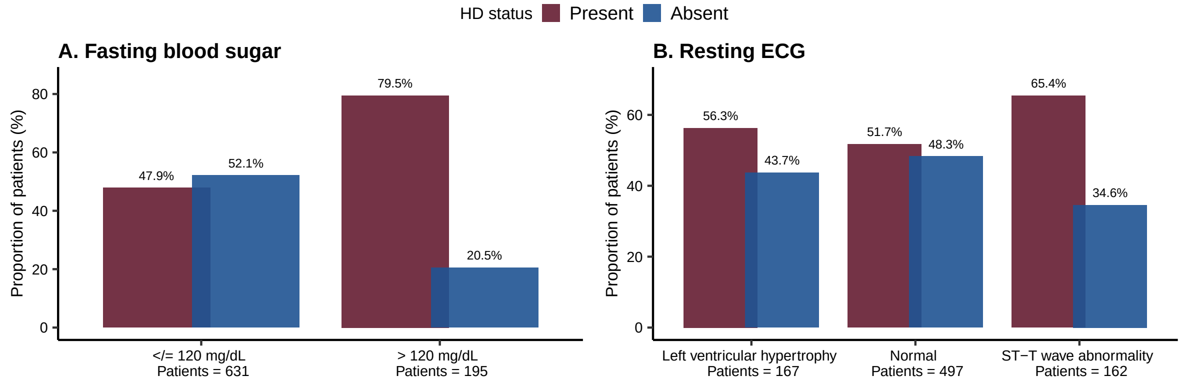
Fasting BS Category	**Diagnosis**	**Interpretation**
< 100 mg/dL	Normal	Healthy range
100-125 mg/dL	Prediabetes (impaired fasting glucose)	At increased risk of developing diabetes
> / = 125 mg/dL	Type 2 DM	At increased risk of heart disease or stroke

In our data, HD was observed to be more prevalent among patients with fasting BS > 120 mg/dL (i.e., patients with Type 2 DM and those at the upper end of the pre-diabetes category) compared to patients with $\leq$ 120 mg/dL (i.e., patients with normal levels of fasting BS and those on the lower end of the pre-diabetes range). Figure 5A shows that among 195 patients with fasting BS > 120 mg/dL, 155 (79.5%) had HD compared to 40 (20.5%) who did not. Among 631 patients with fasting BS $\leq$ 120 mg/dL, 302 (47.9%) had HD compared to 329 (52.1%) who did not.

2.2.2.4 Resting Electrocardiogram Resting electrocardiogram (ECG) is a test that records electrical cardiac activity of a person at rest (Curry et al. 2018). It is a non-invasive tool for screening and diagnosing HD (Larsen et al. 2002). Abnormal resting ECG results, such as ST-segment or T-wave abnormalities and left ventricular hypertrophy (LVH), are associated with increased risk of CVDs (Chou et al. 2011). The *restingECG* column of the *heart* data contained patients' resting ECG results categorised as:

- Normal: Normal ECG;
- ST: ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV);
- LVH: Probable or definite LVH by Estes' criteria.

In our data, HD was observed to be more prevalent among patients with ST-T wave abnormalities and LVH compared to patients with normal resting ECG results. Figure 5B shows that among 162 patients with ST-T wave abnormalities, 106 (65.4%) had HD compared to 56 (34.6%) who did not. Among 167 patients with probable or definite LVH, 94 (56.3%) had HD compared to 73 (43.7%) who did not. Among 497 patients with normal ECG results, 257 (51.7%) had HD compared to 240 (48.3%) who did not.



Source: Heart failure prediction data set

Figure 5: Fasting blood sugar and resting ECG by HD status

2.2.2.5 Exercise-induced Angina Although frequent exercise has been linked with fewer coronary HD events, vigorous physical activity can result in increased risk of heart attacks in vulnerable people (Sports Medicine et al. 2007). Exercise ECG is a test that records electrical cardiac activity during physical activity, such as on a treadmill or a bicycle, usually at a specific exercise intensity (Curry et al. 2018). Exercise-induced angina is chest pain brought on by stress from exercise (Hlatky 1999). The *exerciseangina* column of the *heart* data categorised patients into 2 categories: patients with and without exercise-induced angina based on the results of exercise ECG.

In our data, HD was observed to be more prevalent among patients with exercise-induced angina compared to patients without it. Figure 6A shows that among 337 patients with exercise-induced angina, 287 (85.2%) had HD compared to 50 (14.8%) without. Among the 489 patients without exercise-induced angina, 170 (34.8%) had HD compared to 319 (65.2%).

2.2.2.6 Slope of the Peak Exercise-induced ST-segment Depression (ST Slope) Abnormal exercise ECG results entail changes in the ST segment slope, such as elevations and depressions and have been linked to HD (Lim, Teo, and Poh 2016). The *st_slope* column of the *heart* data contained the patients' slope of the peak exercise ST segment categorised as upsloping, flat, or downsloping, based on exercise ECG results. In our data, HD was observed to be more prevalent among patients with downsloping and flat exercise-induced ST-segment depression slopes. Figure 6B shows that among 414 patients with flat ST-segment depression slopes, 340 (82.1%) had HD compared to 74 (17.9%) who did not. Among 57 patients with downsloping ST-segment depression slopes, 45 (78.9%) had HD compared to 12 (21.1%) who did not. Among 355 patients with upsloping ST-segment depression slopes, 72 (20.3%) had HD compared to 283 (79.7%) who did not.

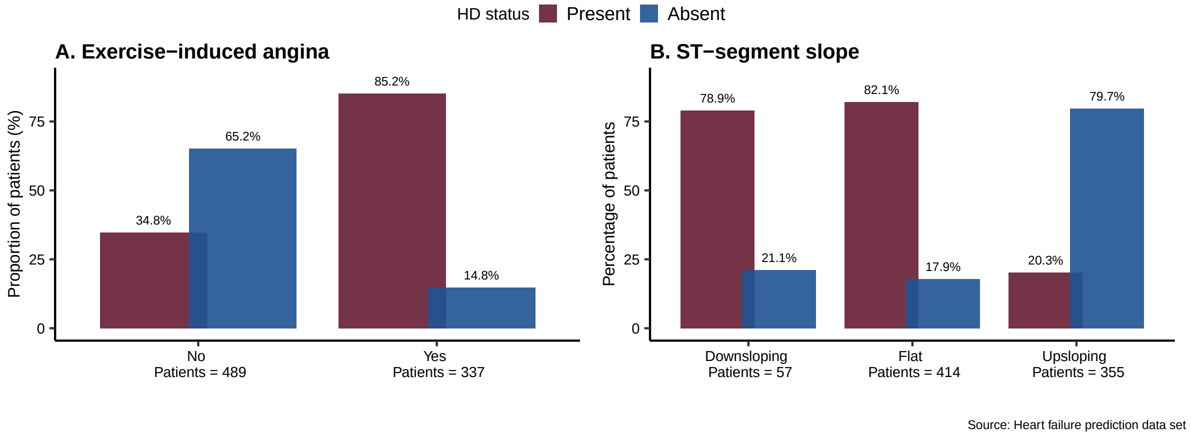


Figure 6: Exercise-induced angina and ST-segment slope by HD status

2.2.3 Predictor Variables - Continuous

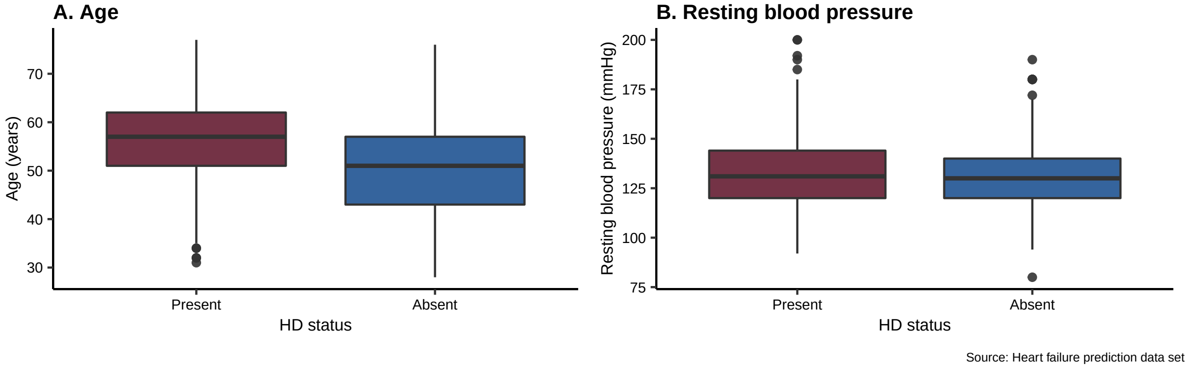
2.2.3.1 Age Age is an essential determinant of heart health, with older age associated with increased risk of CVD (North and Sinclair 2012). The *age* column contained the age of each patient in years. Figure 7A shows that patients with HD tend to be older. Patients with HD have mean and median ages of 55.8 and 57 years respectively, while those without HD have 50.4 and 51 years, respectively (Table 4).

2.2.3.2 Resting Blood Pressure High BP is the leading preventable risk factor for CVD globally (Olsen et al. 2016). According to the 2017 American College of Cardiology (ACC) / American Heart Association (AHA) task force on clinical practice guidelines report, BP is classified as shown in Table 7 (Whelton et al. 2018).

Table 7: Categories of BP in adults

BP Category	Systolic BP		Diastolic BP
Normal	< 120 mm Hg	and	< 80 mm Hg
Elevated	120-129 mm Hg	and	< 80 mm Hg
Hypertension - Stage I	130-139 mm Hg	or	80-90 mm Hg
Hypertension - Stage II	> or = 140 mm Hg	or	> or = 90 mm Hg

The *restingbp* column contained the resting BP of each patient in mm Hg, measured upon admission to hospital. It corresponded to the systolic BP in Table 7. Figure 7B shows no clear difference in resting BP levels between patients with and without HD. Patients with HD have mean and median resting BP levels of 134 and 131 mmHg respectively, while those without HD have 130 and 130 mmHg, respectively (Table 5).



Source: Heart failure prediction data set

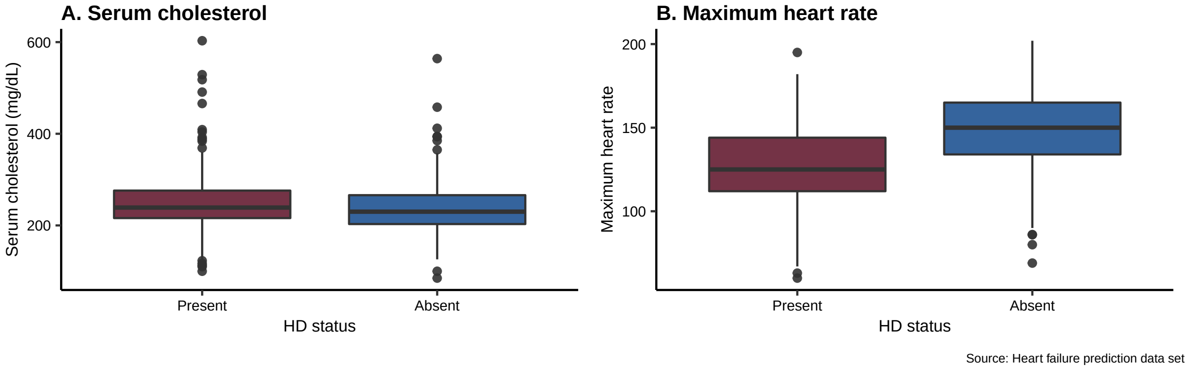
Figure 7: Distribution of age and resting BP by HD status

Table 8: Optimal Cholesterol Levels

Optimal Cholesterol Levels	
Total cholesterol	About 150 mg/dL
LDL ("bad") cholesterol	About 100 mg/dL
HDL ("good") cholesterol	At least 40 mg/dL in men and 50 mg/dL in women
Triglycerides	Less than 150 mg/dL

2.2.3.3 Serum Cholesterol Serum cholesterol, which contains "good" and "bad" cholesterol as well as triglycerises, is related to CVD (Grundy et al. 2019). Optimal cholesterol levels as defined as shown in Table 8 (Grundy et al. 2019). The *cholesterol* column contained patients' serum cholesterol levels in mg/dL. It likely corresponds to total cholesterol in Table 8. Figure 8A shows no clear difference in serum cholesterol levels between patients with and without HD. Patients with HD have mean and median serum cholesterol levels of 249 and 239 mg/dL respectively, while those without HD have 237 and 230 mg/dL, respectively (Table 5).

2.2.3.4 Maximum Heart Rate The *maxhr* column contained the maximum heart rate (HR) achieved by patients following exercise ECG. Figure 8B shows that patients with HD tend to achieve lower maximum HR levels. Patients with HD have mean and median maximum HR levels of 127 and 125 respectively, while those without HD have 148 and 150, respectively (Table 4).



Source: Heart failure prediction data set

Figure 8: Distribution of serum cholesterol and maximum HR by HD status

2.2.3.5 ST Depression Induced by Exercise Relative to Rest (oldpeak) The *oldpeak* column contained patients' ST depression induced by an exercise ECG test relative to rest. Figure 9 shows that patients with HD tend to have higher oldpeak levels. Patients with HD have mean and median oldpeak levels of 1.3 and 1.2, respectively, while those without HD have 0.4 and 0, respectively.

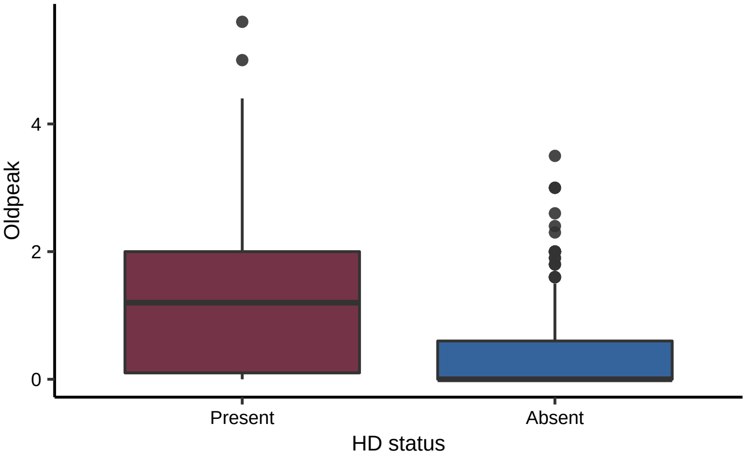


Figure 9: Distribution of ST depression induced by exercise (oldpeak) by HD status

2.2.4 Insights Gained

The data set appears balanced, with HD present in 56.3% of patients and absent in 44.7% of patients. Compared to patients without HD, those with HD were more likely to be older, male, and have lower maximum HR, higher oldpeak values, asymptomatic chest pain type, fasting blood sugar levels greater than 120 mg/dL, abnormal resting ECG results (ST-T wave abnormalities and LVH), exercised-induced angina, and flat or downsloping ST-segment depression slope. No clear differences were observed for resting BP and serum cholesterol levels compared across HD status.

2.3 Data Preparation

2.3.1 Data Partitioning II

2.3.1.1 Creating *train* and *test* Sets Before algorithm development, the imputed *model* data set was split into 2:

i) the *train set*, which was used for training during algorithm development; and
ii) the *test set*, which was used for initial testing during algorithm development.

Splitting was stratified using the outcome variable, *heartdisease*.

2.3.1.2 10-fold Cross-validation The *train* set was then used to create 10 cross-validation folds to be used in hyper-parameter tuning. Sampling was stratified using the outcome variable, *heartdisease*.

2.4 Modeling Approach

2.4.1 Algorithm Development

The ML algorithms selected for this project were Random forest (RF) and k-nearest neighbour (kNN). The R package used for modeling was `tidymodels`. Given that the data set was balanced, the performance metrics used for model comparison was *accuracy*, i.e., the proportion of patients whose HD status was correctly predicted. The prediction of HD status was based on a threshold of 50%. The formula for calculating accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N}$$

where:

TP = true positives,

TN = true negatives,

FN = false negatives,

FP = false positives,

P = positives,

N = negatives.

2.4.1.1 Random Forest RF is a supervised ML algorithm made up of an ensemble of *decision trees* whose results are averaged, resulting in an increased algorithm performance level and more stable predictions. RFs consist of 3 main hyper-parameters that need to be set prior to training data:

- **trees**: number of trees contained in the ensemble;
- **mtry**: maximum number of predictors randomly sampled at each split in the tree; and
- **min_n**: minimum number of observations required for further splitting at a node.

RFs have 2 main advantages:

- They are versatile and can be used for both classification and regression; and
- When enough trees are used, RFs can overcome over-fitting,[2] a common concern with decision trees. However, the use of too many trees is a disadvantage as it slows down computations.

2.4.1.1.1 Variable Engineering Variable engineering entails configuring variables into a format that makes it easier for the model selected to be applied. With `tidymodels`, this step is done using the function `recipe()`, which specifies the pre-processing steps to be carried out on the data set. For the RF model, variable engineering was limited to the kNN-imputation of all improbable values for all predictors that was done in the data cleaning section.

[2] A model that over-fits is one that predicts well the data set used to train it but under-performs on the *test* set or any other new data set.

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         11
```

2.4.1.1.2 Model Specification Specification for the classification RF model was created using the function `rand_forest()` and the computational engine `ranger`. Tuning was specified for all 3 hyper-parameters: *trees*, *mtry* and *min_n*.

```
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   mtry = tune()
##   trees = tune()
##   min_n = tune()
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
```

2.4.1.1.3 Workflow Specification RF's pre-processing recipe and the RF model to be fit were combined in a `workflow` object.

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 0 Recipe Steps
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   mtry = tune()
##   trees = tune()
##   min_n = tune()
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
```

2.4.1.1.4 Hyper-parameter Tuning

Tuning was done using the function `tune_grid()` and the 10 cross-validations folds of the *train* set.[3] The performance metric used to compare results from hyper-parameter tuning was accuracy. The highest accuracy achieved by RF hyper-parameter tuning was 0.879.

2.4.1.1.5 Best Model Selection

The hyper-parameters corresponding to the highest accuracy achieved by RF hyper-parameter tuning were used to create the best RF model. The RF workflow was then updated to reflect the best RF model.

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 0 Recipe Steps
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   mtry = 3
##   trees = 1282
##   min_n = 19
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
```

2.4.1.1.6 Best Model Evaluation

The effectiveness of the best RF model was assessed by fitting it on the entire *train set* and evaluating on the *test* set using the function `last_fit()`. The results are shown in the Results section.

2.4.1.2 k-Nearest Neighbour

kNN is a simple, effective and commonly used ML algorithm. It allows classification of observations in a data set by comparing their attributes to attributes of "similar" labeled observations in the same data set (Bramer 2013) . The number of known observations to be used, *k*, needs to be set prior to training the model.

kNN's advantages include:

- Ease of implementation, with only one main hyper-parameter, i.e., *k*, required to implement it
- Speed: it is faster than other algorithms as it does not require a training period .

kNN's disadvantages include:

[3] As no tuning grid was specified, the function used an in-built default grid, consisting of a semi-random space-filling grid with 10 hyper-parameter combinations.

- Issues with large or high deminsional data sets due to cost of distance calculation;
- Requires variable scaling prior to applying algorithm;
- Sensitive to noisy and missing data and outliers thus requires removal or imputation.

2.4.1.2.1 Variable Engineering In addition to the kNN-imputation of improbable values done in the Data Cleaning section, the function `recipe()` was used to specify the pre-processing steps to be carried out on the data set. For the kNN model this entailed:

- creation of dummy variables for all categorical predictors; and
- normalisation and removal of skewness for all numeric predictors.

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         11
##
## Operations:
##
## Yeo-Johnson transformation on all_numeric_predictors()
## Centering and scaling for all_numeric_predictors()
## Dummy variables from all_nominal_predictors()
```

2.4.1.2.2 Model Specification Specification for the classification kNN model was created using the function function `nearest_neighbor()` and the computational engine `kknn`. Tuning was specified for the kNN hyper-parameter, *neighbours (k)*.

```
## K-Nearest Neighbor Model Specification (classification)
##
## Main Arguments:
##   neighbors = tune()
##
## Computational engine: kknn
```

2.4.1.2.3 Workflow Specification kNN's pre-processing recipe and the kNN model to be fit were combined in a `workflow` object.

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: nearest_neighbor()
##
## -- Preprocessor ----------------------------------------------------------------
## 3 Recipe Steps
```

```
##
## * step_YeoJohnson()
## * step_normalize()
## * step_dummy()
##
## -- Model -----------------------------------------------------------------------
## K-Nearest Neighbor Model Specification (classification)
##
## Main Arguments:
##   neighbors = tune()
##
## Computational engine: kknn
```

2.4.1.2.4 Hyper-parameter Tuning Tuning was done using the function `tune_grid()` and the 10 cross-validations folds of the *train* set.[4] The highest accuracy achieved by kNN hyper-parameter tuning was 0.879.

2.4.1.2.5 Best Model Selection The hyper-parameter corresponding to the highest accuracy achieved by kNN hyper-parameter tuning was used to create the best kNN model. The kNN workflow was then updated to reflect the best kNN model.

```
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: nearest_neighbor()
##
## -- Preprocessor ----------------------------------------------------------------
## 3 Recipe Steps
##
## * step_YeoJohnson()
## * step_normalize()
## * step_dummy()
##
## -- Model -----------------------------------------------------------------------
## K-Nearest Neighbor Model Specification (classification)
##
## Main Arguments:
##   neighbors = 15
##
## Computational engine: kknn
```

2.4.1.2.6 Best Model Evaluation The effectiveness of the best kNN model was assessed by fitting it on the entire *train set* and evaluating on the *test* set using the function `last_fit()`. The results are shown in the Results section.

[4] As no tuning grid was specified, the function used an in-built default grid, consisting of a semi-random space-filling grid with 10 hyper-parameter combinations.

2.5 Results

2.5.1 Best Model Evaluation

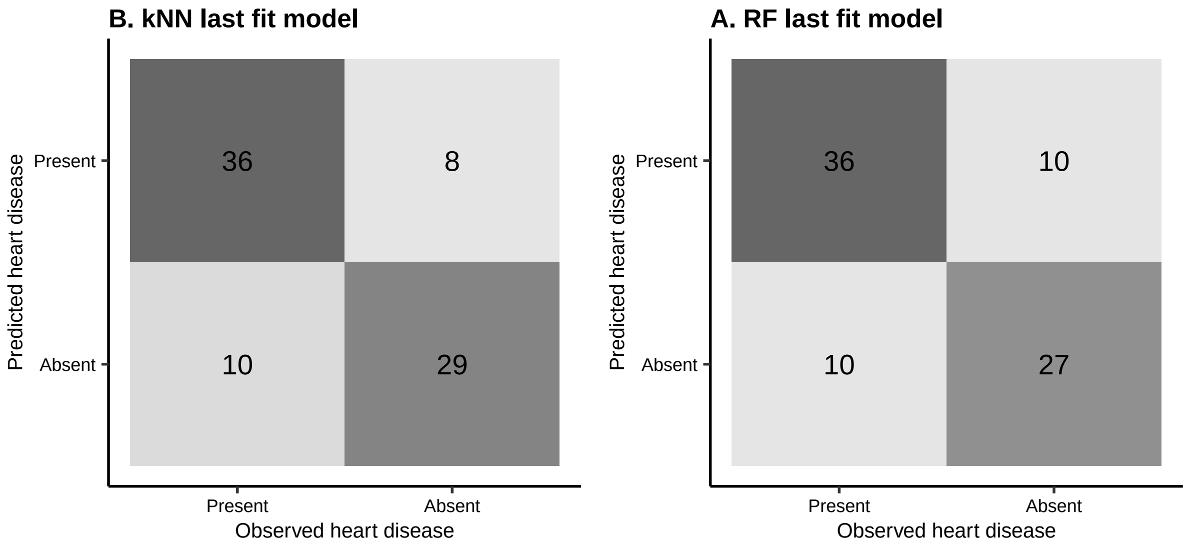


Figure 10: Confusion matrix heatmaps for A) RF and b) kNN last fit models

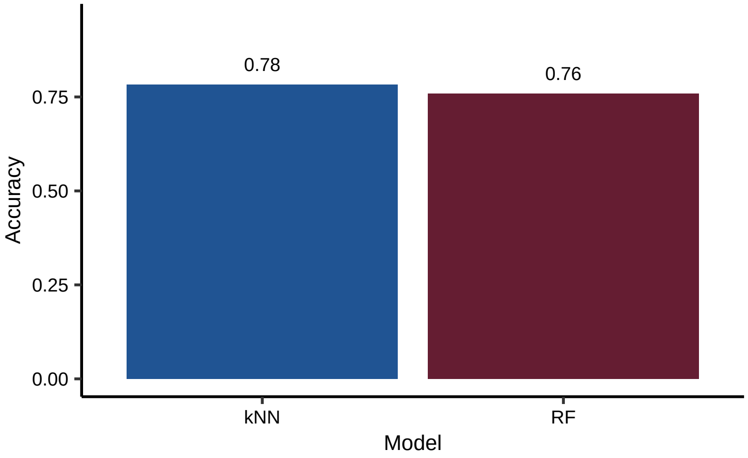


Figure 11: Performance metrics for RF and kNN last fit models

In this section, the effectiveness of the best RF and kNN models were compared in order to select the final model. The results of fitting the best RF and kNN models on the entire *train* set and evaluating on the *test* set were analyzed first. The best kNN model resulted in 10 false negatives and 8 false positives (Figure 10A), for an accuracy of 0.78. The best RF model resulted in 10 false negatives and 10 false positives (Figure 10B) for an accuracy of of 0.76 (Figure 11).

2.5.2 Final Model Evaluation

The Best Model Evaluation section showed that of the 2 ML algorithms used, the kNN model resulted in the highest accuracy when fitted on the entire *train* set and evaluated on the *test* set and was thus chosen

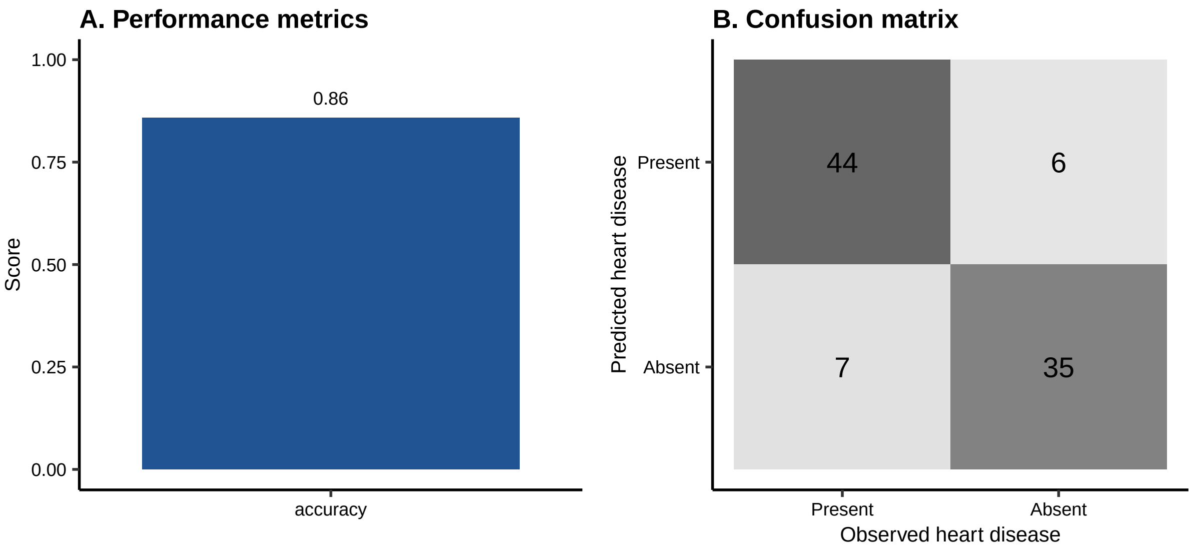


Figure 12: Final kNN model results - trained on entire model data set and evaluated on validation data set

as the final model. In this section, this final model's effectiveness on new data was assessed by fitting it on the entire *model* set and evaluating it on the *validation* (or *final hold-out*) set. The final kNN model resulted in an accuracy of 0.86 (Figure 12). Out of the 92 patients in the validation set, 51 had HD and 41 did not. Of the 51 who had HD, the model correctly predicts presence of HD in 44 and out of the 41 who did not have HD, the model correctly predicts absence of HD in 35 (Figure 12).

3 Conclusion

The aim of this project was to compare the performance of 2 ML algorithms in a binary classification prediction project, which used the heart failure prediction data set, *heart*, to predict the *presence* of HD in patients. When comparing the 2 ML algorithms explored, the kNN model resulted in the highest accuracy of 0.78. This final kNN resulted in an accuracy 0.86 of when evaluated on new data.

3.1 Limitations and Future Work

Some limitations for this project include random tuning of hyper-parameters and lack of other potentially strong predictors of HD, such as obesity, smoking, physical activity and diet. Future work could explore more elaborate and efficient ways to tune hyper-parameters. Moreover, inclusion of stronger predictors of HD would result in higher performing models.

References

AHA. n.d.a. "Hoe to Manage Blood Sugar." https://www.heart.org/-/media/Healthy-Living-Files/LE8-Fact-Sheets/LE8_How_to_Manage_Blood_Sugar.pdf.

———. n.d.b. "What Is High Blood Pressure?" https://www.heart.org/-/media/files/health-topics/answers-by-heart/what-is-high-blood-pressure.pdf.

Akella, Aravind, and Sudheer Akella. 2021. "Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward an Open Source Solution. Future Science OA Volume 7, Number 6, Pages FSO698, 2021."

Benjamin, Emelia J, Salim S Virani, Clifton W Callaway, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Stephanie E Chiuve, et al. 2018. "Heart Disease and Stroke Statistics—2018 Update: A Report from the American Heart Association." *Circulation* 137 (12): e67–492.

Bramer, Max. 2013. "Introduction to Classification: Naive Bayes and Nearest Neighbour." In *Principles of Data Mining*, 21–37. Springer.

Breiman, Leo. 2001a. "Random Forests." *Machine Learning* 45 (1): 5–32.

———. 2001b. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Chou, Roger, Bhaskar Arora, Tracy Dana, Rongwei Fu, Miranda Walker, and Linda Humphrey. 2011. "Screening Asymptomatic Adults with Resting or Exercise Electrocardiography: A Review of the Evidence for the US Preventive Services Task Force." *Annals of Internal Medicine* 155 (6): 375–85.

Curry, Susan J, Alex H Krist, Douglas K Owens, Michael J Barry, Aaron B Caughey, Karina W Davidson, Chyke A Doubeni, et al. 2018. "Screening for Cardiovascular Disease Risk with Electrocardiography: US Preventive Services Task Force Recommendation Statement." *Jama* 319 (22): 2308–14.

Deo, Rahul C. 2015. "Machine Learning in Medicine." *Circulation* 132 (20): 1920–30.

Detrano, ROBERT, JOHN Yiannikas, ERNESTO E Salcedo, G Rincon, RT Go, G Williams, and J Leatherman. 1984. "Bayesian Probability Analysis: A Prospective Demonstration of Its Clinical Utility in Diagnosing Coronary Disease." *Circulation* 69 (3): 541–47.

fedesoriano. 2021. "Heart Failure Prediction Dataset." https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Fuchs, Flávio D, and Paul K Whelton. 2020. "High Blood Pressure and Cardiovascular Disease." *Hypertension* 75 (2): 285–92.

Grundy, Scott M, Neil J Stone, Alison L Bailey, Craig Beam, Kim K Birtcher, Roger S Blumenthal, Lynne T Braun, et al. 2019. "2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Circulation* 139 (25): e1082–1143.

Hlatky, Mark A. 1999. "Exercise Testing to Predict Outcome in Patients with Angina." *Journal of General Internal Medicine* 14 (1): 63.

Larsen, CT, J Dahlin, H e-al Blackburn, H Scharling, M Appleyard, B Sigurd, and P Schnohr. 2002. "Prevalence and Prognosis of Electrocardiographic Left Ventricular Hypertrophy, ST Segment Depression and Negative t-Wave. The Copenhagen City Heart Study." *European Heart Journal* 23 (4): 315–24.

Lim, Yoke Ching, Swee-Guan Teo, and Kian-Keong Poh. 2016. "ST-Segment Changes with Exercise Stress." *Singapore Medical Journal* 57 (7): 347.

Members, Writing Committee, Martha Gulati, Phillip D Levy, Debabrata Mukherjee, Ezra Amsterdam, Deepak L Bhatt, Kim K Birtcher, et al. 2021. "2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: Executive Summary: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines." *Journal of the American College of Cardiology* 78 (22): 2218–61.

NHLBI. n.d. "Know the Differences: Cardiovascular Disease, Heart Disease, Coronary Heart Disease." https://www.nhlbi.nih.gov/sites/default/files/media/docs/Fact_Sheet_Know_Diff_Design.508_pdf.pdf.

North, Brian J, and David A Sinclair. 2012. "The Intersection Between Aging and Cardiovascular Disease." *Circulation Research* 110 (8): 1097–1108.

Olsen, Michael H, Sonia Y Angell, Samira Asma, Pierre Boutouyrie, Dylan Burger, Julio A Chirinos, Albertino Damasceno, et al. 2016. "A Call to Action and a Lifecourse Strategy to Address the Global Burden of Raised Blood Pressure on Current and Future Generations: The Lancet Commission on Hypertension." *The Lancet* 388 (10060): 2665–2712.

Peters, Sanne AE, Paul Muntner, and Mark Woodward. 2019. "Sex Differences in the Prevalence of, and Trends in, Cardiovascular Risk Factors, Treatment, and Control in the United States, 2001 to 2016." *Circulation* 139 (8): 1025–35.

Sajda, Paul. 2006. "Machine Learning for Detection and Diagnosis of Disease." *Annu. Rev. Biomed. Eng.* 8: 537–65.

Sports Medicine, American College of, Paul D Thompson, Barry A Franklin, Gary J Balady, Steven N Blair, Domenico Corrado, NA Mark Estes III, et al. 2007. "Exercise and Acute Cardiovascular Events: Placing the Risks into Perspective: A Scientific Statement from the American Heart Association Council on Nutrition, Physical Activity, and Metabolism and the Council on Clinical Cardiology." *Circulation* 115 (17): 2358–68.

Whelton, Paul K, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, et al. 2018. "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Journal of the American College of Cardiology* 71 (19): e127–248.

WHO. 2021. "Cardiovascular Diseases (CVDs)." *Https://Www.who.int/News-Room/Fact-Sheets/Detail/Cardiovascular-Diseases-(Cvds)*.