ST3189 MACHINE LEARNING COURSEWORK

Unsupervised and Supervised Learning

Contents

1. Unsupervised Learning

1.1 Research Problem

Country development level can vary significantly depending on the socio-economic factors such as the living standard and economic growth. It's important to have a standard measure when classifying the countries into categories. This will help humanitarian aids to be assigned effectively to those countries which need them the most. Hence, using unsupervised machine learning models, we aim to research on the following problems:

1. What are the main factors affecting the level of overall country development?
2. What are the characteristics of the major subgroups of the countries?

1.2 Dataset and Variables

The "Country Data" dataset consists of information on various socio-economic factors collected from 167 countries. There are 10 variables in the dataset, including 9 continuous variables and 1 string variables. Since the purpose of unsupervised learning is to identify patterns and relationship within the data instead of making predictions, the dataset has no labelled variable. From basic exploratory data analysis, there is no missing or null value in the dataset. It's found that some variables are highly correlated to each other, such as life expectancy and child mortality (-0.89), total fertility and child mortality (0.85), per capita GDP and income (0.90).

	Name	**Description**
1	country	Name of the country. (character)
2	child_mort	Death of children under 5 years old per 1000 live births. (numerical)
3	exports	Export of goods and services per capita, given as % of GDP per capita. (numerical)
4	health	Total health spending per capita, given as % of GDP per capita. (numerical)
5	imports	Import of goods and services per capita, given as % of GDP per capita. (numerical)
6	income	Net income per person. (integer)
7	inflation	Annual growth rate of total GDP. (numerical)
8	life-expec	Average years a newborn child would live. (numerical)
9	total_fer	Number of children would be born to each woman. (numerical)
10	gdpp	GDP per capita. (integer)

1.3 Principal Component Analysis (PCA)

PCA is chosen as a dimensionality reduction approach to identify the principal components that explain the most variance of a particular dataset. In other words, this approach helps to extract the most representative variables that collectively explains most of the variability in the original dataset. (James et al, 2017). From the earlier exploratory data analysis, it's found that there are some correlated variables, which suggests that PCA analysis may lead to some useful results.

Before conducting the PCA, standardization of the continuous variables is required to make sure every variable contribute equally to the analysis. This is achieved through ranging the continuous variables to have mean zero and standard deviation of one. Two variables (gdpp, income) with integer data type are converted to numerical data type.

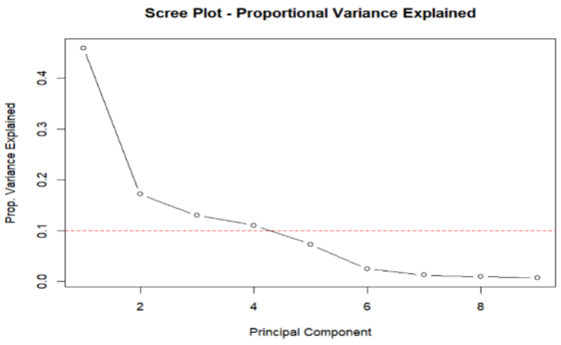


Figure 1. Proportional Variance Explained

From the Scree plot generated after running the PCA analysis, it shows a total of 9 components. It can be observed that there are 4 representative components with each explaining over 10% of the total variance in the original dataset.

1.4 K-Means Clustering

While PCA aims to visualise a low-dimensional representation of the dataset, clustering is another method to simplify the dataset by categorizing the data observations into subgroups to identify common characteristics within the same cluster.

K-Means clustering is one of the clustering methods, the algorithm works by randomly assign a cluster number, 1 to K, to each observation. This is followed by calculating the initial cluster centroid, then reassigning the observations to their closest centroid until no change can be made to the cluster assignment. In other words, the finalised cluster assignment attains the minimal Within-Cluster Variation (WCV).

A few approaches are available to identify the optimal K (number of clusters), such as the Elbow method, Silhouette method, and the Gap Statistic (Nallathambi, J.). Based on the results from R, there is no significant "elbow" can be seen from the scree plot. Instead, the Silhouette method and Gap Statistic show that the optimal K (number of clusters) is 5 and 3 respectively.

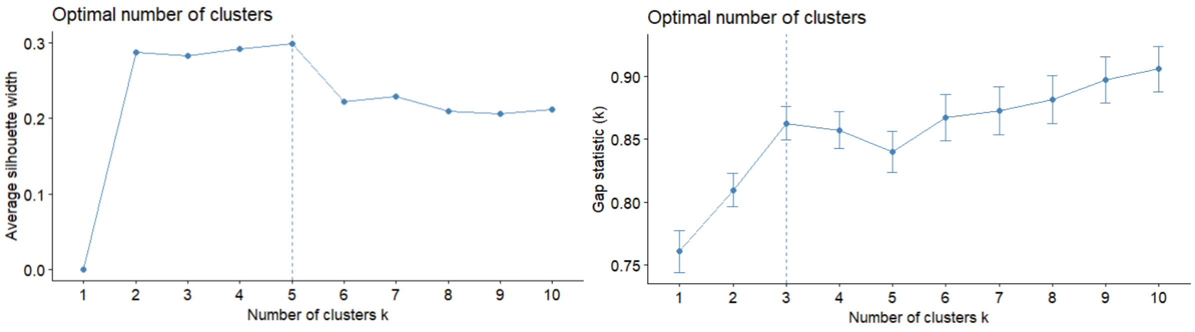


Figure 2. Silhouette method & Gap statistic method for optimal K

1.5 Hierarchical Clustering

Hierarchical clustering takes the bottom-up approach in building the clusters, resulting in a upside down tree diagram, called dendrogram. The bottom of the dendrogram consists of all the observations, which are clusters of their own. They are very similar to each other. Moving up in the dendrogram, similar clusters are merged until one final cluster consisting of all the observations.

In hierarchical clustering, it's important to identify the dissimilarity measure and linkage method. Dissimilarity measure is a way to calculate the distance or correlation between the data observations. Common measures are the Euclidean distance and correlation-based distance. Euclidean distance measure is chosen in this analysis. Linkage method defines how the clusters are being merged. Complete, Single, Average and Ward's Linkage are considered in the analysis.

The clustering criterion is set to be K = 5, according to the optimal number of clusters obtained from K-Means clustering. Based on the dendrogram generated using Euclidean distance measure and Ward's linkage method, the 5 clusters are mostly evenly distributed.

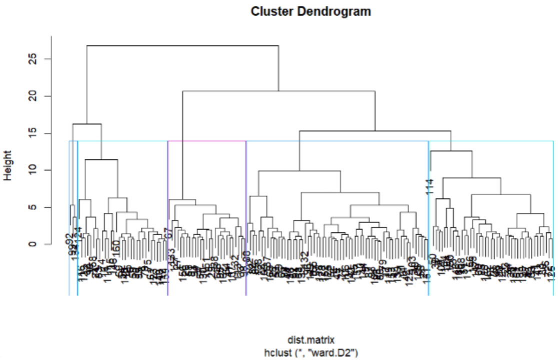
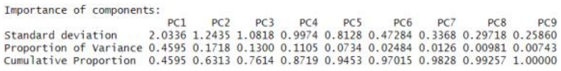


Figure 3. Cluster dendrogram

1.6 Results



```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6    PC7     PC8     PC9
Standard deviation     2.0336 1.2435 1.0818 0.9974 0.8128 0.47284 0.3368 0.29718 0.25860
Proportion of Variance 0.4595 0.1718 0.1300 0.1105 0.0734 0.02484 0.0126 0.00981 0.00743
Cumulative Proportion  0.4595 0.6313 0.7614 0.8719 0.9453 0.97015 0.9828 0.99257 1.00000
```

```
              PC1     PC2     PC3      PC4
child_mort -0.420 -0.1929  0.0295 -0.37065
exports     0.284 -0.6132 -0.1448 -0.00309
health      0.151  0.2431  0.5966 -0.46190
imports     0.161 -0.6718  0.2999  0.07191
income      0.398 -0.0225 -0.3015 -0.39216
inflation  -0.193  0.0084 -0.6425 -0.15044
life_expec  0.426  0.2227 -0.1139  0.20380
total_fer  -0.404 -0.1552 -0.0195 -0.37830
gdpp        0.393  0.0460 -0.1230 -0.53199
```

Figure 4. Principal components and its loadings

There are 4 major factors explaining a cumulative of 87.19% of the total variation in the dataset. In each principal component, the loadings are calculated for each variable. PC1 is relatively more weighted on child_mort, life_expec and total_fer, representing the **Population Well-being**. PC2 is relatively more weighted on exports and imports, representing the **Macro-economic Well-being**. PC3 is relatively more weighted on inflation and health, representing the **Living Standard**. PC4 is more weighted on gdpp, health and income, representing the **Household Economic Well-being**. For example, the linear combination of PC1 will be in this form:

$$z_{i1} = -0.42\text{child\_mort}_i + 0.284\text{exports}_i + \cdots + 0.393\text{gdpp}_i$$

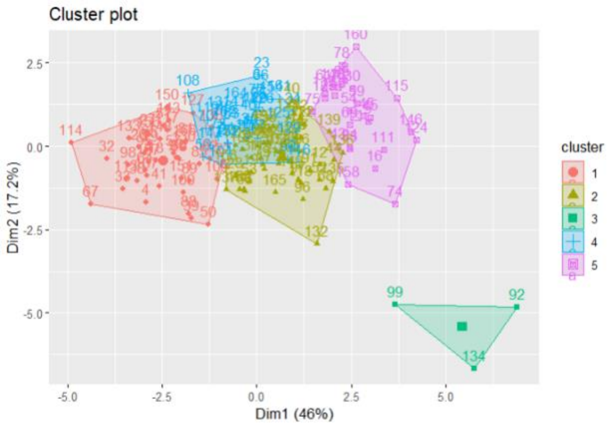
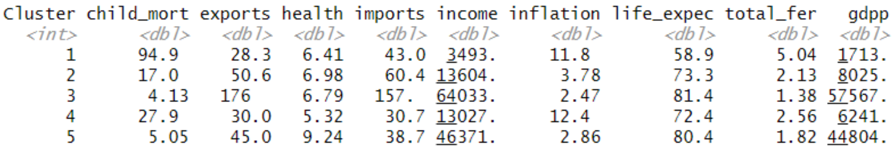


Figure 5. Cluster plot

Meanwhile, the cluster plot shows the cluster distribution based on 2 dimensions across the x and y-axis. This corresponds to the PCA results, where the 1st principal component (Dim1) explaining about 46% of the total variation and the 2nd principal component (Dim2) explaining about 17.2% of the total variation. Hence, it can be understood that the overlapping area between the clusters is due to the lack of the remaining dimensions.



```
Cluster child_mort exports health imports income inflation life_expec total_fer   gdpp
  <int>      <dbl>   <dbl>  <dbl>   <dbl>  <dbl>     <dbl>      <dbl>     <dbl>  <dbl>
      1      94.9     28.3   6.41    43.0  3493.     11.8        58.9      5.04  1713.
      2      17.0     50.6   6.98    60.4 13604.      3.78       73.3      2.13  8025.
      3       4.13   176     6.79   157.  64033.      2.47       81.4      1.38 57567.
      4      27.9     30.0   5.32    30.7 13027.     12.4        72.4      2.56  6241.
      5       5.05    45.0   9.24    38.7 46371.      2.86       80.4      1.82 44804.
```

Figure 6. Cluster mean

When applying the clustering results back to the original dataset, mean values of the variables can be calculated for each cluster. For example, cluster 2 and 4 share common characteristics like **moderate child_mort and total_fer**. Cluster 3 and 5 share common characteristics such as **high income, low inflation, and long life_expec**.

2. Supervised Learning – Regression Task

2.1 Research Problem

Insurance charges are often based on the level of risks associated with the insured individual. At times, companies can find it difficult to correctly assess the amount of money claimed by the policyholder. Therefore, the insurance companies may have set ineffective pricing models. Using regression analysis, we aim to research and analyse the following problems:

1. What are the influencing factors for predicting insurance charges?
2. What is the most suited model to perform regression analysis?

2.2 Dataset and Variables

The "Insurance" dataset consists of information of insurance policyholders on their demographic information and the amount of insurance charges. In this regression analysis, insurance charges will be treated as the target variable Y, while other variables are predictors **X** used for predicting insurance charges. No missing or null values are found in this dataset. Data types for continuous variables are changed to numerical, and categorical variables changed to factor data type. Furthermore, dummy variables are created for sex, smoker, and region, for the convenience of conducting regression analysis.

	Name	**Description**
1	index	Index number. (integer)
2	age	Age of the customer. (integer)
3	sex	Male, Female. (character)
4	bmi	BMI index. (numerical)
5	children	Number of children the customer has. (integer)
6	smoker	Yes, No. (character)
7	region	Southwest, Southeast, Northwest, Northeast. (character)
8	charges	Insurance charges of the customer. (numerical)

2.3 Linear Regression Analysis

Linear regression is used for predictive modelling and understanding the relationship between a target variable and one or more independent variables. The mechanism behind linear regression involves fitting a linear equation to the observed data points with the objective to minimize the errors between the predicted and actual values of the target variable.

To prepare the data for analysis, all continuous variables are standardized except the target variable to ensure each variable contribute equally to the modelling process. A 70:30 train test split is applied onto the dataset. Subset selection is not considered in this analysis since there are not many features in the dataset and the size of the data is not big. Multicollinearity is checked through Variance Inflation

Factor (VIF) and the result does not show significant amount of multicollinearity. Hence, shrinkage methods such as Ridge and Lasso regression are not considered in this analysis (Srinivasan, P.).

Further check on the linear regression assumptions can be done through diagnostic plots (Kim, 2015). The "Residual vs Fitted" plot shows if residuals have non-linear patterns. Ideally, the residuals should spread equally along a horizontal line with no distinct patterns. The "Normal Q-Q" plot shows if residuals are normally distributed by checking if the residuals lie on the straight line. The "Scale-Location" plot is used to check for homoscedasticity by looking if the residuals spread equally across the horizontal line. Lastly, the "Residuals vs Leverage" plot shows if there are influential outliers which will affect the results significantly. Obviously, there are some violations of linear regression assumptions, leading to biased and inconsistent estimation of the model.

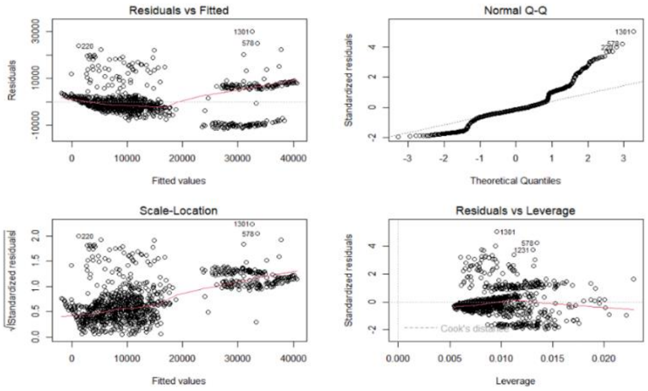


Figure 7. Diagnostic plots

Subsequently, polynomial regression is applied to the dataset based on the variable age, bmi, and children. 10-fold cross-validation has been applied to the degree 1, 2, and 3 polynomial regressions. The cross-validation error for degree 1, 2, 3 are 0.8835, 0.8824, and 0.8858 respectively. Degree 2 polynomial regression has relatively lower cross-validation error, so this would lead to improved modelling results.

2.4 Tree-based models

Regression tree increases the interpretability of the model by splitting the data into various subtrees. Using the "Insurance" dataset, a decision tree with 5 terminal nodes has been created. The cut-points are "smoker.yes = 0", "age < 45", "bmi < 30" and "age < 42". The choices of these cut-points will give the resulting tree the lowest Residual Sum of Squares (RSS).

Initially when no splits have been done, the mean response value is 13000, consisting 100% of the observed data points. After the 1st cut point, the observations are divided into 2 groups, with mean response values of 8318 and 32000. The tree building process stopped when the data has been divided into 5 groups.

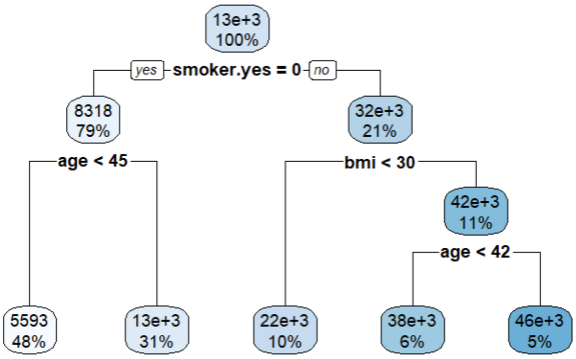


Figure 8. Regression tree

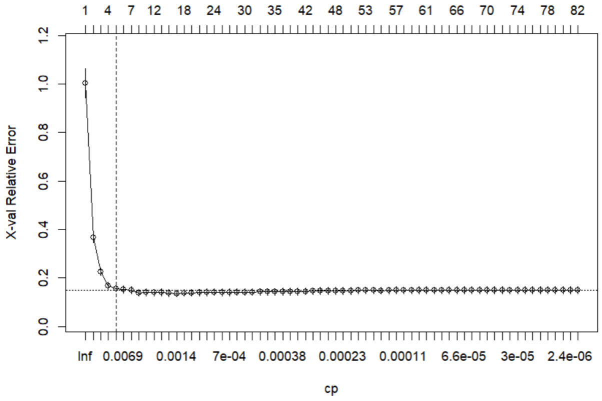


Figure 9. Cross validation error

Looking into the tree building process, a 70:30 train test split is applied to the dataset, 10-fold cross validation is used to train and test the model for 10 times, each time with a different test set. To prevent building a large tree which may lead to overfitting the data, cost complexity pruning is applied to stop the tree building process when the reduction in RSS is high enough.

The result shows that optimal reduction in the RSS is achieved when we have a tree with 5 terminal nodes. At this point, we obtain a simple tree while having a low cross-validation error.

However, building a single tree may sacrifice the prediction accuracy. Ensemble learning is considered in the subsequent modelling process. Gradient boosting is one of the ensemble methods used to build a more robust decision tree, which is obtained by building multiple decision trees, in a sequential manner. The previous tree is used to update the results of the following tree. Note that the tree is not fit to the response values but the residuals. Certain parameters can be adjusted to give the best results, such as the number of boosting rounds, the boosting learning rate, and the boosting complexity. It's resulted that the optimal RMSE is obtained through gradient boosting, compared to linear regression and regression tree.

2.5 Results

Tree-based models perform better on the dataset compared to linear regression models. Linear regression assumptions are violated, thus making the results unreliable. From regression tree, we can identify **three major influencing factors** for predicting insurance charges, which are **smoking**, **age**, and **BMI**. If the insured individual is a smoker, with age greater than 42 and BMI larger than 30, that person will most likely claim more than the rest. Hence, insurance company can adjust its pricing model based on the prediction results to cover the costs.

Root Mean Squared Error (RMSE) is used as the evaluation metrics throughout the regression analysis to measure the average deviation of the predicted values from the actual values. Gradient boosting model yields the best result with average deviation of $4870 from the actual insurance charges. It shows improvement compared to the single decision tree model.

	Linear Regression	Decision Tree	Gradient Boosting
RMSE	6196.277	5241.7	**4870.048**

3. Supervised Learning – Classification Task

3.1 Research Problem

Based on WHO research statistic (WHO, 2021), cardiovascular disease is the leading cause of death globally, representing 32% of global deaths. Among those, 85% of them were due to heart attack and stroke. There are various behavioural risk factors associated with heart attack, such as unhealthy diet and habits. Under classification task, we aim to research on the following problems:

1. What are the factors impacting the occurrence of heart attack?
2. What is the best suited classification model?

3.2 Dataset and Variables

The "Heart Attack Analysis & Prediction Dataset" consists of 14 variables with "output" as target variable. Data types for certain variables are corrected from numerical to categorical, such as age, sex, and cp. From the correlation matrix, the target variable output is moderately correlated to cp (0.43), thalachh (0.42), exng (0.44) and oldpeak (0.43). Stepwise selection is applied to the dataset using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). 7 out of 13 predictors are selected based on the BIC selection criterion, including the top 3 important variables: caa, cp, and thall.

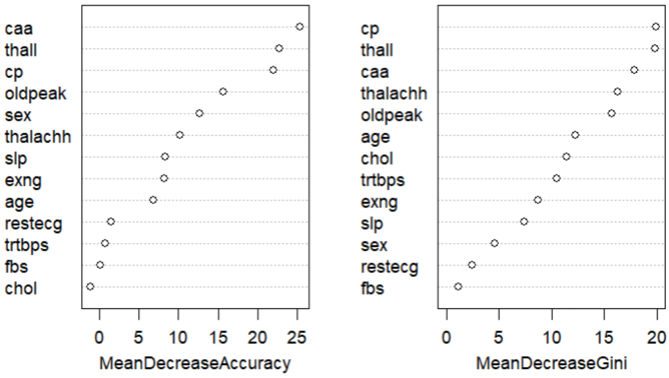


Figure 10. Important variables

	Name	Description
1	age	Age of the patient
2	sex	1 = Male, 0 = Female
3	cp	Chest pain type
4	trtbps	Resting blood pressure
5	chol	Serum cholesterol
6	fbs	Fasting blood sugar
7	restecg	Resting electrocardiographic result

	Name	Description
8	thalachh	Max. heart rate achieved
9	exng	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slp	slope of the peak exercise ST segment
12	caa	Number of major vessels
13	thall	Thalassemia type
14	output	Heart disease

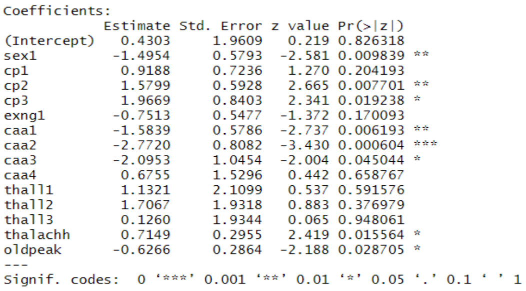
3.3 Modelling

Linear models

Logistic regression (LR) and Linear Discriminant Analysis (LDA) are conducted as part of the linear modelling process. Numerical variables are scaled to help improve the optimization process and make model coefficients more interpretable. 70:30 train test split is applied and the predicted probability are classified into either 0 (no heart attack) when Pr(output < 0.5) or 1 (heart attack) when Pr(output > 0.5). Categorical variables are transformed into dummy variables.

Linear Discriminant Analysis (LDA) is a classification technique that models the distribution of predictors separately in each class and then uses Bayes' theorem to estimate the probability of class membership for new observations.

The logistic regression output shows that certain predictors are statistically significant in predicting the target variable (with asterisks behind). This means those predictors are less likely to have a coefficient of zero. For example, having certain types of chest pains (cp2, cp3) are associated with increased probability of heart attack occurring, while having certain number of major vessels (caa1, caa2, caa3) are associated with decreased probability of heart attack occurring.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.4303     1.9609   0.219 0.826318
sex1          -1.4954     0.5793  -2.581 0.009839 **
cp1            0.9188     0.7236   1.270 0.204193
cp2            1.5799     0.5928   2.665 0.007701 **
cp3            1.9669     0.8403   2.341 0.019238 *
exng1         -0.7513     0.5477  -1.372 0.170093
caa1          -1.5839     0.5786  -2.737 0.006193 **
caa2          -2.7720     0.8082  -3.430 0.000604 ***
caa3          -2.0953     1.0454  -2.004 0.045044 *
caa4           0.6755     1.5296   0.442 0.658767
thall1         1.1321     2.1099   0.537 0.591576
thall2         1.7067     1.9318   0.883 0.376979
thall3         0.1260     1.9344   0.065 0.948061
thalachh       0.7149     0.2955   2.419 0.015564 *
oldpeak       -0.6266     0.2864  -2.188 0.028705 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11. Logistic regression output

Tree-based models

Classification tree works in a similar way to regression tree. It recursively splits the data into subsets, aiming to create homogenous subset with regards to the target variable, which is heart attack output in this case. In this analysis using rpart function, each split is chosen to maximize the subset homogeneity using Gini impurity index. The model stops splitting the tree further when there is minimum improvement on Gini impurity index.

The classification tree gives 7 terminal nodes, 4 groups representing no heart attack (output = 0) and 3 subgroups representing heart attack (output = 1). Based on the 1st cut point, those patients with 1, 2, or 3 major vessels (caa = 1, 2, 3) make up 41% of the training dataset, while those without 1, 2, or 3 major vessels (caa = 1, 2, 3) make up 58% of the training data. Among these two groups of patients, percentage of those having heart attack are 8% and 46% respectively. This shows that having major vessels will prevent suffering from heart attack.

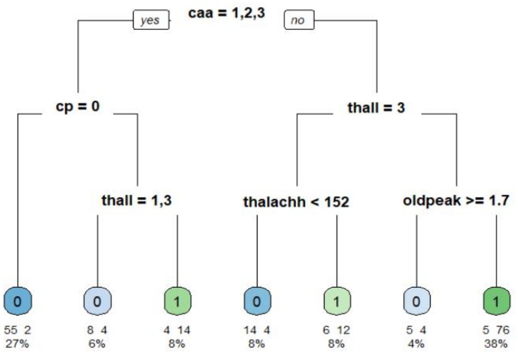


Figure 12. Classification tree

Random forest is applied as an ensemble learning method to enhance the classification tree result by randomly selecting data subsets and features for multiple times and generate multiple trees.

Distance-based models

K-Nearest Neighbour (KNN) and Support Vector Machines (SVM) are considered for building distance-based models. A 70:30 train test split is performed. Since the KNN and SVM are distance-based models, it's necessary to scale the continuous variables to avoid distance sensitive biases.

KNN works by calculating the distance between each point through distance metrics such as the Euclidean distance. A value of k is defined to decide the number of nearest data points to the new observation.

SVM aims to find the optimal hyperplane to separate the data. Initially, it maps each point in a higher dimensional space and aims to find a decision boundary to separate the data points. For example, using the Radial Basis Function (RBF) kernel, the data points are mapped out across the predictor oldpeak and thalachh, which are both continuous variables.

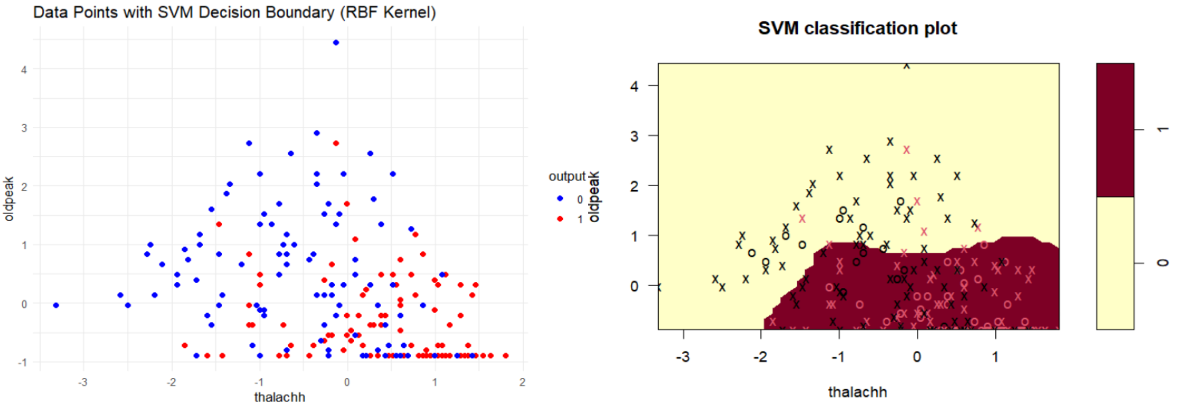


Figure 13. SVM - data points and decision boundary between oldpeak and thalachh

3.4 Results

Confusion matrix is used to evaluate the model performance. Accuracy, with the formula $\frac{True\ Positives + True\ Negatives}{All\ data\ observations}$, representing the proportion of data observations being correctly predicted, are computed based on the results of each model. We can see that Logistic Regression yields the highest accuracy with the selected subset for analysis. However, there is existing research done (Mtambo, G.) to the same dataset and it shows that KNN yields the highest accuracy.

		LR	LDA	CT	RF	KNN	SVM	Naïve Bayes
Accuracy	Modelling results	**0.8242**	0.8132	0.7889	0.8022	0.7802	0.7802	
	Existing research	0.86			0.825	**0.912**	0.842	0.877

Based on the results from logistic regression and classification tree, **caa** (number of major vessels) and **cp** (chest pain type) **have the most impact for predicting the occurrence of heart attack**. These two are also appeared among the top important variables from the Random Forest analysis we performed at the start of the classification task.

Receiver Operating Characteristic (ROC) curve illustrates the trade-off between true positive rate (sensitivity) and false positive rate (1 – specificity). An ideal ROC curve should show high true positive rate and low false positive rate. ROC curves for all classification models are plotted in one diagram, we can see that logistic regression has the best performance, the rest of the models showed similar ROC curves.

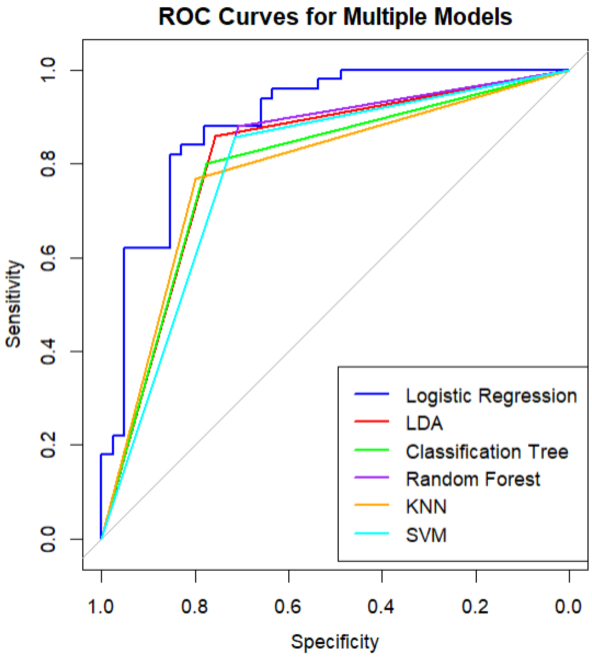


Figure 14. ROC curves for classification models

Datasets links

Unsupervised learning: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

Regression: https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender

Classification: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

References

Cardiovascular diseases (CVDs). (11 6, 2021). Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Gareth James, D. W. (2017). *An Introduction to Statistical Learning with Applications in R.* Springer .

Kim, B. (21 9, 2015). *Understanding Diagnostic Plots for Linear Regression Analysis*. Retrieved from University of Virginia Library: https://library.virginia.edu/data/articles/diagnostic-plots

Mtambo, G. (n.d.). *Heart Attack Analysis & Prediction Using R*. Retrieved from RPubs: https://rpubs.com/GiftMtambo/857615

Nallathambi, J. (18 6, 2018). *R Series — K means Clustering (Silhouette)*. Retrieved from Medium: https://medium.com/codesmart/r-series-k-means-clustering-silhouette-794774b46586

Srinivasan, P. (3 10, 2023). *"Mastering Ridge Regression: Taming Multicollinearity and Overfitting in Linear Models*. Retrieved from Medium: https://medium.com/@psrinivasan028/mastering-ridge-regression-taming-multicollinearity-and-overfitting-in-linear-models-b54b969c65e9