Machine Learning Coursework

ST3189

200552574

3 April 2024

Table of Contents

1. Unsupervised Learning and Classification

1.1. Substantive Issue

For unsupervised learning and classification task we will be using heart disease prediction. Men are more prone to heart diseases than female. According to Dr Manav from New Victoria hospital about 1 in 7 will die due to disease of the heart arteries compared to women which is 1 in 11 (Dr Manav Sohal, 2022). What other factors could be significant for predicting that someone has heart disease. We want to also figure the best classification model for prediction.

1.2. Research Question

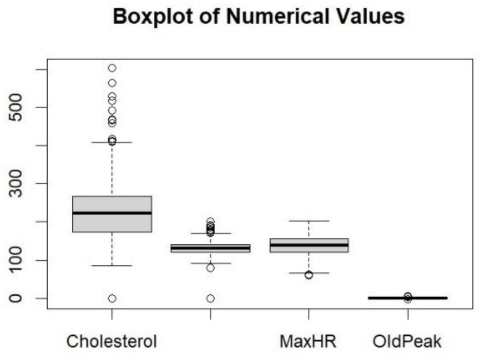
Through this I want to find if there are other significant factors leading to heart disease. These factors could be helpful to predict a person with higher chance of heart disease. Which of classification models will be best at predicting heart disease. this prediction model must be the most accurate so that we can use for future data. in the medical context it is important to be accurate for diagnosing or predicting if someone is at risk due to these factors

1.3. Dataset & Variables

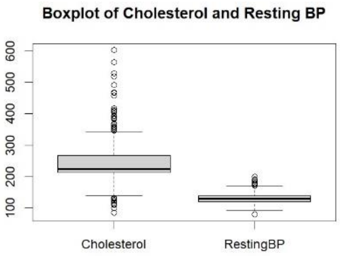
1. Age : age of the patients [years]
2. Sex : sex of the patient [Male = 0, Female = 1]
3. Chest Pain Type: type of chest pain [Atypical Angina (ATA) = 0, Non-Anginal Pain(NAP) = 1, Asymptomatic (ASY) = 2, Typical Angina (TA) = 3]
4. Resting BP : resting blood pressure [mm Hg]
5. Cholesterol : serum cholesterol [mm/dl]
6. Fasting BS : fasting blood sugar [if fasting BS > 120 mg/dl = 1, otherwise = 0]
7. Resting ECG : resting electrocardiogram results[Normal = 1, Having ST-T wave abnormality (ST) = 2, showing probable or definite left ventricular hypertrophy by Estes' criteria (LVH) = 3]
8. Max HR : maximum heart rate achieved.
9. Exercise Angina: exercise induced angina [No (N) = 1, Yes (Y) = 2]
10. Oldpeak : old peak = ST[Numeric value measured in depression]
11. St_Slope : the slope of the peak exercise ST segment [Upsloping (Up) = 0, Flat (Flat) = 1, Downsloping (Down) = 2]
12. HeartDisease : output class/ target variable [Normal = 0, Heart Disease = 1]

1.4. Classification Task

1.4.1. Methodology



For classification we will be using Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN). These three techniques will be used to create a prediction model and logistic regression will also be used to check if sex of the person is a significant figure to predicting that a person has heart disease. The columns that have words as values has been changed to the numerical values as shown in the dataset. The figure on the left shows that both cholesterol and Resting BP has



zero values. However, cholesterol and blood pressure cannot be zero. Instead of removing these rows with zero values, we keep as the other columns could have crucial information for analysis. We replace the zero values with median value as they are less sensitive to outliers.

1.4.2 Analysis

1.4.2.1 Logistic Regression

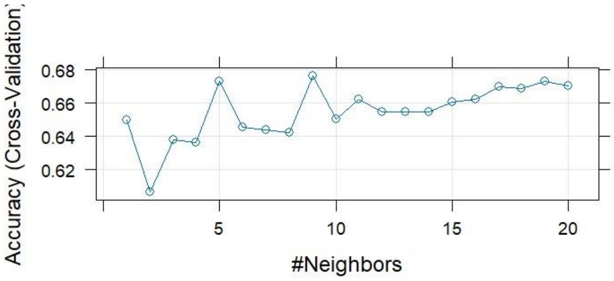
Logistic regression is like linear regression. For logistic regression, y-variable (target) is categorical. In this case y-variable is the heart disease where normal has the value 1 and heart disease has the value as 0. We first split the data into train (70%) and test (30%) set. After splitting, we use the train set to train the model. From the model we can observe that sex is most significant variable to predict the presence of heart disease. There are other factors that are significant to predict. The other variables are chest pain type, fasting blood sugar and ST-slope are also the most significant variables. We found the mean for train set. The mean shows the accuracy of this model which is 86%. Then we use this model for the test set which has an accuracy of 87%. Both train and test set has quite high accuracy and they are about the same. This shows that this model is a good model to use for prediction. We then plot the confusion matrix which will show the count for the prediction the model made against the actual value.

1.4.2.2. Support Vector Machine (SVM)

SVM is effective to ensure better generalisation to unseen data. SVM can be used foe linear and non-linear classification task. For SVM heart disease column should not be factorised. We create dummy variables for data before splitting them train and test set in the ratio of 70:30. We tune the radial using the train set first to get the optimum performance and parameter. After tuning, the best parameter for cost is 10 and for gamma is 0.01. The best performance values are 0.1353846 which gives about 86% accuracy. 86% accuracy is good for SVM. SVM using the best parameter will give the best model for this dataset. We use the model for test set. The accuracy for this model on test set is 86%. This shows that SVM is a good model for heart disease prediction. We plot the confusion matrix for this model.

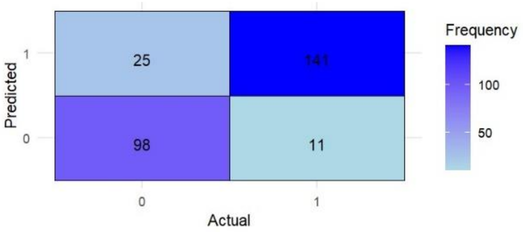
1.4.2.3. K-Nearest Neighbours (KNN)

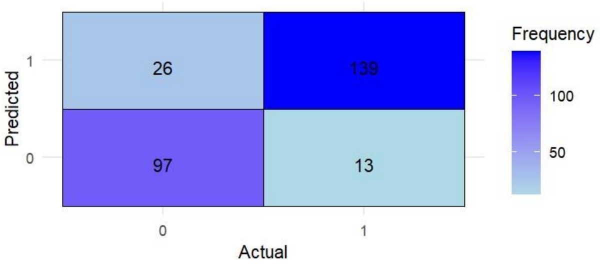
KNN is a non-parametric machine learning technique. It finds the Euclidean distance in the feature space to find K- Nearest Neighbour. KNN does not have training phase as it stores all the training. The downside to this is that prediction can be slow especially for large dataset. We changed the heart disease column values to present and absent. We split the data into train and test in the ratio of 70 to 30. We use the train set to find the best k value. We use the 10-fold cross validation to control for this model. From the figure we can see that the best k value is 8 which has close to 68% accuracy. Next, we use this model with k = 8 for the test set. we plot the confusion matrix for comparison to decide which is a better model to use for prediction.

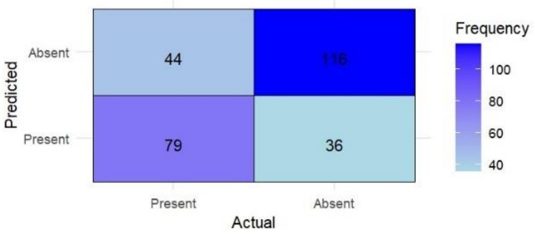


1.4.2. Results

The results of confusion matrix are then used to calculate a few values that is needed to choose the best model. Precision is defined as the proportion of true positive out of all positive. Recall is the proportion of true positive out of the actual positive values, which is the true positive rate. Next value we calculate is the f value. f = 2 * ((precision * recall)/(sum of precision and recall)). Specificity is proportion of actual negative out all negative values. I have changed these values into percentages for easier comparisons. The figure on the left is the confusion matrix for logistic regression. As the from figure, we can observe that there are values that is being predicted correctly compared to falsely predicted. For logistic regression precision = 85%, recall = 93%, f = 89%, specificity = 90%. The next figure that is shown on the right is the confusion matrix of SVM. The precision = 84%, recall = 91%, f = 88%, specificity = 88%. The next figure is shown below is the confusion matrix for KNN model.







We will conclude that KNN is not a suitable model as it has the lowest accuracy amongst the three with 68%. SVM and logistic has accuracy of 86% and 87% respectively. In this context it is important for there to be high recall and specificity. We need high recall, true positive rate, as we need to predict that this person has heart disease or high risk of getting heart disease so that proper treatment can be given. For specificity is important so that we do not end up giving treatment to someone who does not need it due to wrong prediction. Therefore, logistic regression is the best as its recall and specificity is the highest with 93% and 90% respectively. From logistic regression we can also identify that sex is an important factor for prediction of heart diseases.

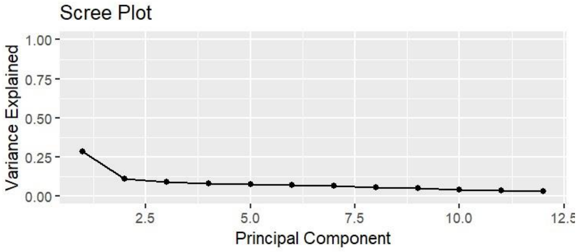
1.5. Unsupervised Learning

1.5.1 Methodology

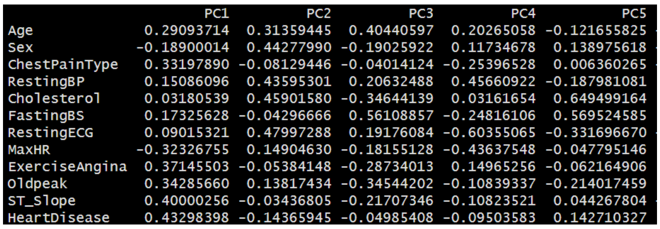
For unsupervised learning we will be using Principal Component Analysis, K-Means Clustering and hierarchical clustering. We used the data that has been wrangled previously for unsupervised learning. Unsupervised learning will focus more on finding the significant factors for prediction of heart disease. We do not use train test set as we are not finding the best model like the classification task.

1.5.2 Analysis

1.5.2.1 Principal Component Analysis (PCA)



PCA is a technique used to reduce the dimensionality. PCA still retains important information even it is being reduced. It identifies the direction the values vary the most and project the data onto the components. We choose the optimum number of components to observe. The optimum number of components to look at would be five as more than 60% of variance is being explained. By looking at the first five components, sex seems to have a positive relationship. Other than PC2, the other four PC values are closer to zero, it does not show significance. Fasting Blood Sugar seems to be more significant.



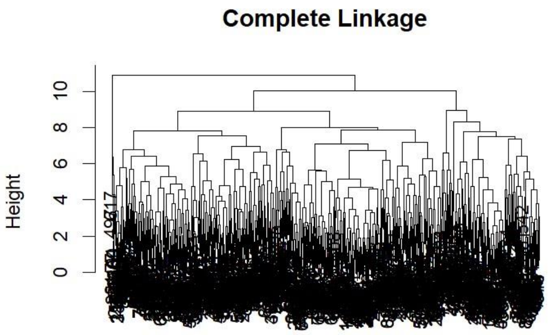
```
                      PC1         PC2         PC3         PC4          PC5
Age            0.29093714  0.31359445  0.40440597  0.20265058 -0.121655825
Sex           -0.18900014  0.44277990 -0.19025922  0.11734678  0.138975618
ChestPainType  0.33197890 -0.08129446 -0.04014124 -0.25396528  0.006360265
RestingBP      0.15086096  0.43595301  0.20632488  0.45660922 -0.187981081
Cholesterol    0.03180539  0.45901580 -0.34644139  0.03161654  0.649499164
FastingBS      0.17325628 -0.04296666  0.56108857 -0.24816106  0.569524585
RestingECG     0.09015321  0.47997288  0.19176084 -0.60355065 -0.331696670
MaxHR         -0.32326755  0.14904630 -0.18155128 -0.43637548 -0.047795146
ExerciseAngina 0.37145503 -0.05384148 -0.28734013  0.14965256 -0.062164906
Oldpeak        0.34285660  0.13817434 -0.34544202 -0.10839337 -0.214017459
ST_Slope       0.40000256 -0.03436805 -0.21707346 -0.10823521  0.044267804
HeartDisease   0.43298398 -0.14365945 -0.04985408 -0.09503583  0.142710327
```

1.2.2.3 K-Means Clustering

K-means clustering is used to partition the dataset into predetermined number of clusters. The clusters are split according to similar data points. It also helps to see if there is any inherent pattern. We need to scale the data before we use it for k-means clustering. We split them into two clusters. Cluster 1 has more data with heart disease than cluster 2. 90% of data in cluster 1 is male while 65% of cluster 2 is male. We performed goodness fit test using chi-square test. It showed that the two cluster are significantly different from each other. We conclude that sex is a significant factor in predicting heart disease.

1.2.2.3 Hierarchical Clustering

Hierarchical clustering is similar to k- mean with the exception that u do not need to specify the number of clusters. First, we used average linkage to see if the data well fitted. We split them into 3 cluster and the second cluster only had 8 cases. Thus, we used complete linkage which 210 cases in the second cluster. We then plot the dendrogram which is shown on the right. 78% in cluster 1 is male,81% in cluster 2 is male and 86% of cluster 3 is male. Cluster 2 has more people with heart disease followed by cluster 3 then cluster 1. We perform the goodness fit test. Cluster 1 and 2 is only two that has statistically similarity while 1 with and 2 with 3 are not statistically similar. We conclude that sex is a significant factor.



1.5.3 Results

PCA is the only technique that does not show sex has significance to predicting heart disease. k-means and hierarchical clustering both show that sex is a significant factor. Logistic regression also shows that sex is a significant factor. This result is supported by Dr Manav from New Victori Hospital (Dr Manav Sohal, 2022). Weidner G also supports this by saying that men are prone to heart disease in 2000 research (Weidner G,2000). Even though it is 24 years ago, it is still relevant today as said Dr Manav. Thus, we can conclude that sex of person is important in predicting heart disease.

2. Regression Task

2.1. Substantive Issue

The current market for housing has increased globally. houses are now more expensive than few years ago. Due to times change the features of the house affecting the price of the house has changed over time. The change in people's preferences has changed over time. Thus, I want to find if the area in square feet of the property is the most significant factor in the price. It is well known that area is one deciding factor of price, but I want to know if it is the most significant factor. Another issue is we want to predict price of the house so that people can save up for their dream house. We want to find the best prediction model to use. We want to know another factor that are significant to affect the price.

2.2. Dataset & Variables

1) Price : Price of property (target variable)
2) Area : total area of the property in square feet
3) Bedrooms : the number of bedrooms
4) Bathrooms : the number of bathrooms
5) Stories : number of stories (floors) in the property
6) mainroad : property located on a main road (No = 0, Yes = 1)
7) Guestroom : property has guestroom (No = 0, Yes = 1)
8) Basement : property has basement (No = 0, Yes = 1)
9) hotWaterheating : property has hot water heating (No = 0, Yes = 1)
10) airconditioning : property has air conditioning (No = 0, Yes = 1)
11) Parking : Number of parking spaces available
12) Prearea : whether property is in a preferred area (No = 0, Yes = 1)
13) Furnishing status : furnishing status of the property (Furnished = 0, Semi-furnished = 1, Unfurnished = 2)

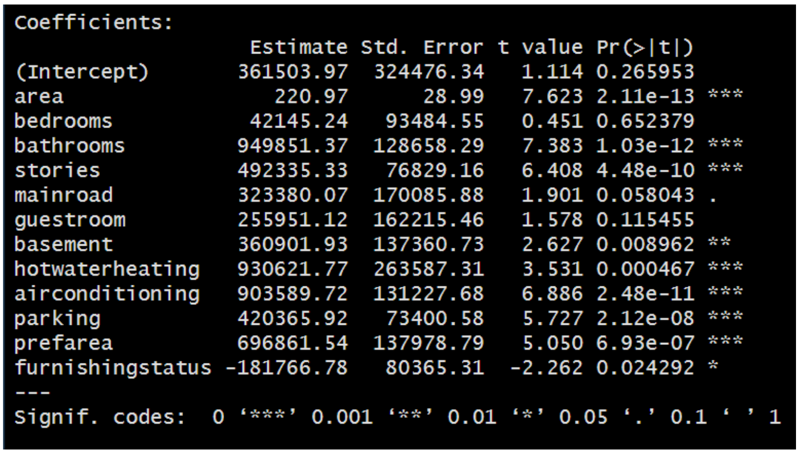
2.3. Methodology

I changed the columns with values as words into binary numbers as shown above in the dataset. For regression task I will be using Linear Regression, Random Forest, and CART. Since we are finding the best model to use for predicting the price of the property, we will split the data in train and set in the ration 70:30.

2.4. Analysis

2.4.1. Linear Regression

As mentioned in logistic regression, linear regression is similar, but we use continuous variable for y instead of categorical. In this scenario, y variable is the price of the property. Linear regression is used to figure out if the x variable, the other feature other than price, is significant variable to find the price of property. We use train set to train our model.

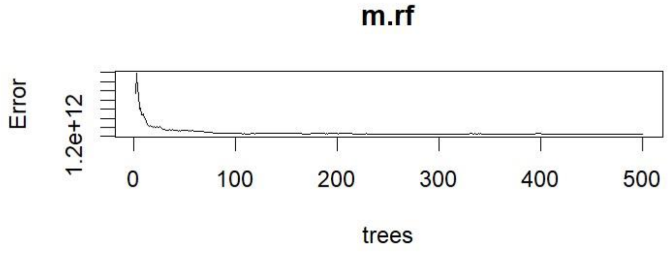


```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       361503.97  324476.34   1.114 0.265953
area                 220.97      28.99   7.623 2.11e-13 ***
bedrooms           42145.24   93484.55   0.451 0.652379
bathrooms         949851.37  128658.29   7.383 1.03e-12 ***
stories           492335.33   76829.16   6.408 4.48e-10 ***
mainroad          323380.07  170085.88   1.901 0.058043 .
guestroom         255951.12  162215.46   1.578 0.115455
basement          360901.93  137360.73   2.627 0.008962 **
hotwaterheating   930621.77  263587.31   3.531 0.000467 ***
airconditioning   903589.72  131227.68   6.886 2.48e-11 ***
parking           420365.92   73400.58   5.727 2.12e-08 ***
prefarea          696861.54  137978.79   5.050 6.93e-07 ***
furnishingstatus -181766.78   80365.31  -2.262 0.024292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The figure on the right the results we get from the linear regression we performed. The number of stars represent the significance of the variable. 3 stars represent the most significance. As the stars decrease the significant decreases. No stars mean it is not significant. From the results we can se that area is a significant factor affecting the price. Bathrooms, stories, hot water heating, air conditioning, parking area and preferred area are also the most significant factor affecting the price. The r squared is 0.68. since it close to 1 it is highly correlated. It also shows that 68% of the dependent variables is explained by this model. Next, we find the root mean squared error (RMSE) for both train and test set.

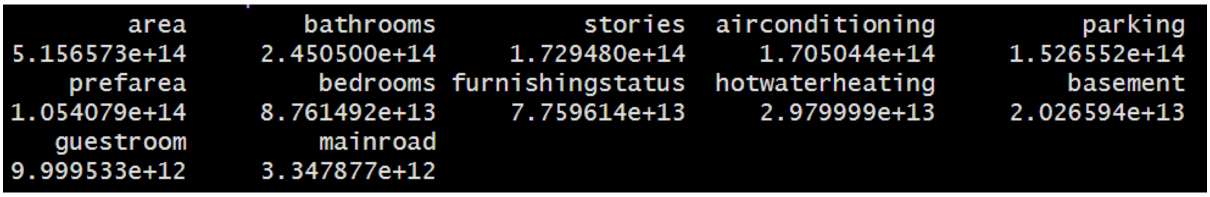
2.4.2. Random Forest (RF)

RF constructs multitude of decision trees when using the training set and give the mean prediction for test set for individual tree. We use train set to train our model. the model explains about 64% of variance of the data. the figure below shows that the error has stabilised before the 500 trees. It has stabilised around 100 trees. We then find RMSE for both train and test set to find the best model.

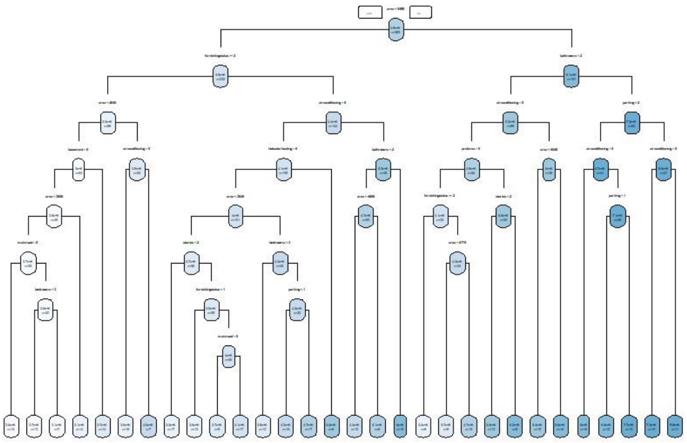


2.4.3. Classification and Regression Tree (CART)

CART is a type od decision tree. CART creates a binary tree by splitting the data into subsets. CART finds the values of input features to maximize the purity. We nee to prune the tree to get the optimum tree for this model. we start with not pruning the tree. We use train set to train the model.

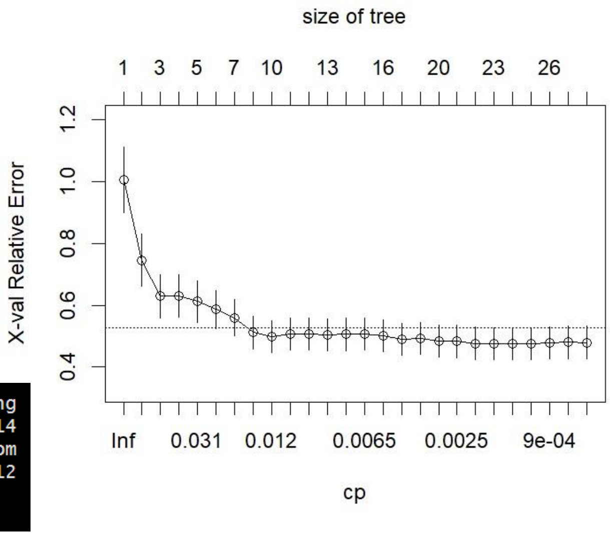


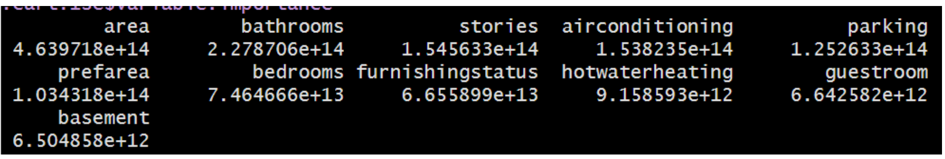
```
         area       bathrooms         stories  airconditioning         parking
5.156573e+14    2.450500e+14    1.729480e+14     1.705044e+14    1.526552e+14
     prefarea        bedrooms furnishingstatus  hotwaterheating        basement
1.054079e+14    8.761492e+13    7.759614e+13     2.979999e+13    2.026594e+13
    guestroom        mainroad
9.999533e+12    3.347877e+12
```



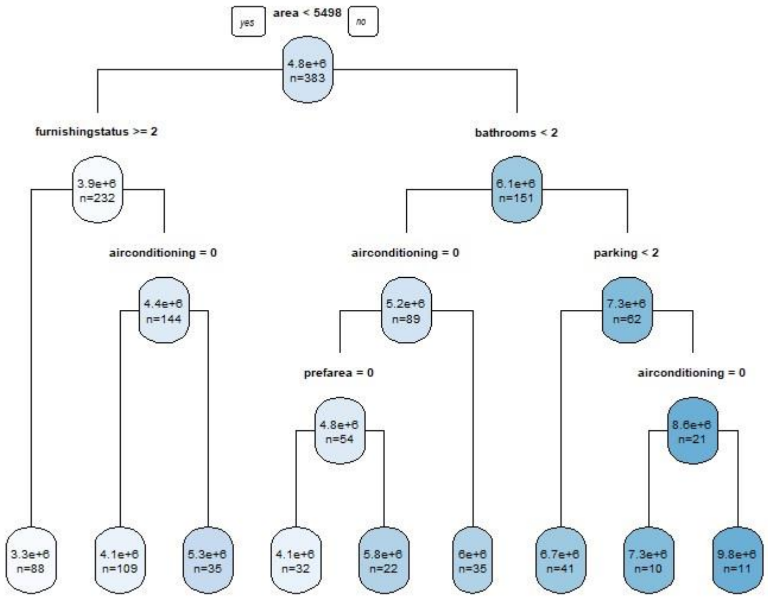
From the figure above we can see that area, bathrooms, stories air-conditioning, parking, preferred area are significant features in predicting price of the property. From the figure on the left shows the tree that has not be pruned. The tree does not show the most important features. Now we need to find the right split that minimise CV error. The minimum CV error is 0.52555. now we find the optimal CP region. We can use through codes or plot a graph and by observation find the right split.

Through the figure we observe that the point where it first is under the line is at tree size 8. Through the code we also find that 8 is the optimal size of tree that minimise CV error. After pruuning the tree we can see that area is still the most significant factor that affects the price. Now we plot the pruned tree nd find the RMSE of using this model for both the train and test set.





```
        area       bathrooms         stories  airconditioning         parking
4.639718e+14    2.278706e+14    1.545633e+14     1.538235e+14    1.252633e+14
    prefarea        bedrooms furnishingstatus  hotwaterheating       guestroom
1.034318e+14    7.464666e+13    6.655899e+13     9.158593e+12    6.642582e+12
    basement
6.504858e+12
```

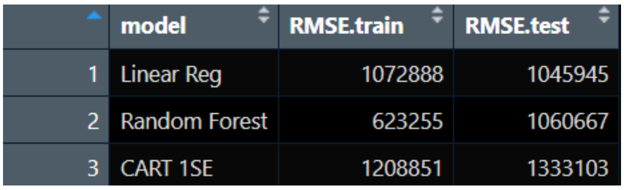


The figure on the left is the tree after pruning and it features the most important factors that affect the price of the property.

2.5. Results

Based on the above analysis I can conclude that area of the property affects the price of the property. The linear regression and CART show that area of the property is the most significant variable that affects the price of the property. According to Gomez, area of the property is one of the eight critical factor that affects the price of the property (Gomez, 2022). What is more surprising factor the number of bedrooms. We would expect number of bedrooms to affect the price of the property, but analysis shows otherwise. Even Padala has proven that number of bedrooms affect the price (Padala, 2021). Through this I can say that the not all housing types has bedrooms affecting the prices. Looking the typical price range in this dataset it could be that they considered to be big houses that bedrooms do not affect its prices.

Now we look at which is the best model to use for predicting the housing prices. We look at the RMSE values to see which model gives the least error to use for prediction.



	model	RMSE.train	RMSE.test
1	Linear Reg	1072888	1045945
2	Random Forest	623255	1060667
3	CART 1SE	1208851	1333103

From the table above we can see that Random Forest gives the lowest train and test set error. Thus, we will use Random Forest to predict prices given the above features. We use RMSE as it is sensitive to prediction errors. Large prediction errors get penalized more heavily compared to small errors due to the squared term. This allows us to minimise the large errors. Thus, we choose the lowest RMSE model to predict the price of the property.

3. References

(Dr Manav Sohal, 2022) *Cardiovascular risk in men - why is heart disease a male problem* (2022) *New Victoria Hospital*. Available at: https://www.newvictoria.co.uk/about-us/news-and-articles/cardiovascular-risk-in-men-why-is-heart-disease-a-male-problem

(Weidner G, 2000) G;, W. (2000) *Why do men get more heart disease than women? an international perspective*, *Journal of American college health : J of ACH*. Available at: https://pubmed.ncbi.nlm.nih.gov/10863872/

(Gomez, 2022) Gomez, J. (2022) *Sell your home the Minute you're ready.*, *Opendoor*. Available at: https://www.opendoor.com/articles/factors-that-influence-home-value

(Padala, 2021) Padala, A. (2021) *Determinants of housing price*, *Medium*. Available at: https://apadala-90574.medium.com/determinants-of-housing-price-bdefc783cf6b

4. Dataset Links

4.1. Unsupervised Learning and Classification

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

4.2. Regression

https://www.kaggle.com/datasets/saurabhbadole/housing-price-data/data