Coding Assignment 1

Kaviyaa Vasudevan
Engineering Science (Data Science) MS
University at Buffalo, Buffalo, NY 14260
kaviyaav@buffalo.edu

1 Linear Regression

1.1 Definition

Simple linear regression model is a method to predict dependent variable Y based on the values of the independent variables X. The two variables are linearly related to each other hence, we try to find a linear function that predicts the values of Y as accurately as possible as a function of the feature or the independent variable.

1.2 Linear Probabilistic Model

The simplest deterministic mathematical relationship between two variables x and y is a linear relationship: $y = \beta 0 + \beta 1x$. So, we assume $Y = \beta 0 + \beta 1x + \varepsilon$, where ε is a random variable. Two variables are related linearly on average if for fixed x the actual value of Y differs from its expected value by a random amount (i.e. there is random error). There are parameters $\beta 0$, $\beta 1$, and $\sigma 2$, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the model equation

$$Y = \beta 0 + \beta 1x + \varepsilon$$

The quantity ε in the model equation is the "error" -- a random variable, assumed to be symmetrically distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma \varepsilon 2 = \sigma 2$
X: the independent, predictor, or explanatory variable (usually known).
Y: The dependent or response variable. For fixed x, Y will be random variable.
ε: The random deviation or random error term. For fixed x, ε will be random variable

1.2 Data set

Mobile Price Prediction is our sample dataset to perform linear regression. Mobile price depends on various factors such as resolution refers to the number of pixels in the display, size is the exact screen size of the mobile phone, weight, Pixels per inch (PPI) is the measure of resolution in a digital image or video display, RAM stores the information for your mobile in a readily accessible way so that information can be easily fetched by the CPU in very little time in GB, battery is the capacity of the battery used in the mobile phone in form of mAh(milliampere/hour) and CPU frequency is that the phone's processor can clock up to speed in GHz, internal memory the manufacturer-installed storage space in GB, CPU core is an element of the processor that implements and executes tasks, front camera refers to the front facing camera in megapixel, thickness is the measure of how thick the mobile is in inches, rear camera refers to the front facing camera in megapixel. The variables in the dataset are continuous as we are tending to perform regression. In this dataset, we want to estimate the price of mobile phones using the above features.

2 Data Preprocessing

The data preprocessing can often have a significant impact on generalization performance of a ML algorithm. The foremost process of data preprocessing is to import the data set by which we develop models. This will be performed using the pandas library. Prior to importing the dataset we are required to import the necessary libraries. The elimination of noise instances is one of the most

difficult problems in inductive ML. Usually the removed instances have excessively deviating instances that have too many null feature values which are also referred to as outliers. Missing data handling is another issue often dealt with in the data preparation steps. Moreover, in real-world data, the representation of data often uses too many features, but only a few of them may be related to the target concept. There may be redundancy, where certain features are correlated so that is not necessary to include all of them in modeling; and interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. Splitting the data set becomes one of the significant processes of data preprocessing. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem

2.1 Fitting the Model and Predicting Results

The linear regression class from sklearn library is used to fit the data set into the model. The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

2.2 Visualization

The final step is to visualize the results for which matplotlib library can be used to make scatter plots of our training set results and test set results to see how close our model has predicted the values.

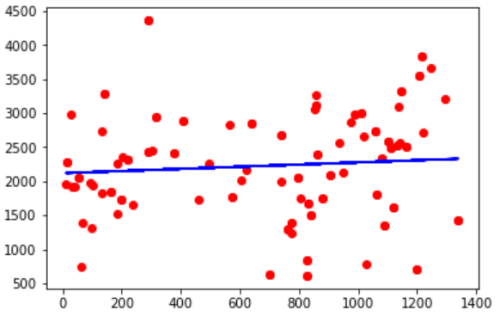


Fig 1.1 Training dataset scatter plot

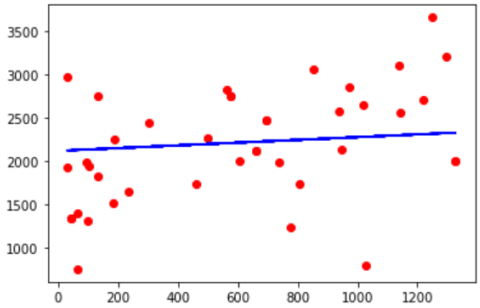


Fig 1.2 Test set scatter plot

The figure 1.2 shows useful feature of scatter plots is that they are easily completed by simple linear regression in the placement of a regression line through the data. This can be done by inserting a trendline. This line is a model that comes from the simple linear regression of the data. The basis of this line is the equation we all learned in high school, which is y=mx+b. The calculation of the line is a little more complex. The values for m and b (slope and intercept, respectively) are calculated from the data using the simple linear regression. The distance from each point to the line is the error. The linear regression line returns the values of m (slope) and b (intercept) that reduce the sum of the errors squared. Another name for simple linear regression is "least squares regression", a name which describes the result of the tool.

Coding Assignment 1

Kaviyaa Vasudevan
Engineering Science (Data Science) MS
University at Buffalo, Buffalo, NY 14260
kaviyaav@buffalo.edu

1 Multiple Linear Regression

1.1 Definition

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. Multiple linear regression makes all of the same assumptions as simple linear regression. It's a statistical variable which uses multiple explanatory variables to predict the response of the outcome variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

1.2 Multiple Regression Analysis

The multiple regression model equation is $Y = b0 + b1x1 + b2x2 + ... + bpxp + \varepsilon$ where $E(\varepsilon) = 0$ and $Var(\varepsilon) = s2$. Again, it is assumed that ε is normally distributed. This is not a regression line any longer, but a regression surface and we relate y to more than one predictor variable x1, x2, …xn. The regression coefficient b1 is interpreted as the expected change in Y associated with a 1-unit increase in x1 while x2,.. xn are held fixed. Analogous interpretations hold for b2,..., bp. Thus, these coefficients are called partial or adjusted regression coefficients. In contrast, the simple regression slope is called the marginal coefficient. Model parameters in a multiple regression model are usually estimated using ordinary least squares minimizing the sum of squared deviations between each observed value and predicted values. It involves solving a set of simultaneous normal equations, one for each parameter in the model. A maximum likelihood estimation of parameters will give the same result if errors are normal.

1.3 Models with categorical predictor

Sometimes, a three-category variable can be included in a model as one covariate, where the best way is to incorporating three unordered categories is to define two different indicator variables.

1.4 Data set

Mobile Price Prediction is our sample dataset to perform multiple linear regression. Mobile price depends on various factors such as resolution refers to the number of pixels in the display, size is the exact screen size of the mobile phone, weight, Pixels per inch (PPI) is the measure of resolution in a digital image or video display, RAM stores the information for your mobile in a readily accessible way so that information can be easily fetched by the CPU in very little time in GB, battery is the capacity of the battery used in the mobile phone in form of mAh(milliampere/hour) and CPU frequency is that the phone's processor can clock up to speed in GHz, internal memory the manufacturer-installed storage space in GB, CPU core is an element of the processor that implements and executes tasks, front camera refers to the front facing camera in megapixel, thickness is the measure of how thick the mobile is in inches, rear camera refers to the front facing camera in megapixel. The variables in the dataset are continuous as we are tending to perform regression. In this dataset, we want to estimate the price of mobile phones using the above features.

2 Data Preprocessing

The data preprocessing can often have a significant impact on generalization performance of a ML algorithm. The foremost process of data preprocessing is to import the data set by which we develop models. This will be performed using the pandas library. Prior to importing the dataset we are required to import the necessary libraries. The elimination of noise instances is one of the most difficult problems in inductive ML. Usually the removed instances have excessively deviating instances that have too many null feature values which are also referred to as outliers. Missing data handling is another issue often dealt with in the data preparation steps. Moreover, in real-world data, the representation of data often uses too many features, but only a few of them may be related to the target concept. There may be redundancy, where certain features are correlated so that is not necessary to include all of them in modeling; and interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. The Dummy Variable trap is a scenario in which two or more variables are highly correlated; in simple terms, one variable can be predicted from the others. Intuitively, there is a duplicate category: if we dropped the male category, it is inherently defined in the female category (zero female value indicate male, and vice-versa). The solution to the dummy variable trap is to drop one of the categorical variables - if there are m number of categories, use m-1 in the model, the value left out can be thought of as the reference. Splitting the data set becomes one of the significant processes of data preprocessing. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem

2.1 Fitting the Model and Predicting Results

The linear regression class from sklearn library is used to fit the data set into the model. The regressor object of linear regression is made and we try to fit the regressor object into our dataset using the fit() method. After that, we will make predictions based on the fitted model. Here, we just have called the predict() method from linear model on dependent variables from partial data frame, and then the system will predict the price values. Then, the accuracy of the linear regression model is calculated with the help of R squared value. The accuracy of the model turned out to be 83.70 %.

2.2 Visualization

The final step is to visualize the results for which matplotlib library can be used to make scatter plots of our training set results and test set results to see how close our model has predicted the values.

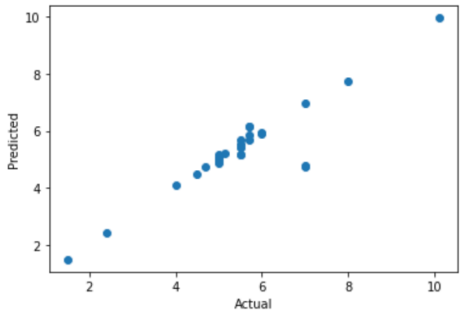


Fig 1.1 Scatter plot

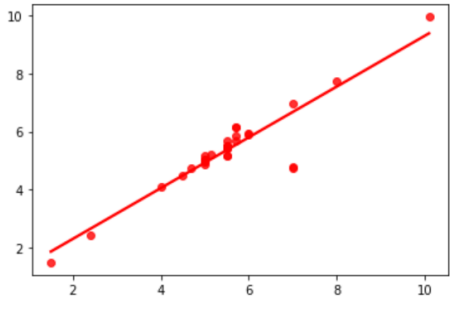


Fig 1.2 SNS scatter plot

The figure 1.2 shows a very strong tendency for X and Y to both rise above their means or fall below their means at the same time. The straight line is a trend line, designed to come as close as possible to all the data points. The trend line has a positive slope, which shows a positive relationship between X and Y. The points in the graph are tightly clustered about the trend line due to the strength of the relationship between X and Y. The figure 1.1 shows that there are some outliers in the dataset. With the above positive results of visualization, it is known that the linear regression model has predicted the prices of mobile phones more accurately. This is also proven by the strong accuracy shown by the model. The price of the cellphone are accurately predicted with the help of all dependent variables in the dataset. The sample of the predicted values looks like the figure 1.3.

	Actual Value	Predicted Value	Difference
0	5.50	5.664491	-0.164491
1	2.40	2.436196	-0.036196
2	6.00	5.903996	0.096004
3	5.50	5.439697	0.060303
4	5.70	5.857168	-0.157168
5	5.00	5.015777	-0.015777
6	5.70	6.153237	-0.453237
7	5.50	5.174203	0.325797

Fig 1.3 Sample of the predicted values

Coding Assignment 1

Kaviyaa Vasudevan
Engineering Science (Data Science) MS
University at Buffalo, Buffalo, NY 14260
kaviyaav@buffalo.edu

1 Logistic Regression

1.1 Definition

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

1.2 Logistic Regression Analysis

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. The response variable Y is a binomial random variable with a single trial and success probability Thus, Y = 1 corresponds to "success" and occurs with probability, and Y = 0 corresponds to "failure" and occurs with probability 1 - π. The set of predictor or explanatory variables = (1, 2,...,) are fixed (not random) and can be discrete, continuous, or a combination of both. As with classical regression, two or more of these may be indicator variables to model the nominal categories of a single predictor, and others may represent interactions between two or more explanatory variables. Together, the data is collected for the ith individual in the vector (z,,, Y), for i = 1,...n. These are assumed independent by the sampling mechanism. This also allows us to combine or group the data, which we do below, by summing over trials for which is constant. In this section of the notes, we focus on a single explanatory variable.

The model is expressed as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

Or, by solving for π;, we have the equivalent expression

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

To estimate the parameters, we substitute this expression for π, into the joint pdf for Y_1,..., Yn

$$\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

to give us the likelihood function L(Bo, B1) of the regression parameters. By maximizing this likelihood over all possible Bo and B1.Extending this to include additional explanatory variables is straightforward.

1.3 Interpreting Logistic Regression

Consider first the simple linear regression where Y is continuous and X is binary. When X = 0, $E(Y|X=0) = \beta_0$ and when X = 1, $E(Y|X=1) = \beta_0 + \beta_1$. And so when interpreting the meaning of β_1, we say it represents the mean difference between the two groups i.e. the mean difference from when X = 0 (reference group) and when X = 1 (comparison group).

Now, let us assume the simple case where Y and X are binary variables taking values 0 or 1.When it comes to logistic regression, the interpretation of $\beta_1$differs as we are no longer looking at means. Recall that logistic regression has model $\log(E(Y|X)/(1-E(Y|X)) = \beta_0 + \beta_1 X$ or for simplification's sake, $\log(\pi/(1-\pi)) = \beta_0 + \beta_1 X$. This is all based on an odds ratio. When looking at what we would get for all possible values of X,

X = 0	$\log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta_0$
X = 1	$\log\left(\frac{\pi_1}{1-\pi_1}\right) = \beta_0 + \beta_1$

If we wish to interpret β_1 from these two above cases, we will analyze it similarly as if it were a simple linear regression. That is, β_1 results from subtracting the result from when X = 1 to that of when X = 0:

$$\beta_1 = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right)$$

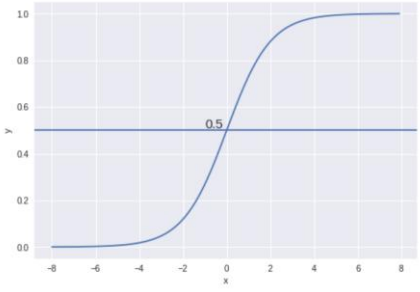
This shows that β_1 is a log odds ratio, and that $\exp(\beta_1)$ is an odds ratio.

1.4 Sigmoid Function

It is a mathematical function having a characteristic that can take any real value and map it to between 0 to 1 shaped like the letter "S". The sigmoid function also called a logistic function.

Y = 1 / 1+e -z

So, if the value of z goes to positive infinity then the predicted value of y will become 1 and if it goes to negative infinity then the predicted value of y will become 0. And if the outcome of the sigmoid function is more than 0.5 then we classify that label as class 1 or positive class and if it is less than 0.5 then we can classify it to negative class or label as class 0.



sigmoid function

Sigmoid Function acts as an activation function in machine learning which is used to add non-linearity in a machine learning model, in simple words it decides which value to pass as output and what not to pass

1.5 Models with categorical predictor

The two commonly used techniques to deal with categorical variables:

One-Hot Encoding - Convert a variable with N classes into N separate variables with binary labels. Repeat for each of the 3 variables.
Label Encoding - Map categorical variables into integers.

Label Encoding works only if there's some inherent order in the variables. Variables like these are called ordinal. Example would be a variable like days of the week. Monday can be one 1, Tuesday can be 2 and so on. Here the classifier would assume that 2 is greater than 1 in some way, which is fine as there is some order in the variable. One hot encoding is for the case where the variables are not ordinal - like names of places. This also gets used a lot in natural language processing.

1.6 Data set

This dataset consists of features that can be used to predict which patients have a high risk of heart disease. The prediction depends on various factors such as age refers to the age of the person, gender is the type of gender of the person, hypertension is in binary values which describes if the person has hypertension or not, ever_married is a Boolean value of the person's marital status, work_type refers to the type of work the person is employed in, residence_type has the value of the area where the person resides, avg_glucose_level is the glucose level of the particular person, BMI(Body Mass Index) , stroke describes if the person has suffered from stroke or not, and smoking status implies whether the person smokes or not. The variables in the dataset are both continuous and categorical as we are tending to perform logistic regression we have to change all the values to continuous variables. In this dataset, we want to predict if the person has any high chances of suffering from heart attack using the above mentioned variables.

2 Data Preprocessing

The data preprocessing can often have a significant impact on generalization performance of a ML algorithm. The foremost process of data preprocessing is to import the data set by which we develop models. This will be performed using the pandas library. Prior to importing the dataset we are required to import the necessary libraries. The elimination of noise instances is one of the most difficult problems in inductive ML. Usually the removed instances have excessively deviating instances that have too many null feature values which are also referred to as outliers. Missing data handling is another issue often dealt with in the data preparation steps. Moreover, in real-world data, the representation of data often uses too many features, but only a few of them may be related to the target concept. There may be redundancy, where certain features are correlated so that is not necessary to include all of them in modeling; and interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. Splitting the data set becomes one of the significant processes of data preprocessing. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem

2.1 Fitting the Model and Predicting Results

There were 5 predictors which are significantly associated to the outcome. These include: age, hypertension, smoking_status, stroke and avg_glucose_level. The coefficient estimate of the variable glucose is b = 0.045, which is positive. This means that an increase in glucose is associated with increase in the probability of being diabetes-positive. However the coefficient for the variable residence_type is b = -0.007, which is negative. An odds ratio measures the association between a predictor variable (x) and the outcome variable (y). It represents the ratio of the odds that an event will occur (event = 1) given the presence of the predictor x (x = 1), compared to the odds of the event occurring in the absence of that predictor (x = 0).For a given predictor (say x1), the associated beta coefficient (b1) in the logistic regression function corresponds to the log of the odds ratio for that predictor. If the odds ratio is 2, then the odds that the event occurs (event = 1) are two times higher when the predictor x is present (x = 1) versus x is absent (x = 0). For instance, the regression coefficient for glucose is 0.042. This indicate that one unit increase in the glucose concentration will increase the odds of being diabetes-positive by exp(0.042) 1.04 times. From the logistic regression results, it can be noticed that some variables – residence_type, work_type, gender - are not statistically significant. Keeping them in the model may contribute to overfitting. Therefore, they should be eliminated.

2.2 Visualization

The final step is to visualize the results for which matplotlib library can be used to make scatter plots of our training set results and test set results to see how close our model has predicted the values.

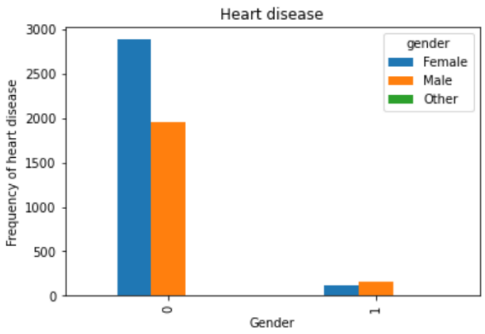


Fig 1.1 Bar plot

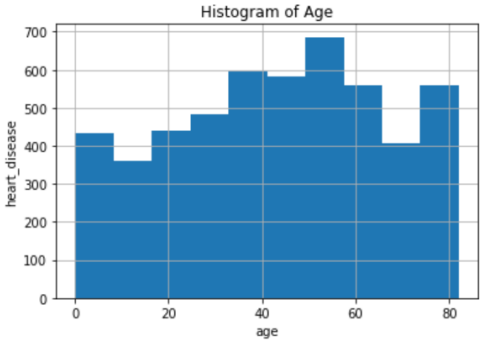


Fig 1.2 Histogram

The figure 1.2 shows people with different age groups those are highly prone to heart attacks. The figure 1.2 is a bar plot visualization between gender and heart disease.

2.3 ROC Curve

ROC curves in logistic regression are used for determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1). Mathematically these are represented as:

Sensitivity = (number correctly identified 1s)/(total number observed 1s)

Specificity = (number correctly identified 0s)/(total number observed 0s)

Given this information, we can put everything together to understand ROC curves. First, we identify the axes of an ROC curve: the Y axis is just sensitivity (or true positive rate), while the X axis is 1-specificity.

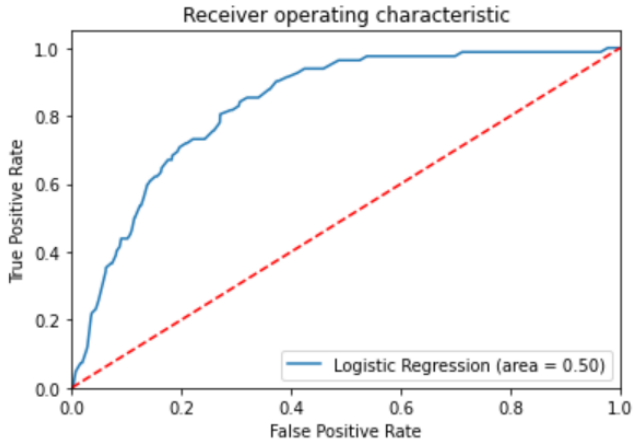


Fig 1.1 ROC Curve

Coding Assignment 1

Kaviyaa Vasudevan
Engineering Science (Data Science) MS
University at Buffalo, Buffalo, NY 14260
kaviyaav@buffalo.edu

1 Decision Tree

1.1 Definition

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. Decision trees look like flowcharts, starting at the root node with a specific question of data, that leads to branches that hold potential answers. The branches then lead to decision (internal) nodes, which ask more questions that lead to more outcomes. This goes on until the data reaches what's called a terminal (or "leaf") node and ends.

1.2 Decision Tree Analysis

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm**.** Begin the tree with the root node, says S, which contains the complete dataset. Find the best attribute in the dataset using Attribute Selection Measure (ASM). Divide the S into subsets that contains possible values for the best attributes. Generate the decision tree node, which contains the best attribute. Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

1. Information Gain:

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree.

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no

2. Gini Index:

Gini index is a measure of impurity or purity used while creating a decision tree in the

CART(Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

Gini Index= 1- ΣjPj2

1.3 Data set

Mobile Price Prediction is our sample dataset to perform linear regression. Mobile price depends on various factors such as resolution refers to the number of pixels in the display, size is the exact screen size of the mobile phone, weight, Pixels per inch (PPI) is the measure of resolution in a digital image or video display, RAM stores the information for your mobile in a readily accessible way so that information can be easily fetched by the CPU in very little time in GB, battery is the capacity of the battery used in the mobile phone in form of mAh(milliampere/hour) and CPU frequency is that the phone's processor can clock up to speed in GHz, internal memory the manufacturer-installed storage space in GB, CPU core is an element of the processor that implements and executes tasks, front camera refers to the front facing camera in megapixel, thickness is the measure of how thick the mobile is in inches, rear camera refers to the front facing camera in megapixel. The variables in the dataset are continuous as we are tending to perform regression. In this dataset, we want to estimate the price of mobile phones using the above features.

2 Data Preprocessing

The data preprocessing can often have a significant impact on generalization performance of a ML algorithm. The foremost process of data preprocessing is to import the data set by which we develop models. This will be performed using the pandas library. Prior to importing the dataset we are required to import the necessary libraries. The elimination of noise instances is one of the most difficult problems in inductive ML. Usually the removed instances have excessively deviating instances that have too many null feature values which are also referred to as outliers. Missing data handling is another issue often dealt with in the data preparation steps. Moreover, in real-world data, the representation of data often uses too many features, but only a few of them may be related to the target concept. There may be redundancy, where certain features are correlated so that is not necessary to include all of them in modeling; and interdependence, where two or more features between them convey important information that is obscure if any of them is included on its own. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. Splitting the data set becomes one of the significant processes of data preprocessing. The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem

2.1 Fitting the Model and Predicting Results

When it comes to actually building a decision tree, we start at the root, which includes the total population of atoms. As we move down the tree, the goal is to split the dataset into smaller and smaller subsets of atoms at each node; hence the popular description, “divide and conquer.” Each subset should be as distinct as possible in terms of the target indicator that is heart disease. For example, if you are looking at high- vs. low-risk person, you would want to split each node into two subsets; one with mostly high-risk people, and the other with mostly low-risk people. This goal is achieved by iterating through each indicator as it relates to the target indicator, and then choosing the indicator that best splits the data into two smaller nodes. As the computer iterates through each indicator-target pair, it calculates the Gini Coefficient, which is a mathematical calculation that is used to determine the best indicator to use for that particular split. The Gini Coefficient is a score between 0 and 1, with 1 being the best split, and 0 being the worst. The computer chooses the indicator that has the highest Gini Coefficient to split the node, and then moves on to the next node

and repeats the process.

2.2 Visualization

The final step is to visualize the results for which matplotlib library can be used to make scatter plots of our training set results and test set results to see how close our model has predicted the values.

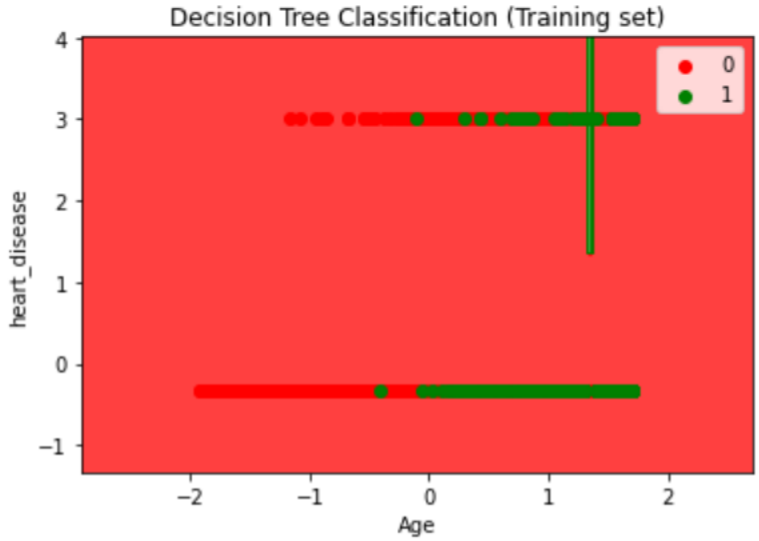


Fig 1.1 Listed color map for training set

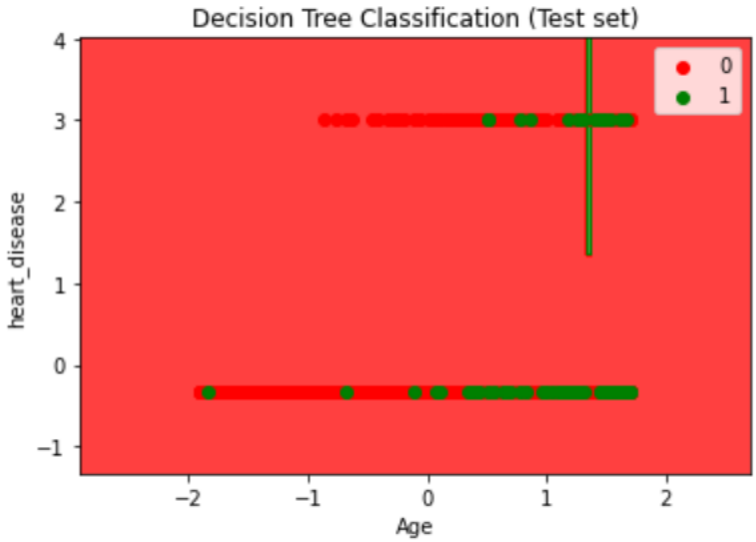


Fig 1.2 Listed color map for test set

The figure 1.2 shows testing result set of people with different age groups those are highly prone to heart attacks. The figure 1.1 shows training result set of people with different age groups those are highly prone to heart attacks.