EVALYNE LWOBA

ANDREW_ID : elwoba

DIAML

ASSIGNMENT 6

Libraries imported.

```
import math
import matplotlib.pyplot as plt

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statistics


#Algorithms
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_validate
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.linear_model import LinearRegression

from sklearn import linear_model
from sklearn import tree
from sklearn import svm

import warnings
warnings.filterwarnings("ignore")
import warnings;warnings.simplefilter("ignore")
```

1. Nonlinearity (25 points)

1.1 Why might it be necessary to consider nonlinear relationships between variables?

Non-linearity is a state where statistically, there isn't a direct relationship between two independent features of a model. A change in the output doesn't affect or alter the changes in the input. It is necessary to consider nonlinearity between variables because it can affect the performance of a model. Non linearity offers a broader range of dependencies thus producing good estimates. It also offers accommodation for a huge range of data.

1.2 Write down the mathematical equation for a nonlinear model and provide an example of an application where it might be appropriate. (5 marks)

Threshold autoregression (TAR)

Formula:

$$y_t = \phi_0^{(i)} + \sum_{j=1}^{p_i} \phi_j^{(i)} y_{t-j} + \varepsilon_t^{(i)} \text{ if } y_{t-d} \in R_i, i = 1,2,\ldots,k$$

Application:
This model can be used where one needs to analyze nonlinear behavior of stocks.

1.3 Can a nonlinear model be more parsimonious than a linear model? Write down mathematical formulae for both the linear and nonlinear models to support your answer. [1]

A Non linear model is more parsimonious than a linear model because a non linear model can take fewer parameters than linear model. There are cases in statistical modelling where a non-linear model with only a small number of parameters fits the data well, and any attempt to apply a linear model neces sitates a lot of factors to get a proper fit. This can happen in a periodic regression model for example.:

$$Y_i = \beta_0 + \alpha \sin(2\pi\phi(x_i - \mu)) + \sigma\varepsilon_i \varepsilon_i \sim \text{IID } N(0,1).$$

Where the frequency factor $0<\phi<1$ joins the model in a non-linear way.there is no matching linear model that accurately achieves this same shape. It is feasible to build a linear regression model that estimates the sinusoidal signal. This also can be a good estimate over the data range if you have a lot of parameters. For example, you might decide to evaluate this signal as a sum of periodic signals with harmonic frequencies, yielding the linear regression model:

$$Y_i \approx \beta_0 + \sum_{i=1}^{m} [\beta_{1s} \sin(2\pi i m x_i) + \beta_{1c} \cos(2\pi i m x_i)] + \sigma\varepsilon_i \varepsilon_i \sim \text{IID } N(0,1).$$

In the above instance where we have a non-linear model with only 4 unknown coefficient parameters and a linear estimate with $1+2m$ unspecified coefficient parameters. If m is reasonably large then the linear approximation is less parsimonious in the sense that it has more parameters. This would be the kind of case where it is sensible to say that **a non-linear model is more parsimonious than its corresponding linear approximation**.

1.4 Surrogate data are used for testing for nonlinearity. What characteristics are typically preserved when generating surrogates?

Characteristics to consider when generating surrogates includes:

a) FFT surrogates preserve the linear correlations only
b) Polished surrogates preserve unconditional distribution and linear correlation
c) IID surrogates preserve the conditional distribution .
d) Surrogates should look like the original

Give the names of two surrogate techniques and describe the approaches for implementing them.(5 marks)

a) Surrogate data testing technique.
This is used to detect nonlinearity in a time series by proof of contradiction. it involves specifying a null hypothesis which illustrates a linear process then coming up with a number of surrogate data sets in line with the null hypothesis.

b) Jvndfkvjh

The approaches for implementing surrogate techniques are:

a) Parametric- this is where a model is fit to the original data and used to come up with surrogate data sets.
b) Non-parametric-this is where there is no model fit on the original data instead, surrogate data sets are created right from the original data.

1.5 Define information, entropy and mutual information using mathematical formulas. Explain how mutual information can be used for feature selection and why it might be better than correlation. (5 marks)

a) Information is given by

$$I(x) = -\log p_x(x)$$

b) Entropy

This is where a random variable is a function which tries to characterize the unpredictability or uncertainty in the value of a variable. Entropy is also not just about sum of likely results but also their rate of recurrence. For instance, z results of a weighted six sided die that happens 90% of the time as a 2 implies that z has a lesser entropy than y representative of a fair 6 sided die. The weighted die is less unpredictable in a way.[1]

Entropy is given by:

$$H = -\Sigma_k\, p(k) \ln p(k).$$

c) **Mutual information**.
This is a quantity that measures a relationship between two random variables sampled simultaneously. It calculates the amount of information conveyed in one random variable about the other in bits. Example, is the roll of a fair 6 sided die, Y is whether the roll is even or odd. The value of Y will infer something about the value of X and vice versa. These two variables share information about each other.
The formula for mutual information is:

$$I(p) = \log(1/p) = -\log(p).$$

Explain how mutual information can be used for feature selection and why it might be better than correlation.

Mutual information can be used for features selection because a larger mutual information implies an existence of a relationship between features it doesn't matter if the data is linear or non linear

Describe how entropy can be used for constructing a feature for measuring regularity and give an example of an application.

To construct a feature for measuring regularity, Approximate entropy (ApEn) technique is applied. ApEn is a modification that allows estimation of entropy using empirical observations and is independent of the volume of data. This is used to quantify the amount of regularity and the unpredictability of fluctuations over a time series data set. This formula necessitates availability of big volume of data that are sensitive to noise.

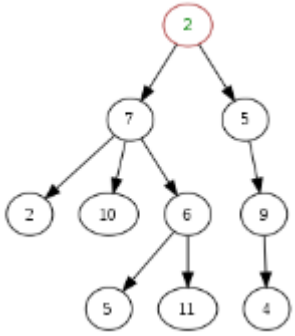
An example of application is measuring regularity of the heart rate where the heartrate of men and women were compared and a conclusion that men are less complex than women was made.

2. Classification using trees (25 points)
2.1. Describe the components (nodes, branches) of a decision tree.

A decision tree is hierarchical representation of abstract data. It is made of linked nodes that represent different aspects of a data set. There are three types of nodes in a decision tree. (a) Root node which is the decision node demonstrating a choice that will lead to a separation of all data into numerous mutually exclusive subsets. (b) internal nodes (chance nodes) that denote one of the likely options offered at that

moment of the tree. It is usually linked to the parent on top and the child at the bottom. (c) leaf nodes(end node) that indicate the end result of a blend of sets of decisions. Branches denote possible outcomes or occurrences that arise form root nodes and internal nodes. A hierarchy or network of branches form a decision tree. All paths from the root through the internal nodes to the leaf showcase a categorization of a decision rule.[2]

A graphical representation of a tree is below:.(McSharry,2021)



The root number 2 is the root node and 7,5 are the children of the root. Nodes 2,10,6 and 9 are children to nodes 7 and 5. nodes 2 and 10 don't have children while nodes 6 and 9 have children.

Why might it be necessary to prune the tree?

One of the most common issues with understanding a decision tree is getting the optimal size of the resulting tree which leads to good accuracy of the model. A tree that has too many branches and layers can lead to **overfitting** of the data. This is where pruning comes in. it helps to avoid overfitting the data in order for the model to generalize well to unseen data. Pruning involves getting rid of subtrees that are redundant and have no relationship with the independent feature and replace it with a leaf node. Pruning can be done in 2 ways i.e. (a) pre-pruning and (b) post-pruning.[3]

Why are decision trees an attractive method for classification in practical applications?

This is because:[3]

a. Decision trees help in predicting feedback to input variables
b. Decision tree can help make a complex system relatively easy to use in practice.
c. Decision trees are free of ambiguity.
d. Decision trees are robust even when there re some missing values
e. Both discreet and continuous variables can be used interchangeably as dependent or independent variables.

2.2. Suppose an organization has built a rule-based classifier using domain knowledge. After collecting a large amount of data, outline the steps required to improve upon the existing approach by constructing a data-driven classifier.

Data driven classifiers are build using the following steps:

a. Data extraction
b. Data cleaning
c. Analyzing data

d. Identify critical features
e. Choose a workable learning algorithm(supervised, unsupervised or reinforcement learning).
f. Evaluate then deploy the model

How would you advise to test the validity of the new model?(5 marks)

To test the validity of a new model, the following can be done:

a. Doing data checks on the model by diving the data into training and test split.. this will help clarify the the missing and outliers in the data sets
b. Check the type of model variables. This is done by the variance information factor that indicates the intercollinearity between the dependent and independent variables. If a VIF is less than 2 in the training set then there is no multicollinearity.
c. Check how the model fits. The area under the receiver operator curve is used to measure the predictiveness power of a model. An AUROC of 1 indicates that the model predictive power is perfect.
d. Check the coefficient stability- this helps evaluate the sign and coefficient stability of a model. If the coefficients are within 95% confidence interval then they are stable.

2.3. Consider the challenge of classifying the likelihood of survival using the Titanic dataset. Construct a decision tree and display the structure of this tree using a graphic. (5 marks)

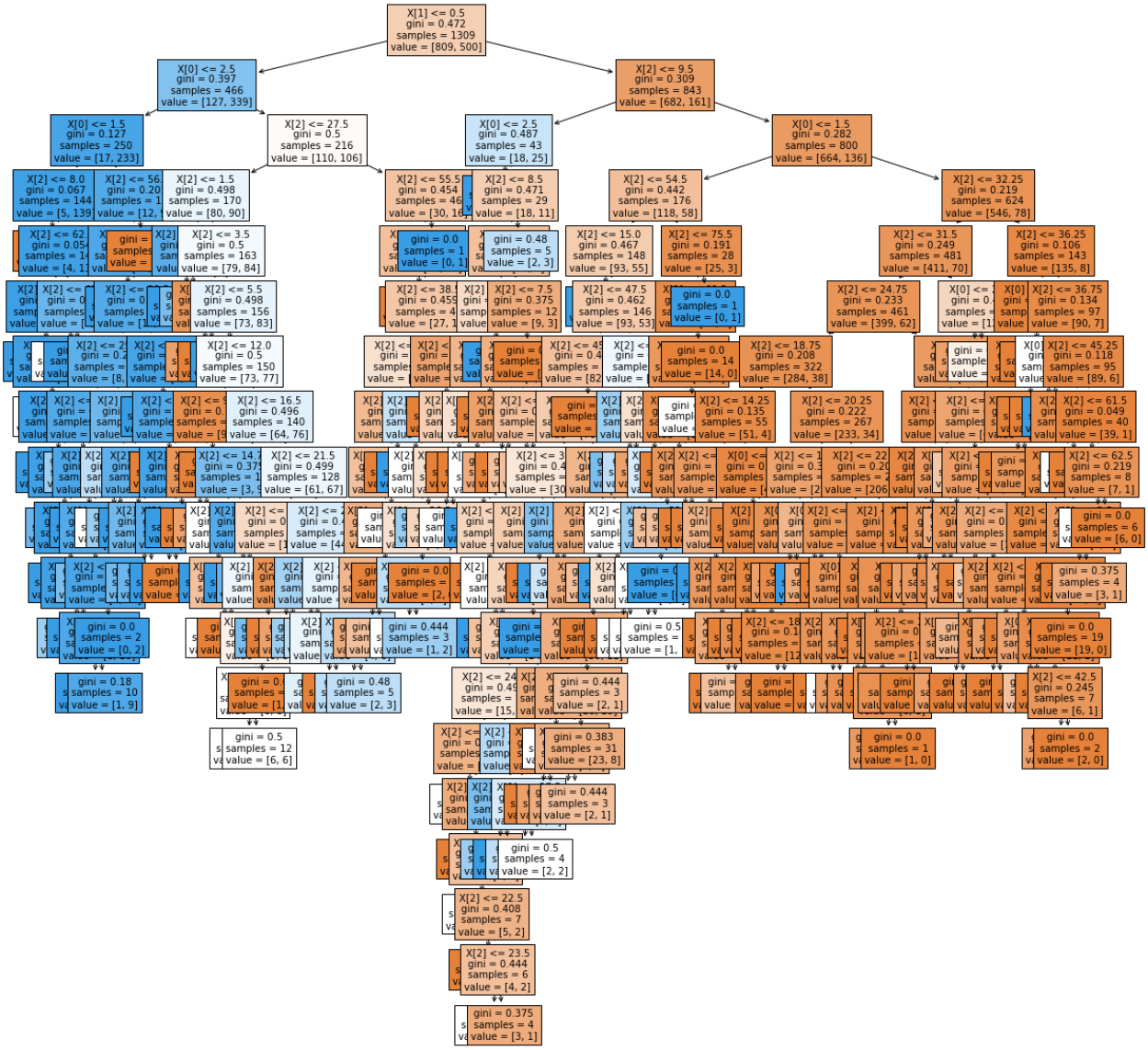
I imported the titanic data into my working envoromnet and extracted the required columns. In order to analyse the data, the sex column values had to be converted from strings to numerical values. I created two columns, female and male, using get_dummies function from pandas which will be used to store the numerical values corresponding to the sex string values. I passed in the sex column of the Titanic dataset as my variable which generated a sub data frame with female being represented by a zero and male represented by one. I then dropped the first column because only one column is enough to represent both men and women. I finally used the concat ()method to join or add the sub data frame to the main data frame Titanic.

The two sub data frames generated(x as predictor and y as dependent) is what was used for all the analysis. Sklearn.tree library has inbuilt functions that help in construction of decision trees. A model was created using DecisionTreeCLassifier library and passed in the x and y variables when fitting the model.I used the tree.plot tree function to plot the default decision tree. Figire 2.1 below is the graphical representation of the decision tree.

To check the accuracy of the default decision tree, I calculated the cross validation score using the cross_val_score from sklearn library. The results are as below:

```
The model accuracy before prunning is:  73.33
```

Result Q2: Decision tree1



Inference

The decision tree has a depth of 18, with the first classifion being x1<=0.5. 809 samples had values less than or equal to 0.5 while 500 samples had values greater than or equal to 0.5. the last leaf node had 4 values that had a gini of 0.375

From the cross validation score, we can infer that the default decision tree classifier is 73.33% accurate.

To compute misclassification I used the decision tree classifier and checked the cross validation score fro insaplde data and subtracted from 100. The error is:

```
The in sample missclassification error is:  26.67
```

2.4. Evaluate the performance of the tree (before and after pruning) and provide results using cross-validation. (5 marks)

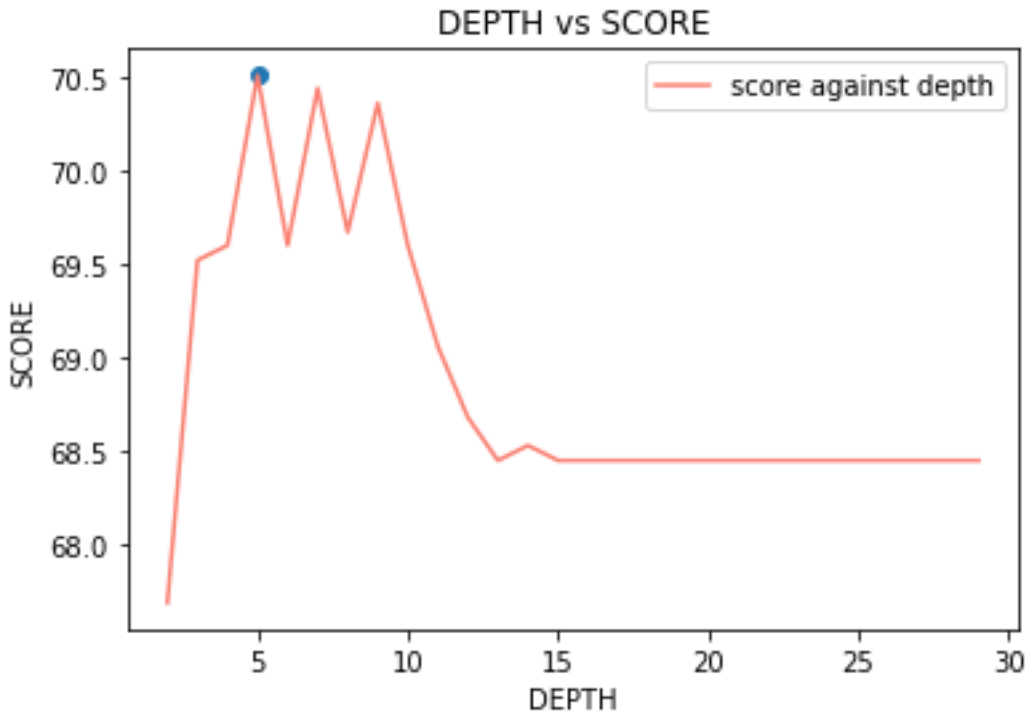
In order to generate the best pruning level, I created a for loop which would loop througha range of different tee depths and calculates accuracy score. The different depths and ranges are stored in 2 empty lists earlier initialized. I set the depths to 2 – 30. The accuracy score was calculated using the cross validation. The lists generated are:

Depths: `[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]`

Accuracy: `[67.69, 69.52, 69.6, 70.51, 69.6, 70.44, 69.67, 70.36, 69.6, 69.06, 68.68, 68.45, 68.53, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45, 68.45]`

I plotted the above for clear visualization as shown in figure 2.4a

Figure 2.4a



Inference

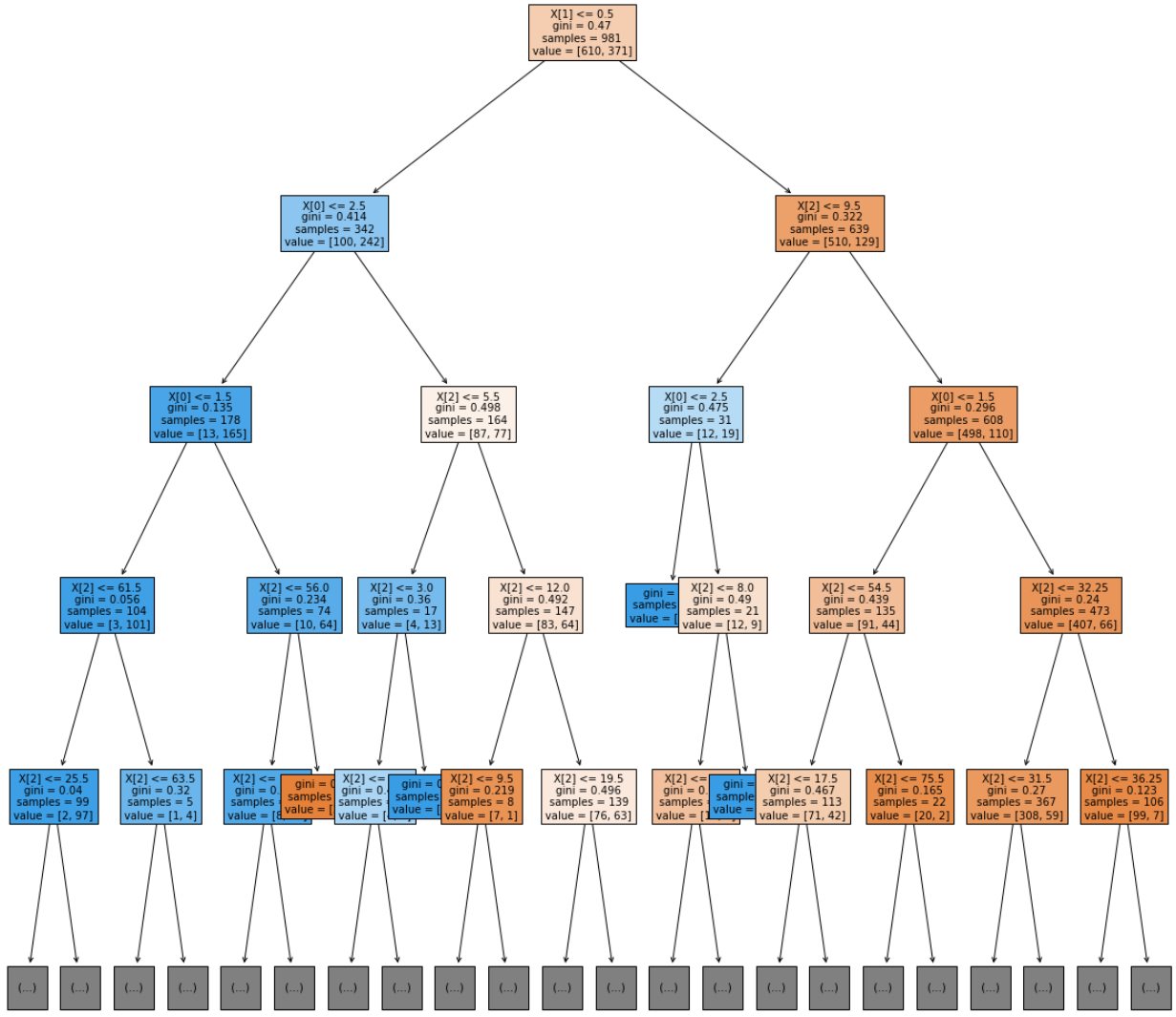
The graph above shows the different ranges of depths of the tree against the calculated scores. We can see that the highest score was 70.51 with a depth of 4. The accuracy score plateaued at 68.45 when the max depth of 12 was attained.

The results obtained from above were used to prune the decision tree. A new tree was generated using the maximum tree depth of 4 as shown **on figured 2.4b**

Using cross validation, I generated the accuracy of the pruned tree which is:

```
Accuracy of the pruned decision tree is 78.223 %
```

Figure 2.4b



Inference

The pruned tree performed better than the default decision tree with an accuracy of 78.2%. This means the depth of 2 was the optimum to evaluate the model for this classifier.

2.5. Compare the final tree with logistic regression and comment on the advantages and disadvantages of both. Which model is best for competing in the Kaggle competition? (5 marks)

Using the LogisticRegression function from sklearn library using the trained x and y data of the titanic data set. I ran a cross validation score to generate the accuracy score. the insample score is **78.6% was generated** while the outsample data score was 76.76.

The accuracy of the default and pruned tree

The decision tree classifier accuracy is **73.33%** which is lower that the logistic regression model. I would use the Logistic regression model in the Kaggle competition because it performed better.

Advantages and disadvantages of decision tree[4]

Advantages	Disadvantages
Requires less effort in preprocessing and data preparation	A small shift in data can cause instability
It doesn't need normalization	Calculations can become complex compared to other algorithms
It doesn't need scaling of data	Takes a lot of time in training the model
Missing values don't affect the constructionof the tree	Very expensive because of time and complexity
It is easy to make inference.	It is not enough to apply regression and predicting continuous values.

Advantages of Logistic regression[5]

Advantages	**Disadvantages**
Simplest algorithim	Can lead to offerting incase of high dimensional datasets
Inference of importance of each feature is easily predicted	Non linear problems cannot be solved with logistic regression.
Model can be easily updated with new data	It is difficult to measure complex relationships using logical regression
Well outlines probabi;ities are given as outputs	Repetitionof information could result in a wrong training of parameters
It is less prone to overfitting in data sets with a small sample size	Logistic regression needs independent variables to be linearly related to log odds
Easy and quick to implement	Models predictive value can be degraded if important features are not used in the model
Very efficient when the dataset are linear features	It requires a large volume of data sets for optimum results

3. Classification using KNN (25 points)

3.1 By focusing on small neighborhoods of state space it is possible to construct parsimonious models. Describe the concept behind this general approach and a step-by-step procedure for implementing such a model. (5 marks)

The concept behind small neighbourhoods is to get the optimum number of points in data set. This can be achieved by setting a range of K values ,then by using the decision tree classifier model, we loop through the range with the k value as the maximum depths. The different scores are recorded in a list and the best obtained accuracy is selected as the optimum. The corresponding depth is what is used as the k value to create a tree.

3.2 Consider the challenge of classifying the likelihood of survival using the Titanic dataset. To construct a KNN classifier, how will you transform the available variables? (5 marks)

The sex column is of the data type string. Inorder to perform an analysis I transformed the data into numerical values. I first generated two columns to represent both gender(female and male) using get_dummies function from pandas. I passed in the sex column of the Titanic dataset as my variable which generated a sub data frame with female being represented by a zero and male represented by one. Since both column alsost infer to the same thing, I dropped the first column and remained with one which represented both male and female.I finally used the concat ()method to join or add the sub data frame to the main data frame Titanic.

3.3 Calculate the performance of the classifier versus the number of neighbors used and provide a graphic to display the result. What is the optimal number of neighbors using cross-validation? (5 marks)

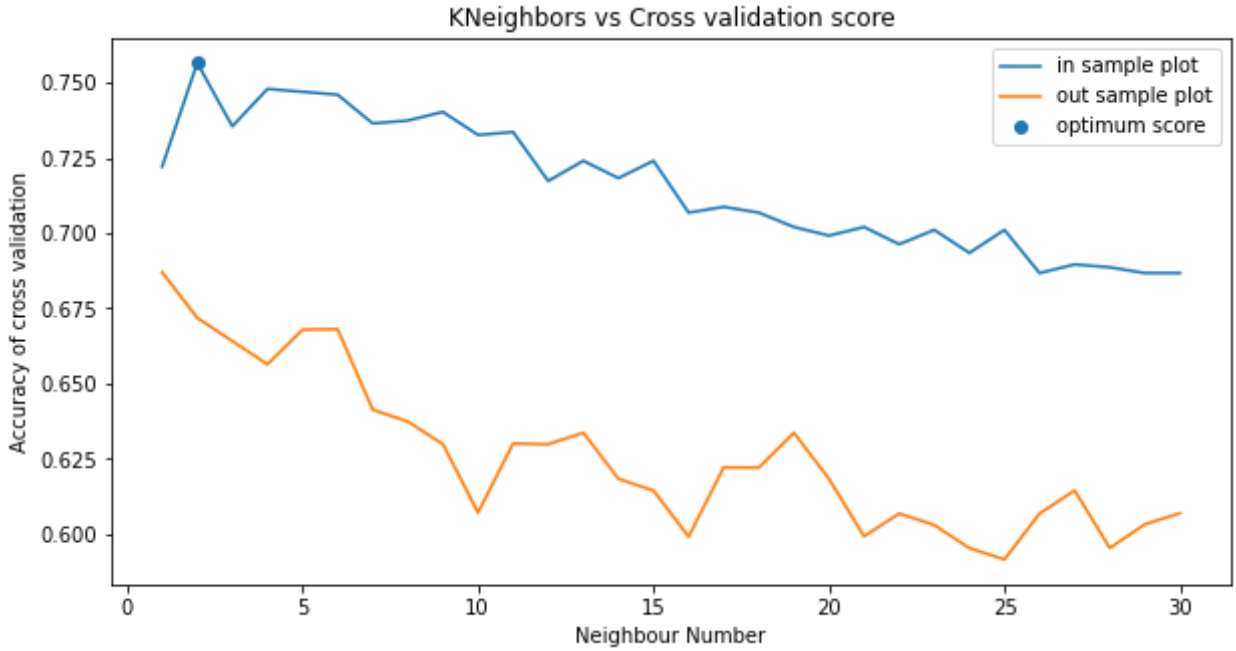
Using the titanic dataset imported in question 2, I extracted the required columns(pclass, gender, age and survived) and created a subdata frame. I split the data before doing any analysis.

To generate the in-sample K neighbours, I used the KNeighbours Classifier for sklearn.neighbours libracy as my model. I passed the in sample data(x and y train datasets) as my variables and fit the model. I calculated a cross validation loss by using the cross validation score function and set my neighbours to 7. I also did the same for the out sample (x and y test datasets) to generate the outsample neighbours and cross validation loss. The results are:

```
Default insample cross validation score is 75.075
Default outsample cross validation score is 65.637
```

I looped through a range of 1-31 neighbours to test the score on different depths of neighbours for both the in and out sample datasets. I stored the resulting scores in a list and generated a graph of the same. The results are as shown in **figure 3.3a**

Figure 3.3a



Inference

The graph above shows the in and out sample xcross validation scores using different ranges of neighbours. The optimum neighbors in the insample **is 2 with a score of 75.64**. Both the grahps have a negative slope which infers that as the number of neighbors increase the cross validation score reduces.

3.4 Explain why some distance metrics are sensitive to the kind of features used. Evaluate the performance using different distance metrics. (5 marks)

Choosing a distance metric is very important aspect in multimedia information retrieval. It is crucial when constructing a hierarchy over points. The Euclidean metric is good in almost all instances and produces a balanced hierarchy however it tends to merge disjoints features when they are close enough. The mahalbois metric takes care of the disjoint by wrapping the spaces or distances along the local normal direction using PCA . It leads to partitions that are more natural and intuitive. It is considered better than Euclidian metric since it is used when the surface is too complicated. The chebychev metric estimates the highest level of coordinate difference in a model. Correlation metric comutes oneminus the sample linear correlation between observations.[5]

The different metrics are sensitive because they influence the importance of inputs in a model and their relevance. These are defined by the number of neighbours, the disctance metric used and their weights.

To evaluate the performance of the KNN classifiers I used 3 different distance metrics, chebyshev, Euclidian and hamming. I created a for loop that applied the different metrics and stored the scores in an array. The results are as below:

```
Distance metric scores are: Chebyshev ['72.01',euclidean '75.45',hamming
'70.48']
```

From the result we can infer that the best metric to be used in this case is **Euclidean with a score of 75.45**

3.5 Compare the best KNN classifier with logistic regression and comment on the advantages and disadvantages of both. Which model is best for competing in the Kaggle competition? (5 marks)

I created two models, logistic regression and KNN classifier and computed their cross validation score. I fit the models using the in sample data. The results are as below:

```
Logistic regression score is 78.22
Best KNN score is 76.024
```

From the results we can infer that Logistic regression model was the best in this case with a score of 78.22.

Advantages and disadvantages of Logistic regression

Advantages	**Disadvantages**
Simplest algorithim	Can lead to offerting incase of high dimensional datasets
Inference of importance of each feature is easily predicted	Non linear problems cannot be solved with logistic regression.
Model can be easily updated with new data	It is difficult to measure complex relationships using logical regression
Well outlines probabi;ities are given as outputs	Repetitionof information could result in a wrong training of parameters
It is less prone to overfitting in data sets with a small sample size	Logistic regression needs independent variables to be linearly related to log odds
Easy and quick to implement	Models predictive value can be degraded if important features are not used in the model
Very efficient when the dataset are linear features	It requires a large volume of data sets for optimum results

Advantages and disadvantages of KNN classifier.

Advantages	**Disadvantages.**
Ther e is no training period required	**Doesn't work well with huge datasets**
New data can be added anytime	**It doesn't do well where the dimensions are high**
Easy to implement	**It needs feature scaling**
	It is very sensitive to noisy data, missing values and outliers.

4. Regression – wine quality (25 points)

4.1 Calculate the average of each feature for the red and white wines separately and make a comparison using a bar graph showing the two wines together. How do the

results relate to common sense (or the intuition of a wine expert) based on the features that are available?(5 marks)

I imported the 2 data frames into my working environment from the provided link and used a separator to properly format the data into the required columns separately. To generate the mean of the two data frames, I passed the mean() method and passed in the 2 data frame separately as created 2 sub data frames of the mean stored in a new column. The means generated are as shown in **figure 4.1a**

Figure 4.1a

Red wine mean

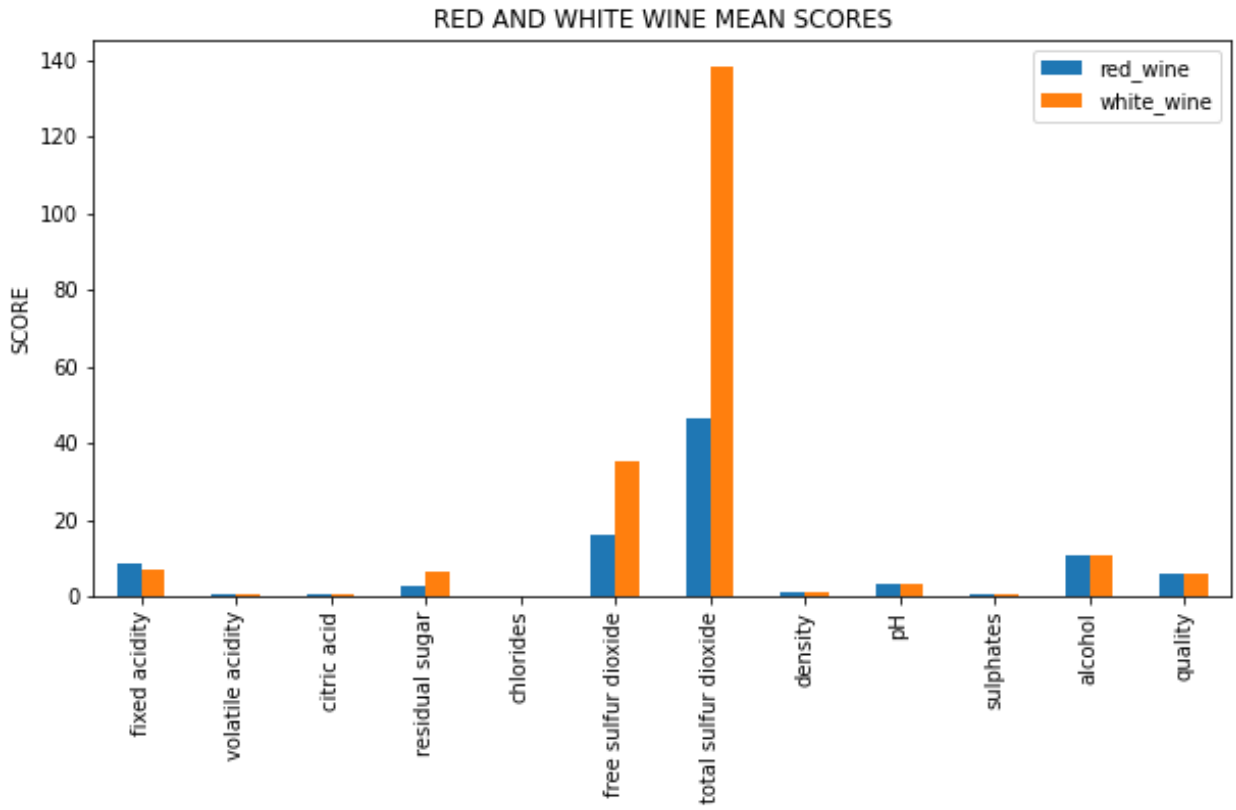
	mean_red
fixed acidity	8.319637
volatile acidity	0.527821
citric acid	0.270976
residual sugar	2.538806
chlorides	0.087467
free sulfur dioxide	15.874922
total sulfur dioxide	46.467792
density	0.996747
pH	3.311113
sulphates	0.658149
alcohol	10.422983
quality	5.636023

White wine mean

	mean_white
fixed acidity	6.854788
volatile acidity	0.278241
citric acid	0.334192
residual sugar	6.391415
chlorides	0.045772
free sulfur dioxide	35.308085
total sulfur dioxide	138.360657
density	0.994027
pH	3.188267
sulphates	0.489847
alcohol	10.514267
quality	5.877909

I plotted the above mean as shown in the figure 4.1b below:

Figure 4.1b



Inference

The graph above illustrates the mean of all the features of the red and white wines. The features with the highest mean is total suphur dioxide for both wines. This makes it the most important feature. Alcohol content is almost the same for both wines. Volatile acidity, chlorides,density sulphates and citric acid have the lowest mean for the two wine samples.

4.2 What is the correlation between each feature and the dependent variable using a separate analysis for white and red wine? Which variable is most relevant for each wine?(5 marks)

to compute correlation of the features, I extracted all the required features and stored in a dta frame for the two data frames. I used pearson correlation to cuompute the correlation. I extracted the correlation values of features against quality. The results are as shown in f**igure 4.2a**

Figure 4.2a

```
Red wine correlation
                       quality
fixed acidity         0.124052
volatile acidity     -0.390558
citric acid           0.226373
residual sugar        0.013732
chlorides            -0.128907
free sulfur dioxide  -0.050656
total sulfur dioxide -0.185100
density              -0.174919
pH                   -0.057731
sulphates             0.251397
alcohol               0.476166
quality               1.000000
```

```
White wine correlation
                       quality
fixed acidity        -0.113663
volatile acidity     -0.194723
citric acid          -0.009209
residual sugar       -0.097577
chlorides            -0.209934
free sulfur dioxide   0.008158
total sulfur dioxide -0.174737
density              -0.307123
pH                    0.099427
sulphates             0.053678
alcohol               0.435575
quality               1.000000
```

Inference

From the figure above we can infer that the most important features for both wines is alcohol because it has the highest correlation with quality.

defines the dependent and independent variables. The wine quality is the y variable(dependent) while the rest of the features were my independent variables. I used the pearsons corr () function to compute the correlation of all the feature

4.3 Use Lasso and cross-validation to provide a plot of MSE against lambda and the parameter estimates versus lambda. How do the features selected by LASSO compare with an approach of setting a threshold on the absolute correlation coefficient? (5 marks)

Figure 4.3a

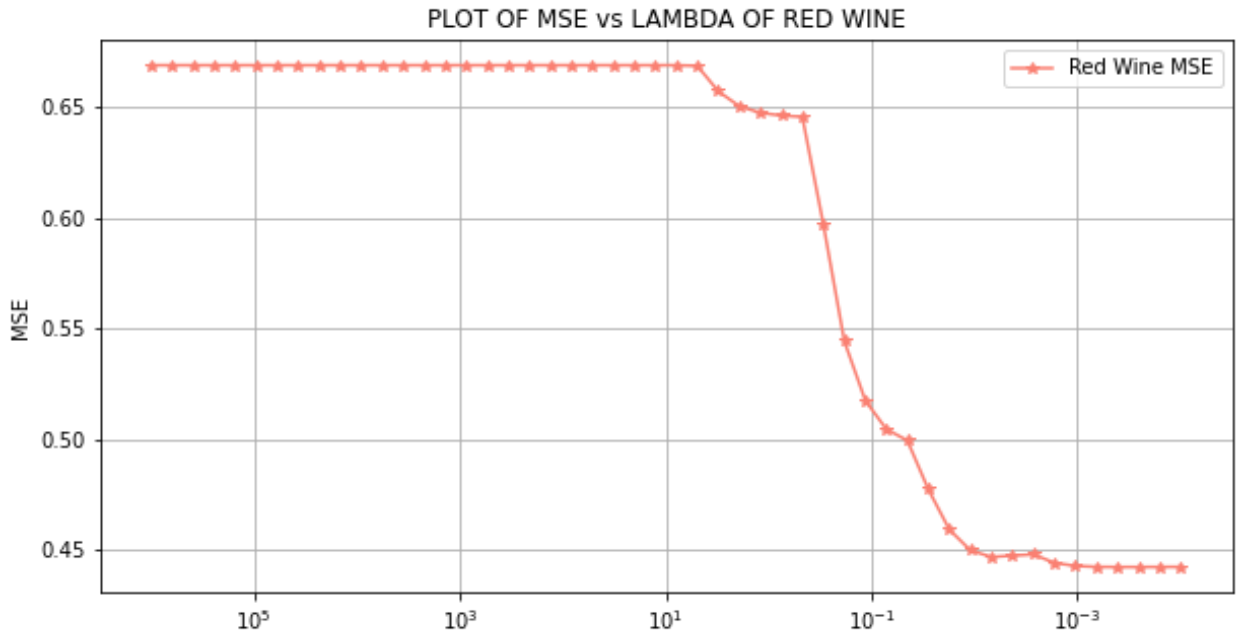
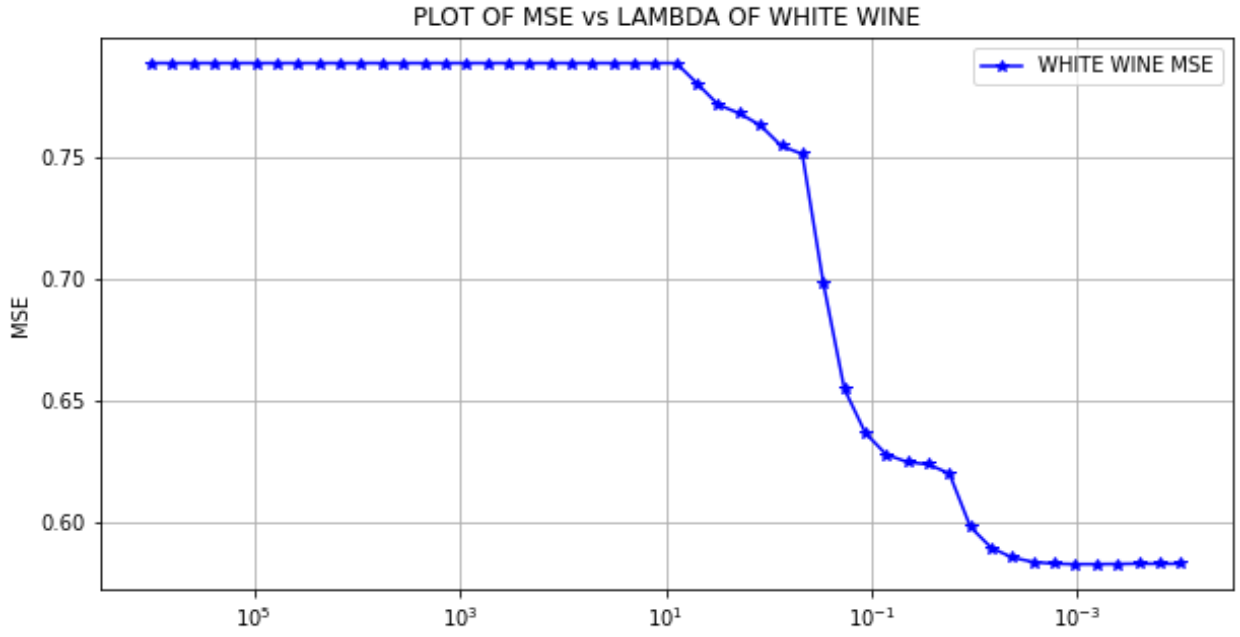


Figure4.3b



How do the features selected by LASSO compare with an approach of setting a threshold on the absolute correlation coefficient?

The model dropped all features that had a zero estimate. The best features selected are;

```
Index(['fixed acidity', 'volatile acidity', 'free sulfur dioxide',
       'total sulfur dioxide', 'sulphates', 'alcohol'],
      dtype='object')
```

Figure 4.3a

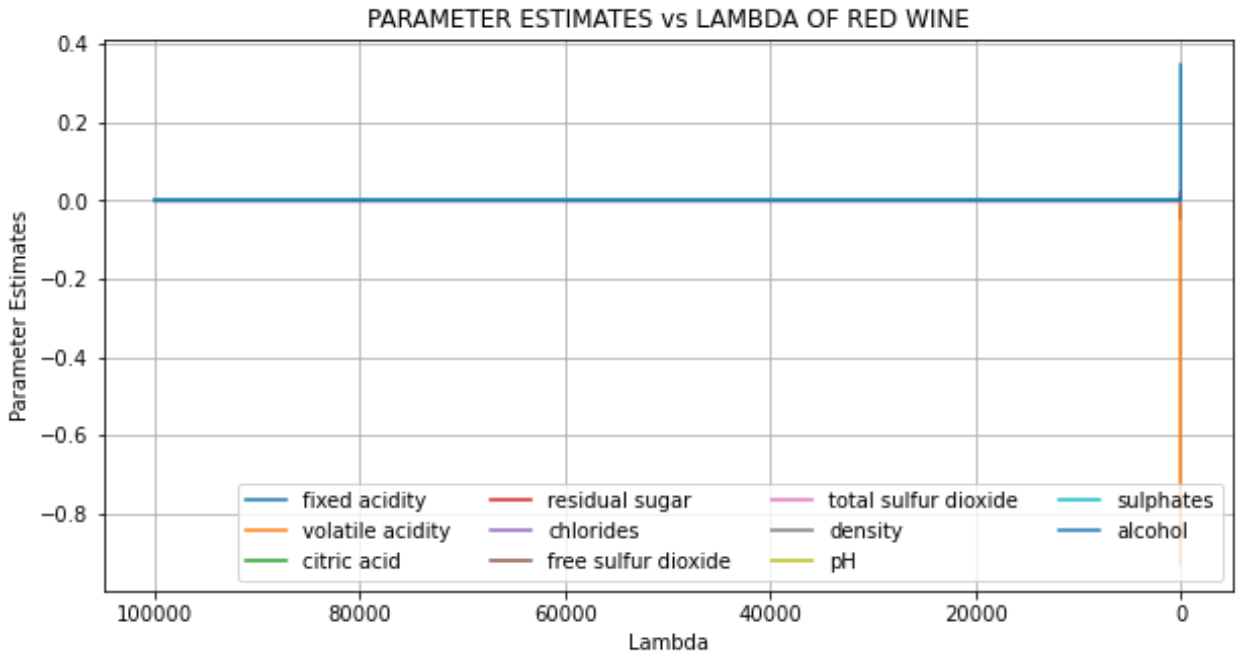
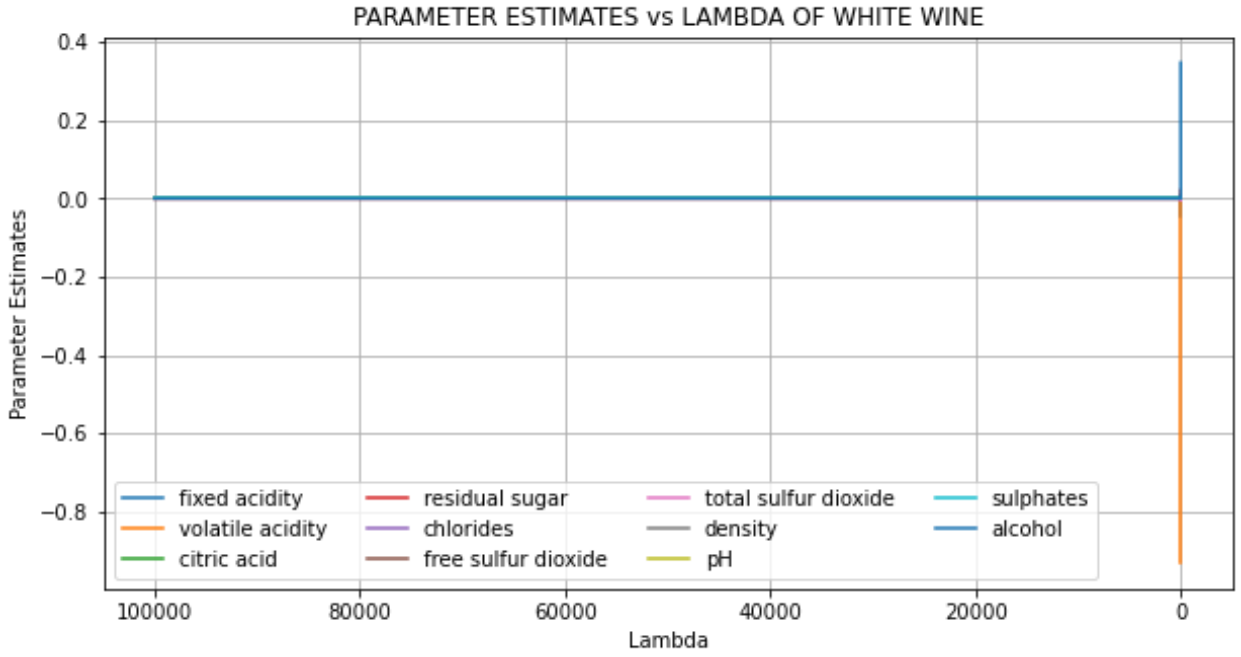


Figure4.3b



4.4 Use the features identified by LASSO to construct a KNN regression model for red wine. (5 marks)

I created a KNN regression model for the best selected feature of red wine using KNeighborsClassifier () method. I fit the best features data selected by the lasso model. I generated a prediction of y variables.

4.5 What is the performance of a linear regression model and the KNN model, measured by MSE and R_2? Describe the advantages and disadvantages of both models.(5 marks)

To compute the performance of the linear regression and KNN model, I created the respective models and measured the accuracy with MSE and R2. The results are as below:

```
MSE score for KNneighbours is 0.545
R square score for KNneighbours is 0.164

MSE score for Linear regression is 0.425
R square score for Linear regression is 0.348
```

The best performing model is Linear regression because it has the lowest **MSE score of 0.425**

Describe the advantages and disadvantages of both models.[7][8]

Linear regression model.

Advantages	Disadvantages
Performs very well when dataset is linear	The model assumes linearity in the data set
Easy to implement	Chances of overfitting are high
Ovefitting can be easily avoided	Multicollinearity can happen in the model
	The model issensitive to outliers

KNN model

Advantages	Disadvantages
Very easy to implement	The model doesn't perform well with big data sets
No training period is required by the model	The model doesn't perform well eith high dimensions
New data can be incorporated into the model easily	There is a requirement for feature scaling

Refrerences.

1. https://stats.stackexchange.com/questions/181298/linear-and-nonlinear-model-parsimony

2. Hastie TJ, Tibshirani RJ, Friedman JH. The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition. Springer; 2009. ISBN 978-0-387-84857-0 [Google Scholar]

3. https://www.educative.io/edpresso/what-is-decision-tree-pruning-and-how-is-it-done

4. https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

5.

6. http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-knn.html

7. https://towardsdatascience.com/k-nearest-neighbours-explained-7c49853633b6

8. https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/