ASSIGNMENT 6

Andrew-ID: parmenin

DATA INFERENCE AND APPLIED MACHINE LEARNING (18-785)

11/21/22

Niyomwungeri Parmenide ISHIMWE

I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: **Niyomwungeri Parmenide ISHIMWE**
Andrew ID: **parmenin**
Full Name: **Niyomwungeri Parmenide ISHIMWE**

The libraries used:

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- from sklearn import tree
- from sklearn.preprocessing import LabelEncoder
- from sklearn.model_selection import cross_val_score
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.linear_model import LogisticRegression, Lasso, LassoCV, LinearRegression
- from sklearn.metrics import r2_score, mean_squared_error, accuracy_score

QUESTION 1: NONLINEARITY

1.1) It may be required to consider non-linear correlations between variables because there are some statistical situations where independent and dependent variables do not have a straight-line or direct relationship. In this situation, the rate of increase or decrease or the slope of the curve showing the relationship can change as one of the variables changes. In a nonlinear relationship, change in any of the inputs is not proportional to changes in the output. Because it is flexible, it can be useful to produce appropriate results when similar future situations or inputs are present, such as choices, categorization, or inferences, a more flexible nonlinear analysis is required. A nonlinear relationship forms a curve when it is plotted on a graph as opposed to a straight line when a linear relationship is presented. One popular example of a non-linear relationship is a relationship between weight loss and the amount of sport exercise a person does.

1.2) Non-linear relationships can be modeled mathematically in various shapes like exponential, logarithmic, logistic, trigonometric, power functions, Lorenz curves, Gaussian, and so on. the following formula is the general equation of non-linear regression.

$Y = f(X, \beta) + \varepsilon$ where **X** = a vector of p predictors, **β** = a vector of k parameters, **f (-)** = a known regression function, and **ε** = an error term [1]. It is applicable for forecasting, predicting, and financial modeling. It can be appropriate for agricultural research purposes because many crops and soil relationships require non-linear relationships to produce reliable results.

1.3) A parsimonious model uses the fewest feasible predictor variables to provide the desired level of explanation or prediction. And so, a nonlinear model can be more parsimonious than a linear model because, in statistical modeling, there are occasions when a non-linear model with fewer parameters fits the data well but using a linear model would require many more parameters to get a decent fit [2].

Even while linear regression is simpler to use and understand, the types of curves it can fit are somewhat limited. There are situations when using fewer data won't allow it to fit the curve. Contrarily, nonlinear regression can fit many more varieties of curves, but it might be more difficult to determine the optimum fit and understand the significance of the independent variables. [3]. The mathematical formula for linear models as provided in the course notes can be written as: **$y = a + bx + \varepsilon$** where **a** is the intercept (also known as constant); **b** is the slope (indicates how y depends on x), and the **ε** is the model errors or residuals. From the previous question, the general equation for non-linear regression can be written as **$Y = f(X, \beta) + \varepsilon$**. This shows that a linear formula's variable can only be raised to the power of 1 only which indicates that its slope is constant, while the one for non-linear can be raised to the power of 2 or more which indicates that its slope is variable [4].

1.4) According to the course notes, the characteristics that are typically preserved when generating surrogates are:

- Specifying a null hypothesis H0 describing a linear process and then generating several surrogate data sets according to H0 using Monte Carlo methods.
- Calculating a discriminating statistic for the original time series and all the surrogate sets.
- If the value of the statistic is significantly different for the original series than for the surrogate set, the null hypothesis is rejected, and non-linearity is assumed.

As listed in course notes, two of the surrogate techniques or methods are random shuffle and random phases. Random shuffle surrogates are generated by randomly shuffling the original data, obtaining the same distribution, and destroying linear correlations. Random phases surrogate data are generated by

the inverse Fourier Transform of the amplitudes of the Fourier Transform of the original data with new (uniformly random) phases. This approach preserves the linear correlations in the data.)

1.5) Mathematically, information is defined as **$I(p) = \log(1/p) = -\log(p)$** and this definition satisfies three properties such as **$I(p) \geq 0$** to mean that information is a non-negative quantity, **$I(1) =$** 0to mean that events that always occur do not communicate information, and **$I(p1\ p2) = I(p1) + I(p2)$** to mean that information due to independent events is additive.

Entropy is a measure of disorder or expected surprise. it measures the average uncertainty in the value of the discrete-valued probability density. The entropy H(X) of a discrete random variable is defined in a mathematical way as **$H = -\Sigma k\ p(k) \ln p(k)$**, where p(k) is the probability of recurrences times, with k=1…, K.

Mutual information given as **$I(x,y) = H(x) + H(y) - H(x,y)$** is a metric that quantifies the amount of information that one variable provides about another variable and is typically measured in units called "bits". A value of zero indicates no dependence and larger values imply the existence of a relationship (which might be linear or nonlinear).

The degree of regularity and unpredictability of variations in time series data can be measured using the approximation entropy, which is also used to estimate entropy. Entropy can be used to determine the regularity in heart rate.

Mutual information is a method for calculating the amount of information that one variable informs another. To choose the best features, one can compare the impact of each feature on the dependent variable using mutual information between each feature and the dependent variable. Mutual selection may be preferable to correlation since a feature may be less correlated with a dependent variable while still offering a wealth of information about the dependent variable.

<u>QUESTION 2:</u> CLASSIFICATION USING TREES

2.1) According to the course notes, a decision tree is a type of model that enables straightforward decision-making based on the observed values of several input features. It is a kind of flowchart that shows the reasoning behind a course of action. Its algorithm is useful in data analytics that uses conditional "control" statements to categorize data. A decision tree has a single node at the beginning or the starting point and branches or splits or leaves in two or more directions from the top node as possible outcomes. Each branch contains a vast range of options that can be reached by several decisions and arbitrary occurrences until a conclusion is reached. Following the decisions in the tree from the root

(starting) node, down to the leaf node that carries the response will therefore allow ascertaining the response.

Because the decision tree is constructed so that it uses variables that are relevant for determining the dependent variable, the first decision nodes highlight the variables that are most important in this nonlinear context. Sometimes, decision trees can get quite complex so they can end up giving irrelevant information and excessive weights. This issue can be solved by using the process called "pruning" which removes any unneeded or low-important branches from the tree, which in turn prevents overfitting and improves the performance of the decision tree. Eventually, variables with little or no information on the dependent variable are removed through this optimal pruning.

Because they are much faster and more effective than other classification algorithms, decision trees are appealing methods for classification in real-world applications. They are simple to comprehend, analyze, and visualize. No data normalization is necessary because decision trees can handle any type of data, including category, numerical, and Boolean data. Any sort of data can use it, although categorical predictors are particularly effective in this regard. Basic customer segmentation may be one of many of its applications. Based on the segments, other predictive models can be created to act on those segments. Another example of the application of a decision tree is the USSD short code options that guide the user to get to a particular outcome or decision. Additionally, making real-life decisions may depend on algorithms like this of a decision tree by considering different factors as nodes [5].

2.2) Rule-based classification uses a set of IF-THEN rules for classifying the data into different categories based on pre-established rules. For instance, IF a certain condition is met, THEN a certain decision is made. Data-driver classification does not depend on any rule set by the user to categorize data, it only depends on the data. Important steps needed in constructing data-driven could be as follows: first, **data preparation**, which entails gathering, cleaning, and preparing data to use in the classifier; second, **data splitting**; data are divided into training and test sets, which are used to fit and train the model; and third, **classification model building**, which involves choosing appropriate model, training it using training data, testing it and evaluating its accuracy and then ask it to make predictions.

The build classifier would be validated by evaluating its accuracy using the testing data to assess its performance in predicting. This allows for determining how accurately the model can produce valid results. Second, by creating a logistic regression model with the same data set as the classification model and evaluating its accuracy, one can create a logistic regression model. In the end, it will be best to compare the accuracy of the two models and determine which has the higher accuracy [6].

2.3)

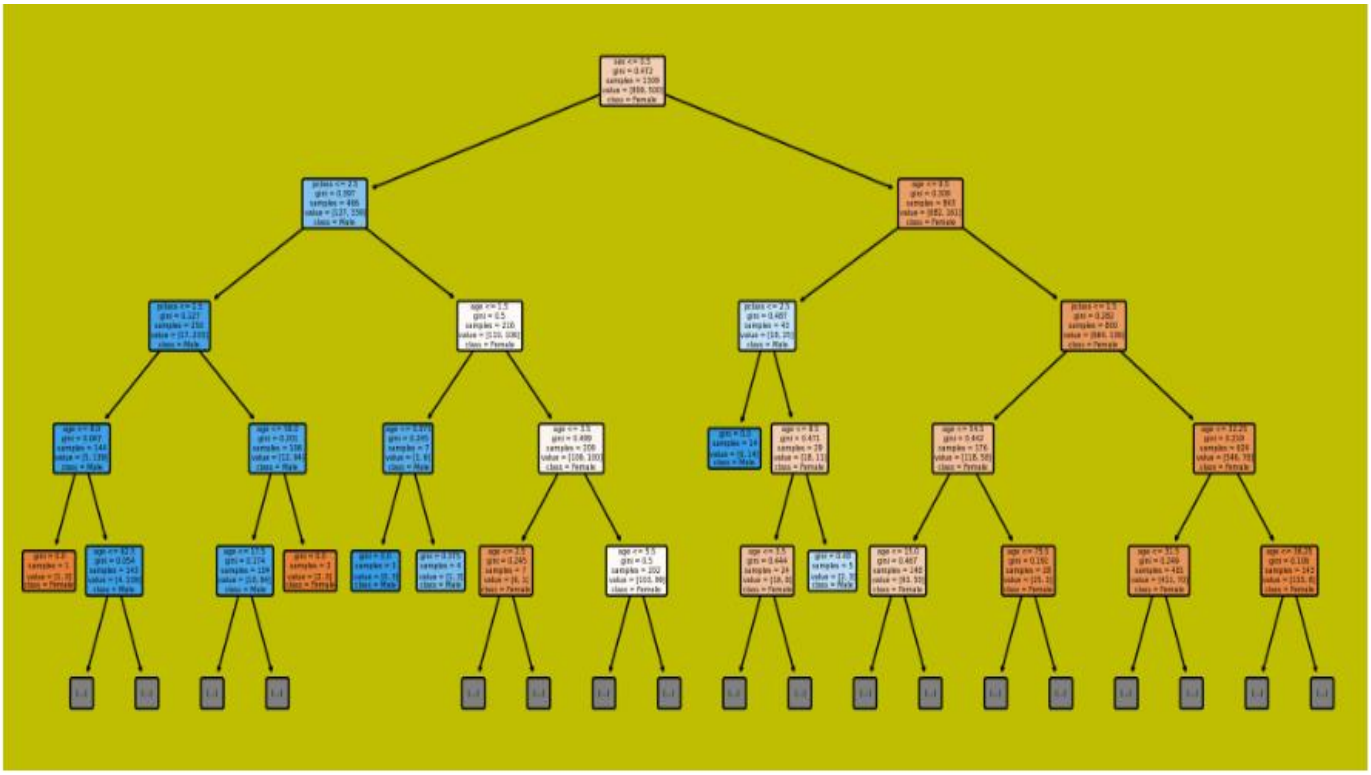

Figure 1: Full decision tree

2.4) To evaluate the performance of the tree (before and after pruning), the decision tree is built and fitted using the sklearn tree's DecisionTreeClassifier and plotted tree's plot_tree function to generate a full-sized tree. By evaluating its performance by cross-validation, a 0.7745977686435701 score is obtained. After that, the tree is pruned or optimized using the sklearn tree's DecisionTreeClassifier setting the maximum tree's depth and optimization criterion. This optimization uses the classification optimization criteria called Gini Index, which is maximum when all observations are equally distributed across the classification classes and minimum when all observations belong to the same class. After pruning, the performance increased a bit reaching a score of 0.7791720493247211.

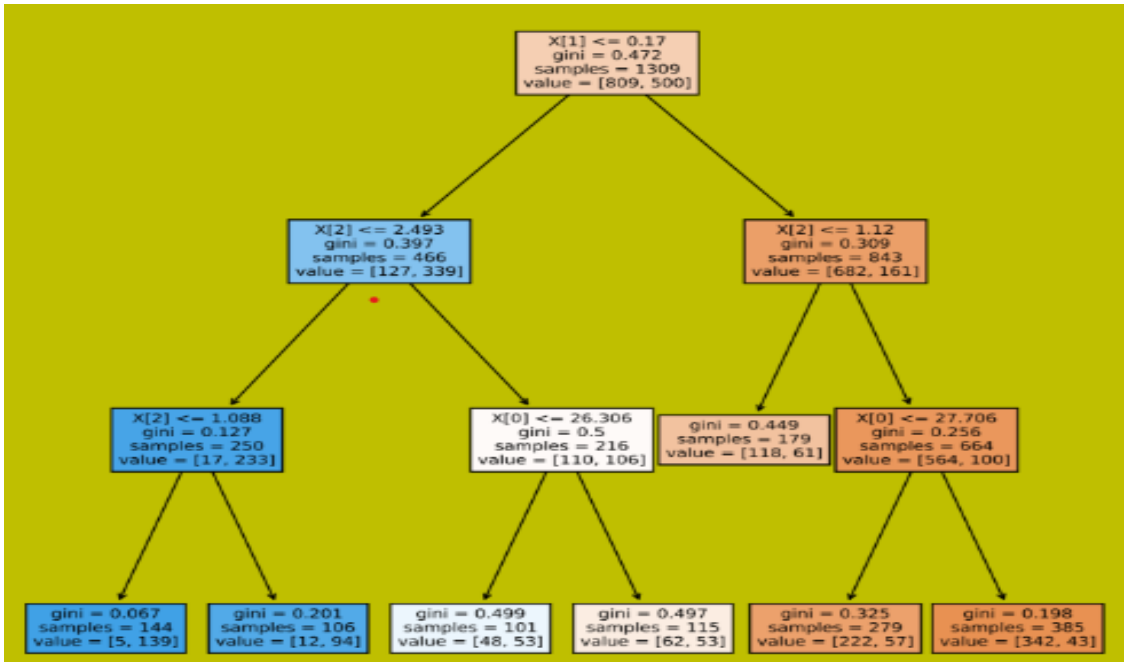

Figure 2: Decision tree after pruning

2.5) After building a logistic regression model using cross-validation techniques (in-sample and out-of-sample datasets), the model shows scores of 0.7532413388138579 which is less than 0.7791720493247211 of the pruned decision tree. This means that the pruned decision tree model has fewer errors than that of logistics. By computing mean absolute errors (MAE), it is seen that the decision tree has an MAE of 0.2223546682325308 and logistics gives 0.2147269524368761.

QUESTION 3: CLASSIFICATION USING KNN

3.1) The three components that define a KNN model are the number of nearest neighbors, the distance metric, and the weighting system. To construct a parsimonious model, small neighbourhoods state space is considered. Implementing a small neighborhood parsimonious model starts with identifying the k nearest neighbors from the training data set. The probability that the provided point belongs to the class to which the k belongs is determined in the second step. The provided point is categorized in the class in which it has the highest probabilities in the third phase.

3.2) With the challenge of classifying the likelihood of surviving the titanic, to construct the KNN classifier the available variables should be transformed. Transforming the data for the sex variable into numerical values and figuring out how to fill in the gaps in the age variable would be very helpful. The classifier wouldn't rely on variables with huge scales if they scaled all the variables to the same level, which would increase accuracy.

3.3) The performance of the classifier versus the number of neighbors is calculated and found that the optimal number of neighbors using cross-validation is 4 since it has a minimum loss of 0.25142102172636527 compared to other numbers of neighbors and this result is depicted in a graphic below.

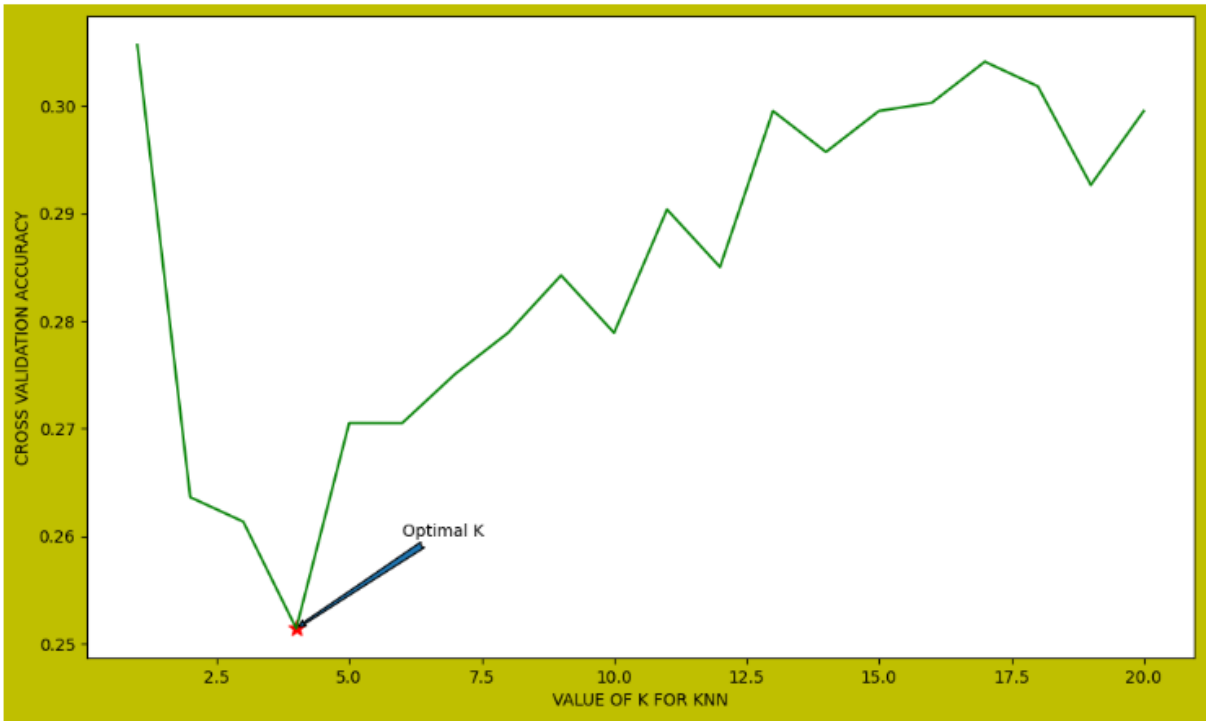

Figure 3: Feature comparison

3.4) The type of features being used can affect some distance measures because feature data have different sizes and in turn, distance measures are dependent on the features used. The distance between two samples may be greatly influenced by features with huge scales, which could change the chosen distance metric.

The performance can be measured using a variety of distance metrics, including establishing a range of distinctive distance metrics, developing a loop where it will build a model for each metric, using cross-validation to find misclassification errors, finding the least error computed and holding its index to be able to handle changes in distance metrics alone by fixing the value of K, and finally evaluating the performance using these metrics, and finally using the index to handle changes in distance metrics alone.

3.5) Comparing the two reveals that Logistic Regression is a parametric model, but KNN is a non-parametric model. KNN is slower than Logistic Regression. KNN also permits non-linear solutions

while Logistic Regression only permits linear ones. Additionally, while Logistic Regression can calculate the confidence level, KNN can only output the labels (about its prediction). [7]. The accuracy for the KNN model using cross-validation is 0.7485789782736347 and for the Logistic Regression, it is 0.7532413388138579. This suggests that since the Logistic model has greater accuracy than the KNN model, it is the best option for entering the Kaggle competition.

QUESTION 4: REGRESSION – WINE QUALITY

4.1) The following chart was created to determine the means or averages of each attribute for the red and white wines separately and compare them using a bar graph that shows both wines at once.

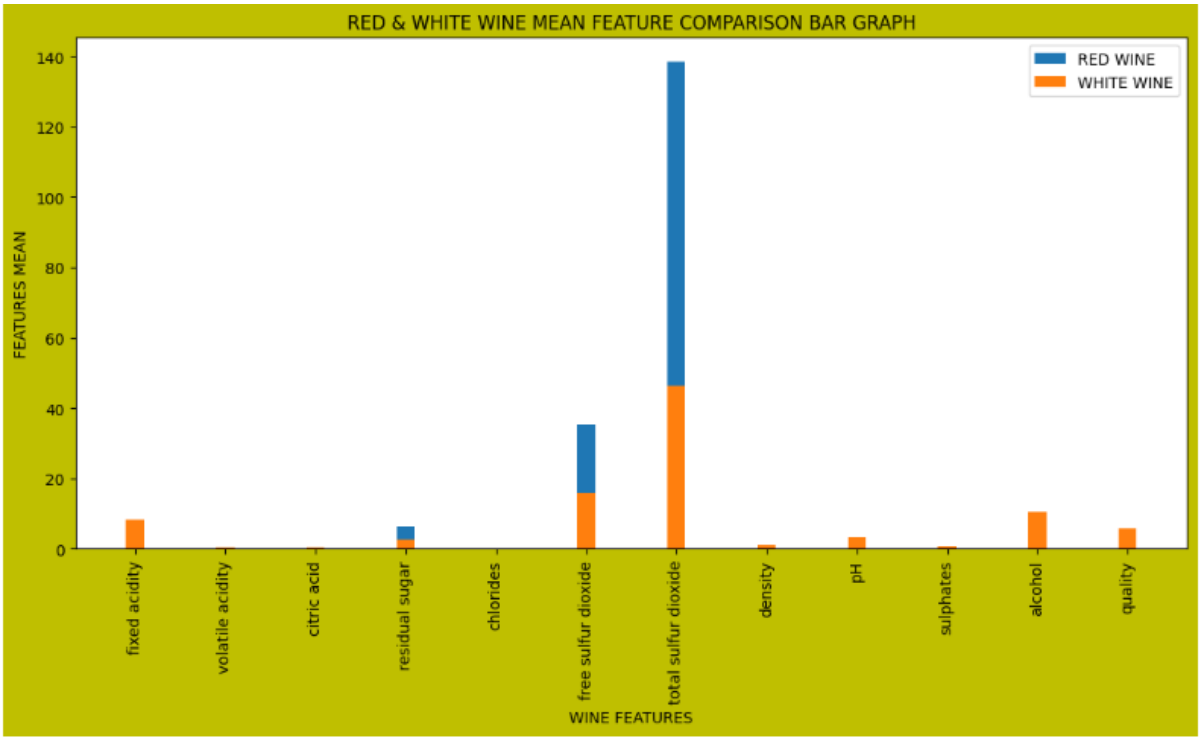

Figure 4: Feature comparison

The results show that the wine is concentrated into more sulfur dioxide and fewer acids in both white and red wine. This indicates that the quality of the red and white wine depends more on the sulfur dioxide, alcohol, fixed acidity, and sugars and at a lower quantity of PH, density, sulphates, density, chlorides, citric acids, and volatile acidity.

4.2) The correlation between each feature and the dependent variable using a separate analysis for white and red wine is calculated and the results are depicted below.

```
The correlation for red wine is:

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
Name: quality, dtype: float64

The relevant feature is alcohol with correlation of 0.4355747154613733
```

Figure 5: Correlation for red wine

```
The correlation for white wine is:

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
Name: quality, dtype: float64

The relevant feature is alcohol with correlation of 0.47616632400113656
```

Figure 6: Correlation for white wine

This clearly shows that the relevant feature for both wines is alcohol with a correlation coefficient of 0.476166 for white wine and 0.435575 for red wine.

4.3) The mean squared errors and parameter estimates are produced and plotted against lambdas as shown below using Lasso and cross-validation to provide a plot of MSE against lambda and the parameter estimates versus lambda.

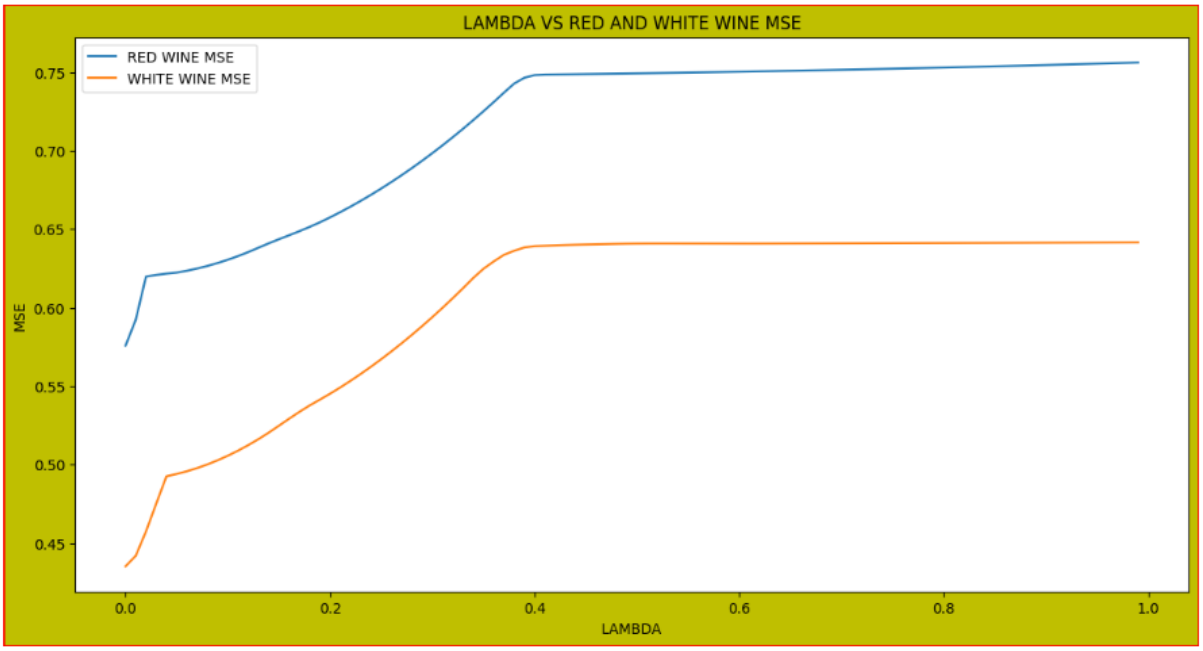

Figure 7: Mean squared errors vs lambda

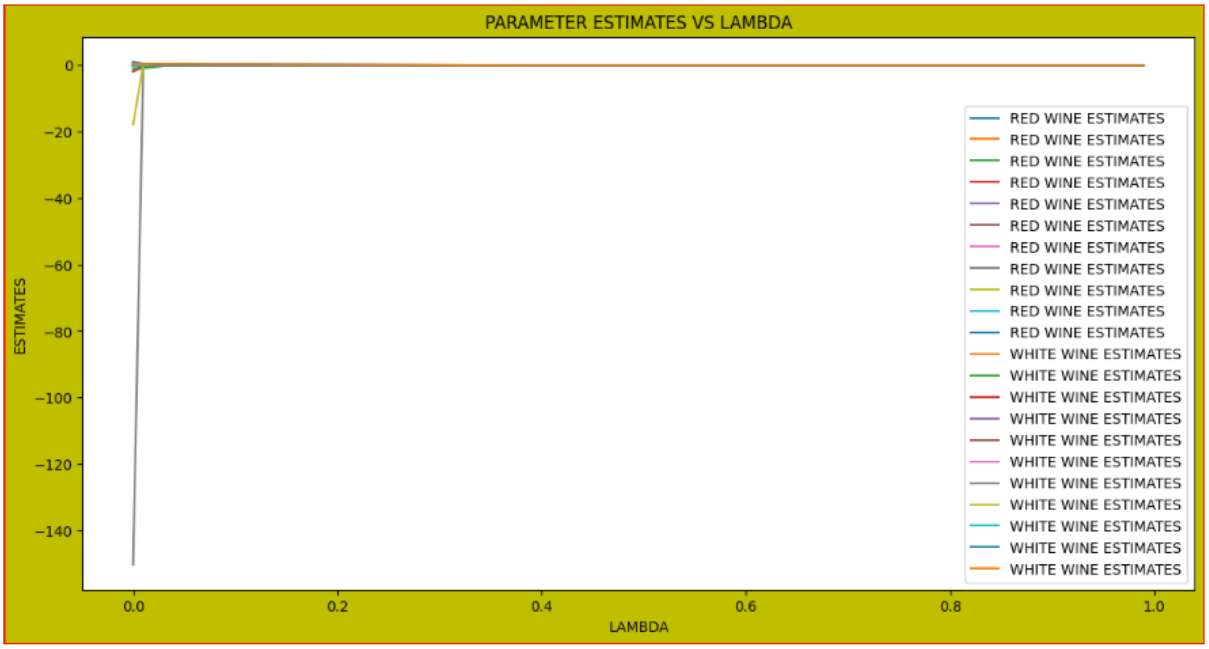

Figure 8: Parameter estimates vs lambda

The number of features selected by LASSO for the red wine is 6 namely, ['fixed acidity', 'volatile acidity, 'residual sugar', 'free sulfur dioxide', 'total sulfur dioxide', 'alcohol'].

fixed acidity	-0.051962
volatile acidity	-1.303798
citric acid	0.000000
residual sugar	0.022013
chlorides	-0.000000
free sulfur dioxide	0.005694
total sulfur dioxide	-0.001104
density	-0.000000
pH	0.000000
sulphates	0.000000
alcohol	0.355502

Figure 9: Selected features

4.4) Using the features identified by LASSO, a KNN regression model for the red wine is constructed.

4.5) With an MSE of -0.580998982395001 and the best R squared of 0.2683552222395802, the linear regression model is determined to have fewer errors than the KNN model, which has an MSE of -1.0226246817745501 and an R squared of 0.20619419334372602. Thus, the linear model is the best compared to KNN.

The linear model has many benefits, including being simpler to use, comprehend, and train. For linearly separable data, it performs remarkably well, and. Its drawbacks include the fact that it frequently suffers from noise and overfitting as well as the assumption of linearity between dependent and independent variables.

On the other hand, KNN has benefits such as an algorithm that is particularly user-friendly, memory-based, and adaptable to various estimations of proximity. Its drawbacks include occasional inefficiency and difficulty choosing the appropriate k value.

REFERENCES

[1] statisticshowto, "Nonlinear Regression: Simple Definition & Examples," [Online]. Available: https://www.statisticshowto.com/nonlinear-regression/.

[2] C. Validated, "Linear and nonlinear model parsimony," [Online]. Available: https://stats.stackexchange.com/questions/181298/linear-and-nonlinear-model-parsimony. [Accessed 13 November 2022].

[3] S. B. Jim, "How to Choose Between Linear and Nonlinear Regression," [Online]. Available: https://statisticsbyjim.com/regression/choose-linear-nonlinear-regression/. [Accessed 13 November 2022].

[4] indeed.com, "Linear vs. Nonlinear Equations: Definitions and Examples," [Online]. Available: https://www.indeed.com/career-advice/career-development/linear-vs-nonlinear. [Accessed 13 November 2022].

[5] educba.com, "Decision Tree Advantages and Disadvantages | Decision Tree Regressor," [Online]. Available: https://www.educba.com/decision-tree-advantages-and-disadvantages/. [Accessed 17 November 2022].

[6] openclassrooms.com, "Build and Evaluate a Classification Model," [Online]. Available: https://openclassrooms.com/en/courses/6389626-train-a-supervised-machine-learning-model/6405911-build-and-evaluate-a-classification-model. [Accessed 17 November 2022].

[7] towardsdatascience.com, "Comparative Study on Classic Machine learning Algorithms," [Online]. Available: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222. [Accessed 18 November 2022].

[8] Wikipedia, "Rule-based machine learning," 14 July 2021. [Online]. Available: https://en.wikipedia.org/wiki/Rule-based_machine_learning. [Accessed 5 November 2022].

[9] HackerNoon, "7 Effective Ways to Deal With a Small Dataset," [Online]. Available: https://hackernoon.com/7-effective-ways-to-deal-with-a-small-dataset-2gyl407s. [Accessed 5 November 2022].

[10] O. i. M. L. W. I. I. a. H. t. P. It, "EliteDataScience," [Online]. Available: https://elitedatascience.com/overfitting-in-machine-learning#how-to-prevent. [Accessed 7 November 2022].

[11] net-informations, "Squared: Coefficient of Determination," [Online]. Available: http://net-informations.com/ds/psa/r-squared.htm#:~:text=The%20R%C2%B2%20is%20calculated%20by,then%20subtract%20it%20from%201..

[12] towardsdatascience, "8 Metrics to Measure Classification Performance," [Online]. Available: https://towardsdatascience.com/8-metrics-to-measure-classification-performance-984d9d7fd7aa.

[13] GitHub, "bio304-class/bio304-book: Bookdown based course notes for Bio 304 at Duke University, taught by Paul Magwene," [Online]. Available: https://github.com/bio304-class/bio304-book. [Accessed 7 November 2022].

[14] MobiDev, "TOP 9 Machine Learning Technology Trends To Impact Business in 2022," [Online]. Available: https://mobidev.biz/blog/future-machine-learning-trends-impact-business. [Accessed 7 November 2022].

[15] T. AI, "The Evolution of Machine Learning in Business," [Online]. Available: https://www.turintech.ai/the-evolution-of-machine-learning-in-business/. [Accessed 7 November 2022].

[16] Medium, "Machine Learning Approaches and Its Applications," [Online]. Available: https://medium.datadriveninvestor.com/machine-learning-approaches-and-its-applications-7bfbe782f4a8. [Accessed 7 November 2022].

[17] u. blog, "Random Forest Vs Decision Tree: Difference Between Random Forest and Decision Tree," [Online]. Available: https://www.upgrad.com/blog/random-forest-vs-decision-tree/#:~:text=A%20decision%20tree%20combines%20some,forest%20model%20needs%20rigorous%20training. [Accessed 7 November 2022].

[18] Javatpoint, "Applications of Machine learning," [Online]. Available: https://www.javatpoint.com/applications-of-machine-learning. [Accessed 7 November 2022].

[19] britannica, "collinearity," [Online]. Available: https://www.britannica.com/topic/collinearity-statistics. [Accessed 7 November 2022].

[20] datacadamia, "Statistics - Forward and Backward Stepwis," [Online]. Available: https://datacadamia.com/data_mining/stepwise_regression. [Accessed 7 November 2022].