**T.R.**
**SAKARYA UNIVERSITY**
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# A HYBRID PREDICTION APPROACH USING MULTIPLE LINEAR REGRESSION AND DECISION TREE

**MSc THESIS**

**Maryam Arif AZEEZ**

**Computer and Information Engineering Department**

**Computer Engineering Program**

**AUGUST 2023**

**T.R.**
**SAKARYA UNIVERSITY**
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# A HYBRID PREDICTION APPROACH USING MULTIPLE LINEAR REGRESSION AND DECISION TREE

**MSc THESIS**

**Maryam Arif AZEEZ**

**Computer and Information Engineering Department**

**Computer Engineering Program**

**Thesis Advisor: Dr.Öğr. Üyesi Kayhan AYAR**

**AUGUST 2023**

The thesis work titled “A HYBRID APPROACH USING MULTIPLE LINEAR REGRESSION AND DECISION TREE” prepared by Maryam Arif AZEEZ was accepted by the following jury on …./…. /……. by unanimously/majority of votes as a MSc /PhD THESIS in Sakarya University Graduate School of Natural and Applied Sciences, Computer and Information Engineering department, Computer Engineering programe.

**Thesis Jury**

**Head of Jury :** **Kayhan AYAR**
**Sakarya University**

**Jury Member :** **Sümeyye KAYNAK**
**Sakarya University**

**Jury Member :** **Muhammed Ali Nur ÖZ**
**Sakarya University of Applied Sciences**

**STATEMENT OF COMPLIANCE WITH THE ETHICAL PRINCIPLES AND RULES**

I declare that the thesis work titled " A HYBRID APPROACH USING MULTIPLE LINEARREGRESSION AND DECISION TREE", which I have prepared in accordance with Sakarya University Graduate School of Natural and Applied Sciences regulations and Higher Education Institutions Scientific Research and Publication Ethics Directive, belongs to me, is an original work, I have acted in accordance with the regulations and directives mentioned above at all stages of my study, I did not get the innovations and results contained in the thesis from anywhere else, I duly cited the references for the works I used in my thesis, I did not submit this thesis to another scientific committee for academic purposes and to obtain a title, in accordance with the articles 9/2 and 22/2 of the Sakarya University Graduate Education and Training Regulation published in the Official Gazette dated 20.04.2016, a report was received in accordance with the criteria determined by the graduate school using the plagiarism software program to which Sakarya University is a subscriber, I accept all kinds of legal responsibility that may arise in case of a situation contrary to this statement.

(……/……/20…..)

signature

Maryam AZEEZ

*To Radhwan and Kenan my little family*

**ACKNOWLEDGEMENT**

During my study to obtain a master's degree, I would like to thank everyone who contributed to my success in completing my studies, which I consider one of the most important achievements I have made for myself.

Thank you to my professors and supervisors, Dr. Mustafa Akpinar and Dr.Kayhan Ayar. thanks to my colleagues in the study, we were a really helpful and supportive team.

And a special thanks to my husband, Radhwan Abdul-Razzaq Khaleel, who was the best companion and helper for me and still is. I would also like to thank my family and my husband's family.

Maryam AZEEZ

**TABLE OF CONTENTS**

**Sayfa**

## ABBREVIATIONS

**DTs** : Decision Trees

**DT** : Decision Tree

**ML** : Machine Learning

**CART** : Classification And Regression Tree

**BFOS** : Leo Breiman, Jerome Friedman, Richard Olshen, And Charles Stone

**MSE** : Mean Squared Error

**$R^2$** : Coefficient Of Determination

**SSE** : Sum Of The Squared Errors

**MLR** : Multiple Linear Regression

**MAPE** : Mean Absolute Percent Error

**RMSE** : Root Mean Square Error

**MRA** : Multiple Regression Analysis

**PCA** : Principal Components Analysis

**TARGET** : Tree Analysis With Randomly Generated And Evolved Trees

**SST** : Sum Of Squared Totals

**SSR** : Sum Of Squared Residuals

# LIST OF TABLES

**Page**

# LIST OF FIGURES

**Page**

## A HYBRID PREDICTION APPROACH USING MULTIPLE LINEAR REGRESSION AND DECISION TREE

## SUMMARY

When you wake up one winter morning, you may wonder whether it will rain or will the weather be fine? In our life we fall into many choices that require prediction and anticipation of the answer before starting work.

In this thesis, a hybrid method was used between decision tree (regression tree) and multiple linear regression based on the CART mechanism. It used three different datasets to test the approach. The first is the advertising data set, which was represented by using (TV, radio, and newspapers) (X) to show the relationship between these advertising methods with sales (Y) in terms of their impact on sales and purchasing power. This dataset is called as "Advertising". The second data set contains (Species of fish, length, height, width), which are the independent variables (X) and their impact on the weight of the fish, which represents the dependent variable. This dataset is called as "Fish". The third dataset is the effect of the car's specifications on its price, which was considered the dependent variable. The car specification was (car name, fuel type, aspiration, door number, car body, drivewheel, engine location, wheelbase, car length, car width, car height, curb weight, engine type, cylinder number, engine size, fuel system, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg). This dataset is called "Car". The datasets were divided into train and test 80% - 20%, respectively.

Where the research steps that represent the study were implemented, by making accurate predictions with the help of linear regression and CART. First, we split datasets using CART. For each leaf, different sub-datasets are filtered and created. The splitting point in the dataset was found with nodes. Our hypothesis is to divide the dataset using CART to increase the accuracy of the estimates. It applied multiple linear regression to filtered datasets. Then, it is compared multiple linear regression estimations using whole data and splitting dataset.

The classification and regression tree (CART) algorithm represents a dataset's connection between the dependent variable and independent factors. It consists of a sequential binary dataset partition based on the variable values. Fitting tree models involves repeatedly splitting the data into homogenous groups. The output is a hierarchical tree of relevant decision rules for classification or prediction.

Splitting is a procedure that divides the tree from its nodes into two or more nodes. The root node represents the entire sample or population and is divided into two or more groups as homogeneous groups. The nodes that sub-nodes are separated into are called parent and child nodes. Nodes that cannot be divided and have reached the minimum division are called leaf nodes. Pruning is the opposite of splitting, removing child nodes from the root node.

In this study, results were compared to predict the value of the dependent variable (Y) using the regression tree method, multiple linear regression, and the particular research method of splitting the regression tree and constructing multiple linear regression models from it in order to select the best method that gives the best prediction based on the $R^2$, MSE, and MAPE values.

It was found in this study that splitting the data using multiple linear regression based on the regression tree gave a good result compared to using the multiple linear regression method alone or using the regression tree only. It was also found that the use of one error measure is not sufficient, but more than one error measure must be added to obtain an optimal model.However, it can add a classification and regression tree to divide the data set and find the best result from the hybrid tree and multiple linear regression model. The depth of the tree in an extensive real-life dataset will be increased to see the effect of height. Furthermore, we will delve into alternative approaches to linear regression in a distinct study. It could be scalable and effective in increasing tree size and powerful machine learning techniques.

It was found in this study that splitting the data using multiple linear regression based on the regression tree gave a good result compared to using the multiple linear regression method alone or using the regression tree only. It was also found that the use of one error measure is not sufficient, but more than one error measure must be added to obtain an optimal model.

# ÇOKLU DOĞRUSAL REGRESYON VE KARAR AĞACI KULLANARAK HİBRİT TAHMİN YAKLAŞIM

## ÖZET

Bir kış sabahı uyandığınızda, yağmur yağacak mı yoksa hava güzel mi olacak diye merak edebilirsiniz. Hayatımızda, işe başlamadan önce cevabı tahmin etmeyi ve tahmin etmeyi gerektiren birçok seçeneğe düşeriz.

Bu tezde, bir karar ağacı (regresyon ağacı) ile CART mekanizmasına dayalı çoklu doğrusal regresyon arasında hibrit bir yöntem kullanılmıştır. Yaklaşımı test etmek için üç farklı veri seti kullandı. Birincisi, bu reklam yöntemlerinin satışlarla (Y) ilişkisini satış ve satın alma gücü üzerindeki etkisi açısından (TV, radyo ve gazeteler) (X) kullanılarak temsil edilen reklam veri setidir. Bu veri kümesine "Reklam" adı verilir. İkinci veri seti, bağımsız değişkenler (X) olan (Balık türleri, uzunluk, yükseklik, genişlik) ve bunların bağımlı değişkeni temsil eden balığın ağırlığı üzerindeki etkilerini içerir. Bu veri kümesine "Balık" adı verilir. Üçüncü veri seti, arabanın teknik özelliklerinin bağımlı değişken olarak kabul edilen fiyatı üzerindeki etkisidir. Araç özellikleri araba adı, yakıt tipi, çekiş, kapı numarası, araba gövdesi, tahrik tekerleği, motor konumu, dingil mesafesi, araba uzunluğu, araba genişliği, araba yüksekliği, boş ağırlık, motor tipi, silindir numarası, motor boyutu, yakıt sistemi, delik oranı, sıkıştırma oranı, beygir gücü, en yüksek devir sayısı, şehir içi mpg ve otoyol mpg'sidir. Bu veri kümesine "Araba" denir. Veri kümeleri, sırasıyla %80 - %20 tren ve test olarak bölünmüştür.

Lineer regresyon ve CART yardımıyla doğru taminler yapılarak araştırma adımları gerçekleştirilmiştir. İlk olarak, CART kullanarak veri kümelerini ayırdık. Her yaprak için farklı alt veri kümeleri filtrelenir ve oluşturulur. Veri setindeki ayrılma noktası düğümler ile bulundu. Hipotezimiz, tahminlerin doğruluğunu artırmak için veri setini CART kullanarak bölmektir. Filtrelenmiş veri kümelerine çoklu doğrusal regresyon uyguladı. Daha sonra, tüm veri ve bölünmüş veri seti kullanılarak çoklu doğrusal regresyon tahminleri karşılaştırılır.

Çoklu doğrusal regresyon (MLR) modelleri, katsayıları basit modellere benzer şekilde tahmin eder. Basit doğrusal regresyonda olduğu gibi, çoklu doğrusal regresyondaki en küçük kareler tahmin edicileri tarafsızdır. Ayrıca yansız tahminciler en küçük varyasyona sahiptir ve tutarlıdır. Bu nedenle, regresyon varsayımları doğru kalırsa, kullanıcılar en küçük kareler tahmincilerini kullanarak çoklu doğrusal regresyon katsayılarını güvenle çıkarabilirler. Regresyon çizgisi, değişken yanıt noktalarına yaklaşır. Nokta tahminini çevreleyen değişkenlik, çıkarım varsayımlarını doğrulamada, sorunlu gözlemleri belirlemede ve güven veya tahmin aralıkları yaratmada da yardımcı olur.

Çoklu doğrusal regresyon, bağımsız ve bağımlı değişkenler arasındaki doğrusal ilişkiyi modellemek için çalışır. Bu bağımsız değişkenler sürekli veya kesikli olabilir. Çoklu doğrusal regresyon, basit doğrusal regresyonu birden fazla açıklayıcı değişken içerecek şekilde genişletir. Bu, bir bağımlı değişken (Y) ve birden fazla bağımsız

değişken (Xi) olduğu anlamına gelir. Yanıt değişkeni, açıklayıcı değişkenlerin doğrusal bir kombinasyonu ile doğrudan ilişkili olduğundan, her iki senaryoda da "doğrusal" terimi kullanılmaktadır.

Bölme, ağacı düğümlerinden iki veya daha fazla düğüme ayıran bir prosedürdür. Kök düğüm, tüm örneklemi veya popülasyonu temsil eder ve homojen gruplar olarak iki veya daha fazla gruba bölünür. Alt düğümlerin ayrıldığı düğümlere ebeveyn ve alt düğümler denir. Bölünemeyen ve minimum bölünmeye ulaşmış düğümlere yaprak düğümler denir. Budama, bölmenin tersidir, alt düğümleri kök düğümden çıkarır.

Eğitim veri setleri kullanılarak sınıflandırma ve regresyon ağaçları oluşturulmuştur. İnşa edilen ağaçların derinliği üçtü. Daha sonra her bir derinlik seviyesi ve düğüm noktası için farklı MLR denklemleri oluşturulmuştur. Yaklaşımımızı test etmek için maksimum derinlik üç olarak seçildi. Bir sonraki adımda, ağaç oluşturulduktan sonra, her düğüm veri kümesi için bir filtre olarak kullanıldı. Her veri seti için toplam MLR modelleri 15 farklı filtre kullanılarak oluşturulmuştur. Sonraki bölümde, üç veri seti değerlendirildi ve geleneksel CART ve MLR modelleriyle karşılaştırıldı.

Bu çalışmada sonuçlar, bağımlı değişkenin (Y) değerini tahmin etmek için regresyon ağacı yöntemi, çoklu doğrusal regresyon ve regresyon ağacını bölme ve ondan çoklu doğrusal regresyon modelleri oluşturma özel araştırma yöntemi kullanılarak karşılaştırıldı. R2, MSE ve MAPE değerlerine dayalı olarak en iyi tahmini veren en iyi yöntem olarak tespit edildi.

Karar verme ve olayları tahmin etme hayatımızın ayrılmaz bir parçasıdır. Regresyon ağaçları, kararlarımızı düzenlemenin yaygın yollarından biridir ve makine öğrenimi yöntemlerinden biri olarak kabul edilir. Çoklu doğrusal regresyon modeli algoritması da tahminde önemli bir yöntemdir. Bu araştırmada, regresyon ağacından çoklu doğrusal regresyon modelleri oluşturmak için regresyon ağacı algoritması ile çoklu doğrusal regresyon algoritmasını birleştiren bir süreç önerilmiştir. Üç tür veriye uygulanmıştır. Birinci veri setinde üçüncü seviyeye (L3) bağlı olarak hatayı azaltarak en iyi sonucu verdiği görülmüştür. İkinci veri setinden farklı olarak trende en ufak hatayı üçüncü seviyede (L3) verirken, testte üçüncü seviyede (L3) en iyi sonucu vermeye yetmedi. Bu durumda optimal seviye L1 olarak bulunmuştur. Bu, çalışılan soruna bağlıdır. Ayrıca, en iyi sonucu seçmek için tek bir ölçüme güvenmek imkansızdır. Bununla birlikte, en iyi sonucu, yani en iyi modeli elde etmek için başka hata ölçüleri eklenmelidir. Üçüncü veri setinin birinci seviyede (L1) en iyi sonucu verdiği, ancak üçüncü seviyenin (L3) de en iyi sonucu verdiği için tek olmadığı not edilebilir. Veri setini bölmenin uygun bir yöntem olduğu söylenebilir çünkü en azından üç veri seti olan L1, L2 ve L3'te L0'a veya tam ağaca kıyasla en iyi sonuçları vermiştir. Ancak çoklu doğrusal regresyon modelini tek başına kullanmak en iyi sonucu vermez. Ancak, veri setini bölmek ve hibrit ağaç ve çoklu doğrusal regresyon modelinden en iyi sonucu bulmak için bir sınıflandırma ve regresyon ağacı ekleyebilir.

Bu çalışmada, regresyon ağacına dayalı çoklu doğrusal regresyon yöntemi kullanılarak verilerin bölünmesinin, tek başına çoklu doğrusal regresyon yöntemi veya yalnızca regresyon ağacı kullanılmasına göre iyi bir sonuç verdiği görülmüştür. Ayrıca, bir hata ölçüsü kullanımının yeterli olmadığı, optimal bir model elde etmek için birden fazla hata ölçüsünün eklenmesi gerektiği görülmüştür. Ancak, veri setini bölmek ve en iyi sonucu bulmak için bir sınıflandırma ve regresyon ağacı ekleyebilir. hibrit ağaçtan ve çoklu doğrusal regresyon modelinden. Kapsamlı bir gerçek yaşam veri kümesindeki ağacın derinliği, yüksekliğin etkisini görmek için artırılacaktır. Ayrıca, ayrı bir

çalışmada doğrusal regresyona alternatif yaklaşımları inceleyeceğiz. Ağaç boyutunu ve güçlü makine öğrenimi tekniklerini artırmada ölçeklenebilir ve etkili olabilir.

Doğrusal olmayan verilerin analizi söz konusu olduğunda, karar vericiler genellikle hataları azaltma zorluğuyla karşı karşıya kalır. Neyse ki, regresyon ve CART'ı birleştiren hibrit bir yaklaşımın bunu başarmak için etkili bir yöntem olduğu kanıtlanmıştır. Doğrusal regresyonun ilişkisel tahminini ve CART'ın gruplandırmasını kullanan bu yaklaşım, büyük miktarda veriyi yönetmek ve veriye dayalı kararların doğruluğunu artırmak için güçlü bir araç sunar.

Regresyon ağacı, kararı karmaşık ve basit hale getirmek için kullanılan istatistiksel araçlardan biridir. Regresyon ağaçları, ayrık değerler yerine sürekli değerlerle çıktıyı tahmin eder. Tepki değişkenine bağlı olarak, yordayıcıların veya ortak değişkenlerin bir vektörüne bakmak. Regresyon ağacı, basit ve çoklu doğrusal regresyon gibi parametrelere ihtiyaç duymaması bakımından diğer geleneksel regresyon yöntemlerinden farklıdır. Güçlü değişken alt bölümleri üzerinde çalışmasında fark yaratır, aykırı değerlerden etkilenmez ve farklı veri türleri üzerinde uygulanabilir.

## 1. INTRODUCTION

In our life, we are constantly faced with several options, our world offers many options, but we have to choose one course of action from among the possibilities. But to choose the correct option, a process of thinking takes place to select a logical option from several options or alternatives. This process is called decision-making. For a correct decision to be made, it is necessary to know the advantages and disadvantages of that decision and to analyze the alternatives that could be reached. Decision-making is a task facing humanity. Therefore, many algorithms were created that perform the decision-making function more accurately, considering all the relevant features without missing a single point. The decision trees (DTs) are a critical method for decision-making.

The decision tree (DT) is one of the most widely used tools for decision-making. It is a drawing of a decision tree with branches and distinct leaves at the end of it. Similar to a decision support tool, it formally uses a tree-like statement to represent decisions and their potential outcomes, including resource costs, event outcomes, and utility [1].

One of the well-known machine learning (ML) models is the decision trees that were studied in 1984, where the DTs are seen as an interpretable machine learning model. Motivated by this view, considerable effort has been expended on learning DTs (and similar logic machine learning models) with interpretability-critical parameters, such as the number of nodes, maximum/average depth, etc. Additionally, work has been conducted on distilling or approximating sophisticated machine learning models using soft decision trees. Nonetheless, current research indicates that interpretability should be related to the depth of DTs [2].

DTs are resistant to mistakes and noise in the actual world data, are simple to display, and are regarded as a straightforward and robust way for forecasting an example category label. Decision trees are useful in many domains, including image processing, health, and finance [3].

The decision tree uses the iteration technique to extract data and develop classification systems based on multiple common variables or predictive algorithms for a target

variable. It consists of a community divided into sections like branches to form an inverted tree with a root, internal, and leaf nodes. The nonparametric algorithm efficiently handles a complex data set without imposing a complex borderline structure. The study data can be divided into training sets and validated if the sample size is large enough. The training data is used to create a decision tree model and a validation data set to determine the size of the optimum tree for the final model. These examples are put into groups based on the value of an attribute that is checked at a certain node and sent down a branch that is linked to that value. Following these branching paths is a group of leaf nodes or examples that all belong to the same class. So, when a new instance is sent down the root of the tree, it will be checked by each attribute as it moves down the branches that are relevant to it. At the end of a branch route in the decision tree, there is a leaf node that has related examples that all share a classification that is used to identify the new example [3,4].

A classification o regression tree is a kind of decision tree that can be used to depict a prediction model for DTs development. There are many algorithms like (C4.5, CART, CRUISE, GUIDE, and QUEST). CART (Classification and Regression Tree) will be taken as the algorithm in this thesis, which will be implemented on the regression tree that is the subject of the research.

Regression trees are a versatile statistical instrument for modeling the conditional distribution of a response variable given a vector of predictors or covariables. Trees provide a number of benefits over typical parametric regression methods: they are nonparametric, have a robust variable subset selection, are resistant to outliers in the covariate space, and may be used effectively to a wide range of data sources.

The regression tree is considered a piecewise constant (A regression tree is piecewise constant because in a regression tree, a constant piecewise function is defined in the input space at each tree [5] or linear estimate of the regression function, formed by recursively partitioning the data and sample space. Many alternative flexible regression (The regression tree is a flexible model because any small change in the data leads to a large change in the model [49]) approaches exist, for example: (i) locally smoothed nonparametric regression models; (ii) piecewise regression splines; (iii) projection pursuit; and (iv) neural networks [16].

When constructing a regression tree, the midpoints found from the prediction variables represent the possible splits of the tree. Initially, the sum of the squares of the differences between the observations and their average are calculated. The result that gives the minimum value represents a node. From this node the observations are divided into two subgroups. The previous process is repeated on each subset. The split continues until one or more stopping criteria is reached [6].

Although regression trees were primarily designed for large datasets, they may also be utilized economically with tiny datasets, such as those from repeated or unreplicated complete factorial experiments. In such cases, regression tree models can give better and more evident interpretations of interaction effects as variations between conditional main effects. Some study demonstrates through simulations that the models can provide less prediction mean squared errors than prior methods. From piecewise constant to piecewise simple and multiple linear, and from least squares to Poisson and logistic regression, the tree models cover a broad range of sophistication [7].

Regression is used to predict a result based on a certain input. The simplest regression technique is known as linear regression, while the most complex is known as multiple regression. It is contingent on the variable. If a single descriptive variable is employed, the approach is known as simple linear regression, and if many descriptive variables are employed, it is known as multiple regression. MLR predicts the future value of a variable ($\hat{Y}$ ) in relation to additional variables ($X_i$). And multiple regression analysis (MRA) is a technique that correlates the behavior or variance of independent variables that reflect elements influencing the objective with one variable representing the dependent variable [12,13].

According to research, multivariate analytic techniques provide the simultaneous examination of several variables from persons or items under investigation, hence assisting users with decision-making. Multiple linear regression models that assess one response variable for two or more predictor variables can be used to model multivariate data. In light of the link between predictor variables and response variables (dependent), the linear regression approach enables the development of mathematical models for predicting response variables (dependent) (independent or explanatory). In environmental sciences, health, and other areas of life, multiple linear regression

models are widely used to evaluate the statistical significance of the relationship between predictor and response variables [14].

Leo Berriman, Jerome Friedman, Richard Olschen, and Charles Stone (BFOS) contributed to the advancement of artificial intelligence, machine learning, nonparametric statistics, and data mining in 1984 with the formation of classification and regression trees. Examination of decision trees has expanded the scope of work and the technological improvements it provides, the complex examples of analyzing data structured by the trees it contains, and the authoritative discussion of big sample theory for the trees it contains. According to the Scientific and Social Indexes citation, the CART has been used 3000 times since its publication, The scholar from Google also lists approximately 8,450 quotes. CART has entered into a number of areas and uses, including credit risk, target marketing, financial market modeling, electrical engineering, quality control, biology, chemistry and clinical medical research. CART also had a significant impact on image reduction through the use of tree-structured vector quantization [8].

CART models are alternatives to conventional statistical techniques like logistic regression, discriminant analysis, and regression analysis. These tree models are gaining popularity due to their interpretability and application ease [10]. And there are two crucial statistical concerns. Each predicts a response variable y (the dependent variable) given the values of a vector of predictor variables x. (the independent variable). Consider X to be the domain of x, and Y to be the domain of y. When y is a continuous or discrete variable with real values (such as the weight of a car or the frequency of accidents), the resulting issue is known as regression. Alternatively, if Y is a finite set of unordered values (such as the type of car or the country of origin), the job is known as classification [11]. CART is a segmentation modeling approach using characteristics in which the hierarchy is referred to as a tree and each part is referred to as a node, with the root node containing the whole database. The root node is successively split into child nodes. When no more data subdivision is possible, the last subgroups are referred to as terminal nodes or leaves [15].

Machine learning approaches for creating prediction models from data include regression and classification trees. Models are created by recursively partitioning the data space and fitting a basic prediction model to each division. Consequently, partitioning may be represented graphically using a decision tree. Classification trees

are utilized to describe dependent variables with a finite number of unordered values, and prediction error is defined by misclassification cost. For dependent variables that take continuous or ordered discrete values, regression trees are utilized, with prediction error often measured as the squared difference between actual and predicted values [9].

In this thesis, will be talking about the Regression Tree, and will be taking multiple linear regression the primary purpose of the thesis is to demonstrate a hybrid use of decision trees with multiple linear regression analysis in the concept of the CART mechanism, which is a statistical technique used to predict ordered/continuous and unordered/categorical variables. Its algorithm creates simple binary nodes from the root node, and finally, one gets the structure of a binary tree and uses it to classify the objects concerning their classes. The tree is reduced based on either standard deviation, variance, the sum of error squares... etc. But the thesis will take mean squared error (MSE) as an approach to splitting a regression tree, which facilitates the prediction of the optimal variables or alternatives, which represents the best decision for the case under study. The R-square (coefficient of determination) and MSE and Mean Absolute Percentage Error (MAPE) will be relied upon in order to determine the best algorithm to follow (the regression tree algorithm, the multiple linear regression model algorithm and the multiple linear regression models generated from the regression tree splitting process) in order to predict the value of the dependent variable Y.

### 1.1. Literature Review

A regression tree is basically a decision tree that is utilized for the task of regression which can be used to predict continuous valued outputs instead of discrete outputs [17]. It is also regarded one of the machine learning models because to its incorporation into a variety of industrial, medical, and life sciences domains, as well as several topics and areas of study in the present and preceding eras. The ML method is a rapidly developing area of predictive modeling that focuses on identifying structure in complicated, sometimes nonlinear data and building accurate prediction models. As ML approaches are not limited to the traditional assumptions (about data characteristics) commonly used with conventional and parametric approaches, ML approaches typically exhibit greater power for resolving complex relationships (i.e., nonlinear, nonmonotonic, multimodal relationships typical of landscape and

ecological applications). Traditional modelling strategies are typically based on more stringent statistical assumptions and data requirements, and they generally employ linear or additive modelling approaches that are inconsistent with the natural processes that occur in the landscape [18]. Enhanced regression tree analysis was utilized to determine the adjustment levels of MCS due to its excellent accuracy in the investigation of underwater sensor networks; its classification accuracy for MCS levels is 99.97% [19]. The CART model is an example of ML techniques that offer alternatives to conventional prediction methods. CART, and RF (random forest) paired with a GIS to predict groundwater spring sites have been effectively deployed in landslide susceptibility and hazard mapping in areas where CART has been implemented [18].

Regression trees have properties that make them desirable to use for data extraction or analysis of complex datasets in general, as they can handle numeric, ordinal, binary and categorical variables with the same ease. The tree handles in an elegant way with missing values. No need to use imputation schemes. Trees are immune to the effects of outliers among the x-variables of prediction, thus easing the burden of data cleaning. The predictive model is stable under strict monotonic transformations of the input variables; Performing a transformation on any variable xj → gj (xj), where gj (xj) is any monochromatic function, produces the same form. Thus, there is no problem in trying to find out the "good" transitions beforehand. If the lowest absolute deviation loss, one of the implemented loss functions, is used, there is also complete immunity to outliers in the output response y, thus providing complete immunity to the effects of outliers. Unlike kernel-based, near-neighbor-based methods, and supporting vector machines, trees are invariant in changing the relative scales of predictor variables, so there is no need to experiment with different scales. With all these useful properties, it is not surprising that obtaining a regression or classification tree has become a very popular tool for predictive learning in data mining. However, trees have one major drawback, which is inaccuracy. Although sometimes competitive, tree-based models usually do not achieve an accuracy close to best in any given application. Fortunately, remediation is made available through so-called boost, and enhancing tree-based models is always This greatly increases its accuracy [27].

Neural networks are very popular tools for data analysis; however, the output of a neural network is difficult to understand in terms of the original covariates or the input

variables. Therefore, regression trees were used as an easy way to understand the results of the neural network [28].

Regression tree analysis (RTA) is a nonparametric statistical method based on a tree diagram that has considerable advantages such as easy interpretation, assumptions of the distribution of the predictor variables are not required, being able to be applied using continuous dependent, nominal and ordinal variables, and not being affected by outliers. For the building of the regression tree, different data mining methods are utilized (CART, QUEST, CHAID, and exhaustive CHAID); nevertheless, prior research have indicated that predictive predictions utilizing the CHAID approach showed models with superior precision. Recently, studies employing the RTA as a tool to predict features of economic value in animal science, such as body weight, fleece weight, weaning weight, and milk output have been expanded; however, few studies have been done to analyze and forecast egg qualities. Therefore, a study was done to predict egg weight from the exterior parameters of guinea fowl eggs using multiple linear regression and statistical regression methods [20].

Based on a set of variables that define socioeconomic characteristics and land use, travel behavior was analyzed using a common application of classification, regression trees (CART), multiple linear regression and principal components analysis (PCA) for a sample of the urban population of San Paolo. It was intended to obtain numerical variables for the use of (PCA) and combining the original data set into a collection of variables and identifying correlations between the new variables and travel behavior [14].

Regression trees are a common alternative to conventional regression methods. Multiple methods exist for constructing regression trees. The majority of these procedures, including CART, are sequential and optimum at every node split. As a result, the final solution to the tree may not be the optimal answer in general; typically, modest changes in the training data lead to substantial changes in the final output as a result of relative instability. These are greedy algorithms for planting trees. This instability is intended to be exploited by cluster techniques such as random forests by constructing a forest of trees from the data and averaging their predictions. The predictive performance of these approaches is enhanced, but the one-tree method's simplicity is lost. The TARGET approach (Tree Analysis with Randomly Generated and Evolved Trees) Regression trees offer greater predictive accuracy than recursive

partitioning techniques like CART and single-tree stochastic search methods like Bayesian CART, according to experimental data. The prediction performance of TARGET is modestly inferior than ensemble techniques such as random forests, but TARGET solutions are considerably more interpretable [21].

Productivity forecasting is a rational and scientific strategy for predicting future agricultural events - the magnitude of production impacts. Its primary objective is to mitigate risk in the decision-making process that influences the return in terms of quantity and quality. The most prevalent models for return forecasting are regression models. This approach permits the examination of the link between independent factors and dependent variables. Multiple linear regression (MLR) is utilized to forecast productivity since its variance is influenced by a large number of independent factors. Plant growth, development, and production are the most often modeled processes using conventional methods. First, the economic significance of the outcomes obtained by such a model cannot be overstated. Second, the resulting models frequently serve as a foundation for the creation of agricultural engineering simulations [31].

In an effort to anticipate the highest daily surface ozone concentration over the following 24 hours in the Greater Athens Area (GAA), multiple linear regression (MLR) models were compared to an Artificial Neural Network (ANN) based prediction model. Where modeling is based on meteorological data and air pollution data collected from thirteen monitoring stations within the GAA (the network of the Greek Ministry of Environment, Energy, and Climate Change) between 2001 and 2005. The results indicate that the ANN model may be utilized to give alerts to the general public and vulnerable groups. Basically. In each of the aforementioned research efforts, the ANN model outperformed the MLR model [30].

The decision tree algorithms C4.5, ID3 and CART algorithm were used to classify hepatitis disease and the correction rate was compared and the effectiveness was compared among them. Therefore, the model derived from Kart along with the extended definition for determining (diagnosing) hepatitis disease provided a good model that depends on the accuracy of classification [22].

The CART algorithm was also introduced into the data mining study to examine the classification of blood donors, where the study method is to identify blood donation

behavior using data mining classification algorithms. The analysis was using a standard transfusion data set and using the CART decision tree algorithm implemented in Weka. The CART-derived model along with the extended definition to identify regular voluntary donors provided a model based on classification accuracy [23].

Since both continuous and discrete predictive variables can be integrated into the models and the outputs are simply understood, CART approaches have shown to be particularly useful in ecological and environmental problems. CART models have several additional advantages over other techniques because they are nonparametric and divide data sets into distinct groups: input data do not need to be normally distributed; predictor variables do not need to be independent; and nonlinear relationships between predictor variables and observed data can be modeled. Various CART enhancements are proposed to increase the robustness [24].

The CART method was used to construct an intrusion detection model, and when compared with the ID3 algorithm and the C4.5 algorithm, it was determined that the split-based CART algorithm is used to simplify the scale of the decision tree and achieve the aim of classification. When utilizing ID3 in choosing attributes, it is possible to accept the nested value of the property and ignore the values of other attributes. In addition, it cannot handle continuous numbers, and the tree branch is seen to be big. While the C4.5 algorithm when utilized in training more continuous attribute data, from the efficiency of the above approach is easy to be altered by continuous numerical estimating [25].

In the field of psychiatry, the CART algorithm has had a role in identifying the highest potential users of services among low-income psychiatric outpatients. It has been used successfully in other disciplines of medicine, for example, to predict the outcome of ischemic hypoxic coma and the risk of death in patients with Caring for the heart admitted to emergency care [26].

In order to better understand the factors that affect the roughness of asphalt pavement, the classification and regression tree analysis method (CART) was used to explore the relationship between pavement roughness and design characteristics, and it was considered as a tree structure that is easy to interpret and apply and provides simple accurate and reasonable results that are more applicable and practical to provide guidance in Pavement design and construction [29].

## 2. METHODOLOGY

This chapter will introduce multiple linear regression and the classification and regression tree (CART) algorithm and how they work. Furthermore, evaluation metrics are given in this chapter.

### 2.1. Multiple Linear Regression

Linear regression is a statistical technique used to describe the correlations between several inputs and a single result. Important steps in the use of this method include estimation and inference, variable selection in model building, and model fit testing [32]. Multiple linear regression models estimate coefficients similarly to simple models. As with simple linear regression, the least-square estimators in multiple linear regression are unbiased. In addition, unbiased estimators have the smallest variation and are consistent. Therefore, users may confidently infer multiple linear regression coefficients using least-square estimators as long as the regression assumptions stay true. The regression line approximates the response variable points. The variability surrounding the point estimate is also useful for validating inference assumptions, identifying problematic observations, and creating confidence or prediction ranges. Multiple linear regression is a valuable technique for administrative decision-making. With the availability of primary and secondary data and statistical analysis tools, everyone has the ability to do linear regression analysis [33]. The work of multiple linear regression is to model the linear relationship between the independent and dependent variables. These independent variables may continuous or discrete [35,36]. Multiple linear regression extends simple linear regression to include more than one explanatory variable. that means there is one dependent variable ($Y$) and more than an independent variable ($X_i$). Both scenarios still utilize the term 'linear' because the response variable is directly related to a linear combination of the explanatory variables.

The objective of decreasing the difference between observed and estimated values is frequently achieved via curve fitting using the regression method. The mathematical expression of the MLR is given in Equation (2.1).

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad (2.1)$$

where:

i: the number of observations

$Y_i$: dependent variable

$x_i$: explanatory variables

$\beta_0$: the average coefficient of the model

$\beta_p$: slope coefficients for each explanatory variable

$\beta_0$ ..... $\beta_p$: the linear coefficient of each explanatory variable.

ε: Represents the residual (fitting error) used to test the overall significance (F-test) of the equation and the importance of each regression coefficient (t-test). When the residual is distributed normally (normal distribution) and independent, i.e., with zero mean and constant variance $\delta^2$ correct results will be obtained. Using residual analysis, this can be checked. This analysis may also eliminate data outliers [35,36].

### 2.2. Regression Tree

The regression tree is one of the statistical tools used to make the decision in an uncomplicated and easy way. Depending on the response variable by looking at a vector of predictors or covariates. The regression tree differs from the rest of the traditional regression methods in that it does not need parameters as in simple and multiple linear regression, it is distinguished in its work on strong variable subdivisions, is not affected by outliers if any, and can be implemented on different types of data [16]. Regression trees predict output with continuous values rather than discrete values.

To construct a regression tree for the best prediction of the dependent variable Y for the independent variable X ,the particular observations of the independent variable are sorted, then the average of the two observations is found ($a_1$, $a_2$) and then these two

values are divided into two groups ( $X \geq a_1$ , $X < a_2$), then the average of the values that fall within these two groups is found concerning the values of the dependent variable Y, and these two values represent the expected output of the decision tree ( $X \geq a_1$ , $X < a_2$). Then the mean squared error (MSE) is calculated for each group. After calculating the average for the first two observations in variable X, the data set was divided accordingly, and the predictions were calculated. The process is repeated for the rest of the observations and finding the MSE. After MSE calculations, the point at which the division of the dataset will be determined is the lowest MSE. Then the node, i.e., the root node, is divided into two branches, right and left, and then the data points that go towards the right branch and the left branch of the root node are repeatedly subjected to the same algorithm for further splitting. In the case of more than one independent variable, the same process mentioned above is performed, where the variables' data are sorted separately. Then the points that reduce the MSE for all variables are calculated. The variable with the lowest estimate is selected among the variables and the points calculated for them [17].

### 2.2.1. CART

The CART algorithm is a representation of the connection between the dependent variable and independent factors in a dataset. It consists of a sequential binary partition of the dataset based on the variable values. Fitting tree models involves repeatedly splitting the data into homogenous groups. The output is a hierarchical tree of relevant decision rules for classification or prediction. CART is a segmentation modelling technique that satisfies the following properties:

1) The hierarchy is referred to as a tree, and each section is a node.
2) The root node includes the whole database.
3) The root node is subdivided in order to produce child nodes.
4) The ultimate subgroups are referred to as terminal nodes or leaves when further data subdivision is impossible.
5) For the creation of the CART, three main components must be determined: a set of questions defining the division of data, a criterion for evaluating the best division, and a mechanism for terminating further subdivisions (stop-splitting rule) [14].

Important Terminology:

- Root node: The root represents the entire sample or population, and the root node is divided into two or more groups as homogeneous groups.
- Splitting: This procedure divides the tree from its nodes into two or more nodes.
- Leaf or Terminal Node: Nodes that cannot be divided, that is, have reached the minimum division, are called leaf nodes.
- Pruning: The process of removing child nodes from the root node is the opposite of splitting.
- Branch or Sub-Tree: A branch or subtree is a subdivision of the overall tree.
- Parent and Child Node: The nodes that sub-nodes are separated into.

The regression problem (Figure 2.1.) shows points in the XY plane. These points are marked with a cross. The regression problem is to find a "good" function y = f (x) whose graph lies close to the given data points (Figure 2.1.). The regression tree method is done by dividing the X-axis region into a number of sections. Then the average value of the dependent variable Y is calculated, which represents a constant value of the regression function at each sub-region resulting from the division.

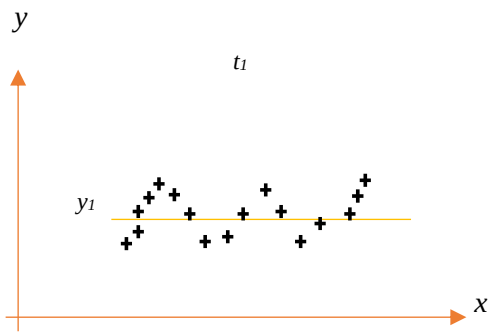


**Figure 2.1**. Given data with X on the x axis


**Figure 2.2.** The regression tree with one node $t_1$

To begin, (Figure 2.1) contains a single region: the entirety of the x-axis. The regression function must be constant in this region. As a result, the average must be the most plausible representative. This average y-value y1 is stored in (Figure 2.2) a one-node tree.

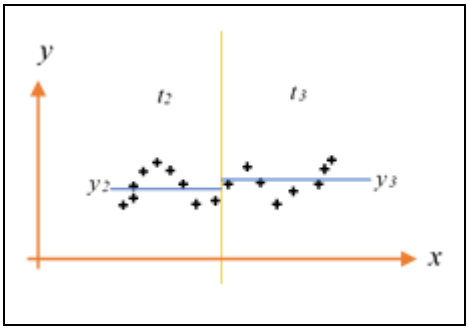


**Figure 2.3**. Splitting $\boldsymbol{t_1}$

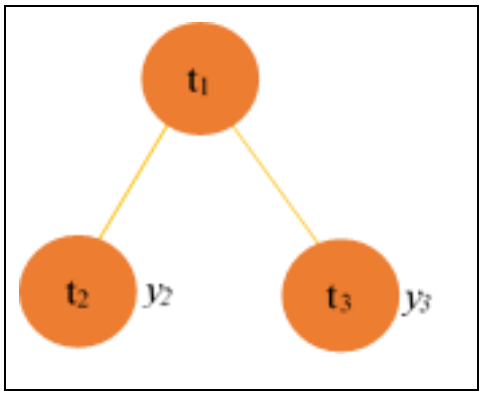


**Figure 2.4.** Regression tree after the split of $\boldsymbol{t_1}$

The region $t_1$ has now been divided into two subregions, $t_2$ and $t_3$. The average y-value of the data for each subregion is $y_2$ and $y_3$, respectively, as shown in (Figure 2.4). The subregions $t_2$ and $t_3$ further subdivided, the average y-value for each smaller subregion are calculated and recorded (Figure 2.6).

Now let us assume that the division halted. The region is then subdivided into four subregions labeled $t_4$, $t_5$, $t_6$, and $t_7$. The average t-value, $y_4$, $y_5$, $y_6$ and $y_7$ are determined for each subregion. Thus, the final regression function y = f (x) is a piecewise constant with values $y_i$ on $t_i$ for i = 4, 5, 6, and 7 [8].

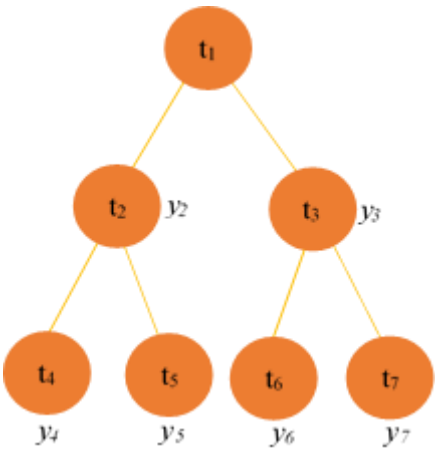


**Figure 2.5.** Regression tree after split of $t_2$ and $t_3$

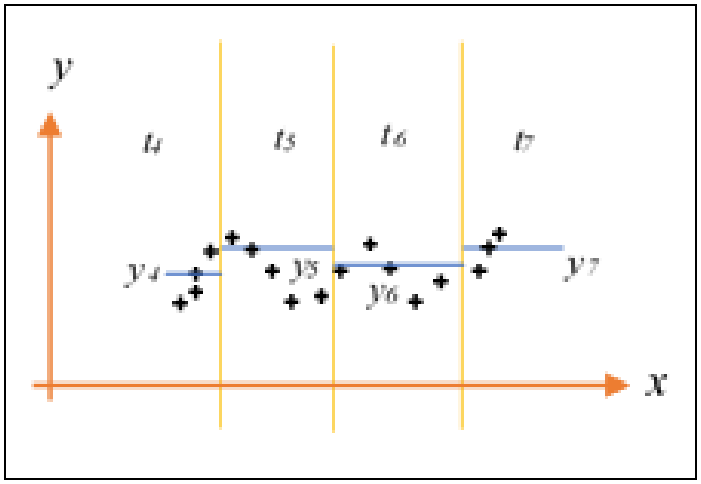


**Figure 2.6.** Splitting of $t_2$ and $t_3$

The procedure of divisions on the tree affects the accuracy of the tree. In this study, a regression tree was used (the goal of a regression tree is to generate a line that best fits the data), and the mean square error (MSE) was used as a split criterion. When making a binary tree, the algorithm will choose a value and divide the data into two subsets. The MSE will be calculated separately for each subset, and the value that gives the smallest MSE is chosen. The MSE partition criterion is calculated for each node after the base model is formed, which is the average of the data points. The mathematical expression of the MSE is given in Equation (2.2).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \qquad (2.2)$$

where $y_i$ is the actual value, $\hat{y}$ is the prediction, n is the number of samples.

The prediction process is done by asking the true or false question of the data point that runs through the entire tree until it reaches the leaf node. The final prediction represents the average value of the dependent variable in that leaf node [17].

In this study, CART was used to choose the best model for multiple linear regression.

### 2.3. Performance Evaluation

#### 2.3.1. Coefficient of determination (R-square - $R^2$)

The coefficient of determination is represented as the variance ratio of the dependent variable that can be expected among the set of independent variables in the multiple regression equation. Its value lies between zero and one. The use of R-square is intended to determine the quality of the model used, whenever the value of R-square is close to one, this indicates the priority of the model, and if it is equal to zero, that is, the value of Y cannot be predicted, and the model followed is not good [38]. The Equation of R-square is

$$R^2=1-\frac{SSR}{SST} \tag{2.3}$$

Where :

SST(sum of squared totals)= $\sum_{i=1}^{n}(y_i-\bar{y})^2$ , SSR (sum of squared residuals) = $\sum_{i=1}^{n}(y_i-\widehat{y_i})^2$

$y_i$= dependent variable, $\widehat{y_i}$= predicted value of the dependent variable, $\bar{y}$= dependent variable mean [39].

#### 2.3.2. Mean absolute percentage error (MAPE)

Choosing the best regression model to give the best prediction of the Y value is the one in which the MSE value is the lowest among the other models. There are other quality measures represented by the mean absolute error percentage (MAPE). In practice, MAPE is used because it gives an intuitive explanation in terms of relative error. For example, it is included in the price calibration of products, that is, the measurement of gains and losses, because they are often given in relative values.MAPE is calculated in the equation (2.4) [43].

$$\text{MAPE}=\frac{\sum_{i}^{n}(\frac{|y_i-\hat{y}_i|}{|y_i|})}{n} \tag{2.4}$$

Where :

$y_i$ = the actual data value.

$\hat{y}_i$ = the forecasted data value.

n = sample size.

# 3. MODELING

The definition of the hybrid technique, which is the primary purpose of the thesis, and the datasets used are explained in this section. First, the proposed technique is mentioned, and then the information on the datasets used is given.

## 3.1. Modeling Approach

### 3.1.1. Datasets

Three sets of data were used in this thesis. The first is the advertising data set, which was represented by using (TV, radio and newspapers) (X) to show the relationship between these advertising methods with sales (Y) in terms of their impact on sales and purchasing power. The number of data is represented by 200 samples [40].

The second data set contains 159 samples represented (Species of fish, length, height, width), which are the independent variables (X) and their impact on the weight of the fish, which represents the dependent variable [41].

The third set of data is the effect of the car's specifications on its price, which was considered the dependent variable. The car specification was (CarName, fueltype, aspiration, doornumber, carbody , drivewheel, enginelocation, wheelbase,carlength, carwidth, carheight, curbweight, enginetype, cylindernumber, enginesize, fuelsystem, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg and highwaympg). The number of samples was 205 samples [42].

The data was analyzed and processed using the Python language, version 3.9.7 using the Jupyter Notebook tool, based on the scikit-learn library. The data were divided into training and test 80% - 20%, respectively, and Excel program.

### 3.1.2. CART properties

The mean square error was chosen as a criterion for dividing the tree and finding its branches. As for the parameters that were used and the values that were given, they were as follows:

The maximum tree depth is a stop limit for further splits of nodes. The max_depth value is an integer value. The max_depth of the tree was used equal to three[45]. When searching for the best division, the number of selected features is taken into account. If this feature is not set, the tree will take all the available features into consideration. In this study, the number of features was not determined. max_features =None. The min_samples_split used to specify the number of samples to be leaf nodes. This parameter is also used to limit tree growth, min_samples_leaf =1. And min_samples_split represents the minimum number of samples the node must contain to split. through which the tree can be organized. The default value of min_samples_split is two. The tree searches for features of the split by means of a parameter splitter whose default value is “best”, which means that for each node the algorithm takes into account all the features and chooses the best split. If the splitter parameter is set to "random", then a random subset of the features will be taken. Then the best feature will be partitioned within the random subset [44]. max_leaf_nodes Used to determine the number of leaf nodes in a tree, max_leaf_nodes=None[46]. min_impurity_decrease this parameter is used to split a node in the event that the impurity of this node decreases as the number of splits increases so that the value of the impurity is greater or equal to the value of this parameter, min_impurity_decrease =0.0 [47]. min_weight_fraction_leaf It represents the portion of the input samples to be placed in the leaf node. The weights are determined based on the weight of the sample, min_weight_fraction_leaf =0.0 [48].

From the observation, the max depth parameter equal to three was used, so for the huge of the tree formed in the absence of specifying the number of tree levels, and also for the ease of application and explanation of the concept of the research study. Tree graphs were created export_graphviz library. The same scale and parameters were applied to all data represented by advertising data, fish and vehicle.

## 4. RESULTS

In this thesis, results were compared to predict the value of the dependent variable (Y) using the regression tree method, multiple linear regression, and the special research method of splitting the regression tree and constructing multiple linear regression models from it in order to select the best method that gives the best prediction based on the R-square value. Three types of data were dealt with (Advertising, Fish and Car).

After creating a regression tree for the three data, the results were as follows.

### 4.1. Advertising Data Results

In Figure 4.1 the size of the tree was chosen to consist of three levels (max_depth= 3), where notice that the TV and radio variables are dominant in the tree, being the most influential data on the dependent variable sales, where the TV advertisement splits with 130.25.(where 130.25: is the decision threshold for that node (root node),mse: represent the likely error make by the node in splitting samples in to left and right side, samples: is the tatal number data points that split in to two, value: in a decision tree regressor is the value that the tree would predict for a new example falling in that node, To view, see pages 13, 14 and 15).The second splits were equal to 30.05 in the TV advertisement and 26.85 on radio advertisement. It is considered TV advertisements are more effective then the radio advertisement. Also noticeable in the third level of the tree, which is called by the leaves, that the smallest sample is 7 and the largest sample is 31.

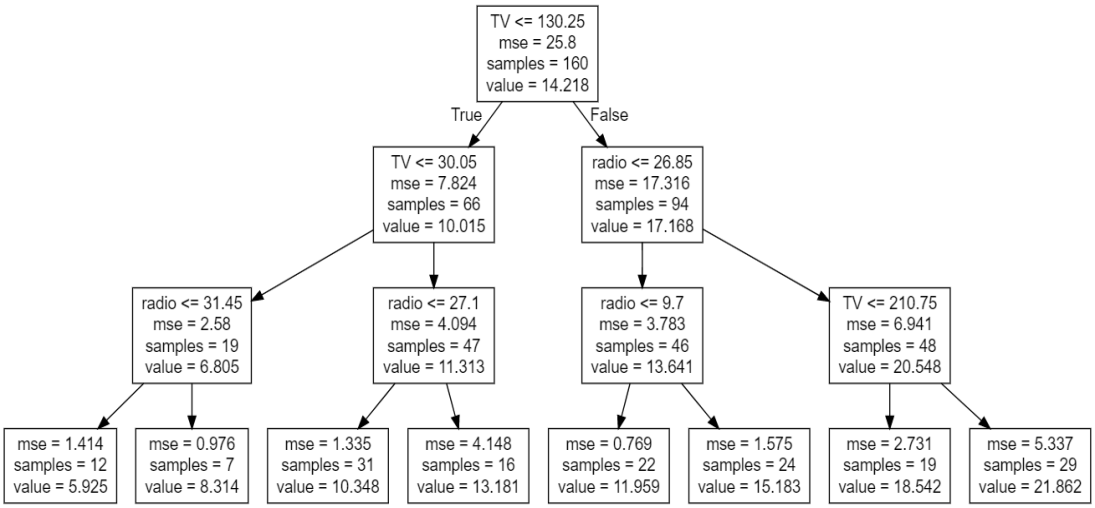


**Figure 4.1.** Advertising max depth 3 regression tree

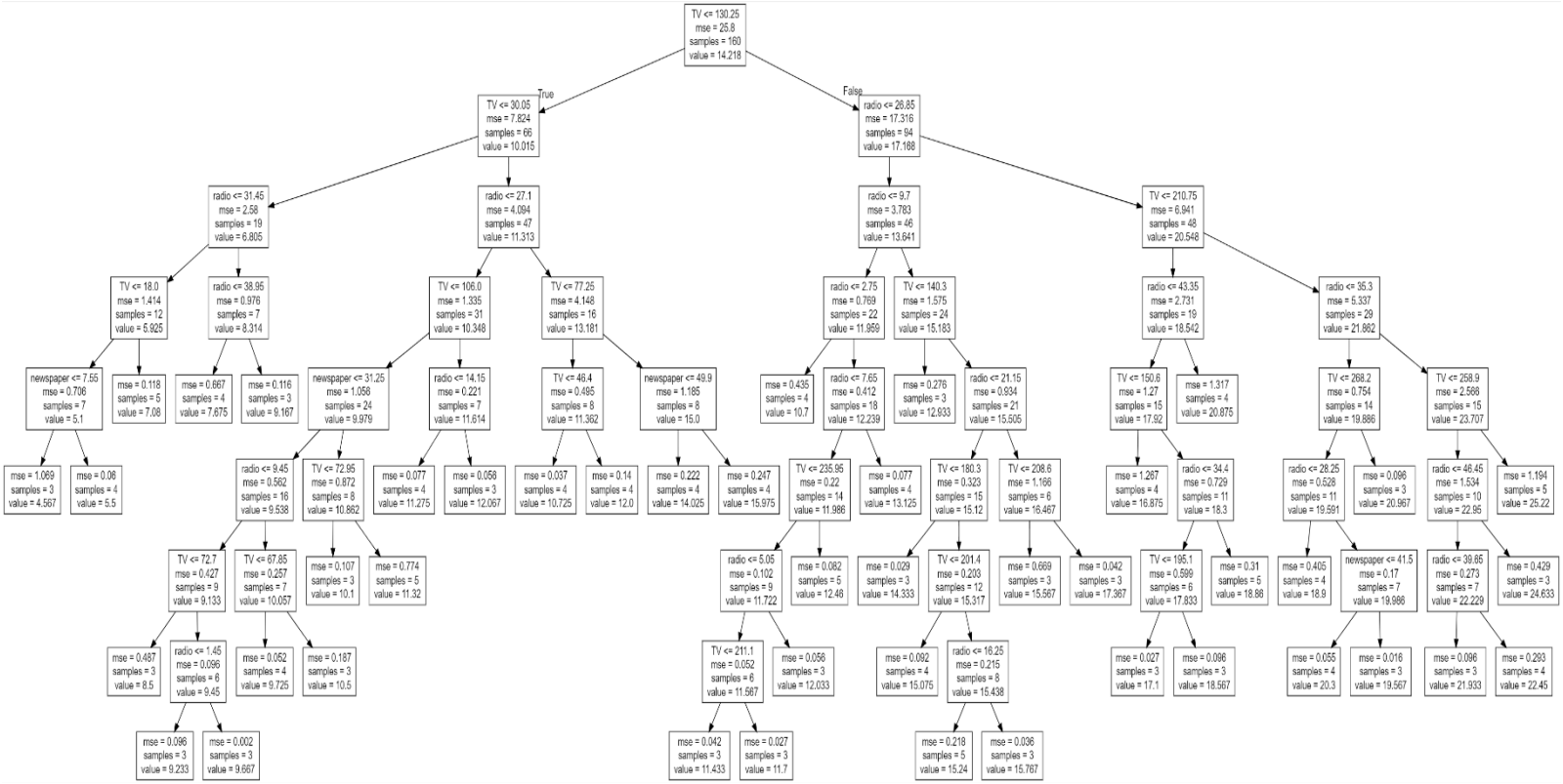


**Figure 4.2**. Full Advertising regression tree

In figure 4.2 represents the full tree of all advertising data with a tree size of eight levels (max _depth= 8). As mentioned in Figure 4.1, the television and radio were the most important variables with the values of ( 30.05), (26.85 ) respectively, while the importance of newspapers appeared at the fourth level of the tree with the value of (31.25). The smallest sample size was equal (min_sample=3) and the largest sample size was equal (max_sample=5).

In Table 4.1 It shows that the data was siplited into three levels, represented by the following symbols (L0, L1, L2, L3). At level (L0) represents a multiple linear regression model for all data. While the first level (L1) is the level that was siplited on the basis of the root node and consists of two linear regression equations as seen in Figure 4.1. While the second level (L2) was calculated into four linear regression equations. As for the last level in the tree with a max_depth equal to three, it consists of eight linear regression equations.

When reading Table 4.1 The value of the coefficient of the variable TV multiple linear regression model (0.0445840201199643), and at the first level in the right branch its value (0.0345823663101959) and the second level in the right-left branch (0.0204044149386929), while its value at the third level in the left-left-right branch (0.152197800802342) shows that The values of variable TV coefficient decreases or increases according to the level it is in. The same results can also be observed for the radio variable which a multiple linear regression model value (0.196497034155405), and at the first level in the right branch its value (0.264812054442296) and the second level in the right-left branch (0.254001958373097), while its value at the third level in the left-left-right branch (0.144874029642748). And newspaper variable which a multiple linear regression model value (-0.00278146398192599), and at the first level in the right branch its value (-0.00013184004) and the second level in the right-left branch (-0.002767044), while its value at the third level in the left-left-right branch (-0.000586751). As for the value of p, it was found that its value also decreases or rises according to the level and depth of the tree. But the p-value is expected to decrease as the depth of the tree increases.

A p-value of the Tv variable of the multiple linear regression model is (4.45749438814932), and at the first level in the right branch its value (1.57586079879491) and the second level in the right-left branch (5.25634911539797), while its value at the third level in the left-left-right branch

(0.00180069782084902). A p-value (where P-value is The P-value is a statistical number to conclude if there is a relationship between the dependent variable and an independent variable, if P-value (< 0.05) means that the coefficient is likely not to equal zero and if P-value (> 0.05) means that cannot conclude that the explanatory variable affects the dependent variable but if it's zero there is no relationship [50], and the regression coefficient represents the extent to which the average of the dependent variable changes when one unit changes in the independent variable while keeping the other independent variables at the model constant. When the value of the coefficient is positive, this indicates that when the value of the independent variable increases, the average of the dependent variable tends to increase, and in the event that it is negative, the average of the dependent variable tends to decrease. [51]) of the radio variable of the multiple linear regression model is (1.14899328505819), and at the first level in the right branch its value (3.31823930330662) and the second level in the right-left branch (1.4040996600736), while its value at the third level in the left-left-right branch (0.0160975065617443). And a p-value of newspaper variable of the multiple linear regression model is (0.652809785378682), and at the first level in the right branch its value (0.598148099937265) and the second level in the right-left branch (0.447299196919234), while its value at the third level in the left-left-right branch (0.0160975065617443).

**Table 4.1.** Advertising tree levels

| | | **intercept** | | **TV** | |
|---|---|---|---|---|---|
| | | coefficient | P -value | coefficient | p-value |
| **L0** | | 2.994893 | 4.65E+00 | 0.04458402 | 4.457494388 |
| **L1** | left | 3.571018326 | 6.850299417 | 0.064994601 | 7.158134399 |
| **L1** | right | 3.305860686 | 5.399467389 | 0.034582366 | 1.575860799 |
| **L2** | left -left | 2.513210161 | 1.586497815 | 0.16114265 | 1.993317505 |
| **L2** | left-right | 4.503562114 | 1.582761308 | 0.050786318 | 6.0000025 |
| **L2** | right-left | 6.480825039 | 3.852163488 | 0.020404415 | 5.256349115 |
| **L2** | right-right | -0.48032 | 0.260431599 | 0.048198127 | 2.137088554 |
| **L3** | left-left-left | 2.780446895 | 0.000203945 | 0.160756657 | 9.95752E-05 |
| **L3** | left-left-right | -0.04555468 | 0.966760978 | 0.152197801 | 0.001800698 |
| **L3** | left-right-left | 5.47914619 | 2.66917087 | 0.039702032 | 4.311163717 |
| **L3** | left-right-right | 3.356731419 | 0.001993394 | 0.0662364 | 2.593247294 |
| **L3** | right-left-left | 7.683138705 | 9.293370754 | 0.013047501 | 3.554682115 |
| **L3** | right-left-right | 5.063502158 | 1.697627382 | 0.026204213 | 9.247471772 |
| **L3** | right-right-left | 0.469709392 | 0.516937682 | 0.050120482 | 1.064262277 |
| **L3** | right-right-right | -0.218273671 | 0.766939701 | 0.044856924 | 3.747060905 |

**Table 4.1. (Continued)** Advertising tree levels

| | | radio | | newspaper | |
|---|---|---|---|---|---|
| | | coefficient. | p-value | coefficient | p-value |
| **L0** | | 0.196497034 | 1.148993285 | -0.00278146 | 0.652809785 |
| **L1** | left | 0.105833316 | 1.468126835 | 0.002979577 | 0.5981481 |
| **L1** | right | 0.264812054 | 3.318239303 | -0.00013184 | 0.970780723 |
| **L2** | left -left | 0.065575253 | 1.132068466 | 0.007245233 | 0.143687071 |
| **L2** | left-right | 0.117098762 | 4.202453778 | 0.007556891 | 0.084177185 |
| **L2** | right-left | 0.254001958 | 1.40409966 | -0.00276704 | 0.447299197 |
| **L2** | right-right | 0.287877474 | 5.598730405 | -0.00131818 | 0.552447222 |
| **L3** | left-left-left | 0.05523426 | 0.00512405 | 0.000172803 | 0.987313415 |
| **L3** | left-left-right | 0.14487403 | 0.016097507 | -0.00058675 | 0.016097507 |
| **L3** | left-right-left | 0.109269663 | 5.417694318 | 0.009893344 | 0.026968077 |
| **L3** | left-right-right | 0.119674606 | 0.000054428 | 0.002944859 | 0.604963801 |
| **L3** | right-left-left | 0.302945495 | 7.944296246 | 0.002570154 | 0.43359834 |
| **L3** | right-left-right | 0.26700997 | 9.014985909 | -0.00242429 | 0.624610316 |
| **L3** | right-right-left | 0.253009533 | 4.437693211 | -0.00097284 | 0.742120374 |
| **L3** | right-right-right | 0.302240557 | 1.151158853 | -2.68147 | 0.991684836 |

**Table 4.2**. Advertising results

| | MAPE | | RMSE | | R2 | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| L0 | 0.114312 | 0.246109 | 1.551391 | 2.098123 | 0.906711 | 0.860432 |
| L1 | 0.054789 | 0.150971 | 0.758873 | 1.238536 | 0.977678 | 0.952996 |
| L2 | 0.027445 | 0.145521 | 0.407861 | 1.264379 | 0.993552 | 0.950862 |
| L3 | 0.020385 | 0.130592 | 0.311288 | 1.118594 | 0.996244 | 0.962266 |
| Tree | 0.458673 | 0.559148 | 6.821821 | 7.683423 | 0.00282 | 0.03134 |
| Tree Full | 0.463218 | 0.627452 | 6.956178 | 8.21254 | 0.003228 | 0.030563 |

In Table 4.2 shows the mean absolute percentage error, mean square error and coefficient of determination to measure the quality of the model. Where it was found that the value of the absolute error rate in the train MAPE and test MAPE decreases as the level of the tree increases, and this is expected when dividing the data in the form of multiple linear regression models, which makes the solution more accurate. It is noticeable that the third level gave the best model compared to the rest of the levels and the multiple linear regression model for all data, as well as for the value of the full tree.

### 4.2. Fish Data Results

In the fish data set, it also depended on the depth of a tree consisting of three levels (max_depth = 3). From Figure 4.3 where notice that the Width and length3 variables are dominant in the tree, being the most influential data on the dependent variable weight, where the width splits with 5.154. The second splits were equal to 41.55 in

the length3 of the right branch of the tree while the value of the length3 of the left branch is equal to 27.95. It is considered width more effective than length3. Also noticeable in the third level of the tree, is that the smallest sample is 2 and the largest sample is 31.

The tree in Figure 4.4 represents the full tree of fish data. It consists of eight levels (max_depth = 8). It can be noted the importance of the independent variables that affect the weight of fish, where the length3 was considered in the first degree of importance with the values of (41.55 ),(27.95 ), then the length1 at the second level with a value (46.55), followed by the height at the third level with a value (6.61) and the length2 with a value (32.9). The smallest value for the sample size (min_sample=3) and the maximum value for the sample size (max_sample=5).

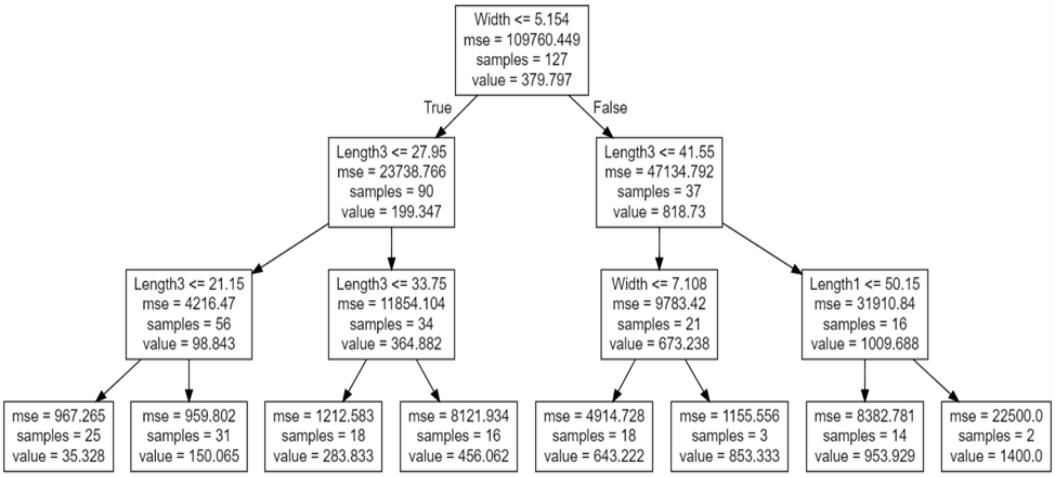


**Figure 4.3.** Fish max depth 3 regression tree

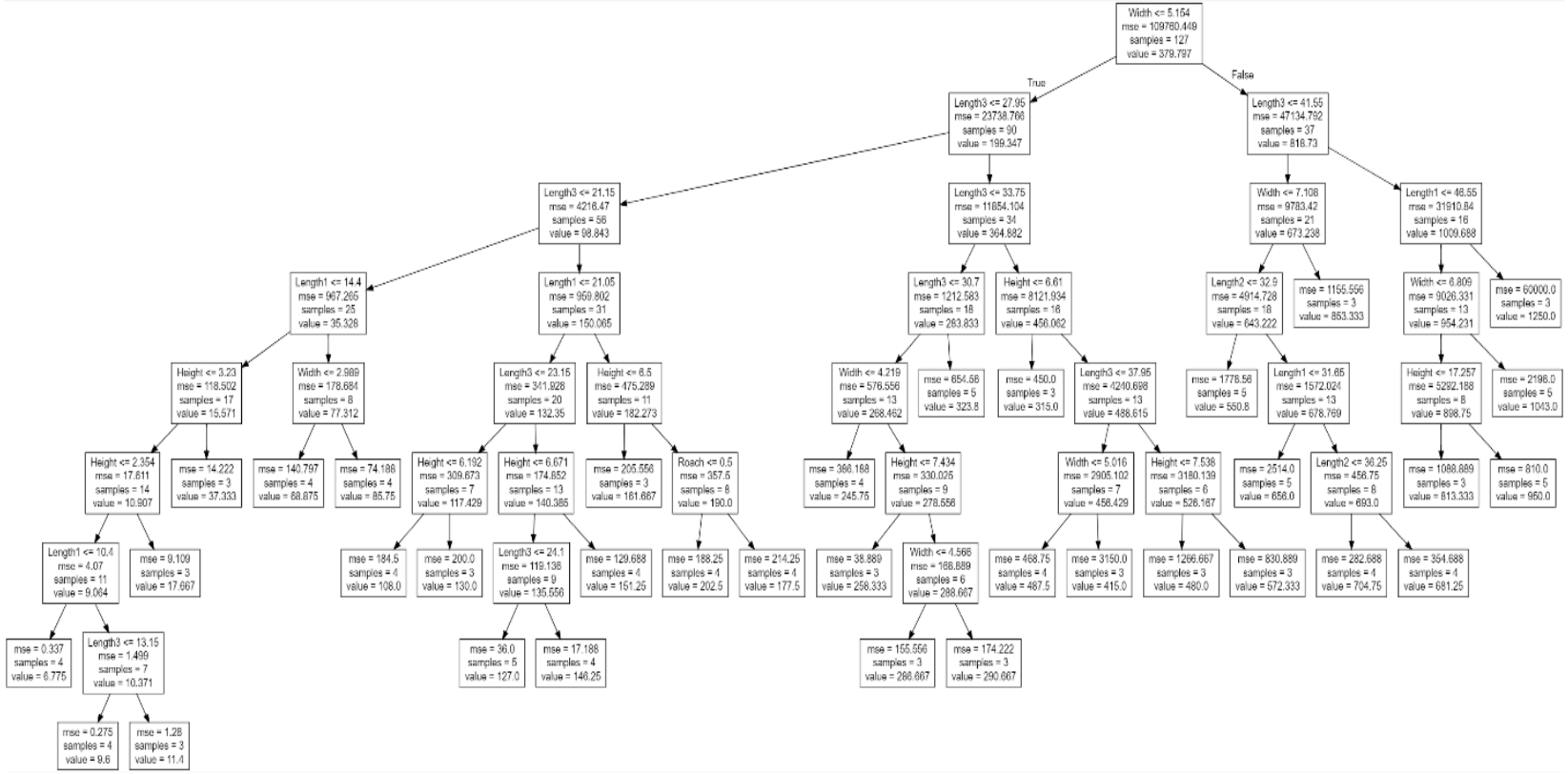


**Figure 4.4.** Full fish regression tree

**Table 4.3.** Fish tree levels

| | | Whitefish | | Parkki | | Pike | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -116.2916353 | 0.07408502 | -105.8356565 | 0.153583999 | -158.7650767 | 0.07241655 |
| **L1** | left | 47.03676799 | 0.12781536 | -79.09817584 | 0.033258817 | 138.6516953 | 0.007228342 |
| **L1** | right | 5.774624391 | #NUM! | 0 | #NUM! | -219.2246656 | #NUM! |
| **L2** | left -left | 0 | #NUM! | -37.07042942 | #NUM! | 0 | #NUM! |
| **L2** | left-right | -4.394399174 | 0.942636793 | -100.7635059 | 0.247253041 | 36.90894792 | 0.673416714 |
| **L2** | right-left | -66.00032376 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L2** | right-right | 0 | #NUM! | 0 | #NUM! | -483.928788 | #NUM! |
| **L3** | left-left-left | 0 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L3** | left-left-right | 0 | #NUM! | -29.74024074 | #NUM! | 0 | #NUM! |
| **L3** | left-right-left | 102.0008366 | 0.083215174 | 128.0332786 | 0.140186016 | 0 | #NUM! |
| **L3** | left-right-right | 0 | #NUM! | 0 | #NUM! | 109.2015751 | #NUM! |
| **L3** | right-left-left | -172.4939639 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L3** | right-left-right | 0 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L3** | right-right-left | 0 | #NUM! | 0 | #NUM! | 444.9484266 | #NUM! |
| **L3** | right-right-right | 0 | #NUM! | 0 | #NUM! | 0 | #NUM! |

| | | intercept | | Bream | | Roach | |
|---|---|---|---|---|---|---|---|
| | | coefficient | P -value | coefficient | p-value | coefficient | p-value |
| **L0** | | -700.7939486 | 1.000673623 | -229.6958212 | 0.072271184 | -33.59370862 | 0.072271184 |
| **L1** | left | -205.228072 | 4.95E+00 | 26.93935183 | 0.652226699 | 3.083808394 | 0.8841098 |
| **L1** | right | -1285.158475 | 1.25993E-06 | -44.18675269 | 0.880242362 | 0 | #NUM! |
| **L2** | left -left | -119.4132972 | 1.85357E-10 | 0 | #NUM! | -29.42679154 | #NUM! |
| **L2** | left-right | -562.658319 | 1.65666E-05 | -122.6781559 | 0.366861868 | -12.62603174 | 0.839524832 |
| **L2** | right-left | -671.493322 | 0.014186487 | 15.32279558 | 0.948419023 | 0 | #NUM! |
| **L2** | right-right | -1409.733348 | 0.005787456 | -843.8811868 | 0.067964264 | 0 | #NUM! |
| **L3** | left-left-left | -50.40699228 | 0.000295845 | 0 | #NUM! | -27.76449425 | #NUM! |
| **L3** | left-left-right | -217.297353 | 0.000947936 | 0 | #NUM! | -17.56321434 | #NUM! |
| **L3** | left-right-left | -447.7343041 | 0.002429337 | 208.0837076 | 0.144809121 | 138.738701 | 0.071188079 |
| **L3** | left-right-right | -483.7553394 | 0.403985361 | 0 | #NUM! | 0 | #NUM! |
| **L3** | right-left-left | -55.70758472 | 0.921386934 | -473.9865367 | 0.379914096 | 0 | #NUM! |
| **L3** | right-left-right | 998.5365854 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L3** | right-right-left | 53.79017663 | 0.914759909 | -111.6154318 | 0.735498778 | 0 | #NUM! |
| **L3** | right-right-right | -2915.116279 | #NUM! | 0 | #NUM! | 0 | #NUM! |

**Table 4.3. (Continued)** Fish tree levels

| | | Smelt | | Length1 | | Length2 | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 302.5883268 | 5.06844E-10 | -71.73958808 | 0.067663026 | 76.67271467 | 0.106831958 |
| **L1** | left | 60.40892895 | 0.008889484 | 60.18467653 | 0.004676044 | -18.68381749 | 0.404004629 |
| **L1** | right | 0 | #NUM! | -46.59054016 | #NUM! | 61.23580423 | 0.504983296 |
| **L2** | left -left | 1.295649255 | #NUM! | 51.94391456 | 0.00068853 | -50.83364644 | 0.000857843 |
| **L2** | left-right | 0 | #NUM! | 7.263888629 | #NUM! | 15.47205961 | 0.714684662 |
| **L2** | right-left | 0 | #NUM! | 100.3051193 | #NUM! | -119.7960936 | 0.076203025 |
| **L2** | right-right | 0 | #NUM! | -180.9867786 | #NUM! | 72.71026829 | 0.622214392 |
| **L3** | left-left-left | 0.461513884 | #NUM! | -19.24694003 | 0.234344638 | 2.351529043 | 0.834872001 |
| **L3** | left-left-right | 0 | #NUM! | 40.65372115 | #NUM! | -28.24363009 | 0.302168232 |
| **L3** | left-right-left | 0 | #NUM! | -10.74034063 | #NUM! | 91.71255948 | 0.018863665 |
| **L3** | left-right-right | 0 | #NUM! | 24.68509529 | #NUM! | 12.53290145 | 0.907073941 |
| **L3** | right-left-left | 0 | #NUM! | 63.38091543 | #NUM! | -167.2009412 | 0.147187298 |
| **L3** | right-left-right | 0 | #NUM! | -4.87804878 | #NUM! | 0 | #NUM! |
| **L3** | right-right-left | 0 | #NUM! | 89.39683637 | #NUM! | 123.972658 | 0.337013914 |
| **L3** | right-right-right | 0 | #NUM! | 0 | #NUM! | 0 | #NUM! |

| | | Length3 | | Height | | Width | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 16.28970884 | 0.605342814 | 32.13498485 | 0.032473606 | 13.20175457 | 0.603989463 |
| **L1** | left | -30.50389572 | 0.030849936 | 39.88838599 | 2.03531E-05 | 7.927202885 | 0.637311506 |
| **L1** | right | 20.10271609 | 0.758169805 | 14.67404681 | 0.559861337 | 62.13867841 | 0.146970791 |
| **L2** | left -left | 6.153136343 | 0.582592797 | 15.33210762 | 0.022203147 | 27.12511459 | 0.027634911 |
| **L2** | left-right | -8.462221856 | 0.752810298 | 46.24808278 | 0.004725603 | 25.44348805 | 0.459387941 |
| **L2** | right-left | 38.50284494 | 0.378920641 | 8.06569331 | 0.729924771 | 112.2709705 | 0.001252214 |
| **L2** | right-right | 142.549956 | 0.248142335 | 54.7622247 | 0.192469808 | -81.02254437 | 0.251386968 |
| **L3** | left-left-left | 16.04814265 | 0.157613054 | 14.87680508 | 0.108639257 | 5.692389074 | 0.580880149 |
| **L3** | left-left-right | -3.493620485 | 0.789182849 | 18.46761839 | 0.011251052 | 37.067307 | 0.027936678 |
| **L3** | left-right-left | -64.44603708 | 0.06941591 | 12.14119408 | 0.326965548 | 52.59477042 | 0.019594901 |
| **L3** | left-right-right | -23.50891706 | 0.699586709 | 43.70100643 | 0.212751225 | 20.95920368 | 0.8024212 |
| **L3** | right-left-left | 109.1103124 | 0.24154254 | 35.83520268 | 0.472293263 | 7.666960772 | 0.90740322 |
| **L3** | right-left-right | 0 | #NUM! | 0 | #NUM! | 0 | #NUM! |
| **L3** | right-right-left | -205.0395508 | 0.191984187 | 124.3046545 | 0.050472531 | -19.82475278 | 0.719210271 |
| **L3** | right-right-right | 69.76744186 | #NUM! | 0 | #NUM! | 0 | #NUM! |

As previously mentioned about the symbols of the tree levels in the advertising data set, the same mechanism was dealt with the fish data set. From reading Tables 4.3 and 4.4, it was found that the value of the coefficient of the linear regression model for the variables (Bream= -229.6958212, Roach= -33.59370862, Whitefish= -116.291635283559, Parkki= -105.835656534886, Pike= -158.765076662994, Smelt=

302.588326825736, Length1= -71.7395880841768, Length2= 76.672714674299, Length3= 16.289708843742, Height= 32.1349848470633, Width= 13.2017545656934), while the values of the coefficient of the linear regression for the three levels increase or decrease according to the level in which it is, as it was observed in the advertising data set, or it is equal to zero, that is, there is no linear relationship or influence between the independent variables and the dependent variable.

The p-value of the multiple linear regression model for all data (Bream= 0.0722711843532193, Roach= 0.0722711843532193, Whitefish= 0.0740850203548798, Parkki= 0.153583999276752, Pike= 0.0724165497916362, Smelt= 5.06844012584746, Length1= 0.0676630263147532, Length2= 0.106831958131172, Length3= 0.605342813860757, Height= 0.0324736063588774, Width= 0.603989463006968). While the value of p when dividing the data, it can be seen in Table 4.3 when the variable Roach is that it decreases as the depth of the tree increases and this is expected (L1-left= 0.884109869981005,L2-left-right= 0.839524831920473,L3-left-right-left= 0.0711880789199419), although there are some values in the variable Roach that are not applicable. While the rest of the variables, the values of p increase or decrease depending on the level at which it is.

**Table 4.4.** Fish results

| | MAPE | | RMSE | | R2 | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| L0 | 0.383754 | 0.30317 | 71.66277 | 121.5637 | 0.956551 | 0.937677 |
| L1 | 0.185975 | 0.140321 | 44.64082 | 64.83715 | 0.981868 | 0.97922 |
| L2 | 0.123813 | 0.109866 | 32.23212 | 54.35301 | 0.990538 | 0.985438 |
| L3 | 0.078388 | 0.17237 | 25.63684 | 91.51419 | 0.994022 | 0.960422 |
| Tree | 5.383027 | 2.453518 | 462.5981 | 584.213 | 7.76E-05 | 0.000137 |
| Tree Full | 5.434081 | 2.413886 | 463.6655 | 575.1419 | 0.000134 | 6.44E-05 |

It is noticeable in Table 4.5 that the train MAPE value at the third level gave the best result, while the test MAPE value at the second level gave the best result. In this case, the third level cannot be considered the level that represents the best model.

### 4.3. Car Data Results

In the car data set at Figure 4.5 of a three-level tree(max_depth=3), it can be seen that the enginesize, curbweight and peakrpm variables are dominant in the tree, being the

most influential data on the dependent variable price, where enginesize splits with 182. The second splits were equal to 2544 in thecurbweihkt and 4425 on peakrp. It is considered enginesize more effective than curbweight and peakrpm. Also noticeable in the third level of the tree the smallest sample is 1 and the largest sample is 50.

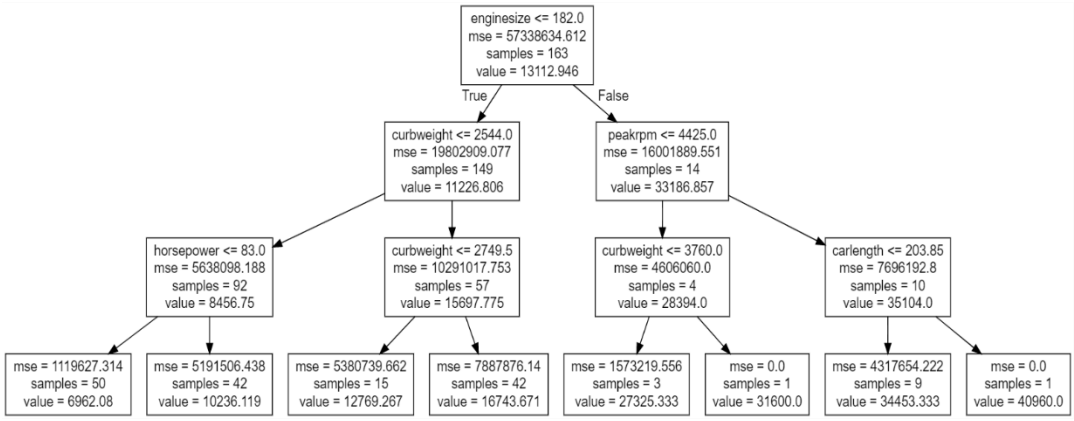


**Figure 4.5.** Car max depth 3 regression tree

the Full tree data of the car, consists of eight levels (max_depth= 8).The two independent variables horsepower and curbwieght with values 139 , 2544 respectively ,represent the most influential variables, then follow this in the second level, carwidth= 70.55 and at the third level carlength=184.65, peakrpm=5450,citympg= 22 and the highway mpg= 29.5 and at the fourth level, the drivewheel_rwd= 0.5 and enginsize=114 and at the fifth level the wheelbase=94.4, stroke= 3.29 and boreratio=3.585 and at the sixth level carbody_hatchback=0.5 and at the seventh level doornumber=0.5. Minimum sample value (min_sample=3) and maximum sample value (max_sample=5 ).

**Table 4.5.** Car tree levels

| | | intercept | | fueltype | | aspiration | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -2.33E+04 | -4.30E+04 | -1.84E+04 | -2.89E+04 | 1904.2679 | -193.332 |
| **L1** | left | -2.16E+04 | -3.73E+04 | -1.03E+04 | -1.94E+04 | 1797.9546 | 2.827 |
| **L1** | right | | | -134.8273 | -434.655 | 22.9391 | -261.985 |
| **L2** | left -left | -4.47E+04 | -6.60E+04 | -1.63E+04 | -28100 | -671.767 | -2302.773 |
| **L2** | left-right | 271.7626 | -41300 | -7816.0254 | -29500 | 9619.2476 | 3995.16 |
| **L2** | right-left | | | -1.68E-10 | nan | -0.0002 | nan |
| **L2** | right-right | | | -15.1329 | -36.313 | 2.41E-10 | -1.03E-10 |
| **L3** | left-left-left | | | -1.57E+04 | -3.08E+04 | 374.2283 | -2029.104 |
| **L3** | left-left-right | | | -9.85E+04 | -5.34E+05 | -255.8169 | -2731.236 |
| **L3** | left-right-left | 2.2035 | -108.8839 | -569.746 | 160.2785 | -2187.053 | 157.0955 |
| **L3** | left-right-right | 3.81E+04 | 4463.7935 | -52000 | 1.56E+04 | -3.29E+04 | 482.4198 |
| **L3** | right-left-left | | | -4.52E-11 | nan | 0.0052 | nan |
| **L3** | right-left-right | | | 1.42E-16 | 1.41E-16 | 0.001 | 0.001 |
| **L3** | right-right-left | | | -0.8155 | -3.175 | 5.53E-11 | -2.43E-10 |
| **L3** | right-right-right | | | 0.0011 | 0.001 | 0 | 0 |

| | | doornumber | | carbody_convertible | | carbody_hatchback | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -86.852 | -1284.231 | 3043.114 | 392.806 | -749.5364 | -1968.261 |
| **L1** | left | -93.553 | -1087.931 | 4107.62 | 1604.264 | -196.5993 | -1223.059 |
| **L1** | right | 4213.4034 | -5856.687 | 4260.8892 | -4.13E+03 | -2.90E-12 | -7.18E-12 |
| **L2** | left -left | -234.5092 | -965.588 | 1820.8278 | -734.056 | 159.8933 | -580.083 |
| **L2** | left-right | 1502.2533 | -900.761 | 584.4795 | -3157.984 | 1172.987 | -1042.516 |
| **L2** | right-left | 394.9986 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 192.6502 | -143.457 | 2023.4232 | -2496.827 | -1.25E-12 | -3.02E-12 |
| **L3** | left-left-left | -85.9069 | -808.159 | 9.12E-10 | 1.41E-10 | -378.5542 | -1064.087 |
| **L3** | left-left-right | -108.5202 | -1659.588 | 1325.2743 | -3383.154 | 253.1355 | -891.843 |
| **L3** | left-right-left | -3730.075 | 430.0559 | -28300 | -128.9648 | -2883.39 | 711.9237 |
| **L3** | left-right-right | 5547.949 | -999.2193 | -3622.296 | 1481.6391 | -7987.508 | -913.3701 |
| **L3** | right-left-left | 108.0787 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | -25.9032 | -303.463 | 1021.4458 | -2725.85 | 4.80E-13 | -1.55E-12 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 4.5**. (Continued) Car tree levels

| | | carbody_wagon | | carbody_hardtop | | drivewheel_rwd | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -1791.5757 | -3089.047 | -788.356 | -3041.9 | 1350.0475 | -104.04 |
| **L1** | left | -1106.235 | -2239.801 | -1228.6257 | -3366.078 | 1246.5296 | 16.373 |
| **L1** | right | 2714.9204 | -1369.204 | 770.3041 | -4877.755 | -111.8882 | -2.60E+02 |
| **L2** | left -left | -910.3977 | -2017.096 | -1131.6638 | -2984.395 | -69.159 | -1773.894 |
| **L2** | left-right | -2533.756 | -4932.208 | -255.2132 | -3522.68 | 2452.2335 | -506.326 |
| **L2** | right-left | -116.9748 | nan | 394.9986 | nan | -0.0002 | nan |
| **L2** | right-right | -2.95E-14 | -2.88E-13 | -1696.9626 | -6222.563 | -15.1329 | -36.313 |
| **L3** | left-left-left | -624.6154 | -1587.603 | 641.9672 | -898.647 | -348.9154 | -2064.873 |
| **L3** | left-left-right | -1128.195 | -4164.855 | 2609.2984 | -31700 | -5853.684 | -30200 |
| **L3** | left-right-left | -17600 | -143.9955 | -16500 | -263.8799 | -25700 | 49.1911 |
| **L3** | left-right-right | -2506.606 | 4.48E-10 | -4.06E+03 | -1898.459 | -4.16E-10 | -6380.1099 |
| **L3** | right-left-left | 9.1756 | nan | 108.0787 | nan | 0.0052 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0.001 | 0.001 |
| **L3** | right-right-left | -3.63E-14 | -1.38E-13 | -1092.027 | -5195.781 | -0.8155 | -3.175 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0.0011 | 0.001 |

| | | drivewheel_4wd | | enginelocation | | wheelbase | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -731.0067 | -2869.784 | 1.19E+04 | 6.60E+03 | 57.1853 | -124.648 |
| **L1** | left | -490.6882 | -2320.555 | 1.67E-11 | 4.09E-12 | -5.1106 | -173.763 |
| **L1** | right | -4.19E-13 | -1.01E-12 | 2567.5899 | -1549.571 | 977.9964 | -1406.533 |
| **L2** | left -left | 231.0134 | -1291.11 | -2.91E-10 | -4.30E-10 | -51.056 | -253.119 |
| **L2** | left-right | -6292.2788 | -13600 | 3.43E-12 | -1.29E-10 | 227.4835 | -94.732 |
| **L2** | right-left | 0 | nan | 0 | nan | -525.8807 | nan |
| **L2** | right-right | -7.19E-28 | -3.55E-27 | 403.7947 | -252.585 | 1644.1829 | -877.024 |
| **L3** | left-left-left | -673.9578 | -2383.087 | -6.41E-10 | -1.18E-09 | -155.0577 | -382.679 |
| **L3** | left-left-right | -1851.8687 | -7798.93 | 2.27E-09 | -7.52E-09 | -147.7018 | -1369.212 |
| **L3** | left-right-left | -13900 | -1.48E-13 | -2.38E+03 | 96.2658 | -3.6E-12 | -849.9295 |
| **L3** | left-right-right | -8927.519 | -4.90E-10 | -1.76E+04 | 266.2375 | -1.5E-09 | -203.7673 |
| **L3** | right-left-left | 0 | nan | 0 | nan | -356.0912 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0.1099 | 0.11 |
| **L3** | right-right-left | 0 | 0 | -116.5857 | -496.583 | -605.4966 | -1916.764 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0.1389 | 0.139 |

**Table 4.5. (Continued)** Car tree levels

| | | carlength | | carwidth | | carheight | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -54.3584 | -162.939 | 751.5903 | 249.42 | 162.6435 | -99.682 |
| **L1** | left | 22.9428 | -75.77 | 549.3004 | 105.834 | 61.7689 | -159.713 |
| **L1** | right | 10.5319 | -1893.787 | -845.0754 | -3776.602 | -638.3781 | -3082.525 |
| **L2** | left -left | -47.4166 | -142.937 | 392.7214 | -77.838 | 44.113 | -151.468 |
| **L2** | left-right | -233.4137 | -550.806 | 327.8398 | -816.896 | 548.0507 | -247.563 |
| **L2** | right-left | 281.6733 | nan | 194.3936 | nan | -917.1285 | nan |
| **L2** | right-right | -1293.6943 | -3256.801 | 867.5925 | -1073.861 | -129.6007 | -1546.093 |
| **L3** | left-left-left | -63.0415 | -158.699 | 678.0375 | 32.33 | 100.4006 | -149.108 |
| **L3** | left-left-right | -21.4448 | -331.319 | -274.8241 | -1045.018 | 99.5152 | -865.599 |
| **L3** | left-right-left | -3810.652 | 298.9207 | -9588.841 | 917.6177 | -16300 | 9.5706 |
| **L3** | left-right-right | -317.941 | 811.0261 | -674.892 | -758.2084 | -1043.106 | 17.6382 |
| **L3** | right-left-left | -366.4809 | nan | 0.3634 | nan | -152.4476 | nan |
| **L3** | right-left-right | 0.1926 | 0.193 | 0.0681 | 0.068 | 0.0535 | 0.054 |
| **L3** | right-right-left | -81.7602 | -770.614 | -55.9286 | -1982.126 | 332.8347 | -1153.299 |
| **L3** | right-right-right | 0.2391 | 0.239 | 0.0824 | 0.082 | 0.0651 | 0.065 |

| | | curbweight | | enginetype_dohc | | enginetype_ohcv | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 6.243 | 1.93 | -2026.68 | -3929.219 | -6655.9592 | -9258.684 |
| **L1** | left | 4.4755 | -0.124 | -2488.2727 | -4508.198 | -3543.7643 | -6216.472 |
| **L1** | right | 5.8958 | -18.003 | -1551.9798 | -3647.075 | 1304.7855 | -1511.97 |
| **L2** | left -left | 10.3422 | 4.904 | -5628.0215 | -8436.912 | -1.10E-10 | -1.62E-10 |
| **L2** | left-right | 17.3888 | 3.809 | 4786.5652 | 148.398 | 2291.4631 | -2897.533 |
| **L2** | right-left | 20.556 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 0.6731 | -18.274 | -99.8057 | -492.788 | 109.7794 | -111.522 |
| **L3** | left-left-left | 11.0741 | 2.595 | -5.78E-11 | -1.05E-10 | -7.11E-11 | -1.29E-10 |
| **L3** | left-left-right | 8.6819 | -8.906 | -6935.1668 | -18900 | -2.54E-09 | -1.34E-08 |
| **L3** | left-right-left | -17700 | 430.0559 | -224.676 | 0 | -2883.39 | 0 |
| **L3** | left-right-right | -2267.352 | 4268.616 | 1.268 | 1898.6449 | -10500 | -7.68E-11 |
| **L3** | right-left-left | 12.8605 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 3.5831 | 3.583 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | 37.1938 | -23.878 | -169.9568 | -591.275 | 90.6825 | -121.993 |
| **L3** | right-right-right | 4.4802 | 4.48 | 0 | 0 | 0.0011 | 0.001 |

**Table 4.5. (Continued)** Car tree levels

| | | enginetype_dohcv | | enginetype_l | | enginetype_rotor | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 5.64E-12 | -1.69E-11 | -5468.7979 | -7975.613 | -1324.363 | -7688.054 |
| **L1** | left | 1.33E-11 | -1.97E-13 | -2360.0264 | -4843.436 | -7017.3654 | -12600 |
| **L1** | right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L2** | left -left | 1.36E-10 | 7.25E-11 | 5.41E-11 | 2.97E-11 | -2.80E+04 | -4.05E+04 |
| **L2** | left-right | 1.21E-11 | -6.70E-11 | -1.14E+04 | -1.77E+04 | 1.93E-12 | -1.03E-11 |
| **L2** | right-left | 0 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | left-left-left | 1.48E-12 | -5.77E-12 | 4.65E-11 | 6.51E-12 | -3.03E-11 | -5.56E-11 |
| **L3** | left-left-right | 6.43E-10 | -2.02E-09 | 1.94E-09 | -6.33E-09 | -6.45E+04 | -3.23E+05 |
| **L3** | left-right-left | 0 | 0 | 0 | 0 | 0 | 49.1911 |
| **L3** | left-right-right | -1.18E+04 | -1.56E+04 | -2.51E-10 | -1.25E-11 | -2.67E+04 | 1.12E-11 |
| **L3** | right-left-left | 0 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

| | | enginetype_ohcf | | cylindernumber_four | | cylindernumber_six | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -489.8347 | -3190.855 | -8002.9811 | -12400 | -4920.6476 | -8361.294 |
| **L1** | left | -2235.3223 | -4586.517 | -7452.2235 | -11600 | -1895.4968 | -6383.854 |
| **L1** | right | 2567.5899 | -1549.571 | 0 | 0 | -1439.6128 | -4172.147 |
| **L2** | left -left | -5868.8819 | -8809.511 | -1.67E+04 | -2.63E+04 | -2.79E-12 | -5.27E-12 |
| **L2** | left-right | -2143.8896 | -13900 | -3383.7902 | -17400 | 3195.492 | -11200 |
| **L2** | right-left | 0 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 403.7947 | -252.585 | 0 | 0 | -124.9123 | -354.465 |
| **L3** | left-left-left | -760.6781 | -3343.044 | -4.31E+04 | -7.89E+04 | 2.98E-11 | 5.14E-12 |
| **L3** | left-left-right | -1.77E+04 | -5.20E+04 | -3.40E+04 | -2.12E+05 | -6.97E-11 | -3.62E-10 |
| **L3** | left-right-left | 0 | 213.9394 | -2376.236 | 0 | -2926.433 | -211.7359 |
| **L3** | left-right-right | -4.25E-11 | 1.13E+04 | -7.69E-12 | 2.25E+04 | -2.51E+04 | 4361.5493 |
| **L3** | right-left-left | 0 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | -116.5857 | -496.583 | 0 | 0 | -91.4981 | -303.89 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 4.5. (Continued)** Car tree levels

| | | cylindernumber_five | | cylindernumber_three | | cylindernumber_twelve | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -7878.6331 | -12700 | 6.14E-13 | -7.37E-12 | -2173.0198 | -11400 |
| **L1** | left | -5187.0432 | -10100 | -4.71E-12 | -9.57E-12 | 4.48E-12 | -2.04E-12 |
| **L1** | right | 22.9391 | -261.985 | 0 | 0 | -817.7899 | -1866.576 |
| **L2** | left -left | -2.64E-11 | -3.77E-11 | -1.35E-11 | -2.08E-11 | -2.13E-12 | -3.05E-12 |
| **L2** | left-right | 460.0609 | -13800 | -6.28E-13 | -6.68E-12 | -4.06E-12 | -2.86E-11 |
| **L2** | right-left | -0.0002 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 0 | 0 | 0 | 0 | -133.8104 | -351.038 |
| **L3** | left-left-left | 2.05E-11 | 3.53E-12 | 0 | 0 | 0 | 0 |
| **L3** | left-left-right | -5.15E-10 | -2.74E-09 | -1.29E-10 | -6.97E-10 | -2.33E-10 | -1.25E-09 |
| **L3** | left-right-left | 0 | 0 | -2882.487 | 0 | 0 | 0 |
| **L3** | left-right-right | -10500 | 5.94E-12 | -2.67E+04 | -7.52E-12 | -5.77E-12 | 1.38E-11 |
| **L3** | right-left-left | 0.0052 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 0.001 | 0.001 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | 0 | 0 | 0 | 0 | 44.678 | -65.93 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

| | | cylindernumber_eight | | enginesize | | fuelsystem_spdi | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 973.4391 | -4803.764 | 78.4111 | 22.697 | -2307.7933 | -4534.878 |
| **L1** | left | -1.41E-12 | -3.36E-12 | -45.3831 | -107.441 | -980.5384 | -2884.998 |
| **L1** | right | 2122.5754 | -1515.161 | -85.9715 | -388.122 | 0 | 0 |
| **L2** | left -left | 5.26E-12 | 1.83E-12 | -299.5302 | -441.475 | -1432.8464 | -3021.881 |
| **L2** | left-right | -5.31E-12 | -2.36E-11 | -33.1789 | -141.776 | -5019.2744 | -1.29E+04 |
| **L2** | right-left | 0 | nan | -0.0432 | nan | 0 | nan |
| **L2** | right-right | 243.5898 | -63.005 | 72.1313 | -175.327 | 0 | 0 |
| **L3** | left-left-left | 0 | 0 | -157.4424 | -371.391 | 0 | 0 |
| **L3** | left-left-right | -9.41E-12 | -4.19E-11 | -891.826 | -3779.729 | -430.6148 | -2450.861 |
| **L3** | left-right-left | 0 | -84.5759 | | 0 | 0 | 0 |
| **L3** | left-right-right | -2.18E-11 | -98.6434 | | -1.45E-11 | -1.45E-11 | -1.82E+04 |
| **L3** | right-left-left | 0 | nan | 0.9459 | nan | 0 | nan |
| **L3** | right-left-right | 0 | 0 | 0.1739 | 0.174 | 0 | 0 |
| **L3** | right-right-left | 46.0045 | -82.733 | -372.0996 | -1045.403 | 0 | 0 |
| **L3** | right-right-right | 0.0011 | 0.001 | 0.3538 | 0.354 | 0 | 0 |

**Table 4.5. (Continued)** Car tree levels

| | | fuelsystem_2bbl | | fuelsystem_mfi | | fuelsystem_1bbl | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | 292.334 | -907.995 | -2423.3222 | -6942.285 | 639.6163 | -1376.988 |
| **L1** | left | -499.2554 | -1529.466 | -6.5813 | -3800.603 | 951.0981 | -739.433 |
| **L1** | right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L2** | left -left | -860.4205 | -1819.418 | 0 | 0 | -166.5711 | -1601.067 |
| **L2** | left-right | 1.42E-12 | -2.26E-12 | -4551.1538 | -12400 | 0 | 0 |
| **L2** | right-left | 0 | nan | 0 | nan | 0 | nan |
| **L2** | right-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | left-left-left | -7796.8411 | -15100 | 0 | 0 | -7921.5671 | -15700 |
| **L3** | left-left-right | -234.2888 | -3913.645 | 0 | 0 | 1515.0197 | -3092.627 |
| **L3** | left-right-left | -1309.754 | 0 | 0 | -770.4929 | 0 | 0 |
| **L3** | left-right-right | -618.703 | 0 | -4.11E+04 | 0 | 0 | 0 |
| **L3** | right-left-left | 0 | nan | 0 | nan | 0 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-left | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

| | | fuelsystem_spfi | | fuelsystem_4bbl | | fuelsystem_idi | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -800.2546 | -5247.937 | -2006.406 | -6892.619 | -4960.3383 | -19300 |
| **L1** | left | -1639.5491 | -5424.564 | -2302.6585 | -6.36E+03 | -1.12E+04 | -22900 |
| **L1** | right | 0 | 0 | 0 | 0 | 22.9391 | -261.985 |
| **L2** | left -left | 0 | 0 | -2748.6081 | -6.22E+03 | -2.84E+04 | -4.25E+04 |
| **L2** | left-right | -8895.4236 | -16200 | 0 | 0 | 8087.7881 | -24200 |
| **L2** | right-left | 0 | nan | 0 | nan | -0.0002 | nan |
| **L2** | right-right | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | left-left-left | 0 | 0 | 0 | 0.00E+00 | -2.74E+04 | -51500 |
| **L3** | left-left-right | 0 | 0 | -6471.2126 | -36800 | 0 | 0.00E+00 |
| **L3** | left-right-left | 0 | 111.0874 | 0 | -1320.5134 | -20800 | -1406.5093 |
| **L3** | left-right-right | -40500 | 3.37E+04 | 0 | -2961.5599 | 0.00E+00 | 597.4426 |
| **L3** | right-left-left | 0 | nan | 0 | nan | 0.0052 | nan |
| **L3** | right-left-right | 0 | 0 | 0 | 0 | 0.001 | 0.001 |
| **L3** | right-right-left | 0 | 0 | 0 | 0 | 0 | 0 |
| **L3** | right-right-right | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 4.5. (Continued)** Car tree levels

| | | boreratio | | stroke | | compressionratio | |
|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | #REF! | -5943.338 | -3278.7398 | -5103.283 | -980.948 | -2143.718 |
| **L1** | left | 1579.1461 | -1389.106 | -2004.5914 | -3607.28 | 85.8556 | -934.236 |
| **L1** | right | 444.4062 | -132.118 | -2330.8356 | -6051.728 | 151.1878 | -414.522 |
| **L2** | left -left | 1.69E+04 | 8962.983 | 3454.9781 | -1078.973 | 855.4403 | -231.578 |
| **L2** | left-right | 660.6602 | -3974.427 | -1923.0281 | -4622.238 | -2246.6435 | -5290.604 |
| **L2** | right-left | -0.0008 | nan | -0.0009 | nan | -0.0051 | nan |
| **L2** | right-right | 73.5208 | -49.45 | -233.1178 | -582.149 | -16.8882 | -273.039 |
| **L3** | left-left-left | 7159.3427 | -797.602 | 5602.285 | 100.566 | 128.1422 | -1396.289 |
| **L3** | left-left-right | 6.01E+04 | -1.51E+05 | 1.36E+04 | -70800 | 2347.3194 | -56.858 |
| **L3** | left-right-left | 0 | 597.9757 | -1468.305 | 202.2434 | -20200 | 7.6599 |
| **L3** | left-right-right | 0 | -3871.0427 | -30200 | -393.4773 | -21900 | 0.0787 |
| **L3** | right-left-left | 0.0185 | nan | 0.0188 | nan | 0.1111 | nan |
| **L3** | right-left-right | 0.0034 | 0.003 | 0.0035 | 0.003 | 0.0204 | 0.02 |
| **L3** | right-right-left | -29.577 | -97.732 | -119.6923 | -380.158 | -28.2241 | -288.703 |
| **L3** | right-right-right | 0.0044 | 0.004 | 0.0038 | 0.004 | 0.0092 | 0.009 |

| | | horsepower | | peakrpm | | citympg | | highwaympg | |
|---|---|---|---|---|---|---|---|---|---|
| | | coefficient | p-value | coefficient | p-value | coefficient | p-value | coefficient | p-value |
| **L0** | | -3.2812 | -53.112 | 1.7277 | 0.447 | -82.2907 | -388.421 | 140.6467 | -138.417 |
| **L1** | left | 28.0497 | -20.703 | 0.0172 | -1.133 | -167.8874 | -423.424 | 119.3761 | -113.325 |
| **L1** | right | 12.8674 | -282.976 | 8.5328 | -8.411 | -611.2595 | -2516.258 | -604.7472 | -1831.936 |
| **L2** | left -left | 74.3598 | 2.786 | 0.2836 | -0.943 | 5.3657 | -212.006 | 11.7114 | -179.28 |
| **L2** | left-right | -227.6085 | -402.097 | -1.7922 | -6.578 | -449.1268 | -1812.902 | 548.0866 | -208.635 |
| **L2** | right-left | -0.029 | nan | -1.026 | nan | -0.0052 | nan | -0.0059 | nan |
| **L2** | right-right | -35.4456 | -300.64 | 13.1706 | -4.731 | -516.3765 | -1201.262 | -844.1453 | -2052.247 |
| **L3** | left-left-left | 41.1551 | -76.019 | -0.6167 | -2.314 | 4.499 | -167.142 | 47.1414 | -87.565 |
| **L3** | left-left-right | 156.9843 | 1.703 | -2.4577 | -9.251 | 417.771 | -1974.412 | -635.2474 | -1898.049 |
| **L3** | left-right-left | -23100 | -888.8245 | -5787.936 | 1208.5534 | -2071.547 | -8229.486 | -118.589 | -10100 |
| **L3** | left-right-right | -3196.428 | 568.6304 | -8492.935 | 34.9107 | -738.111 | -1426.238 | -10.901 | -927.84 |
| **L3** | right-left-left | 0.6358 | nan | 22.4843 | nan | 0.1137 | nan | 0.1292 | nan |
| **L3** | right-left-right | 0.1169 | 0.117 | 4.1343 | 4.134 | 0.0209 | 0.021 | 0.0238 | 0.024 |
| **L3** | right-right-left | 288.7903 | -266.989 | 1.0925 | -12.075 | -69.98 | -359.32 | -217.7739 | -1082.31 |
| **L3** | right-right-right | 0.2114 | 0.211 | 5.1695 | 5.169 | 0.0161 | 0.016 | 0.0184 | 0.018 |

In Table 4.6, it can be seen that the value of the regression coefficient and the p-value (-23300,-43000) respectively, gave a negative value where the negative value in the regression coefficient indicates a decrease in the rate of the dependent variable and the negative value of p-value indicates that there is a negative relationship between the dependent and the independent variable that is they go in opposite directions and also for the case of variable fueltype while with the variable aspiration and it's find that the regression coefficient gave a value of (1904.2679) while the value of p-value is negative (-193.332) there is a significant effect between the independent variable and the dependent variable but in opposite directions It can also be noted that in.The other levels L1, L2, and L3 gave the same results, although there were some values in some variables that were not applicable.

In the advertising data set and the fish data set, the results were for the values of the multiple linear regression model as well as the regression models after dividing the data into three levels that were decreasing or increasing depending on the level in which it was and some values were equal to zero for the coefficients of the independent variables, meaning there is no effect on the dependent variable, and some of them are not applicable as well As for the values of p, it can be observed that the same results for the car data set gave the same result. As in Tables 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12.

**Table 4.6.** Car results

| | MAPE | | RMSE | | R2 | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| L0 | 0.112945 | 0.165514 | 1967.917 | 15140.32 | 0.93945 | 0.887091 |
| L1 | 0.093532 | 0.150159 | 1562.944 | 2814.508 | 0.959427 | 0.903426 |
| L2 | 0.09231 | 0.305106 | 1681.753 | 6448.438 | 0.954849 | 0.569338 |
| L3 | 0.079998 | 0.538321 | 1800.315 | 13054.29 | 0.956967 | 0.153827 |
| Tree | 0.617766 | 0.584156 | 9826.857 | 11765.35 | 0.015332 | 0.00907 |
| Tree Full | 0.633201 | 0.581689 | 9989.75 | 11778.57 | 0.01431 | 0.019246 |

The train MAPE is in Table 4.13 gave the best result at the third level (L3), while the test MAPE gave the best result at the first level(L1). It can be seen from the table mentioned earlier that train, Test R-Square, and RMSE have the best result in the first level(L1). Therefore, it cannot be considered that the third level (L3) gives the best model. Consequently, must use more than one measure to obtain an optimal model.

## 5. CONCLUSION

Decision-making and prediction of events are an integral part of our lives. Regression trees are one of the common ways of arranging our decisions and are considered one of the machine learning methods. The multiple linear regression model algorithm is also an important method in prediction. In this research, a process of combining the regression tree algorithm and the multiple linear regression algorithm has been proposed to form multiple linear regression models from the regression tree, and it was applied to three types of data. In the first data set, it is clear that, depending on the third level (L3), it was found that it gave the best result by reducing the error. Unlike the second data set, it gives the least error in train at the third level (L3), while in the test it was not enough to give the best result at the third level (L3). In this case, the optimal level must be found L1, L2, L3...etc. This depends on the problem under study. In addition, it is not possible to rely on one measure to choose the best result, but other measures of error must be added to obtain the best result, i.e. the best model. It can also be noted that the third data set gave the best result at the first level(L1), but not the only one, as the third level (L3) also gave the best result.It can be said that splitting the data set is a good method because at least in the three data sets L1, L2, L3, it gave the best results compared to L0 or the full tree. However, using the multiple linear regression model alone does not give the best result, but can add a Classification and Regression Tree to divide the data set and find the best result from the multiple linear regression model and the full tree.

## RESOURCES


[1] Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A., "Overview of use of decision tree algorithms in machine learning," In 2011 IEEE control and system graduate research colloquium (pp. 37-42). IEEE, 2011

[2] Izza, Y., Ignatiev, A., & Marques-Silva, J.," On explaining decision trees," arXiv preprint arXiv:2010.11034, 2020

[3] Dufour, D., "Finding cost-efficient decision trees," (Master's thesis, University of Waterloo),2014

[4] Song, Y. Y., & Ying, L. U., "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, 27(2), 130, 2015

[5] Shi, Y., Li, J., & Li, Z., "Gradient boosting with piece-wise linear regression trees, " arXiv preprint arXiv:1802.05640, 2018

[6] White, D., & Sifneos, J. C., "Regression tree cartography," Journal of Computational and Graphical Statistics, 11(3), 600-614,2002

[7] Loh, W. Y., "Regression tree models for designed experiments," In Optimality (pp. 210-228). Institute of Mathematical Statistics,2006

[8] Steinberg, D., "CART: classification and regression trees," In The top ten algorithms in data mining (pp. 193-216), Chapman and Hall/CRC, 2009

[9] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J., "Classification and regression trees," Belmont, CA: Wadsworth. International Group, 432, 151-166,1984

[10] Gray, J. B., & Fan, G., "Classification tree analysis using TARGET," Computational Statistics & Data Analysis, 52(3), 1362-1372,2008

[11] Loh, W. Y., "Classification and regression tree methods," Encyclopedia of statistics in quality and reliability, 1, 315-323,2008

[12] Izzah, A., Sari, Y. A., Widyastuti, R., & Cinderatama, T. A.," Mobile app for stock prediction using Improved Multiple Linear Regression," In 2017 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 150-154). IEEE,2017

[13] Shyti, B., & Valera, D.," The regression model for the statistical analysis of Albanian economy," Int. J. Math. Trends Technol, 62(2), 90-96,2018

[14] Pitombo, C.S., Sousa, A., & Filipe, L.N.," Classification and egression tree, principal components analysis and multiple linear regression to summarize data and understand travel behavior," Transportation Letters, 1, 295 – 308,2009

[15] Machado, M.D., Tommaselli, A.M., Tachibana, V.M., Martins-Neto, R.P., & Campos, M.B., "Evaluation of multiple linear regression model to obtain DBH of trees using data from a lightweight laser scanning system ON-BOARD a UAV," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,2019

[16] Leblanc, M., "Regression trees," Wiley StatsRef: Statistics Reference Online,2014

[17] Ashwin P., "Regression Trees | Decision Tree for Regression | Machine Learning," Retrieved from Analytics Vidhya ,website: https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047,Aug.8.2021

[18] Naghibi, S. A., Pourghasemi, H. R., & Dixon, B., "GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran," Environmental monitoring and assessment, 188(1), 1-27,2016

[19] Alamgir, M. S. M., Sultana, M. N., & Chang, K., "Link adaptation on an underwater communications network using machine learning algorithms: Boosted regression tree approach," IEEE access, 8, 73957-73971,2020

[20] Portillo-Salgado, R., Cigarroa-Vázquez, F. A., Ruiz-Sesma, B., Mendoza-Nazar, P., Hernández-Marín, A., Esponda-Hernández, W., & Bautista-Ortega, J.,"Prediction of Egg Weight from External Egg Traits of Guinea Fowl Using Multiple Linear Regression and Regression Tree Methods," Brazilian Journal of Poultry Science, 23,2021

[21] Fan, G., & Gray, J. B., "Regression tree analysis using TARGET," Journal of Computational and Graphical Statistics, 14(1), 206-218,2005

[22] Sathyadevi, G., "Application of CART algorithm in hepatitis disease diagnosis," In 2011 International Conference on Recent Trends in Information Technology (ICRTIT) (pp. 1283-1287). IEEE,2011

[23] Santhanam, T., & Sundaram, S., "Application of CART algorithm in blood donors classification," Journal of computer Science, 6(5), 548,2010

[24] Bel, L., Allard, D., Laurent, J. M., Cheddadi, R., & Bar-Hen, A, "CART algorithm for spatial data: Application to environmental and ecological data. Computational Statistics & Data Analysis," 53(8), 3082-3093,2009

[25] Li, M., "Application of CART decision tree combined with PCA algorithm in intrusion detection," In 2017 8th IEEE international conference on software engineering and service science (ICSESS) (pp. 38-41). IEEE,2017

[26] Boerstler, H., & de Figueiredo, J. M., "Prediction of use of psychiatric services: Application of the CART algorithm," The journal of mental health administration, 18(1), 27-34,1991

[27] Friedman, J. H., & Meulman, J. J., "Multiple additive regression trees with application in epidemiology," Statistics in medicine, 22(9), 1365-1381,2003

[28] Faraggi, D., LeBlanc, M., & Crowley, J.," Understanding neural networks using regression trees: an application to multiple myeloma survival data," Statistics in medicine, 20(19), 2965-2976.,2001

[29] Wang, X., Montgomery, D. C., & Owusu-Antwi, E. B., "Application of regression trees to LTPP data analysis," In International Conference on Highway Pavement Data, Analysis and Mechanistic Design Applications, 2003, Columbus, Ohio, USA,2003

[30] Moustris, K. P., Nastos, P. T., Larissi, I. K., & Paliatsos, A. G., "Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area, Greece," Advances in Meteorology, 2012.

[31] Piekutowska, M., Niedbała, G., Piskier, T., Lenartowicz, T., Pilarski, K., Wojciechowski, T.,& Czechowska-Kosacka, A.,"The application of multiple linear regression and artificial neural network models for yield prediction of very early potato cultivars before harvest," Agronomy, 11(5), 885,2021

[32] Eberly, L. E., "Multiple linear regression," Topics in Biostatistics, 165-187,2007

[33] Rivera, R., "Principles of managerial statistics and data science," John Wiley & Sons,2020

[34] Tranmer, M., & Elliot, M., "Multiple linear regression," The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5,2008

[35] M.R. Braun, H. Altan, S.B.M. Beck, "Using regression analysis to predict the future energy consumption of a supermarket in the UK," Applied Energy, vol. 130, pp. 305--313, 2014.

[36] Hayes, "How Multiple Linear Regression Works," 2021

[37] Srinidhi and S. (n.d.), "Backward Elimination for Feature Selection in Machine Learning. Medium,"Retrieved from Towards Data Science, website: https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4,Nov.15.2019

[38] Seb,"The Coefficient of Determination and Linear Regression Assumptions | classical Machine Learning | Machine Learning," Retrieved from Programmathically, website: https://programmathically.com/the-coefficient-of-determination-and-linear-regression assumptions/?utm_source=rss&utm_medium=rss&utm_campaign=the-coefficient-of-determination-and-linear-regression-assumptions, Aprl.26.2021

[39] Statistics Dictionary," Coefficient of Multiple Determination," Retrieved from Stat Trek, website:https://stattrek.com/statistics/dictionary.aspx?definition= coefficient-of-multiple-determination,2022

[40] Ashish, "Advertising Dataset," Retrieved from Kaggle, website: https://www.Kaggle.com/datasets ashydv/advertising -dataset, 2019

[41] Aung Pyae, "Fish market," Retrieved from Kaggle, website:https://www.kaggle.com/datasets/aungpyaeap/fish-market,sep.9 2019

[42] Nehal B., Nishant V., Nikhil K. , "Vehicle dataset," Retrieved from kaggle website https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho, Nov.4.2020

[43] ZACH, "How to Calculate Mean Absolute Percentage Error (MAPE) in Excel," Retrieved from statiscs.simplified, statology, website: https://www.statology.org/mape-excel/ , Fep.27. 2020

[44] [44] Frank C., " Scikit-Learn Decision Trees Explained," Retrieved from Towards Data Science, website: https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d , Feb.23.2019

[45] IBM, "Maximum tree depth, " Retrieved from IBM, website: https://www.ibm.com/docs/en/db2/9.7?topic=classification-maximum-tree-depth,march.1.2021

[46] Karan P., "A Comprehensive Guide to Decision trees," Retrieved from Analytics Vidhya, website: https://www.analyticsvidhya.com/blog/2021/07/a-comprehensive-guide-to-decision-trees/ ,Sep.26.2022

[47] Stev A., "sklearn min_impurity_decrease explanation," Retrieved from stack over flow, website: https://stackoverflow.com/questions/54812230/sklearn-min-impurity-decrease explanation, Jan.1.2020

[48] Mukesh M., " How to tune a Decision Tree," Retrieved from Towards Data Science , website: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680 ,Nov.12.2019

[49] George C., "Regression Tree vs Linear Regression," Retrieved from QUANTIFYING HEALTH, website: https://quantifyinghealth.com/regression-tree-vs-linear-regression/, Jan.2.2020.

[50] "Data scince-Regression Table:p-value," Retrieved from w3schools, website: https://www.w3schools.com/datascience/ac.asp, Sep.9.2023.

[51] Jim F., "How to Interpret P-values and Coefficients in Regression Analysi," Retrieved from statistic from jim, website: https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/, Jan.1.2018.

**CURRICULUM VITAE**

Name Surname :Maryam Arif AZEEZ

**EDUCATION:**

- **Undergraduate** : Graduation year, University, Faculty, Department2020-2023, Sakarya University, Computer and Information engineering.
- **Graduate** : 2014, Mosul University, Operation research and intelligence technique.

**PROFESSIONAL EXPERIENCE AND AWARDS:**

Photoshop designer, Life skills teacher, photographer, mathematic teacher and Letters of thanks from the school where I teach**.**