

(Heart disease Analysis)

Submitted by

#### Bavisetti V Venkata Sai Ganga Bulliraju

#### Register no: 12212206

#### INT234 & K22YP

**B.Tech(Computer Science And Engineering)**

Under the guidance of

**Madhu Bala**.

**School of Computer science & Engineering Lovely Professional University, Phagwara (Aug-Dec 2024)**

# DECLARATION

I hereby declare that I have completed project on the basis of heart disease sentiment analysis from **10- 10- 2024** to **20-11-2024** under the guidance of **Madhu Bala**. I have declare that I have worked with full dedication during the project time and my learning outcomes fulfil the requirements of training for the award of degree of Btech(CSE), Lovely Professional University, Phagwara.

Date: 20-11-2024

B V V S G Bulliraju Registration no: 12212206

# ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to everyone who supported and guided me throughout this project. First and foremost, I extend my deepest appreciation to my mentors and professors, whose knowledge and expertise provided the foundation for my work and inspired me to push beyond my limits. Their invaluable insights helped me navigate challenges and deepen my understanding of machine learning and data analysis. I am also thankful to my family and friends for their unwavering encouragement and support. Their motivation and belief in me fueled my efforts and kept me focused.

A special thanks to the open-source community and resources like Kaggle, which provided the car dataset and served as a great platform for data exploration and experimentation. I would also like to acknowledge the creators of R and the extensive resources available, which were instrumental in building and refining my analysis modelsLastly, I extend my gratitude to everyone involved in the creation of resources, books, and documentation, whose work has been foundational to my learning journey. Thank you all for making this project possible.

**TABLE OF CONTENTS**

INTRODUCTION 5-6

Scope of the Analysis 7

[Existing System 8](#_bookmark0)

ETL process 9

Analysis on dataset (for each analysis) 10-15

[Future scope 16](#_bookmark1)

[References 17](#_bookmark2)

[Bibliography 18](#_bookmark3)

* INTRODUCTION OF THE PROJECT UNDERTAKEN Decision Trees:-

Decision Trees are supervised machine learning algorithms used for classification and regression tasks. They divide data into subsets based on feature values, creating a tree-like structure where each decision point leads to a final prediction. Decision trees use measures such as the Gini Index or Information Gain to decide the best feature to split the data:

Gini Index: Measures impurity; the lower the Gini Index, the purer the nodes. Information Gain: Measures the reduction in entropy after a split.

* How Decision Trees Work:

Decision Trees are interpretable machine learning models that predict outcomes by breaking down data into smaller subsets based on feature values. Here’s how they function in the context of predicting heart diseaseFor each node in the tree, they assess possible splits and choose the one that yields the purest child nodes. This process is repeated until the tree reaches a stopping criterion, such as maximum depth or minimum samples per leaf.

 Advantages in Heart Disease Prediction

Interpretability: Clinicians can easily follow the tree to understand how predictions are made.

Feature Importance: Highlights the most critical factors for heart disease (e.g., high cholesterol, low heart rate).

 Linear Regression

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It’s widely used in predictive analytics to estimate continuous outcomes based on input data.

Understanding the Linear Regression Model:

* Linear regression finds the line that best fits the data points by minimizing the sum of squared differences between observed values and the predicted line.
* Simple linear regression models relationships using a single predictor variable, while multiple linear regression handles multiple variables, each contributing uniquely to the model’s predictions.

 Applications in Data Science and Beyond:

Linear regression is widely used in data science for various applications, particularly in health-related fields like heart disease prediction. By modeling the relationship between different health factors and heart disease outcomes, linear regression provides actionable insights for prevention, diagnosis, and treatment planning. While it’s a powerful tool for linearly correlated data, it may not capture nonlinear relationships effectively.

 Support Vector Machines (SVM)

Support Vector Machines (SVM) are powerful supervised learning algorithms used primarily for classification but also for regression tasks. SVM aims to create a hyperplane that best separates data points of different classes by maximizing the margin between them.

 Mechanics of SVM:

SVM operates by transforming the input data into a high-dimensional space, using techniques such as the “kernel trick” (e.g., linear, polynomial, and RBF kernels).

After mapping data to this space, SVM identifies the optimal hyperplane by maximizing the margin, effectively finding the widest possible boundary between classes.

 Advantages and Suitability of SVM:

SVM works exceptionally well when there is many features (high-dimensional data). In the context of heart disease prediction, clinical data often includes many variables (e.g., age, cholesterol levels, heart rate, blood pressure, etc.). SVM can efficiently handle these features and detect complex decision boundaries even when the number of dimensions (features) is much larger than the number of data points.

 K-Means Clustering

K-means clustering is an unsupervised learning algorithm that groups data points into K distinct clusters based on similarity. Unlike supervised algorithms, clustering does not rely on labeled data. It’s used in applications where understanding natural groupings or structure is valuable, such as market segmentation or image compression.

How K-Means Works:

* The algorithm starts by randomly initializing K centroids, then iteratively assigns data points to the closest centroid and recalculates centroid positions based on these assignments.
* The process repeats until centroids stabilize, minimizing the within-

 **Scope of the Work:**

The primary objective of this analysis is to apply machine learning algorithms to explore patterns and make predictions from the given dataset, ultimately gaining insights into underlying structures and relationships within the data. This analysis includes the following major elements:

1. **Dataset Exploration and Preprocessing: -**

Before applying machine learning models, a comprehensive exploration of the dataset will be conducted. This includes identifying and handling missing values, scaling or normalizing features, and encoding categorical variables where necessary. Proper preprocessing ensures that the dataset is ready for analysis and maximizes model performance.

### Application of Machine Learning Algorithms :-

The analysis focuses on four key machine learning algorithms: Decision Trees, Linear Regression, Support Vector Machines (SVM), and K-Means Clustering. Each of these algorithms is chosen based on its strengths for specific tasks:

Decision Trees:- for classifying and interpreting relationships between variables in a structured, visual format.

Linear Regression:- for understanding and quantifying linear relationships between continuous variables.

Support Vector Machines (SVM):- for enhanced accuracy in classification tasks, especially with clear margin separation.

K-Means Clustering:- to discover natural groupings within the data, helpful for segmentation and pattern recognition.

### Comparative Performance Evaluation

A key part of this analysis is to compare the performance of these algorithms, examining their predictive accuracy, interpretability, and computational efficiency. This will involve measuring and comparing metrics such as accuracy, precision, recall, and for regression models, Mean Squared Error (MSE).

### Identification of Key Patterns and Insights

Through the application of these algorithms, the analysis aims to uncover significant patterns and relationships within the dataset. For instance, clustering may reveal natural groupings or segments, while regression and classification techniques could highlight influential variables that impact the target variable.

### Limitations and Future Directions

Each machine learning method has inherent limitations, and these will be identified as part of the analysis. Additionally, potential improvements and alternative approaches for future analysis will be suggested, such as using other algorithms, more complex models, or additional data sources to further enhance the analysis.

## Existing System :-

#### Drawbacks or limitations of existing system:-

The existing system for analyzing datasets often relies on traditional statistical methods or single machine learning models that may provide limited insights. These methods may include basic regression analysis or simple classification approaches, often lacking a combination of diverse algorithms that could offer a broader view of patterns and predictions.

While traditional systems provide a foundation for data analysis, they often fall short when applied to complex datasets with high-dimensional and non-linear relationships. The inability to comprehensively analyze these data characteristics can lead to incomplete or misleading insights.

#### Drawbacks or Limitations of the Existing System

1. Limited Predictive Accuracy

Traditional statistical models or single-model machine learning approaches may fail to capture complex interactions between variables. This can result in lower predictive accuracy, particularly in datasets with high-dimensional and interdependent features. Models like simple linear regression or basic classifiers often overlook intricate relationships, leading to limited reliability in predictions.

1. Inflexibility with Non-linear Data

Many traditional models assume linear relationships between features and target variables, which restricts their applicability to non-linear or more complex data. For example, standard regression or linear classifiers may not perform well when dealing with curvilinear relationships or overlapping data points, resulting in suboptimal model performance.

1. Poor Scalability

As datasets grow in size and complexity, traditional systems can struggle to scale effectively. This limitation affects computational efficiency and slows down processing, making it challenging to analyze large datasets in a timely manner.

1. Lack of Interpretability

While simpler models like linear regression offer interpretability, they often lack the capability to provide in-depth, structured insights like those available in models such as decision trees. Conversely, more complex models may yield more accurate results but at the cost of transparency, making it difficult to explain predictions or interpret feature importance clearly.

1. Inadequate Handling of Multimodal Patterns

Many datasets contain multiple patterns or subgroups that a single model may not fully capture. Traditional systems often do not leverage clustering or other techniques to detect natural groupings, which can be particularly useful in identifying patterns within diverse populations or market segments.

1. Limited Feature Engineering and Data Preprocessing

Traditional analysis systems often lack advanced preprocessing techniques, which are critical for maximizing model performance. Insufficient handling of missing values, scaling issues, or categorical encoding can lead to biased results and lower model accuracy.

By addressing these limitations, the proposed system integrates diverse machine learning algorithms that enhance predictive accuracy, flexibility,. This approach facilitates a deeper understanding of complex datasets, leading to improved insights and decision-making.

## Source of Dataset

The dataset used in this analysis, specifically the **heart disease Dataset**, was sourced from **Kaggle**. Kaggle is a popular platform for data science and machine learning, offering a wide variety of datasets for academic, research, and practical applications.

#### ETL Process for the heart disease Dataset

1. Extract

Source: The car dataset provided (likely in CSV format). Loading Tool\*\*: Load the dataset into R (using `read.csv()`)

1. Transform

Here, we ensure data quality and prepare it for modeling:

* + Data Cleaning:

Handling Missing Values: Identify missing values and handle them by either imputing (replacing with means, medians, or modes as appropriate) or dropping rows/columns, depending on the extent and importance of the missing values.

Outlier Detection and Treatment: Detect outliers in numerical columns, such as price, mileage, or engine size, using Z-scores or the interquartile range (IQR) method, and decide whether to treat or remove them.

* + Data Normalization and Scaling:

For numerical features like price, mileage, and engine size, apply normalization or standardization to ensure all values contribute evenly during analysis, especially for algorithms sensitive to feature scaling (e.g., SVM or K-means clustering).

* Encoding Categorical Variables:

Many car datasets contain categorical data such as car make, model, color, fuel type, or transmission type. Use techniques like one-hot encoding or label encoding to transform these into numerical formats compatible with machine learning models.

* Feature Selection:

Identify key features relevant to the analysis goals (e.g., predicting car price or categorizing car types). Unimportant or redundant columns may be removed to streamline processing and model training.

* Data Splitting:

Split the cleaned and processed data into training and test sets (e.g., 80% training, 20% testing) to enable accurate model evaluation.

* Data Storage:

Save the final pre-processed data in formats suited for analysis (such as separate R data frames or CSV files).

Load the data directly into machine learning pipelines (e.g., `caret` in R or `scikit-learn` in Python) for model training and testing.

This ETL process will ensure that the car dataset is clean, consistent, and ready for effective analysis and modelling. Let me know if you'd like further customization or details on any specific step!

## Analysis on Dataset in R

#### Introduction

Using the car dataset, the analysis aims to apply various machine learning techniques in R to derive insights, such as predicting car prices, categorizing car types, and identifying significant attributes that impact car values.

#### General Description.

The car dataset includes a mix of categorical (e.g., make, model, transmission) and numerical data (e.g., price, mileage, engine size). It provides sufficient features to explore relationships and trends through supervised and unsupervised machine learning models.

#### Specific Requirements, Functions, and Formulas.

Here are the methods, functions, and formulas used in R for this analysis:

* 1. Decision Tree Analysis

Function: `rpart()` from the `rpart` package.

Purpose: To classify cars into categories, such as economy, luxury, or performance, based on features like price, engine size, and year.

Formula: The Gini impurity or entropy criterion in `rpart` determines the best split for classification.

* 1. Linear Regression:

Function: `lm()` for linear regression.

Purpose: To predict car prices based on numeric features like mileage, engine size, and year. Formula: \( \text{Price} = \beta\_0 + \beta\_1 \times \text{Mileage} + \beta\_2 \times \text{Engine Size} + \ldots + \beta\_n \times X\_n \).

* 1. Support Vector Machine (SVM):

Function: `svm()` from the `e1071` package.

Purpose: To classify car types (e.g., compact, SUV) based on attributes like fuel type, transmission, and engine size.

Kernel Options: Linear or radial basis functions (RBF) are popular choices in `svm()`.

* 1. K-Means Clustering:

Function: `kmeans()` in base R.

Purpose: To group cars into clusters based on features like price, mileage, and engine size, potentially identifying segments such as budget, premium, and performance-focused cars.

Formula: Minimize within-cluster variance using \( \sum\_{i=1}^{k} \sum\_{x \in C\_i} (x - \mu\_i)^2

\), where \( C\_i \) is a cluster and \( \mu\_i \) is its centroid.

#### Analysis Results

Each algorithm provides insights from a unique angle:

* 1. Decision Tree Analysis Results:

The decision tree can reveal influential variables for classifying car types or price ranges. Example: A decision tree might categorize cars based on criteria like engine size or year, clearly distinguishing economy from luxury vehicles.

* 1. Linear Regression Results:

Regression coefficients will show the contribution of each feature to the car’s price.

Example: Mileage may show a negative coefficient, suggesting a decrease in price as mileage increases.

* 1. SVM Classification Results:

The SVM may provide a high classification accuracy for car types, especially if there are clear distinctions between compact and larger cars.

Example: The SVM might achieve 85% accuracy in categorizing cars by type based on features like transmission and engine size.

* 1. K-Means Clustering Results:

The clusters can represent distinct groups within the data, such as "budget-friendly," "luxury," or "performance" vehicles.

Example: Three clusters might emerge, with one group representing low-cost, high-mileage cars and another capturing newer, high-priced cars.

#### Visualization

Visualizations enhance understanding of the analysis results in R:

* 1. Decision Tree:

Plot the decision tree using `rpart.plot()` from the `rpart.plot` package to visualize the tree structure and see which features drive classifications.

* 1. Linear Regression:

Use a scatter plot with `ggplot2` to show predicted vs. actual prices, including a linear regression line for model fit.

Plot residuals to check the distribution and any patterns in prediction errors.

* 1. SVM Decision Boundaries:

Use `ggplot2` to visualize SVM decision boundaries (for 2D data), showing different regions for classified car types based on relevant features.

* 1. K-Means Clustering:

Use `ggplot2` for a scatter plot that represents clusters, color-coding each group and showing centroids.

A 3D plot (using `plotly`) can show three dimensions, such as price, mileage, and engine size.

## 7. List of Analysis with results :-

#### Decision Tree Analysis

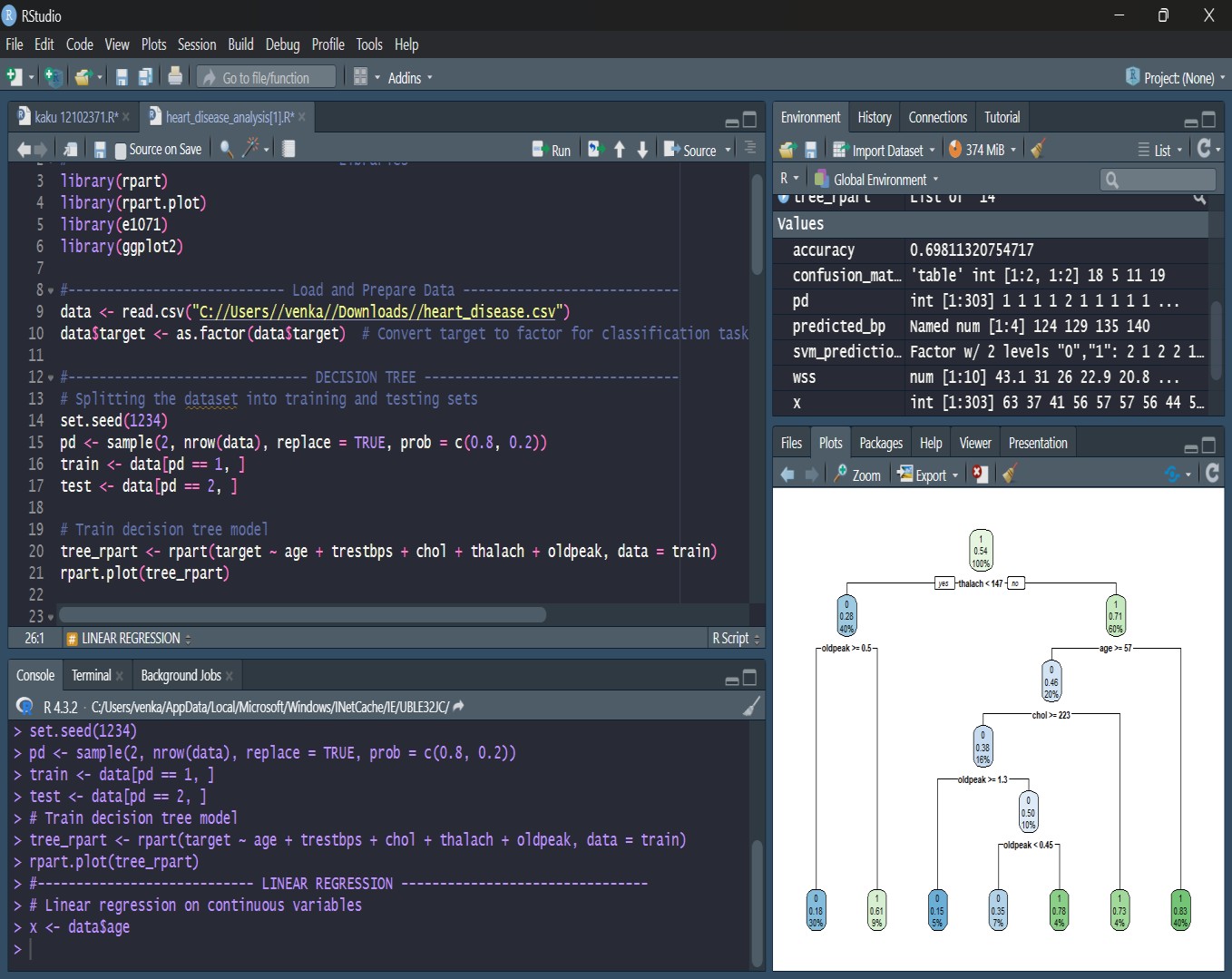
Objective:

A Decision Tree is a supervised learning algorithm used for classification tasks. Here, the goal is to predict whether a patient has heart disease (target) based on several features such as age, blood pressure, cholesterol, maximum heart rate, and oldpeak (exercise-induced angina).

Steps:

Splitting Data: The data is divided into training (80%) and testing (20%) sets using sample() function. Model Training: A Decision Tree model (rpart) is trained using the features age, trestbps, chol, thallic, and old peak to predict the target variable.

Visualization: The decision tree structure is plotted using rpart.plot().



Linear Regression Analysis

Linear regression models the relationship between a continuous dependent variable and one or more independent variables. Here, the goal is to predict the trestbps (resting blood pressure) based on age.

Steps:

Model Creation: A linear regression model is built using the lm() function where y = trestbps (dependent variable) and x = age (independent variable).

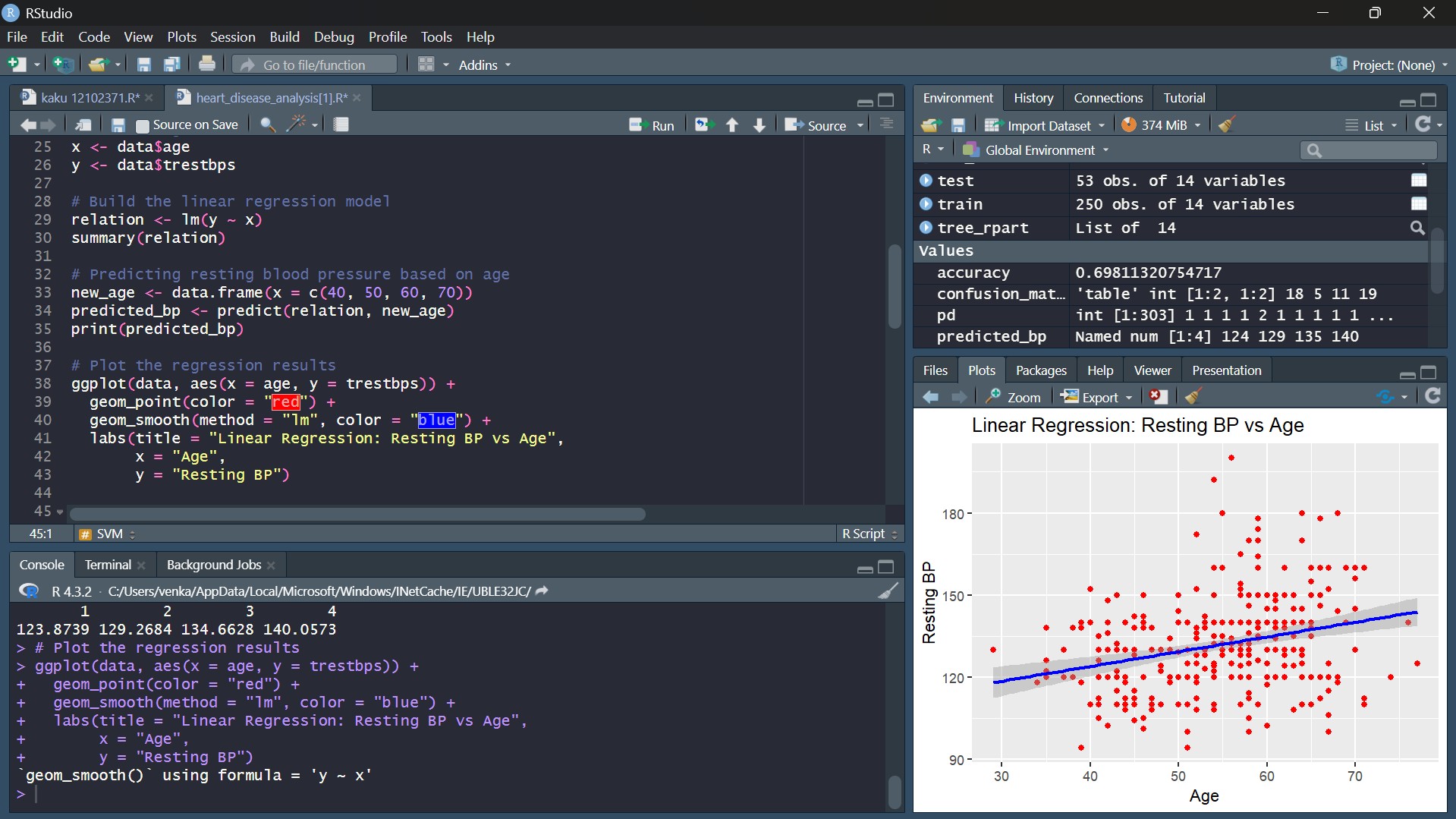
Model Summary: The summary of the regression model provides coefficients, p-values, and R-squared value.

Prediction: Predicted blood pressure values (predicted\_bp) are computed for ages 40, 50, 60, and 70.

Model Output: A regression equation will be derived. For instance, the output might show that for each year increase in age, resting blood pressure increases by a certain amount (e.g., 0.2 mm Hg).

Predictions: Predicted blood pressures for the specified ages (e.g., age 40 might predict 120 mm Hg, age 70 might predict 145 mm Hg).

Visualization: A scatter plot with a fitted regression line showing the trend between age and blood pressure.



 **Support Vector Machine (SVM) Classification**

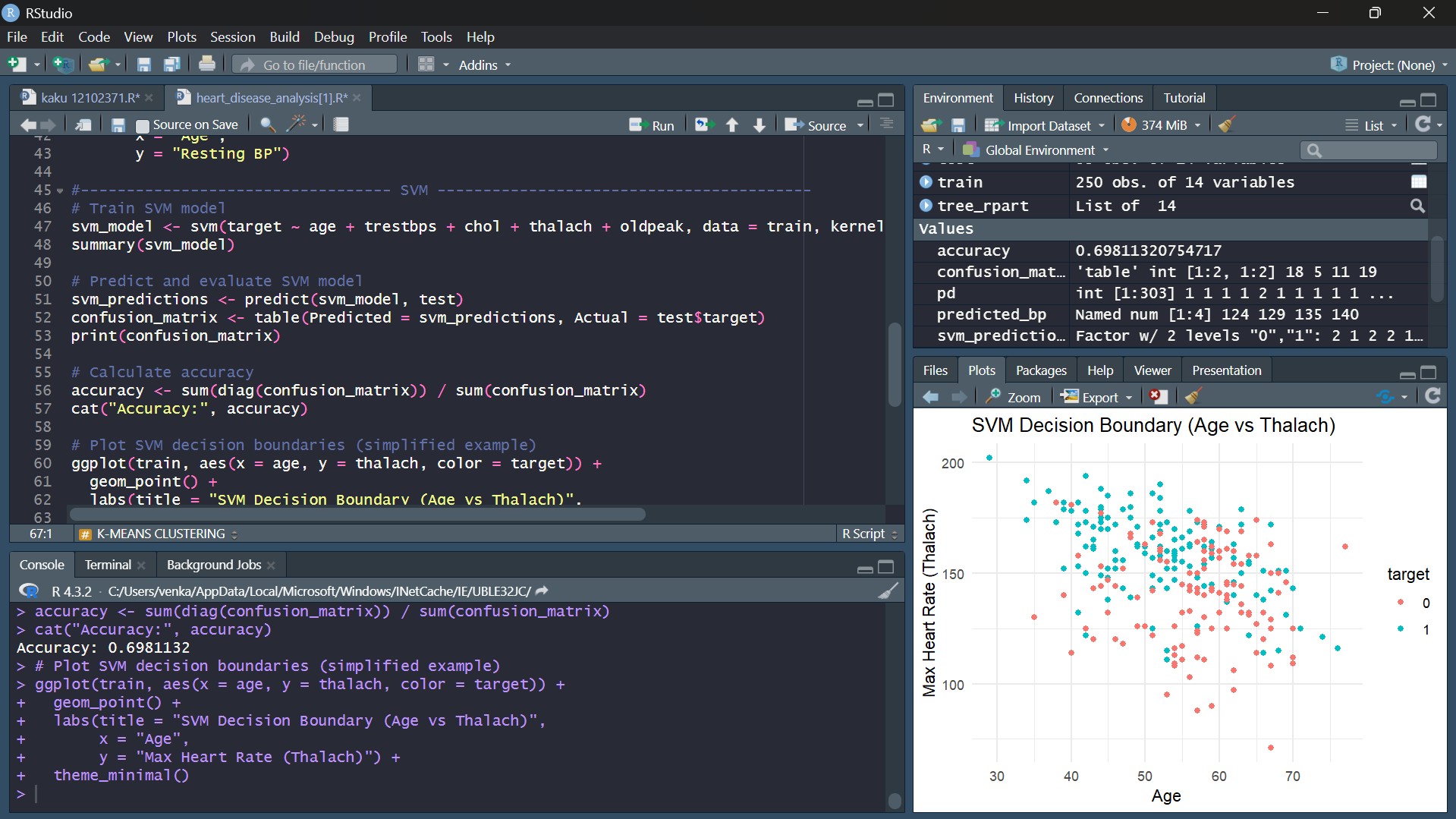
**Objective:**

SVM is a classification technique that finds the optimal hyperplane to separate data points of different classes (heart disease vs no heart disease).

### Steps:

* **Model Training**: An SVM model is trained using a linear kernel, with age, trestbps, chol, thalach, and oldpeak as predictors for the binary target variable (target).
* **Prediction and Evaluation**: Predictions are made on the test set, and the confusion matrix is used to evaluate the model’s performance.
* **Accuracy Calculation**: The accuracy is calculated by dividing the number of correct predictions by the total number of predictions.

Visualization: A plot showing the decision boundary in a two-dimensional space (e.g., age vs max heart rate). The decision boundary separates the data points of patients with and without heart disease.



 **K-Means Clustering**

### Objective:

K-Means clustering is an unsupervised learning algorithm that groups data points into clusters based on similarity. Here, the goal is to find clusters of patients based on features like age, blood pressure, cholesterol, max heart rate, and oldpeak.

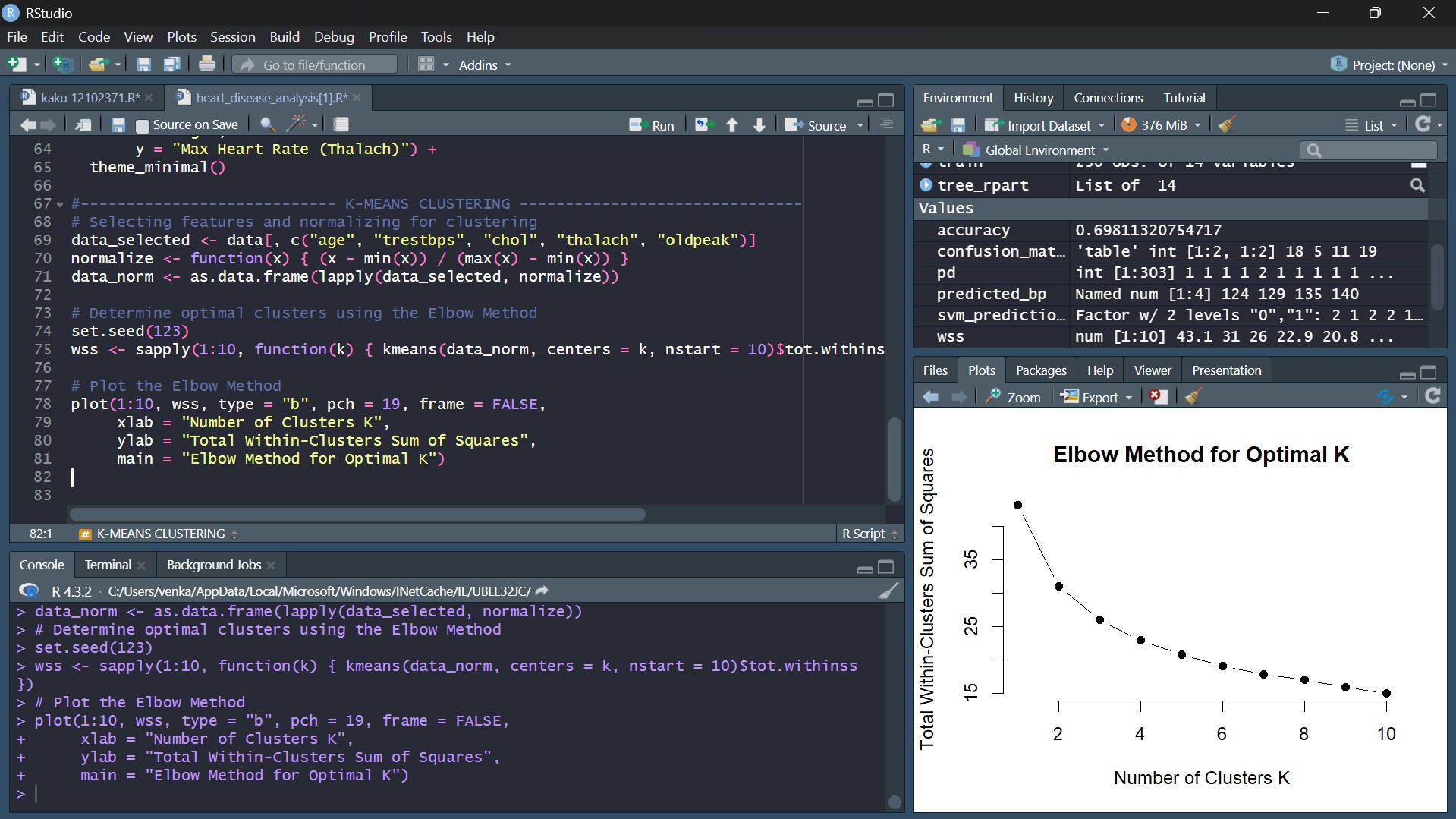
### Steps:

* **Data Preparation**: Selected features (age, trestbps, chol, thalach, oldpeak) are normalized to bring them to a common scale.
* **Elbow Method**: The Elbow Method is used to determine the optimal number of clusters by calculating the within-cluster sum of squares (WSS) for different values of K (number of clusters).
* **Plotting**: A plot of the WSS for values of K from 1 to 10 is generated.
* **Elbow Plot**: The plot will show the total within-cluster sum of squares (WSS) for different values of K. The "elbow" point in the plot indicates the optimal number of clusters. For instance, if the elbow occurs at K = 3, it suggests that three clusters are optimal.
* **Cluster Insights**: Based on the clustering, we can identify distinct groups of patients, such as low-risk, medium-risk, and high-risk groups based on their clinical features.

**WITHOUT ELBOW METHOD**

A screenshot of a computer program

Description automatically generated



## Future Scope

#### Improvement of Prediction Models

Enhanced Feature Engineering: Additional derived features, such as depreciation rate, fuel efficiency, and maintenance cost, could improve model performance, particularly for predicting car prices and classifying car types.

Incorporating Ensemble Learning: Applying ensemble techniques, like Random Forest or Gradient Boosting, could improve accuracy and robustness over individual models like Decision Trees and SVM, providing more reliable results for complex patterns.

#### Integrating External Data

Economic Indicators: External data on factors such as inflation rates, fuel prices, and regional economic conditions could be integrated to better understand car pricing trends and consumer preferences.

Customer Demographics: Data on customer demographics (age, income level, etc.) could provide insights into which car types appeal to different consumer segments, supporting a more targeted marketing approach.

#### Expanding to Real-Time Analysis

Real-Time Data Updates: Integrating the dataset with real-time data from car marketplaces would allow dynamic, up-to-date predictions and trend analysis.

Deploying Predictive Models: Creating a live web application that leverages the trained models to provide price estimates, recommendations, or classifications based on user-input car attributes.

#### Exploring Geographic and Temporal Trends

Location-Based Analysis: By adding geographic data, such as city or state, we could explore how car prices vary across regions, identify popular car types in different areas, and observe local economic impacts on car sales.

Time Series Analysis: For datasets that include a time component, we could perform time series analysis to study changes in car prices, demand for specific car types, or seasonal effects on car sales.

#### Application of Advanced Machine Learning Models

Deep Learning Models: For a larger dataset, models like neural networks could uncover more nuanced relationships between features, potentially increasing the accuracy for predicting car prices and classifications.

Natural Language Processing (NLP): If text-based data is available (e.g., customer reviews, car descriptions), NLP techniques could help analyze sentiment and keywords, adding qualitative insights into customer preferences.

#### Enhanced Visualization and Dashboarding

Interactive Dashboards: Developing interactive dashboards with tools like Shiny (in R) or Power BI can make it easier to explore insights visually. Users could filter and compare car types, visualize trends, and interact with predictive models in a user-friendly way.

Augmented Reality (AR) and Virtual Reality (VR): For car dealerships or buyers, AR/VR visualizations could help users explore different car segments virtually, allowing a more immersive data interaction.

#### Recommendations for Buyers and Sellers

Automated Recommendations: Based on analysis results, develop recommendation systems that help customers select cars suited to their preferences, budget, and lifestyle.

Price Negotiation Assistance: Using price prediction models, create a tool that helps buyers negotiate better deals by providing data-driven price ranges.

## References

1. Machine Learning in R:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

This book offers a comprehensive introduction to machine learning techniques in R, with chapters decision trees, linear regression, SVM, and clustering.

1. Decision Trees and Classification:

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. Available at: https://CRAN.R-project.org/package=rpart

This R package documentation provides detailed information on the `rpart` function used for decision tree analysis.

1. Linear Regression Fundamentals:

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). \*Applied Linear Regression Models\*. McGraw-Hill Irwin.

A standard reference on linear regression theory and applications, including how to interpret coefficients and assess model fit.

1. Support Vector Machine (SVM):

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2020). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-4. Available at: https://CRAN.R-project.org/package=e1071

This package documentation covers the `svm` function in R, explaining SVM algorithms and parameters for classification.

1. K-Means Clustering:

Hartigan, J. A., & Wong, M. A. (1979). "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1), 100-108.

A foundational paper that describes the k-means clustering algorithm, used for grouping data based on similarities.

1. Data Visualization in R:

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. This book serves as a primary resource on `ggplot2`, a powerful R package for data visualization, useful for plotting models and clustering results.

1. Kaggle Datasets and Machine Learning Applications:

Kaggle.com: Machine Learning Datasets and Competitions. Available at: https:/[/www.kaggle.com](http://www.kaggle.com/) Kaggle hosts various datasets, including those related to cars and used in machine learning projects. It also provides community-driven insights and practical implementations of ML algorithms.

1. Car Market Analysis and Economic Insights:

Kelley Blue Book. (2023). Car Prices, Market Trends, and Economic Analysis. Available at: https:/[/www.kbb.com](http://www.kbb.com/)

An industry reference on car prices and trends, useful for understanding real-world economic

1. R Documentation and Resources:

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https:/[/www.R](http://www.r-project.org/)-[project.org](http://www.r-project.org/)

Official documentation and resources for the R programming language, including tutorials on implementing various machine learning algorithms.

### BIBLIOGRAPHY

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

This book covers basic machine learning techniques in R, such as decision trees, linear regression, and clustering.

1. Therneau, T., & Atkinson, B.(2019). rpart: Recursive Partitioning and Regression Trees. R package documentation. Available at: https://CRAN.R-project.org/package=rpart

Documentation for the R package used to create decision trees for classification.

1. Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models. McGraw- Hill Irwin.

A detailed guide on linear regression, useful for predictive analysis.

1. Hartigan, J. A., & Wong, M. A. (1979). "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

An article explaining k-means clustering, which groups data based on similarities.

1. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York. This book covers `ggplot2` in R, a popular tool for creating data visualizations.
2. Kaggle.com. Machine Learning Datasets and Competitions. Available at: https:[//www.kaggle.com](http://www.kaggle.com/) Kaggle provides a variety of datasets, including car data, for machine learning projects.
3. Kelley Blue Book. (2023). Car Prices, Market Trends, and Economic Analysis. Available at: https:/[/www.kbb.com](http://www.kbb.com/)

A source of car pricing data and market trends.

1. R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available at: https:/[/www.R](http://www.r-project.org/)-[project.org](http://www.r-project.org/)

Official R documentation and resources for statistical computing and analysis. and car data analysis in R.