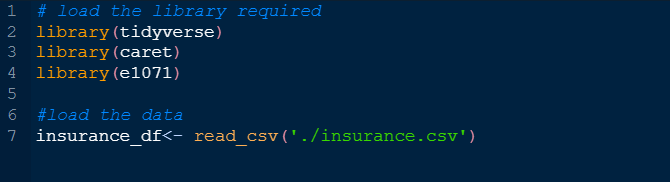
**UTILIZING INSURANCE DATA FOR DATA ANALYTICS AND MACHINE LEARNING WITH R**

**INTRODUCTION**

Businesses and organizations in today's data-driven world rely on data to make wise decisions. Organizations can leverage the useful insights and patterns they discover through data analysis to enhance their operations and stay one step ahead of the competition. The employment of algorithms and statistical models, on the other hand, enables computers to learn from data and make predictions or judgments without the need for explicit programming. Organizations now extract insights from data in a completely new way thanks to the convergence of data analytics and machine learning (Boehmke and Greenwell, 2019). In this assignment, we'll look at how to use R, a well-liked programming language and environment for statistical computation and graphics, to perform data analytics and machine learning. Using different machine learning techniques, we will use an insurance dataset to forecast the purchase of insurance policies based on demographic and other pertinent variables (Lantz, 2019a).

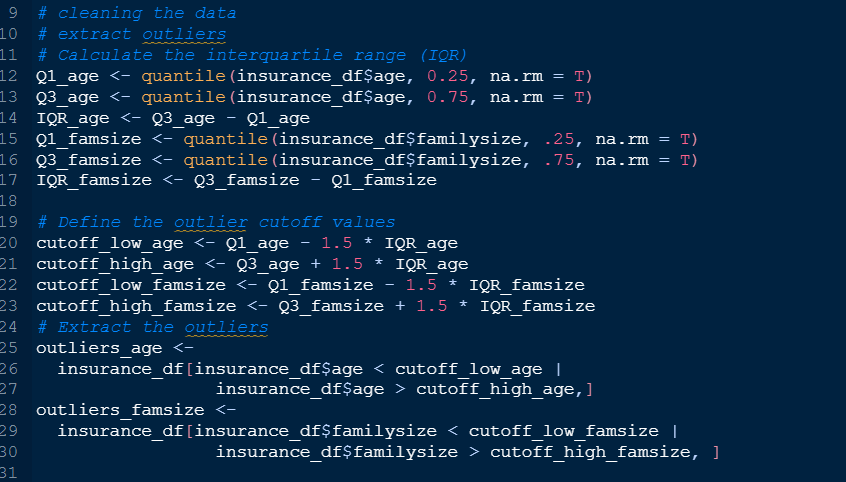
**DATA LOADING**

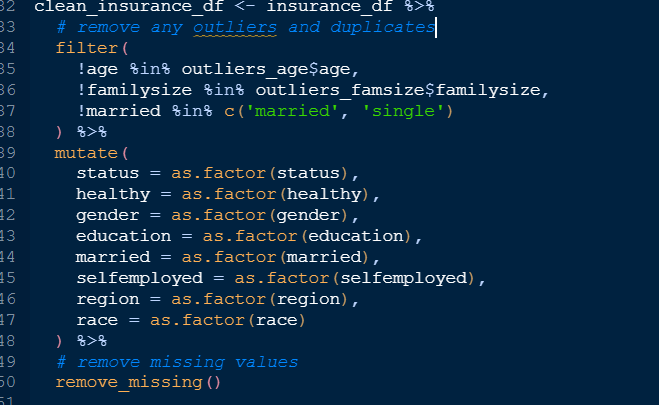
In any data analysis project, importing the data is a key step. It entails gaining access to and reading data from a variety of sources, including files, databases, and APIs (Balduzzi, Rücker and Schwarzer, 2019). There are numerous packages and functions in R that can be utilized for data loading. In this part, we'll examine a few popular approaches of data loading in R, along with their benefits and drawbacks. We will also go through how to deal with corrupt or missing data when the file is being loaded. To do efficient data analysis in R, one must have a firm grasp of data loading. We load data with library() function and if the packages are not installed, install them using install.packages(). Loading the file will require a function that can handle csv files.in this case read\_csv() function will accomplish this and its shown below.



**DATA CLEANSING USING R**

Data cleaning, which involves finding and fixing mistakes, inconsistencies, and missing values in a dataset, is a crucial stage in the data analysis process. Data cleaning tools and methods are widely available in R, a prominent computer language for data analysis. Analysts may verify the accuracy and integrity of their data by utilizing R for data cleaning, which is essential for generating trustworthy insights and findings(Ridzuan and Zainon, 2019). We will create R codes that will clean our data and it is shown below.

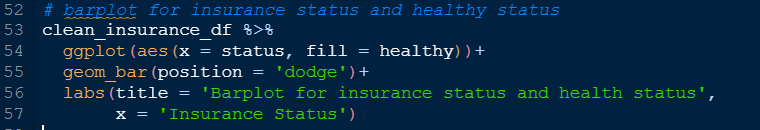


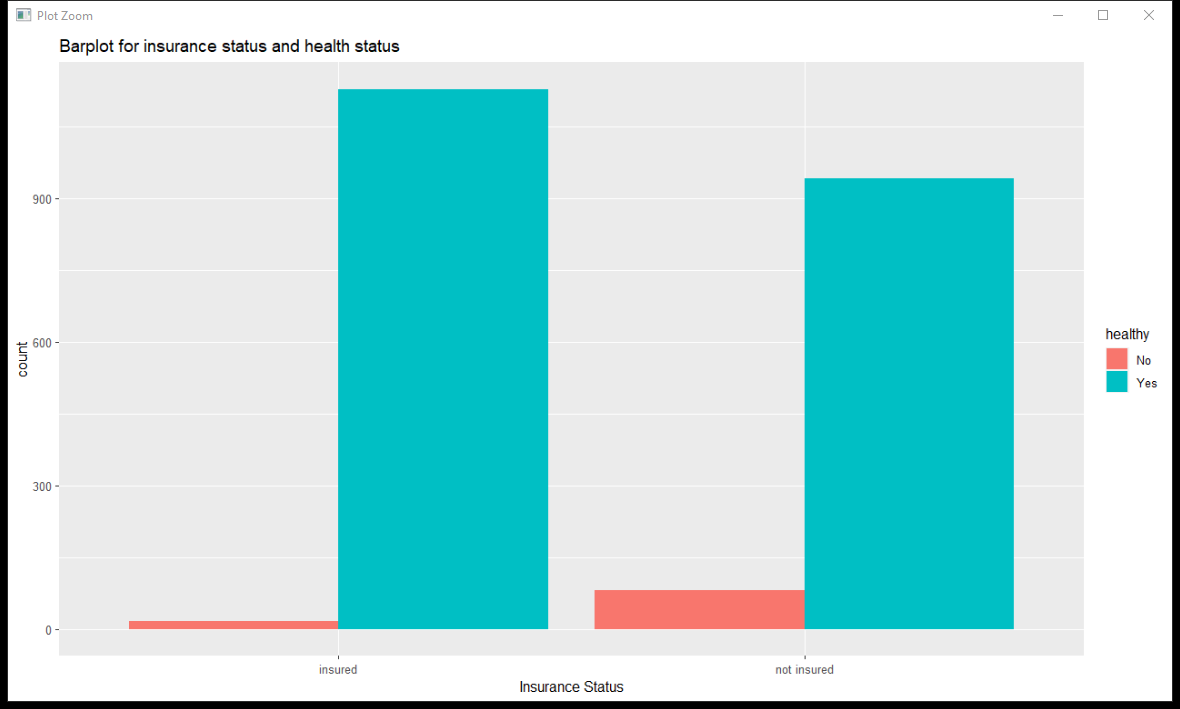


In lines of code above we first filtered out the outliers and used the outliers to create a new data frame by filtering the outliers. Also, we removed the duplicates and factored out all variables of character data type.

**ANALYSING DATA AND CREATING VISUAL REPRESENTATION OF THE FINDINGS**

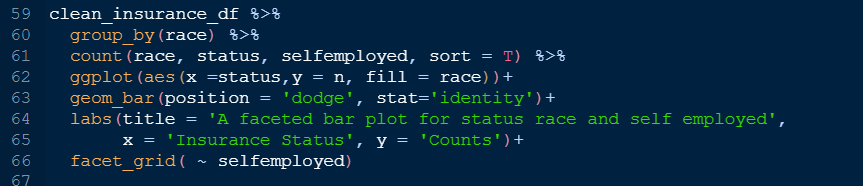
The primary objective of this analysis is to determine potential buyers of health insurance policies, with a particular emphasis on the "status" variable. This variable comprises different levels, such as "insured" or "not insured," and can be represented visually using a bar plot. Additionally, we can incorporate another variable, "healthy" to provide further information about whether the person is healthy or not. We would want to know if health is a factor influencing buying health insurance covers. A plot is created using the following codes.

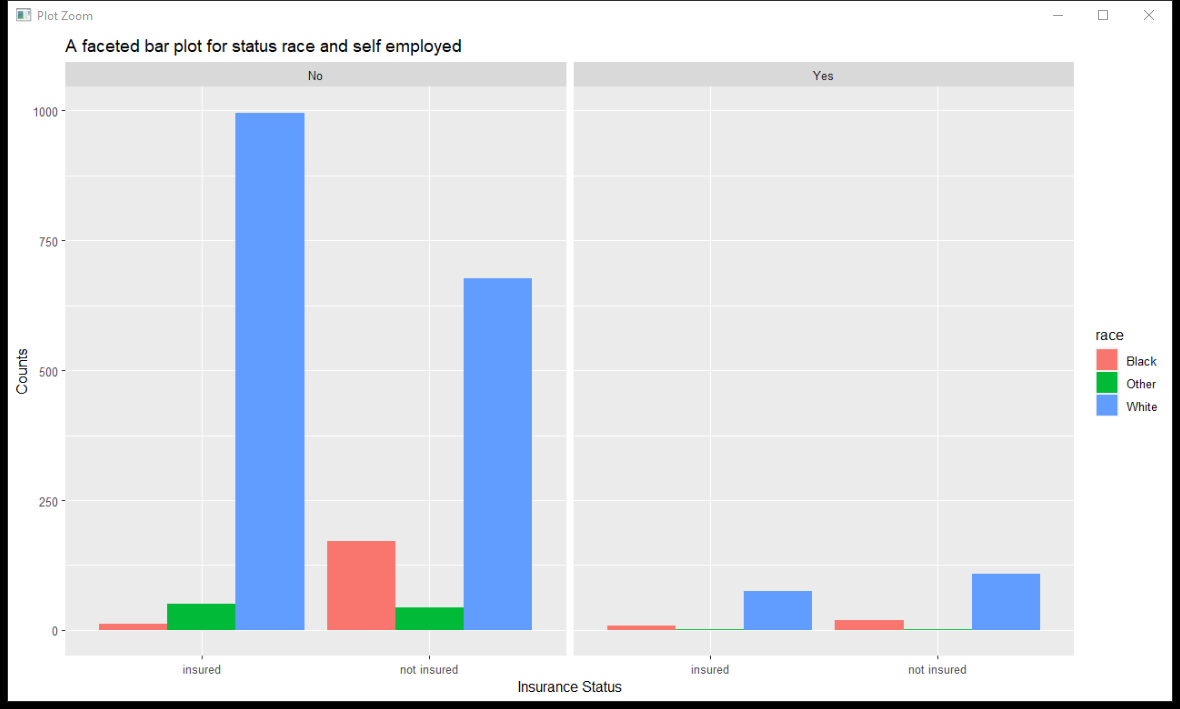




Although many healthy people have obtained health insurance policies, the barplot above demonstrates that the proportion of unhealthy persons having insurance is rather low. This conclusion could be explained by a number of variables, including the chance that healthy people will get insurance, variations in the price of insurance policies, and coverage for various medical conditions. The number of people who haven't protected themselves against certain threats is high for healthy people but lower for those who aren't healthy. This might be because individuals who are healthy see less need for insurance, while those who have pre-existing conditions have trouble getting coverage or have few options. Disparities in insurance coverage and access to healthcare depending on variables including location, age, and income may also be important. Thus we can conclude that health is a factor influencing purchasing health insurance policy.

Next, we can create another plot and look at another variable and see if it’s a factor influencing insurance status. We can look at which race purchases health insurance mostly based on whether they are self-employed or not thus a facetted bar plot will present this information well.





The facetted bar plot reveals that White non-self-employed people have the highest percentage of insurance coverage and the highest percentage of no insurance coverage. This raises the possibility of a connection between race, independence, and insurance protection. For instance, self-employed people may be less likely to have insurance coverage than people of other races, or White people may be more likely to have health coverage than people of other races. It's also plausible that White people who aren't self-employed are more likely to be covered by insurance than White people who are. In general, the complex interaction between race, self-employment, and insurance coverage necessitates more research.

**SUPERVISED MACHINE LEARNING IN R**

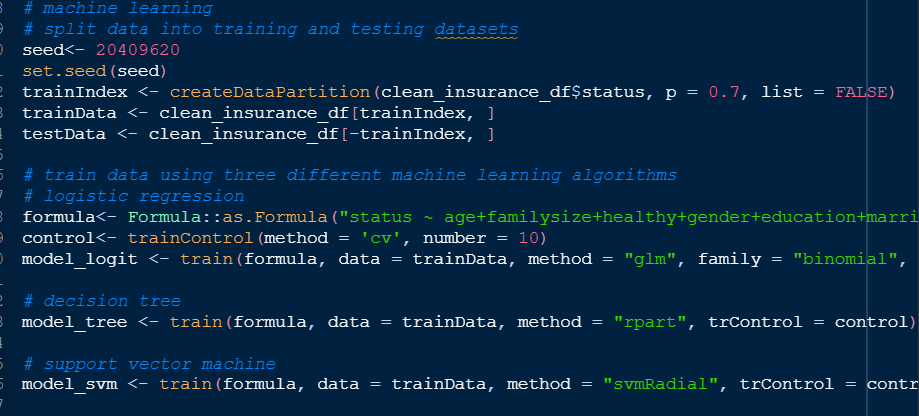
Using data, supervised machine learning is a potent method for creating prediction models. On a labeled dataset, where the input data is coupled with known output data, a model must be trained. This enables the model to understand the connections between the input and output data and utilize this understanding to anticipate the behavior of new, unobserved data (Maurya and Srivastava, 2021).

We are going to create a supervised machine learning that will find a likelihood of an individual to buy an insurance cover.

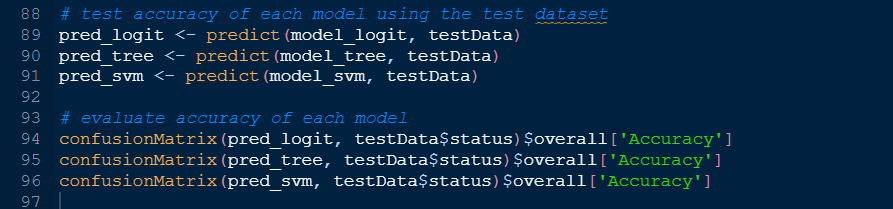
Supervised machine learning is suitable for predicting if a person will buy health insurance, using a labeled dataset to train an algorithm. The algorithm can then make predictions on new data to target marketing efforts and increase sales. For this problem, classification is the appropriate method as it predicts category labels based on input features. Regression is more suitable for predicting continuous values, such as the price of a house based on its attributes.

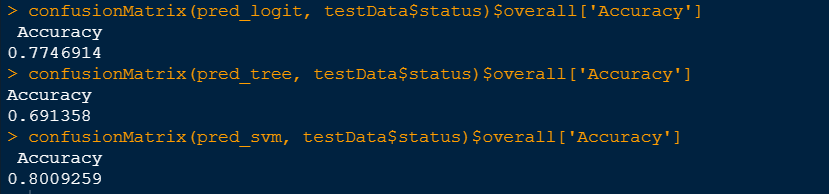
**Creating the model using caret and e1071**

We will start by creating three different models using caret and e1071 packages. This is shown below.



**Test the accuracy of the models**





The logistic regression model has the highest accuracy of 77.47%, followed by the decision tree model with an accuracy of 69.14%. However, accuracy alone may not be the only metric to consider when evaluating the performance of machine learning models. The suitability of the chosen algorithms depends on the specific problem and dataset. To improve accuracy, feature engineering, hyperparameter tuning, and using more advanced algorithms could be explored.

Using the aforementioned analysis to predict whether people will purchase insurance plans raises ethical concerns. Discrimination may occur if the results are used to unfairly target individuals based on demographics such as age or gender, and privacy concerns arise due to the use of personal information. It is important to gather and store personal data securely with the individuals' consent.

**REFERENCES**

Balduzzi, S., Rücker, G. and Schwarzer, G. (2019) ‘How to perform a meta-analysis with R: a practical tutorial’, *BMJ Ment Health*, 22(4), pp. 153–160.

Boehmke, B. and Greenwell, B.M. (2019) *Hands-on machine learning with R*. CRC press.

Lantz, B. (2019a) *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Lantz, B. (2019b) *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd.

Maurya, P. and Srivastava, N. (2021) ‘Performance Evaluation of the Supervised Machine Learning Algorithms Using R’, in *Data Engineering and Intelligent Computing: Proceedings of ICICC 2020*. Springer, pp. 397–406.

Ridzuan, F. and Zainon, W.M.N.W. (2019) ‘A review on data cleansing methods for big data’, *Procedia Computer Science*, 161, pp. 731–738.