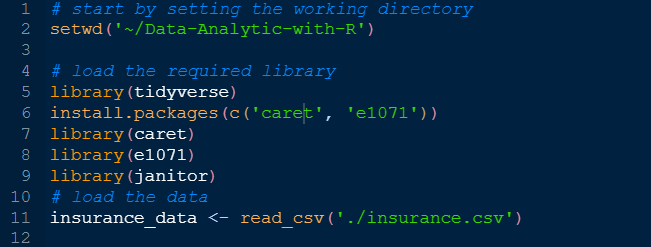
**UTILIZING R FOR DATA ANALYSIS AND MACHINE LEARNING**

**INTRODUCTION**

Extraction of beneficial knowledge and predictions from data requires both data analysis and machine learning techniques, which have become indispensable. In this paper, we'll look at how these techniques were used using insurance data that was gathered as part of the 1996 US Medical Expenditure Panel Survey. In order to obtain knowledge about the variables that influence medical costs and to create prediction models that can be helpful for future decision-making, we will undertake data analysis and machine learning on this information

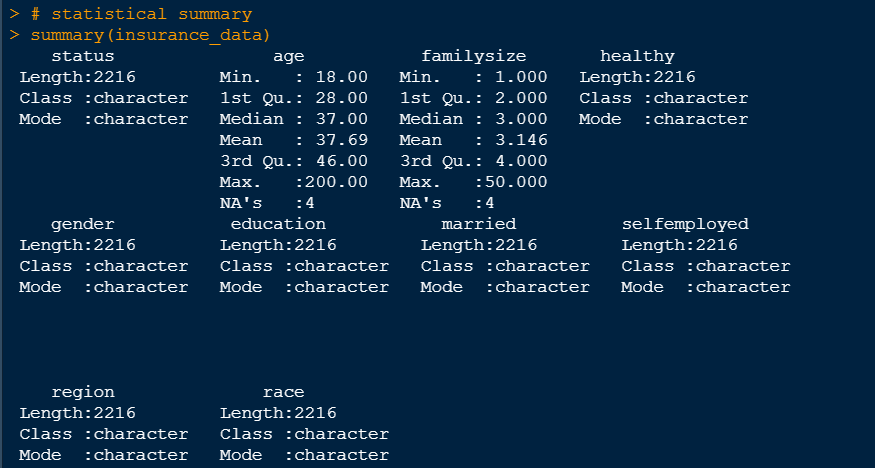
**DATA ANALYSIS USING R LANGUAGE**

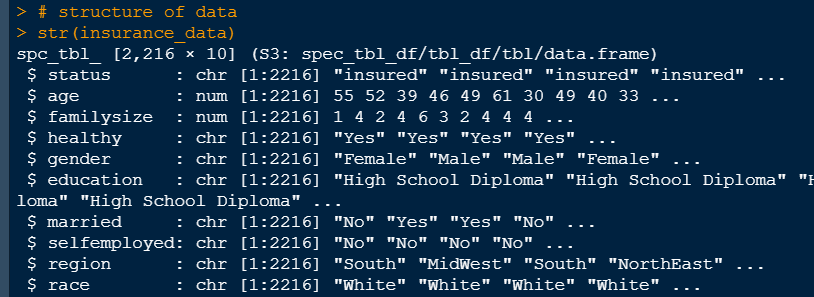
**Loading the data**

To start data analysis in R, we must first load the required datasets and libraries. This involves loading appropriate datasets and libraries to ensure efficient analysis. In this tutorial, we will use the "Insurance" dataset, which we can load using the read\_csv() function from the readr package. To load packages, we use the library() function. If the packages are not installed on the computer, we can install them using the install.packages() function. See the screenshot below for an example.

We start by setting the working directory to where the file is located and then load the packages using library() and install\_package() functions then using the read\_csv() file loaded the file.

We can check out the structure of the data so as to understand what we are dealing with. This is important since we will understand the data before cleaning it incase it need to be cleaned. We also need a statistical summary of the data. These two steps can be done using the str() and summary() functions respectively and the results are displayed below.



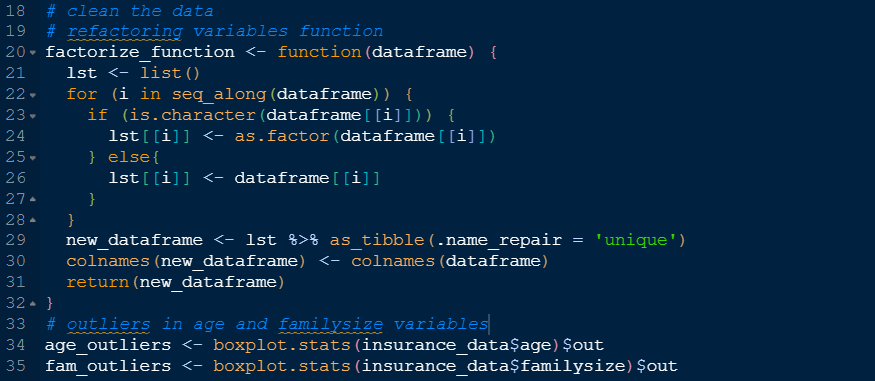


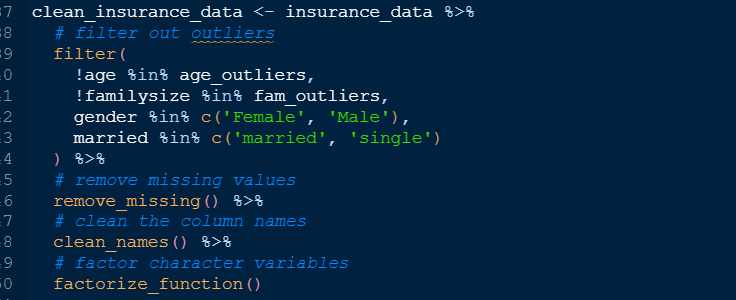
The data has 10 columns and contains 2,216 rows. The summary statistics is seen in the screenshot above where some variable has missing values and outliers which will be dealt with below. The character data variables are to be converted to factors.

**Cleaning of data in R**

Cleaning datasets entails locating and fixing mistakes, inconsistencies, and missing data. It is a crucial stage in the data analysis process.(Kabacoff, 2022). R offers a variety of tools and methods for cleaning datasets, including removing duplicates, dealing with missing information, and fixing data types. Researchers can improve the accuracy and dependability of their findings, which can result in more insightful analyses and better decision-making, by thoroughly cleaning databases (De Jonge and Van Der Loo, 2013).

Cleaning is done using the following codes.

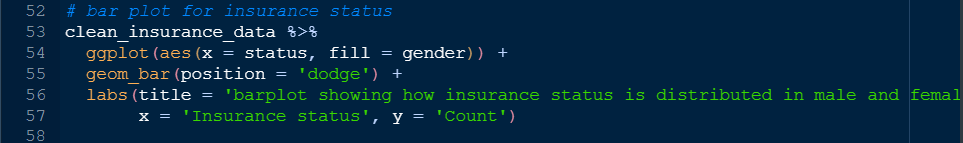


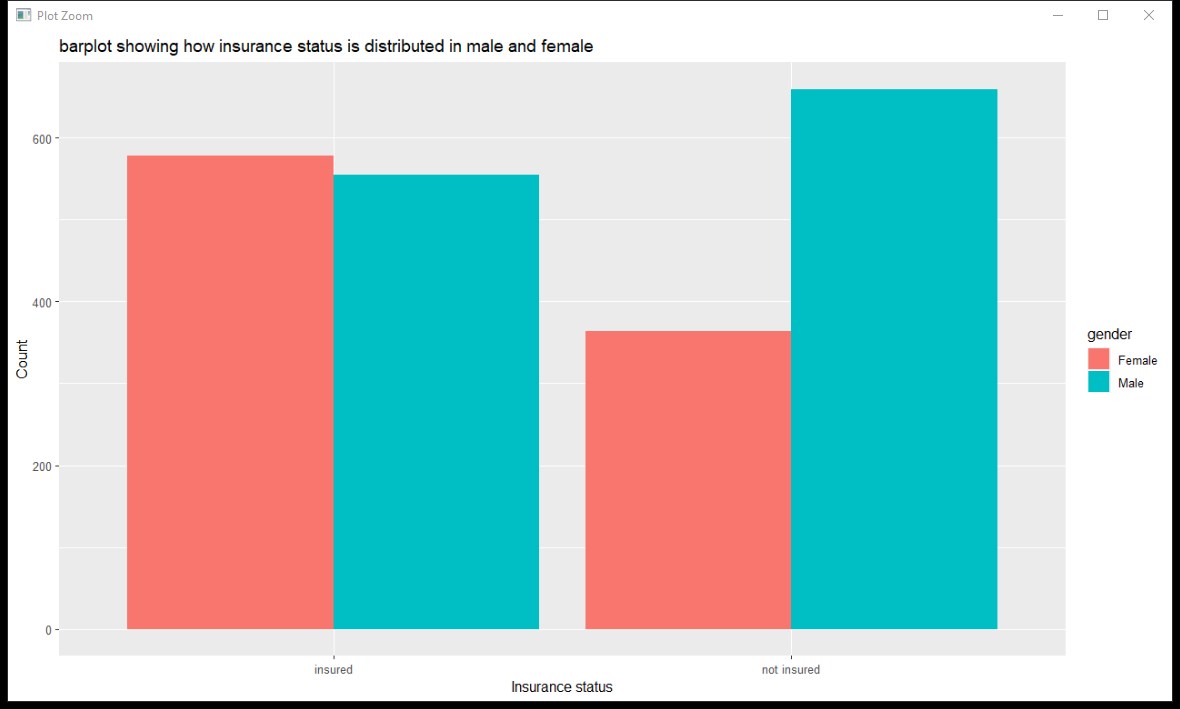


I started by first creating a function that factor out character variables, then filtered out outlier values, cleaned the names and then removed all missing values. The data is clean for analysis.

**Data visualizations and exploratory data analysis**

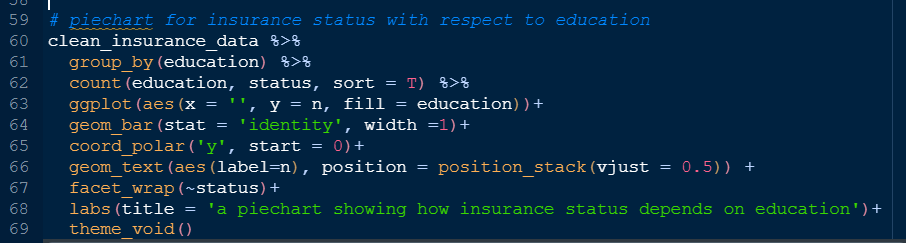
Data analysis must include both data visualization and exploratory data analysis (EDA). In order to help users, comprehend the underlying patterns, trends, and relationships within the data, data visualization refers to the graphical depiction of the data. It enables users to immediately spot trends and insights that are difficult to spot in raw data formats (Lekkala and Maddineni, 2020). EDA, on the other hand, entails the procedure of analyzing, condensing, and displaying a dataset's primary properties. It is frequently the initial phase in any project involving data analysis and is used to comprehend the structure of the data and spot any patterns, trends, or abnormalities. We are going to focus on the insurance status against the other variables. A bar plot is created below with the following codes (Indrakumari, Poongodi and Jena, 2020).

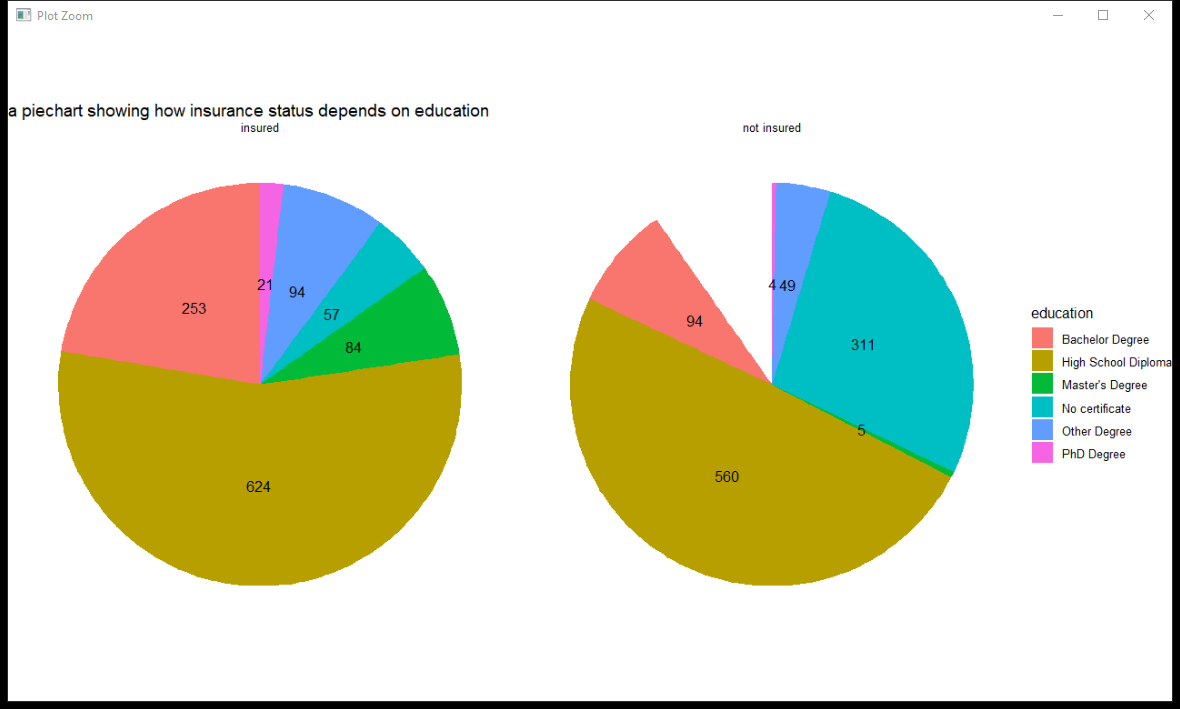




The bar plot shows how insurance status is distributed between men and women. The fact that more women than men have insurance coverage raises the possibility that this might be because women are more health-conscious or because insurance plans are more pertinent to them. On the other side, there are more men than women in the "not insured" group, which may be because insurance is expensive or men believe they are less at risk than women. In conclusion, gender seems to be an important consideration for determining insurance status.

Lets look at the insurance status with education variable and try to see whether insurance status is dependent to education level. A pie chart is created below.





The graph demonstrates that policyholders have greater levels in education, with the majority having a high school diploma or a bachelor's degree. On the other hand, the percentage of uninsured people is higher among those lacking a certificate or with only a high school diploma. According to the data, the likelihood of obtaining insurance coverage rises with education levels. As a result, it would seem that education has a substantial role in determining the status of insurance.

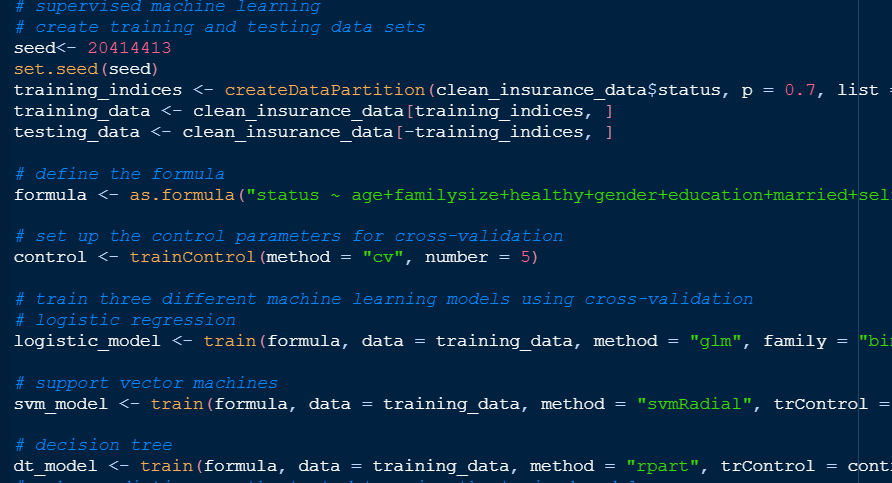
**MACHINE LEARNING USING R**

**Supervised machine learning in R**

On the basis of labeled data, supervised machine learning develops a model to predict the target variable from fresh data (Maurya and Srivastava, 2021). Our goal is to predict a person's likelihood of getting health insurance using this method. We'll employ algorithms that learn from labeled data to forecast results for fresh, unlabeled data, such as decision trees, logistic regression, random forests, and support vector machines. Regression is not appropriate because our insurance status variable is categorical. Our supervised learning algorithm will be written in R.

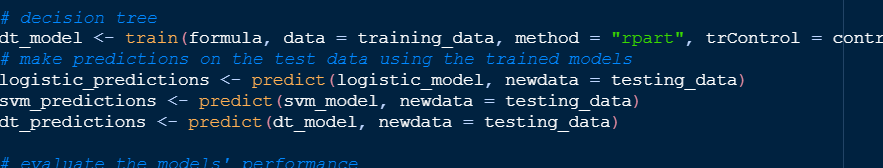
**Training 3 models**

We are going to train three model, that is, logistic regression, support vector machine and decision tree as shown below.



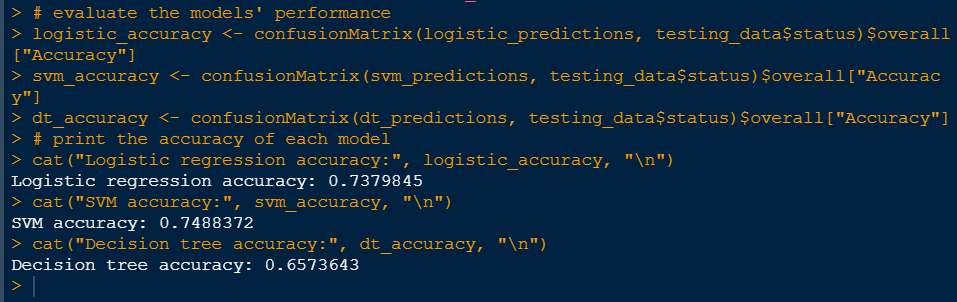
**Prediction based on the models**

Using the models created we make prediction as shown below.



**Evaluate models performance**

We then use the models to evaluate their performance and different performance was given below.



The decision tree came in at 65.73%, followed by logistic regression at 73.79% and the support vector machine at 74.88% accuracy. We can employ sophisticated algorithms or include more pertinent features in the dataset to increase forecast accuracy. Although straightforward, logistic regression may not be able to capture intricate nonlinear relationships. Although they can handle nonlinear interactions, support vector machines may be computationally expensive. Decision trees are straightforward, but they may overfit and have poor predictive ability. In general, the selection of an algorithm is based on certain demands and requirements. Based on performance and flexibility, support vector machines are the best option for this issue.

**Improving the accuracy of the prediction**

Prediction accuracy can be improved by:

1. Increase the dataset's number of pertinent features.
2. Improve the hyperparameters of the models.
3. Using ensemble techniques like random forest or gradient boosting, combine the predictions of various models.

The dataset and model selection, together with the use of appropriate feature engineering, hyperparameter tweaking, and ensemble methods techniques, are all necessary for increasing accuracy (Jiang, Gradus and Rosellini, 2020).

When using the above analytic results to predict the likelihood of individuals to purchase health insurance policies, ethical issues should be considered, such as potential bias and the need for transparency. Bias can occur if the training data is biased, and steps should be taken to ensure a diverse and representative dataset. Individuals should be informed about the use of predictive models and the algorithms and data used should be transparent.

**REFERENCE**

De Jonge, E. and Van Der Loo, M. (2013) *An introduction to data cleaning with R*. Statistics Netherlands Heerlen.

Indrakumari, R., Poongodi, T. and Jena, S.R. (2020) ‘Heart disease prediction using exploratory data analysis’, *Procedia Computer Science*, 173, pp. 130–139.

Jiang, T., Gradus, J.L. and Rosellini, A.J. (2020) ‘Supervised machine learning: a brief primer’, *Behavior Therapy*, 51(5), pp. 675–687.

Kabacoff, R. (2022) *R in Action: Data Analysis and Graphics with R and Tidyverse*. Simon and Schuster.

Lekkala, H.R. and Maddineni, V. (2020) ‘An Third Eye of Data Visualization Using R’, *Available at SSRN 3654792* [Preprint].

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.