DEEPTHI CHRISTINA VICTOR SAGAYANATHAN

Word Count: 2200

Table of Contents

Abstract

The objective is to help SmartphoneTech make sense of the data they have collected by conducting advanced data analysis techniques, such as cluster analysis to notice the pattern in the behaviour and demographics of the users and perform classification to predict the type of of smartphone used by each user.

Introduction

In today's rapidly evolving technological landscape, companies like SmartphoneTech are keen to better understand smartphone users and their behaviours to gain a competitive edge in the market. The valuable insight from analysis can inform business decisions and provide a competitive advantage in the crowded smartphone market. Ultimately, the goal is to provide actionable insights that help SmartphoneTech make informed decisions that will drive growth and success in the future.

Dataset Explanation

The dataset consists of 532 instances recorded. There are 12 variables included in the dataset:

Variable	Description
Smartphone	It indicates the type of the smartphone an individual has. It can be one of the following 3 types – Android, iPhone, or Blackberry.
Gender	gender of the individual. Values Male, Female or other.
Age	Age of the individual. Takes numbers in years as values.
Honesty-Humility	(HEXACO) Personality Studies show that smartphone use is very much related to the personality traits of an individual. These take values 1-5 depicting 1 as minimum and 5 as maximum.
Emotionality	
Extraversion	
Agreeableness	
Conscientiousness	
Openness	
Phone as status object	It takes values from 1-5 with 1 as minimum and 5 as maximum depicting the minimal consideration of phone as status object to maximal.
Socioeconomic status	It takes values from 1-10 with 1 as minimal consideration of phone as socio economic status and 5 as maximum.
Time owned current phone	It takes numeric values. (In months).

Data Cleaning

Data cleaning (Han, J., Pei, J., Kamber, M., 2011, p. 85) is an essential process that involves identifying and resolving inaccuracies, inconsistencies, and errors in datasets. The goal of data cleaning is to improve the quality and accuracy of the data, making it suitable for analysis and decision-making processes.

The first step in data cleaning is **data inspection**, where the data is examined to identify any errors, inconsistencies, or missing values. For example, if the age variable has a non-numeric value such as 's', it is replaced with a null value. If outliers are found, they are replaced with appropriate alternatives.

The second step in data cleaning is **data correction**, (Han, J., Pei, J., Kamber, M., 2011, pp. 88-89) where errors are identified and corrected. This may involve replacing missing values with appropriate estimates or imputations, correcting inaccurate or inconsistent data entries, and removing duplicates. For instance, if the age variable has missing values, they are replaced with the mean value of the column. The data related to Blackberry smartphone is omitted as we are interested to study people with android or iPhone users.

The next step in data cleaning is **data standardization**, which involves ensuring that data is presented in a consistent format suitable for analysis. For example, the categorical data like Smartphone and gender and converted to numerical. (Gender_Number = 1 as female, 0 as male; Other_Gender_Number = 1 as other, else 0)

The final step is **data transformation** (Han, J., Pei, J., Kamber, M., 2011, pp. 86-87). It involves transforming the datatype or format of the columns. For an instance, if numerical variables are found to be in character format, they are standardized to the appropriate format. In our case all the data are formatted to numeric type.

Exploratory Data Analysis

The Exploratory data analysis is done on a dataset called SmartphoneTech. The dataset contains information about smartphone users, including their gender, age, length of time they have owned their phone, Personality, and type of smartphone they use. The code has been organized into several sections, and each section focuses on answering a particular question or comparison.

The first section of the code plots the correlation coefficients of data with missing values and anomalies. From the above plot, it is inferred that the iPhone users generally exhibit emotionality as their personality trait whereas android users exhibit honesty-humility and openness as their personality trait. iPhone users seem to hold it as a status object. Mostly female users opt for iPhone and vice versa for Android.

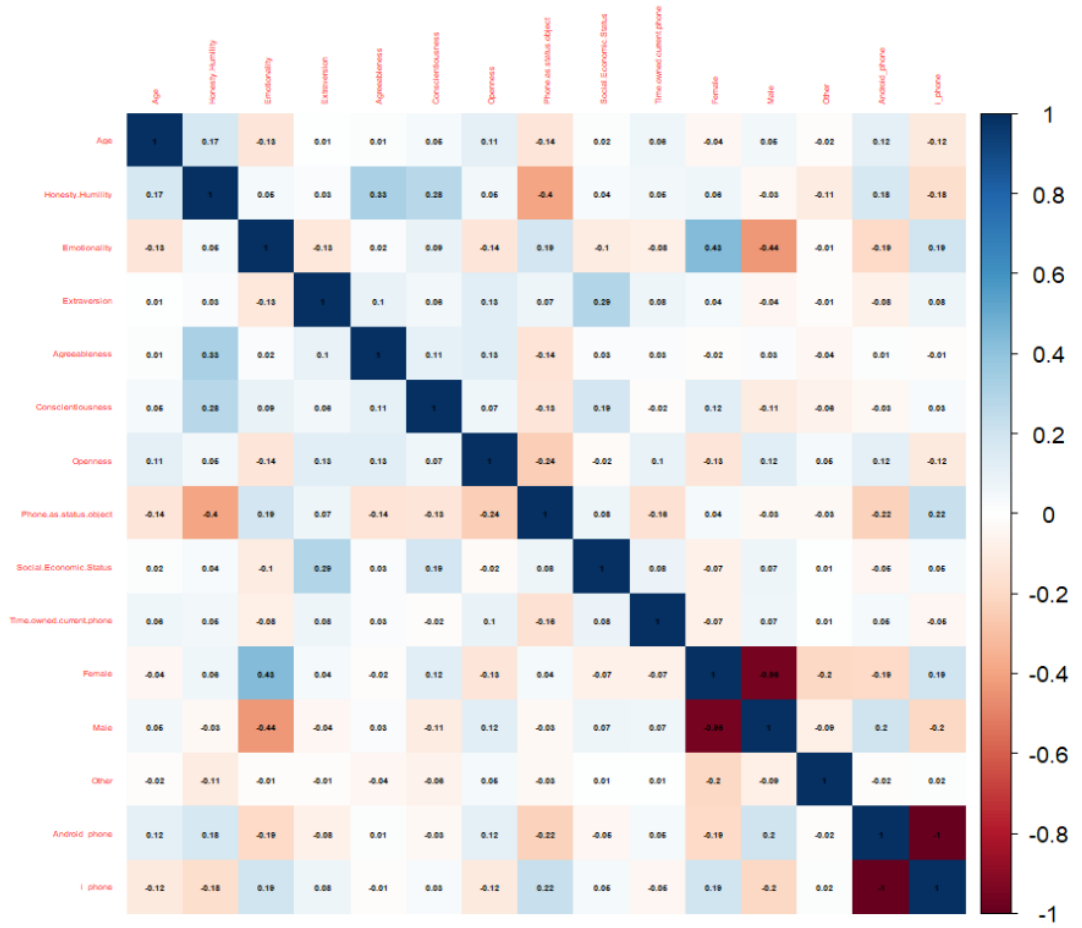

Fig.1: Correlation plot with NA values in the data

The second section of the code uses ggplot to plot a bar graph to show the count of Android and iPhone users. The plot shows that the number of iPhone users is slightly higher than the number of Android users.

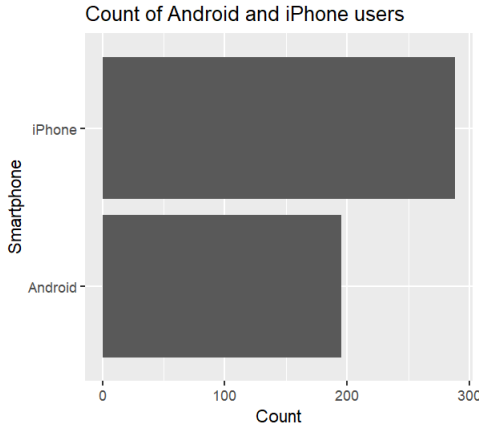

Fig.2: bar graph to show the count of Android and iPhone users

The third section of the code groups the SmartphoneTech data by gender and smartphone type and computes the count of each group. This section aims to answer whether female users have more iPhones than male users. The resulting table shows that more male users own Android phones,

while more female users own iPhones. Additionally, the section uses ggplot to create a bar graph that displays the number of Android and iPhone users for each gender.

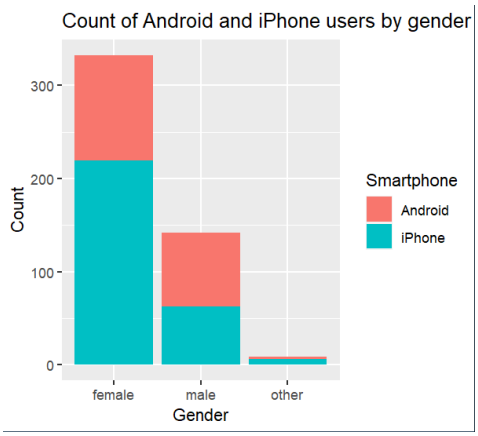

Fig.3: Bar graph that shows count of phone users by gender

The fourth section of the code creates a new column called Age_Range and populates it with age groupings based on the age of the smartphone users. This section aims to answer whether age influences the type of phone a user has. The resulting bar graph shows that there are more iPhone users in 16-25 age group.

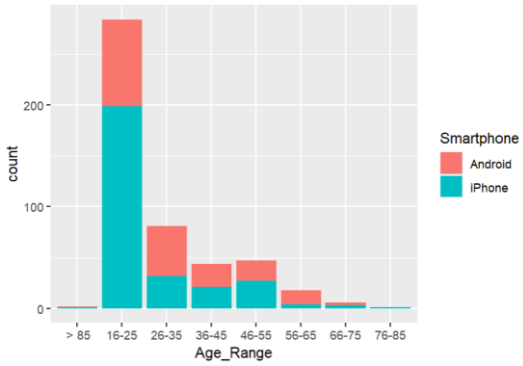

Fig.4: Bar graph that shows count of phone users by age range.

The fifth section of the code creates a new column called Time_Owned_By_Users_Range and groups the data based on how long users have owned their current phones. This section aims to answer what the distribution of the length of time individuals have owned their current smartphones is and how it is segmented by smartphone brand. The resulting bar graph shows that most users have owned their phone for two years or less, and Android users tend to own their phones for a longer time.

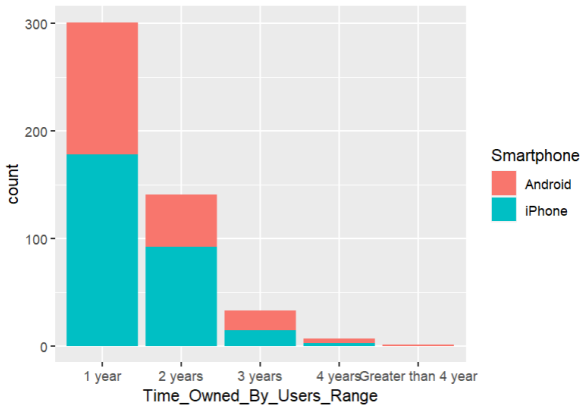

Fig.5: Bar graph that shows count of phone users by Time they owned the phones

The sixth section of the code provides a statistical summary of the entire SmartphoneTech dataset, including the count, mean, and standard deviation for each variable.

The seventh section depicts the histogram interpretation of age and personality traits. The age factor is right skewed (Mode<Median<Mean). Similarly, the personality traits Conscientiousness and Extraversion are also right skewed.

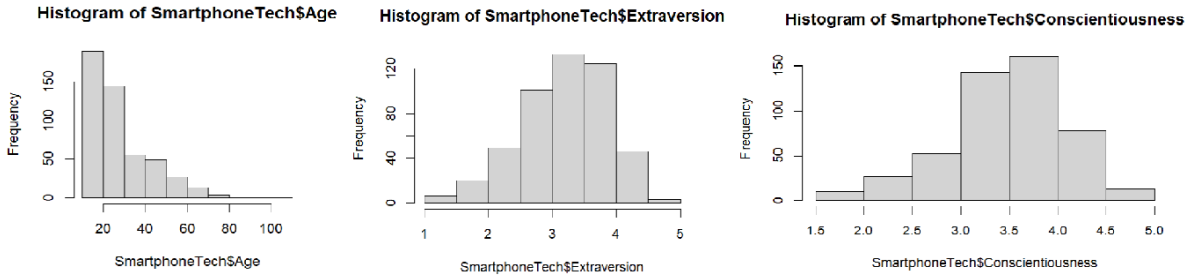

Fig.6: Histogram of Age, Extraversion and Conscientiousness

The Eighth section of the code creates a correlation matrix to show the relationship between each variable in the SmartphoneTech dataset i.e., without null values. The resulting plot shows the same inferences as the previous one.

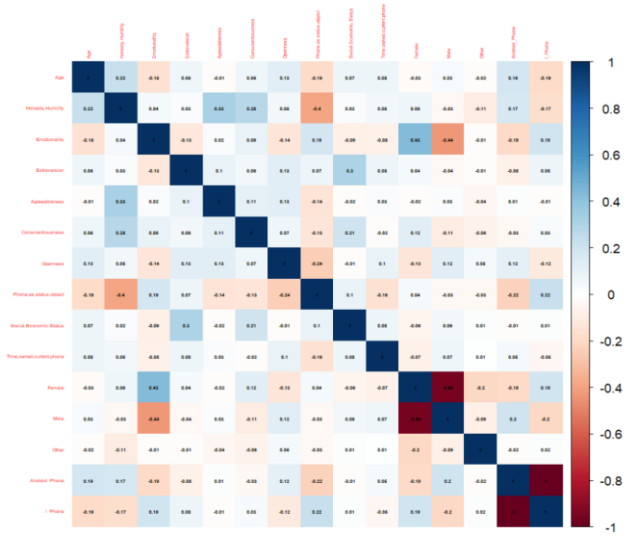

Fig.7: Correlation plot of the dataset after data correction

In conclusion, the provided code performs some data analysis on the SmartphoneTech dataset to answer several research questions related to smartphone users. The code uses various visualization techniques to present the results of the analysis, making it easy to understand and interpret the findings.

Clustering

Clustering is a machine learning technique that groups similar data points together based on their features (Han, J., Pei, J., Kamber, M., 2011, pp.444). The goal is to divide a set of data points into clusters so that points within the same cluster are more similar to each other than to points in other clusters .

Many applications use heuristic methods for clustering, such as the popular k-means (partitioning methods) algorithm (Han, J., Pei, J., Kamber, M., 2011, pp. 451-454). They are effective in finding spherical-shaped clusters in databases of small to medium size. These methods, while widely used, do have their limitations. They may not be effective for datasets with irregular shapes and can be computationally expensive for larger datasets.

The **k-means algorithm** (Han, J., Pei, J., Kamber, M., 2011, pp. 451-454) is a Centroid based technique used for clustering. It involves defining the centroid of a cluster as the mean value of the points within the cluster and proceeds by randomly selecting k objects in a dataset, each of which represents a cluster centre. Then, the remaining objects are assigned to the cluster to which they are most similar based on the Euclidean distance between the object and the cluster centre.

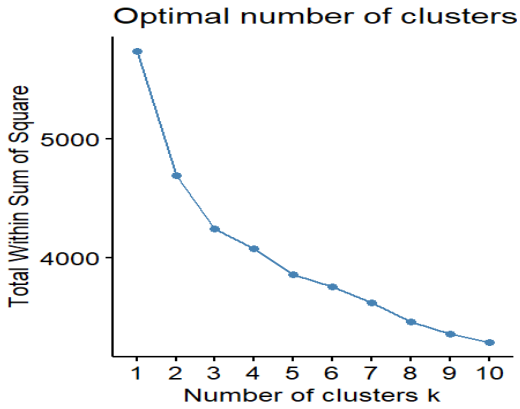

Fig.8: Elbow point

To get the k value, the elbow method is being used. From the plot of WCSS vs. k, From the plot, look for the "elbow point" (Han, J., Pei, J., Kamber, M., 2011, pp. 486-487). on the plot where adding more clusters is not improving the clustering significantly. The k means is done for clusters 3 to 7 out of which the one with k=3 performed well.

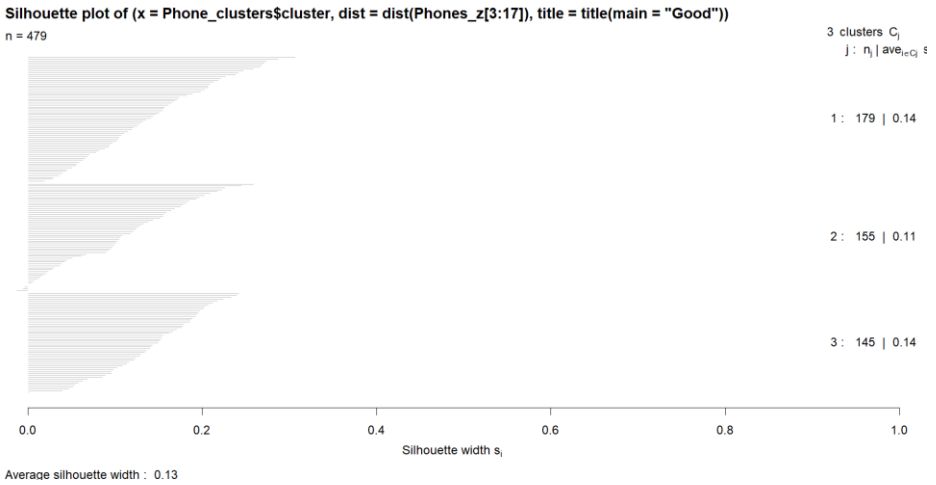

Fig.9:Silhouette analysis

Silhouette analysis is used to calculate the average silhouette width (Han, J., Pei, J., Kamber, M., 2011, p. 489). Average silhouette width evaluates clustering quality by measuring how well data points are assigned to their clusters. Silhouette width ranges from -1 to 1 (Han, J., Pei, J., Kamber, M., 2011, p. 490), with positive values indicating good matches and negative values indicating poor matches. The Average silhouette width for k means clustering with k=3 is 0.13.

Key Findings

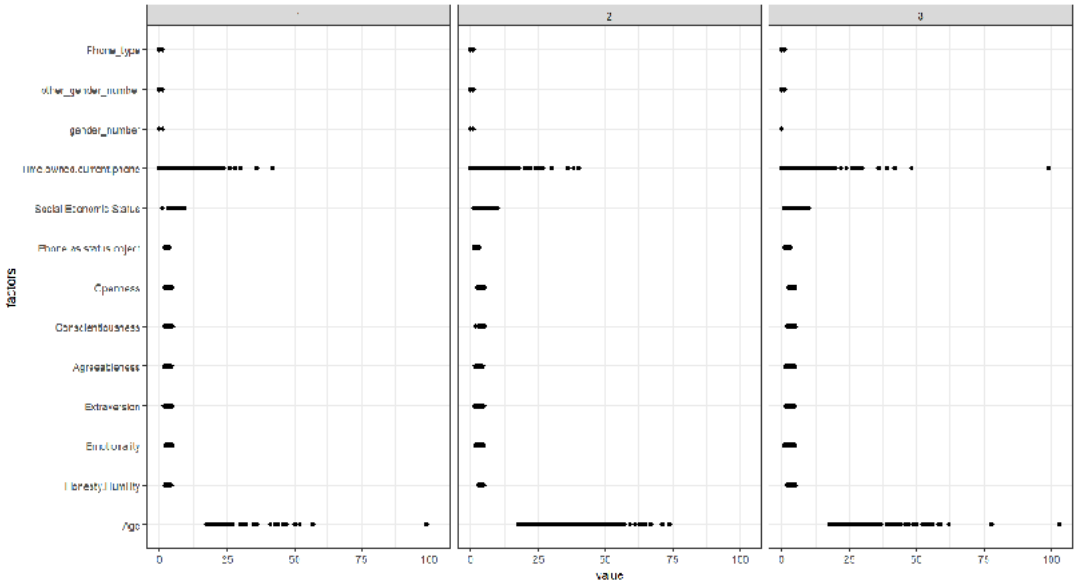

Fig.10: Cluster Analysis

1. Cluster 1 has personality traits of emotionality and hence we can see that there are a greater number of iPhones preferred. phones are used as a status object.
2. Cluster 3 has personality traits of Openness and hence we can see that there are a greater number of Androids preferred. Android phones are more likely to be a liability when comes to getting data from mobile phones. People who generally don't consider it as a liability uses more android phones.
3. Cluster 2 has personality traits of conscientiousness and hence we can see that both iPhones and androids preferred. Users who are more conscious about the cost may prefer androids

whereas users who are more conscious about their data or status prefer iPhones. This cluster shows more honesty-humility personality.

4. Cluster 2 has users who are of age between 25 – 35 which can depict that this age group has conscious use of phones.

Classification

Data Classification two steps: first, a classification model is constructed, and then it is used to predict the class labels of new data (Han, J., Pei, J., Kamber, M., 2011, pp.327-330).

To construct decision trees, the recursive partitioning heuristic is commonly used (Yobero, C., 2018). This involves selecting the feature with the highest predictive power for the target class at each node and then partitioning the data based on the feature's distinct values. The process continues until a stopping criterion is met (Yobero, C., 2018).

C5.0 is a popular implementation of decision trees that is easy to use and performs well in most scenarios (Yobero, C., 2018). It uses entropy, a measure of the impurity of segments, to select the best feature to split on (Yobero, C., 2018). Entropy is 0 for pure segments and 1 for maximum disorder. The algorithm calculates the information gain to evaluate the effectiveness of a feature in creating homogeneous groups after a split. It weighs each partition's entropy by the proportion of records falling into that partition to calculate the total entropy across all partitions (Yobero, C., 2018).

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

Fig.11: Entropy

In our case, the prediction of smartphone is done on the testing data and a confusion matrix is constructed (Han, J., Pei, J., Kamber, M., 2011, pp.364-365). The following figure shows the confusion matrix.

Phone_prediction	Android	iPhone
Android	14	11
iPhone	25	45

Fig.12: Confusion Matrix of test data

		Predicted class: yes	Predicted class: no	
Actual class	yes	TP	FN	P
	no	FP	TN	N
	Total	P′	N′	P + N

Fig.13: Confusion Matrix

The F-measure assesses classifier precision and recall using the harmonic mean (Han, J., Pei, J., Kamber, M., 2011, pp.368-369), with a range of 0-1. A higher score indicates better model performance. The model used for the prediction has a F-score of 0.43. The accuracy of the model is 0.62.

$$precision = \frac{TP}{TP+FP} \qquad F = \frac{2 \times precision \times recall}{precision + recall}$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{P}.$$

Fig.14: F-Score

Discussion on Results

The Decision tree has classified 59 out of 95 samples correctly. (accuracy = 62%). In training model, it is known that out of 384 case 313 are classified correctly. This may occur due to the more occurrences of iPhone in the given dataset. Equal distribution of the data would help in good classification. The following picture shows the importance of the attributes. Age is of 100% variable importance.

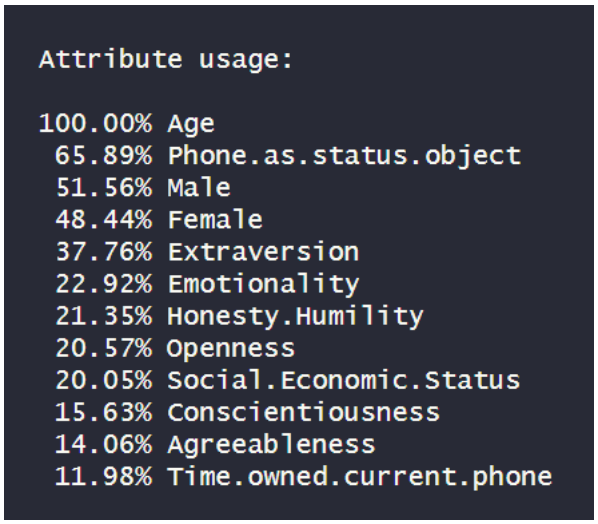

Fig.: attribute usage for classification

Conclusion

The Data mining and machine learning techniques like EDA, clustering and classification is being carried out on the dataset . The model used to cluster is K means which has a limitation of working on clusters of different size or dataset with outliers mean (Han, J., Pei, J., Kamber, M., 2011, p. 454). It is also a necessity that the user provides the k-value, and this is often seen as a disadvantage mean (Han, J., Pei, J., Kamber, M., 2011, p. 454 The model used to classify is C5.0 which has a limitation of overfit or underfit the model (Yobero, C., 2018), small changes in training data can result in large changes to decision logic and often biased towards splits on features having a large number of levels leading to a huge tree which makes the interpretation difficult (Yobero, C., 2018). The results shown after clustering or classification may vary depending on iterations or parameter change. But this is considered as one of the possible results.

References

1. Han, J., Pei, J., Kamber, M., 2011. *Data Mining: Concepts and Techniques.* 3rd ed. Saint Louis: Elsevier Science (The Morgan Kaufmann series in data management systems).
2. Yobero, C., 2018. *Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees* [online]. Available from: https://rpubs.com/cyobero/C50 [Accessed 03 May 2023].