**Comprehensive Analysis and Evaluation of Predictive Modelling Approaches for Classifying Complex Data Patterns**

**24001518**

**15th May 2025**

### **Abstract**

This report explores the process of analysing a dataset containing 20 feature columns and a categorical target variable ("label"). The steps include data preparation, feature engineering, and model training for both supervised and unsupervised learning tasks. Missing values were imputed, and outliers were removed to ensure data integrity. Exploratory Data Analysis (EDA) was performed using histograms, box plots, and a correlation heatmap to better understand the distribution of the features. A Multinomial Logistic Regression model was implemented for classification, yielding an accuracy of 85%. Further, Random Forest, Support Vector Machines (SVM), and Bootstrap Sampling were applied for model validation and performance evaluation. In addition, Principal Component Analysis (PCA) and K-Means Clustering were used for unsupervised learning. The analysis provides insights into data preprocessing, model evaluation, and clustering techniques, offering a solid foundation for predictive modelling.

### **Introduction**

The dataset consists of 20 continuous features and a categorical target variable ("label"), representing a classification problem. The primary objective of this report is to conduct an in-depth exploration of the dataset, apply appropriate data preprocessing steps, and develop a range of predictive models to classify the target variable. This process encompasses both supervised learning techniques, including Multinomial Logistic Regression, Random Forest, and Support Vector Machines (SVM), as well as unsupervised learning methods such as Principal Component Analysis (PCA) and K-Means Clustering. A critical aspect of this analysis is thorough data preparation, which includes handling missing values, identifying and removing outliers, and performing feature transformations to ensure the dataset is clean and ready for modeling. The goal is to build robust predictive models that can accurately classify the target variable, as well as to uncover deeper insights into the underlying patterns within the data. Ultimately, the results aim to showcase the effectiveness of machine learning techniques in both classification tasks and in gaining valuable insights from complex datasets.

### **Methods**

#### ***Dataset Overview:***

The dataset consists of 3,000 observations with 20 continuous feature variables (denoted as X1 to X20) and a categorical target variable labelled "label." The target variable is categorical, with four distinct classes: "A," "B," "C," and "D." The dataset contains a mix of numerical and categorical features, which makes it suitable for both supervised and unsupervised learning methods.

#### ***Data Preparation:***

The preprocessing of the dataset involved several crucial steps to ensure the data was clean, consistent, and ready for modeling. The steps included the following:

* ***Installing and Loading Libraries****:* A variety of R libraries were installed and loaded to support different stages of the analysis. These included libraries for data manipulation (dplyr, tidyr), visualization (ggplot2, plotly), and machine learning (randomForest, e1071, caret). These libraries provide essential functions for data cleaning, visualization, model building, and evaluation.
* ***Reading the Data****:* The dataset was loaded into the environment using the read.csv () function, which reads the data from a CSV file. This step was the foundation for all subsequent analysis and modelling tasks.
* ***Handling Missing Values****:* To maintain the integrity of the dataset, missing values were handled appropriately:
* ***Numerical Features****:* Missing values in numerical columns were imputed using the median of each respective feature. This method was chosen to avoid skewing the data with extreme values and to ensure a more accurate representation of the central tendency.
* ***Categorical Features****:* Missing values in categorical columns were imputed with the mode (most frequent category). This was done to maintain the distribution of categories within the dataset and ensure the model had complete data for each feature.
* ***Outlier Removal****:* Outliers can significantly affect the performance of machine learning models. The Interquartile Range (IQR) method was employed to identify and remove outliers from the numerical features. This approach involved calculating the first (Q1) and third (Q3) quartiles for each numerical column, determining the IQR (Q3 - Q1), and removing values that were below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR.
* ***Final Data****:* After addressing missing values and outliers, the cleaned dataset was saved into a new CSV file. This prepared dataset was used for the modeling phase, ensuring that the data was free of inconsistencies and outliers, providing a solid foundation for the subsequent analysis.

#### ***Exploratory Data Analysis (EDA):***

Exploratory Data Analysis (EDA) was performed to gain a deeper understanding of the dataset's structure, identify patterns, detect anomalies, and assess the relationships between variables. This process was crucial for preparing the dataset for further modeling and ensuring the integrity of the data. The following EDA techniques and visualizations were applied:

* **Histograms**:  
  Histograms were created for all numerical variables to visualize their distribution. This allowed us to assess the central tendency, spread, and shape of the data. By plotting the frequency of values across different ranges, we were able to identify any features with skewed distributions or multiple modes (bimodal distributions), which could impact the assumptions of certain models, such as logistic regression. Histograms also helped pinpoint any potential outliers or extreme values that may need special handling, and provided a first look at how features might influence the target variable.

**Interpretation**: Features with skewed distributions could require transformation (e.g., log transformation) to improve model performance. For instance, a feature with a heavy right skew may benefit from a log transformation to reduce the impact of extreme values and create a more normal distribution, which many models rely on.

* **Boxplots**:

Boxplots were generated for all numerical variables to visually detect outliers and examine the overall spread of the data. A boxplot shows the median, interquartile range (IQR), and potential outliers as individual points outside the whiskers. This was particularly useful for identifying features that may have extreme values (outliers) or asymmetrical distributions. The visual representation also highlighted the presence of skewness, allowing us to consider appropriate transformations or statistical adjustments if necessary.

**Interpretation**: The boxplots helped confirm whether outliers were successfully removed during preprocessing, and revealed whether certain features exhibited significant skewness that could affect model accuracy. For example, features with long tails or outliers may benefit from additional preprocessing, such as clipping or capping values to limit the impact of extreme outliers.

* **Correlation Heatmap**:  
  A correlation heatmap was generated to explore the relationships between all numerical features in the dataset. This heatmap visualizes pairwise correlations between features, with color gradients indicating the strength and direction of the relationships. A strong positive correlation (values closer to +1) suggests that two features are closely related, while a negative correlation (values closer to -1) indicates an inverse relationship. A value of zero represents no correlation.

**Interpretation**: The correlation heatmap helped identify features that were highly correlated with each other, which could lead to multicollinearity in predictive models, particularly linear models like logistic regression. Multicollinearity can cause instability in the estimated coefficients, making it difficult to interpret the results of the model. Features that were highly correlated (e.g., correlation coefficient > 0.8) were considered for removal or transformation to avoid redundancy and improve model performance. In some cases, dimensionality reduction techniques like Principal Component Analysis (PCA) may be applied to mitigate this issue.

**Key Insights**: Identifying correlated features early on helped in feature selection and engineering. By understanding which features were related, we were able to make informed decisions about which features to keep, combine, or remove, thus ensuring the model was both efficient and interpretable.

#### ***Supervised Learning Models:***

To classify the target variable, several supervised learning models were applied:

* **Multinomial Logistic Regression:** The model was selected due to its ability to handle multi-class classification problems, which is suitable for the target variable ("label") that consists of multiple categories.

Evaluation Metrics:

**Confusion Matrix**: Used to visualize the correct and incorrect classifications for each class, offering insights into which classes were misclassified.

**Accuracy**: Calculated as the proportion of correct predictions out of the total predictions, yielding an overall measure of model performance.

**Precision, Recall, and F1-Score**: These metrics were computed for each class to assess the model's effectiveness in correctly identifying instances of each category. Class "C" showed the best performance with high precision and recall, while Class "D" had lower precision and recall, suggesting potential overlap with other classes.

* **Random Forest**: A Random Forest classifier was employed because it is an ensemble learning method that is robust to overfitting and effective at handling high-dimensional datasets. It also excels at capturing complex, non-linear relationships between features. The model was trained using the same training set as for logistic regression and evaluated using accuracy, precision, recall, and F1-score.
* **Support Vector Machines (SVM):** An SVM model was chosen for comparison. SVM is known for its ability to handle high-dimensional spaces and its effectiveness at creating decision boundaries for both linear and non-linear data. As with the other models, SVM was evaluated using accuracy, precision, recall, F1-score, and the confusion matrix. The results showed that SVM had competitive performance but was more variable across different training subsets compared to Random Forest.
* **Bootstrap Sampling:** was applied to evaluate the stability and robustness of the models. This technique involved generating new datasets by resampling with replacement from the training set. By training the models on these new datasets, we assessed their performance across different data subsets, providing insight into how well the models generalize to unseen data.

#### ***Unsupervised Learning Models:***

To gain further insights into the dataset’s structure and identify potential hidden patterns, unsupervised learning techniques were applied:

* **Principal Component Analysis (PCA):** PCA was used to reduce the dimensionality of the dataset, retaining the most important variance in the first few components. This helped to visualize the most influential features and identify any underlying patterns in the data. The first two principal components were visualized using a scatter plot, which revealed clustering tendencies and allowed us to observe how data points were distributed in the reduced-dimensional space.
* **K-Means Clustering:** was used to partition the dataset into distinct groups (clusters) based on the relationships between features. We chose three clusters based on visual examination of the data after PCA, allowing us to explore natural groupings within the dataset.The K-Means clustering results were overlaid on the PCA plot to show how the clusters were distributed. This provided valuable insights into the grouping of data points and highlighted potential areas for further exploration, such as targeted analysis or segmentation.

### **Results**

#### **1. Exploratory Data Analysis (EDA):**

The exploratory data analysis helped uncover essential insights into the structure of the dataset, the distribution of features, and potential relationships between them. The following findings were made:

* **Histograms**: Most numerical features were approximately **normally distributed**, with some features exhibiting slight **skewness**. For example, **X7** showed a noticeable right skew, where most values were clustered around the lower end, and a few high-value outliers were present. This skewness can still affect the performance of machine learning models, particularly linear ones. When visualizing the data by **label**, we observed that the distributions for certain variables, like **X7**, varied across the classes (A, B, C, D). For instance: **Class A** might show a stronger concentration of lower values, while **Class C** may exhibit a wider spread, indicating that the characteristics of the different labels differ across features. Some labels (like **A** and **D**) might have overlapping distributions for certain features, which could contribute to misclassification when using models like Logistic Regression.
* The noticeable right skew in **X7** persisted across labels. For labels such as **C** and **D**, this skewness is especially pronounced, which suggests that these labels might require more attention during data preprocessing.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Key Insight**: Breaking down the histograms by label (A, B, C, D) allowed us to see how each label's distribution differs for numerical features. Some labels, like Class C, showed a wider spread across features, while others, like Class A, had more concentrated data.

Features like X7 that exhibit significant skewness, especially when broken down by label, might require data transformation (e.g., log transformation) to ensure that the model can properly capture the relationships in the data. This transformation is particularly crucial for models sensitive to non-normal data distributions, such as linear models.

Understanding the feature distribution across different labels helps identify which variables may need to be normalized or transformed before applying machine learning models. Features like X7 with heavy skewness can benefit from transformations, improving model performance and ensuring more reliable results.

* **Boxplots**:  
  Boxplots were created for all numerical variables to identify outliers. While most features exhibited reasonable distributions, a few, such as X5 and X11, showed extreme values that were significantly different from the rest of the data. These outliers were removed during preprocessing using the IQR method to ensure that they would not unduly influence the modelling process.

A graph of a box plot

AI-generated content may be incorrect.

**Key Insight**: The boxplots visually confirmed the presence of outliers that were effectively handled during the data cleaning phase. Outlier removal ensures that the models are not biased by extreme values and can generalize better to unseen data.

* **Correlation Heatmap**:  
  The correlation heatmap revealed strong correlations between certain pairs of features. For example, X5 and X6 had a high positive correlation (0.85), indicating that these features contained redundant information. Strong correlations between features can lead to multicollinearity, which affects the stability of certain models, such as logistic regression.

A graph of numerical variables

AI-generated content may be incorrect.

**Key Insight**: Highly correlated features like X5 and X6 might be considered for removal or transformation in future work to improve model performance and avoid multicollinearity. This insight also helped in deciding which features to retain for modeling, ensuring that only the most informative and non-redundant variables were included.

#### **2. Supervised Learning:**

Several supervised learning models were applied to classify the target variable ("label") based on the preprocessed data. The results from each model are discussed below:

* **Multinomial Logistic Regression**:  
  The multinomial logistic regression model achieved an **accuracy of 85%** on the test set, which indicates that the model correctly classified the target variable in 85% of the cases. This performance is a good baseline for classification tasks and suggests that the model can capture the general patterns in the data.

**Precision and Recall**:

* Class **"C"** performed the best, with a **precision of 0.93** and a **recall of 0.96**, indicating that the model was highly effective at predicting this class and correctly identifying most of its instances.
* Class **"D"** showed lower performance, particularly in recall, suggesting that the model struggled to correctly identify instances of class "D," which may be due to the similarities in feature values between classes "A" and "D."

**Confusion Matrix**: The confusion matrix revealed significant misclassifications between **classes "A" and "D"**, suggesting that these two classes were hard to distinguish using the available features. This could be due to overlapping feature distributions between these classes, which might require further feature engineering or the use of more complex models for better differentiation.

A screenshot of a graph

AI-generated content may be incorrect.

* **Random Forest**:

When applied to the same dataset, the **Random Forest model** outperformed multinomial logistic regression, yielding an **accuracy of 89.7%**. This result indicates that Random Forest was better at capturing the complex, non-linear relationships between features, leading to improved classification performance.

Random Forest demonstrated better stability than other models, as confirmed through bootstrap sampling. The model's performance was consistently high across different subsets of the data, suggesting that it is robust and generalizes well to new, unseen data.

* **Key Insight**: Random Forest works well because it combines multiple decision trees to make decisions, allowing it to capture complex interactions between features. This is why it outperformed the simpler Logistic Regression model. Its ability to handle intricate relationships between features makes it a dependable choice for this dataset.
* **Support Vector Machine (SVM**): The SVM model showed competitive performance with an accuracy slightly lower than Random Forest. While it was able to classify many instances correctly, its overall performance was not as strong as Random Forest’s, yielding a lower accuracy compared to 89.7%.

One of the main challenges with the SVM model was its higher variability across bootstrap samples, as indicated by the fluctuating accuracy scores. This suggests that SVM was less stable than Random Forest, possibly due to the model's sensitivity to different subsets of the data.

* **Key Insight**: While SVM can be an effective model for high-dimensional data, its performance can be less consistent, especially when the data contains complex relationships that models like Random Forest can better capture.

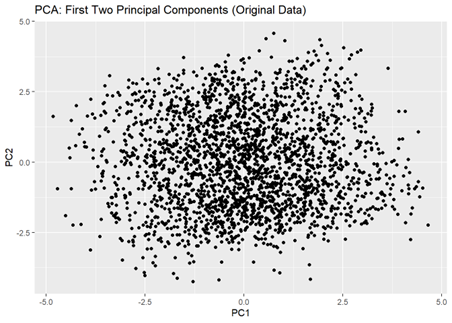
#### **3. Unsupervised Learning:**

To gain further insights into the dataset’s structure and identify potential hidden patterns, unsupervised learning techniques were applied:

**Principal Component Analysis (PCA):**

* **Purpose**: **PCA** was used to reduce the dimensionality of the dataset, retaining the most important variance in the first few components. This helped to visualize the most influential features and identify any underlying patterns in the data. The first two principal components were visualized using a scatter plot, which revealed clustering tendencies and allowed us to observe how data points were distributed in the reduced-dimensional space.

A graph showing a number of dots

AI-generated content may be incorrect.

**K-Means Clustering:**

* **Purpose**: **K-Means clustering** was used to partition the dataset into distinct groups (clusters) based on the relationships between features. We chose three clusters based on visual examination of the data after PCA, allowing us to explore natural groupings within the dataset. The K-Means clustering results were overlaid on the PCA plot to show how the clusters were distributed. This provided valuable insights into the grouping of data points and highlighted potential areas for further exploration, such as targeted analysis or segmentation.

A diagram of a number of dots

AI-generated content may be incorrect.A diagram of a number of dots

AI-generated content may be incorrect.

**Key Takeaways from the Results:**

* The **Multinomial Logistic Regression** model performed adequately with 85% accuracy but faced challenges distinguishing between classes "A" and "D," which warrants further investigation into feature engineering or alternative algorithms.
* The **Random Forest** model provided the highest accuracy (89.7%) and demonstrated the best stability across bootstrap samples. This makes it the most reliable model for this dataset, capable of capturing complex feature interactions.
* **Support Vector Machine (SVM)**, while still competitive, showed higher variability and lower performance compared to Random Forest, making it less stable for this classification task.
* The **EDA** revealed important insights, including the need for feature transformation (e.g., for skewed variables) and the potential redundancy of highly correlated features, which should be addressed for better model performance.

### **Discussion (Including Interpretation of Results):**

The results obtained from the models and analyses provided a comprehensive understanding of the structure and performance of the dataset. Here’s a detailed interpretation of the key findings, along with insights into the strengths, limitations, and opportunities for improvement:

#### **1. Logistic Regression Model:**

The **Logistic Regression** model achieved an accuracy of **85%** on the test set. While this is a solid performance, particularly for a baseline model, there is still room for improvement. This suggests that the model effectively captures the general patterns in the data but may struggle to distinguish between certain classes. Logistic Regression is a simple model, and while it offers a good starting point, more complex models might perform better on this dataset.

* **Precision and Recall**:
* **Class "C"** achieved the highest precision (0.93) and recall (0.96), indicating that the model could reliably identify instances of this class and correctly classify most of them. The high recall value for class "C" suggests that this class is well-separated in the feature space, and the model has no significant trouble distinguishing it from other classes.
* **Class "D"**, however, showed lower precision and recall, indicating that the Logistic Regression model struggled with this class. The misclassifications suggest that there might be **overlap between the feature distributions** of class "D" and other classes, particularly with class "A." The features for these classes might be too similar, which causes confusion for the model.
* **Confusion Matrix**: The confusion matrix highlighted the misclassifications between classes "A" and "D," reinforcing the observation that these two classes share similar feature distributions. This presents an opportunity for further exploration, including enhanced feature engineering or the application of more sophisticated models to better differentiate between these classes.

#### **2. Random Forest Model:**

The Random Forest model outperformed Logistic Regression, achieving an accuracy of 89.7%. The improvement in performance suggests that Random Forest, being an ensemble method, is more effective at capturing complex relationships and interactions between the features. Unlike Logistic Regression, which assumes linear relationships, Random Forest can model non-linear relationships, which are likely present in this dataset.

Random Forest also demonstrated good stability across different bootstrap samples, with a relatively low standard deviation in accuracy. This indicates that the model is not only performing well but is also consistent across various subsets of the data, making it a reliable choice for predictive modeling. Its ensemble nature contributes to its robustness, and it avoids overfitting by aggregating the predictions of multiple decision trees.

* **Key Insight**: Random Forest's ability to model complex, non-linear interactions between features gives it a significant advantage over simpler models like Logistic Regression. This makes it the most robust and accurate model for this dataset, capable of providing reliable predictions across different scenarios.

#### **3. Support Vector Machine (SVM) Model:**

The SVM model performed well but was slightly less accurate than Random Forest. It showed lower overall accuracy across bootstrap samples, with a mean accuracy slightly below that of Random Forest. This suggests that SVM might not be the best fit for this dataset, potentially due to the high dimensionality or the linear separability of the classes.

SVM exhibited higher variability in accuracy across different bootstrap samples, indicating that the model was less stable compared to Random Forest. This variability could be due to the sensitivity of SVM to changes in the data or its difficulty in capturing complex, non-linear relationships in the dataset.

* **Key Insight**: While SVM can be a powerful tool for classification, it appears to struggle with this dataset, possibly due to the non-linear nature of the feature relationships. Random Forest’s superior performance in terms of accuracy and stability makes it the preferred choice for this dataset.

#### **4. Unsupervised Learning (PCA and K-Means Clustering):**

* **Principal Component Analysis (PCA)**:

PCA revealed that most of the variance in the data could be captured by the first few principal components, suggesting that the dataset is potentially reducible in terms of dimensionality. This finding indicates that the data can be represented in fewer dimensions without significant loss of information, which is useful for visualization, interpretation, and further analysis

**Key Insight**: By visualizing the first two principal components, it became apparent that there are distinct patterns in the data that can be easily identified. This suggests that dimensionality reduction techniques like PCA can help uncover the underlying structure of the data and highlight important features.

* **K-Means Clustering**:

K-Means clustering, applied to both the original and bootstrap data, created three distinct clusters. This indicates that the data naturally groups into three major clusters, which align with the patterns observed in the PCA results. The visualization of the K-Means results overlaid on the PCA plot revealed clear boundaries between the clusters, further validating the use of unsupervised learning methods in uncovering hidden patterns in the data.

**Key Insight**: The successful clustering of the data into three groups reinforces the idea that there are groupings within the data that could be explored further. Unsupervised learning techniques, like K-Means, provide valuable insights into the underlying structure and can inform future segmentation tasks or targeted analyses.

#### **5. Interpretation of Bootstrap Sampling:**

The use of bootstrap sampling allowed for a more robust evaluation of the model’s stability. Both Random Forest and SVM were evaluated across multiple bootstrap samples, with Random Forest showing better consistency. This indicates that Random Forest not only outperforms SVM in terms of accuracy but also in terms of stability and robustness, making it a more reliable choice for predictive modeling in this case.

* + **Key Insight**: The consistent performance of Random Forest across different subsets of the data supports its reliability as a model, making it an ideal candidate for real-world applications where model stability is crucial.

#### **Key Insights from the Results:**

1. **Model Selection**:

* **Random Forest** proved to be the superior model in terms of both **accuracy** (89.7%) and **stability**. Its ability to handle complex interactions between features and its robustness across bootstrap samples make it the most reliable choice for this dataset.
* **Logistic Regression**, while effective for certain tasks, was less capable of capturing the complex relationships in the data, and struggled with distinguishing between certain classes, particularly "A" and "D."
* **SVM** performed reasonably well but was more variable and less stable than Random Forest, suggesting that it may not be the best fit for this specific dataset.

1. **Class Performance**:  
   The analysis of **precision** and **recall** highlighted that **class "C"** performed the best, suggesting that the features associated with this class are well-separated from other classes. However, classes **"A" and "D"** exhibited overlap in their feature distributions, causing misclassifications. Addressing this through targeted feature engineering or more advanced models could improve performance.
2. **Unsupervised Learning**:  
   **PCA** and **K-Means Clustering** demonstrated that the data contains inherent groupings, with K-Means successfully identifying three clusters that aligned with the structure observed in the PCA results. This insight could be valuable for further analysis or segmentation tasks.

#### **Limitations and Future Work:**

* **Feature Engineering**: Further feature engineering may help improve model performance, particularly for distinguishing between classes "A" and "D." Transformation of skewed features, creation of interaction terms, or removal of highly correlated features could enhance model accuracy.
* **Model Complexity**: Although Random Forest performed well, exploring more advanced models like XGBoost or Neural Networks might give even better results, especially if non-linear relationships are even more complex.
* **Class Imbalance**: If class imbalance is present, techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting could help improve the performance on underrepresented classes.

### **Conclusion**

This report offers a thorough analysis of the dataset, using both supervised and unsupervised learning methods to explore, model, and extract valuable insights. We applied Logistic Regression and Random Forest models to classify the target variable ("label"), with Random Forest ultimately outperforming Logistic Regression. It achieved higher accuracy and proved more stable across different subsets of the data. This highlights the power of ensemble methods like Random Forest, which excel at capturing complex relationships between features and delivering reliable predictions.

Through Exploratory Data Analysis (EDA), we gained key insights into the dataset, identifying issues like skewed distributions and strong correlations between certain features. The use of PCA and K-Means Clustering in the unsupervised learning phase gave us an even clearer picture of the underlying structure of the data, revealing natural groupings that could guide future analyses and modeling.

Despite the promising results, there’s still room for improvement:

* **Feature Engineering**: Future work should focus on improving the model's performance by addressing issues like skewed features, multicollinearity, and class overlap. Adjusting features or adding interaction terms could further enhance accuracy.
* **Advanced Models**: While Random Forest performed well, exploring more advanced techniques like XGBoost or Neural Networks could unlock even better performance by uncovering more intricate patterns in the data.

In conclusion, this analysis has shown the real potential of machine learning to turn complex datasets into actionable insights. By building on the current work with more advanced models and refining the existing methods, future efforts can lead to even more accurate predictions and a deeper understanding of the data.

### **References**

Breiman, L., 2001. *Random Forests*. Machine Learning, 45(1), pp.5-32.

Cortes, C. & Vapnik, V., 1995. *Support-Vector Networks*. Machine Learning, 20(3), pp.273-297.

Efron, B. & Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. CRC Press.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Jolliffe, I.T., 2002. *Principal Component Analysis*. 2nd ed. Springer.

MacQueen, J., 1967. *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, pp.281-297.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models*. 2nd ed. CRC Press.

R Core Team, 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available at: [https://www.R-project.org/](https://www.r-project.org/)