ST3189

MACHINE LEARNING

COURSEWORK

STUDENT NUMBER: 210421794

Table of Contents

1. Unsupervised Learning

1.1 Substantive Issue

In today's competitive marketplace, conducting a customer personality analysis is crucial for businesses to better understand and effectively engage with their diverse customer base. This analysis seeks to identify the company's target customer by exploring their specific needs and behaviours across various segments. Such insights allow a company to modify its products and marketing strategies precisely, ensuring they are closely aligned with the target customer's needs and expectations. Therefore, we will apply unsupervised learning techniques to analyse and categorize customers into distinct segments for more targeted and effective business strategies.

Based on the issue above, we have identified the following research questions (RQs):

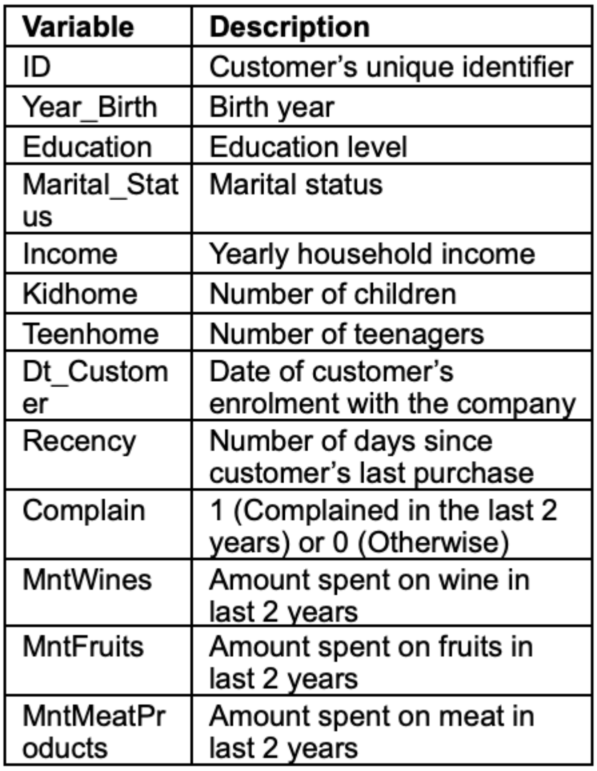
- RQ1: What is the optimal number of clusters to segment the customers?
- RQ2: What is the interpretation of the identified clusters?

1.2 Methodology

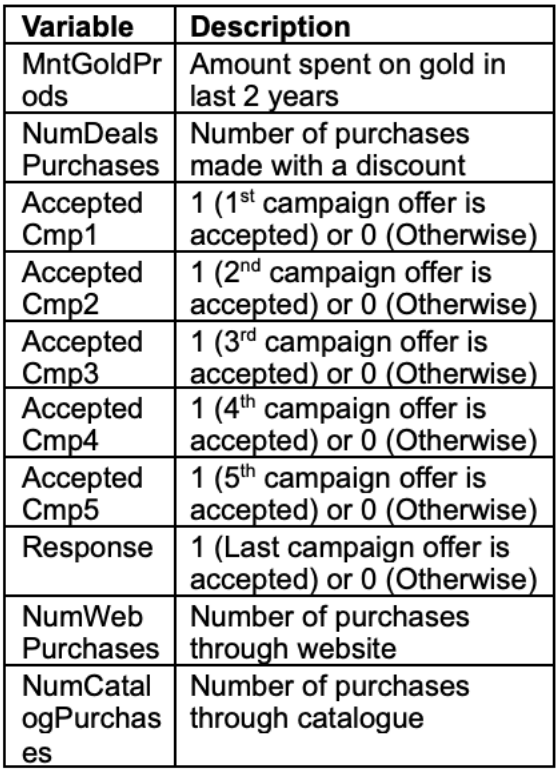
Prior to applying any techniques, we will undertake Exploratory Data Analysis for data cleaning, preparation, and exploration purposes. Following this, unsupervised learning techniques such as Principal Component Analysis (PCA), K-Means Clustering, and Hierarchical Clustering will be used to analyse customer data with the goal of categorizing customers into distinct segments according to their behaviours. To determine the optimal number of clusters for our analysis, we will utilize both elbow and silhouette methods.

1.3 Dataset and Variables

This customer dataset initially comprises 29 variables and 2240 observations but contains missing values and extreme outliers that will be removed. We will also introduce new variables, 'Age', and 'Child', to facilitate easier data interpretation. Additionally, we will eliminate any unnecessary or redundant columns, resulting in the dataset now consisting of 24 variables with a total of 2212 observations.



Variable	Description
ID	Customer's unique identifier
Year_Birth	Birth year
Education	Education level
Marital_Status	Marital status
Income	Yearly household income
Kidhome	Number of children
Teenhome	Number of teenagers
Dt_Customer	Date of customer's enrolment with the company
Recency	Number of days since customer's last purchase
Complain	1 (Complained in the last 2 years) or 0 (Otherwise)
MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years



Variable	Description
MntGoldProds	Amount spent on gold in last 2 years
NumDeals Purchases	Number of purchases made with a discount
Accepted Cmp1	1 (1st campaign offer is accepted) or 0 (Otherwise)
Accepted Cmp2	1 (2nd campaign offer is accepted) or 0 (Otherwise)
Accepted Cmp3	1 (3rd campaign offer is accepted) or 0 (Otherwise)
Accepted Cmp4	1 (4th campaign offer is accepted) or 0 (Otherwise)
Accepted Cmp5	1 (5th campaign offer is accepted) or 0 (Otherwise)
Response	1 (Last campaign offer is accepted) or 0 (Otherwise)
NumWeb Purchases	Number of purchases through website
NumCatalogPurchases	Number of purchases through catalogue

MntFishProducts	Amount spent on fish in last 2 years	NumStorePurchases	Number of purchases through store
MntSweetProds	Amount spent on sweets in last 2 years	NumWebVisitsMonth	Number of visits to website in the last month
Z_CostContact	-		
Z_Revenue	-		

Table 1: Customer Dataset Variables Description

1.4 Analysis

In preparing our customer dataset for unsupervised learning, the Exploratory Data Analysis (EDA) began by addressing missing values in the 'Income' column, removing 24 observations. We then introduce a new 'Age' variable and combine 'Kidhome' and 'Teenhome' into a 'Child' variable for clearer interpretation. 'Marital_Status' and 'Education' are re-categorized and converted into numerical data types. After dropping unnecessary and redundant columns, we visually inspect the distributions of numerical and categorical variables using boxplots and bar charts, respectively. Outliers in 'Income' and 'Age' are identified and removed, refining our data to 2212 observations and 24 variables. Finally, a correlation matrix is created to see the relationship between variables.

1.4.1 Principal Component Analysis (PCA)

PCA is a machine learning technique used to reduce the complexity of high-dimensional data while retaining trends and patterns. It works by transforming potentially correlated variables in a dataset into a smaller set of uncorrelated variables known as principal components. In applying PCA to the customer dataset, we must ensure that all variables are centred and scaled. Based on the Kaiser Criterion, we select the first seven principal components, each with a variance greater than 1, as evident in Figure 1. However, note that these principal components only account for 62.19% of the total variance (cumulative proportion) of the data. The contributions of variables are illustrated through biplots, as exemplified in Figure 2. This PCA process in the customer data helps us to understand customer patterns for subsequent clustering analysis.

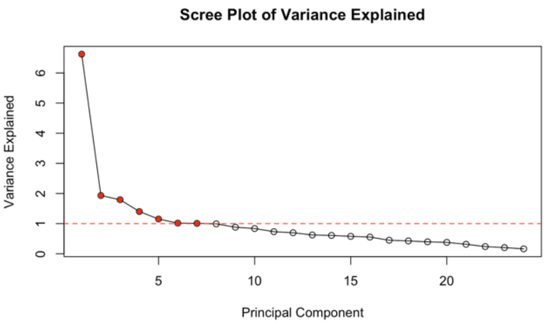


Figure 1: Scree Plot of Variance Explained

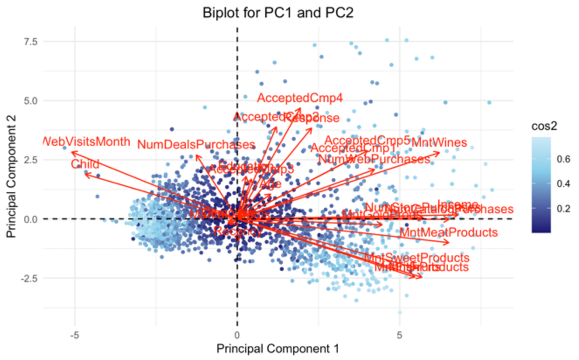


Figure 2: Biplot for Principal Components 1 and 2

1.4.2 K-Means Clustering

K-Means Clustering groups data points into k clusters based on their proximity to the cluster's central point, or centroid, aiming to minimize the sum of squared distances between each data point and its centroid. This process begins with scaling the data and determining the optimal number of clusters using the Elbow and Silhouette methods. Despite a difference in the optimal cluster number suggested by each method, with the Elbow suggesting three clusters and the Silhouette two, we opt to proceed with three clusters for further analysis. Analysis of the clustering results reveals three distinct clusters: “Middle-Income Family with Children” (1st cluster, indicated by pink circles in Figure 4), “High-Income Childfree Family (2nd cluster, indicated by blue triangles), and “Low-Income Family with Children” (3rd cluster, indicated by green squares). These labels are based on economic status, spending and purchasing behaviours, and family composition, as derived from the variable distributions within each cluster.

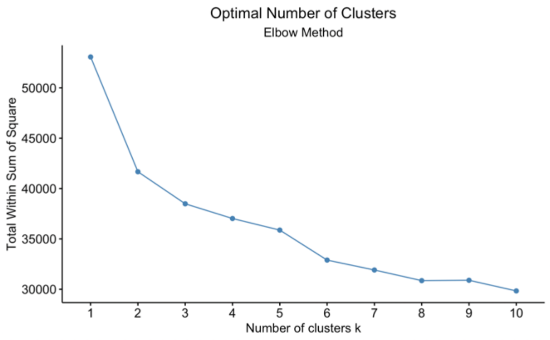


Figure 3: Elbow Method

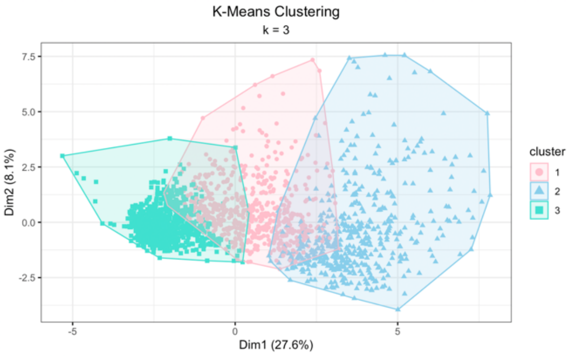


Figure 4: K-Means Cluster Plot (k=3)

1.4.3 Hierarchical Clustering

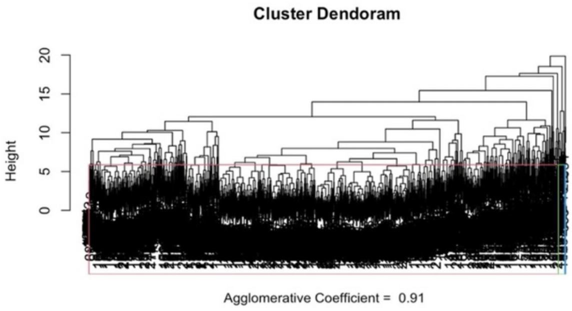


Figure 5: Hierarchical Cluster Plot

In Hierarchical Clustering, data points are grouped into a tree-like structure based on similarity, aiming for creating a hierarchy of clusters. After scaling the data, we apply hierarchical clustering using the Euclidean distance and explore three common linkage methods: complete, single, and average. The agglomerative coefficient, a measure used to assess the degree of similarity among points in the same cluster, indicates that Complete linkage method has the strongest clustering structure (indicated by the value closest to 1). Thus, we select the Complete linkage method as the final model, cutting the tree at k = 3 as recommended by the Elbow method. However, the result shows a highly uneven distribution of data points among these clusters, suggesting that hierarchical clustering may not be the best clustering method for this customer dataset.

1.5 Results

RQ1: What is the optimal number of clusters to segment the customers?

- Based on the elbow method used in both K-Means Clustering and Hierarchical Clustering, three is identified as the optimal number of clusters to segment the customers.

RQ2: What is the interpretation of the identified clusters?

- According to K-Means Clustering, the 3 identified clusters are: Cluster 1 labelled as “Middle-Income Family with Children”, Cluster 2 labelled as "High-Income Childfree Family", and Cluster 3 labelled as "Low-Income Family with Children". Cluster 1 represents a stable household with moderate income level, high overall spending and purchases, often including children in their family structure. Cluster 2 is highlighted by the highest average income and spending levels, along with a notable absence of children in the household. Cluster 3 is characterized by a lower economic status, with corresponding spending and purchasing levels. Notably, this cluster has the youngest average age, possibly indicating they are at the beginning of their family life and career path.

2. Regression

2.1 Substantive Issue

Secondary education serves a crucial phase in an individual's educational journey, fostering intellectual, emotional, and social development. It prepares students for higher education and professional careers by equipping them with essential skills and knowledge. Mastering the national language, in this case Portuguese for students in Portugal, is essential. It is a fundamental skill that facilitates effective communication and critical thinking, making it a key to personal and professional growth. Therefore, our aim is to develop a robust model that accurately predicts student achievement in Portuguese language courses, uncovering significant factors that influence students' grades in Portuguese.

Based on the issue above, we have identified the following research questions (RQs):

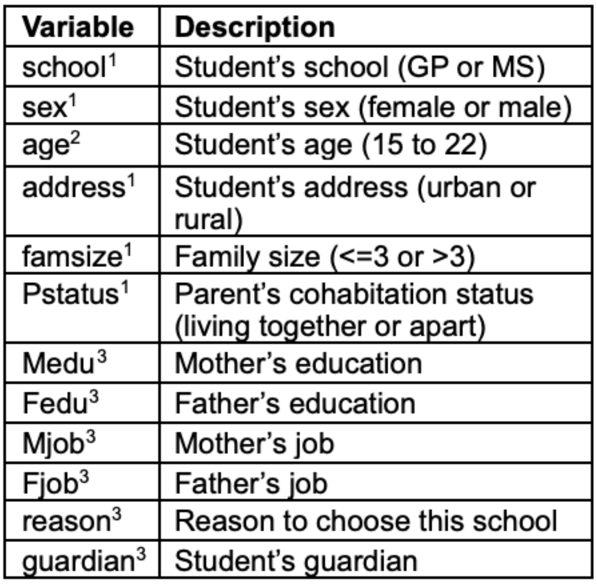
- RQ1: What is the best machine learning model for predicting students' grades in Portuguese (G3)?
- RQ2: What are the most important factors that influence students' grades in Portuguese (G3)?

2.2 Methodology

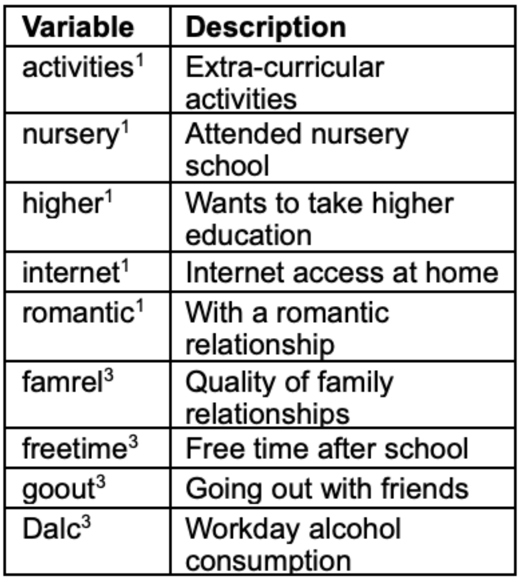
Prior to performing regression tasks, we will conduct Exploratory Data Analysis (EDA) to further investigate the data and uncover correlations among variables, particularly between predictors and the target variable. Our approach will encompass four distinct regression techniques: Linear Regression, Regularization, Random Forest, and Classification and Regression Tree (CART). Subsequently, we will assess the models' performances by comparing their test set Root Mean Squared Error (RMSE) values and determine the most important predictor variables.

2.3 Dataset and Variables

The dataset captures student performance in Portuguese language subject in secondary education of two Portuguese schools (Cortez and Silva, 2008). It consists of 649 observations and 33 variables, with no missing and duplicated values. The dataset encompasses both numerical and categorical variables.



Variable	Description
school[1]	Student's school (GP or MS)
sex[1]	Student's sex (female or male)
age[2]	Student's age (15 to 22)
address[1]	Student's address (urban or rural)
famsize[1]	Family size (<=3 or >3)
Pstatus[1]	Parent's cohabitation status (living together or apart)
Medu[3]	Mother's education
Fedu[3]	Father's education
Mjob[3]	Mother's job
Fjob[3]	Father's job
reason[3]	Reason to choose this school
guardian[3]	Student's guardian



Variable	Description
activities[1]	Extra-curricular activities
nursery[1]	Attended nursery school
higher[1]	Wants to take higher education
internet[1]	Internet access at home
romantic[1]	With a romantic relationship
famrel[3]	Quality of family relationships
freetime[3]	Free time after school
goout[3]	Going out with friends
Dalc[3]	Workday alcohol consumption

[1] Binary (2 Categorical) Variable

[2] Numerical Variable

[3] Categorical Variable (more than 2 categories)

traveltime[3]	Home to school travel time
studytime[3]	Weekly study time
failures[3]	Number of past class failures
schoolsup[1]	Extra educational support
famsup[1]	Family educational support
paid[1]	Extra paid classes
Walc[3]	Weekend alcohol consumption
health[3]	Current health status
absences[2]	Number of school absences
G1[2]	First period grade
G2[2]	Second period grade
G3[2]	Final grade

Table 2: Student Performance Dataset Variables Description

2.4 Analysis

We start by importing the dataset, preparing the data, and conducting initial data exploration through Exploratory Data Analysis. This process includes converting dataset variables to appropriate formats for regression analysis, numeric data type for continuous variable and factor data type for categorical variable. Our goal is to predict student's final grade (G3) in Portuguese, which is identified as our target variable. Notably, 'G3' is strongly correlated with 'G1' and 'G2' (grades from first and second periods), as shown in the correlation plot (Figure 6). Given this, we opt to develop two separate models for each regression method: one including 'G1' and 'G2' as predictors (Model A) and another excluding them (Model B), to evaluate the impact of other variables, which might be overshadowed by the strong influence of 'G1' and 'G2'. Finally, we split the dataset into a train set and test set at a 70:30 ratio before proceeding with regression analysis.

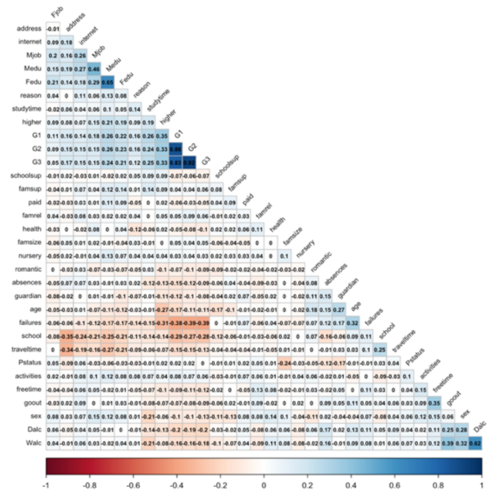


Figure 6: Correlation Plot

2.4.1 Linear Regression

In Multiple Linear Regression model, we estimate the relationship between the dependent (target) variable and two or more independent (predictor) variables. Our aim is to predict the target variable's outcome using the predictors' values. To refine our model, we employ backward elimination to select the most significant variables for a multiple linear regression model. It involves starting with all variables and iteratively removing the least significant variable, until the model no longer shows improvement, as indicated by the Akaike Information Criterion (AIC). After applying backward elimination on models A and B, the resulting regression coefficients show that in model A , a larger number of variables have been eliminated compared to model B. This suggests that 'G1' and 'G2' are dominant explanatory variables, overshadowing the influence of other factors. When 'G1' and 'G2' are not included, the impact of the remaining variables become more significant.

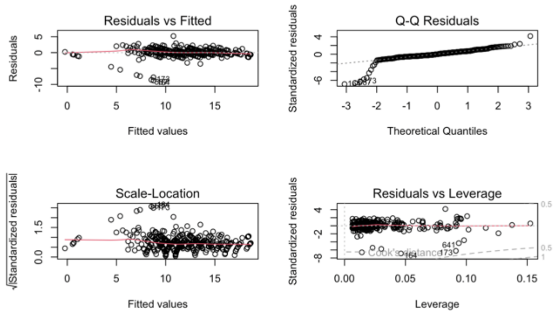


Figure 7: Diagnostic Plot Model A

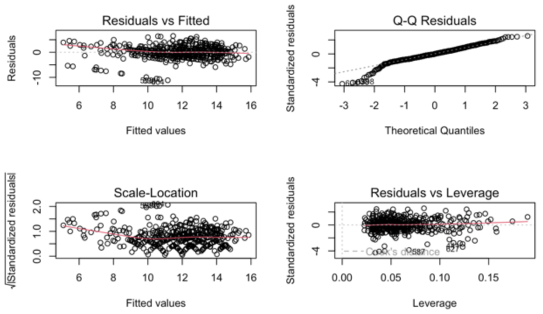


Figure 8: Diagnostic Plot Model B

The diagnostic plots shown in Figure 7 & Figure 8 are based on models A and B after conducting backward elimination to optimize the models. These plots help us examine the linear regression model's assumptions. In the Residual vs Fitted plots for both models, we observe a distinct pattern that may indicate the models do not perfectly capture a linear relationship between 'G3' and predictors. The Normal Q-Q plots exhibit slight deviations of residuals from the dotted straight line but, on the whole, suggest that we can reasonably assume the residuals are normally distributed. For model A, the Scale-Location plot displays a consistent residual spread across the range of predictors, supporting the assumption of homoscedasticity. However, model B's Scale-Location plot shows increased spread with larger fitted values, hinting at potential heteroskedasticity. In the Residuals vs Leverage plots, no data points exceed Cook's distance, implying the absence of influential outliers. The RMSE value for model A is 1.112621, while model B's RMSE value is 2.728180.

2.4.2 Regularization

Regularization techniques, like Ridge and Lasso Regressions, are employed to prevent overfitting in a model with large variance by imposing a penalty to shrink less significant features' coefficients. Ridge Regression shrinks these coefficients towards zero, but hardly hit zero. It still includes all predictors in the final model but with reduced impact. Whereas, Lasso Regression can reduce coefficients of less significant variables to be exactly zero, performing a feature selection, resulting in a simpler model. Both regressions use lambda as a tuning parameter that controls the strength of the penalty applied to the coefficients, determined through cross-validation. Ridge Regression's RMSE for models A and B are 1.180274 and 2.681700, respectively, while Lasso Regression's RMSE are 1.080560 and 2.682352, respectively.

2.4.3 Random Forest

Random Forest is a machine learning technique that constructs multiple decision trees to make predictions by combining the outcomes. It involves building each tree from a randomly selected subset of data and choosing a random subset of features at each split. This approach not only prevents overfitting but also enhances the model's robustness and accuracy. In model A, the RMSE value is 1.121698, with 'G2', 'G1', and 'absences' being the most important variables. Meanwhile, model B has an RMSE of 2.616848 with 'failures', 'school' and 'higher' as the top three predictors.

2.4.4 Classification and Regression Trees (CART)

CART model divides the data into smaller subsets, creating tree-like structure decisions consisting of nodes and branches. It starts from a root node, representing the input variable, and branches into decision nodes, with branches representing possible outcomes. Initially, we grow the tree to its maximum by setting the complexity penalty (CP) to zero. We then determine the optimal CP using two approaches: one based on the minimum cross-validation error (xerror), and another using the 1 Standard Error (SE) rule, choosing a CP that achieves a xerror just below the sum of the minimum xerror and one standard error.

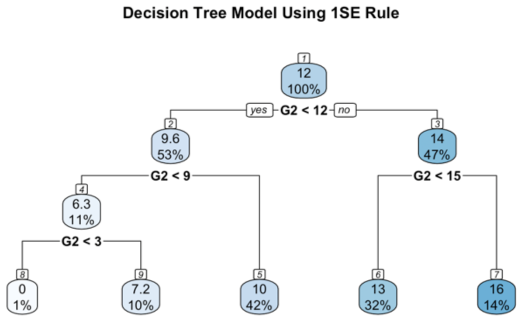


Figure 9: Decision Tree using 1SE Rule (Model A)

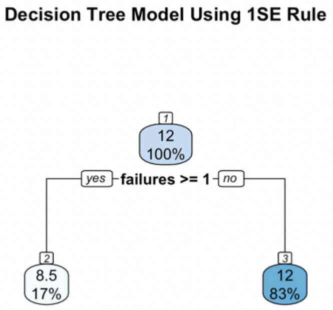


Figure 10: Decision Tree using 1SE Rule (Model B)

After identifying the optimal CP values, we prune the tree accordingly and evaluate the models based on their RMSE values. For model A, using the 1 SE rule results in a lower RMSE compared to the minimum xerror approach. Meanwhile, for model B, both approaches yield the same CP, leading to identical RMSE values, indicating a balance of complexity and accuracy. The models no longer benefit from increased complexity, as shown by stable cross-validation performance. Figures 9 and 10 illustrate the optimal trees for models A and B derived using the 1 SE rule, respectively. For model A, the most important predictors are ‘G2’, ‘G1’, and ‘school’, whereas, for model B, they are ‘failures’, ‘age’, and ‘guardian’.

2.5 Results

RQ1: What is the best machine learning model for predicting students’ grades in Portuguese?

- Model A: The Lasso Regression is the best model to predict ‘G3’ using all variables as it has the lowest RMSE value compared to other regression techniques.
- Model B: The Random Forest is the best model to predict ‘G3’ when ‘G1’ and ‘G2’ are excluded as it has the lowest RMSE value compared to other regression techniques.

Note that the RMSE values observed in model B (excluding G1 and G2) are higher than those in model A (including G1 and G2). This is due to the strong correlation between ‘G1’ and ‘G2’ with the target variable ‘G3’, highlighting their importance as predictors. Without these variables, model B loses the valuable information, leading to less accurate predictions and higher RMSE values.

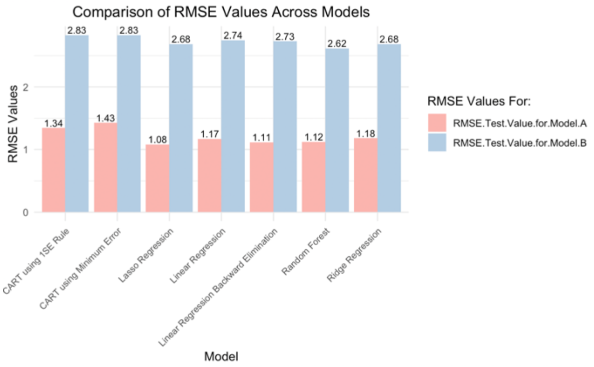


Figure 11: Regression Model Comparison

RQ2: What are the most important factors that influence students’ grades in Portuguese?

Given that Random Forest has shown robust performance for both models A and B, evidenced by its RMSE values being among the lowest compared to other techniques, we will use Random Forest variable importance analysis to address this question.

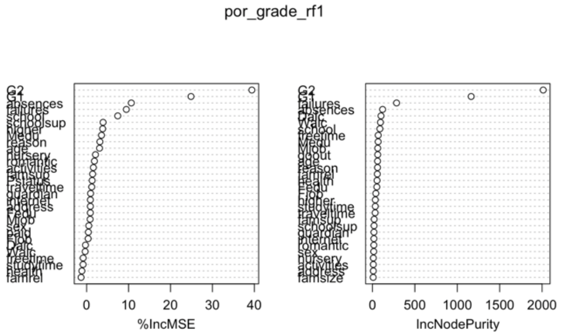


Figure 12: Variable Importance Model A

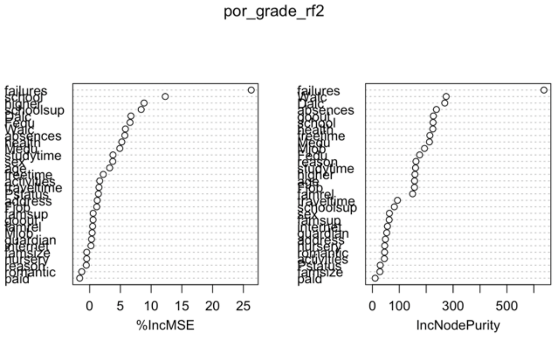


Figure 13: Variable Importance Model B

- Model A: The most important variables that influence ‘G3’ are second period grade (G2), first period grade (G1), and number school absences (absences).
- Model B: The most important variables that influence ‘G3’ are number of past class failures (failures), student’s school (school), and whether the student wants to take higher education (higher).

3. Classification

3.1 Substantive Issue

Diabetes, a chronic disease, presents a significant challenge to global public health. This dataset aims to use diagnostic measurements to accurately predict the presence of diabetes in patients, focusing particularly on female Pima Indians population aged 21 and over. Given the complex interplay of genetic, environmental, and lifestyle influences on diabetes, developing a predictive model based on this dataset could offer substantial benefit to the health industry. This model is not just focused on improving health outcomes through the early detection of potential diabetes cases but also provides valuable insights on the significant factors influencing the risk of diabetes.

Based on the issue above, we have identified the following research questions (RQs):

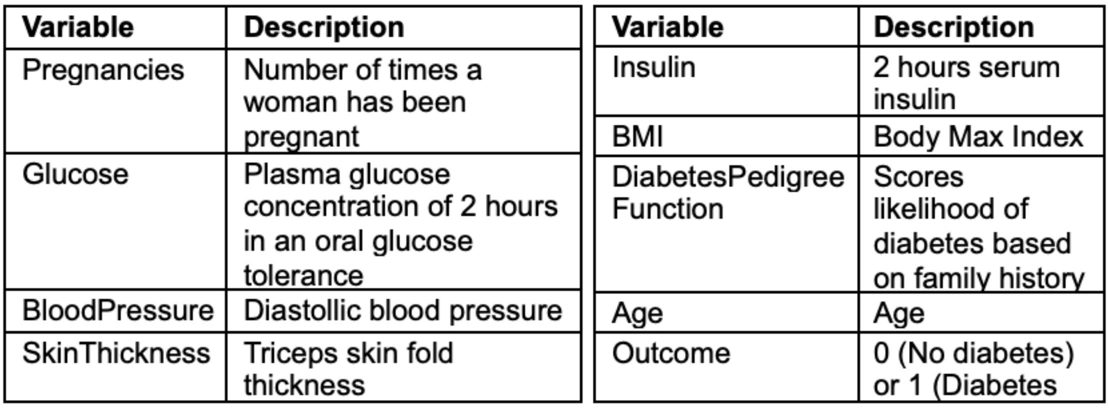
- RQ1: What is the best machine learning model for predicting the presence of diabetes (Diabetes)?
- RQ2: What are the most important diagnostic measurements that influence the prediction of diabetes (Diabetes)?

3.2 Methodology

Exploratory Data Analysis is initially conducted to prepare, clean, and explore the data, focusing on understanding the distribution of variables and their correlation between each other. To identify the best predictive model, we will employ various classification techniques, including Logistic Regression, Random Forests, Decision Tree (CART), and K-Nearest Neighbours (KNN).The performance of each model will then be assessed based on metrics such as accuracy, precision, sensitivity, F1 score, and Area Under the Curve (AUC) values.

3.3 Dataset and Variables

The diabetes dataset comprises diagnostic measurements from patients, including variables like BMI and age, and a target variable labelled ‘Outcome’, which will be renamed to ‘Diabetes’ for clearer interpretation. This dataset encompasses 768 observations and 9 variables, with all variables being numerical.



Variable	Description
Pregnancies	Number of times a woman has been pregnant
Glucose	Plasma glucose concentration of 2 hours in an oral glucose tolerance
BloodPressure	Diastollic blood pressure
SkinThickness	Triceps skin fold thickness
Insulin	2 hours serum insulin
BMI	Body Max Index
DiabetesPedigree Function	Scores likelihood of diabetes based on family history
Age	Age
Outcome	0 (No diabetes) or 1 (Diabetes

Table 3: Diabetes Dataset Variables Description

3.4 Analysis

Upon importing the dataset, we start with Exploratory Data Analysis, cleaning the data by identifying and addressing missing values and zero values in variables that are medically impossible to be zero like ‘Glucose’, ‘BloodPressure’, ‘SkinThickness’, ‘Insulin’, and ‘BMI’. Imputation for these variables is based on their distribution, using the mean for normally distributed variables and the median for skewed ones. The analysis of predictor variables points to a consistent pattern: individuals with diabetes generally exhibit higher median values in these diagnostic measurements compared to non-diabetic individuals, indicating a potential association with the likelihood of a diabetes diagnosis. This observation is supported by the scatter plot matrix comparing predictor variables with ‘Diabetes’, as shown in Figure 14. Here, we observe positive trends where higher diagnostic measurements’ values are associated with the presence of diabetes. Additionally, a heatmap analysis reveals moderate correlations between variables, with no values high enough to cause concern for multicollinearity.

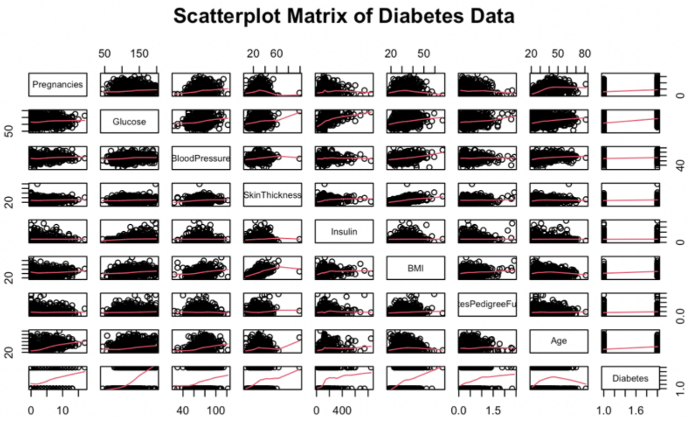


Figure 14: Diabetes Scatter Plot

3.4.1 Logistic Regression

We divide the dataset into train and test sets with a 70:30 split ratio and build a Logistic Regression model targeting the binary outcome of the ‘Diabetes’ variable. This model predicts the probability of each data point belonging to one of two classes, thus outputting the likelihood of a particular class given the input variables. To enhance the model’s accuracy, we utilize stepAIC method with a ‘both’ direction approach, enabling the addition and removal of variables through forward selection and backward elimination, based on the Akaike Information Criterion (AIC). This method ensures we retain only statistically significant variables, as evidenced by P-values < 0.1 and 95% confidence intervals for each variable’s odds ratio that exclude one. We evaluate the model’s performance using a confusion matrix, plotting the Receiver Operating Characteristic (ROC) curve, and calculating the Area Under the Curve (AUC). Notably, a threshold 0.5 is applied to transform the predicted probabilities into class labels. The AUC assesses the model’s ability to distinguish between classes, while accuracy measures the proportion of correct predictions (true positives and true negatives). This model achieves an accuracy of 0.7913043 and an AUC value of 0.8446667, indicating a strong predictive capability.

3.4.2 Random Forest

For Random Forest model, we use the identical train and test sets used in the Logistic Regression model. We choose to use 500 trees to enhance the model’s stability. After training the data, we make predictions on the test set, both in terms of class labels and class probabilities to assess the model’s performance. The accuracy of this model is 0.7869565, while the AUC value is 0.8397917. Furthermore, the variable importance analysis highlighted ‘Glucose’, ‘BMI’, and ‘Age’ as the most important predictors.

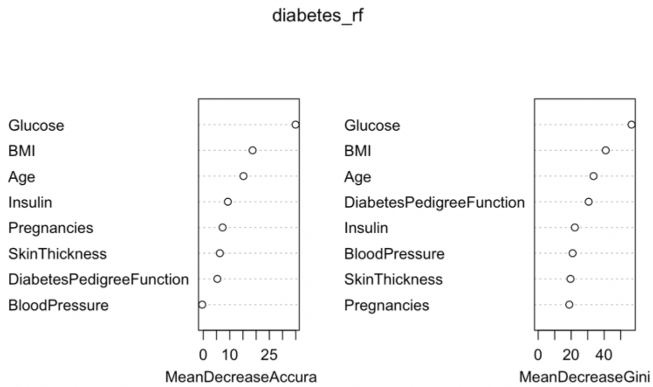


Figure 15: Random Forest Variable Importance

3.4.3 Decision Tree (CART)

Our Decision Tree model uses the CART algorithm, constructing a tree by splitting the data at decision nodes based on predictors to forecast outcomes. To refine the model, we select a complexity parameter (CP) that minimises cross-validation error, leading to the pruning of the model to its optimal structure. The optimal decision tree model outlines the explicit rules the model uses to make decisions and determine the presence of ‘Diabetes’. Evaluating this model on test set yielded an accuracy of 0.7695652 and an AUC of 0.8397917. Similar to our findings from Random Forest analysis, the CART model’s variable importance analysis also identified ‘Glucose’, ‘BMI’, and ‘Age’ as the top 3 most important predictors.

3.4.4 K-Nearest Neighbours (KNN)

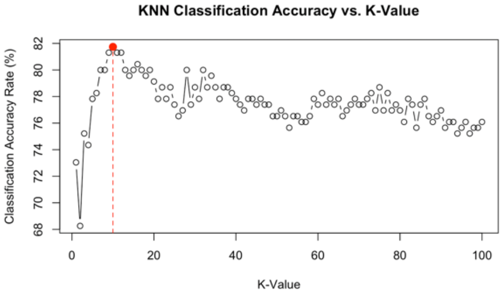


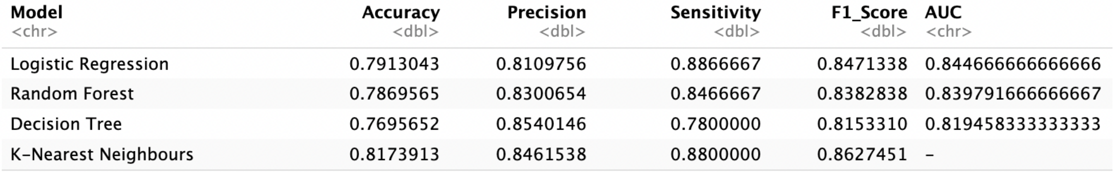
Figure 16: Optimal K-Value

KNN classifies new data points based on the majority class among its k nearest neighbours in the training data. Since it uses a distance metric, Euclidean in this case, to measure similarity between data points, scaling both training and testing datasets is crucial to ensure that all features contribute equally. We initiate our model with k = 1 and explore various k values up to 100, finding k = 10 as the optimal k value that gives the model the highest percentage of correct classifications (PCC), as shown in Figure 16. A higher PCC indicates greater model accuracy in predicting ‘Diabetes’. Utilizing k = 10, we achieve the best KNN model performance with an accuracy of 0.8173913. Notably, the AUC value is not computed due to the limitation of the ‘knn()’ function we used to build our model in R, which outputs only class labels without the probability estimates necessary for AUC calculation.

3.5 Results

RQ1: What is the best machine learning model for predicting the presence of diabetes (Diabetes)?

- The model performance is evaluated using accuracy, precision, sensitivity, F1 score, and AUC value. For ‘Diabetes’ prediction, we prioritise accuracy and AUC. Accuracy gives us a clear indication of the model’s overall correct classification rate, while AUC reflects the model’s ability to distinguish between patients with and without diabetes across various thresholds. The KNN model generates the highest accuracy at 0.8173913, as seen in Figure 17. Meanwhile, the Logistic Regression model gives the highest AUC of 0.8446667, with its ROC curve nearest to the top-left corner compared to others.



Model <chr>	Accuracy <dbl>	Precision <dbl>	Sensitivity <dbl>	F1_Score <dbl>	AUC <chr>
Logistic Regression	0.7913043	0.8109756	0.8866667	0.8471338	0.844666666666666
Random Forest	0.7869565	0.8300654	0.8466667	0.8382838	0.839791666666667
Decision Tree	0.7695652	0.8540146	0.7800000	0.8153310	0.819458333333333
K–Nearest Neighbours	0.8173913	0.8461538	0.8800000	0.8627451	–

Figure 17: Classification Model Comparison

RQ2: What are the most important diagnostic measurements that influence the prediction of diabetes (Diabetes)?

- According to the Random Forest and Decision Tree models, ‘Glucose’, ‘BMI’, and ‘Age’ are the most important diagnostic measurements that influence in predicting ‘Diabetes’.

4. References

Surles,W. (2017). Unsupervised Learning in R. [online] RPubs. Available at: <https://rpubs.com/williamsurles/310847> [Accessed 28 March 2024].

Sidoti, S.A. (2019). An Introduction with Examples in R Principal Component Analysis. [online] RPubs. Available at: < https://rpubs.com/carabidus/465971> [Accessed 28 March 2024].

Mahendru,K. (2019). How to determine the Optimal K for K-Means?. [online] Medium, Analytics Vidhya. Available at: < https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb#:~:text=The%20Elbow%20Method,-This%20is%20probably&text=Calculate%20the%20Within%2DCluster%2DSum,Errors%20sounds%20a%20bit%20complex.> [Accessed 28 March 2024].

Analytix Labs. (2023). What are Lasso and Ridge Techniques?. [online] Analytix Labs. Available at: < https://www.analytixlabs.co.in/blog/lasso-and-ridge-regression/#:~:text=Lasso%20and%20Ridge%20Regression%20are%20two%20popular%20regularization%20techniques%20used,while%20Ridge%20uses%20L2%20regularization> [Accessed 28 March 2024].

Geeks For Geeks. (2024). Random Forest Algorithm in Machine Learning. [online] Geeks For Geeks. Available at: < https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/> [Accessed 28 March 2024].

Narkhede,S. (2018). Understanding AUC - ROC Curve. [online] Towards Data Science. Available at: < https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5> [Accessed 28 March 2024].

5. Dataset Links

- The link for Customer Dataset (Unsupervised Learning): https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

- The link for Student Performance Dataset (Regression): https://www.kaggle.com/datasets/larsen0966/student-performance-data-set?select=student-por.csv
or can be found:
https://archive.ics.uci.edu/dataset/320/student+performance (only the Portuguese language dataset)

- The link for Diabetes Dataset (Classification): https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database