# STAT462 – Data Mining

## Assignment 10 – Report 2

| Attribute | Details |
|---|---|
| **Individual/Group** | Group |
| **Course Learning Outcomes (Relevant to this assessment)** | The course Learning Outcomes demonstrated by successful completion of the task below include:<br>• Identify suitable algorithms to address specific research questions in classification.<br>• Implement, adapt and apply algorithms in R to solve research questions, and accurately interpret their results.<br>• Critically compare and reflect on the limitations and performance of the algorithms under consideration.<br>• Compose comprehensive scientific and technical reports using R Markdown to communicate your analysis and findings effectively. |
| **Conditions** | • Your report must be a single .pdf document created from R Markdown.<br>• Once the due date and time has passed no resubmissions are allowed, without permission from the facilitator. |
| **Submission** | See LEARN for specific requirements and deadline. |
| **Length** | No specific length is expected but there is a strict maximum of 20 pages in pdf form. The report is allowed to be much shorter than this without penalty. |
| **Weighting** | 15% |
| **Maximum grade** | 100 (See detailed guidance at the bottom of this document for how your report will be assessed) |

## Assessment task

The objective of this assessment is to investigate a dataset and predict a categorical response (classification). You will load datasets, do some cleanup and feature engineering, train and run classification algorithms, and interpret your findings in a report. You will be graded on all parts of this procedure.

## Practical application

You'll practice "starting from scratch" in this assessment: Given just a dataset, you need to set up everything you need, and in the end present a well-written and illustrative report. This will teach you an important workflow in your future career as a data scientist.

## Detailed instructions

### A. Automatically detecting seed types (Logistic Regression)

In this question you are going to training a logistic regression for automatic classification of two types of pumpkin seeds, Çerçevelik and Ürgüp Sivrisi, based on geometric (measured) features of these seeds.

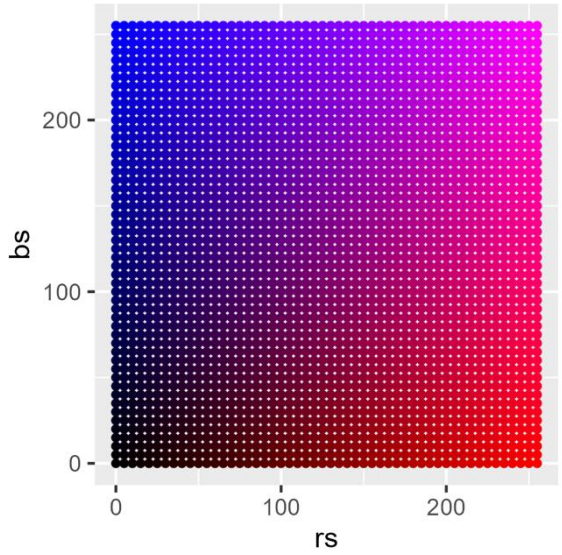
1. Load the datasets `seeds_training.csv` and `seeds_test.csv`. Since the class labels (the seed types) contain characters that R struggles with, encode the class features as a factor variable in both training and test set.
2. Fit a multiple logistic regression classifier using all of the features and record its accuracy on the test set.
3. Provide a confusion matrix for your model based on its predictions on the test set. Why does the concept of sensitivity and specificity not make a lot of sense here?
4. Consulting the summary of your `glm` model, which three features are most significant?
5. Train another `glm` model, but this time using only the three most significant features from your first model run. Compute test set accuracy and a confusion matrix in the same way as for your first model and briefly compare the two models. Which model would you pick, and why?

### B. Predicting color name from RGB values (Discriminant Analysis)

Colors can be coded via their RGB (red, green, blue) value in the form `(r,g,b)`, where `r`, `g`, and `b` are integers between 0 and 255. For example, `(255,0,0)` is pure red, and `(128,200,128)` is a shade of green.

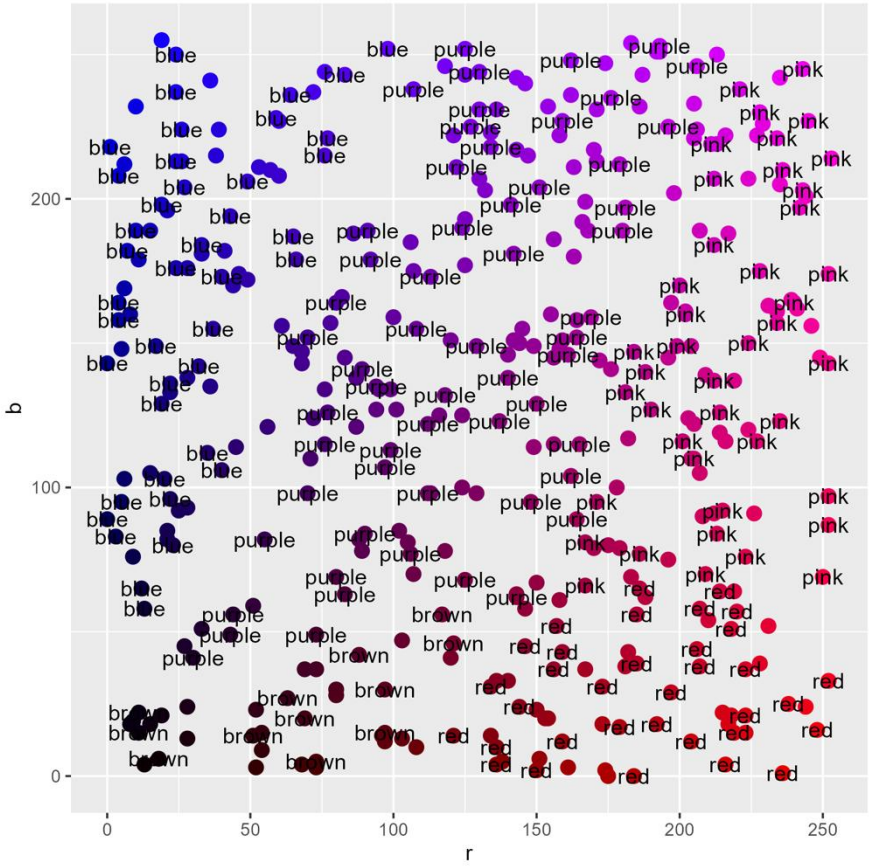
In this exercise we will map `(r,g,b)` values to their color names. For example, we want `(255,0,0)` to be classified as red.

To make things a bit easier, we focus on the part of colour space where `g=0`, i.e. there is no green component. This means the feature space is all combinations `(r,0,b)`, where `r` and `b` are between 0 and 255. For orientation, here is a visualisation of some of these colors, with each circle having the




color of its `r-b-coordinate`. We can clearly see that the coordinate `(255,0,0)` corresponds to bright red. Our goal is to create a classification algorithm that takes an `r` value, and a `b` value, and outputs the name of the color this corresponds to.

Thankfully, we have a dataset where some of these labels have been entered. This is visualised below




1. Load the dataset `colors_train.csv`. How many classes are there in the dataset?
2. Fit a QDA algorithm to this classification problem and compute the confusion matrix.
3. Visualise the decision boundaries in a suitable way.
4. Test your algorithm on `(200,0,200)`. What colour is this being called by your algorithm?

## Recommended structure

All answers need to be referenced to the original question, and we recommend you go through the questions in the order in which they are presented above.

## Research and referencing

Additional research and referencing is not required to complete this assessment, but any ideas that are not your own need to be cited using APA 7th edition.

## Presentation guidelines

- Your report must be a single .pdf document created from R Markdown.
- Your report needs to contain and display all R code that needs to be run to get your results.
- All code needs to be sufficiently commented
- Full report (including all figures and code) must not exceed 20 pages in .pdf, but it can be much shorter (the instructor's model solution is about 12 pages long, and there will be no penalty for shorter reports with the same content)

## How your report will be assessed

This rubric provides a general guide to how your work will be evaluated. Not all criteria apply to every question—some may be more relevant than others. Focus on producing clear, well-documented, and accurate work.

## Grading Overview

**Grade range 50%-65%**

- You have attempted the assignment, but in a limited way that might not fully demonstrate understanding of the concepts involved.

**Grade range 65%-79%**

- You have completed most of what was required and did nearly everything correctly.

**Grade range 80%-100%**

- You have done all that was required and everything is correct.

## General Guidelines

- **Code quality and documentation**: Your code should be readable, well-commented, and functional. Errors will result in reduced marks.
- **Statistical accuracy**: Ensure calculations (e.g., regression coefficients, $R^2$, confidence intervals) are correct and appropriately rounded (in the context of this course, accuracy of 3 decimal places is sufficient unless stated otherwise).
- **Interpretation of results**: Provide logical conclusions supported by data.
- **Data visualisation**: Figures should be clear, well labelled, and appropriately formatted.
- **Report quality and presentation**: The report should be well-structured and professional, with proper formatting, grammar, and clear explanations, and not longer than necessary.
- **Model comparison (if applicable)**: If required, discuss performance and insights. If a specific model is assigned, implement and explain it clearly.

## Detailed grading expectations

| Skills | Expectations for 50-65% | (Additional) Expectations for 65-79% | (Additional) Expectations for 80-100% | Approximate Weighting |
|---|---|---|---|---|
| **Communicating clearly via text** | Text is in understandable English. For most questions, a few short sentences will be adequate unless stated otherwise. | Uses adequate technical jargon and concise language. "Generic garbage" as sometimes produced by GenAI will lead to detractions here. | Writing is clear, concise, and well-structured. Uses precise terminology and presents complex ideas in an accessible way, demonstrating strong communication skills. | Necessary prerequisite. Work cannot be marked if not communicated clearly. |
| **R Markdown and format of submission** | Submission is a single file in either .html or .pdf format, generated by R Markdown. | – | – | Necessary prerequisite. |
| **R programming skills** | Some attempt to implement or use a statistical algorithm is made. | Code does what it is intended to do. Code is readable, uses comments, and applies suitable naming conventions. Syntax is correct, and code runs without errors. | Code is reasonably efficient and modular, demonstrating best practices in programming. dplyr, ggplot, and statistical learning modules are used appropriately. | **30%** |
| **Knowing and applying statistical learning algorithms** | Basic statistical learning task is addressed. | Algorithm results are (when required) put into context, correctly explained, and relevant implications identified. | Demonstrates deep understanding of model limitations and improvements. | **35%** |
| **Interpreting and evaluating algorithm results and output** | Some interpretation of the findings is provided, even if not correct. | Correct choice (within constraints of admissible modules) and application of algorithm. | Provides critical analysis, discussing model strengths and weaknesses. Compares results with alternative approaches, considers broader implications, and supports claims with strong reasoning. | **25%** |
| **Communicating clearly via figures** | An attempt at visualisation (if required) is made. | Axes are labelled, figure is suitably scaled, and the intention is clearly communicated. Interpretation of the figure is given in either an expressive caption or in the accompanying text. | Choice of visualisation type and colour scheme are highly effective in conveying information. | **10%** |

## Final Notes

- Not all criteria apply to all questions—focus on what's relevant to each task.
- Marks will be assigned based on correctness, clarity, and depth relative to the question's requirements.
- Follow the instructions carefully and justify key decisions where relevant.

This rubric is a guide to help you understand expectations. If you're unsure about something, refer to the assignment instructions or ask for clarification.