End of Module Test (Module 4)

Stats & ML with R

Tutor: George K Agyen

Day 20: Break Day Mini Project

Take a Break!

Congratulations on completing 20 days of learning R! Take some time to relax and review what you've learned.

Instructions For Submission

- Answer all the questions provided in an R script
- Save the finished work with your last name and module number separated by an underscore (e.g. if your name is **Benjamin Owusu**, you save the script file as `Owusu_Module4.R`)
- Submit the saved script to me via email at `gkagyen@ghana-rusers.org` with the subject **Module 5 Test**
- You have one day to complete this test and submit by end of day on `Thursday May 1, 2024`

Stats & ML Mini Project

This mini project is designed to test your understanding through a realistic healthcare analytics scenario using the **Pima Indians diabetes** dataset `pima` from the `pdp` package.

Overview

Objective: Analyse diabetes risk factors using statistical methods and build a predictive model

Part 1 Statistical Analysis

1. Clean the dataset and perform some EDA on the dataset
2. Perform Some hypothesis tests
    - perform a t-test to compare mean glucose- levels between diabetic/non-diabetic groups
    - conduct a chi-square test between diabetes and hypertension (pressure > 90)
    - run a one-way ANOVA comparing BMI across age groups (create age categories)
3. Create 3 plots using ggplot2 showing

- Boxplot of insulin by diabetes status
- Bar plot of diabetes prevalence by `age_group`
- scatterplot of `glucose` v `mass`

Part 2 - Machine Learning

1. Perform a Linear regression to predict bmi (`mass`) using:
    - `glucose` and `insulin` only
    - `glucose`, `pressure` and `age`
    - make predictions with the two models and compare their evaluation metrics (R-squared, MAE and RMSE)
2. Build a diabetes classifier (logistic regression model) using:
    - `glucose` and `insuling` only
    - predictors from the regression models that were significant
    - make predictions on the test data and evaluate performance using a confusion matrix

Deliverables

1. R script with all codes
2. A pdf report containing:
    - key statistical findings
    - model performance summaries
    - visualisations and interpretations