CM52058 – Statistical Data Science
Final Coursework: Diabetes Dataset Analysis

Professor Nello Cristianini

Set: Wednesday 26/11/2025 (week 9) **Due:** Friday 12/12/2025 (week 11), 8 pm
Percentage of overall unit mark: 100%
Submission Location: Moodle **Submission Components**: **Submission Format:** .pdf file named '*your email*'.pdf (to be marked) and a single python notebook (not to be marked)
Anonymous Marking: yes
Generative AI Assessment Categorisation: Type A/B/C [guidance] ⇒ **B**

This coursework forms 100% of your final mark.
The maximum marks available for each question is shown.

Instructions. Analyse the provided dataset, and compile a report with the required results. Report your results in a pdf file with readable plots, essential metrics, and a short description of each method, plot or table, and brief conclusions drawn. Use at most 10 pages in font 11, keep the figures small (not full page, please).
Use: Python (any setup, e.g., Colab/Jupyter) with pandas, numpy, matplotlib, and scikit-learn.

Deliverable (Two files, only the pdf file is marked). A 10-page (maximum) PDF report with results and plots. Include concise captions so figures and tables are interpretable on their own. Also upload the notebook (a single notebook with the code), which will not be marked.

Upload it on Moodle **by 8pm, Friday 12th December, 2025.**

Dataset Provided. We will analyse the dataset "Diabetes" (Efron, Hastie, Johnstone & Tibshirani, Annals of Statistics, 2004), that can be found here:
https://hastie.su.domains/CASI_files/DATA/diabetes.html
In this file, you can assume this convention for patients (0=female and 1=male).

INDIVIDUAL TASKS

1) **Load the Diabetes data** from the provided CSV file, and build a DataFrame — Then report the number of rows, columns, names of columns. State which method you followed. (Method 1 can result in at most 5 marks, method 2 in at most 2 marks, and there are no extra marks for doing both).

- Preferred Option: Load from the provided CSV into a pandas DataFrame with readable column names. [5 marks]

- Alternative option: load dataset from from sklearn.datasets.load_diabetes()
  Note that in this version of the file every feature is rescaled, and the features / target variables may have different names in the two versions of the data. [2 marks]

2) **Basic dataset information**: Report number and name of columns, data types contained in them, first few rows, any missing values, and identify the target (y) and its data type. Try to find a simple, compact, readable way to present this information.

[10 marks]

3) **Descriptive statistics** (properties of the sample) — For each predictor (feature) , compute and report: mean, standard deviation, range (min, max). Do the same for the target. You may also propose other descriptive statistics that might be useful, and why they would add information.

[5 marks]

4) **Statistical test.** Plot separate histograms of “disease progression” for male and female patients. (Assume 0=female and 1=male) . Optional: you may also explain what is a reasonable number of bins in the histogram, for this specific dataset.

[10 marks]

Perform a statistical test (using python libraries) to test if disease progression is significantly different in those two classes of patients. Report p-value, test description (test statistics, alternative and null hypothesis, significance threshold), and draw a conclusion that is supported by the data.
(note: for this exercise, two-sample t-test can be used) [10 marks]

5) **Pearson correlation heatmap** (only for predictors / features) — Compute the 10×10 Pearson correlation matrix among predictors and plot a heatmap using matplotlib, or other reasonable library. Briefly explain any choices you made.

[5 marks]

6) **Correlation with target (bar chart)** — Compute Pearson correlation between each predictor and (y). Plot a bar chart. Identify the top 2–3 predictors, and briefly summarise in one sentence what this means.

[5 marks]

7) Classification exercise (swap target and one feature) — We want to simulate a 2-class classification problem, so for this exercise we turn the feature “sex” into the “target” and the target “disease progression” into a feature (the terms ‘target’ and ‘label’ are equivalent in this

context).

a) Preparation: Create a new dataframe from the original one, by swapping two columns. In the new dataframe, the feature “sex” should be the target and the former target “disease progression” should be a feature (keep other features as they are). Summarise output of “df.info()” for this new dataframe (just column names and first rows, no need for missing values). (assume 0=female and 1=male in the data)
Make sure that the new target is an integer.

[5 marks]

b) Model: Train Linear Discriminant Analysis(LDA) on the full set. Plot a bar chart of the learned coefficients. (Note: this is a descriptive task only, not a learning task; use python library)

[5 marks]

c) Quality of separation. Report the confusion matrix and accuracy (Note: again, this is computed on the same sample used for training, descriptive only). [5 marks]

d) Hold-out evaluation. Perform a train/test split (e.g., ~80/20). Train fisher LDA on train set; evaluate on test. Report the confusion matrix and accuracy computed on the test set. State the proportion train/test you used. (Note: here we train on a subset of the data, and we test on the remaining data, so we are generalising). Explain: under which conditions you can expect this performance to generalise to new data?

[5 marks]

8) Correlation analysis.

Produce a scatter plot of BMI vs disease progression. [5 marks]
Is there a linear relation between BMI and disease Progression? [5 marks]
(Pearson correlation coefficient)
Is this statistically significant? (report p-value and test specs)
Briefly explain the assumptions behind this statistical test. [5 marks]

9) **Short written report** —

This report should be at most 10 pages long, pdf file, in font 11.
Address each question in turn, include data descriptions, results or tables, plots with caption (the caption explains what the plot shows). [5 marks]
In a few sentences, report conclusions based on the following question: “*Is disease progression predictable from these indicators?*”.
List key findings supporting your conclusions, how the data informs your conclusions. You may refer back to plots and tables computed in the previous questions.

[10 marks]

Include a brief statement about use of GenAI, see instructions above.

FINAL CHECKLIST

- Single PDF file, at most 10 pages long, with labeled, readable plots and concise captions, font 11.
- Notebook with code, not marked, but to be uploaded.
- Methods stated once (e.g., your chosen train/test split, name the algorithm used, eg its corresponding python command) and used consistently
- Figures and numbers sufficient to understand conclusions without reading code
- Statement about use of GenAI

Marking and Feedback

All questions are marked first on correctness and clarity, and then for demonstrating understanding that goes above and beyond standard bookwork. For each question, you may expect to pass by providing a correct answer. Demonstrating clarity and understanding in the answer will bring you in the range of a high second or low first. Demonstrating a deeper understanding, potentially above and beyond bookwork, can lead to a higher first.

You will receive **summative feedback** on your work within 3 semester weeks of the submission deadline. The feedback will discuss your performance based on the criteria for marking.

Academic Integrity

Your work will be checked to ensure that you have not plagiarised. For more information about the plagiarism policy at the University, see: https://library.bath.ac.uk/referencing/plagiarism

Remember that the published work that you refer to in your report should be clearly referenced in your text and listed in a bibliography section given at the end of your report. For more information, see https://library.bath.ac.uk/referencing/new-to-referencing

Any code submitted can be checked.

> **This is a coursework of Type B:** you may use generative AI to help you find useful python commands or libraries, or check IDE error messages, but you need to understand every command you use, and cannot have the GenAI solve your assignment: just to provide you with the information you need about specific commands, functions, libraries, error messages.
>
> If you have made use of GenAI to debug some code or interpret error messages from the IDE or look up some syntax, please declare so at the end of your report.
>
> It is not allowed to use GenAI for the writing of this report.

For more information, see:

https://teachinghub.bath.ac.uk/guide/genai-assessment-categorisation/.