SBEG209 - Biostatistics
Final Project.
Estimated Working Hours: 10-14 hours (including the report and presentation).
Deadline: Wednesday – December 31st, 2025, at 02:00pm.
Presentation (online): Wednesday – December 31st, 2025, at 04:00 pm.

- You can submit before the deadline. Please arrange with me for that.
- No late submissions.
- Only one member from each team should submit the project files to Google Classroom.

Teams: Each team should have 3-4 members.

Project Description:

You need to pick up only ***ONE*** idea from the below to work on.

Idea One.

You are required to conduct a binary classification experiment using a Naive Bayes (NB) classifier.

- Select a tabular dataset from the Kaggle platform for binary classification problems.
- For the variables/columns of your dataset, mention which ones are quantitative and which ones are categorical.
- Use any statistical method to remove outliers from the data, if existing.
- Calculate a set of descriptive statistics to quantitatively describe the data:
  - Measures of central tendency,
  - Measures of dispersion.
- Standardize the features using your calculations for the descriptive statistics.
- Split the data randomly into 2 partitions with a 80%-20% proportion:
  - The 80% partition is called the training data,
  - The 20% partition is called the testing data.
- For each feature/column in the training data:
  - Plot the histogram/distribution.
  - Comment on the type of each distribution (Gaussian, exponential, uniform, etc).
  - Statistically test if a feature/column is normally distributed. You need to search for a statistical test and explain its null and alternative hypotheses.
  - Plot the conditional distributions of each feature on each target class (label).
- Apply the Naïve Bayes (NB) calssifier:
  - Implement the NB classifier from scratch.
  - Train the NB classifier on your training data.
  - Use the trained NB model to predict the classification of the test data.
  - Calculate the model accuracy.
  - Compare your results to the case of using the NB classifier from standard Python packages.

Idea Two.

You are required to conduct an association experiment using linear regression analysis.

- Select a tabular dataset from the Kaggle platform for linear regression problems.
  - Your data must have more than three predictors (mutivariable linear regression).
  - Make sure that the number of data points per feature/predictor (n) is more than the number of predictors (p) by at least 5; $n \geq p + 5$.

- For the variables/columns of your dataset, mention which ones are quantitative and which ones are categorical.
- Use any statistical method to remove outliers from the data, if existing.
- Calculate a set of descriptive statistics to quantitatively describe the data:
  - Measures of central tendency,
  - Measures of dispersion.
- Standardize the features using your calculations for the descriptive statistics.
- For each feature/column in the training data:
  - Plot the histogram/distribution.
  - Comment on the type of each distribution (Gaussian, exponential, uniform, etc).
  - Statistically test if a feature/column is normally distributed. You need to search for a statistical test and explain its null and alternative hypotheses.
  - Compute a correlation coefficient between the response variable and each predictor.
- Apply the linear regression analysis on the response against each predictor individually.
  - Implement from scratch the method of obtaining the regression coefficient (RC).
  - Compare your results to the case of getting the RC from standard Python packages.
- Apply a multivarialbe linear regression analysis on the response against all the predictors simultaneously.
  - You do NOT need to implement this approach from scratch. You can use any standard Python packge.
- Expalin how can you statistically assess the multivarialbe regression quality in terms of:
  - The individual regression coefficients,
  - The regression model as a whole.

General notes for all the ideas.

- Support your findings/results/conclusions with figures.
- You have to deliver the following:
  - All the code scripts you used for your analysis,
    - Comments are a must.
  - Project report:
    - It should look like a research paper. It should have the following sections:
      - Introduction,
      - Methods: describe all the steps carefully and include all the used software packages,
      - Results and Discussion: report your results in details and discuss them,
      - Conclusion: list the overall findings of your analysis.
      - **Members Contribution: list in details what each member in your team did in this project. Each member in the team may receive a different grade based on the contribution weight.**
  - Presentation:
    - Submit the presentation slides to Google Classroom. These slides are part of the evaluation process.
    - You will be given a few minutes to represent your work online.
    - Prepare yourself for discussing your analysis and findings.

Good luck!