Task Data Scientist

General Instructions

1. Create a git code repository on either github/gitlab, and share the instructions to access you solution
2. Please do not plagiarize/copy-paste any publicly available solution. You are allowed to use code-assist tools and research if required.
3. As a solution, create a Solutions document that explains your approach and steps

Assignment

1. Preliminary questions.
    a. In X city, there is an increasing number of young adults refusing to go to enter the workforce. To solve the problem, the government hosts a program that gives monetary aid for young adults who found work. A control group consisting of young adults who found work without the program is also given. Throughout 10 months, these are the result of unemployed young adults who found work:

Using aids	Not Using aids
31	18
13	20
23	39
46	16
20	18
14	38
17	13
47	20
13	39
45	7

Compute the mean and std of the data.

a. What is the null hypothesis of the experiment? What would be the alternate hypothesis?
b. Compute the t-test for this experiment. Interpret your answer elaborately.

c. Is there a significant difference between the two groups?

2. Using the Structural Protein Sequences (kaggle.com) dataset, create a protein classifier. Write the solution into one Jupyter Notebook file, you may use SQL and Python to do this. The solution must include answers to the following questions:
    1. Exploratory Data Analysis
        a. What does each column represent? Explain your answer as detailed as possible
        b. Create a visualization of each column. State the reason for your choice of visualization method.
        c. What can you infer about the dataset?
    2. Data Preprocessing
        a. What are the types of data preprocessing that you know, and how would they be useful in the given dataset?
        b. Show that the preprocessing method of your choice works on the dataset.
        c. Given a set of columns, we could try to make a prediction on a certain aspect of the dataset. However, there are times when the same column would provide a radically different result from the previous training iteration. What is that phenomenon called, and how can we analyze and mitigate that?
    3. Feature Engineering
        a. What are features? How do you make one?
        b. How many features can you think of to classify the proteins?
        c. Based on the previous point, create features based on what you thought of, and explain the logic behind it.
    4. Modelling
        a. What kind of model are you planning to use, and why?
        b. Create a model to train the classifier. Can you extract the top important features?
    5. Evaluation
        a. What are the metrics typically used in the classifier you have made? And what is the metric of your choice here?
        b. Provide a visualization of the metrics of your choice.
        c. Based on what you have done, what do you think can be improved from your model?