Home test – Taliaz LTD

Question 1 (Maximum half a page)

Given a set of 24 genes that are suspected to be co-regulated by a transcription factor (TF):
A. Suggest ways to reveal which TF regulates them.
B. Find additional genes that may be regulated by this TF.

Question 2 (Maximum 6 lines)

Given 150 genomic sequences with overlaps, suggest methods or tools to:
A. Find the overlapping regions (intersection)
B. Find the union of these regions (assembly)

Question 3 (Maximum half a page)

When handling data sets with a large number of features (more than 1 million features) and a relatively small number of samples (less than 1000):
A. What problems do you expect to encounter?
B. How would you try to handle them (which methods/algorithms)? Please explain.

Question 4 (Maximum one page)

You are investigating a specific disease with a genetic background. Your aim is to develop a genetic-based predictive model. For this mission you have recruited 2000 individuals: 1050 with the disease and 950 healthy individuals. Each individual's DNA is analyzed using a genetic chip that tests 500,000 SNPs. You wish to analyze your results using machine learning approaches.
A. What is the best method to code the genetic data into machine learning features? Please explain and write a pseudocode for this coding.
B. You have found a predictive model based on five SNPs only. Suggest ways to validate its reliability.
C. Suggest a way to understand the biological mechanisms that may be explain your findings.

Question 5 (Maximum 4 lines)

You have created two decision trees: the first is relatively big, with an accuracy of 92%, the second is much smaller and has an accuracy of 90%. Which of the trees would you choose? Please explain your reasons for choosing it?