# Question 1 – Case Study

A school is interested in understanding the factors that contribute to student academic performance. It has collected data on a group of students, including their study habits, time spent studying, participation in extracurricular activities, and scores in Mathematics, Science, and English. The goal is to analyze this data to identify key predictors of overall academic success and potentially develop strategies to improve student outcomes.

The dataset “[Student Performance Dataset.csv](https://learning.online.hw.ac.uk/services/api/courses/h11bd/files/Student%20Performance%20Dataset.csv)” includes the following variables for each student:

* student\_id: Unique identifier for each student.
* gender: Gender of the student (Male/Female).
* study\_time: Time spent studying per week (in hours).
* extracurricular: Participation in extracurricular activities (Yes/No).
* math\_score: Score in Mathematics (out of 100).
* science\_score: Score in Science (out of 100).
* english\_score: Score in English (out of 100).
* overall\_score: Average score across all subjects (out of 100).

Using R or python, please create a model to predict the overall score. Demonstrate and explain each step you take to develop this model (screenshot should be presented). Based on the model you have developed, suggest interventions that can improve the overall score.

**CASE STUDY 2:**

An insurance company is interested in improving its customer acquisition process. It has collected data on potential customers, including demographic information, prior insurance history, interaction details, and financial indicators. The company wants to leverage this data to build machine learning models that can predict the likelihood of a customer converting (i.e., purchasing a policy). Accurately predicting conversion rates will allow the company to optimize its marketing efforts, personalize customer interactions, and ultimately increase sales.

The “[insurance\_data.csv](https://learning.online.hw.ac.uk/services/api/courses/h11bd/files/insurance_data.csv)” dataset contains the following variables for each potential customer:

* Age: Age of the customer.
* Is\_Senior: Whether the customer is a senior citizen (Yes = 1 /No = 0).
* Marital\_Status: Marital status of the customer.
* Married\_Premium\_Discount: Discount applied due to marital status.
* Prior\_Insurance: Whether the customer had prior insurance (Yes = 1 /No = 0).
* Prior\_Insurance\_Premium\_Adjustment: Premium adjustment based on prior insurance.
* Claims\_Frequency: Number of claims filed by the customer (if applicable).
* Claims\_Severity: Severity of claims filed by the customer (if applicable).
* Claims\_Adjustment: Adjustment to premium based on claims history.
* Policy\_Type: Type of insurance policy.
* Policy\_Adjustment: Adjustment to premium based on policy type.
* Premium\_Amount: Premium amount.
* Safe\_Driver\_Discount: Discount applied for safe driving.
* Multi\_Policy\_Discount: Discount applied for multiple policies.
* Bundling\_Discount: Discount applied for bundling policies.
* Total\_Discounts: Total discounts applied.
* Source\_of\_Lead: Source of the customer lead (e.g., online, referral).
* Time\_Since\_First\_Contact: Time since the first contact with the customer.
* Conversion\_Status: Whether the customer converted (Yes = 1 /No = 0)
* Website\_Visits: Number of website visits.
* Inquiries: Number of inquiries made by the customer.
* Quotes\_Requested: Number of quotes requested by the customer.
* Time\_to\_Conversion: Time taken for conversion (if applicable).
* Credit\_Score: Credit score of the customer.
* Premium\_Adjustment\_Credit: Premium adjustment based on credit score.
* Region: Region of the customer.
* Premium\_Adjustment\_Region: Premium adjustment based on region.

Using R or Python please build machine learning models for predicting customer conversion (Conversion\_Status). Demonstrate (using screenshot) and justify each step in the model development. From your analysis, please suggest the best model to deploy in a real-world scenario. Interpret the model's coefficients or feature importances to gain insights into the factors that drive customer conversion.

Question 3:

You have been provided with a sample of average bank balance own by people living in the United States. Describe the pattern in the dataset and please propose a model to predict the average bank balance of an individual.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Independent variable** | | | | **Dependent variable** |
| **City Location** | **Median Years Education** | **Median Income** | **Median Home Value** | **Average Bank Balance** |
| Y | 14.80 | $ 91,033 | $ 183,104 | $ 38,517 |
| Y | 13.80 | $ 86,748 | $ 163,843 | $ 40,618 |
| N | 13.80 | $ 72,245 | $ 142,732 | $ 35,206 |
| N | 13.20 | $ 70,639 | $ 145,024 | $ 33,434 |
| N | 13.20 | $ 64,879 | $ 135,951 | $ 28,162 |
| N | 13.70 | $ 75,591 | $ 155,334 | $ 36,708 |
| N | 14.40 | $ 80,615 | $ 181,265 | $ 38,766 |
| N | 13.90 | $ 76,507 | $ 149,880 | $ 34,811 |
| N | 16.10 | $ 107,935 | $ 276,139 | $ 41,032 |
| Y | 15.10 | $ 82,557 | $ 182,088 | $ 41,742 |
| Y | 14.20 | $ 58,294 | $ 123,500 | $ 29,950 |
| Y | 15.80 | $ 88,041 | $ 194,369 | $ 51,107 |
| Y | 12.90 | $ 64,597 | $ 119,305 | $ 34,936 |
| Y | 13.10 | $ 64,894 | $ 141,011 | $ 32,387 |
| Y | 16.10 | $ 61,091 | $ 194,928 | $ 32,150 |
| Y | 13.60 | $ 76,771 | $ 159,531 | $ 37,996 |
| Y | 13.50 | $ 55,609 | $ 123,085 | $ 24,672 |
| N | 12.80 | $ 74,091 | $ 143,750 | $ 37,603 |
| N | 12.90 | $ 53,713 | $ 112,649 | $ 26,785 |
| N | 12.70 | $ 60,262 | $ 126,928 | $ 32,576 |
| N | 16.10 | $ 111,548 | $ 230,893 | $ 56,569 |
| Y | 12.80 | $ 48,600 | $ 105,737 | $ 26,144 |
| Y | 12.70 | $ 51,419 | $ 104,149 | $ 24,558 |
| Y | 12.80 | $ 51,182 | $ 106,898 | $ 23,584 |
| Y | 12.80 | $ 60,753 | $ 95,869 | $ 26,773 |
| Y | 13.80 | $ 64,601 | $ 103,737 | $ 27,877 |
| Y | 13.20 | $ 62,164 | $ 114,257 | $ 28,507 |
| N | 12.70 | $ 46,607 | $ 94,576 | $ 27,096 |
| N | 12.70 | $ 61,446 | $ 122,619 | $ 28,018 |
| N | 12.80 | $ 62,024 | $ 134,430 | $ 31,283 |

# Question 4:

You are an HR analyst working for a company, and you have been tasked with analysing the factors that influence employee salaries. The dataset you have collected contains information about the number of years of experience, level of education, and department of each employee, along with their corresponding salaries. Please select the most appropriate predictive model for this task and develop it using R or Python (demonstrate your work using screenshots). Interpret your model and suggest policies implications from your analysis.

*Table 3*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Employee** | **Years of Experience** | **Education Level** | **Department** | **Salary (in $)** |
| 1 | 3 | Bachelor's | Sales | 50000 |
| 2 | 5 | Master's | Marketing | 60000 |
| 3 | 10 | PhD | HR | 80000 |
| 4 | 2 | High School | IT | 45000 |
| 5 | 7 | Bachelor's | Sales | 70000 |
| 6 | 4 | Master's | Marketing | 55000 |
| 7 | 8 | Bachelor's | HR | 75000 |
| 8 | 6 | PhD | IT | 65000 |
| 9 | 1 | High School | Marketing | 40000 |
| 10 | 9 | Master's | Sales | 85000 |