5.4: Intro to Data Mining

One of the data analysts for the sales team recently left Pig E. Bank, so you've agreed to take their place and help out with a customer retention project.

To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. You've created a table of client attributes that you believe could indicate whether customers will leave—for example, age, estimated salary, etc. You're going to use this information to identify the top risk factors that contribute to client loss and model them in a decision tree.

Create a new text document and call it "Answers 5.4"; then, based on the data mining techniques covered in this Exercise and the course, complete the steps below:

1. Download Pig E. Bank's client data set (.xlsx). Open the data set in Excel and take a moment to familiarize yourself with the data.
2. To understand the data, you'll first need to assess the quality of the data, by checking for missing values, errors, and inconsistencies.
    - You'll also need to clean your data, using the techniques that you learned in previous Achievements. Fix any inconsistencies in the table and/or any errors, as far as it is possible.
    - Document your processes for assessing the data quality and cleaning the data, and note down any missing values or errors.
3. Now that you've cleaned the data, you're ready to calculate some basic descriptive statistics to understand the data. Remember, your goal is to identify the risk factors that have contributed to customers leaving the bank.
    - Separate the clients into 2 groups: one for those who have left the bank and a second for those who have stayed (hint: "1" in the "ExitedFromBank" column represents customers who have left).
    - Use pivot tables and other Excel functions to identify the top 3 to 4 factors that lead to clients leaving.
    - Gather and analyze statistical information on both groups (e.g., find averages, means).

- Determine the leading factors that contribute to client loss, based on your analysis of the data provided.
- Document your results and how you reached them.

4. Using the information you've uncovered so far, create a decision tree to determine the probability of customers leaving the bank.
    - Pick which tool you'll use to create your decision tree. You can either create your own template using Excel or Powerpoint, for example, or download a decision-tree template.
    - Determine which decision node will have the greatest impact and place it at top of the tree. For example, if you decide that an estimated salary below 15,000 USD is the biggest risk factor, then you would put this at the top and build your tree from there. Make sure that your decision tree includes the top 3 to 4 risk factors you identified in step 3.
5. Combine your decision tree and answers document into one PDF and upload it here for your tutor to review.