# The Machine Learning Track

### **Problem 1: Strategizing and Predicting a Classification Task**

### Saturday, 27th April 2024

### Overview

Machine learning needs no introduction; it is the science of generating predictions that can survive after deployment, i.e., generalize well-enough for extended periods of time. In this challenge, the idea is to put your ML strategy skills to a comprehensive test.

### Scenario: Predict Investment Failure

The task at hand is to develop a classification model that can accurately predict the "Investment Failure." You'll be provided with a dataset containing features such as investor demographics, risk tolerance, investment details, past performance, and market sector. Participants are expected to preprocess the data, engineer relevant features, select appropriate algorithms, and fine-tune model parameters to achieve the best possible predictive performance.

### Columns

The description of the 18 columns are as follows:

1. Investor\_ID: A unique alphanumeric code assigned to each investor.
2. Age: The investor's age in years.
3. Annual\_Income: The investor's total income earned throughout a year.
4. Investment\_Amount: The sum of money invested in a specific opportunity.
5. Risk\_Tolerance\_Score: A numerical representation of the investor's willingness to accept risk (higher scores indicate a greater appetite for risk).
6. Months\_of\_Investing\_Experience: The duration of the investor's active participation in investing (in months).
7. Number\_of\_Active\_Investments: The count of different investments the investor currently holds.
8. Potential\_Return\_Rate: The estimated percentage return on an investment.
9. Investment\_Duration: The intended length of time (in months) the investment will be held.
10. Investment-to-Income\_Ratio: The ratio of the investment amount compared to the investor's annual income.
11. Education: The investor's highest level of educational attainment.
12. Employment\_Status: The investor's current employment situation (e.g., full-time, part-time, unemployed).
13. Marital\_Status: The investor's marital status (e.g., single, married, divorced).
14. Owns\_Property: Indicates whether the investor owns real estate (Yes/No).
15. Has\_Dependents: Indicates whether the investor has others who financially rely on them (Yes/No).
16. Investment\_Sector: The industry or market the investment is focused on (e.g., technology, healthcare, real estate).
17. Has\_Investment\_Advisor: Indicates whether the investor utilizes a professional investment advisor (Yes/No).
18. Investment\_Failure: Indicates whether a past investment resulted in a loss (Yes(1)/No(0)).

### Evaluation Metrics

The performance of the models will be evaluated using standard classification metric ROC-AUC curve.

### Kaggle Instructions

The competition is hosted on Kaggle.

* Competition link: <https://www.kaggle.com/competitions/iba-datathon-classification>
* Competition link 2: <https://www.kaggle.com/t/b60fdce106dc455fb8d658f20e15f424>
* To Join the competition **one team member** will use his email to open Kaggle and click the Join competition button to download the dataset (train\_data, test\_data).
* Each team is only allowed a maximum of 15 submissions
* Once you submit a prediction file you can view the leaderboard to see your overall rank/position.
* Make sure you join the competition with correct team\_name.

### Submission Requirements

Email the following at [tmahmood@iba.edu.pk](mailto:tmahmood@iba.edu.pk) by 6 pm (late submissions will not be accepted):

1. Executed Python notebook with comments.
2. Transformed dataset file.