

# Assignment 2: Predicting Healthcare Employee Attrition

**Assigned: 23/11/2023**

**Due: 22/12/2023**

## **Competition Overview**

In the face of unprecedented challenges posed by the COVID-19 pandemic, the healthcare industry is grappling with a surge in employee attrition, particularly among nurses. Addressing this issue is critical for maintaining the stability and quality of healthcare services. The Kaggle competition, "Predicting Healthcare Employee Attrition", focuses on leveraging machine learning and analytics to develop robust models that can accurately predict whether an employee is likely to leave their position.

## **Dataset Details**

The dataset provided for this competition is a synthetic compilation, drawing inspiration from the IBM Watson dataset for attrition. To align the data with the healthcare domain, employee roles and departments have been strategically modified. Additionally, some known outcomes for specific employees have been altered to enhance the challenge and foster the development of advanced machine learning models.

## **Objective**

The overarching goal of this competition is to encourage data scientists, machine learning practitioners, and analysts to employ cutting-edge techniques in predictive modelling. By accurately forecasting employee attrition, participants can contribute valuable insights to healthcare organizations, enabling them to implement targeted strategies for employee retention and ultimately enhance the stability and efficiency of healthcare services during these challenging times.

## **Competition Entrance**

<https://www.kaggle.com/t/7f63a0b7b5ae35ac492309e0758ed09d>

### Tasks 1 (60 Marks)

1. Create an account on <https://www.kaggle.com/>. You MUST set a **TEAM NAME** with the format in (**Student ID**). Example: 2036673.
2. Create a notebook on Kaggle. Conduct exploratory data analysis and data processing, train and validate your models, and generate your 'submission.csv' on test data using the notebook on Kaggle. Add necessary comments to your notebook. Download the notebook from Kaggle and **submit it to Learning Mall**, with the name in (Name\_Student ID). (30 Marks)
3. Submit your predictions ('submission.csv') for the test solution to Kaggle. Also, you are required to include your Kaggle score in your report (see below in Task 2). (30 Marks)

### Tasks 2 (40 Marks)

Write a 1-page report, which **must contain** 2 or 3 tables or figures.

- Name your report with Name\_Student ID.
- **Submit your report to Learning Mall.**

The report must cover:

- **Introduction:** (4 Marks)  
What is the background of this project? How is it related to Big Data?
- **Methodology:** (8 Marks)  
A. Data Preprocessing  
What are the steps of data pre-preprocessing explored before training? Data visualization, data cleaning and reduction, normalization and discretization, feature selection, imbalanced data, etc. No need to cover all of them.  
B. Classification Algorithm  
How does it work? Explain the algorithm or framework.
- **Results:** (10 Marks)  
Are there other competitor models for this project? How does it compare to your technology?
- **Discussion:** (4 Marks)  
What are the good aspects, and what are the bad aspects? Be sure to add a

sentence on “**contributor thoughts:**” What are your own unique thoughts on the pros and cons of the technology? Do you envision an extension that might be helpful?

- **Conclusion:** (4 Marks)

Summarize the 2 to 4 points you think are most important.

**Concise, information-rich content.** For each of the sections above, you will not simply be graded on having content but on the quality of the content and how well it answers the questions in concise, clear, and engaging terms.

**Style.** (10 Marks)

In order to make your report consistent and visually appealing, as well as to make the evaluation of your work fairly, each page should be conformed to the following specifications:

- Margins: approx. 0.5” on all 4 sides.
- Columns: 2 with approx. 0.3in margin; justified text
- Fonts:
  - Body text: Times New Roman, 11pt.
  - Section headings: Calibri 13pt bold-Italic
  - Within captions, tables, figures, or images: Calibri 9-11pt.
- Line Spacing:
  - Body text: Single (1.0)
  - Section headings: 6pt spacing above heading

**Academic Honesty.** Copying chunks of code or problem-solving answers from other students, online or other resources is prohibited. You are responsible for both (1) not copying others’ work, and (2) making sure your work is not accessible to others. Assignments will be extensively checked for copying of others’ work. Problem-solving solutions are expected to be original, using concepts discussed in the book, class, or supplemental materials but not using any direct code or answers. Please see the syllabus for additional policies.