Assignment 2: Will your employees leave?

Assigned: 18/11/2022 **Due: 16/12/2022**

The main focus of INT303, the class, is to give you the fundamental knowledge of big data such that you can tackle a variety of situations yourself, but you shouldn't always need to reinvent the wheel from the basics when others have been perfecting the wheel you need potentially for years or decades.

Goals

- Programming language Python and its libraries NumPy (to perform matrix operations) and SciKit-Learn (to apply machine learning algorithms)
- Practice summarizing a potential complex topic into usable information, distilling it down to the important points.
- Determining which modern big data libraries and tools are available for their project goals.
- Several machine learning algorithms (decision tree, random forests, extra trees, linear regression).
- Feature Engineering techniques.

Problem

Employee attrition has become a focus of researchers and human resources because of the effects of poor performance on organizations regardless of geography, industry, or size. The goal of the project was to predict if an employee is likely to quit from the job based on a set of data. We used the Kaggle competition " Will your employees leave? " (see https://www.kaggle.com/competitions/int303-big-data-analysis-2223-s1/data) to retrieve necessary data and evaluate the accuracy of our predictions. An IBM's fictional dataset has been split into two groups, a 'training set' and a 'test set'. For the training set, we are provided with the outcome (whether or not an employee quit). We used this set to build our model to generate predictions for the test set. For each employee in the test set, we have to predict whether or not the employee quit from the job. Our score was the percentage of correct predictions.

Competition Entrance

https://www.kaggle.com/competitions/int303-big-data-analysis-2223-s1/overview

Tasks 1 (40 Marks)

1. Create an account on https://www.kaggle.com/. You MUST create an ID with the format in (XJTLU_Student ID). Example: XJTLU_1921013.

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). (*10 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 (60 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:** (*6 Marks*)
  Why should we care about this technology? How is it related to Big Data?
- **Methodology:** (*14 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:** (*14 Marks*)
  Are there benchmarks for its use? How does it compare to similar technology?
- **Discussion:** (*8 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:** (*8 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.