CSE140: Introduction to Intelligent Systems

Lab assignment

Due date: 01-Sep-2021 11:59PM

1. The dataset (attached as 'Data.xlsx') provides anonymized biodata information for graduates along with their respective skill scores. Specifically, the following information is available for every engineer:
    a. Scores on Aspiring Minds' AMCAT - a standardized test of job skills. The test includes cognitive, domain and personality assessments.
    b. Personal information like gender and date of birth.
    c. Pre-university information like 10th and 12th grade marks, board of education and 12th grade graduation year.
    d. University information like GPA, college major, college reputation proxy, graduation year and college location.
    e. The outcome whether the engineer, at the end of graduation gets a "High Salary"(labelled as 1) or "Low Salary" (labelled as 0) in the last column.
2. The task is to make a ML-based prediction system to predict if the graduates will get "high salary" or not. Specifically, you will have to do the following:
    a. Study the given data thoroughly and then modify it/prepare it considering the end goal at hand - that is classifying whether the student will get a "high salary" or not.
    b. Divide it into training and testing sets.
    c. Using the sklearn library, train a logistic regression model using the training data.
    d. Predict the outcome for the testing data.
    e. Using the predicted labels and actual labels, find out accuracy, confusion matrix and class-wise accuracies.
    f. Also, you need to experiment with data preparation - analyze what modifications produce better results and why. Try out different ratios of train-test splits (60-40, 70-30, 90-10 etc.), shuffling the data etc.
    g. You do not have to connect this data or the prediction system with your career advisory system that you have already built in assignment #1. This can be independent of that. The aim is to get you to build a ML model for prediction of high salary based on training data.
3. Demonstrate the program to your assigned TAs.