Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 6

Due: March 29, 2023(100 points)

Team Member: Lingyi Zheng, WeihongYang, Wei Zhang

Repository: https://github.com/EloiseZLY/HeartAttack_datamining

Your assignment for this week will be to send us a 2-page project proposal. The Google Docs version of this proposal is available here. You can choose the format you want to use, but make sure to include the following information to your document:

(To complete your proposal submission, create your project's repository in your own Github namespace, and upload your provide your URL in Gradescope. For example, my Git handle is kni-neu, and my project repository is: https://github.com/kni-neu/project You can use my template or any other template that you might find appropriate. (Or, it is acceptable to use no template at all.) An example set of organized proposal sections is shown below.)

- What problem are you going to be tackling on your project?

The problem I am going to tackle on this project is predicting whether a patient will die or not based on their age, sex, diagnosis, length of stay, and medical charges. The goal of this project is to choose an accurate model that can predict patient mortality using readily available clinical data, which can help healthcare professionals identify patients who are at high risk of mortality and take proactive measures to prevent it.

- Why is that an interesting/useful application of data mining?

This is an interesting and useful application of data mining because it can help healthcare professionals identify patients who are at high risk of mortality and take proactive measures to prevent it. By doing this project, we can help patients make better predictions on their health and increase their life's healthiness and lower their anxiousness.

- What models/techniques (clustering/classification/etc.) are you envisioning to apply?
    - Naive Bayes
    - KNN
    - Decision Trees

    - Logistics Regression
    - Support Vector Machine
    - Neural Networks

- Where are you going to get the data?

The data for this project is from the New York State Department of Health's Hospital Inpatient Discharges (SPARCS De-Identified) dataset, specifically the subset of patients with heart disease discharged from New York hospitals in 1993.

1 Problem

Make sure you make clear what problem are you going to solve in as concise and unambiguous manner. This section is typically 2-3 sentences.

Answer:

The goal of this project is to build several models to forecast the dying rate for the patients who have heart attack. We will choose an accurate model that can predict patient mortality using readily available clinical data, which can help healthcare professionals identify patients who are at high risk of mortality and take proactive measures to prevent it.

2 Background

This is where you tell us why solving this problem is important. What will people be able to do once you've solved this problem? How could it conceivably help people? What makes this application of data mining useful?

Answer:

Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease were men in 2015. Patients who are suffering a heart attack, but did not have surgery might face a higher likelihood of death. According to the clinical data on Heart Attack Patients discharged from all of the hospitals in New York State in 1993, where the admitting diagnosis was an Acute Myocardial Infarction (AMI), also called a heart attack, who did not have surgery, there are several factors that would influence patients' death. It could help medical staff to predict the patients' death. Our project will use different models to compare and discuss which model's predicted result is more reliable and accurate, and comes to recommended forecasting for future clinical trials.

3 Approach

Here, you can start detailing some specifics. Be sure to cover:

- Where are you going to get data?
- What data mining techniques will you use?

Answer:

- We got data from http://wiki.socr.umich.edu/index.php/SOCR_Data_AMI_NY_1993_HeartAttacks

- The work methodology aims at understanding the factors related to demographics (sex, age) and hospital codes (diagnosis and length of stay) affecting the dying in the hospital. The statistical analysis was based on Python by using existing algorithms and methods of machine-based data analysis. The Python language was applied for importing and manipulating the variable data and training different models, combined with using several packages to visualize the final result.