CICS 397A – Fall 2019

Homework 4

Due Thursday, December 5th at 11:59pm

You are encouraged to discuss the assignment in general with your classmates, and may collaborate closely with 1-2 other students on the design and logic of your solutions. If you choose to do so, you must indicate with whom you worked. In addition, the code you submit must be entirely your own; two students submitting the same code will be considered plagiarism.

Code must be written in a reasonably current version of Python (>3.0), and be executable from a Unix command line. You are free to use Python's standard modules for data structures and utilities. In addition, this assignment will require you to have both the pandas and scikit-learn libraries installed.

Objective

The goal of this assignment is to give you experience working with the scikit-learn toolkit (https://scikit-learn.org). You'll be asked to flesh out code that preprocesses data, builds models, and evaluates their predictive power. In order to complete the assignment, you will need to familiarize yourself with some of the core pieces of the toolkit. You are encouraged to consult the documentation as well as online resources such as StackOverflow.

The Data

The exercises are based on a data set consisting of 64 financial indicators associated with 5,910 Polish companies collected from 2000-2012. All data is stored in a csv file called `polish_bankruptcy_data.csv`. The last column holds the class value, with a 1 indicating that the company went bankrupt within a 5-year period and a 0 indicating that it did not.

The Code

Some stub code can be found in `polish_bankruptcy.py`. The `main()` function goes through a series of steps, building classifiers and printing out their accuracy. For each step, you'll need to fill in the bodies of one or more of the functions defined further down in the code. Pay attention to the comments, as they'll provide you with hints on the functionality of each method as well as the data types of the return values. To run the code from a command line, you'll need to supply the path to the data file as an argument:

```
python polish_bankruptcy.py /path/to/polish_bankruptcy_data.csv
```

Grading

We will run your modified `polish_bankruptcy.py` and test the output of each function for correctness. Your grade will be determined by how many of the exercises achieve the correct output, with partial credit being awarded.

What to Submit

You should submit:

- A modified `polish_bankruptcy.py`, containing your code
- A readme.**pdf**, containing
    - Your name
    - The most and least interesting topics we've covered this semester
    - Anyone with whom you worked with on the assignment (see note above)
    - Notes or warnings about you code (what you got working, what is partially working, and what is broken)