**Drexel University**

**College of Computing and Informatics**

**INFO 371 – Data Mining Applications**

**Course Project**

**Due Date: Sunday, Dec. 8, 2019 (End of the Week 11)**

**A. Requirements**

**TEAM (up to three members) assignment**: Please work with another student or two on this project. Please inform the instructor of your team membership before working on the project.

**B. Objective**

This project presents to you a real-world problem (credit card defaults) that requires deep understanding of the problem itself, the ability to deal with messiness of real data, and creativity to combine your understanding with machine learning skills to solve the problem.

**C. Data Source**

The data is given in the CSV format, available at:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

According to the site:

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

**D. Problem (Concept)**

Build machine-learning models to predict credit card defaults based on demographic or payment status history data.

**E. Exploratory Data Analysis**

* **E.1. Demographic Variables**

Discuss how demographic variables such as education have any impact on payment default. Use these demographic variables: **SEX, EDUCATION, MARRIAGE, AGE.** Produce a distribution for each variable and color code the distribution with default.payment.next.month. Explain what you see in the distribution and whether the variable can help predict default.payment.next.month.Be specific about using the variables, their data type, potential impact on default.payment.next.month and provide your reasoning.

* **E.2. Payment (History) Status Variables**

Discuss how payment status data have any impact on payment default. Use the payment status variables here: **PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6.** Produce a distribution for each variable and color code the distribution with default.payment.next.month. Explain what you see in the distribution and whether the variable can help predict default.payment.next.month.

Be specific about using the variables, their data type, potential impact on default.payment.next.month and provide your reasoning.

**F. Data Classification**

* **F.1. Naive Bayes Classification**

Build Naive Bayes models using different subsets of the attributes as predictors. Evaluate and summarize the results in tables, figures, or other well-structured formats.

* **F.2. Decision Tree**

Build decision trees using different subsets of the attributes as predictors. Evaluate and summarize the results in tables, figures, or other well-structured formats.

**G. Discussion**

Compare the results from the different models, with different parameters, and using different variables. Which model appears to be better (best) given the stated problem and identified data? Which one performs better on your two testing data rows? If it predicts correctly, what (in the model and/or in the data) are helpful? If it does not predict the correct answer, what (in the model and/or in the data) are missing?

**H. What to Hand In**

1. A well-structured report documenting the steps, data analyses, results, and discussions in MS Word or PDF format. Your report must contain the following content:
   1. For each step, describe the requirements and show the results including necessary screenshots. For each table and figure, you must add sufficient commentary to explain what the table or the figure is about.
   2. Detailed descriptions about the design of the models. Discussion about their pros and cons.
   3. Necessary intermedia steps or raw data to allow any readers to reproduce your results.
   4. Performance evaluation including confusion matrix, accuracy, precision, recall, and F1-Measure.
   5. A discussion section including the comparisons, analyses, and future considerations.
   6. A conclusion section describing lessons learned from the project.

**I. How to Hand In**

1. Please name your report file as **INFO371-project-yourFirstName-yourLastName.docx**.
2. Submit your report file through the course website in the **Blackboard Learn** system.

**J. When to Hand In**

1. Submit your report no later than **11:59pm** in the due date.
2. There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

**K. Written Presentation Requirements**

Images must be clear and legible. Assignments will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be either 11 or 12 point.