Data Mining (Section 001 and 900)
Instructor: Sharma Chakravarthy
Project I: Classification

Made available on:	**9/01/2020**
Milestone 1 Due on:	**9/15/2020 (11:59 pm) –** *canvas is setup for this.*
Submit to:	*Canvas (uta.instructure.com)*
	1 zip folder as indicated below. *No late submissions!*
Complete Project Due on:	**9/29/2020 (11:59 PM)**
Submit to:	**Canvas (uta.instructure.com)**
	1 zipped folder containing all the files/sub-folders
	Late submissions have a penalty as indicated!
Weight:	**15% of total**
Total Points:	**100**

This project is about understanding the classification approach to mining (supervised learning) using several widely-used techniques. For all the project you will use R and RStudio (https://www.rstudio.com/products/rstudio/download/).
You need to install R as part of RStudio installation on your machine. Downloads are available for windows, Mac, and Linux. There is also a wikibook that explains the algorithms available in R (https://www.webpages.uidaho.edu/~stevel/517/Data%20Mining%20Algorithms%20In%20R.pdf) which you are likely to use, unless you want to develop your own code/algorithms! Hence you should familiarize yourself with both R and R Studio ASAP. You should also be able to pre-process the data set given to you as needed for analysis. Partitioning, random sampling, cleaning, may have to be done before you start applying the techniques. Some packages may be available in R. You will be developing snippets of code in R for your customization.

Submission of the milestone is mandatory without which you will lose assigned points. I have added this mainly based on the feedback received asking me to introduce milestones. Hopefully, this will encourage you to start immediately and finish the project on time. My goal is for everyone to complete all projects as specified and do well in the course. I hope that is your goal as well! If need be, I will also be asking the TA to make a presentation in the class on this project to make sure you have no problems with RStudio and can complete the milestone by the deadline.

I. Problem Statement:

In this project, you will be given a bank data set (about 41,000+ rows) for analysis using two classification approaches: decision trees and Naïve Bayes. First, you should become comfortable with both the approaches by reading the text book in addition to the material presented in the class. As the analysis aspect is what you will be graded on, playing with small data sets (e.g., data set shown in the class) to understand the algorithm and the parameters will be important.

The data set has a number of attributes for a client and has an outcome field for "has the client subscribed to a term deposit? (binary: 'yes','no')". The goal is to develop a model using training data and classify the test data into a correct outcome label.

A sample row of input looks like (*Not all columns are shown here*):

ousing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
yes	no	cellular	sep	wed	...	1	5	1	success	-1.1	94.199	-37.5	0.886	4963.6	yes
no	yes	telephone	jun	fri	...	4	999	0	nonexistent	1.4	94.465	-41.8	4.967	5228.1	no

Below are the descriptions of the features for the data set:

Feature description

Bank client data:

- 1 - age (numeric)
- 2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- 3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- 4 - education (categorical: basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- 5 - default: has credit in default? (categorical: 'no','yes','unknown')
- 6 - housing: has housing loan? (categorical: 'no','yes','unknown')
- 7 - loan: has personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

- 8 - contact: contact communication type (categorical: 'cellular','telephone')
- 9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- 10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

- 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- 14 - previous: number of contacts performed before this campaign and for this client (numeric)

- 15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes

- 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
- 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
- 18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- 19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
- 20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- 21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

You are asked to study the data set, count the number of unknown values, and clean the data set to "eliminate/drop tuples with unknown values". After that, you need to drop [*marital, default, housing, loan, contact*] attributes. Finally, identify a few attributes (3 to 5) based on histogram analysis (for relevance). After that you will take a random subset (specified for each team separately using the given seed) of the data for your analysis (as specified). You will be applying both the approaches on this data and do the analysis, visualization etc. as specified in this project description later.

After random selection of a subset, you will use a ratio on that subset for training and testing the model generated by the training data. Again, the splits for training and testing will be given to each team.

Understanding classification mining: Understand the nuances of each approach by using a small data set and studying the output. The data set should be small enough to check it manually and allow small changes to see the results of those changes on the classification (error, height of the tree, how attribute values are used for splitting, etc.)

II. Installing RStudio

- Install RStudio on your machine. Install R when asked.
  - For Decision tree, you can use rpart library, more details could be found on- https://cran.r-project.org/web/packages/rpart/rpart.pdf
  - For Naïve Bayes, you can use (e1071) and "caret" library.
- Understand R, understand the packages available and which you want use and why.
- Understand the parameters to be set, their role in classification etc.
- Analyze the output, summaries that can be produced
- See how you can visualize various aspects of the technique and benefit from that.
- Always analyze a small data set to understand the details. You can use the data set shown in class or another data set. The key is that it should be small to be checked manually and instructive enough to understand the nuances of the techniques used

III. What you are asked to do:

1. Clean the data set, take a random sampling as specified and run the decision tree algorithm using the specified splits (again randomly). You are welcome to try other splits, but you are required to submit the analysis of the splits specified.
   a) Pre-processing (eliminating attributes with unknown values, and identifying 3 to 5 attributes as not relevant or contributing) (10 pts, milestone)
   b) Do the classification using Information Gain
   c) Withhold one column of the data set and repeat the above classification. You need to explain as part of the analysis how you determined the column to withhold and the reason for change in accuracy.
   d) Do the classification using GINI index (discussed in the class and text book)
   e) Withhold one column of the data set and repeat the above classification. You need to explain as part of the analysis how you determined the column to withhold.
2. Use the same sample and splits and apply Naïve Bayes for classification

3. For the above, present the following analyses as part of your report. For each of the analysis indicated above, where appropriate, include the confusion matrix, precision, recall, F1 score, and accuracy (also any other information to make the analysis interesting!). Optionally show the ROC curve.

   a. Individual results of GINI, IG, and Naïve Bayes for assigned splits
   b. Compare Information Gain, GINI, and Naïve Bayes results
   c. Basis for dropping attributes
   d. After building the classification model (for each training set) check which attributes have more significance and compare 'with vs without' the columns that you identified earlier based on attribute analysis. Explain whether you identified the correct attributes or not?
   e. Explanation of withholding attribute and its results

IV. Project Report

Please include the following sections in a **REPORT {.doc or .pdf format}** file that you will turn in along with your code. **Please limit your report to less than 10 pages with at least 11 font size. Anything beyond 10 pages will not be evaluated.**

- **Overall Status**
  Give a *brief* overview of how you went about doing this project. If you were unable to finish any portion of the project, please give details about what is completed and your understanding of what is not. (This information is useful when determining partial credit.)
- **File Descriptions**
  List any new files you have created and *briefly* explain their major functions and/or data structures. If you have added additional test cases, please summarize them.
- **Division of Labor**

Describe how you divided the work, i.e., which group member did what. Please also include how much time the team spent on this project. (This has no impact on your grade whatsoever; we will only use this as feedback in planning future projects -- so be honest!)

- **Problems encountered and how you handled them**

  List at least 3 problems you encountered (logical and not syntactic) during the completion of the project. Pick those that challenged you. This will provide us some insights into how we can improve the description and forewarn students for future assignments.
- **Most important, detailed comparisons and Analysis, as indicated above**

V. What to Submit

a. For the milestone (10 pts)

- **Report for:**
  - Process used for pre-processing
  - Basis for dropping attributes
  - Answering questions on the process (5 mins)

b. For completed project (90 pts)

- After you are satisfied that your project is complete, you upload it to canvas for grading. Please submit your **project report** and a table of routines/algorithms used/developed **in a zipped folder**. It may have sub-folders (one for decision tree and one for Naïve Bayes, for example.)
- All the above files should be placed in a single zipped folder named as '**Proj1Fall20_team_<teamNo>**'. **Only one zipped folder should be uploaded using canvas**
- You can submit your zip file at most 3 times. The latest one (based on timestamp) will be used for grading. So, be careful in what you turn in and when!
- **Only one person per team should turn in the zip file!**
- To discourage late submissions, a penalty of 20% per day (no partial penalty) will be imposed. This means that if your submission is delayed by more than 5 days, do not even bother submitting. We certainly do not want this delay to hurt your next project!

VI. Coding Style:

If you write any code in R, please follow the coding guidelines for R.

VII. Grading Scheme for the Completed Project:

The project will be graded using the following scheme (out of 80 points). The report should contain a section/para of analysis for each item below and the team should be able to answer questions on how they arrived at this analysis:

1. GINI classification using decision tree: 10
2. IG classification using decision tree 10

3.	Comparison after withholding an attribute (IG and GINI)		10
4.	Classification using Naïve Bayes approach:		10
5.	Comparison of IG, GINI with Naïve Bayes		15
6.	Analysis of training/test splits	10	
7.	Analysis of dropping attributes	5	
8.	Discussion of 3 problems encountered and solved		5
9.	Answering questions during the demo	5	
10.	Visualization and presentation	10	
	TOTAL (100)		90 (+10 for milestone 1)

Accuracy matrix and ROC curves are not considered as visualization. I want you to figure out what to visualize and how on your own to help understand the results of this analysis.

Five days after the due date, the submission will be closed.

You are welcome to use your laptop (windows, apple, or Linux). It is your responsibility to have it working in your environment. You cannot debug code and fix problems during the demo! Any code you have written for the project should be included in the uploaded folder as a src folder.

… All the Best …