Assignment 2

Q1. This assignment requires understanding the concepts explained in data mining, predictive analytics and machine learning sections.

(a) For this exercise, your goal is to build a model to identify inputs or predictors that differentiate risky customers from others (based on patterns pertaining to previous customers) and then use those inputs to predict new risky customers. This sample case is typical for this domain. The sample data to be used in this exercise is CreditRisk.xlsx .

The data set has 425 cases and 15 variables pertaining to past and current customers who have borrowed from a bank for various reasons. The data set contains customer-related information such as financial standing, reason for the loan, employment, demographic information, and the outcome or dependent variable for credit standing, classifying each case as good or bad, based on the institution's past experience.

Take 400 of the cases as training cases and set aside the other 25 for testing. Build a decision tree model to learn the characteristics of the problem. Test its performance on the other 25 cases. Report on your model's learning and testing performance. Prepare a report that identifies the decision tree model and training parameters, as well as the resulting performance on the test set.

You can use either R (and related packages e.g., rattle Package) or a GUI-based software Weka.

To use Weka go through Learning Resource for Weka decision tree

See R resources posted in the blackboard.

(b) Using the same dataset also develop a Neural Network (NN) model using either R or Weka (Multilayer Perceptron)

(c) Compare and evaluate the model performances of decision tree and NN. (use 10-fold cross validation and Leave-one-out for classification assessment). Also generate ROC plots. Explain and discuss the results.

(d) How can you improve the prediction accuracy? What are the pre-processing or post- processing steps required to improve the accuracy? Finally, implement them to show that they really improve accuracy?

Report everything in a .pdf file with descriptions of preprocessing steps, model development, model output (including plots) interpretation, explanations, and validations of the models and their significance.

Q2. In this exercise you'll use R package tidyverse (see chapter 4 of *Introduction to Data Science Data Analysis and Prediction Algorithms with R* by Rafael A. Irizarry. You need to go through chapter 4 before attempting the following questions. Also, see my lecture video in the blackboard (**Data wrangling with tidyverse)**.

Using dplyr functions (i.e., filter, mutate ,select, summarise, group_by etc.) and "murder" dataset (available in dslabs R package) and write appropriate R syntax to answer the followings:

a. Calculate regional total murder excluding the OH, AL, and AZ

b. Display the regional population and regional murder numbers.

c. How many states are there in each region?

d. What is Ohio's murder rank in the Northern Central Region (Hint: use rank(), row_number())

e. How many states have murder number greater than its regional average.

f. Display 2 least populated states in each region

g. Find the state in each region whose population is closest to its regional average population.

Use pipe %>% operator for all the queries. Show all the output results.

Q3.

The solutions to the following problems require understanding the concept of mathematical optimization and linear programming and Goal Seek described in chapter 8 (page 477 in Section 8.6). You need to use Excel Solver Add-in to develop the model for some of the problems. See related resources in the blackboard.

If Excel Solver is not already loaded, follow these steps to load Solver. (Only for PC, not Mac).

Use separate Excel worksheet to develop a model for each problem. Rename the worksheets to reflect the problem number. Upload one single Excel file in the Blackboard.

Problem a.

At a chip manufacturing plant, four technicians (A, B, C, and D) produce three products (products 1, 2, and 3). The chip manufacturer can sell 80 units of product 1 this month, 50 units of product 2, and at most 50 units of product 3. Technician A can make only products 1 and 3. Technician B can make only products 1 and 2. Technician C can make only product 3. Technician D can make only product 2. For each unit produced, the products contribute the following profit: product 1, \$6; product 2, \$7; product 3, \$10. The time (in hours) each technician needs to manufacture a product is as follows:

Product	Technician A	Technician B	Technician C	Technician D
1	2	2.5	Cannot do	Cannot do
2	Cannot do	3	Cannot do	3.5
3	3	Cannot do	4	Cannot do

Each technician can work up to 120 hours per month. How can the chip manufacturer **maximize** its monthly profit?

Problem b.

Suppose that each day, northern, central, and southern California each use 100 billion gallons of water. Also assume that northern California and central California have available 120 billion gallons of water, while southern California has 40 billion gallons of water available. The cost of shipping one billion gallons of water between the three regions is as follows:

	Northern	Central	Southern
Northern	\$5,000	\$7,000	\$10,000
Central	\$7,000	\$5,000	\$6,000
Southern	\$10,000	\$6,000	\$5,000

We will not be able to meet all demand for water, so we assume that each billion gallons of unmet demand incurs the following shortage costs:

	Northern	Central	Southern
Storage cost/billion gallon short	\$6,000	\$5,000	\$9,000

How should California's water be distributed to **minimize** the sum of shipping and shortage costs?

Problem c.

A student is trying to determine how many hours of studying to devote to each of his subjects to **maximize** his overall grade point average this semester. To do so, the student predicts the grade average he will receive for studying different amounts of time in each of his classes. The table below displays his predictions.

Hours per Week	**Calculus**	**Chemistry**	**Physics**	**Economics**
1-5	75	76	65	85
6-10	84	87	81	92
11-15	93	94	91	97

The student wants to study no more than a total of 35 hours per week. He estimates that the amount of time he should study physics is at least double the amount of time he should study economics, and the amount of time he should study calculus is in between those two values. Formulate this problem as an optimization model and use the solver to find the optimal solution. (Hint: Use Excel's VLOOKUP function; Select "Evolutionary" for solving method in Solver)

Problem d.

The growth of Internet users is given in the following table:

Year	Number of Users (In Million)
1995	16
1996	36
1997	70
1998	147
1999	248
2000	361
2001	531
2002	584
2003	719
2004	817
2005	1018

2006	1093
2007	1316
2008	1574
2009	1804
2010	1971
2011	2267
2012	2497
2013	2802

Fit a nonlinear growth model $\boldsymbol{y} = \frac{\boldsymbol{K}}{(\boldsymbol{1}+\boldsymbol{\alpha} \boldsymbol{e}^{-\boldsymbol{\beta t}})}$ (where, y = number of Internet users, t = time; K, $\boldsymbol{\alpha}, \boldsymbol{and}$ β are model parameters) of the growth of the Internet by estimating the model parameter alpha, beta and K using Excel Solver by **minimizing** the overall error term. Also, plot the scatterplot graph of the actual growth and the fitted curve (continuous).

Upload one single pdf file with the answers of Q1 & Q2. For Q3, just upload the Excel file.

This is a group assignment. **Do not share or interact between the groups**.