Assignment 03

Task 01:

In this exercise, you will perform tasks faced by a data scientist working for a game developer company. The company develops games for social networking platforms. Your job is to make sense of a dataset describing users from different perspectives. The dataset (sn game.csv) contains the following variables:

(a) gender : the gender of the user
(b) age: the age of the user
(c) edu: the highest education of the user
(d) salary: monthly salary of the user
(e) sn.conn: number of connections that the user has on the social network site
(f) sn.min: number of minutes spent on the social network site
(g) game.min: number of minutes spent on the social network site playing the game
(h) game.purchase: amount of money spent on purchasing extra features, upgrades, power-ups etc. in the game

You need to perform the following tasks:

• Exploratory data analysis: try to understand the different variables in the data. Identify the variables, based on exploratory data analysis methods, that you think have an effect on the money spent in the game

• Develop a regression model that the company can use to predict a new users' future spending in the game. The model should contain only the variables that were found as potentially important in the previous step.

Task 02:

In this exercise, you will have to analyze a dataset (haberman.csv) that contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on patients who had undergone surgery for breast cancer. The task is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 year (negative). More details on the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.names

- Build a logistic regression classification model using all the three predictor variables (Age, Year of operation, Number of positive axillary nodes detected). Divide the data set into training (70 %) and test set (30 %), use random state = 0, and follow the process of building a classification model as discussed in the course.

- Create the confusion matrix, calculate classification performance measures, and check the accuracy for the test set.

- Perform the previous steps but now using only the two variables Age, and Number of positive axillary nodes detected. How did the accuracy of the model on the test set change? Based on this, do you think Year is an important predictor variable?

Task 03:

In this exercise, you have to work with a data of customers, who bought one of two possible products. Your task is to consider this as a classification problem and build a model that will predict which one of the two products a customer would buy using the available data (i.e. which product should the company recommend to a new customer). The variables included in the data are the following:

- gender: the gender of the customer
- age: the age of the customer
- edu: the highest education of the customer
- salary: monthly salary of the customer
- previous.orders: number of orders by the customer
- previous.purchase: amount of money spent by the customer
- favourite.genre: favourite genre of the customer
- recommendation: the ID of a product chosen by the customer

Build a logistic regression model to predict the product that is more likely to be purchased by the customer. Regarding the categorical columns, create dummy variables (one-hot encoding) for gender, edu, and favourite.genre to be used in the model. First, create a training set and test set, with test size = 0.20 (set random state to be 0), and after building a model, evaluate predictions for both the test and training sets. In particular check the difference of accuracy for the training and test; do you see a result that you would expect there? Second, perform cross-validation with cv = 5 and scoring = 'accuracy'. What is the minimum and maximum score you observe after building the models?

Task 04:

In this exercise you have to work with the data in the file 'ItalianWineSamples.csv', that contains 13 chemical measurements on 178 Italian wine samples. More information about the data can be found here: https://archive.ics.uci.edu/ml/datasets/wine Your task is to perform K-Means clustering to the dataset; in the model building process, do not use the column 'Type'. After scaling the variables, determine the optimal number of clusters using the elbow-method introduced in the course, and then perform the K-Means analysis using the optimal cluster number you determined (set random state = 0). Based on the results, answer the following questions:

- How many observations are there in each cluster?
- What is the average of each variable in each cluster (the original, not the scaled variables)?
- Can you identify some variables that clearly have different average values for each cluster?
- Compare your K-Means Clustering solution with the 'Type' variable, which describes the wine varietal each wine sample belongs to out of three possible types. If you consider the created clusters, do any of them correspond to one of the categories in Type (i.e., most of the wines belonging to the cluster have the same Type), or are the clusters just a mix of all the categories in Type?