|  |  |
| --- | --- |
|  | **Introduction to Business Data Analytics** |

|  |  |
| --- | --- |
| **Homework #3 Part 2** |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(put your names above (incl. any nicknames))

Note: This is a team homework assignment. Discussing this homework with your classmates outside your MSBA team is a **violation** of the Honor Code. If you **borrow code** from somewhere else, please add a comment in your code to **make it clear** what the source of the code is (e.g., a URL would sufficient). If you borrow code and you don’t provide the source, it is a violation of the Honor Code.

Total grade: \_\_\_\_\_\_\_ out of \_\_\_100\_\_\_ points

***ATTENTION: HW3 has two parts. Please first complete the Quiz “HW3\_Part1” on Canvas. Then, proceed with Part 2 in the following page. You will need to submit (a) a PDF file with your answers and screenshots of Python code snippets and (b) the Python code.***

**(100 points) [Mining publicly available data] Use Python for this Exercise.**

**Please use the dataset on breast cancer research from this link:** [**http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data**](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data) **We have worked with this dataset in HW2. The description of the data and attributes can be found at this link:** [**http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names**](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names) **. Each record of the data set represents a different case of breast cancer. Each case is described with 30 real-valued attributes: attribute 1 represents case id, attributes 3-32 represent various physiological characteristics, and attribute 2 represents the type (benign or malignant). If the dataset has records with missing values, you can filter out these records using Python. Alternatively, if the data set has missing values, you could infer the missing values.**

[We have seen this data before – No need to explore the data for this exercise]

1. **We would like to perform a predictive modeling analysis on this same dataset using the a) decision tree, b) the k-NN technique and c) the logistic regression technique. Using the nested cross-validation technique, try to optimize the parameters of your classifiers in order to improve the performance of your classifiers (i.e., f1-score) as much as possible. Please make sure to always use a random state of “42” whenever applicable. What are your optimal parameters and what is the corresponding performance of these classifiers? Please provide screenshots of your code and explain the process you have followed.**

[part a is worth 25 points in total:

7 points for correctly optimizing at least two parameters for the Decision Tree and providing screenshots/explaining what you are doing and the corresponding results

7 points for correctly optimizing at least two parameters for the kNN and providing screenshots/explaining what you are doing and the corresponding results

7 points for correctly optimizing at least two parameters for the Logistic Regression and providing screenshots/explaining what you are doing and the corresponding results

4 points for contrasting their performance of all three algorithms and discussing which one would you prefer to use]

1. **Build and visualize a learning curve for the logistic regression technique (visualize the performance for both training and test data in the same plot). Please provide screenshots of your code and explain the process you have followed.**

[part b is worth 25 points in total:

8 points for correct visualization of learning curve for in-sample sample performance – show the performance for 10 different sizes - provide screenshots of your code and explain the process you have followed.

8 points for correct visualization of learning curve for out-sample sample performance – show the performance for 10 different sizes - provide screenshots of your code and explain the process you have followed.

9 points for discussing what we can learn from this specific learning curve – what are the insights that can be drawn]

1. **Build a fitting graph for different depths of the decision tree (visualize the performance for both training and test data in the same plot). Please provide screenshots of your code and explain the process you have followed.**

[part c is worth 25 points in total:

8 points for correct visualization of fitting graph for in-sample sample performance – show the performance for 15 different values- provide screenshots of your code and explain the process you have followed

8 points for correct visualization of fitting graph for out-of-sample performance – show the performance for 15 different values- provide screenshots of your code and explain the process you have followed

9 points for discussing what we can learn from this specific fitting graph – what are the insights that can be drawn]

1. **Create an ROC curve for k-NN, decision tree, and logistic regression. Discuss the results. Which classifier would you prefer to choose? Please provide screenshots of your code and explain the process you have followed.**

[part d is worth 25 points in total:

5 points for correct visualization of ROC graph for kNN – use optimal kNN from part a

5 points for correct visualization of ROC graph for Decision Tree – use optimal Decision Tree from part a

5 points for correct visualization of ROC graph for Logistic Regression – use optimal Logistic Regression from part a

2 points for showing all the ROC graphs in one single plot

3 points for showing AUC estimators in the ROC graph

5 points for discussing and correctly identifying which classifier you would use]