Marmara University – Faculty of Engineering – Department of Computer Engineering

Fall 2020 – CSE4063 Fundamentals of Data Mining
Project #1

Due: 16|21|23.12.2020.Wed|Mon|Wed 23:59

1) Dataset

a) Click the link "View All Data Sets" on URL http://archive.ics.uci.edu/ml. to get the data sets assigned:

No	Group Members	Dataset [Rows x Columns]	Presentation Slot
1	Cem Güleç * Buğra Akdeniz Kadir Hızarcı	Anuran Calls (MFCCs) [7,195x22]	17.12.2020 **Thursday 12.00 – 12.17**
2	Mehmet Nusret Odabaşı * Abbas Kutay Orhan Fatih Bayazıt	Early stage diabetes risk prediction dataset. [520x17]	17.12.2020 **Thursday 12.20 – 12.37**
3	Emin Kağan Kadıoğlu * Mert Mengü	Turkiye Student Evaluation [5,820x33]	17.12.2020 **Thursday 12.40 – 12.57**
4	Ayberk Ömer Altuntabak * Abdulhalik Şensin Amela Karmaj	Phishing Websites [2,456x30]	17.12.2020 **Thursday 13.00 – 13.17**
5	Diala Jassem M.B.J. * Münevver Sueda Kocatürk Nurhande Akyüz	Census Income [48,842x14]	17.12.2020 **Thursday 13.20 – 13.37**
6	Osman Mantıcı * Buse Batman Fatmanur Özdemir	MAGIC Gamma Telescope [19,020x11]	17.12.2020 **Thursday 13.40 – 13.57**
7	İlker Fener * Doğukan Deniz Halil İbrahim Şimşek	Letter Recognition [20,000x16]	22.12.2020 **Tuesday 12.00 – 12.17**
8	Zahide Gür Taştan * Merve Ayer Zeynep Naz Akyokuş	EEG Eye State [14,980x15]	22.12.2020 **Tuesday 12.20 – 12.37**
9	Halid Seyfullah Sert * Dilek Dündar Mert İlik	South German Credit (UPDATE) [1,000x21]	22.12.2020 **Tuesday 12.40 – 12.57**
10	Ayşenur Yılmaz * Belgin Taştan Kevser İldeş	Iranian Churn Dataset [3,150x13]	22.12.2020 **Tuesday 13.00 – 13.17**
11	Sedanur Kara * Berke Şahin Sinem Onal	Australian Sign Language signs [6,650x15]	24.12.2020 **Thursday 12.00 – 12.17**
12	Deniz Arda Gürhizin * Can Berk Durmuş Tarkan Batar	Firm-Teacher_Clave-Direction_Classification [10,800x20]	24.12.2020 **Thursday 12.20 – 12.37**
13	Furkan Akman * Burak Fidan Mustafa Sertaç Öztürk	Image Segmentation [2,310x19]	24.12.2020 **Thursday 12.40 – 12.57**
14	Ferihan Çabuk * Ali Berat Çetin Muhammed İsa Akbaba	Page Blocks Classification [5,473x10]	24.12.2020 **Thursday 13.00 – 13.17**
15	Ahmet Enes Gündüz * Hakan Yalçın Muhammed Fethullah Eroğlu	Shill Bidding Dataset [6,321x13]	24.12.2020 **Thursday 13.20 – 13.37**
16	Yasin Orhan * Abdullah Yazar Ahmet Hakan Ekşi	Pen-Based Recognition of Handwritten Digits [10,992x16]	24.12.2020 **Thursday 13.40 – 13.57**

a) The first students indicated by * sign are the group representatives.

b) Learn / Get information about your data.

2) Python Platform & Environment

a) Get a platform/environment for python work on, if you do not have any. Install it on your computer.

b) You may use any libraries you want; however, you should have complete understanding to use and explain it in demo sessions.

c) Implement your work with your own code as possible as you can.

3) Model Construction: Classification

a) Do the data preprocessing steps, if required.

b) Training & Test

i) Decide how you will partition your data into training and test sets.

ii) Use holdout method for each of your classifiers separately.

iii) In addition to holdout method, use cross-validation for at least one of your classifiers.

iv) In addition to above, implement bagging ensemble method for your classifiers.

v) In addition to above, implement boosting ensemble method for your classifiers

c) Make the required type node settings.

d) Use your dataset to construct 6 classification models as follows:

i) Decision tree using gain ratio.

ii) Decision tree using gini index.

iii) Naïve Bayes.

iv) Artificial neural networks with 1 hidden layer.

v) Artificial neural networks with 2 hidden layers.

vi) Support vector machines.

4) Implementation & Model Evaluation

a) Implement 6 algorithms above on your dataset using python.

b) Compare the performance and the results of 6 classifiers on your test dataset.

c) Compare the performances of your classifiers with performances of the relevant papers given on the site.

5) Presentation

a) You are going to present your work done online in 12 minutes at the time slot reserved for your group. Group members should equally participate the presentation. See the table above.

b) Prepare a presentation file discussing the details of your work done and results of the classifiers.

c) Your presentation should contain the following parts at least:

i) Problem definition

ii) Dataset

(1) Information about the dataset.

(2) Number of instances, columns, etc.

iii) Data preprocessing, cleaning

(1) Missing values, and how you conduct on these.

(2) Transformations and normalizations.

(3) Training and test dataset.

iv) Python implementation for each of the 6 classifiers

(1) IDE/environment used.

(2) Implementation details.

(3) Libraries used.

v) Model evaluation & performance results

(1) Confusion matrices.

(2) Values of accuracy, recall, precision etc.

(3) Comparison of all 6 classifiers.

vi) Conclusion

6) Demo with Presentation

a) You are going to demonstrate your work done online in 5 minutes after your presentation. See the table above.

b) You are going to have 17 minutes in total for your group's session (12 minutes for presentation, and 5 minutes for demonstration).

c) Please keep in mind that all the presentation and demo sessions will be recorded.

d) All the students are expected to attend all sessions.

7) Related Questions & Answers

a) Prepare 5 questions and answers related to your topic. These questions may be asked to other students.

b) Question types can be multiple choice (single or multiple selection), fill in the blanks, matching, essay, etc.

c) Prepare a presentation file with 11 slides consisting these 5 questions and answers. First slide will be used for your topic and group members' info. Use 1 slide per each question, and 1 slide per each answer.

8) Evaluation

a) Your grade related to project #1 will cover 10% of your total grade at least; may increase subject to coronavirus issues.

b) Evaluation will be done out of 100 points:

i) [4 pts] Data set understanding.

ii) [4 pts] Data preprocessing.

iii) [4 pts] Training & test set partitioning.

iv) [4 pts] Models construction: Classification.

v) [36 pts] Implementation.

vi) [10 pts] Model evaluation, test and results, comparison.

vii) [20 pts] Presentation quality.

viii) [8 pts] Demo quality.

ix) [10 pts] Questions & answers quality.

9) Submission

a) You are going to submit the followings:

i) Python codes implemented.

ii) Presentation file.

iii) Questions & answers presentation file.

b) Write the following sentence in a text file: "We hereby swear that the work done on this project is totally our own; and on our honor, we have neither given nor received any unauthorized and/or inappropriate assistance for this project. We understand that by the school code, violation of these principles will lead to a zero grade and is subject to harsh discipline issues." Rename it as "we_swear.txt" and include this file in the zip submission file.

c) Only one of the group members (i.e. group representative, in short "GrRep") is going to submit the project using GrRep's info all the time. However, all group members should have a complete and comprehensive understanding of all the work done for all tasks and steps of the project.

d) Zip all your documents into a single file using filename GrRepStudentNumber_P1.zip (e.g. 150118123_P1.zip) and submit it to the site http://ues.marmara.edu.tr before deadline.

e) In case of any form of copying and cheating on solutions, all parts will get ZERO points. You should submit your own work. In case of any forms of cheating or copying, both giver and receiver are equally culpable and suffer equal penalties. All types of plagiarism will result in zero points from the homework.

f) If case of using your handwriting, your handwriting should be readable, clear and neat. If possible, do not use any handwriting.

g) Do not send project submissions through e-mail. E-mail attachments will not be accepted as valid submissions.

h) You are responsible for making sure you are turning in the right file, and that it is not corrupted in anyway. We will not allow resubmissions if you turn in the wrong file, even if you can prove that you have not modified the file after the deadline.

i) Grade evaluation may be done on selected parts of the project, so try to complete all parts of your project successfully.

j) No late submissions will be accepted.