

HOMEWORK 2

For this homework, you will exercise the visualization and preprocessing stages of a Machine Learning Project.

Using a data set related to your final project:

- create interesting questions related to the objective you aim to achieve with your project, and try to answer them and to find insights using summary statistics and visualization techniques
- search for common data problems and techniques (hint: related works) used for data preparation on your type of data/application
- Check for similar data problems in your data set and apply some of these techniques (cleaning/feature engineering/reduction) on your data

If possible, students should use a data set related to their research project (master's or PhD). If not possible, try to find a suitable public data set. Some possible sources are: Kaggle, UCI ML Repository, Papers with Code, any other you may find. The data set must be documented: number of instances, features and their descriptions, target variable(s), data source etc.

Students will be free to find or reuse a data set that is related to their course project. To find a suitable data set, the following platforms can be used: Kaggle, UCI ML Repository, Papers with Code. The data set must be documented: number of instances, features and their descriptions, target variable(s), data source etc.

After the analyses are completed, students will have to present their findings in a report that will be delivered via Moodle.

Deliverables:

1. Report in PDF format, following the IEEE or ACM conference format
2. Source code of the implementations, either via direct upload or github URL
3. If possible, the data set used in the analyses (or a fake sample if the data is protected)
4. A presentation with no more than 10 slides. Some projects will be presented during the synchronous activities.

Deadline:
05/12/2021