Home

The aim of the project is to get an **overview of how data science works**, the basic concepts and techniques of data analysis, understand how they work and gain intuition for their appropriate application in order to discover knowledge in data. They will also get an idea of what questions we can answer using data analysis and apply **basic machine learning approaches**. Emphasis is placed on data analysis and preprocessing, the use of machine learning methods, and ways to evaluate and compare data.

The project shall be developed **in pairs of** acceptable quality. The solution uses the **Python** programming language and available libraries for data science such as **pandas, numpy, scipy, statsmodels, scikit-learn**, etc... At each stage and activity, an executable **Jupyter Notebook** is uploaded to AIS containing all the transformations performed on the data with appropriate documentation. The submitted notebook must contain not only the code but also the results (calculated values, statements, visualizations, etc.) along with comments on the results obtained and the resulting decisions for the next steps of the data process. The ability to communicate and present relevant results well is an important component of the evaluation.

For each stage in the handed in notebook, indicate the **percentage of work done** by the members of the pair.

Data

https://drive.google.com/drive/folders/1vLIh5f3ix4KGQm0qX1cj2uGUVg0G7r5T?usp=share_link
(each pair has one dataset under the number you have on the exercise)

Mobile devices are now an integral part of Human-Computer Interaction (HCI), providing users with fast and convenient access to information and applications. This kind of interaction is intuitive and efficient, but it also opens the door to threats such as malware. Malware, or malicious software, can infiltrate mobile devices through infected apps or malicious websites, stealing sensitive data or gaining control of the device. To keep mobile devices secure, it is crucial for smart anti-malware software to detect it quickly and accurately, and then alert users in the shortest possible time. The core of such software is built on the basis of the provided data called logs and the detection is done using machine learning.

In the records (dataset for you) there is a dependent variable named "***mwra***" indicating *malware-related-activity* in one time interval. The dataset is backed up using the Rapid7 agent (https://www.rapid7.com) deployed on android mobile devices (https://developer.android.com). The dataset is partially preprocessed for the IAU project learning.

Project task

The QUEST

Each pair will work with their assigned dataset from week 2 onwards. **Your task** is to predict the dependent values of the variable "***mwra***" (predicted variable) using machine learning methods. In doing so, you will have to deal with several problems present in the data such as data formats, missing, skewed values and many more.

The expected **outcome of** the project is:

1. **the best** machine learning **model;**
2. **data pipeline** to build it based on the input data.

Phase 1 - Exploratory analysis: 15% = 15 points

1.1 Basic description of the data together with their characteristics (5b)

EDA with visualisation

- (A-1b) Analysis of data structures such as files (structures and relationships, number, types, ...), records (structures, number of records, number of attributes, types, ...)
- (B-1b) Analysis of individual attributes: for the selected significant attributes (min 10), analyse their distributions and basic descriptive statistics.
- (C-1b) Pairwise Data Analysis: Identify relationships and dependencies between pairs of attributes.
- (D-1b) Pairwise data analysis: identify the relationships between **the predicted** variable and other variables (potential predictors).
- (E-1b) Document your initial thinking to address the project assignment, e.g., are any attributes interdependent? Which attributes does the predicted variable depend on? whether records from multiple files need to be combined?

1.2 Problem identification, integration and data cleansing (5b)

- (A-2b) Identify and initially resolve problems in data e.g.: inappropriate data structure, duplicate records (rows, columns), inconsistent formats, missing values, skewed values. There may be other problems in the data not listed here.
- (B-2b) Missing values: try solving the problem with at least 2 techniques
  - removing observations with missing data
  - replacing a missing value by e.g. median, mean, ratio, interpolation, or kNN
- (C-1b) Outlier detection, try to solve the problem with min. 2 techniques
  - elimination of skewed or outlying observations
  - replacing the skewed value by the distribution bounds (e.g. 5%, 95%)

1.3 Formulating and statistically testing hypotheses about the data (5b)

- (A-4b) Formulate **two hypotheses** about the data in the context of the given prediction problem. Test the formulated hypotheses with appropriately chosen statistical tests.
  Example wording:

  > *android.defcontainer has on average a higher weight in the malware-related-activity state than in the normal state*

- (B-1b) Verify that your statistical tests have sufficient support from the data, i.e., that they have strong enough statistical power.

In the submitted report (Jupyter notebook) you should thus answer the questions:
Does the data have a suitable format for further processing? What problems are there in them? Are any attributes taking on inconsistent values? How do you solve these problems you have identified?

The report is due in the 5th week of the semester. The pair will present the completed phase to their practitioner in the Jupyter Notebook as appropriate for the exercise. Indicate in the notebook **the percentage of work done** by the members of the pair. The report will then be submitted electronically **by one member of the pair** to the **AIS** system by **23:59** on Sunday **20 October 2024**.

Phase 2 - Pre-processing of data: 15 points

At this stage, you are expected to perform **data preprocessing** for machine learning. The result will be a dataset (csv or tsv) where one observation is described by one row.

- **scikit-learn** can only handle numeric data, so something needs to be done with non-numeric data.
- Replicability of preprocessing on training and test datasets, so that you can repeat the preprocessing multiple times according to your need (iteratively).

When the preprocessing may have changed the shape and characteristics of the data, you need to implement EDA repeatedly according to your need. We will not score the techniques again. Document changes to the chosen techniques. You may solve the data problem iteratively at each stage and at all stages as needed.

2.1 Implementation of data preprocessing (5b).

- (A-1b) Divide the data into training and test sets according to your predefined ratio. Next, work only **with the training dataset**.
- (B-1b) Transform the data into a suitable format for ML i.e. one observation must be described by one line and each attribute must be in numeric format (encoding). zIteratively integrate also the data preprocessing steps from the first phase (missing values, outlier detection) as a whole.
- (C-2b) Transform attribute data for machine learning according to available techniques at a minimum: scaling (2 techniques), transformers (2 techniques), and others. The goal is that you test for effects and combine appropriately in the data pipeline (starting in Section 2.3 and in Phase 3).
- (D-1b) Justify your choices/decision for implementation (i.e., documenting)

2.2 Attribute selection for machine learning (5b)

- (A-3b) Identify which attributes (features) in your ML data are informative to the predicted variable (at least 3 techniques with inter-comparison).
- (B-1b) Rank the identified attributes in order of importance.
- (C-1b) Justify your choices/decisions for implementation (i.e. documenting)

2.3 Replicability of preprocessing (5b)

- (A-3b) Modify your code implementing the preprocessing of the training set so that it can be reused **to preprocess the test set** in a machine learning context without further modifications.
- (B-2b) Take advantage of the possibilities of **the sklearn.pipeline**

The report is due in the 7th week of the semester. The pair will present the completed phase to their practitioner in a notebook as appropriate for the exercise. Indicate the percentage of work done by the members of the pair. The report will then be submitted electronically **by one member of the pair** to the **AIS** system by **23:59** on Sunday **03/11/2024**.

Phase 3 - Machine Learning: 20 points

In data analysis, our goal may not only be to extract the knowledge contained in the actual data, but also to train a model that will be able to make reasonable **predictions** for new observations using **machine learning** techniques.

3.1 A simple classifier based on dependencies in the data (5b)

- (A-3b) Implement a simple **ID3** classifier with depth min 2 (including root/root).
- (B-1b) Evaluate your ID3 classifier using accuracy, precision, and recall metrics.
- (C-1b) Determine if your ID3 classifier has an overfit.

3.2 Training and evaluation of machine learning classifiers (5b)

- (A-1b) Use one **tree algorithm** in scikit-learn for training.
- (B-1b) Compare with one other **non-tree algorithm** in scikit-learn.
- (C-1b) Compare the results with the ID3 from the first step.
- (D-1b) Visualize the trained rules for **at least** one algorithm of your choice
- (E-1b) Evaluate trained models using accuracy, precision and recall metrics

3.3 Optimization aka hyperparameter tuning (5b)

- (A-1b) Try different hyperparameter settings (tuning) for the selected algorithm to optimize performance (without underfitting**)**.
- (B-1b) Try combinations of models (ensemble) for the chosen algorithm to optimize performance (without underfitting**)** .
- (C-1b) Use **cross** validation on the training set.
- (D-2b) Prove that your set best model is without overfitting**.**

3.4 Evaluation of the impact of the chosen solution strategy on classification (5b)

Evaluate your chosen solution strategies in terms of classification accuracy, whether they are effective for your dataset:

- (A-1b) Strategies to address missing values and outliers
- (B-1b) Data transformation (scaling, transformer, ...)
- (C-1b) Attribute selection, algorithm selection, hyperparameter tuning, ensemble learning
- (D-1b) Which is your **best** model for deployment?
- (E-1b) What is the **data pipeline** for building it based on your dataset **in production**?

Support all assessments with evidence. The best model should be stable, without overfit and without underfit. Its data pipeline should come with metadata, if that metadata is needed and produced in development.

The report is due in the 10th week of the semester. The pair will present the completed phase to their practitioner in the Jupyter Notebook as appropriate to the exercise. Indicate in the notebook the percentage of work done by the members of the pair. The report will then be submitted electronically **by one member of the pair** to the **AIS** system by **23:59** on Sunday **24 November 2024**.

Exercise activity: 10 points

The QUEST (just one of the two, either Q1 or Q2)

- Q1 (Image classification): classification task by number of classes
- Q2 (Time-series forecasting): predict the upcoming situation as accurately as possible based on historical data.

You don't have to use the whole dataset, just as much as you need for modelling with the justification that you have enough data.

Choose only one of the datasets to solve "*the quest*" according to the exercise block

Alqnatri

- Q1 Age Detection Dataset 50 MB, Age recognition, 3 classes: old, middle, young
- Q2 Household Electric Power 131 MB, Time-Series Forecasting

Bakonyi

- Q1 Periocular 14 MB, Facial recognition with medical mask, 2 classes
- Q2 S&P 500 Stocks 202 MB, Time-Series Forecasting

Kollár

- Q1 Covid-19 Image Dataset 166 MB, 3 classes: COVID-19, Viral Pneumonia, Norma
- Q2 Household Electric Power Consumption 131 MB, Time-Series Forecasting

Lytvyn

- Q1 Head CT - hemorrhage data 26 MB, Tumor detection, 2 classes: normal, hemorrhage
- Q2 Netflix 10+ Year Stock 303 KB, Time-Series Forecasting

Nguyen

- Q1 Covid-19 Image Dataset 166 MB, 3 classes: COVID-19, Viral Pneumonia, Norma
- Q2 S&P 500 Stocks 202 MB, Time-Series Forecasting

4.1 EDA and data preprocessing (5b)

- (A-4b) EDA and data preprocessing for your selected characteristics from the dataset
- (B-1b) Justify the choice of ML/DL methods in relation to your chosen dataset for 4.2

4.2 Modelling and evaluation (5b)

- (A-4b) Models your selected characteristics using appropriate ML/DL methods. The result of the modelling is the best model.
- (B-1b) Evaluate your approach and the result obtained

Support all assessments with evidence

The best model should be stable, without overfit and without underfit.

The report is due in the 12th week of the semester. The pair will present the completed phase to their practitioner in the Jupyter Notebook as appropriate for the exercise. Indicate in the notebook the percentage of work done by the members of the pair. The report will then be submitted electronically **by one member of the pair** to the **AIS** system by **23:59** on Sunday **08/12/2024**.