Assignment lab 4 - Clustering trials and multidimensional features visualisation

The purpose of this assignment

The goal of this lab is to check that the student has knowledge in the following topics:

- Dataset preprocessing.
- k-means clustering.
- Multidimensional features visualisation.

Business problem description

There is need to develop clustering model to segment credit card holders. Background and possible segmentation is described in blog post Perfect Credit Card Clustering with Machine Learning Models.

Dataset description

Kaggle dataset Credit Card Dataset for Clustering is used. See dataset webpage for features description and histograms.

Task

1. Install Python Anaconda distribution (or Python with required modules) if it was not installed before.
2. Create software project in GitLab. Use one of to https://gitlab.cs.ttu.ee or https://gitlab.com. See class 1 material for details.
3. Print out python and available modules versions.
4. Read dataset file to pandas data frame. **See lab1 for CSV file handling**
5. Save dataset description to file in results directory. **See lab3 for guideline and implementation.**
6. Preprocess dataset by removing identifier (unique for each customer) and replace missing values with feature mean value. **See lab1 and class 3 materials for details.**
7. Select desired number of clusters with help of elbow method. **WCSS plot shall be saved to results folder for review. See lab1 how to save plot to file.**
8. Visualise dataset with help of t-SNE dimensions reduction to 2 dimensions. **See class 9 materials and examples for details.**
9. Find clusters with k-means by using number of clusters defined in task step 7. **See class 8 materials and examples for details.**
10. Visualise dataset with found clusters with help of t-SNE dimensions reduction by adding different colour and symbol to each cluster. **See class 9 materials and classes 8, 9 examples for details.**

Guidelines

Project repository structure and files

Project shall consist of following files (excluding directories `.git` and also `builds` if local gitlab-runner is used).

```
.
├── .gitignore
├── .gitlab-ci.yml
├── .pylintrc
├── common
│   ├── describe_data.py
│   └── test_env.py
├── data
│   └── cc_general.csv
├── lab4.py
└── results
    └── .placeholder
```

`lab4.py` shall be created by student.

For `.gitignore`, `.gitlab-ci.yml`, `pylintrc`, `data` and `common` files from lab4 template shall be used.

NB! Be aware that if you want to use different file names you need to modify CI configuration and tests accordingly.

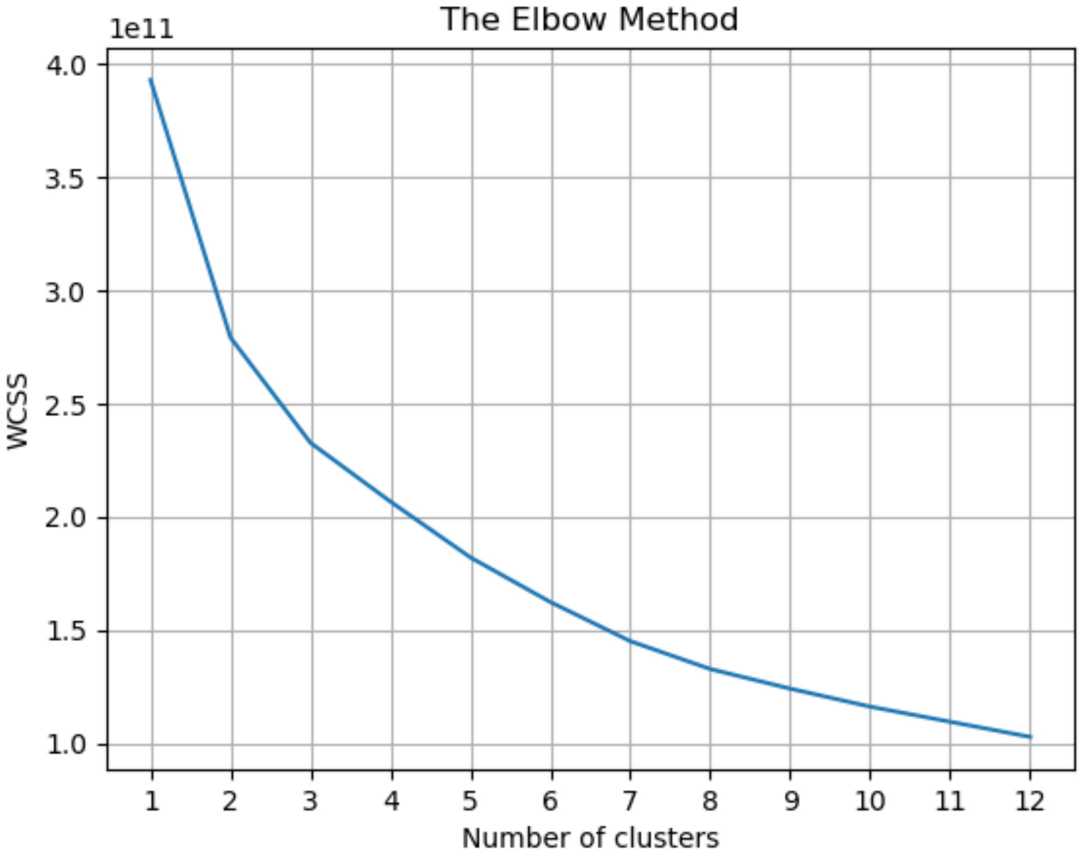
Producing comparable plots

In order to get plots comparable with example, set random state to 0 when creating both TSNE and KMeans objects.

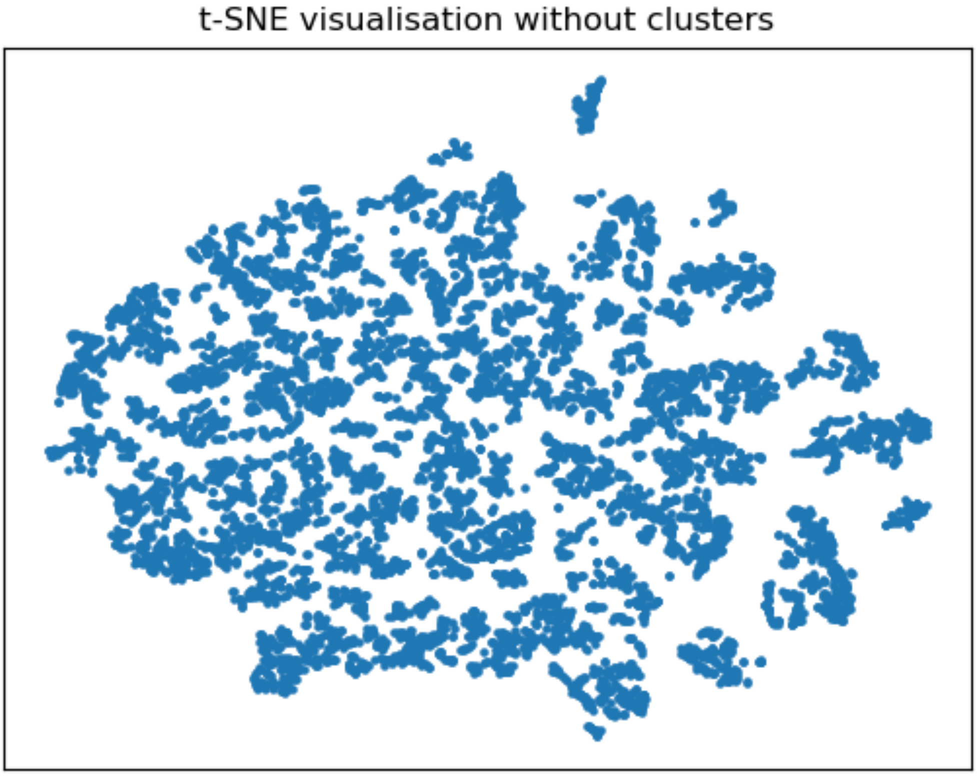
Plots examples

Your program shall create following plots to results directory and those shall be saved and downloadable as pipeline artefacts from GitLab.

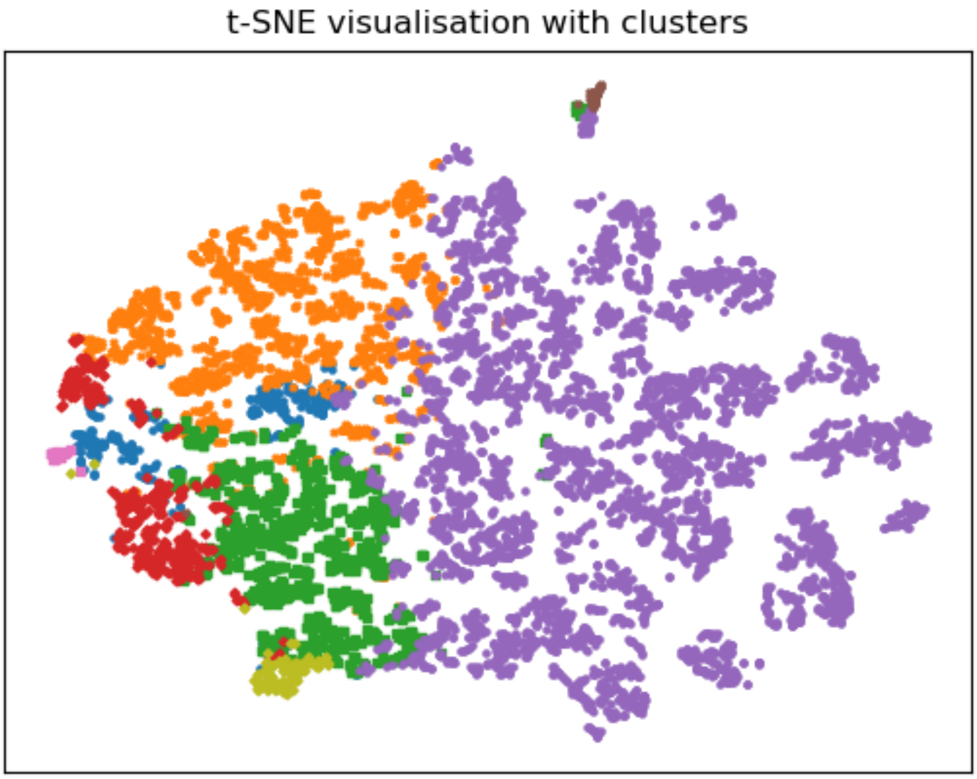
1. The Elbow Method plot (file: results/cc_wcss_plot.png)



2. t-SNE visualisation without clusters (file: results/cc_tsne_no_clusters.png)



3. t-SNE visualisation with clusters (file: results/cc_tsne_X_clusters.png) (X is number of clusters)



See Moodle for example downloadable archive for bigger images.

Automation and GitLab CI stages

- Check-files
  - Tests existence of required files and fail if all files are not present.
  - List repository files excluding `.git` and `build` directories.
- Lint
  - Test `lab4.py` formatting with `pep8`.
  - Lint `lab4.py` with `pylint` by using configuration from file `.pylintrc`.
- Run-lab
  - Run `lab4.rb`

Content of results directory is archived as build artefacts and can be downloaded.

Formatting and lint

autopep8 is used to test code formatting. autopep8 is supported by VS Code. For other editors it can be installed with conda:

```
$ conda install -c conda-forge autopep8
```

To run formatter from command line:

```
$ autopep8 --in-place lab4.py
```

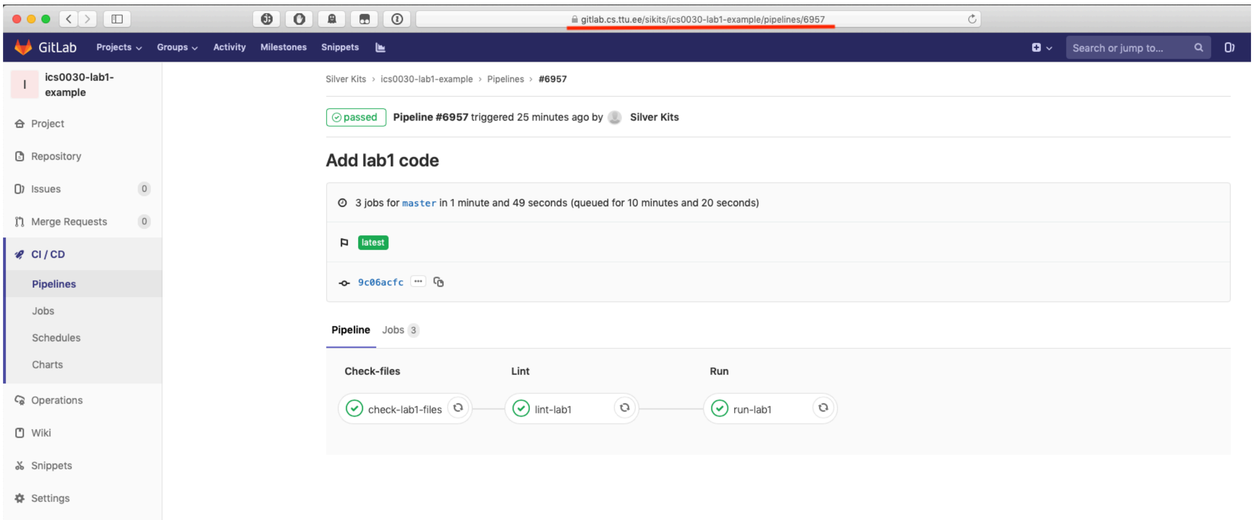
pylint is used for lint. Project template contains `.pylintrc`. Settings in this file are inline with VSCode default settings.

To run pylint from command line:

```
$ pylint lab4.py
```

Submission instructions

1. Be sure that your pipeline succeeds before submitting assignment in Moodle.



gitlab.cs.ttu.ee/sikits/ics0030-lab1-example/pipelines/6957

Silver Kits › ics0030-lab1-example › Pipelines › #6957

passed **Pipeline #6957** triggered 25 minutes ago by **Silver Kits**

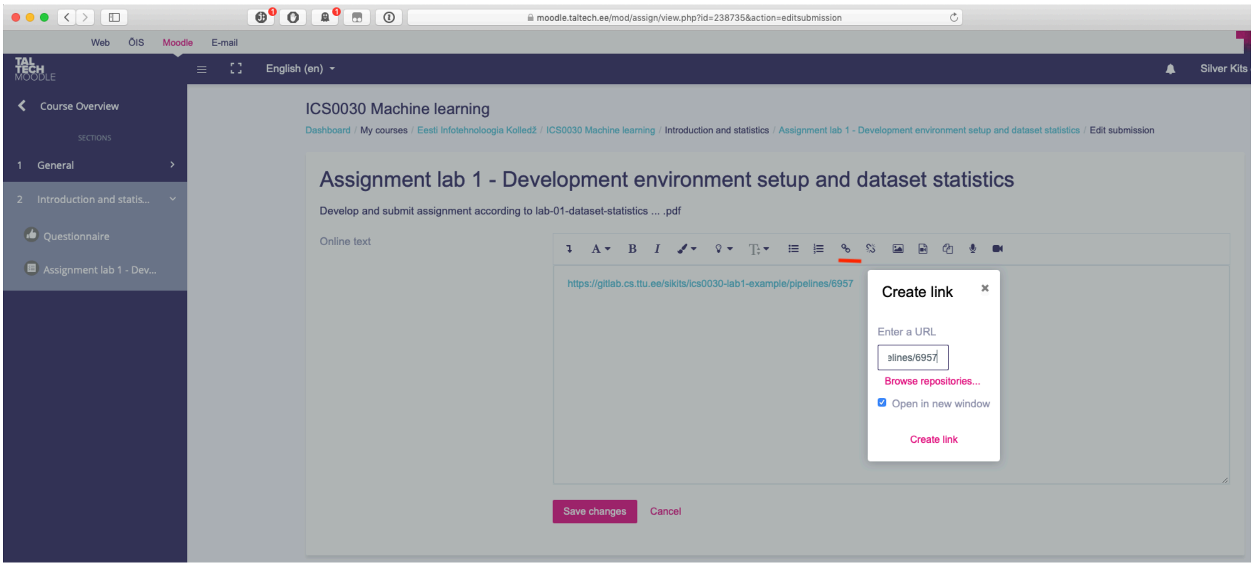
Add lab1 code

3 jobs for master in 1 minute and 49 seconds (queued for 10 minutes and 20 seconds)

latest

9c06acfc

2. Submit link to the pipeline as an answer in Moodle. **Please make link HTML URL!**



moodle.taltech.ee/mod/assign/view.php?id=238735&action=editsubmission

ICS0030 Machine learning

Dashboard / My courses / Eesti Infotehnoloogia Kolledž / ICS0030 Machine learning / Introduction and statistics / Assignment lab 1 - Development environment setup and dataset statistics / Edit submission

Assignment lab 1 - Development environment setup and dataset statistics

Develop and submit assignment according to lab-01-dataset-statistics ... .pdf

Online text

https://gitlab.cs.ttu.ee/sikits/ics0030-lab1-example/pipelines/6957

Create link

Enter a URL

ɘlines/6957

☑ Open in new window