Data Science (Info Management 2)

Introduction

Prof. Dr. Hendrik Meth
HdM, Stuttgart

Agenda

- Project Case Introduction
- Anaconda Introduction
- Warm Up Lab using Anaconda & Jupyter Notebooks

Schedule Info Mgmt. 2– Bachelor

#	Slot	Thursday, 14:15 – 15:45
1	24.03.2022	Case Introduction + ANACONDA and Jupyter Notebooks + Warm Up Labs
2	31.03.2022	PANDAS 1+2
3	07.04.2022	PANDAS 3 + Data Exploration (incl. Visualization)
4	14.04.2022	Build and Evaluate a Clustering Model
5	21.04.2022	Derive and Evaluate Association Rules
6	28.04.2022	Build and Evaluate a Classification Model
	05.05.2022	-
7	12.05.2022	Build and Evaluate a Regression Model

Online via Zoom

HOCHSCHULE DER MEDIEN

Course Grading

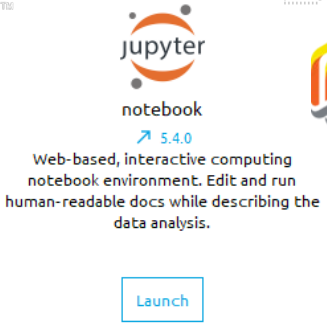
Element	Description	Contribution	Exam / Due Date
Exam	60 minutes exam consisting of questions covering lecture material and labs	60%	Within regular exam period
Project	Analysis case study to be solved with RapidMiner and Python in teams of five or six. Teams can be chosen by yourself. Data will be provided by 7th of April 2022 Individual contribution to coaching sessions AND team result will be graded	40%	Due Date: Sunday, 3rd of July, 2022

Case Study Organization

- Form teams of six and document them in Moodle (team selection will be made available tomorrow)
- Perform and document analysis until due date: Sunday, 3rd of July, 2022
- Upload one zip-file per team to Moodle containing
  - A PDF presentation of your results and how you achieved them (about 20 slides), presentation should include the parts
    - Data Exploration (RapidMiner)
    - Model Baseline & Data Preparation (RapidMiner)
    - Model Optimization (Python)
  - The according Rapidminer process(es) and Python/Jupyter notebooks

Case- Infrastructure

Student notebooks



Local files

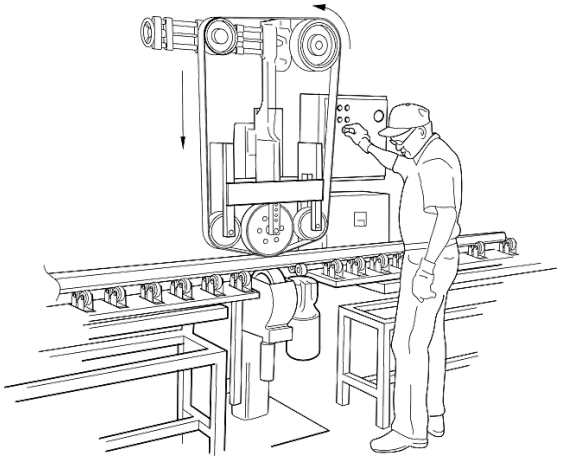
HOCHSCHULE DER MEDIEN

Picture Source: Pixabay

Dataset

- Structured data describing productivity data of different teams in a plant
- Method to predict label: Regression
- Format:
  - 1.511 rows × 21 columns
  - 20 features
  - 1 label: actual_productivity

Picture: pixabay



HOCHSCHULE DER MEDIEN

Dataset

#	Attribute	Description
1	date	Date in MM-DD-YYYY
2	day	Day of the Week
3	quarter	A portion of the month. A month was divided into four quarters
4	department	Associated department with the instance
5	team_no	Associated team number with the instance
6	no_of_workers	Number of workers in each team
7	no_of_style_change	Number of changes in the style of a particular product
8	targeted_productivity	Targeted productivity set by the Authority for each team for each day.
9	smv	Standard Minute Value, it is the allocated time for a task
10	wip	Work in progress. Includes the number of unfinished items for products
11	hsu	Hours of sunshine
12	over_time	Represents the amount of overtime by each team in minutes
13	incentive	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action.
14	idle_time	The amount of time when the production was interrupted due to several reasons
15	idle_men	The number of workers who were idle due to production interruption
16	work_cond_f1	Unspecified feature describing further working condition
17	work_cond_f2	Unspecified feature describing further working condition
18	work_cond_f3	Unspecified feature describing further working condition
19	work_cond_f4	Unspecified feature describing further working condition
20	work_cond_f5	Unspecified feature describing further working condition
21	actual_productivity	The actual % of productivity that was delivered by the workers. It ranges from 0-1.

Task 1: Data Exploration: RapidMiner (RM)

- Explore the provided data sets using descriptive statistics (e.g. mean values, standard deviations, min/max values, missing values) and visualizations (e.g. histograms, boxplots)
- Point out which data quality issues you identified in terms of
  - Missing values
  - Outliers
  - Features to be transformed (e.g. normalization) transformation
  - Features to be removed (feature selection)
- This task shall be performed using RapidMiner

Task 2: Model Baseline and Data Preparation (RM)

- Create a baseline linear regression model using the originally provided training dataset with minimal preprocessing and evaluate it with your test dataset based on accuracies (MAE)
- Preprocess the original datasets to address the identified data quality issues
  - Missing values
  - Outliers
  - Features to be transformed (e.g. normalization) transformation
  - Features to be removed (feature selection)
- Within the test dataset, you are not allowed to:
  - Remove examples
  - Change label values
- Create a new linear regression model using the preprocessed training dataset and evaluate it with your test dataset based on accuracies (MAE)
- Export the pre-processed training and test dataset to a CSV
- This task shall be performed using RapidMiner

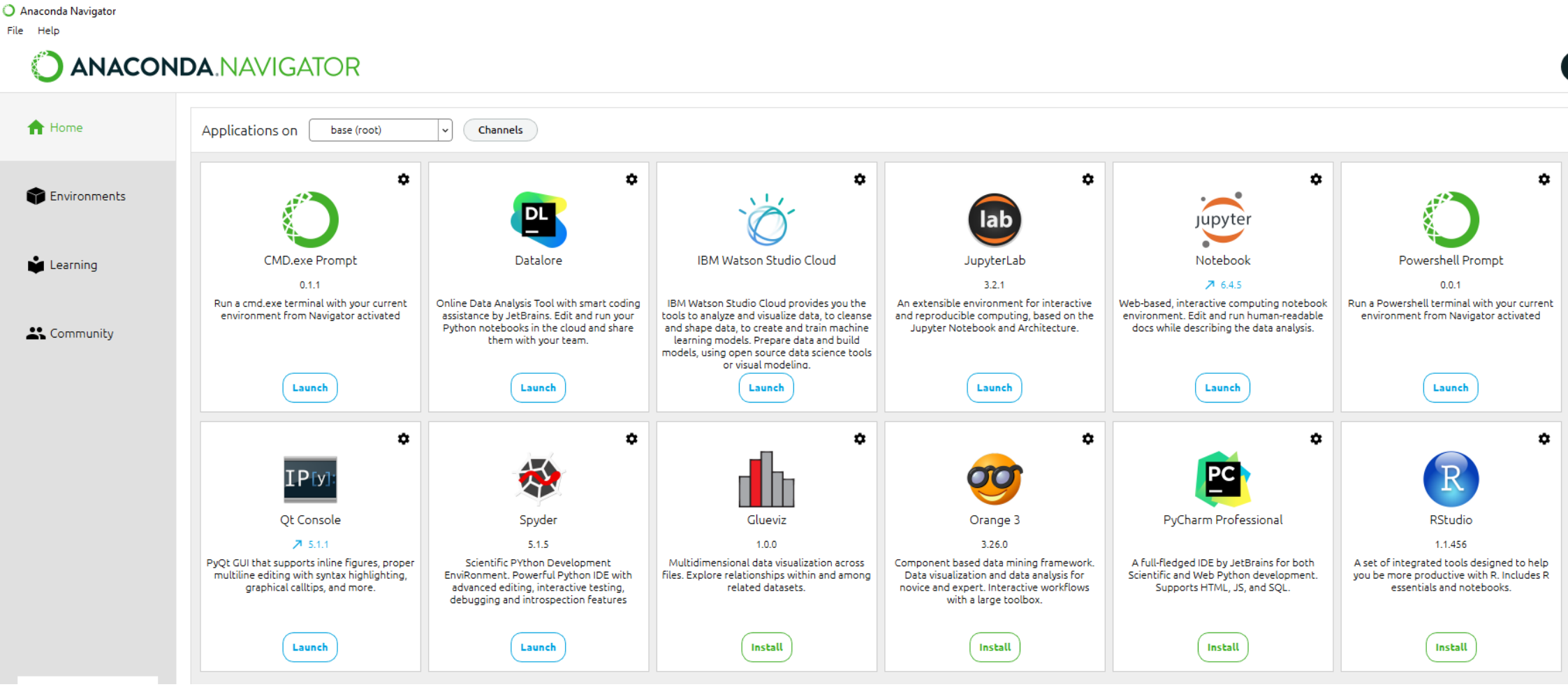
Task 3: Modeling Optimization (Python)

- Create a new baseline linear regression model using the preprocessed training dataset and evaluate it with your test dataset based on accuracies (MAE)
- Optimize your model / create further model versions
  - Algorithm Selection: Experiment with different regression algorithms, e.g. linear regression, polynomial regression, regression trees etc.
  - Hyper-parameter Tuning: Change the hyper-parameters of your algorithms (e.g. „degree“ in case of polynomial regression)
- Evaluate each model version based on accuracies (MAE) using the test data and store evaluation results in a data frame
- Create an overview of your evaluation data frame…
  - By generating an overview table
  - By generating an appropriate visualization

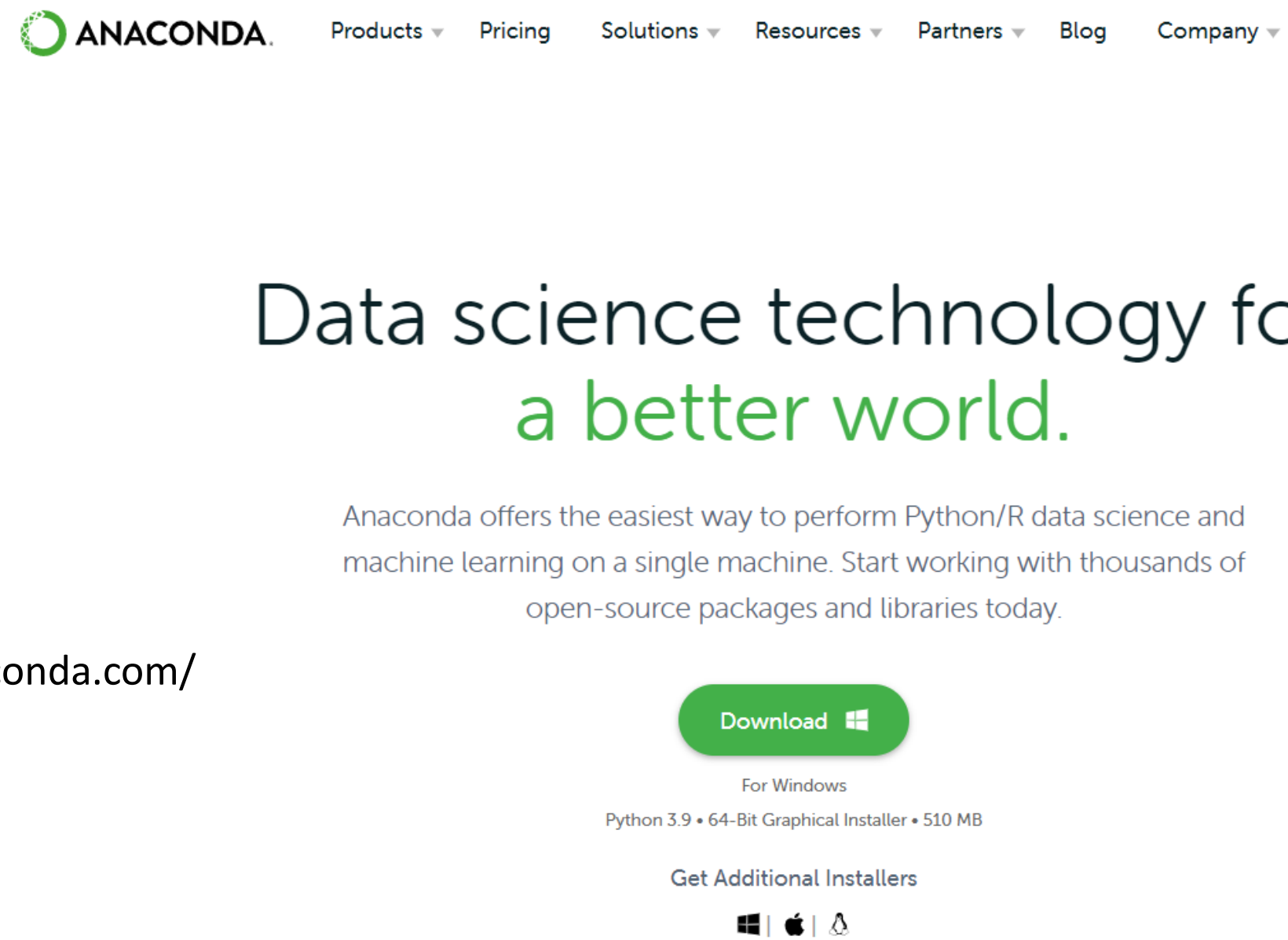
Questions?

Picture: pixabay

Anaconda Demo



Anaconda Installation



https://www.anaconda.com/

Warm Up Lab

jupyter PYCRASH-CHAPTER3 Last Checkpoint: 06.10.2021 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Markdown

Lists in Python

In [ ]:
```
#working with lists
#create lists and access list elements

#bikes = 'trek'
bikes = ['trek', 'canyon', 'cube']
print(bikes)
```

In [3]:
```
#access specific elements of list
print(bikes[0].title())
print(bikes[0])
```

```
Trek
trek
```

In [ ]:
```
# last element in list
print(bikes[-1].title())
```

In [ ]:
```
# concatentation and list values
print("My favourite bike brand is " + bikes[-1].title())
```

In [ ]:
```
# change list values
print(bikes)
bikes[1] = 'basso'
print(bikes)
```

In [ ]:
```
#adding list elements
bikes.append('giant')
print(bikes)
```

In [ ]:
```
#build list from scratch
```

jupyter PYCRASH-CHAPTER10 Last Checkpoint: letzten Montag um 11:18 Uhr

File Edit View Insert Cell Kernel Widgets Help

Run Markdown

File Handling in Python

In [1]:
```
#Reading from a File
#pi_digits.txt
#3.1415926535
#8979323846
#2643383279

#file_dir = 'C:\\tmp\\python_test_data\\'
#file_path= file_dir+file_name
file_path= 'pi_digits.txt'

with open(file_path) as file_object:
    contents = file_object.read()
    print(contents)
```

```
3.1415926535
8979323846
2643383279
```

In [2]:
```
#Reading Line by Line
filename = 'pi_digits.txt'

with open(filename) as file_object:
    for line in file_object:
        print(line)
```

```
3.1415926535
```

HOCHSCHULE DER MEDIEN