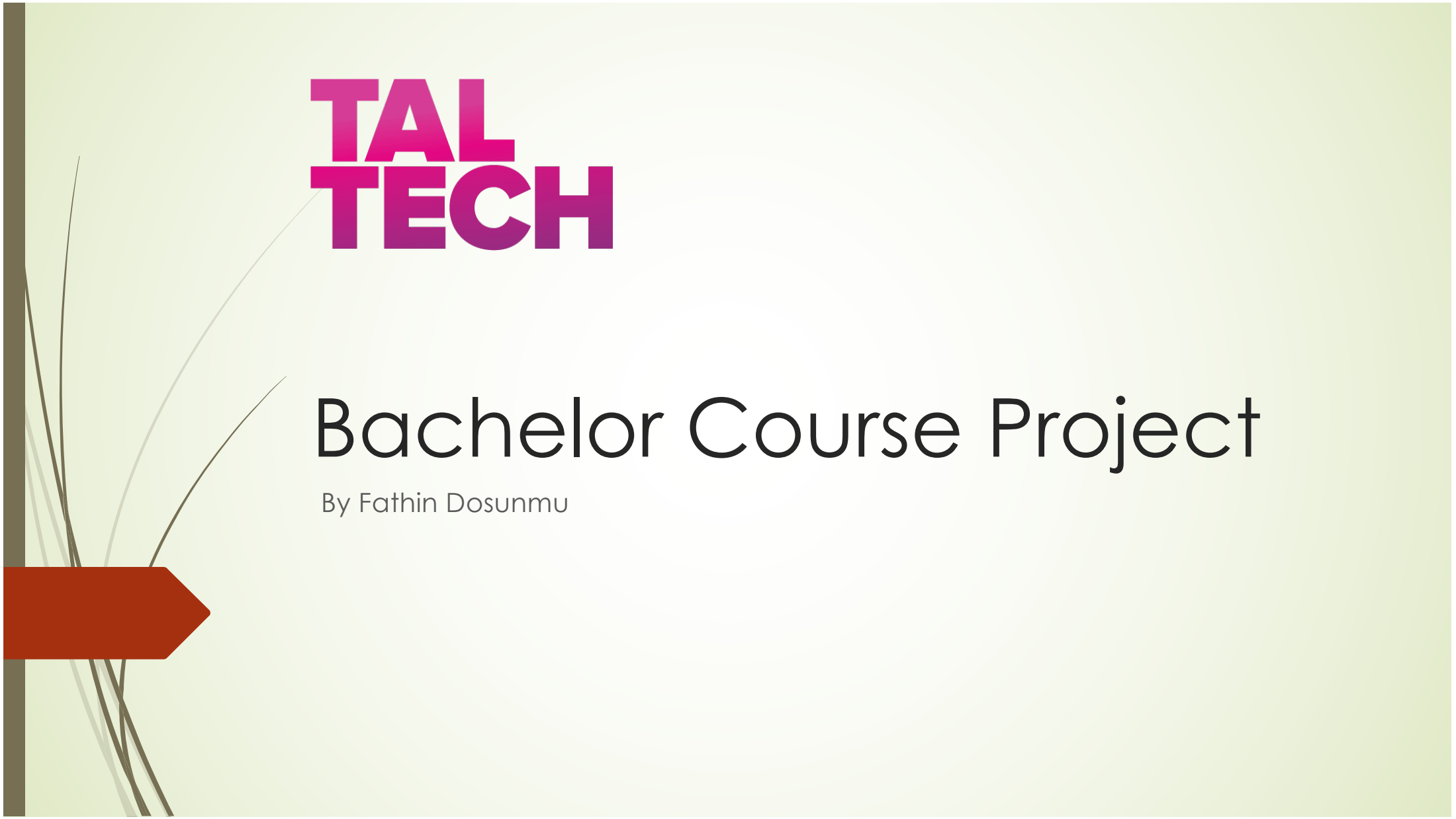


Bachelor Course Project

By Fathin Dosunmu

What is my Thesis about?

- **Goal 1**: The goal is to develop a machine learning model that would predict if a student is likely to continue studies at Tallinn University of Technology after the 3rd semester of studying based on statistics of university students collected from Tallinn University of Technology.

- **Goal 2:** Compare various machine learning algorithms, compare the pros and cons, and conclude on what algorithm is suitable to solve our problem.

- **Goal 3:** (Future Goal) **……**would be really nice to integrate my ML model with a software application and make it interactable for people. Perhaps sell the product and allow other universities to input their data and obtain their own predictions….

TAL TECH

Main Techincal and software tools used for Project

- **Python programming Language**: I choose this language because:
  - it has a very broad and strong deveoper community.
  - It is a flexible language and easily understandable by humans
  - Since **Python** is a general-purpose language, it can do a set of complex **machine learning** tasks and enables us to build prototypes quickly that allow us to test our product for **machine learning** purposes.
  - It has almost all machine learning algorithms implemented in libraries.
- **PyCharm**: a powerful iDE for python development
- **Anaconda**: Anaconda is the standard platform for Python data science, leading in open source innovation for machine learning.

TAL TECH

Structure of my project

- Since my projecat is mainly code and scriptBased, everthing is pretty much impelemented in python scrtipts.
- These scripts contain functions. My code is divided into functions that perform their own specidic tasks

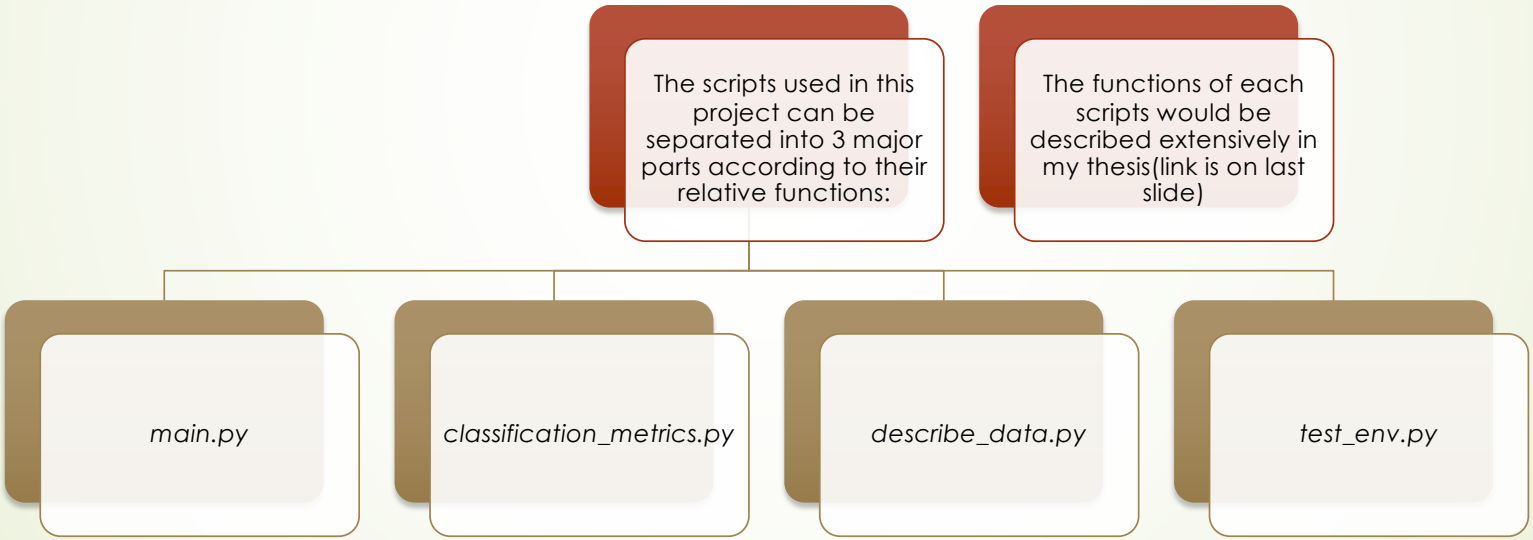
TAL TECH

hierarical structure of my project.

- As seen, these are the main files contained in my project. Further details about their respective functios are described in my thesis :

```
.
├── .gitignore
├── .gitlab-ci.yml
├── .pylintrc
├── common
|   ├── describe_data.py
|   └── test_env.py
├── data
|   └── students.xlsx
├── main.py
└── results
    └── .placeholder
```

TAL TECH

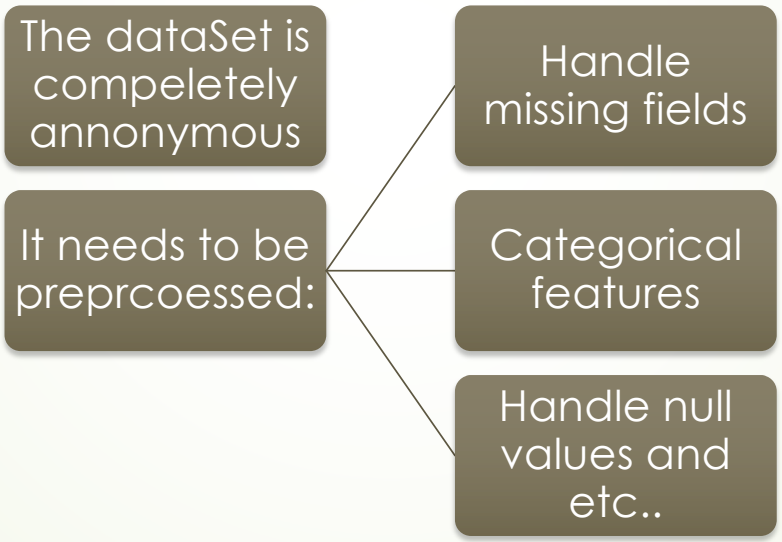

The scripts used in this project can be separated into 3 major parts according to their relative functions:
The functions of each scripts would be described extensively in my thesis(link is on last slide)
main.py
classification_metrics.py
describe_data.py
test_env.py

Studnets.xlsx File

- This is the dataSet I am working with for my project.
- A glimpse of part of the data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Faculty	Paid tuition	Study load	Previous school level	Previous school study language	Recognition	Estonian language exam points	Estonian as second language exam points	Mother tongue exam point	Narrow mathematics exam point	Wide mathematics exam point	Mathematics exam point	Study language	Foreign student	In university after 4 semesters	
2	School of Information Technologies	Yes	Full	General secondary education (310)	English								English	Yes	Yes	
3	School of Information Technologies	Yes	Full	General secondary education (310)	Estonian		63				67		Estonian	No	No	
4	School of Information Technologies	No	Full	General secondary education (310)	Estonian		97				90		Estonian	No	No	
5	School of Information Technologies	No	Full	General secondary education (310)	Estonian			67			67		Estonian	No	Yes	
6	School of Information Technologies	No	Full	General secondary education (310)	Estonian		58				59		Estonian	No	No	
7	School of Information Technologies	No	Full	General secondary education (310)	Estonian			66			68		Estonian	No	No	
8	School of Information Technologies	No	Full	General secondary education (310)	Estonian		86				56		Estonian	No	No	
9	School of Information Technologies	No	Full	General secondary education (310)	Estonian		58				58		Estonian	No	Yes	
10	School of Information Technologies	No	Full	General secondary education (310)	Estonian		65					58	Estonian	No	No	
11	School of Information Technologies	No	Full	General secondary education (310)	Estonian			89			80		Estonian	No	Yes	
12	School of Information Technologies	No	Full	General secondary education (310)	Estonian		55				67		Estonian	No	Yes	
13	School of Information Technologies	No	Full	General secondary education (310)	Estonian			76			82		Estonian	No	No	
14	School of Information Technologies	No	Full	General secondary education (310)	Estonian		49				62		Estonian	No	Yes	
15	School of Information Technologies	No	Full	General secondary education (310)	Russian			61				64	Estonian	No	No	
16	School of Information Technologies	No	Full	General secondary education (310)	Estonian		47				68		Estonian	No	Yes	
17	School of Information Technologies	No	Full	General secondary education (310)	Estonian		55				92		Estonian	No	No	
18	School of Information Technologies	No	Full	General secondary education (310)	Estonian		54				72		Estonian	No	Yes	
19	School of Information Technologies	No	Full	General secondary education (310)	Estonian		71				86		Estonian	No	Yes	

About the DataSet



TAL TECH

Main Algorithms to be used

- Logistic Regression
- K-Nearest neighborhood
- Support vector machines
- Naive Bayes
- Decision Tree
- Random Forest

- NB: Further explanations would be made in my thesis writeup about the algorithms.

Conclusions

- There is still a lot more to be discussed about my project. In terms of the logic, methods and workflow to be carried out.
- The points stated in this presentation are just the main pillar points for an introduction to the project.