Report - Project C16:

2018, 2019 & 2020 math exam results in Estonia

Brandon Loorits

Tauno Tamm

Business goals

Background

At the moment is very complicated times due to COVID-19 virus. We have to wear masks, keep distance, wash hands. It is not recommended to visit public places like theater, cinema, pubs, restaurants and as well as schools (as less as possible). Now in the autumn schools are opened longer and students visit lectures, labs and classes, but in the spring, when we had emergency situation in April to May schools were on the distance study. Students had to learn at the home and can communicate with teachers with Skype, Microsoft Teams or Zoom. Students had not study in classical ways they had to adopt with new situation and study more independently.

Business goals

Our goal is to find out is distance study influenced students performance in exams or not and if it had impact on results how it influenced results. If we found it did not had impact on results, then we have knowledge that it is not so important to avoid distance study like we do it in our country at the moment. The result of data analysis can give important information to government of Estonia how to act in these situations.

Business success criteria

Our main goal is to find out what is predicted results and compare them with actual results. Also it should answer the question is it important to save our health and go at distance study as soon as it seems necessary or keep schools opened as long as possible, beacause at distance study makes studing harder and makes results low. We can give information to government of Estonia about high school and basic school, how they should act in unusual situations like COVID-19 pandemic in 2020.

Assessing situation

Inventory of resources

We are going to use Anaconda3 Jupyter Notebook to make data analysis and predictions(kernel is python3)

We have public data from Eksamite infosüsteem, it is based on schools and participants(we do not have exactly every student ID but we have schools sum of participants and school mediaan and so on)

We use for predicting math advanced exam results, math simple exam results and Estonian language exam results from 2018 and 2019:

Estonian-2018.csv

Estonian-2019.csv

Math-advanced-2018.csv

Math-advanced-2019.csv

Math-simple-2018.csv

Math-simple-2019.csv

And for comparing predicted results we use math advanced exam results, math simple exam results and Estonian language exam results from 2020:

Estonian-2020.csv

Math-advanced-2020.csv

Math-simple-2020.csv

Requirements, assumptions, and constraints

We have constraints that we do not know every students result, we have only schools results. And their results average,mediaan and standard deviation.

Risks and contingencies

Our risk is that we do not know when some school went on distance study or finished it and how they organized it. We could have wrong results due to participants number is in some school 60 and in another school 6 so the average could vary more.

Terminology

Keskmine – School average result

Mediaan – School results mediaan

Sugu – sex(M-male/F-female)

Kool – school, where exams have performed

Kooli tüüp – is it basic, high or trade school

Aasta – year

Õppekeel – studing language

Sooritus keel – exam language

Õppevorm – studing form, for example statsionary or not-stationary

Sooritajaid –participants

Min

Max

Costs and benefits

We can not estimate costs and benefiits in euros or dollars. We could gather information how to act in these situations and how different studing impact exam results. We could have benefiits in saving our health and häving better scores in exams.

Defining your data-mining goals

Data-mining goals

We are going to use diferent predicting models for example KNN, random forest, basic lineaar classifier and so on. Definetely we use some visualisatsions and diagrams to show which schools had better results and which worse. Which schools results changed more and where less.

Data-mining success criteria

Every team māte have to work about 30 hours with project to achieve a goal. We need to pay attention on visualisatsion to make it easy to understand for ohter people. We want to show how results have changed and so we have to find best classifier t omake predictions, which we compare to actual values.

Gathering data

We got our data for the Project from EIS (https://eis.ekk.edu.ee/eis), which is public source where is located Estonian Basic schools and high schools examination results. Examination infosystem has a searchengine, which gives you data about each year of your chosen exam. It can be downloaded directly to computer in csv file format.

Data is in .csv files and Jupyter Notebook is working well with .csv file and files are easily to read in.

Verify data availability

Data is public in Eksamite Infosüsteem and we have pre-downloaded it and pushed it to our repository.

Data files names:

Estonian-2018.csv

Estonian-2019.csv

Math-advanced-2018.csv

Math-advanced-2019.csv

Math-simple-2018.csv

Math-simple-2019.csv

Estonian-2020.csv

Math-advanced-2020.csv

Math-simple-2020.csv

All these files are added to our repository manually to avoid possbile mistakes made with web scraping.

Define selection criteria

We use Estonia national Examination information system, which gather every year results about exams. 2018 and 2019 results is for making test model and making predictions and 2020 results is for comparing predictions with acutal values. Also we can predict the examination results for the next year with the influence of the COVID-19 pandemic and predict the examination result if the pandemic should stop spreading.

Describing data

Our data is originated in Eksamite infosüsteem - EIS (Examination infosystem - https://eis.ekk.edu.ee/eis/). It is governments lawful information system, where shown data is reliable and correct. Our Project is examines Estonian math examination results over 2018, 2019 and 2020. Every examination year has a separate csv file which is downloaded from EIS manually. In csv file, there is described 22 different variables – for example year when the exam was taken, Test ID, Sex, school type, school's language, taken examination language, province, score, max score, min score, etc.

Exploring data

Because of the data variables variety, it's possible to explore data from different angles. We can have a great overview where the best exams were taken, where the worst exams were taken, predict how the exam result would have been without COVID-19 pandemic. This wide overview is quite good information for the government to take measures to improve Estonian education system. We are using all of the data from the csv files to make analysis as comprehensive as possible. As we can see

Verifying data quality

The data is originated in Estonian examination infosystem infosüsteem - EIS (Examination infosystem - https://eis.ekk.edu.ee/eis/), which is administrated by the governments authorities. It means the data is already strictly controlled and we can assume it is correct.

Task	Deadline	Assignee	Effort
Gathering data in one file	29.Nov	Brandon	2h
Making data correct (delete unneccesary attributes/make new attributes)	30.Nov	Brandon	2h
Control is data correct	1.Dec	Tauno	1h
Gathering and comparing classifier and choose best ones	4.Dec	Tauno/Brandon	15h
Working with classifiers and aslo make predictions (report AUC)	10.Dec	Tauno/Brandon	50h
Verify done work, check is everything working correctly	11.Dec	Brandon	5h
Making visaulisations	14.Dec	Tauno	10h
Improving visualisations with comments etc.	15.Dec	Brandon	5h
Making presentation and presntation talk	16.Dec	Tauno	5h
Present work	17.Dec	Brandon/Tauno	1h