UiO : University of Oslo

Supporting Future Educational Data Miners through a Summer Research Internship

John M. Aiken[1], Gabriel Sigurd Cabrera[1], Lucas G. G. Charpentier[1], Rachel Henderson[2], Nils Johannes Mikkelsen[1],
Matthew Ring[1], Fu-Anne Wang[1], Zhen Xu[1], Nicholas Young[2], Linrui Zhang[1], Marcos (Danny) Caballero[1,2]
[1] Center for Computing in Science Education, Department of Physics, University of Oslo, Oslo, Norway
[2] Department of Physics and Astronomy, Michigan State University, East Lansing, MI

MICHIGAN STATE UNIVERSITY

A Summer Internship in Data Mining

Educational data mining is an emerging field that leverages large-scale data collected by university registration systems, online learning portals, and other digital data collection systems. Text data and modern machine learning methods are used to gain insight into undergraduate student learning, participation, and other outcomes. These data can be connected to traditional PER data such as surveys and concept inventories and be used to produce new insights in the PER space. In 2019, the University of Oslo (UiO) and Michigan State University offered a nine week summer internship program for bachelor's and master's students. Eight students attended from China, Norway, and the United States. Students learned how to use SQL databases and python to clean and parse dirty data, visualize that data, and build machine learning models to predict outcomes. This poster presents an outline of the summer program and the projects the students pursued.

Summer Program Design

The program is 9 weeks long and includes:

- Intensive 1-week workshop in the beginning on machine learning and extracting data from databases
- Research project designed to investigate a specific research question guided by seasoned PER researchers
- Weekly lectures from academic scientists, industry data scientists, and career specialists
- Cultural exchange
- Access to state of the art computing resources at UiO
- Stipend, travel, and housing allowances

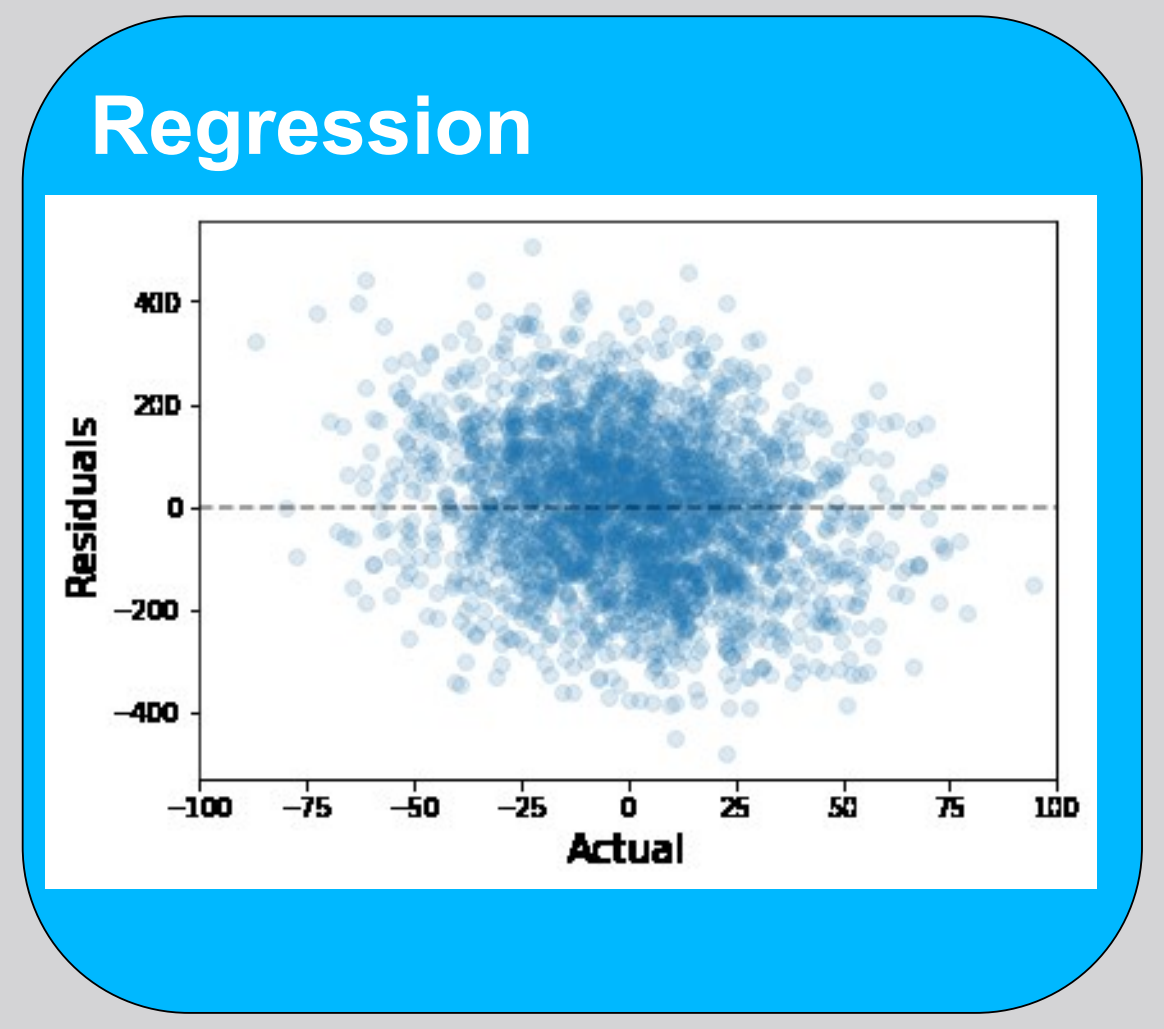
The crew: (Top) Nils Johannes Mikkelsen, John M. Aiken (*facilitator*), Matthew Ring, Danny Caballero (*facilitator*); (Middle) Robert Solli (*facilitator*), Gabriel Sigurd Cabrera, Alyssa Waterson, Rachel Henderson (*facilitator*), Fu-Anne Wang; (Bottom) Lucas G. G. Charpentier, Joseph Wilson, Xu Zhen (Jenn), Zhang Linrui (Rachel).

Intensive 1-week workshop in machine learning techniques

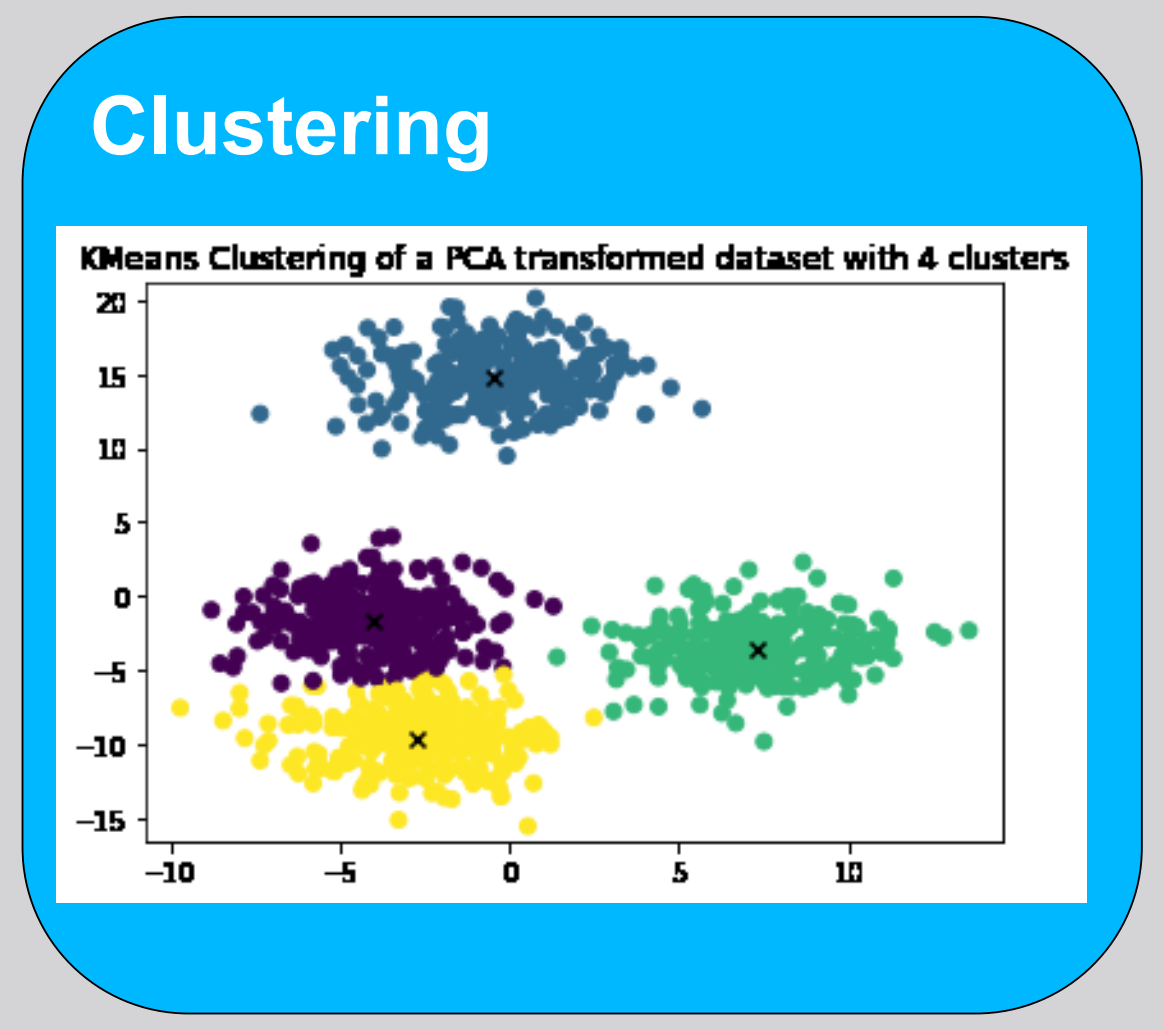
- Introduces students to machine learning concepts such as supervised and unsupervised learning, and model evaluation.
- Students gain experience with the programming languages Python and SQL.
- Students learn software development tools such as jupyter notebooks, IDEs, and git/github.
- The students learn to query databases and create data pipelines with SQL and python.
- The workshop materials are designed to be both an introduction to machine learning and a framework for education research projects .

workshop github repository

Regression



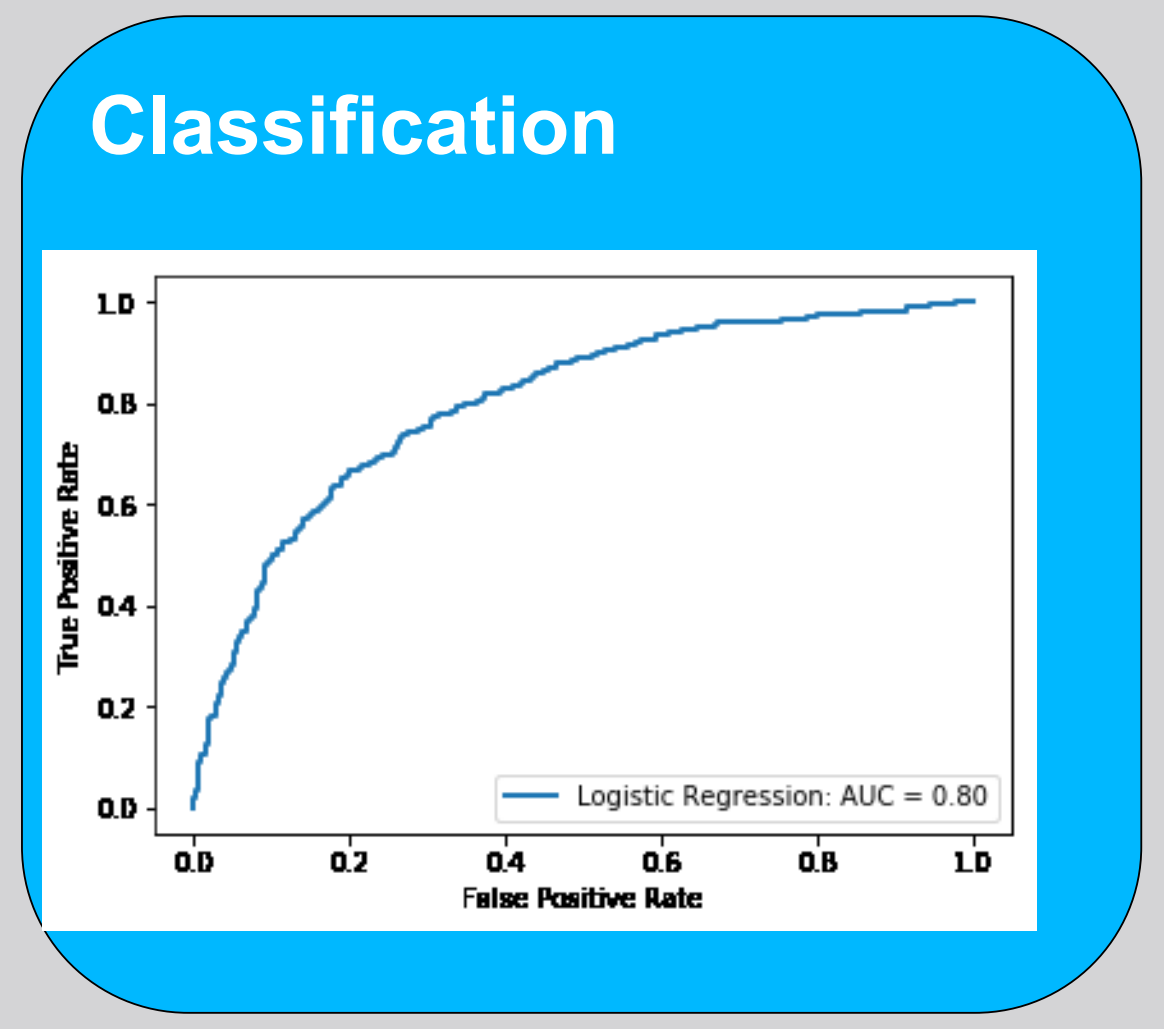
Clustering



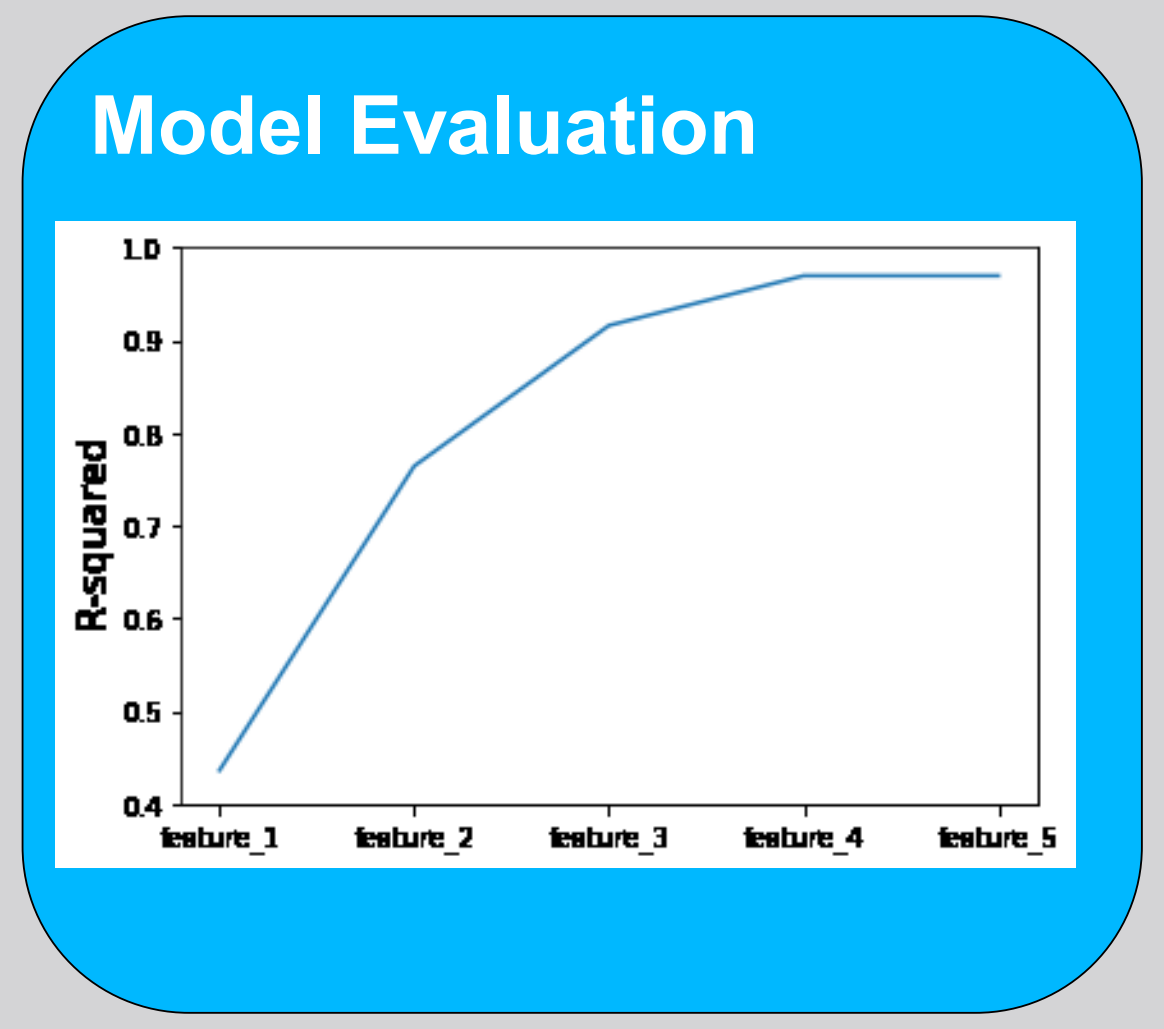
Research and development tools

jupyter

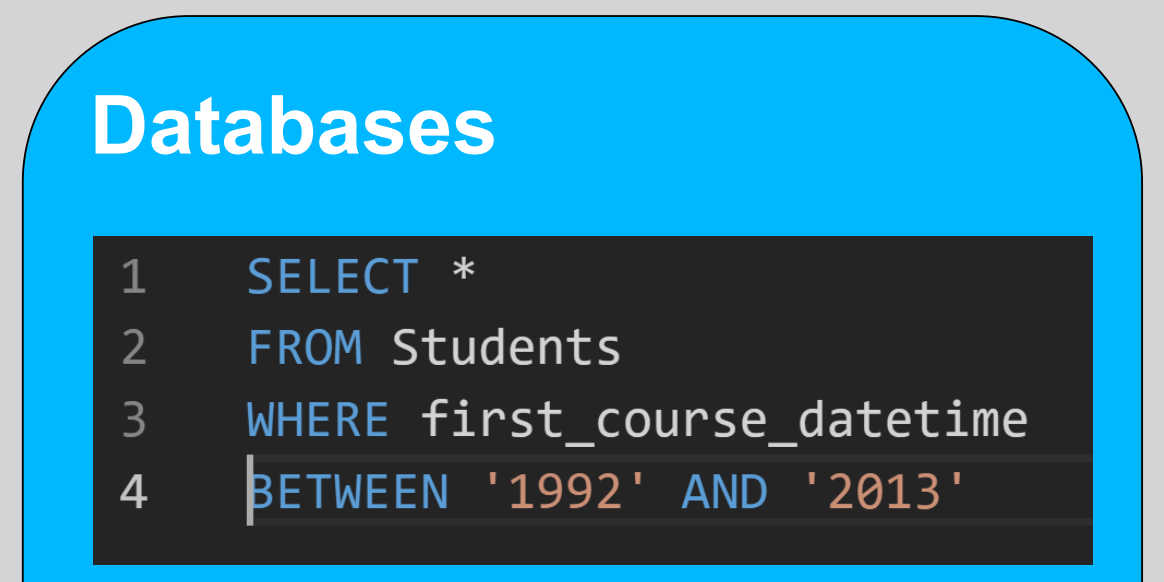
Classification



Model Evaluation



Databases

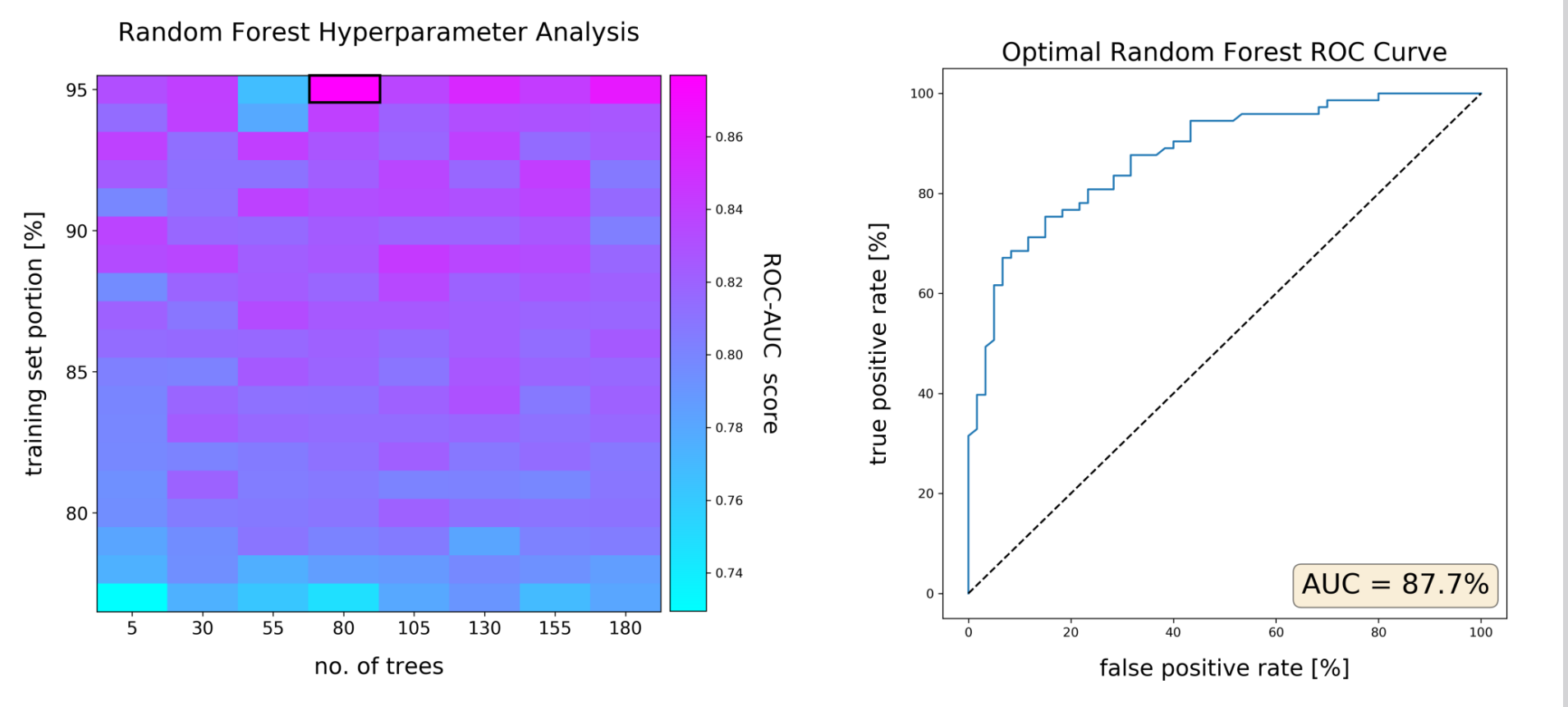


```
SELECT *
FROM Students
WHERE first_course_datetime
BETWEEN '1992' AND '2013'
```

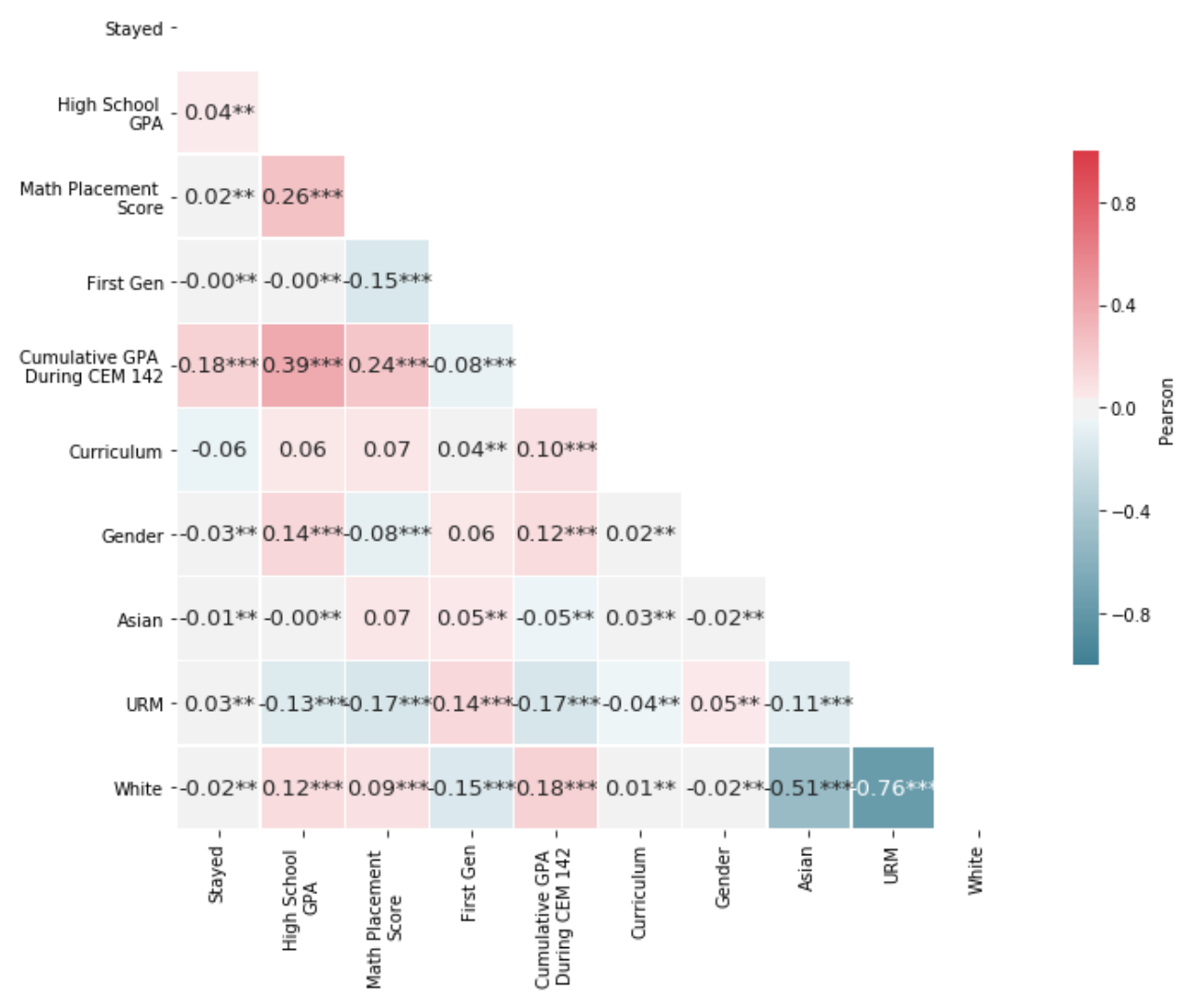
Microsoft SQL Server

Projects

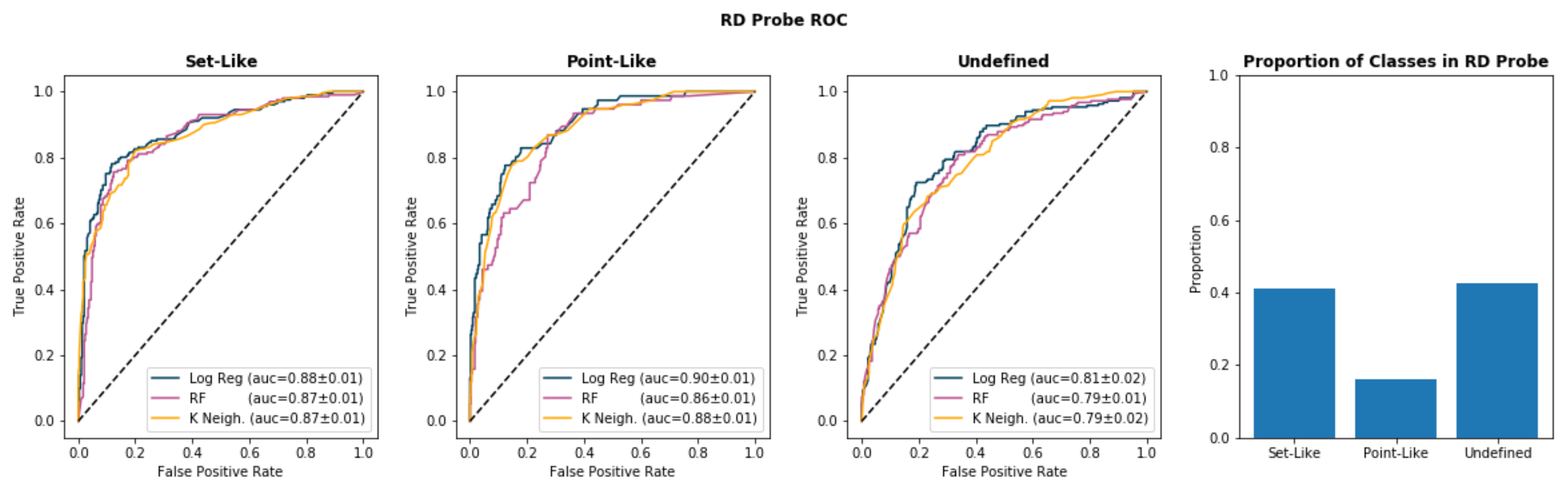
The projects concerned several relevant topics to PER such as network analysis, natural language processing, and students writing codes for numerical simulations. In each project the students start with a well defined PER research question and attempt to answer it via machine learning skills they learned from the workshop.



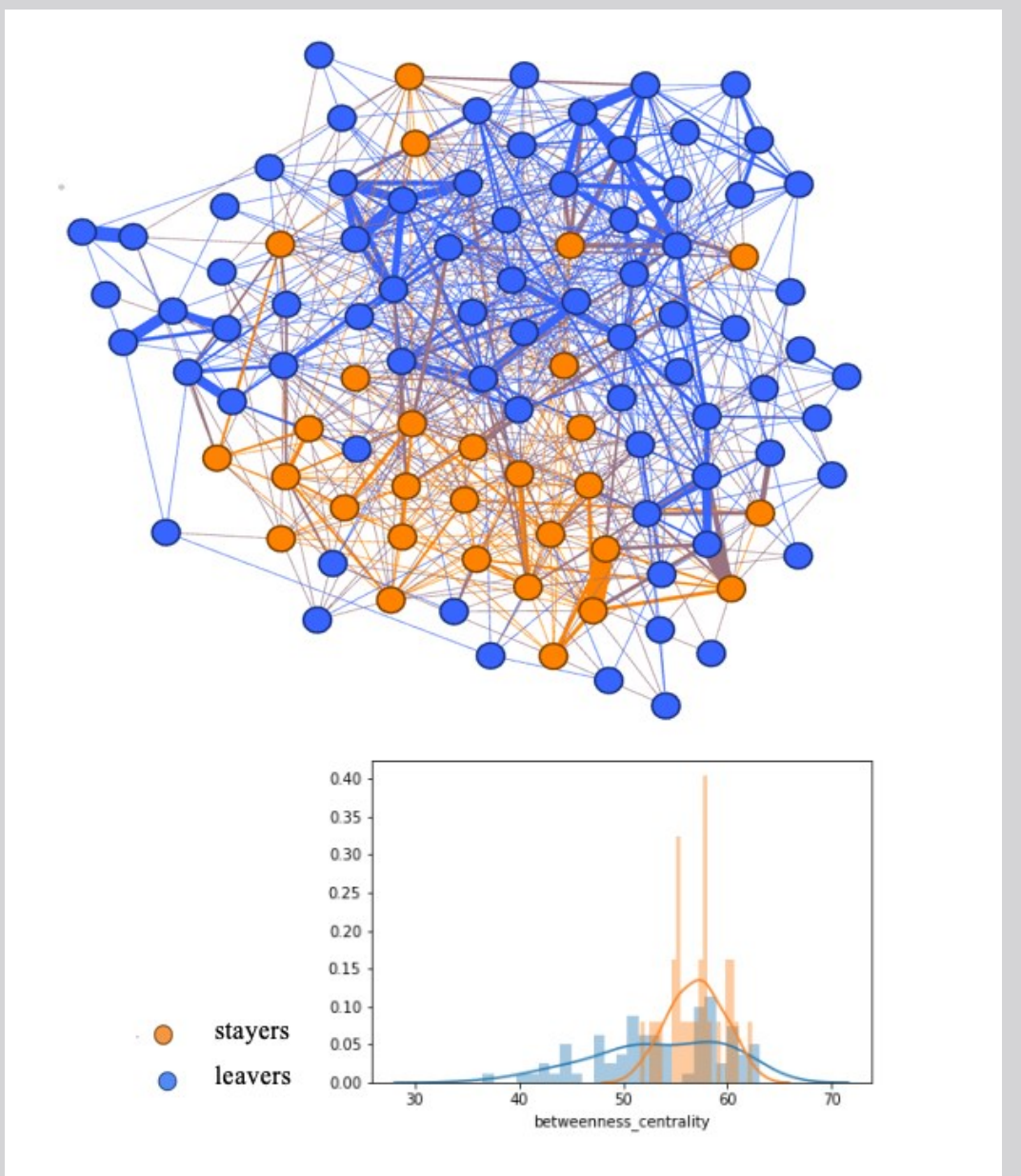
Understanding how prior preparation influences physics GRE scores (Nils Johannes Mikkelsen): Hyperparameter analysis of random forest classifiers predicting above/below average PGRE scores (N=2649). Model features include institutional-level features like Barron's category, and student-level features like undergraduate GPA, gender, ethnicity.



Persistence in STEM degrees due to curricular changes (Matthew Ring): Data from Michigan State University's pathways dataset is used in this study to assess student persistence. The data includes demographic and performance information for around 1,200 students after preprocessing of the raw data.

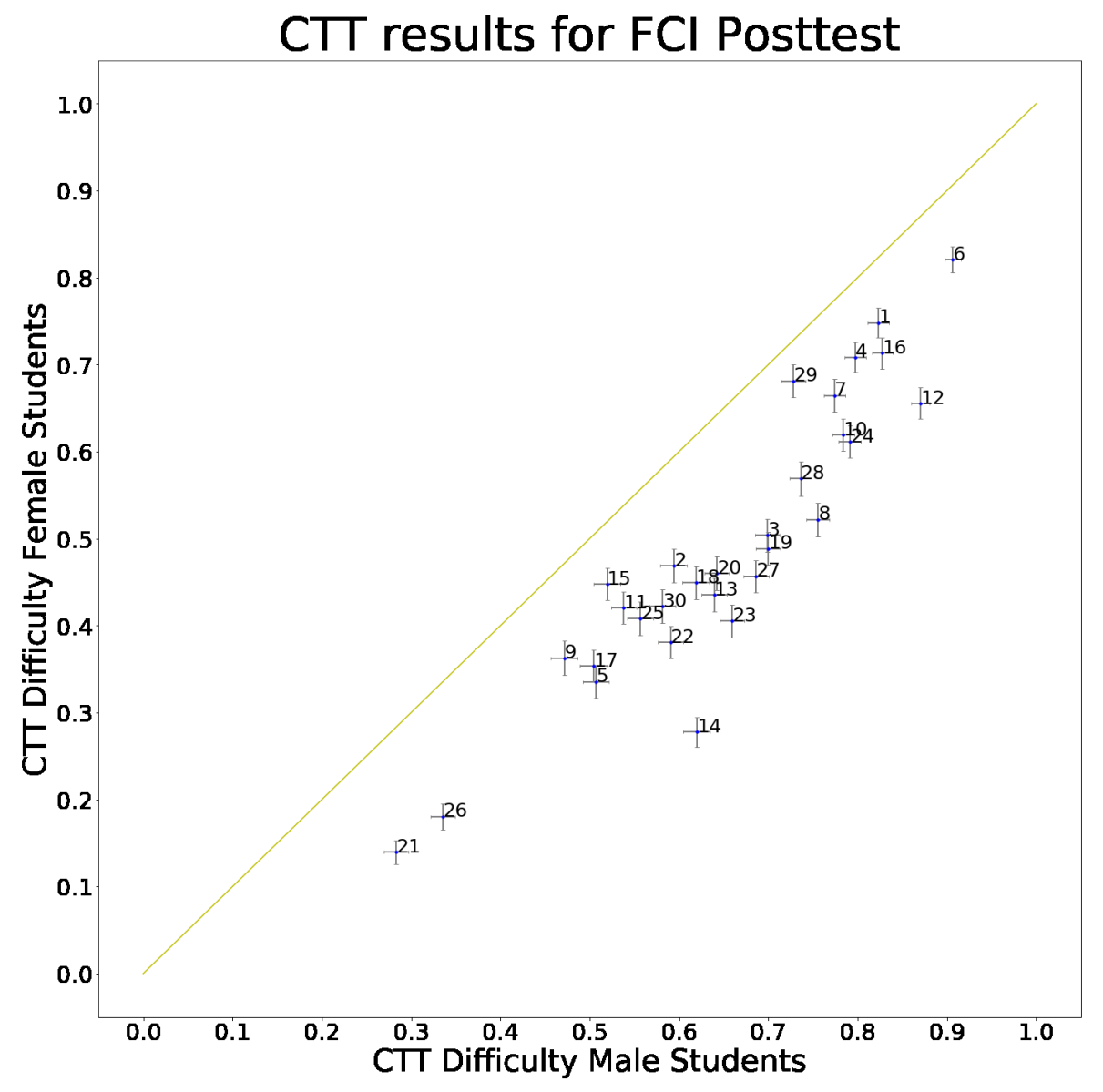


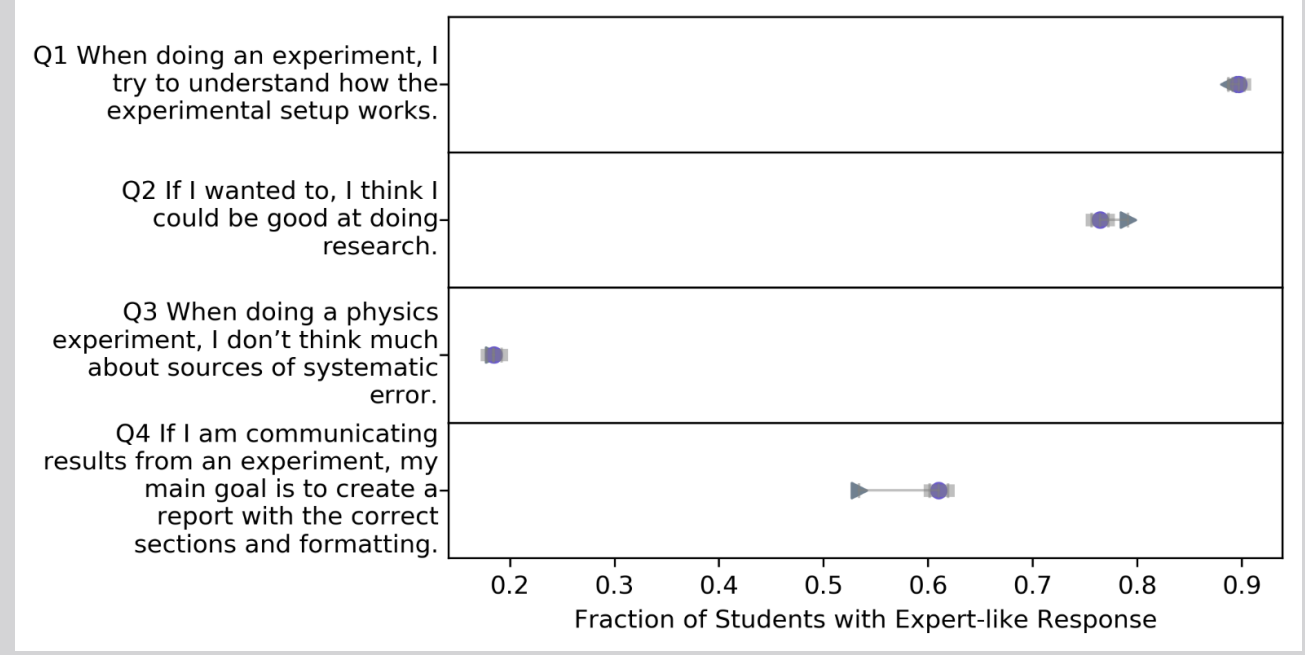
Natural Language Processing of the Physics Measurement Questionnaire (Joseph Wilson): ROC curves demonstrating the success of different machine learning models assessing the content of student essays. The essays are taken from the 'Repeating Distance' probe on the PMQ. This item asks students if they should repeat an experiment and subsequent measurement once, many times, or not at all.



Cohort effects on physics major attrition (Xu Zhen): Students who graduate with physics bachelors are much more likely to be connected with other physics students in courses other than physics in all the years than those who leave the physics department. Most students who leave the physics department do so for other science and engineering programs with similar math requirements, thus students who leave the physics department do so for other departments that share general electives.

Gender biased items on the FCI measured within the LASSO data set (Zhang Linrui): This study employed data collected at two US universities enrolling 3,387 students. The sample was collected 2015-2017. After screening, the data set contains 1,785 complete post-test responses (36.9% female). A diagonal line represents perfectly fair questions. From this figure: (1) almost all items are biased against females (via CTT difficulty); (2) items 14, 22, 23, and 27, stand out as substantially unfair to women which are almost matched with a large amount of previous research result.





Exploring the E-CLASS using Item Response Theory (Fu-Anne Wang): Fraction of students with expert-like responses for E-CLASS items 1-4. Circles indicate the pre-instruction fraction while the arrow indicates the post-instruction fraction and points in the direction of the shift from pre- to post-instruction. Shaded bars indicate the 95% confidence interval on the pre-instruction fraction. IRT models will be created for each construct in the E-CLASS survey.

This study was funded by the Olav Thon Foundation, the Norwegian Agency for Quality Assurance in Education (NOKUT) that supports the Centre for Computing in Science Education, and the Research Council of Norway, INTPART ES634020.