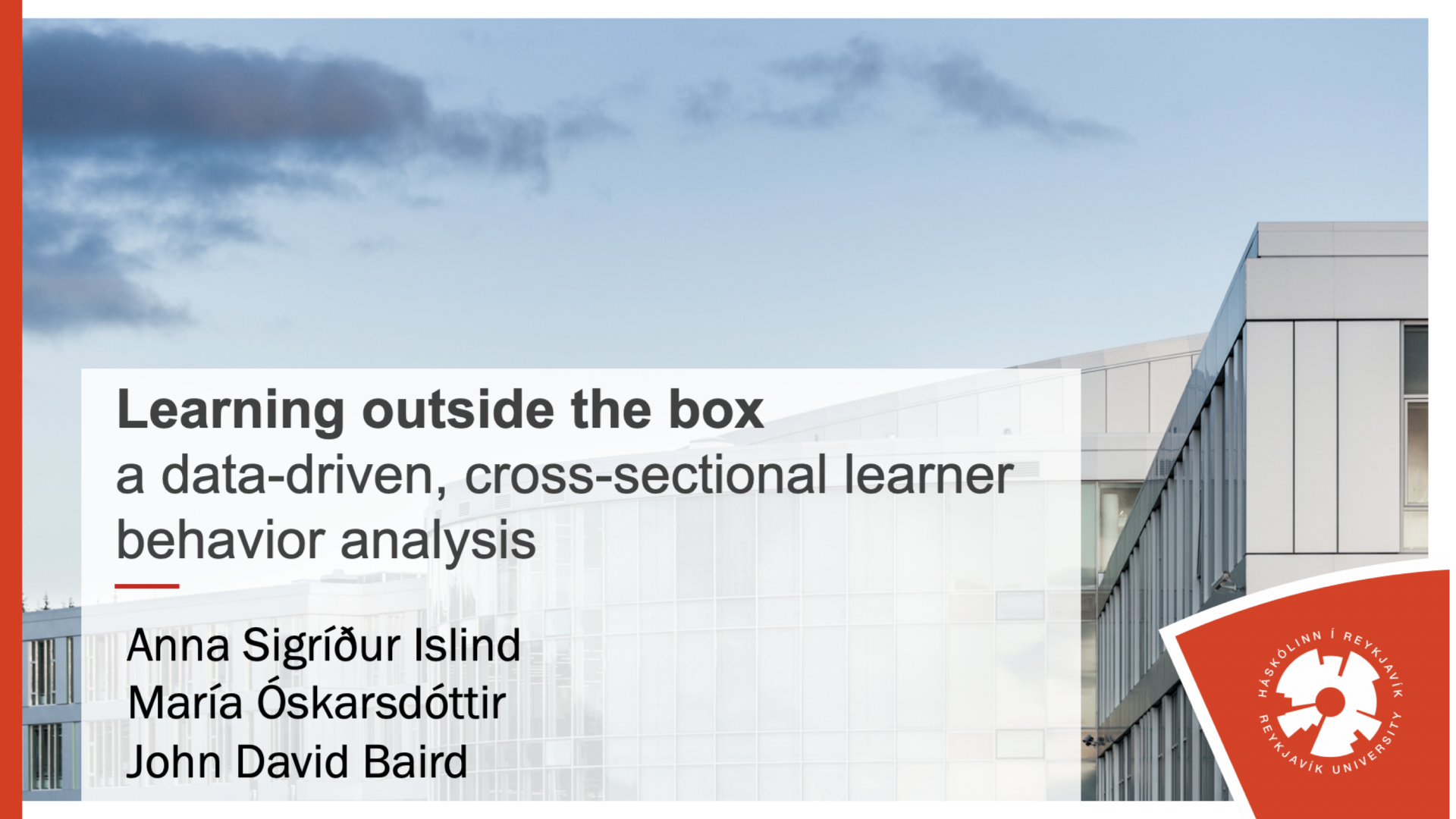


Learning outside the box

a data-driven, cross-sectional learner behavior analysis

Anna Sigríður Islind
María Óskarsdóttir
John David Baird

HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

Introduction

- Learning analytics is a rapidly growing research field
- On the border between learning and analytics
- Learning analytics should be embedded in education sciences
- Student data is rich and highly usable to understand learning behavior
- Should consider ‘third variable type factors’
    - First language, interest in course of study, time of delivery, educational maturity

Introduction

- Cross sectional study across and within 3 years of undergraduate computer science degree at Reykjavík University
- Mixed methods, data driven approach
- Machine learning techniques applied to student data
- Use qualitative data to augment quantitative data
    - Interviews, surveys, LMS click logs

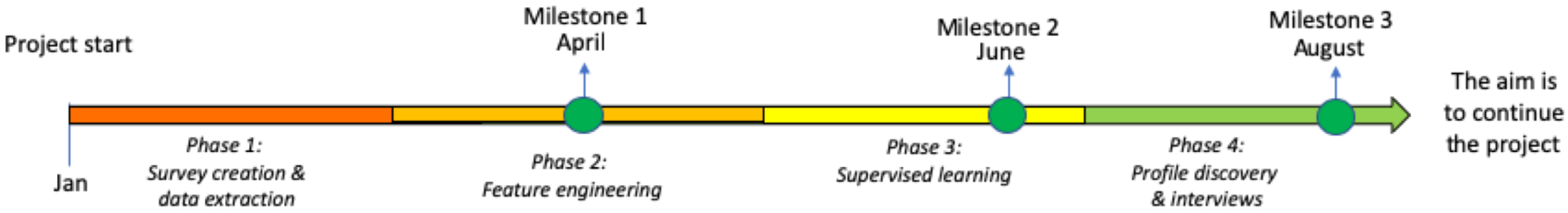
Facilitates *out the box thinking* for learners and teachers

Goals

- Use data from the learning management system to discover learning behavior profiles
- Identify variables that correlated with performance outcome
- Identify how lecture capture can better support learning
- Enhance understanding through semi-structured interviews

Project timeline

- Four phases
- Three milestones
  a. April 2020: Progress report
  b. June 2020: Presentation slides
  c. August 2020: Conference presentation



Phase 1: Survey Creation and Data Extraction

- Survey
  - Based on Yani et al.
  - Focus on non-subject based educational maturity

- Data Extraction
  - Click logs from the LMS
  - 6 courses in computer science

Course name	Study semester	Number of students
Discrete Mathematics I	1	171
Calculus and Statistics	3	156
Databases	2	215
Data structures	2	191
Software requirements & design	1	237
Computer Networks	5	181

These features afford

- Heterogeneous student population
- Scope to generate an extensive and varied dataset

Phase 2: Feature engineering

- Student interaction with LMS is stored in logs
- Unstructured data - clickstreams
- Define numerous variables from the logs
- Variables are used in Phases 3 and 4

Phase 3: Supervised learning

- Supervised learning approach to predict academic performance
- Final grade is the target variable
- Use both interpretable and black box models
- Measure model performance in accuracy
- Investigate which variables are most predictive of the target

Phase 4: Unsupervised learning - Profile discovery

- Unsupervised learning approach to discover student learning profiles
- Partition the data into clusters using k-means
  - Observations in the same cluster are similar
  - Observations in different clusters are dissimilar
- Analysis of clusters gives student profiles

- Semi-structured interviews with teachers to gain in depth understanding of results

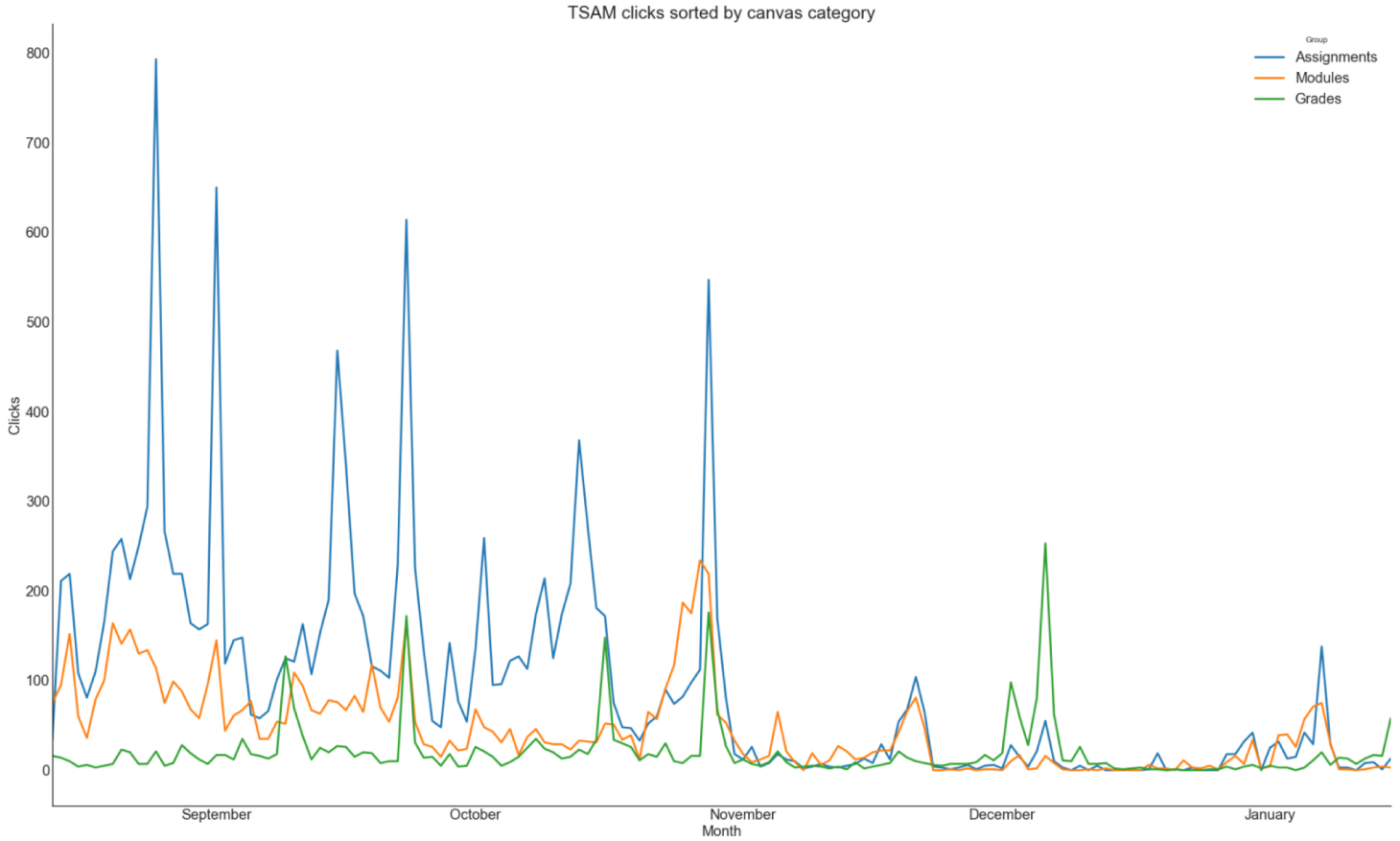
Results

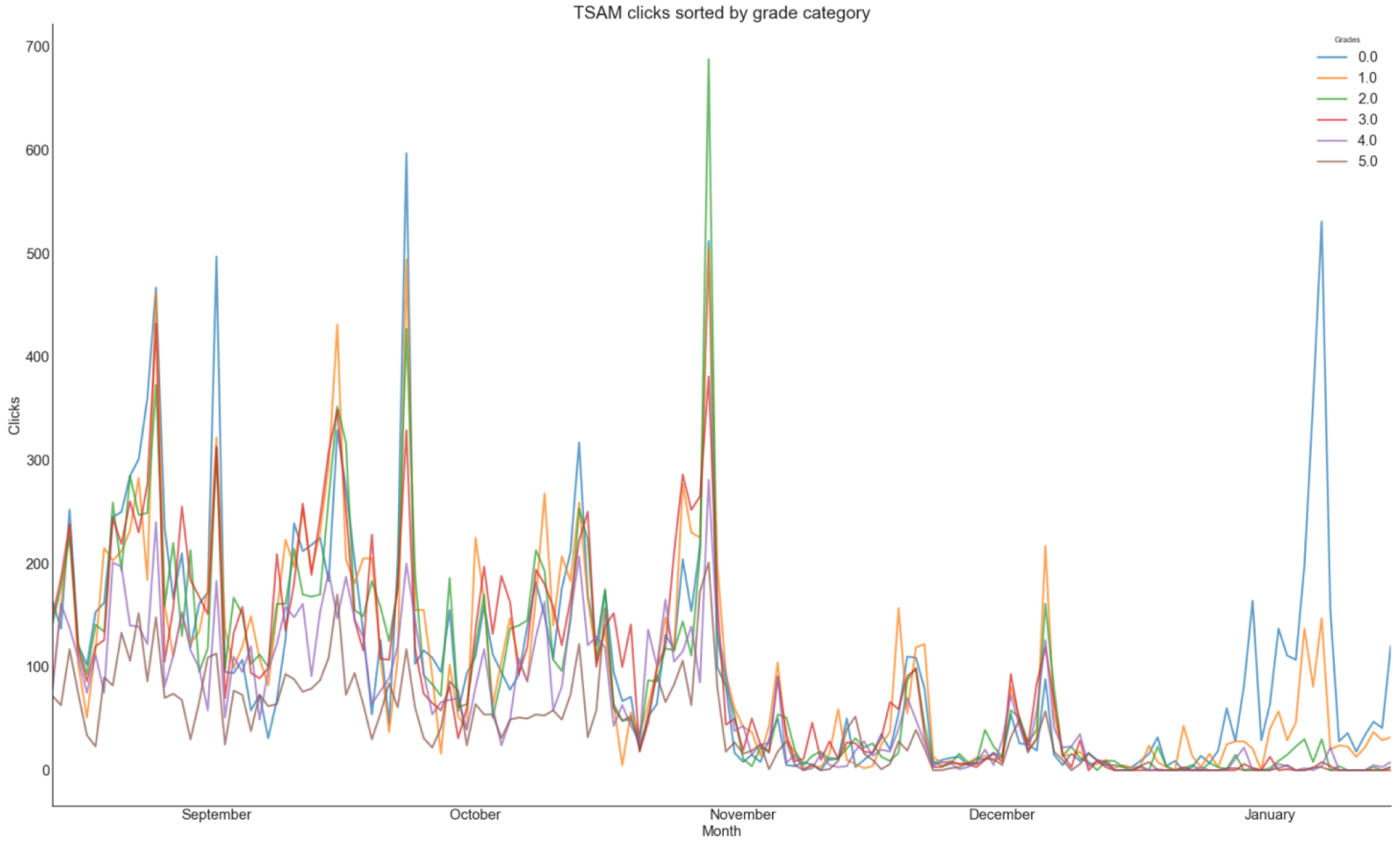
Survey

- 42 questions
  - Institutional Support, Educational Values, Goals, Self- Efficacy, Academic Apathy, University Related, Social Network
- Sent out to 678 individuals, 273 students responded 178/92 (M/F)

Main results

- 93.41% of students show high self-esteem, belief in themselves and their capabilities
- Just under 50% of participants attend live lectures, around 50% watch online
- High correlation between methodical individuals and those who organize their study times to best achieve their goals


TSAM clicks sorted by canvas category
Group
Assignments
Modules
Grades
Clicks
800
700
600
500
400
300
200
100
0
September
October
November
December
January
Month


TSAM clicks sorted by grade category
Grades
0.0
1.0
2.0
3.0
4.0
5.0
Clicks
700
600
500
400
300
200
100
0
September
October
November
December
January
Month

Feature engineering

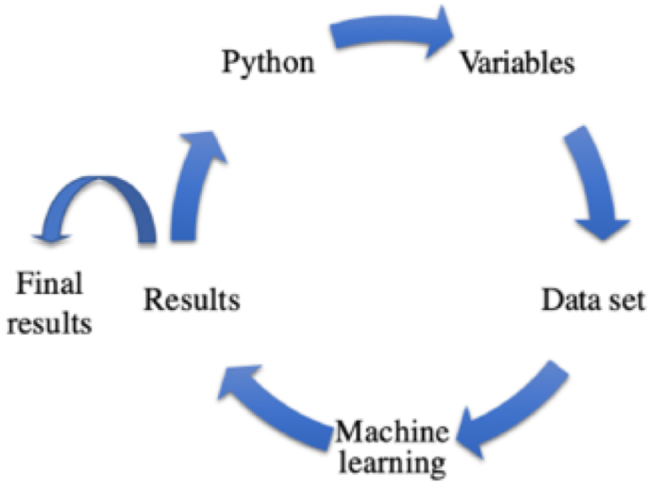
- Convert LMS click logs to variables that describe usage and engagement
- Clicks on LMS components
  - Modules, asigments, grades
- Clicks by week
  - Week 1, week 2, ...
- Click by part of week
  - Work days vs. weekend
- Click by part of day
  - Moring, afternoon, evening, night
- Clicks by location
  - At school vs. not at school
- Click statistics
  - Average, min, max, std by week

Student features

Student ID	Date count	Total clicks	Assignment clicks	Average clicks per week	Weekend use ratio	Night use ratio	On campus use ratio	Week 1 clicks	Grade category
1	100	2467	222	45	0.23	0.05	0.89	100	A
2	47	5553	875	23	0.67	0.3	0.32	503	D
3	28	793	353	67	0.88	0.54	0.1	200	C
4	84	555	375	47	0.53	0.01	0.67	10	B
...	...	...	...	...	...	...	...	...	...

Supervised machine learning

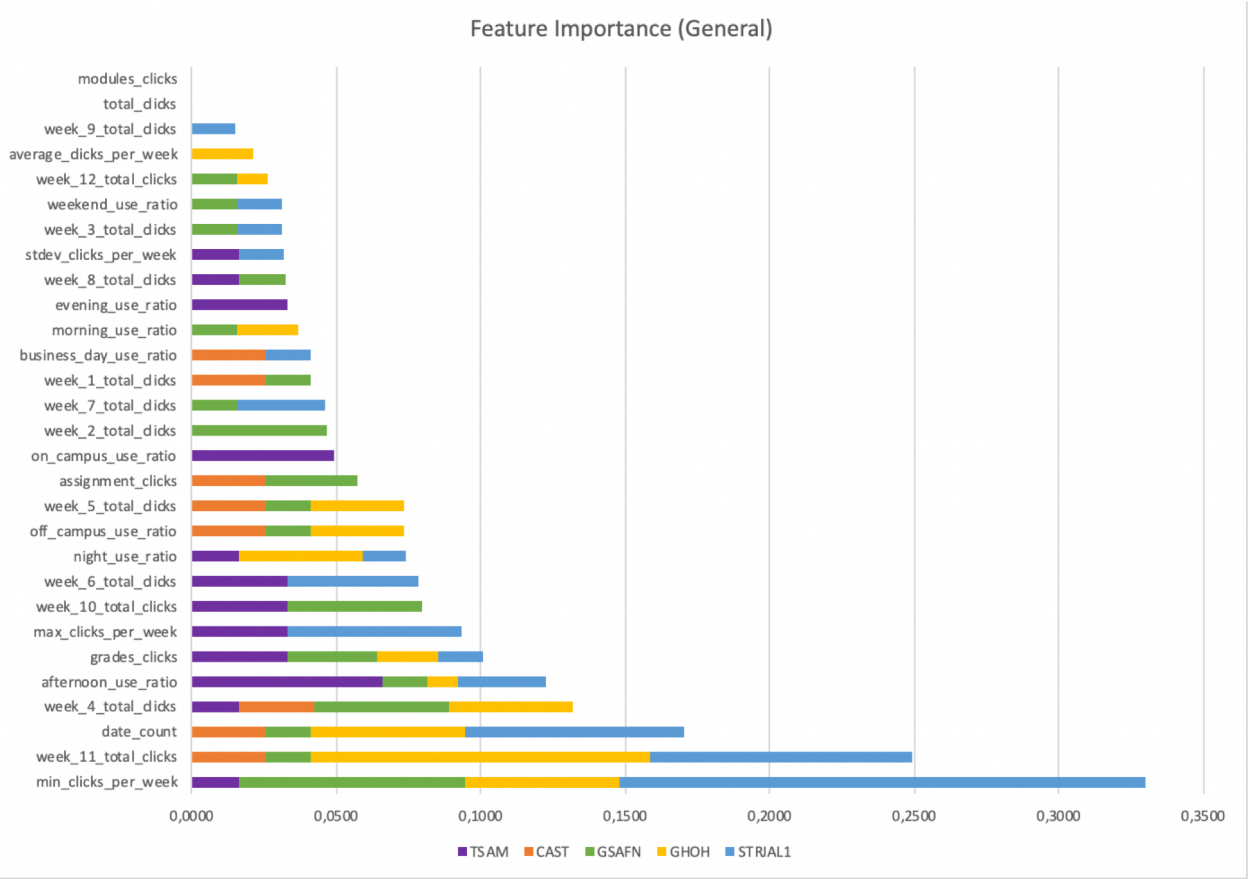
- Target variable is divided into four groups: A,B,C,D
  - Multiclass classification model
- Build model for each course
  - Decision jungle
  - Decision forest



Model accuracy

Course	Multiclass Decision Jungle		Multiclass Decision Forest	
	Overall Accuracy	Average Accuracy	Overall Accuracy	Average Accuracy
Discrete Mathematics I	0.485	0.742	0.394	0.697
Software Requirements and Design	0.523	0.761	0.521	0.761
Calculus and Statistics	0.333	0.667	0.344	0.671
Computer Networks	0.459	0.729	0.377	0.689
Databases	0.453	0.727	0.484	0.742

Feature importance



Ripple effects from Bluenotes faculty research grant

- Proof of concept
- Reward
- Student salaries
- Started with Computer Science and now we have the whole university on board
- Hired a PhD student
- Applied for a national research grant to continue project
- Would like to do the same in other universities (seeking collaborations with universities that use Canvas)

Conclusion

- Survey results give understanding of students' academic maturity
- The clickstream logs clearly show student engagement with study material
- Students that are engaged throughout the semester perform best

Next steps

- Profile discovery
- Longitudinal study
- Present results to the teachers, and get their feedback

Thank you!

And thank you BNG2020 for the opportunity!

Anna Sigríður Islind
islind@ru.is

María Óskarsdóttir
mariaoskars@ru.is