Predicting students' digital confidence in schools across Europe

Based on the European Commission's 2nd Survey of Schools: ICT in Education

Courtney Froehlig
Data Science Capstone Project
December 2, 2020

Backround

- European Commission’s 2nd Survey of Schools: ICT in Education, conducted in 2018 (first survey conducted in 2011-2012)
- Survey of head teachers, teachers, students and parents from EU28, Norway, Iceland and Turkey
- Provided detailed information related to:
  - Access to, use of, and attitudes towards use of digital technologies
  - Digital activities and digital confidence of teachers and students
  - ICT related teacher professional development
  - Digital home environment of students
  - Schools’ digital policies, strategies and opinions

Students' confidence in their digital competence

- Provision, access and connectivity do not, in themselves, lead to ICT competence in learning and teaching (international comparative study of pedagogy and ICT use in schools, *SITES 2006)*
- Students' digital self-efficacy and confidence has gained considerable attention in research about students' learning and outcomes
- Survey confidence questions: students were asked to rate their level of confidence in their ability to perform 20 ICT-related tasks using a Likert scale ranging from 'none' to 'a lot'

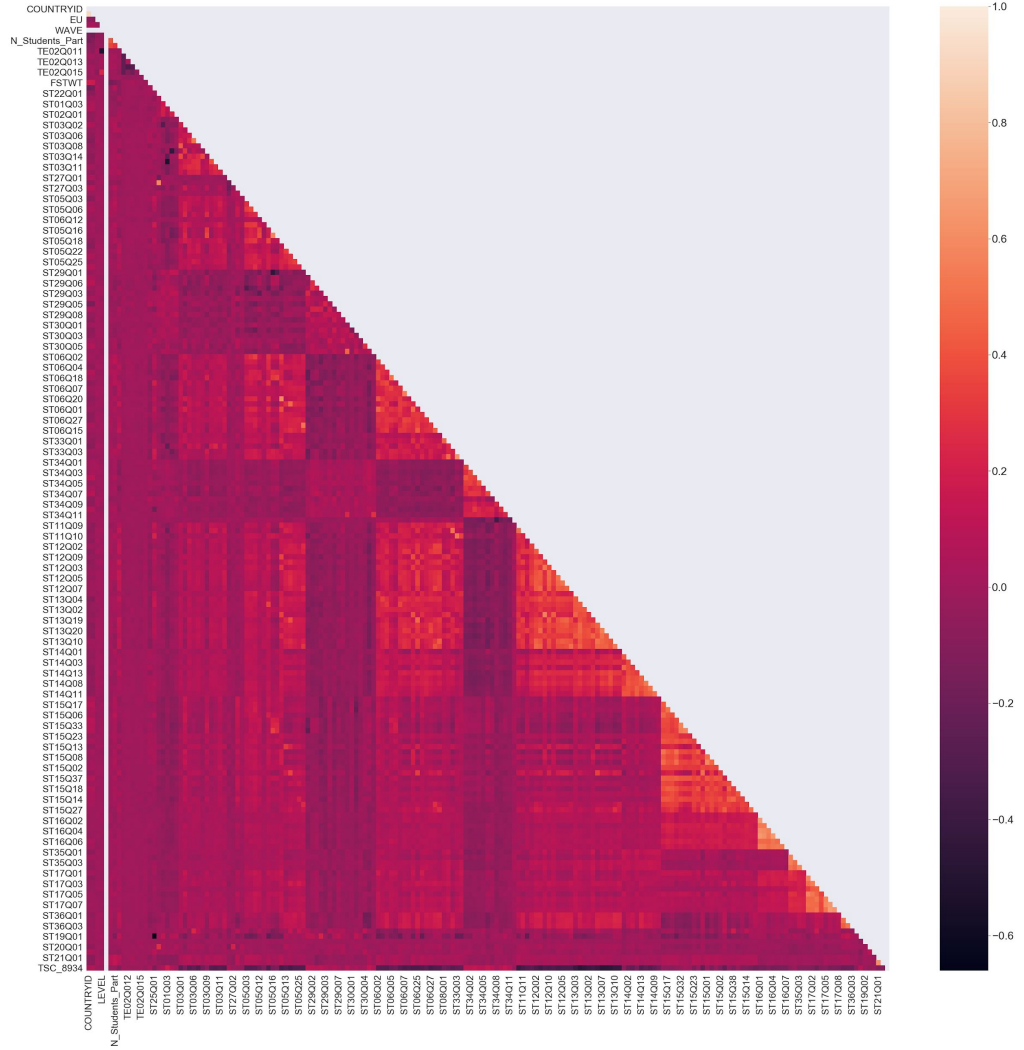
Goals of the project

- ➔ Use student survey questions to predict student confidence in an exploratory way
- ➔ Add in teacher student survey questions to see if the model improves
- ➔ Perform a cluster analysis to better understand students' digital confidence (with more fine-grained detail than just low/high confidence)
- ➔ Prioritise parsimonious models to examine the most important features

Data

Student survey data:

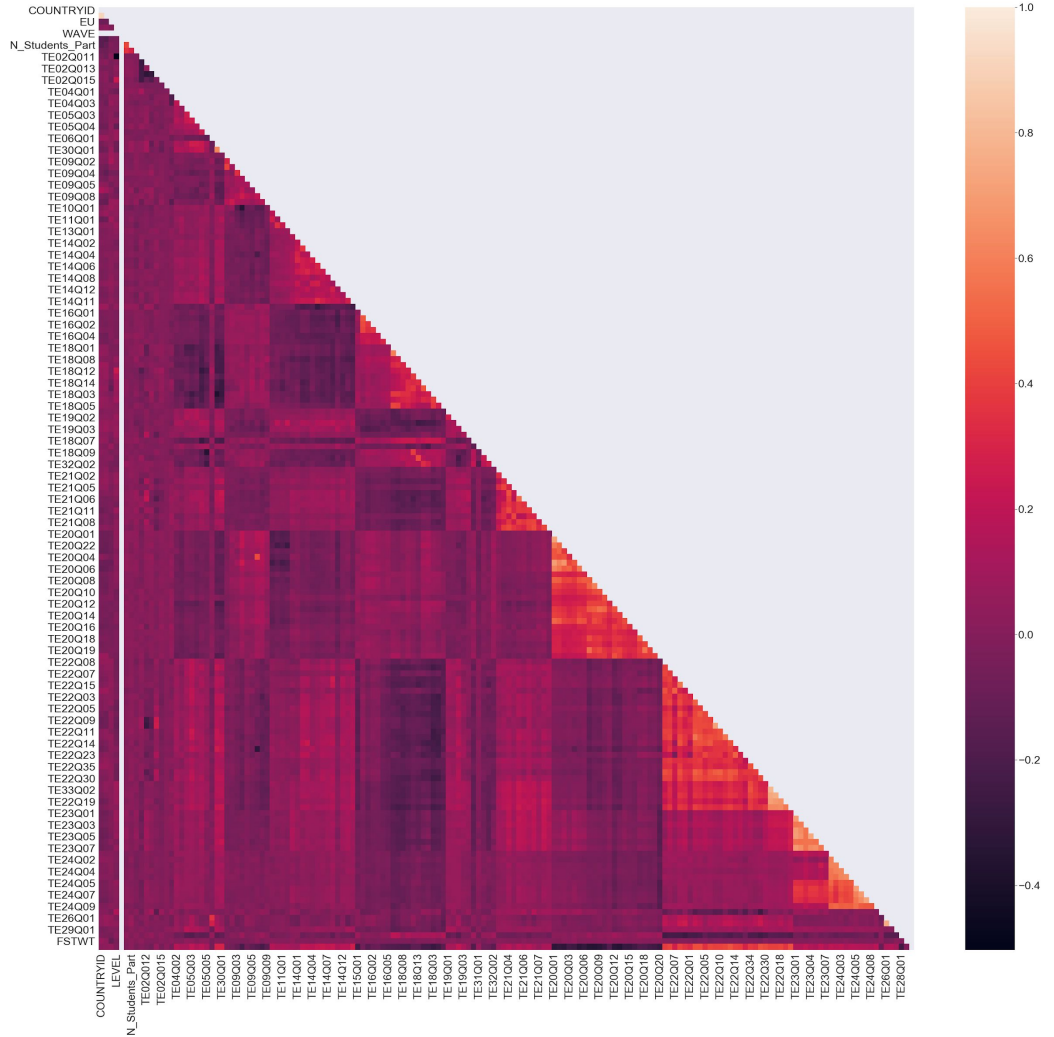
- 11.3 % missing values, imputed with mean
- 48835 rows
- 181 columns
- Variables are all scales (except country code), some skew on the variables with higher ranges but very few outliers due to small scale ranges
- Decided to treat variables as continuous, except for country
- Multicollinearity - can be treated with L2 regularization



Data

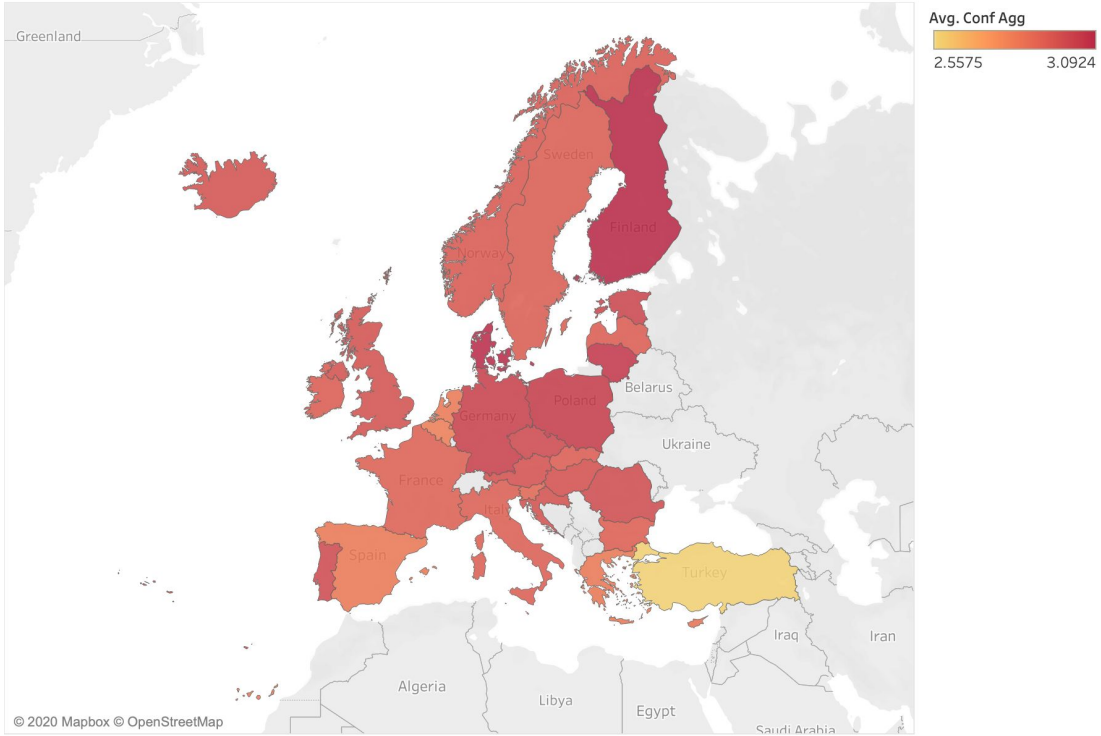
Teacher survey data:

- 6.6% missing data, imputed with mean
- 9927 rows
- 151 columns
- Variables are all scales (except country code), some skew on the variables with higher ranges but very few outliers due to small scale ranges
- Decided to treat variables as continuous, except for country
- High multicollinearity! Need to be cautious in interpreting results

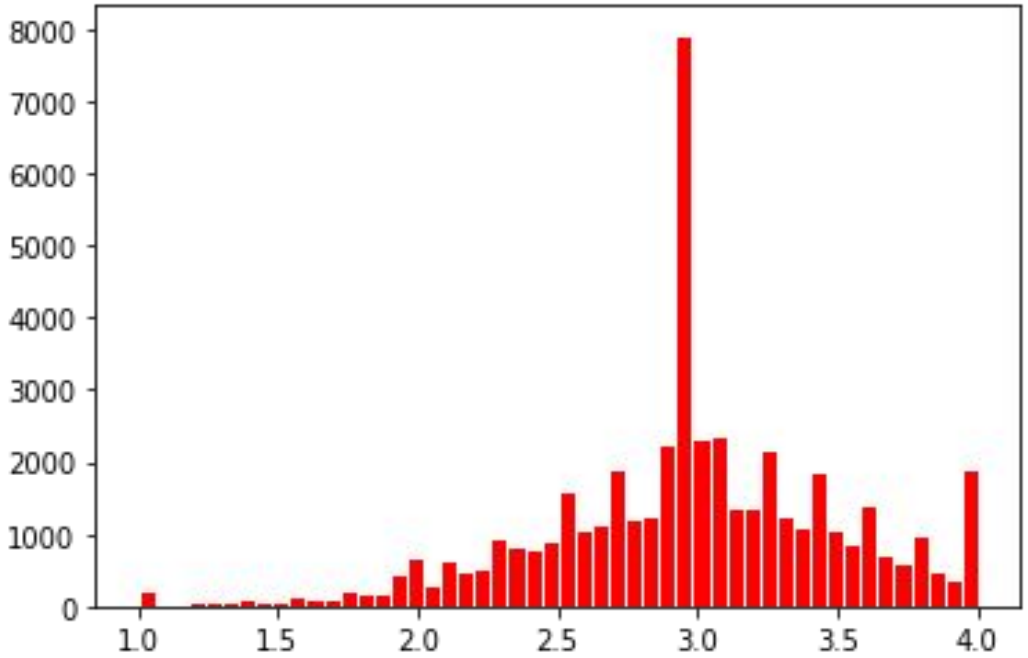


Mapping students' digital confidence

Students' digital confidence across Europe



Target Variable

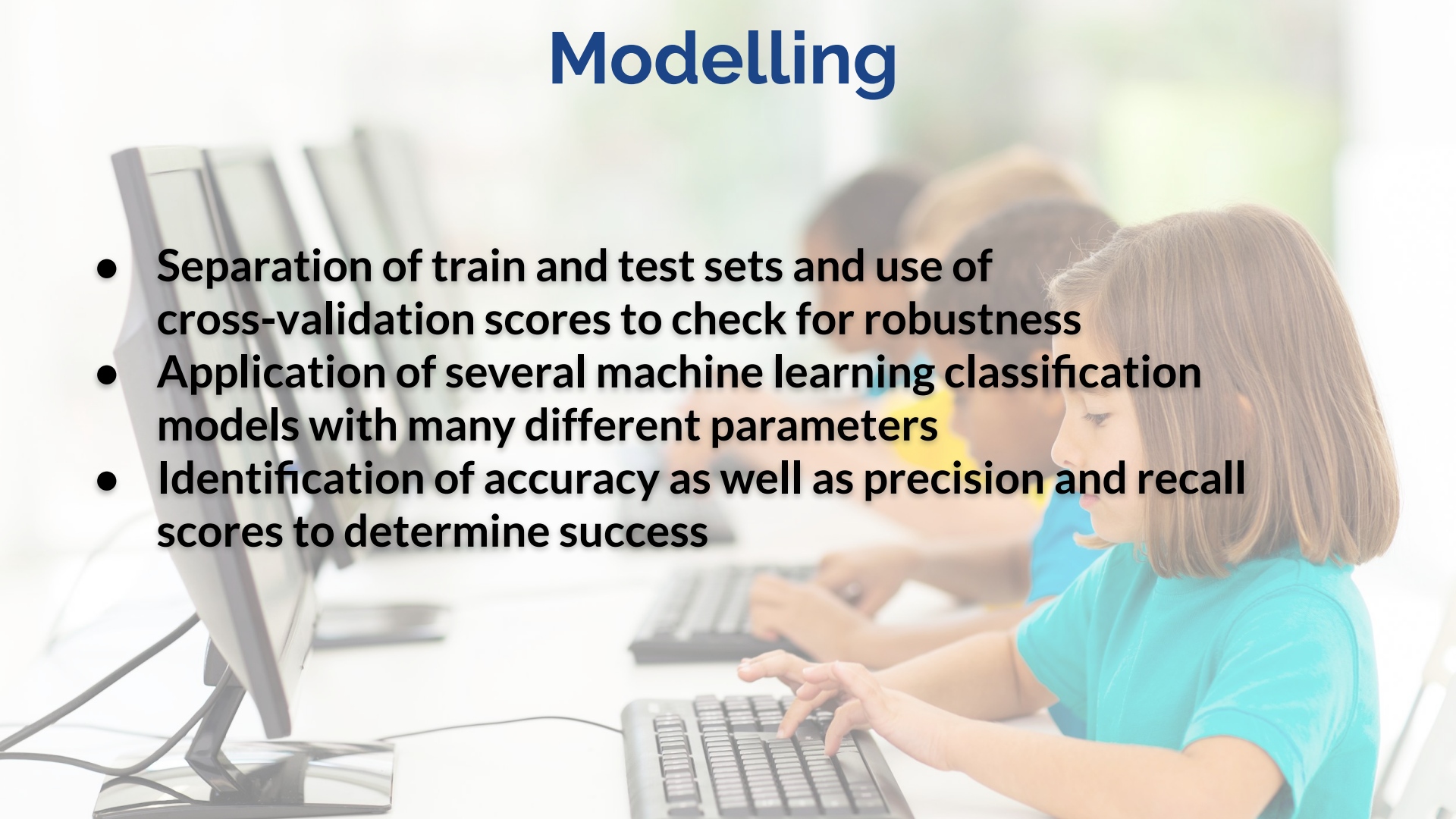


Confidence divided into two (relatively) equal groups of 'low' and 'high' confidence around the median:

Baseline:
0 0.54
1 0.45

Uneven groups due to the the high proportion of data points that were positioned exactly at the median

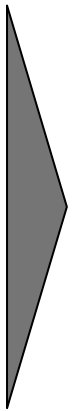
Modelling



- Separation of train and test sets and use of cross-validation scores to check for robustness
- Application of several machine learning classification models with many different parameters
- Identification of accuracy as well as precision and recall scores to determine success

Classifying low and high confidence classes

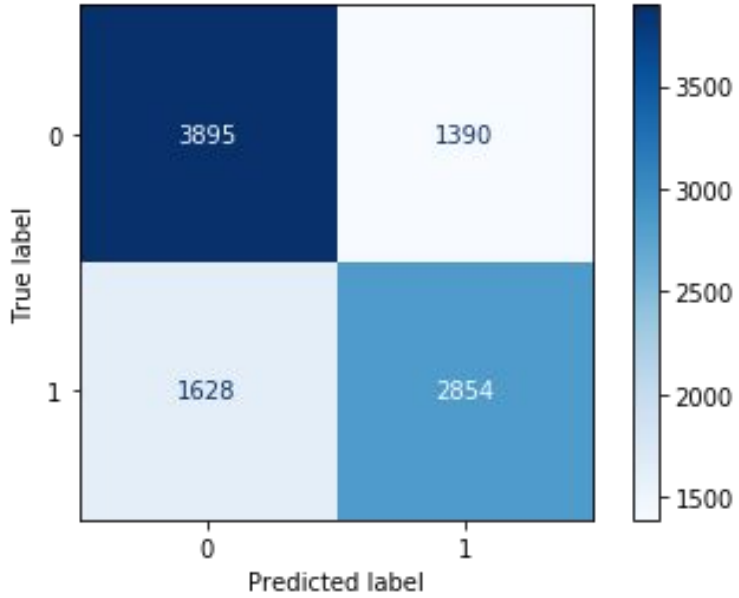
Model	Parameters	Best accuracy score for test set	Mean CV Score
Logistic Regression	penalty='l1' penalty='l2'	0.691 0.690	0.694 0.694
Naive Bayes	BernoulliNB	0.67	.66
Decision Tree Classifier	criterion='gini', max_depth=6,	0.692	.688
KNN	N neighbors = 200	0.566	.565
Bagging Classifier (Decision Tree)	Max depth = 5	0.712	0.69



Accuracy of logistic regression and Bagging classifier were 15 percentage points better than baseline

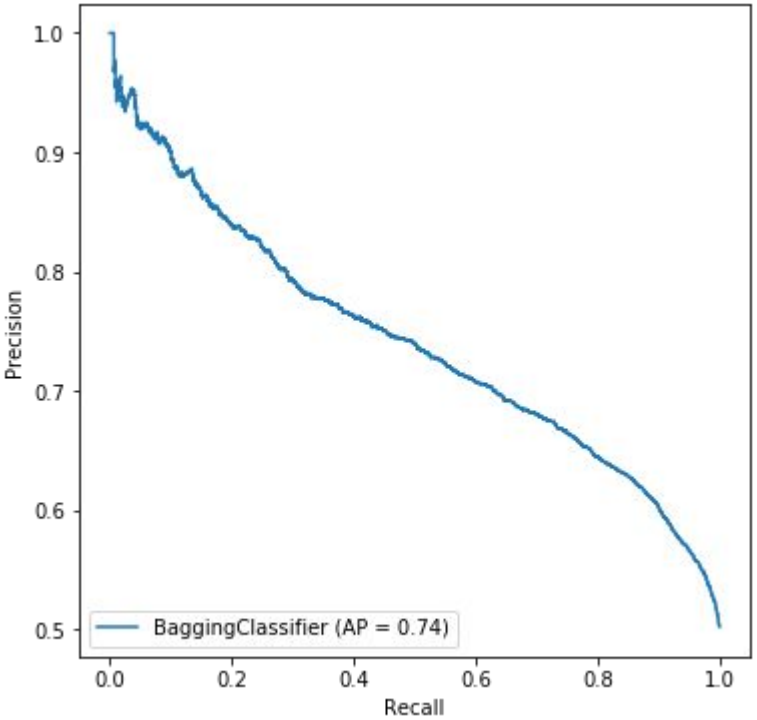
Precision and Recall Scores: Bagging Tree Classifier

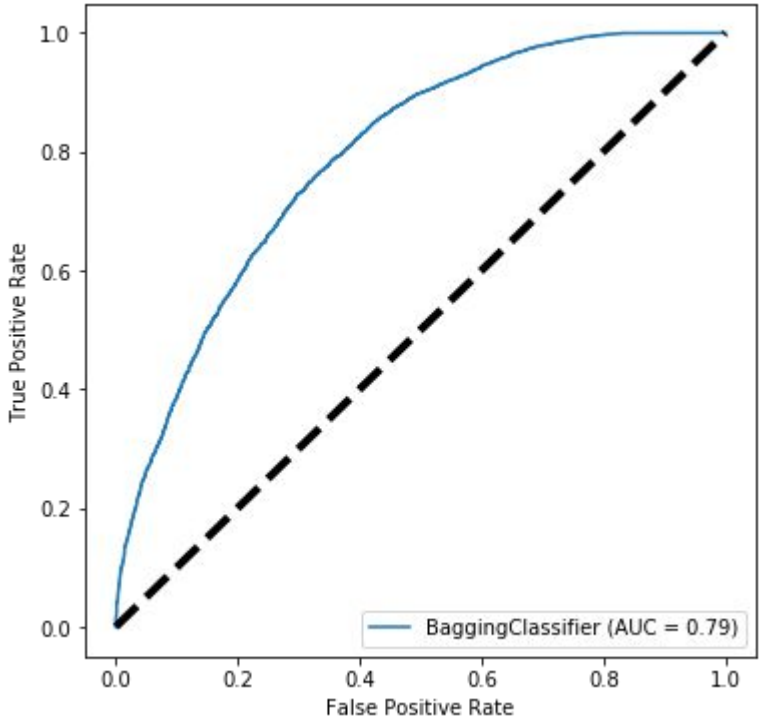
	precision	recall	f1-score	support
0	0.7426	0.7179	0.7300	5285
1	0.6799	0.7066	0.6930	4482
accuracy			0.7127	9767
macro avg	0.7113	0.7122	0.7115	9767
weighted avg	0.7138	0.7127	0.7130	9767



Precision-recall and ROC

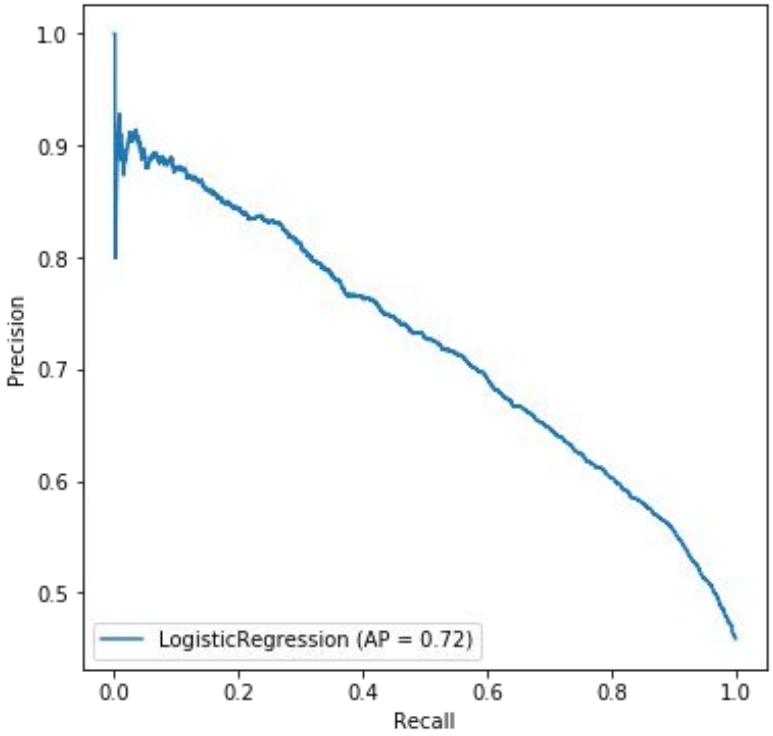
Bagging Tree Classifier

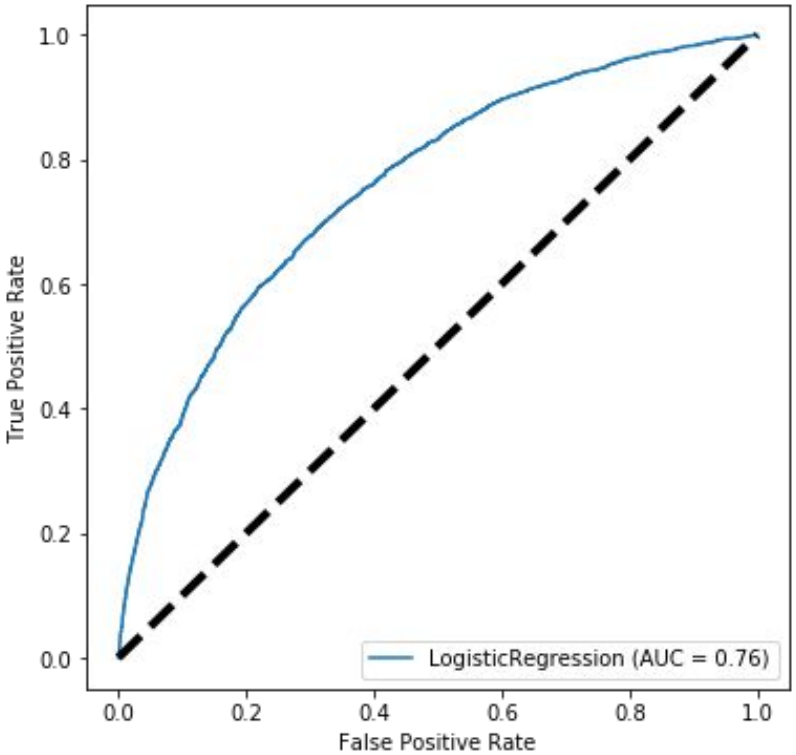




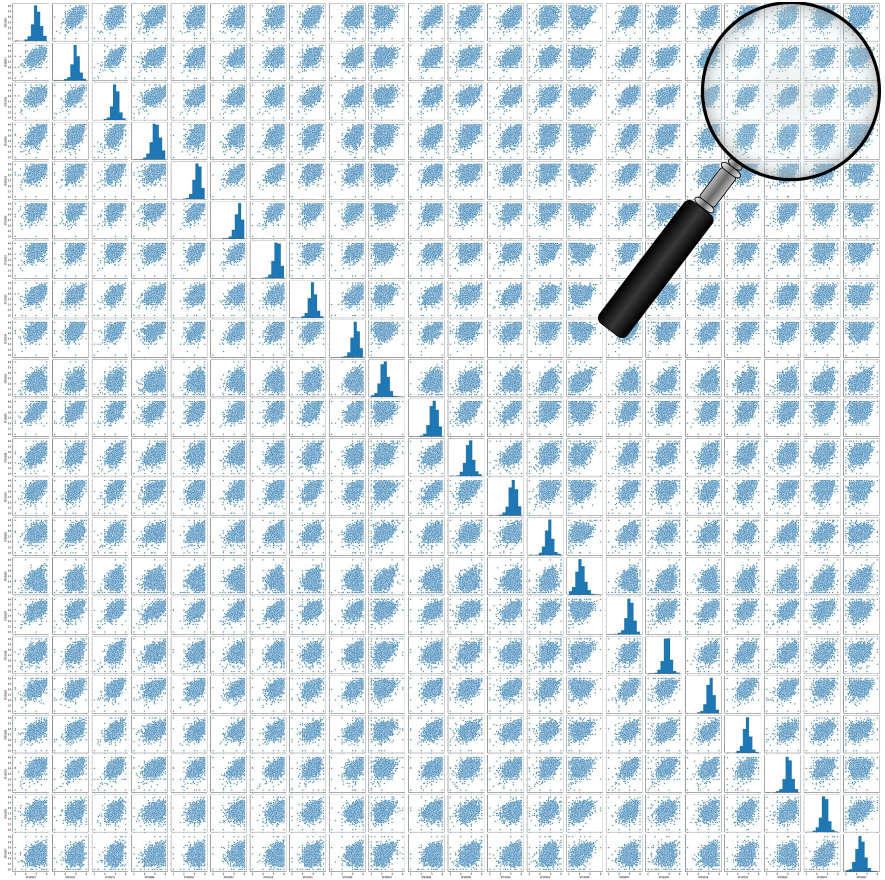
Precision-recall and ROC

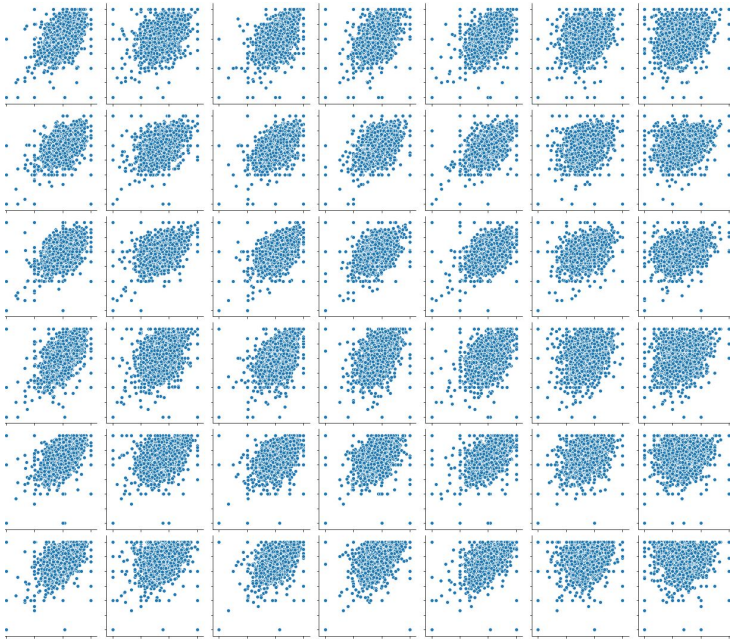
Logistic regression





Rethinking the target classes

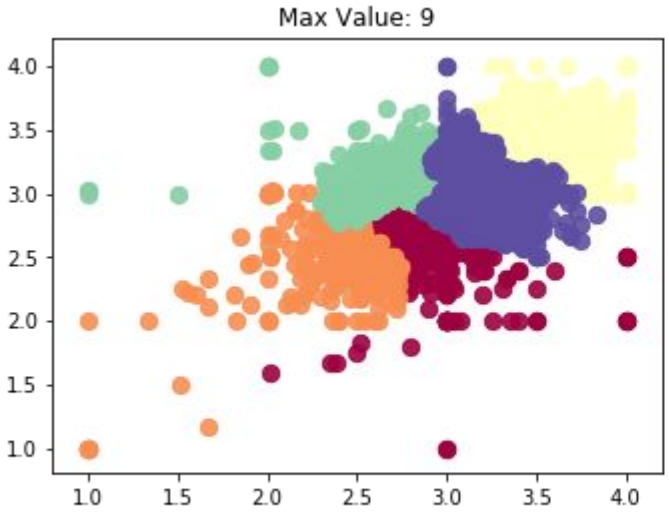


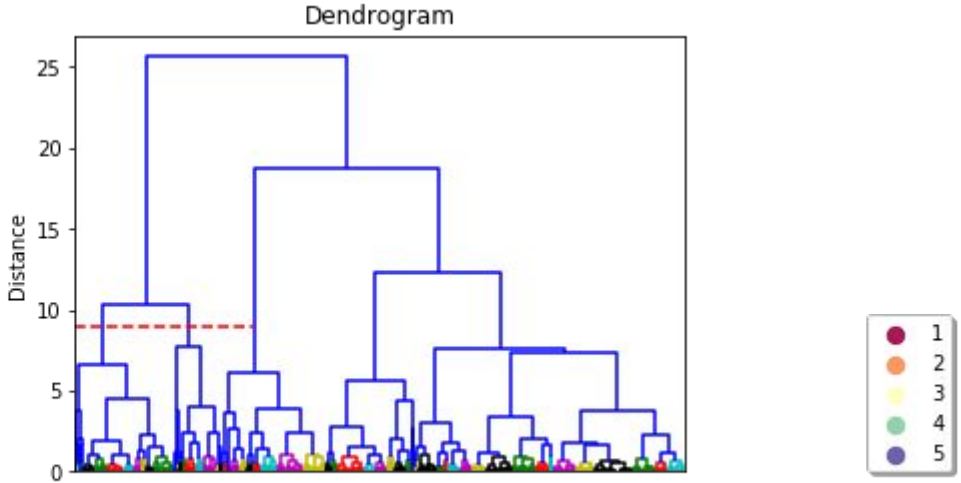


- Not always a linear relationship between individual confidence variables so low/high might not be the best way to classify confidence
- Solution: capture the complexity of the construct of confidence through a hierarchical cluster analysis

Hierarchical Cluster Analysis: Visualization

of clusters with two plotted variables*





Class dist:

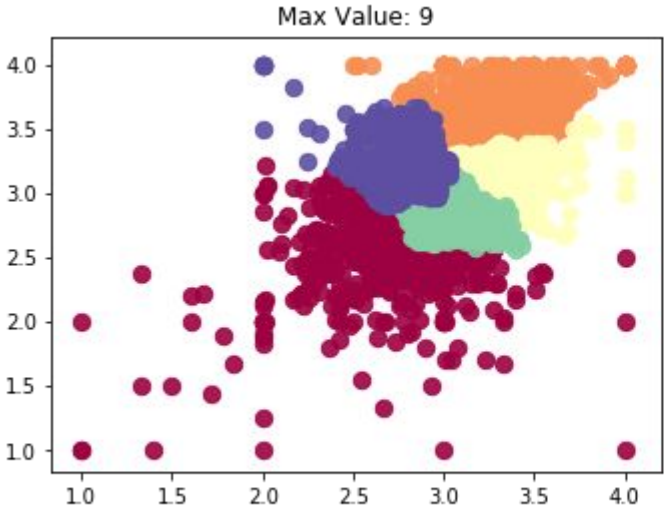
0	.312
1	.118
2	.147
3	.043
4	.379

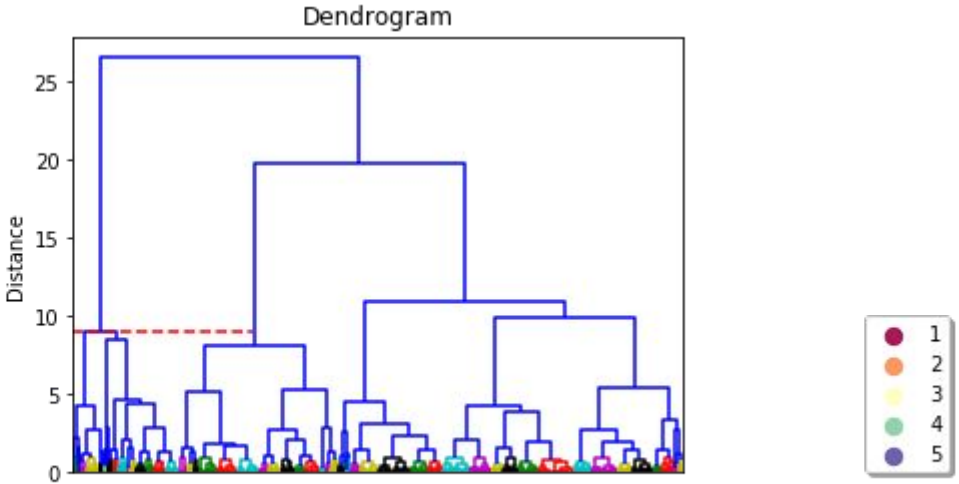
→ Baseline accuracy

**ST15Q07 (confidence with filing electronic documents in computer folders and sub-folders) and ST15Q017 (confidence with identifying online sources of reliable information)*

Hierarchical Cluster Analysis:

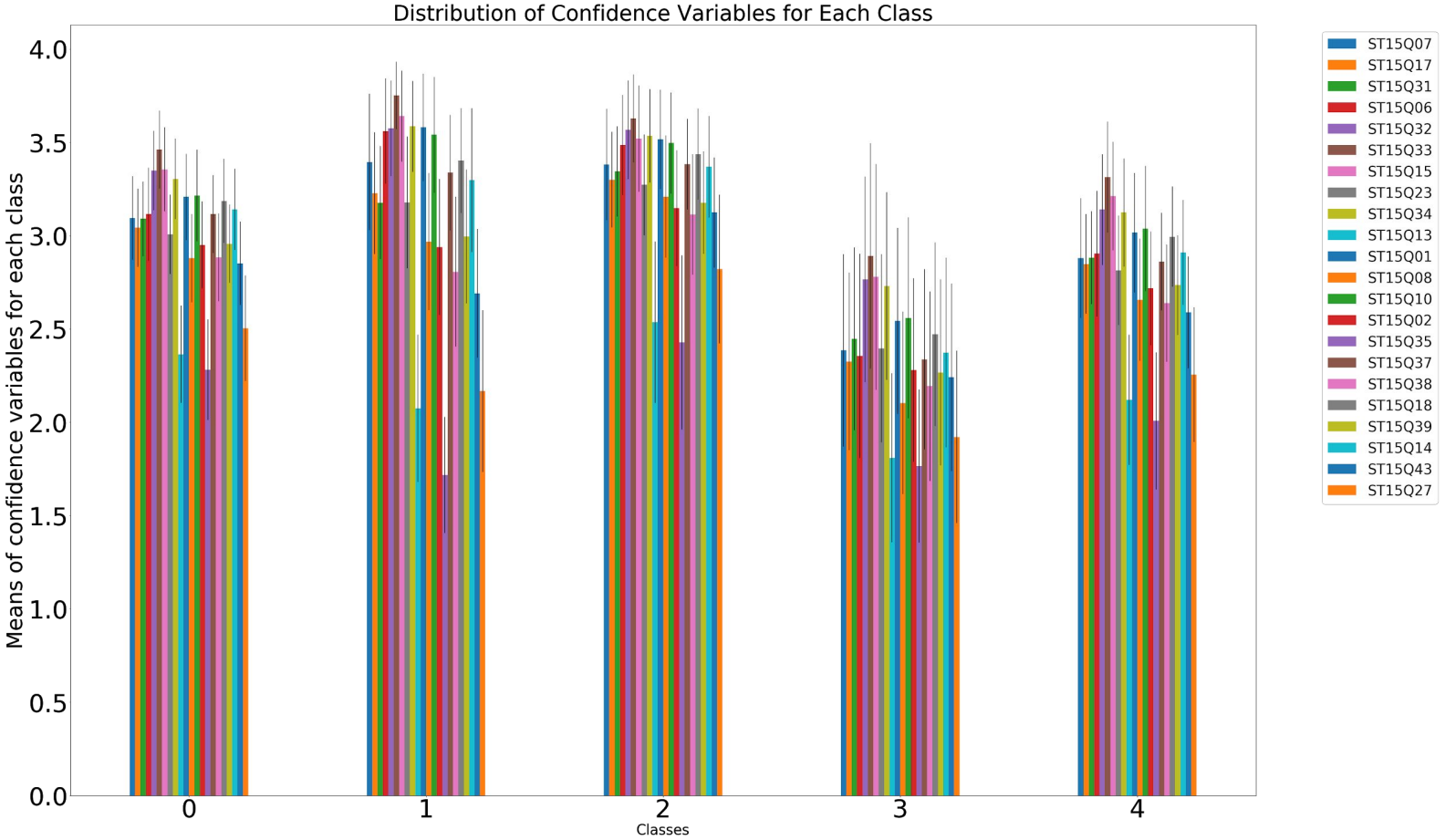
Visualization of clusters with two plotted variables*





ST15Q31 (confidence with checking if the information that found online is true) and ST15Q06 (confidence with emailing a file to someone/another student or teacher)

Examining the clusters



Comparing means of confidence variables across clusters

- **Group 1** - higher variability between lows and highs, very low confidence with coding & creating websites, highest on using mobile applications/social media
- **Group 2** - consistently high among variables, higher coding and creating websites
- **Group 3** - consistently low overall - low on both coding/programming/creating websites and social media/communication

One way f-test to compare means across three groups

Question	Statistic	P-value
ST15Q07	481.691814	7.314798e-149
ST15Q17	633.090984	1.110168e-180
ST15Q31	323.708339	4.552192e-110
ST15Q06	801.901418	1.434286e-211
ST15Q32	503.809265	8.461130e-154
ST15Q33	512.063654	1.294923e-155
ST15Q15	228.639928	4.722715e-83
ST15Q23	455.885252	5.807146e-143
ST15Q34	366.409364	3.161158e-121
ST15Q13	538.482506	2.503740e-161
ST15Q01	610.234575	3.728315e-176
ST15Q08	376.698490	7.788718e-124
ST15Q10	397.450650	5.252531e-129
ST15Q02	471.348880	1.622968e-146
ST15Q35	402.788740	2.565050e-130
ST15Q37	455.850696	5.915162e-143
ST15Q38	489.290674	1.431887e-150
ST15Q18	386.745124	2.361833e-126
ST15Q39	677.385723	3.338961e-189
ST15Q14	329.976796	9.667689e-112
ST15Q43	388.024294	1.134027e-126
ST15Q27	414.525637	3.568106e-133

Classifying five target classes

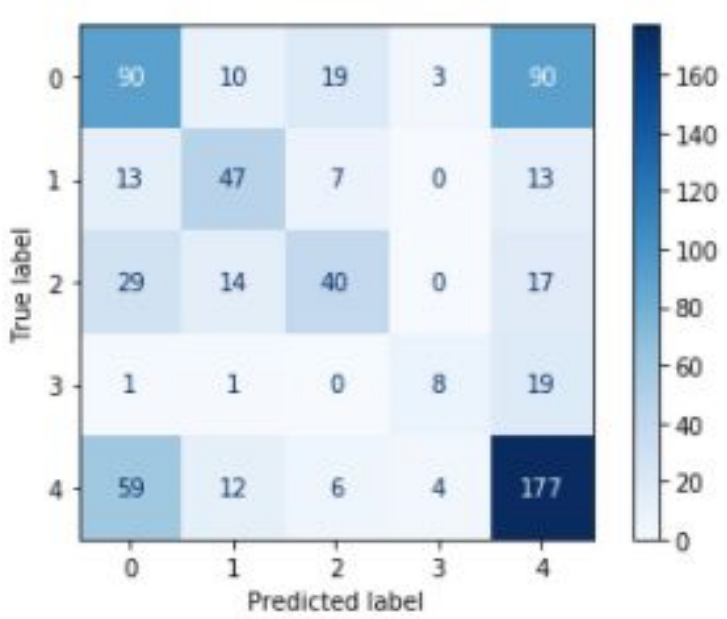
Model	Parameters	Best test accuracy	Mean CV Score
Logistic Regression	Penalty = l1 Penalty = l2 Penalty = ElasticNet	0.533 0.534 0.532	0.507 0.501 0.506
Naive Bayes	BernoulliNB	0.377	0.374
Decision Tree Classifier	criterion='gini', max_depth=5	0.43	0.436
Support Vector Machine	Rbf kernel	0.564	0.53
Bagging Classifier (Decision Tree)		0.5007	0.501
Gradient Boosting Classifier	Max depth = 3	0.512	0.477
Ada Boost (Decision Tree)	Max depth = 3	0.445	0.465
Neural Network	4 folds, each with 100 nodes	0.5596	.547

The support vector machine performed 47% better than baseline

Precision and Recall:

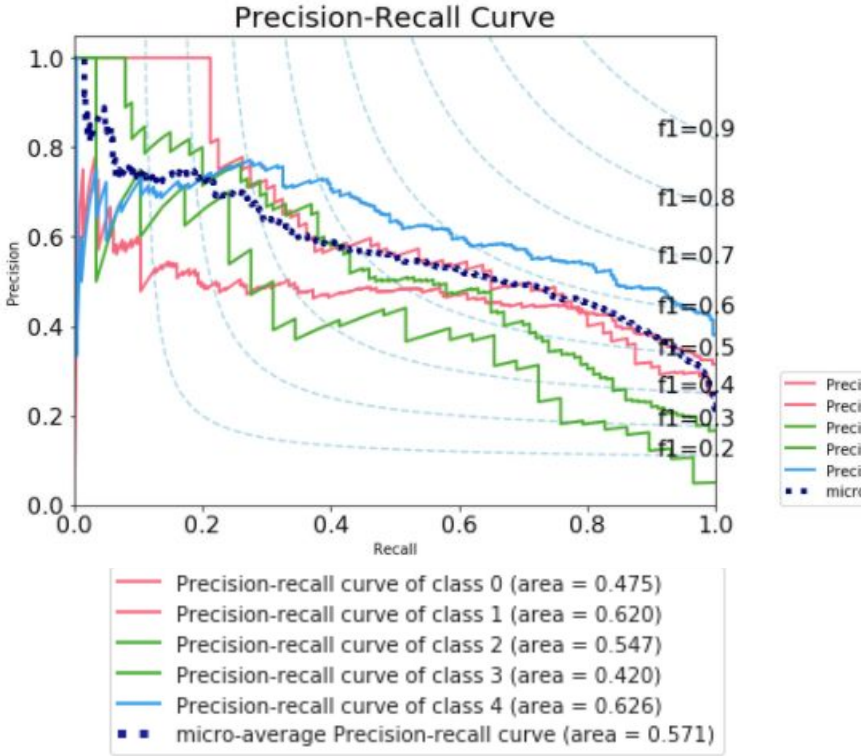
Logistic regression

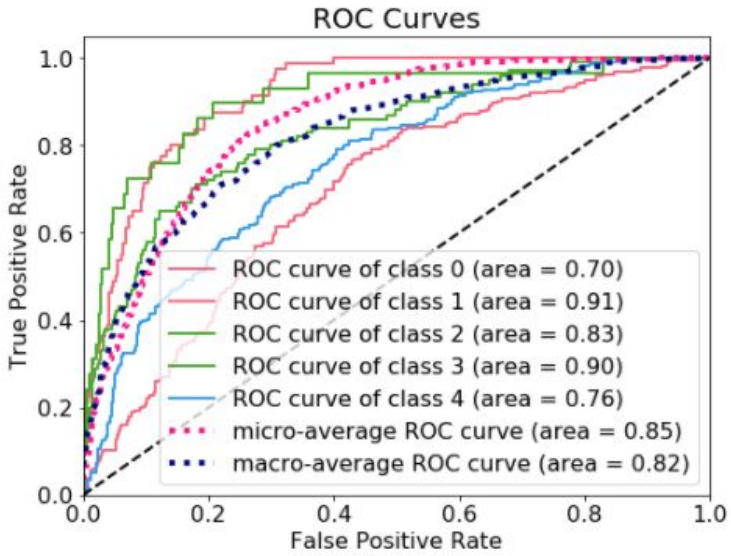
	precision	recall	f1-score	support
0	0.4688	0.4245	0.4455	212
1	0.5595	0.5875	0.5732	80
2	0.5556	0.4000	0.4651	100
3	0.5333	0.2759	0.3636	29
4	0.5601	0.6860	0.6167	258
accuracy			0.5331	679
macro avg	0.5355	0.4748	0.4928	679
weighted avg	0.5297	0.5331	0.5250	679



Precision-recall and ROC

Logistic regression

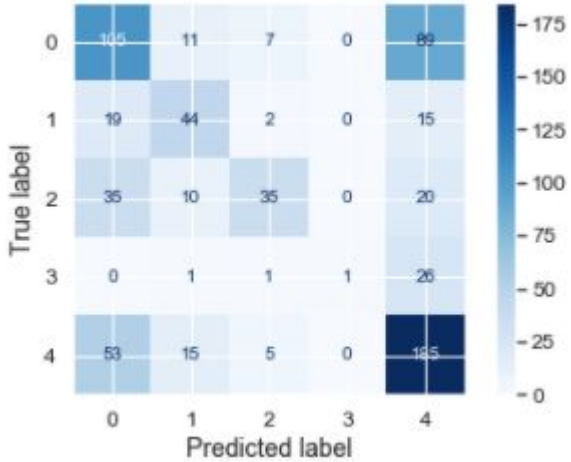




Precision and Recall:

Support Vector Machine

	precision	recall	f1-score	support
1	0.5067	0.5330	0.5195	212
2	0.5556	0.6250	0.5882	80
3	0.6731	0.3500	0.4605	100
4	0.7778	0.2414	0.3684	29
5	0.5836	0.6899	0.6323	258
accuracy			0.5641	679
macro avg	0.6193	0.4879	0.5138	679
weighted avg	0.5778	0.5641	0.5553	679



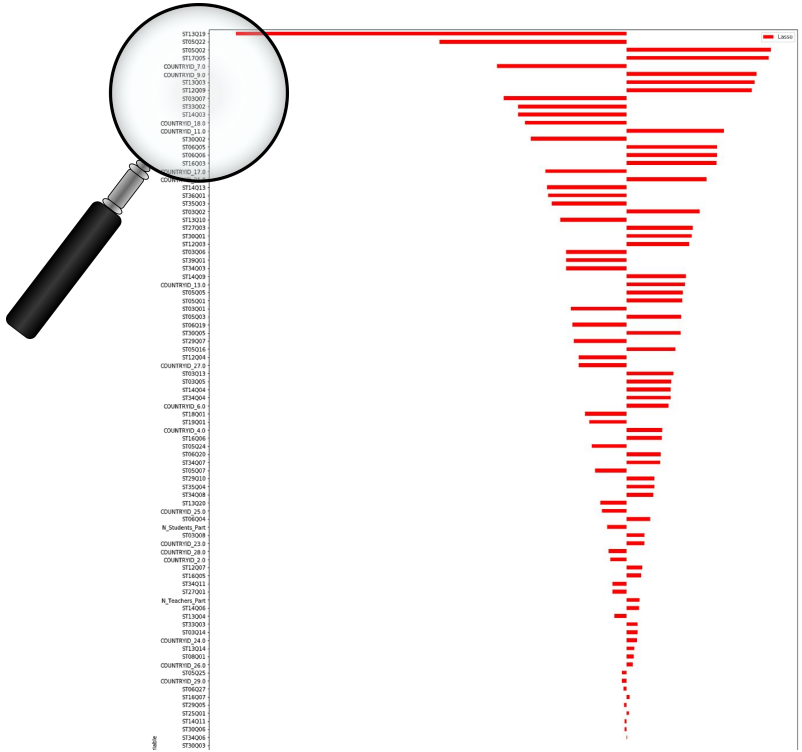
Coefficients for predicting Class 1 (class with the most varied confidence)

Top negative coefficients:

- ST13Q09: How often are you engaged in Code/programming apps, programmes and/or robots in classroom
- ST05Q22: How often do you take part in the following activities in your free time, at home or any place other than school? Coding and programming apps, programmes and/or robots
- Finland

Top Positive coefficients:

- ST05Q02 How often do you spend time Chatting online (e.g. WhatsApp, Viber, Google Hangouts, Facebook messenger, Skype messenger, etc.) in your free time?
- Denmark



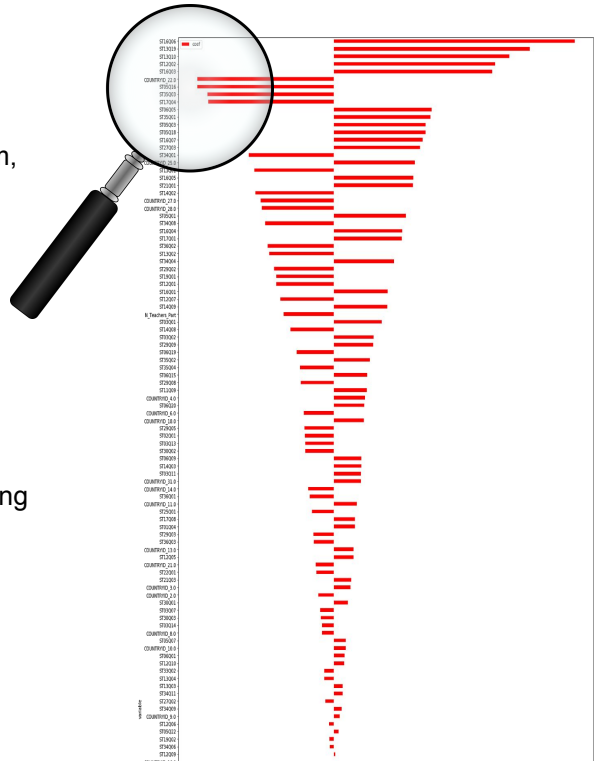
Coefficients for predicting Class 2 (class with highest overall confidence)

Top negative coefficients:

- Country 22: Norway
- ST05Q16: Participate in social networks (e.g. Facebook, Instagram, Twitter, Snapchat, Ask.fm, etc.)
- ST35Q03: My school encourages me to use ICT to learn by doing instead of just listening to lectures
- ST17Q04: I lose track of time when I'm learning with the computer

Top positive coefficients:

- ST1606: ICT enables you to work better with other students on tasks
- ST13Q19/Q10: Code/programme apps, programmes and/or robots/Participate in online training programmes during lessons
- ST12Q02: Use software, online quizzes and tests
- ST16Q03: Does ICT make you feel more independent in your learning?
- ST06Q05 Use other online tools on a computer (e.g. Viber, Google Hangouts, Facebook, Skype, etc.) to contact other students about schoolwork
- ST35Q01 My school encourages me to use my digital skills in a variety of learning activities
- ST05Q03 How often do you read and watch the news online at home?



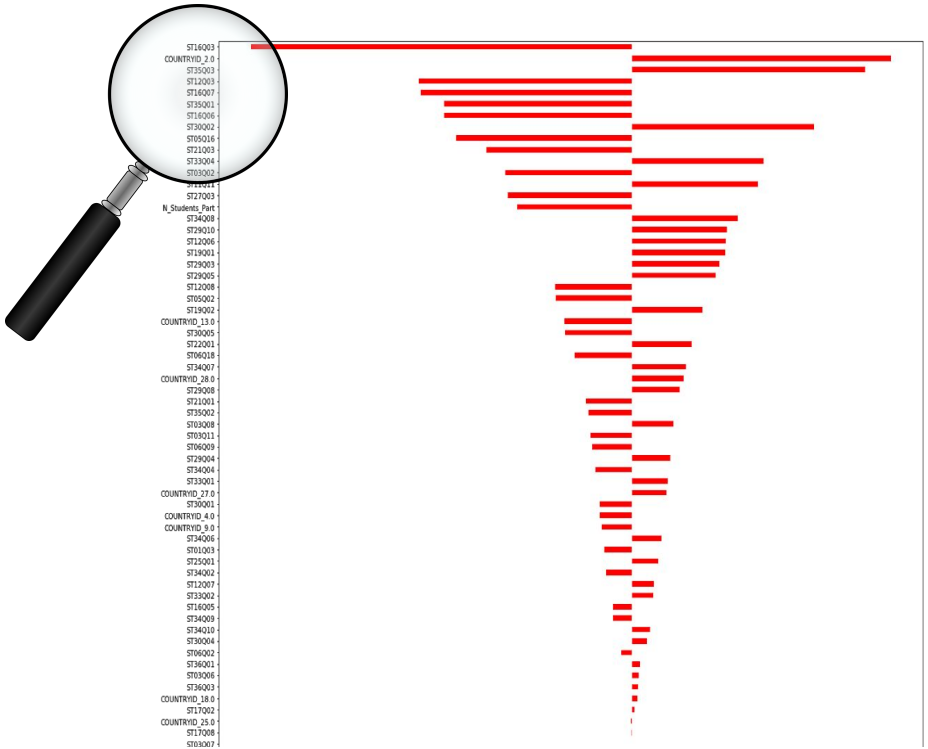
Coefficients for predicting Class 3 (the lowest overall confidence class)

Top negative coefficients:

- ST16Q03: using technology during lessons helps me to feel more independent in my learning
- ST12Q03 How often do you use the following in lessons: Multimedia production tools (e.g. PowerPoint, video editing, digital recording)
- ST16Q07 ICT improves the atmosphere in class (students are more engaged, there is less disruption)
- ST35Q01 My school encourages me to use my digital skills in a variety of learning activities
- ST16Q06: ICT enables you to work better with other students on tasks

Top positive coefficients:

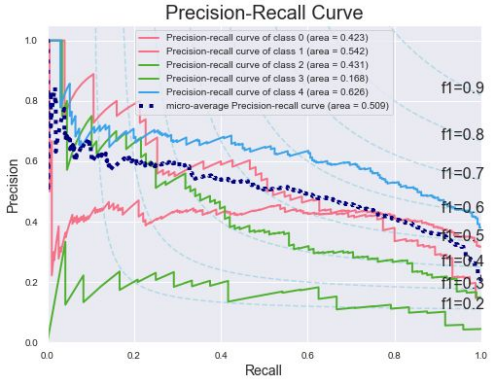
- ST35Q03 My school encourages me to use ICT to learn by doing instead of just listening to lectures, positively associated with being a part of this group
- ST30Q02 How do you know the information you read is reliable? I double check with another source

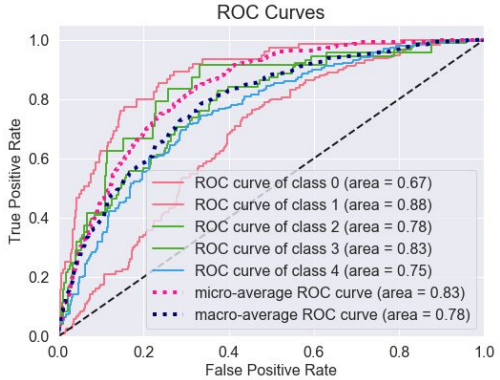


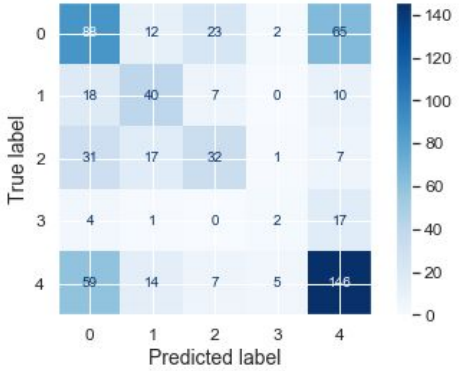
Models with added teacher survey data

Model	Key parameters	Best R2 score for test set	Mean CV Score
Logistic Regression	Penalty = l1 Penalty = l2 Penalty = ElasticNet	0.506 0.495 0.506	.0479 0.477 0.481
Decision Tree Classifier	criterion='gini', max_depth=6	0.43	0.42

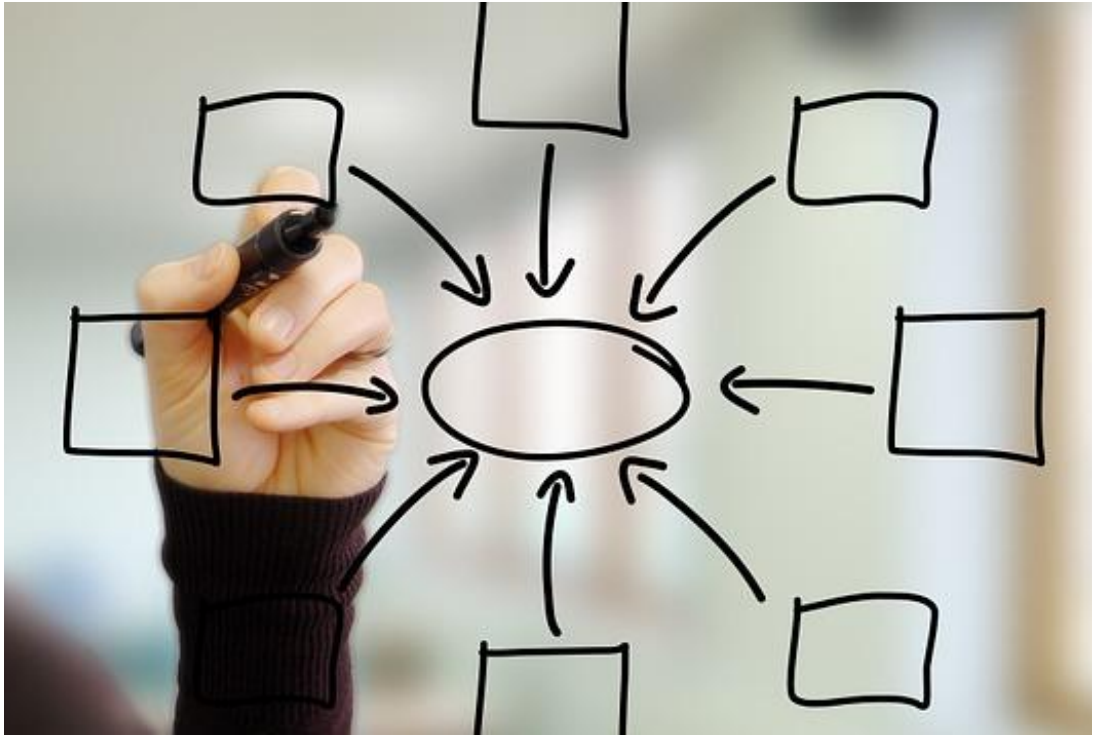
The model did improve slightly by adding in teacher data!







Conclusions



Findings and Next Steps

- Important to think about different types of digital competence: operational use and social media use (descriminating using social media for communication with other students about school work vs. leisure) , in addition to safe and responsible use of the internet
- Teacher data did not improve the model: teachers' attitudes/confidence/skills may not play an influential role in students' digital confidence
- Students reported overall lower confidence on programming/creating websites, but they reported higher confidence when they participated in online training programmes and used ICT to collaborate with other students
- Moving forward: look at differences between age groups/school levels, measure actual digital competence/skills in order develop a better understanding of how outcomes relate to confidence; examine students' experiences with ICT in schools and at home as predictors of digital skills and competence

Thank you

Courtney Froehlig
courtfroehlig@gmail.com