**Proposed list of contributions**

**Contribution 1: Different types of students revealed through clustering (Week 4) – exploratory data analysis**

* **Density plot (i.e. saliency map) showing** 
  + Final scores vs number of slides completed

**Results**

* + There are categories of students
    - Students who perform well but do not complete any slides
    - Students who perform well and complete slides (linear relationship)
    - Students who do not perform well and complete slides
    - Students who do not perform well and do not complete any slides

**Contribution 2: Course dropout prediction**

* Defined three types of dropout behaviours: early dropout, late dropout, no dropout
* Dropout distributions, as measured by weeks and types are relatively spread out.
* Course dropout prediction was framed as a classification task
  + **Classifier:** Multinomial logistic regression
  + **Validation:** 10-fold cross-validation with stratification
  + **Featurs:** Only used features from the first 1, 2 and 3 weeks of the course.
    - Aggregate
      * Number of problems completed in first 1,2 and 3 weeks.
      * Number of slides completed in first 1,2 and 3 weeks.
      * Number of slides attempted but not completed in first 1, 2 and 3 weeks
      * Number of problems attempted but not completed in first 1, 2 and 3 weeks
  + **Conclusions**
    - Student’s interactions for the first week can predict course dropout relatively well, with accuracy that is comparable to if we had taken a longer subset (e.g. up to three weeks) or even the entire interaction sequence.
    - **What this really means:** Student course dropout behaviour is highly determined by how they interact with slides in the first few weeks of the challenge.
    - **What does this mean for educators:** For educators, this emphasises the importance of intervening early to ensure that students can progress throughout the course.
  + **Work that needs to be done** 
    - Correlative analysis between the different features esp. between slides and problems.

**Contribution 3: Module outcome prediction (tonight)**

Predict the outcome of the final problem in the module, using all previous slides (and then previous problems too i.e. include Ns, Fs

**Contribution 4: Slide value importance**

Quantified the value of each slide in determining whether a student would successfully pass the last problem of each module, and whether they would dropout (no submission for the last problem). The educational value of this is that it helps educators and course content creators understand how their content affects performance for that module. This section also compares and validates these results with Grok designers.

**To do**

* Compare with different feature importance methods, and choose the most appropriate method – also use CFS (correlative feature selection).
* Rank the different features using the different methods, and try to reconcile among different methods.
* Information gain
* Evaluate the slide importance with Grok designers
* Compare this with what students complete

**Contribution 5: Evaluate across all challenges and analyse differences/similarities**

* Detail what was required to make it extensible to all challenges.

**Contribution 6: Conference Paper**

**Updates**

* **Procedure**
  + Isolated each module, and used the interaction sequence representing all slides besides the last problem slide to predict the outcome of the problem slide.
  + In the case of having two paired problem slides, the interaction sequence did not include the first problem slide.
  + Repeated the above for all four challenges (challenge-beginners-2018, challenge-newbies-2018, challenge-intermediate-2018).
  + Still need to do challenge-beginners-blockly-2018 challenge-advanced-2018)
* **Conclusions**
  + Performance for all three classifiers (multinomial linear regression, naive Bayes, decision trees) is higher than the base classifier performance, and are generally high and consistent with each other.
  + Multinomial linear regression tends to slightly outperform the other two classifiers across all challenges.

challenge-newbies-2018

|  |  |
| --- | --- |
| challenge-newbies-2018 |  |
| w1p1 | Base score: 0.65  **Linear regression: 0.86**  Naive Bayes: 0.62  Decision tree: 0.85 |
| w1p2 | Base score: 0.65  **Linear regression: 0.82**  Naive Bayes: 0.77  Decision tree: 0.80 |
| w2p1 | Base score: 0.69  **Linear regression: 0.86**  Naive Bayes: 0.83  Decision tree: 0.84 |
| w2p2 | Base score: 0.63  **Linear regression: 0.86**  Naive Bayes: 0.76  Decision tree: 0.85 |
| w3p1 | Base score: 0.62  **Linear regression: 0.86**  Naive Bayes: 0.77  Decision tree: 0.84 |
| w3p2 | Base score: 0.58  **Linear regression: 0.83**  Naive Bayes: 0.27  Decision tree: 0.81 |
| w4p1 | Base score: 0.53  **Linear regression: 0.86**  Naive Bayes: 0.85  Decision tree: 0.85 |
| w4p2 | Base score: 0.56  **Linear regression: 0.86**  Naive Bayes: 0.31  Decision tree: 0.85 |
| w5p1 | Base score: 0.50  **Linear regression: 0.85**  Naive Bayes: 0.42  Decision tree: 0.83 |
| w5p2 | Base score: 0.50  **Linear regression: 0.86**  Naive Bayes: 0.46  Decision tree: 0.85 |

|  |  |
| --- | --- |
| challenge-beginners-2018 |  |
| w1p1 | Base score: 0.80  **Linear regression: 0.95**  Naive Bayes: 0.90  Decision tree: 0.95 |
| w1p2 | Base score: 0.82  **Linear regression: 0.95**  Naive Bayes: 0.92  Decision tree: 0.94 |
| w2p1 | Base score: 0.82  **Linear regression: 0.95**  Naive Bayes: 0.93  Decision tree: 0.94 |
| w2p2 | Base score: 0.80  **Linear regression: 0.93**  Naive Bayes: 0.92  Decision tree: 0.93 |
| w3p1 | Base score: 0.79  **Linear regression: 0.92**  Naive Bayes: 0.90  Decision tree: 0.91 |
| w3p2 | Base score: 0.78  **Linear regression: 0.92**  Naive Bayes: 0.91  Decision tree: 0.92 |
| w4p1 | Base score: 0.60  **Linear regression: 0.66**  Naive Bayes: 0.76  Decision tree: 0.77 |
| w4p2 | Base score: 0.65  **Linear regression: 0.90**  Naive Bayes: 0.88  Decision tree: 0.90 |
| w5p1 | Base score: 0.62  **Linear regression: 0.92**  Naive Bayes: 0.90  Decision tree:0.92 |
| w5p2 | Base score: 0.49  Linear regression: 0.85  Naive Bayes: 0.85  Decision tree: 0.85 |

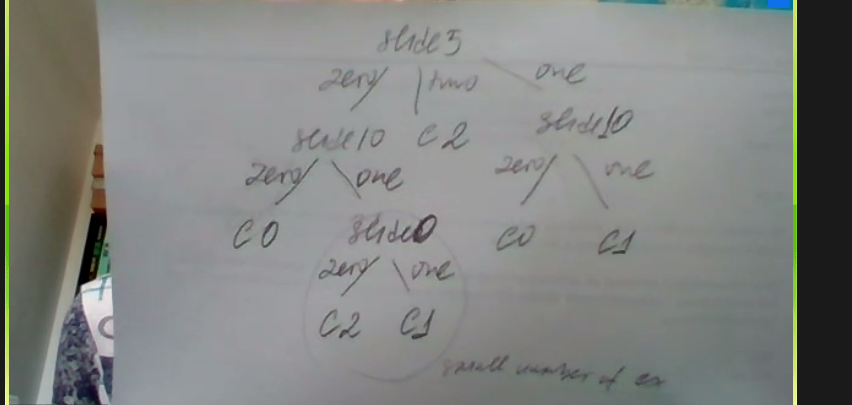
|  |  |
| --- | --- |
| challenge-intermediate-2018 |  |
| w1p1 | Base score: 0.77  **Linear regression: 0.91**  Naive Bayes: 0.91  Decision tree: 0.91 |
| w1p2 | Base score: 0.84  Linear regression: 0.93  Naive Bayes: 0.92  **Decision tree: 0.94** |
| w2p1 | Base score: 0.77  Linear regression: 0.87  Naive Bayes: 0.87  **Decision tree: 0.88** |
| w2p2 | Base score: 0.75  Linear regression: 0.88  Naive Bayes: 0.89  **Decision tree: 0.89** |
| w3p1 | Base score: 0.76  Linear regression: 0.88  Naive Bayes: 0.89  **Decision tree: 0.89** |
| w3p2 | Base score: 0.73  Linear regression: 0.90  Naive Bayes: 0.90  **Decision tree: 0.91** |
| w4p1 | Base score: 0.63  **Linear regression: 0.85**  Naive Bayes: 0.85  Decision tree: 0.85 |
| w4p2 | Base score: 0.54  **Linear regression: 0.80**  Naive Bayes: 0.80  Decision tree: 0.80 |
| w5p1 | Base score: 0.71  **Linear regression: 0.95**  Naive Bayes: 0.94  Decision tree: 0.95 |
| w5p2 | Base score: 0.65  **Linear regression: 0.86**  Naive Bayes: 0.87  Decision tree: 0.87 |

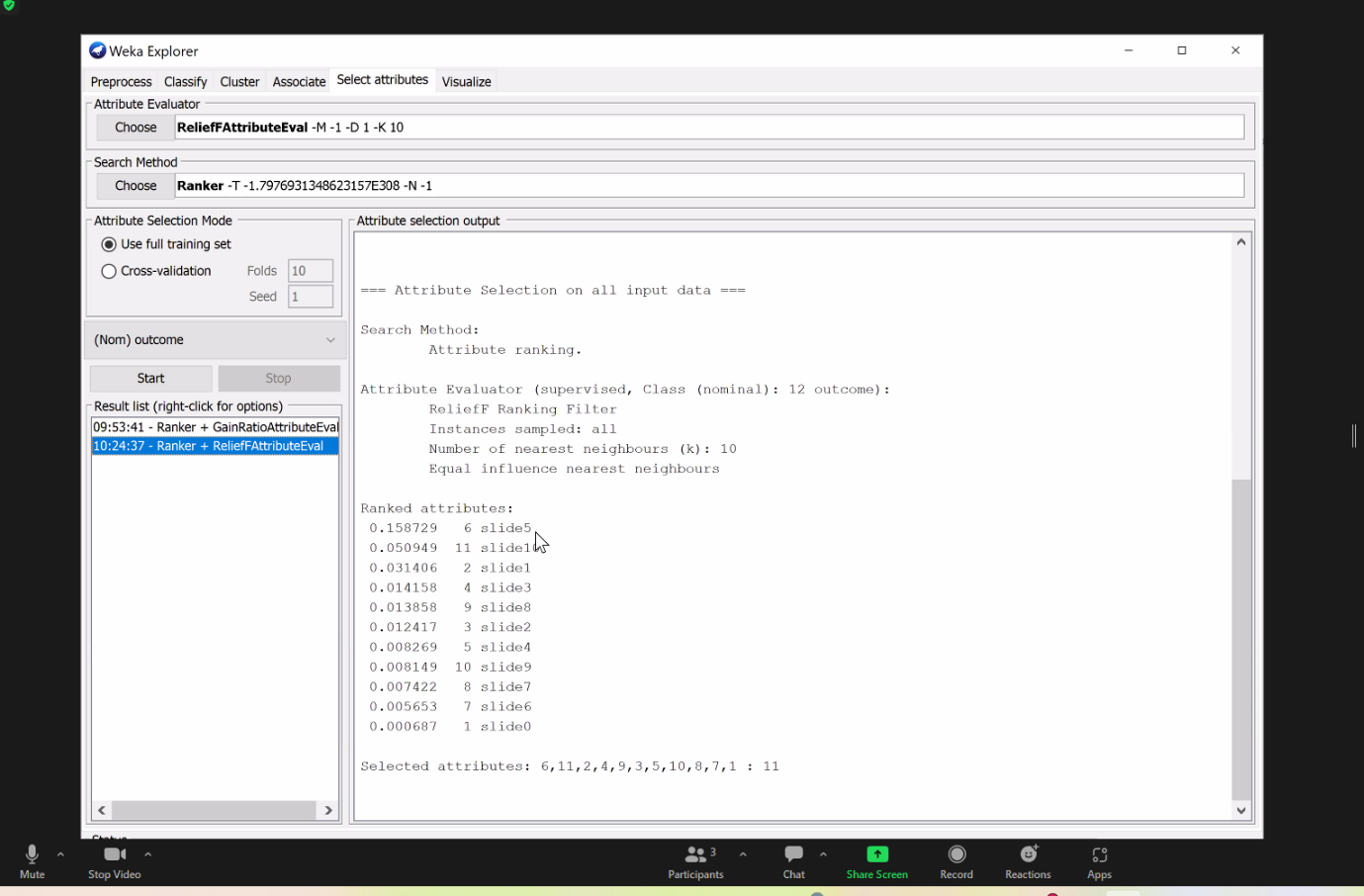
**Things to do**

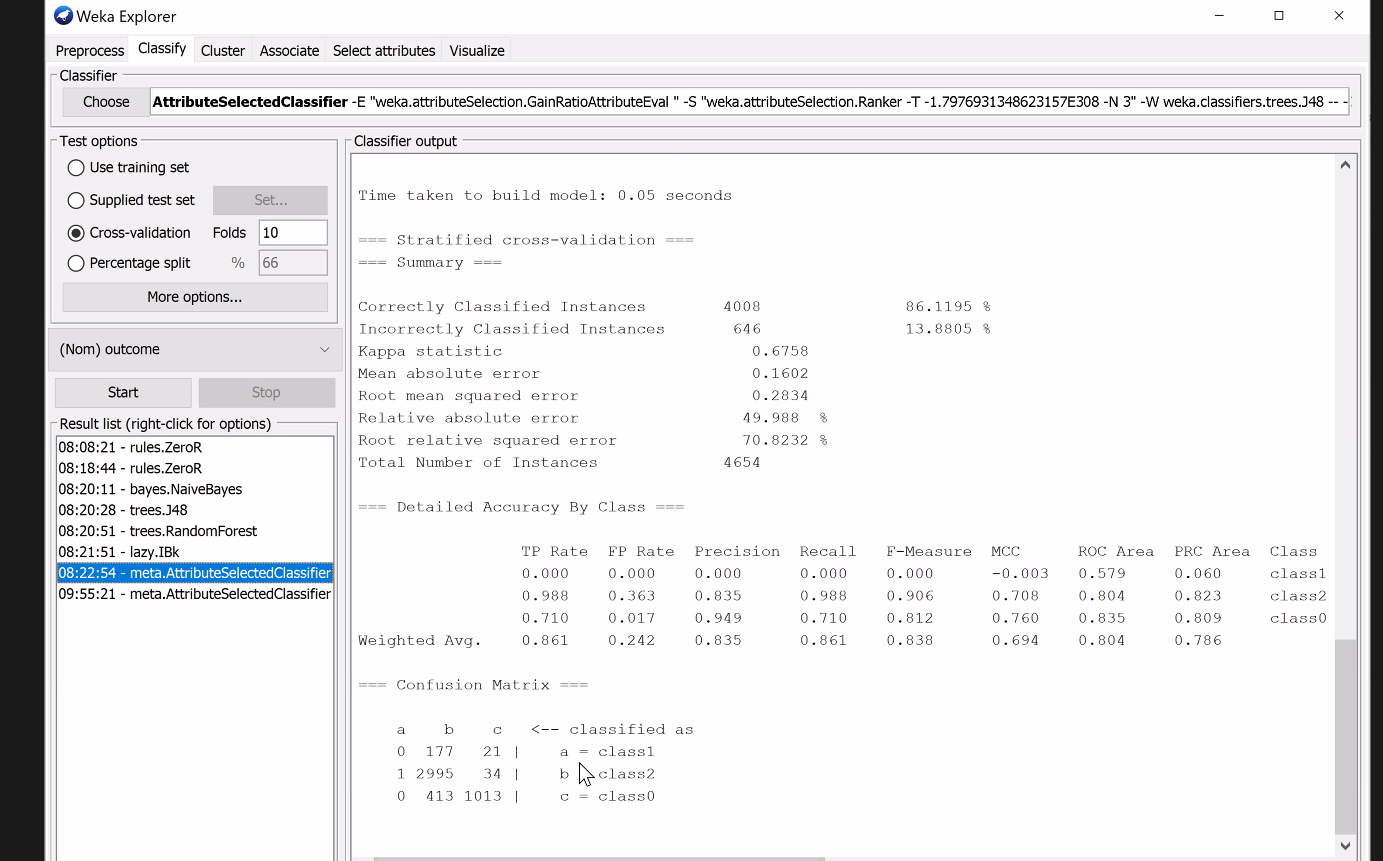
* Paths for failure and success
  + Look at how many students follow the branch or path down a decision tree
  + Useful to predict when someone will fail because success is the majority class
  + Behaviours that you don’t do e.g. completing a problem slide, completing this slide and so on, can have an effect.
  + We can tell students in future cohorts, and explain to them these rules – extract these rules at the very end, and make them user friendly.
* Weka
  + Note: Weka has a good suite of feature selection algorithms that are less available on Python
  + Irena’s results
    - We can build a simple decision tree with 2 levels (better than baseline)
    - Slides 0, 4, 5, 10
      * Slide 0 is an interesting differentiator – if they don’t visit slide 0, it suggests that they are jumping to the problem directly, which suggests that they might be very confident.
      * Slide 4 is also a good predictor. The slide is about syntax errors.
      * Slide 5 is a problem slide, and hence them attempting it does affect the last slide. Passing this first problem slide is more determinative of module performance than completing interactive slides.
      * Slide 10 is about variable assignment.
    - Caveats
      * Note: For educators, this does NOT mean that we should throw away slides that receive less feature importance. They can still be important.
      * Applications
        + Identify pathways for success and failure
        + Help design better content
    - Procedure
      * Ran information gain
  + To do
    - Categorise and label everything – no submission instead of 0
    - Prune trees based on % of students going down a set path
    - Identify important modules, and the important slides within those modules and send them to Grok for validation
    - Accuracy rates might not be best measure for end user. We want to focus more on failure rates. How well is the model in predicting failure.
    - Run Weka analysis on all challenges.
    - Validity of rules analysis
    - Run IG and CFS
      * Note:

Thesis has two main conclusions:

1. Overall course
   1. First week matters. We can predict dropout based on performance and behaviours in first week.
   2. This is important to educators to tell them to intervene early, and for students to focus on developing foundational skills – give a few examples.
2. Pathways to failure and success
   1. Shows educators when to intervene and how to intervene – guide students from paths of failure to paths of success
3. Conference paper
4. End product that could be useful for educators







Doing very badly for class 1, but do very well for passing and pretty well for failure, sometimes misclassify people who don't do well as passing

Look at distribution of classes in weka

We don't have many class 1 - unsuccessful submission, most people if they attempt will eventually pass

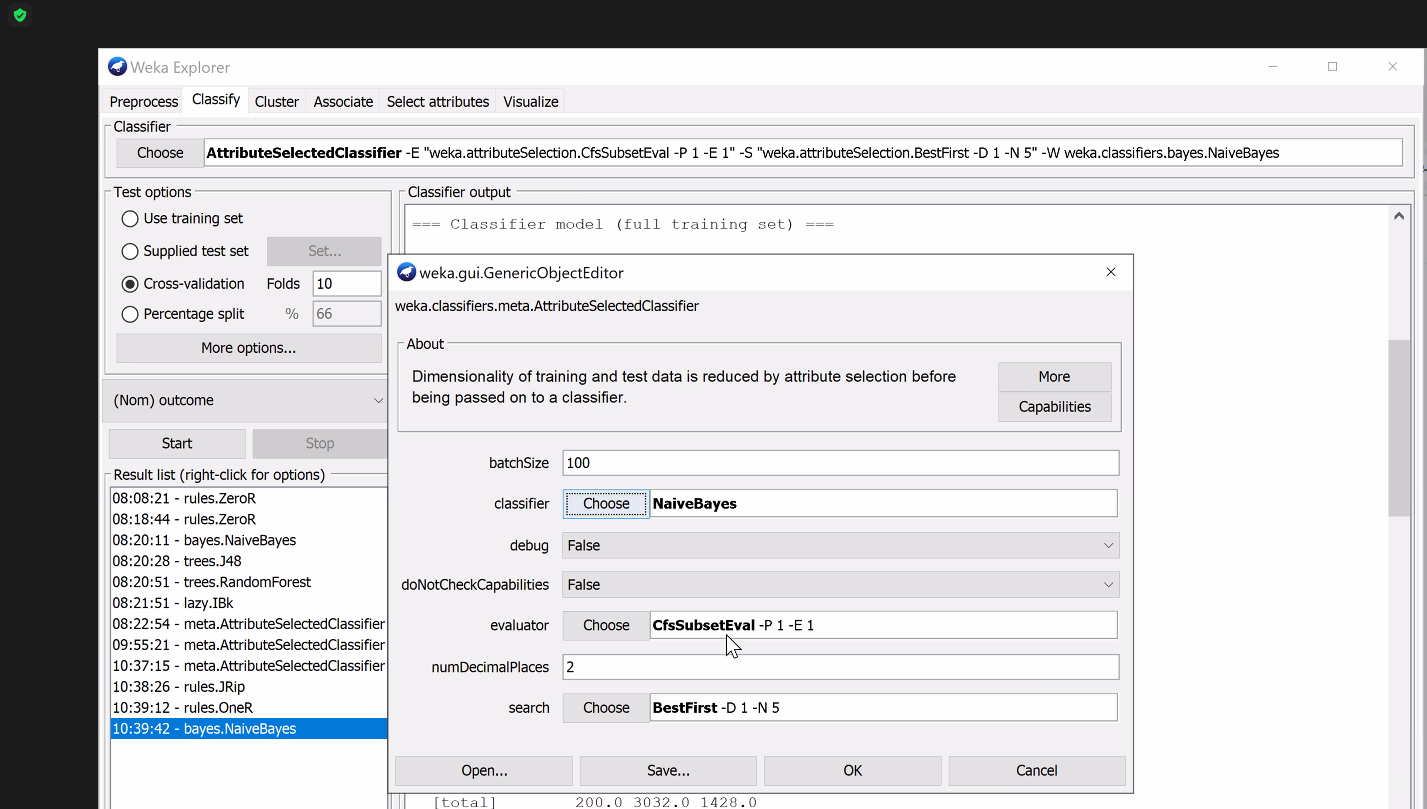
Distrinution should change across weeks - do other weeks so we can provide interesting feedback

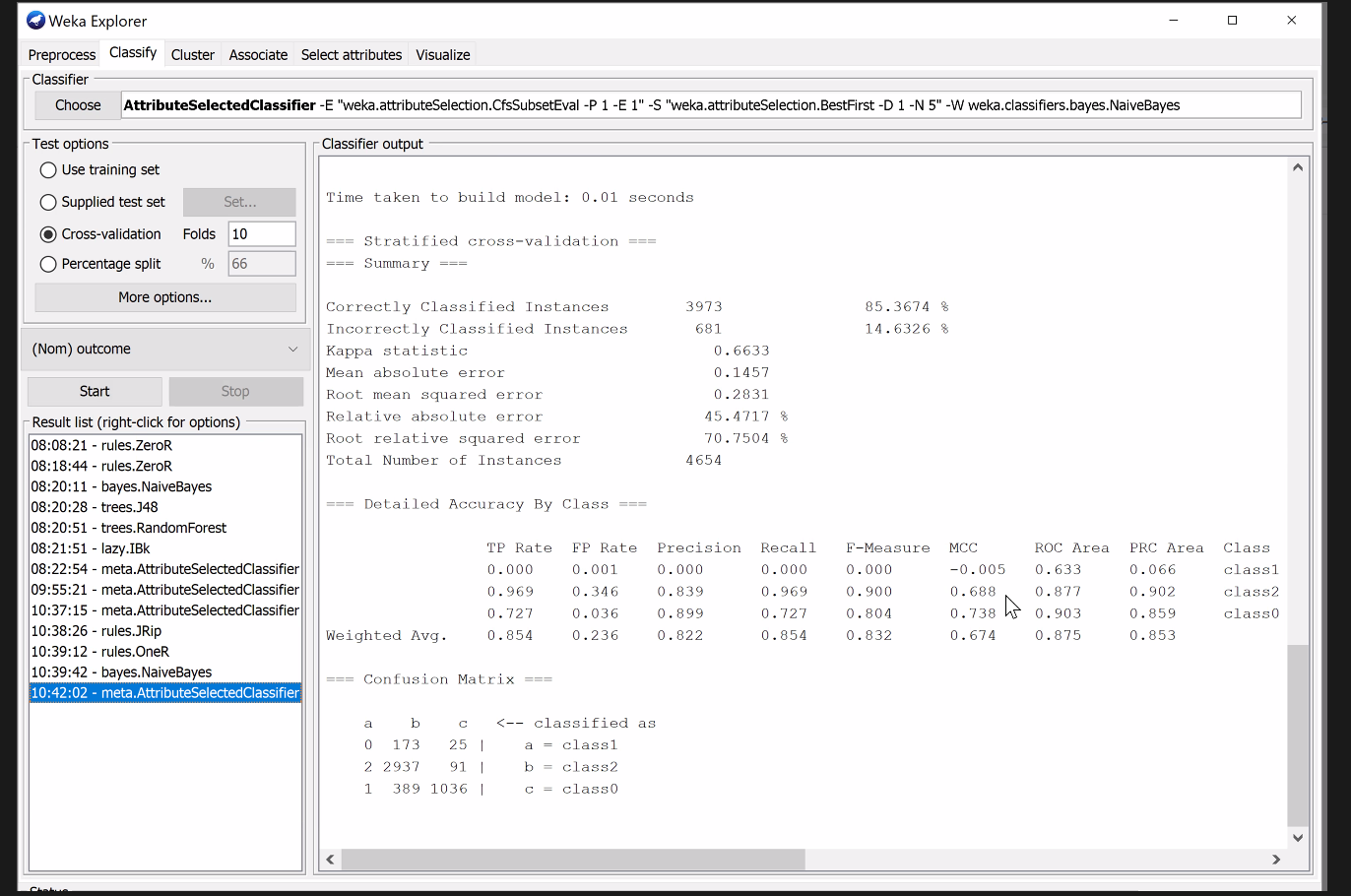
Can run rule-based algorithm like PART, JRIP, OneR

kNN, random forest does well - use statistics to understand why NB doesn't do well

NB - most attributes are strongly correlated which is why NB might not work

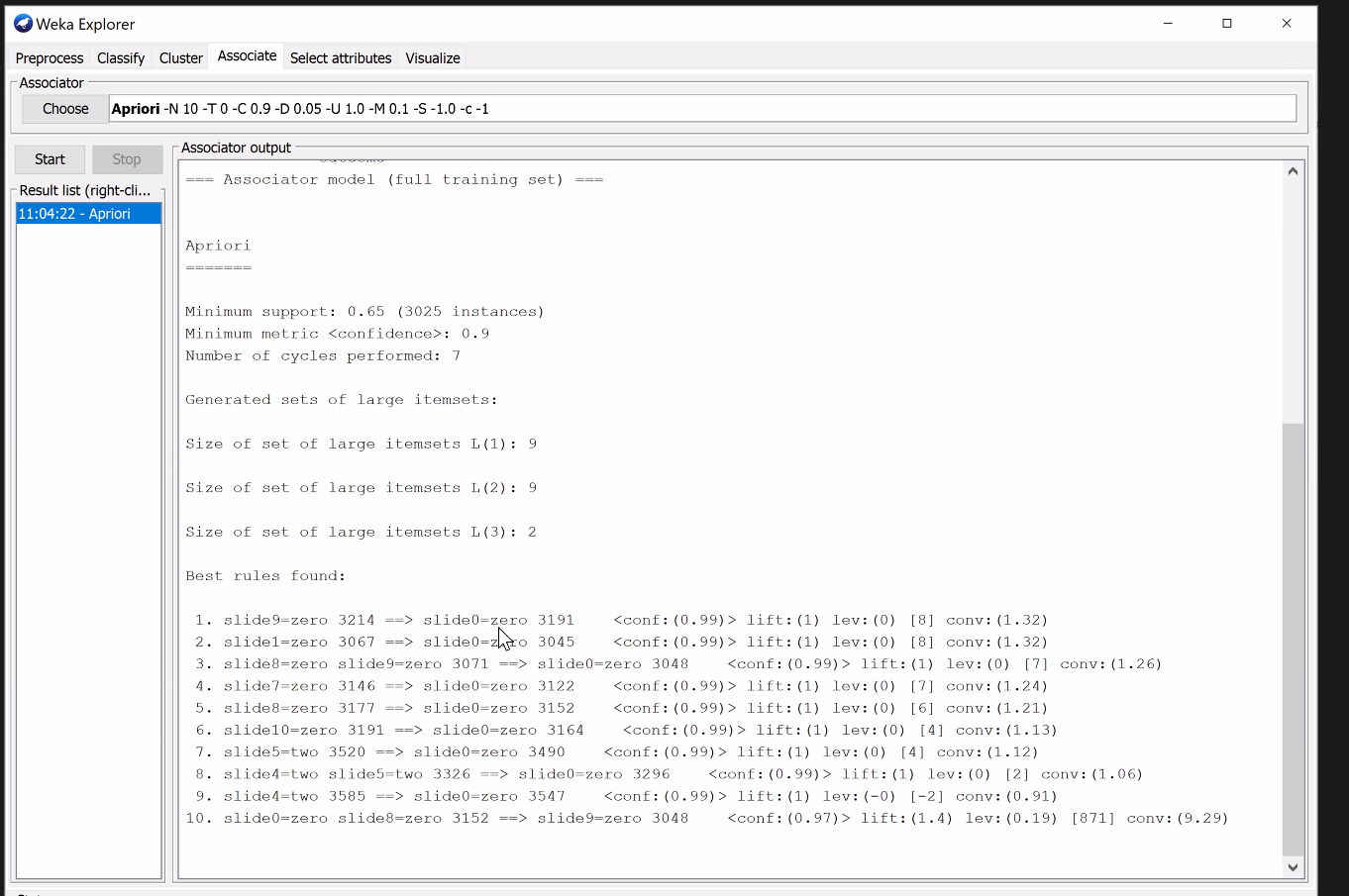
But then try NB with feature selection - meta - attributeselectedclassifier





This says more about data than algorithm - pretty much 85% achievable as long as you don't have correlated variables - use 3 slides, slide 0, 5, 10

STICK WITH DECISION TREE + CFS + IG - trees are very short



Procedure

1. Use select attributes to select a subset of useful attributes. Select attributes based on the decrease in accuracy – after 3 attributes, accuracy decreases a lot, so choose the top 3 etc.
2. Run decision tree algorithm using meta Attribute Selected Classifier – CFS – using those specified attributes. Note: Default is numattributes is -1, change this to 3 e.g. from above.
3. Run information gain.