CENTENNIAL COLLEGE

Predicting Students At-Risk Using Demographic and eCentennial Data

RISES 2018 Research, Innovation, Scholarship and Entrepreneurship Summit

CENTENNIAL COLLEGE

- **ILIA NIKA**, Ph.D., Software Programs Coordinator, ICET/SETAS, Centennial College
- **Huizi Zhao**, Director, Institutional Research Office, Centennial College
- **Paul Armstrong**, Ph.D., Institutional Research Analyst, Centennial College
- **Yun Kui Pan**, Machine Learning Developer, Software Engineering Technology Student

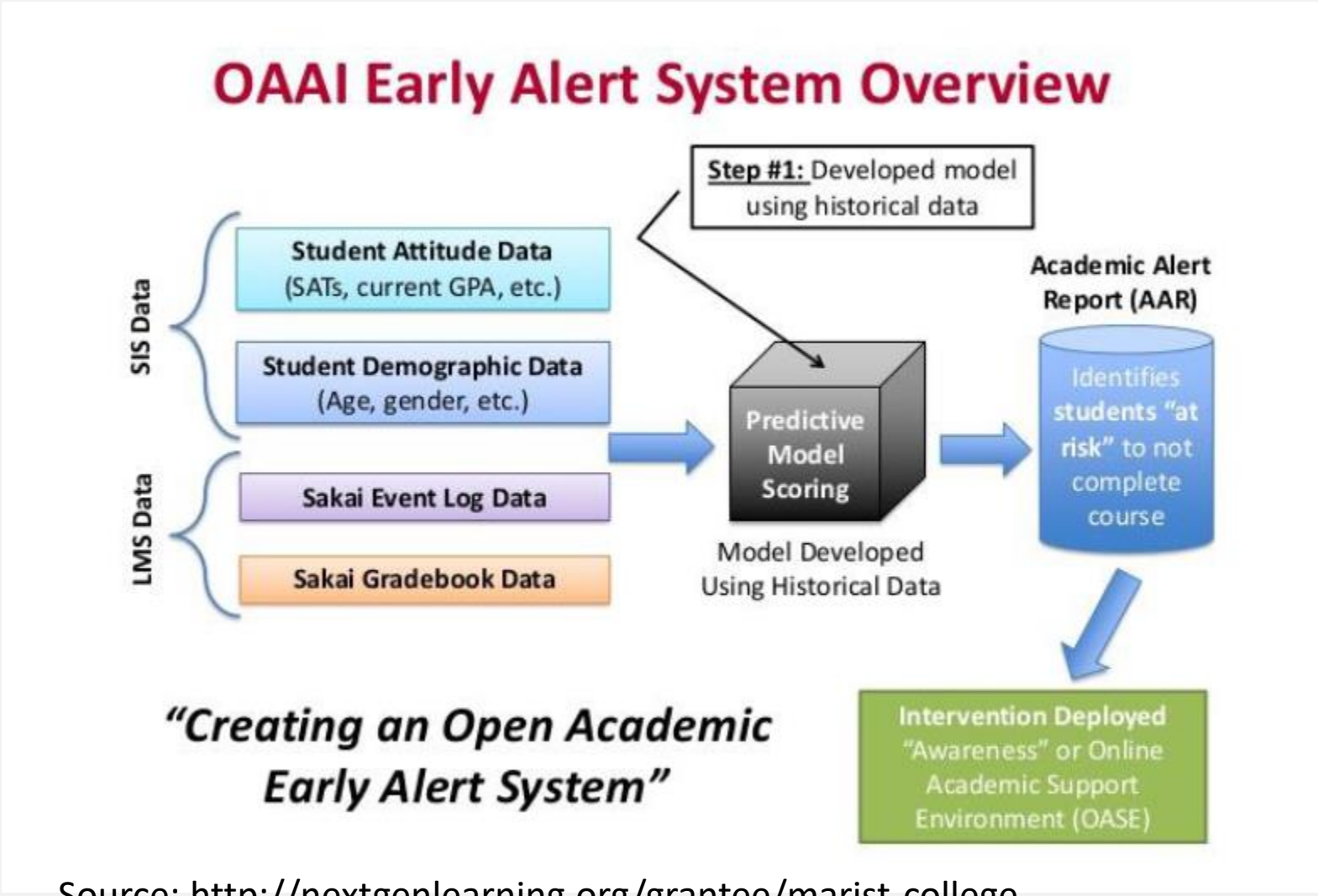
Predicting Students At-Risk Using Demographics and eCentennial Data

- ❑ Project goals
- ❑ Research Methodology
- ❑ Prediction Model
- ❑ Evaluation and Results

Project Goals

- Provide a tool and the metrics for the ongoing monitoring of the Centennial Advising and Pathways Services (CAPS) model efficacy
  - use machine learning to predict students at-risk based on demographic and eCentennial data
  - develop the necessary models and an application that based on demographic and eCentennial data, predicts students at–risk in week 7 with high accuracy

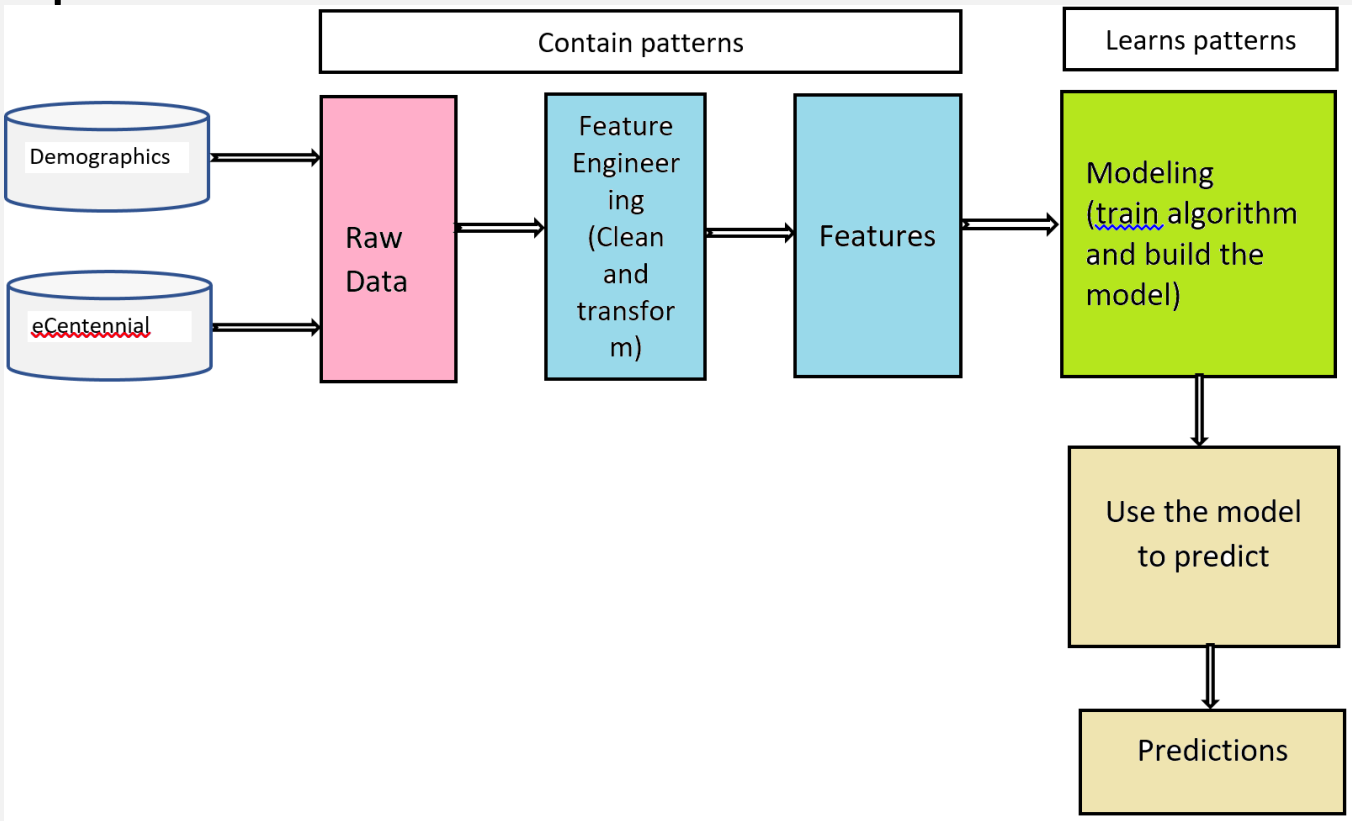
Marist College's Early Alert System



Source: http://nextgenlearning.org/grantee/marist-college

Research Methodology

- Mainly based on the CRISP-DM (Cross Industry Standard Process for Data Mining)
- Our implementation:



Data Preprocessing and Transformation

- Data cleaning
- Data filtering
- Feature Engineering (Calculated variables, scaling, etc.)
- File merging (A special tool was developed for that)
- Friendly input file was created for using the application (to be improved further)

Features

CENTENNIAL COLLEGE

Demographic Data

Student background
13 features

eCentennial Data

LearnerUsage

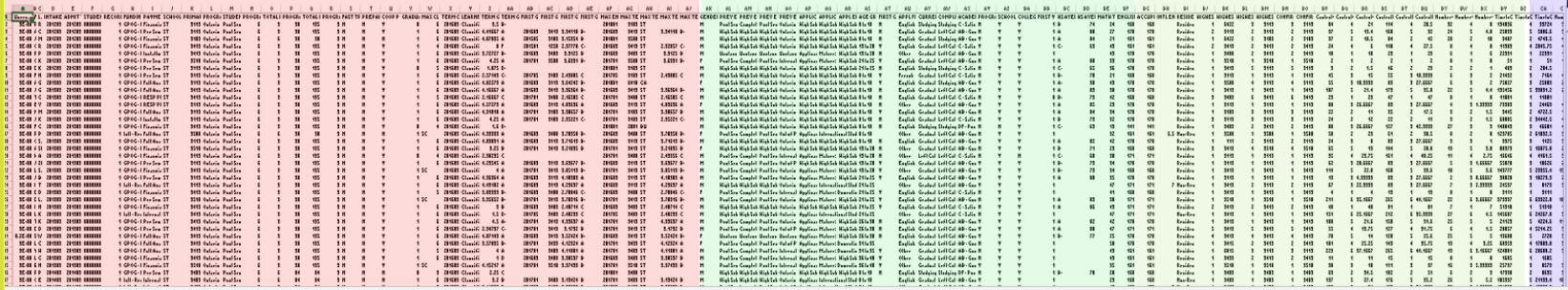
Frequency
5 features

AllGradesExport

Courses
5-8 features

Outcome

Risk Level



Features

1. FUNDING SOURCE NAME
2. COOP PROGRAM IND
3. GENDER
4. PREV EDU CRED LEVEL NAME
5. AGE GROUP LONG NAME
6. APPL FIRST LANGUAGE DESC
7. HS AVERAGE GRADE
8. MATH TEST SCORE
9. ENGLISH TEST SCORE
10. RESIDENCY STATUS NAME
11. CONFIRM PROG CHOICE CODE
12. CONTENTCOMPLETED_AGGAVG
13. CONTENTREQUIRED_AGGAVG
14. NUMBEROFDROPBOXSUBMISSIONS_AGGAVG
15. TIMEINCONTENT_AGGAVG
16. NUMBEROFLOGINSTOTHESYSTEM_AGGAVG
17. COMP100_CCRATE
18. COMP120_CCRATE
19. MATH175_CCRATE

Dependent Variable (Outcome)

CENTENNIAL COLLEGE

Risk Level

- Computed variable
- Based on the number of courses student failed

Algorithms Tested To Build The Prediction Model

- Single Model:
  1. Logistic Regression
  2. SMO
  3. LibSVM
  4. NaiveBayes
  5. J48
  6. Voted Perceptron
  7. Multi Layer Perceptron
- Ensemble Model:
  1. Bagging
  2. Randomization
  3. Boosting
  4. Stacking

Experiments

- Data:
  - ICET, Software programs
  - Demographic + D2L
- Train model on 2016F intake data
- Predict students at-risk in 2017F
  - Model Accuracy
  - Factual Accuracy (computed based on the true results in 2017F)

Evaluation of Models for 2016F (124 records: 31 | 93)

Algorithm	Accuracy	Sensitivity(Recall)	Specificity
Randomization	84.21052631578948	0.555555555555...	0.9310344827586...
Boosting	84.21052631578948	0.444444444444...	0.9655172413793...
Bagging	78.94736842105263	0.555555555555...	0.8620689655172...
Stacking	76.3157894736842	0.555555555555...	0.8275862068965...

Algorithm	Accuracy	Sensitivity(Recall)	Specificity
NaiveBayes	84.21052631578948	0.666666666666...	0.8965517241379...
Logistic	81.57894736842105	0.777777777777...	0.8275862068965...
VotedPerceptron	81.57894736842105	0.777777777777...	0.8275862068965...
MultiLayerPerceptron	76.3157894736842	0.555555555555...	0.8275862068965...
LibSVM	73.6842105263158	0.666666666666...	0.7586206896551...
SMO	73.6842105263158	0.555555555555...	0.7931034482758...
J48	71.05263157894737	0.444444444444...	0.7931034482758...

Predictions for 2017F
(175 records - 32 | 143)

Algorithm	Accuracy	Algorithm	Factual
Randomization	84.21052631578948	Randomization	65.63 |
Boosting	84.21052631578948	Boosting	71.88 |
Bagging	78.94736842105263	**Bagging**	**78.13 |**
Stacking	76.3157894736842	Stacking	71.88 |
NaiveBayes	84.21052631578948	NaiveBayes	43.75 |
Logistic	81.57894736842105	Logistic	59.38 |
VotedPerceptron	81.57894736842105	VotedPerceptron	65.63 |
MultiLayerPerceptron	76.3157894736842	MultiLayerPerceptron	59.38 |
LibSVM	73.6842105263158	LibSVM	65.63 |
SMO	73.6842105263158	SMO	59.38 |
J48	71.05263157894737	**J48**	**75.0 |**

Evaluation of Models for 2016F (FailedCourse Threshold=2, 124 records - 20|104)

Algorithm	Accuracy	Sensitivity(Recall)	Specificity
Randomization	97.36842105263158	0.8	1.0
Boosting	97.36842105263158	0.8	1.0
Bagging	94.73684210526316	0.8	0.9696969696969...
Stacking	94.73684210526316	0.8	0.9696969696969...

Algorithm	Accuracy	Sensitivity(Recall)	Specificity
NaiveBayes	97.36842105263158	0.8	1.0
SMO	94.73684210526316	1.0	0.9393939393939...
J48	94.73684210526316	0.8	0.9696969696969...
MultiLayerPerceptron	94.73684210526316	0.8	0.9696969696969...
Logistic	86.84210526315789	0.8	0.8787878787878...
LibSVM	84.21052631578948	1.0	0.8181818181818...
VotedPerceptron	84.21052631578948	1.0	0.8181818181818...

Predictions for 2017F, thr2 (FailedCourse Threshold=2, 175 records - 14 | 161)

Algorithm	Accuracy
Randomization	97.36842105263158
Boosting	97.36842105263158
Bagging	94.73684210526316
Stacking	94.73684210526316
NaiveBayes	97.36842105263158
SMO	94.73684210526316
J48	94.73684210526316
MultiLayerPerceptron	94.73684210526316
Logistic	86.84210526315789
LibSVM	84.21052631578948
VotedPerceptron	84.21052631578948

Algorithm	Factual
Randomization	78.57 |
Boosting	**85.71 |**
Bagging	**85.71 |**
Stacking	78.57 |
NaiveBayes	28.57 |
SMO	**85.71 |**
J48	78.57 |
MultiLayerPerceptron	**92.85 |**
Logistic	71.43 |
LibSVM	78.57 |
VotedPerceptron	78.57 |

Conclusions

- We can predict students at risk, based on demographic & eCentennial data with reasonable accuracy.

Future Work

CENTENNIAL COLLEGE

1. Creating UI for input file
2. Testing models with all schools and programs
3. Improving overall friendliness of the application
4. Automating the feature selection

Acknowledgement

- This work was partially supported by:
  - ARIF funding
  - Student Services funding
- Centennial College CPO provided the datasets used in the research
- ICET, SETAS

Questions

CENTENNIAL COLLEGE