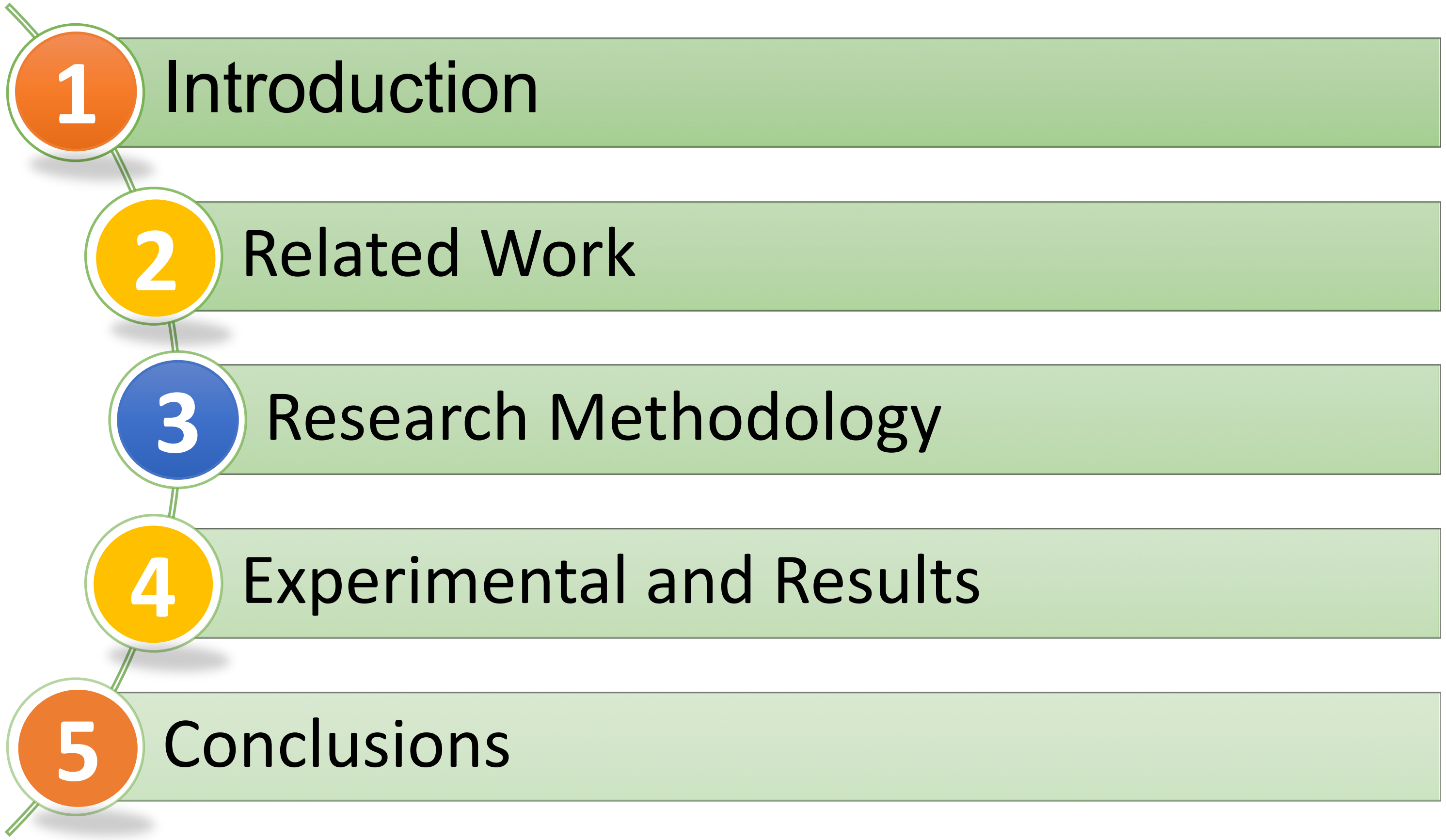
Improving Student Academic Performance Prediction Models using Feature Selection

Wongpanya Nuankaew and Jaree Thongkam

Mahasarakham University, Thailand

Outline



1. Introduction
2. Related Work
3. Research Methodology
4. Experimental and Results
5. Conclusions

Introduction

- In the last two decades, data mining techniques have played **an important role in the analysis of educational data for the development an efficient educational system, and to manage teaching and learning to achieve maximum academic results.**

- The application of education data mining to predict student achievement is a concrete concept for improving student quality, which is gaining widespread attention. For this reason, many researchers have proposed a predictive **function to improve the quality of learning and student performance.**

- The information in the education system consists of both **quality data and poor-quality data**. Therefore, **the preprocessing data will help eliminate noise data for pattern learning and increase model performance**. In addition, the selection features and the elimination of noise from the data will be of help in cleaning data.

- This research proposes methods **to improve student academic performance prediction models and compares the performance of prediction methods.**

Related Work

Predicting Academic Performance

2014

- **Aziz** proposed academic performance prediction models for Bachelor of Computer Science students' at Universiti Sultan ZainalAbidin. Their result showed that Decision Tree and Rule-Based provided better accuracy than Naïve Bayes.

2017

- **Costa** presented introductory programming courses available at a Brazilian Public University. The preprocessing used the Information Gain algorithm to selected attributes and the Synthetic Minority Oversampling Technique (SMOTE) for class imbalanced problems. Their results showed that Support Vector Machine performed better than the other methods.

2020

- **Sokkhey and Okazaki** presented a hybrid approach of principal component analysis (PCA) to solve the misclassification problem for improving the prediction academic performance using feature extraction. Their models provided very satisfactory results.

Related Work

Feature Selection

2011

- **Smith and Martinez** proposed the PRISM (Preprocessing Instances that Should be Misclassified) method that identified and removed instances that should be misclassified. PRISM helped the classification accuracy up to 1.3% and the non-outlier classification accuracy up to 1.9%.

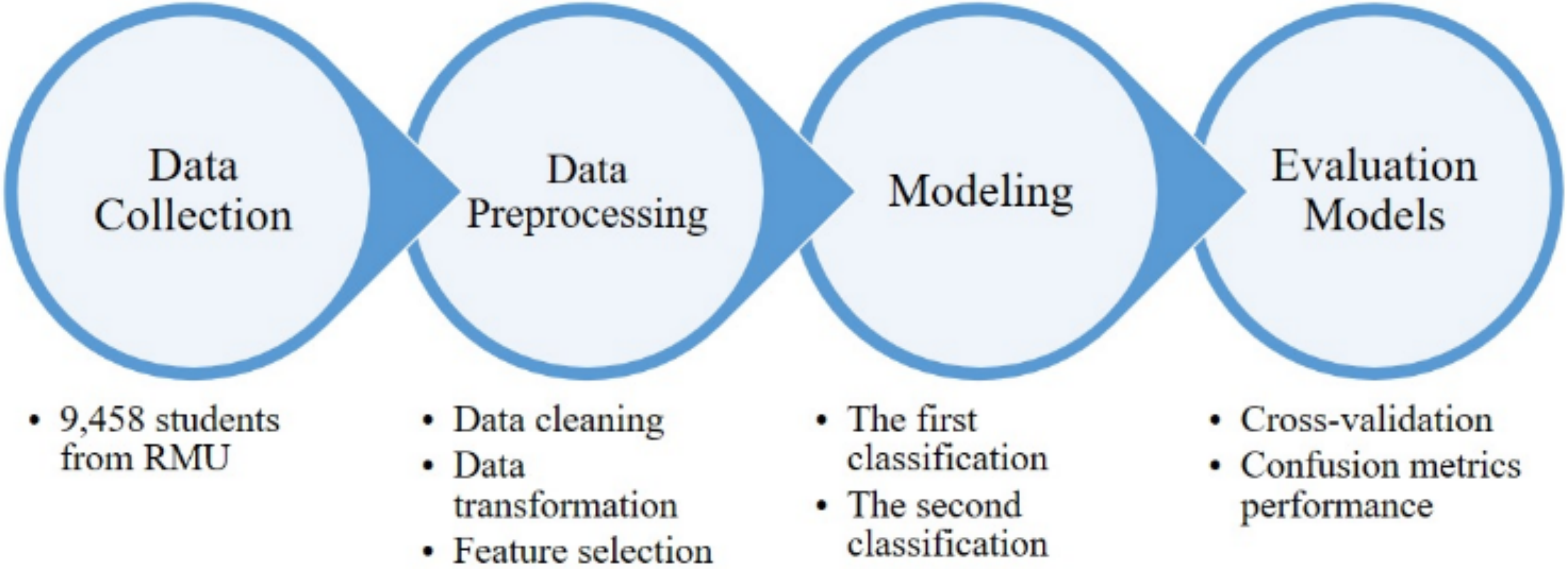
2018

- **Geetharamani** proposed to identify and remove the misclassified instances applied to the Best First Tree (BFTree) for improving classification accuracy and compared it to eighteen classification methods. Their results showed that the best accuracy, up to 32.5%.

2018

- **Abdrabo** used Fuzzy Rough Nearest Neighbor to remove the misclassified instances for improving the classification performance. Their model helped to increase the classification accuracy to 89.2%.

Research Methodology



Data Collection

- 9,458 students from RMU

Data Preprocessing

- Data cleaning
- Data transformation
- Feature selection

Modeling

- The first classification
- The second classification

Evaluation Models

- Cross-validation
- Confusion metrics performance

Research framework

Research Methodology

Data Collection

- The data collected were 9,485 students from Rajabhat Maha Sarakham University during 2015-2018 which include demographics, environment, and GPA.
- Three GPA class are low (2.00-2.74), medium (2.75-3.24), and good (3.25-4.00).
- Two datasets are dataset A and dataset B.

Research Methodology

Data Preprocessing

- Data cleaning
- Data transformation
- Feature selection :
  - Gain ratio
  - C4.5 decision tree

Research Methodology

Modeling

- Naïve Bayes (NB)
- Sequential Minimum Optimization (SMO)
- Artificial Neural Network (ANN)
- K-nearest Neighbor (KNN)
- Reduced Error Pruning Tree (REPTree)
- Partial Decision Trees (PART)
- Random Forest (RF)

Research Methodology

Modeling

- **The first classification part:**
  - Comparison the classification methods predict academic performance.
- **The second classification part:**
  - C4.5 decision tree is applied for identifying the misclassified instances. These instances are removed to clean the dataset that resulted from dataset A in 6,272 and dataset B in 6,281 clean instances.
  - Removing the misclassified instances from the dataset makes the class imbalance which is solved using the Synthetic Minority Over-sampling Technique (SMOTE) to tackle the class imbalance.
  - Comparison the classification methods predict academic performance.

Research Methodology

Evaluation Models

- 10-Fold Cross Validation
- Confusion Metrics Performance of the model, including Precision, Recall, and F-Measure

Experimental and Results

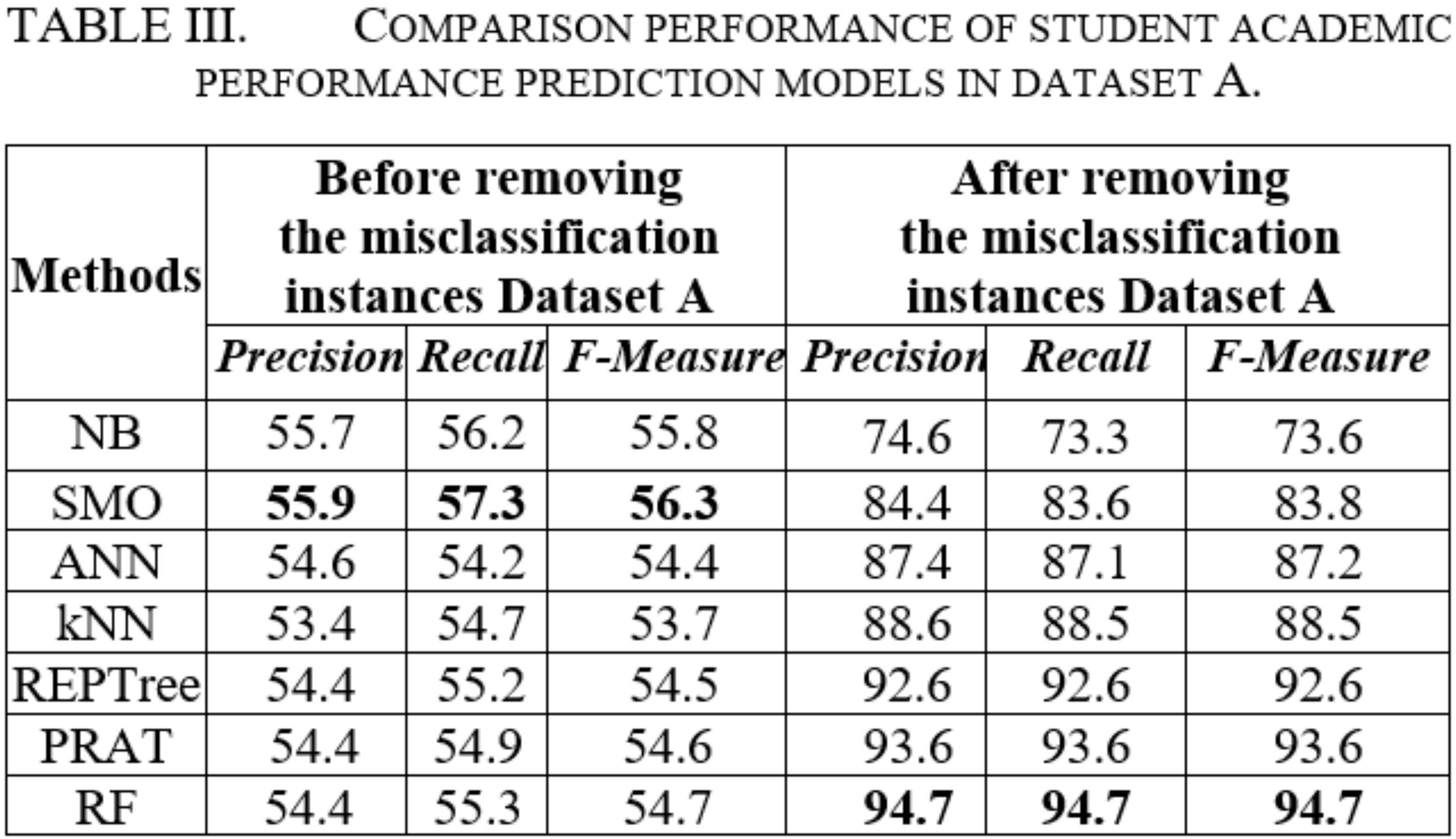


TABLE III. COMPARISON PERFORMANCE OF STUDENT ACADEMIC PERFORMANCE PREDICTION MODELS IN DATASET A.

Methods	Before removing the misclassification instances Dataset A			After removing the misclassification instances Dataset A		
	Precision	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*
NB	55.7	56.2	55.8	74.6	73.3	73.6
SMO	**55.9**	**57.3**	**56.3**	84.4	83.6	83.8
ANN	54.6	54.2	54.4	87.4	87.1	87.2
kNN	53.4	54.7	53.7	88.6	88.5	88.5
REPTree	54.4	55.2	54.5	92.6	92.6	92.6
PRAT	54.4	54.9	54.6	93.6	93.6	93.6
RF	54.4	55.3	54.7	**94.7**	**94.7**	**94.7**

Experimental and Results

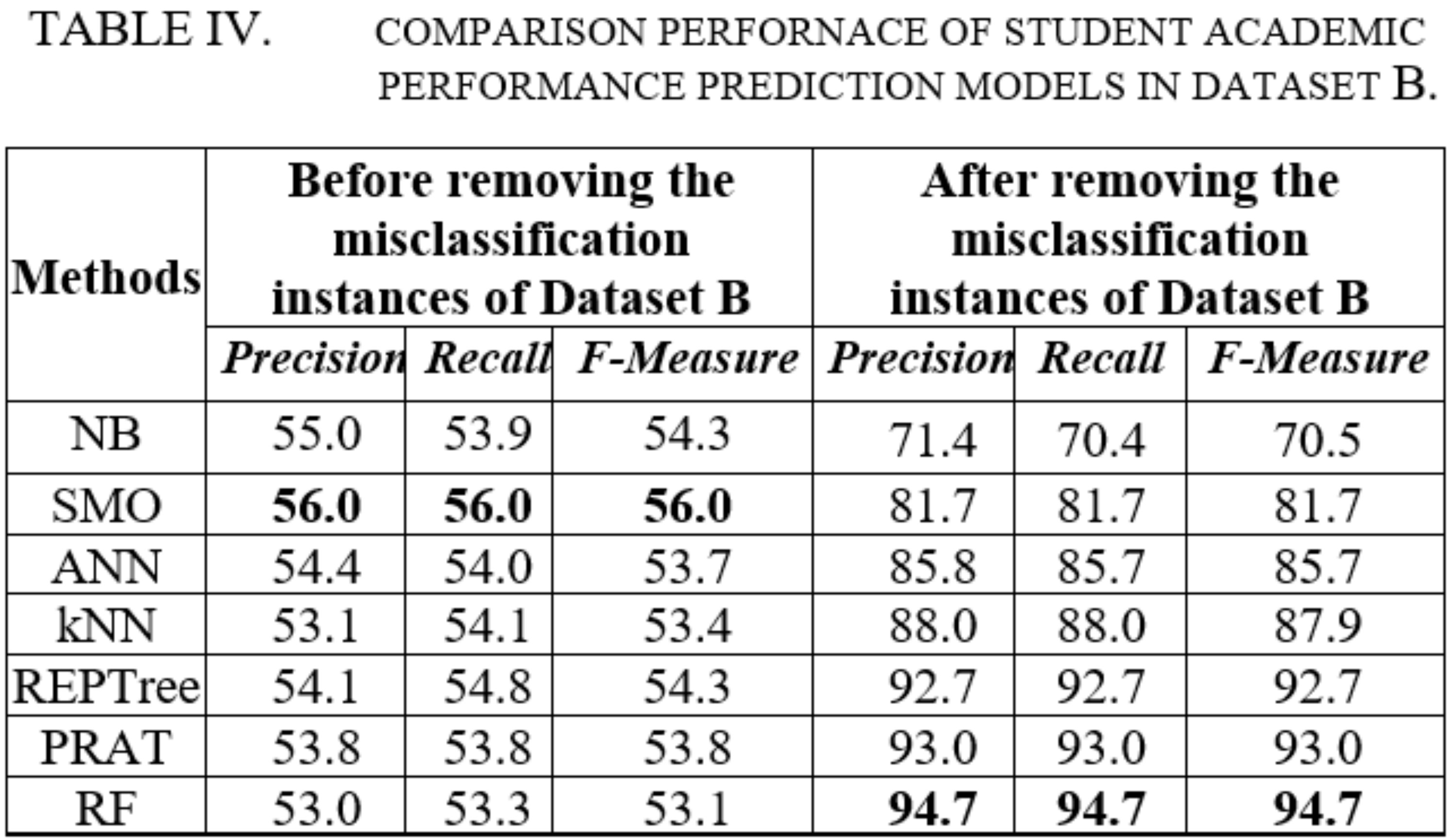


TABLE IV. COMPARISON PERFORNACE OF STUDENT ACADEMIC PERFORMANCE PREDICTION MODELS IN DATASET B.

Methods	Before removing the misclassification instances of Dataset B			After removing the misclassification instances of Dataset B		
	Precision	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*
NB	55.0	53.9	54.3	71.4	70.4	70.5
SMO	**56.0**	**56.0**	**56.0**	81.7	81.7	81.7
ANN	54.4	54.0	53.7	85.8	85.7	85.7
kNN	53.1	54.1	53.4	88.0	88.0	87.9
REPTree	54.1	54.8	54.3	92.7	92.7	92.7
PRAT	53.8	53.8	53.8	93.0	93.0	93.0
RF	53.0	53.3	53.1	**94.7**	**94.7**	**94.7**

Conclusions

- This paper presents to improve the prediction of student academic performance using feature selection.
- C4.5 decision tree was applied for identifying the misclassified instances.
- SMOTE was used to solve the class imbalance problems.
- The result of student academic performance prediction models increase in all classifiers.
- The random forest produce the highest performance.
- This approach significantly improves the student academic performance prediction models with precision up to 41.70%, recall up to 41.40% and F-measure up to 41.60%.
- In future work, researchers plan to consider more factors and other feature selection methods.

Thank you ☺

Questions and Answers