10TH MALAYSIA STATISTICS CONFERENCE
"Looking Beyond GDP: Towards Social Well-being and Environmental Sustainability"
25TH-26TH SEPTEMBER 2023

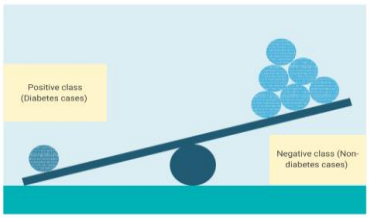
Diabetes Prediction using Synthetic Oversampling Approaches for Imbalanced Classification

Nuryasmin Wahida Binti Hamil; Adilah Abdul Ghapor; Yong Zulina Zubairi

Diabetes Mellitus

A chronic metabolic disorder characterized by elevated blood glucose levels.
It has become a global health concern with a significant impact on public health systems

CLASS IMBALANCED PROBLEM (CIP)



The CIP triggers the data scientist community for decades as one of the major problems in data mining classification process.

OBJECTIVE

To improve the classifier performance in predicting diabetes cases.

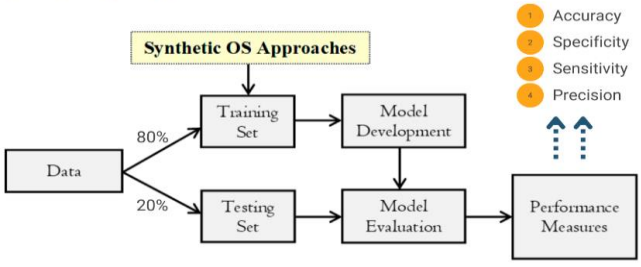
SYNTHETIC OVERSAMPLING

1. **SMOTE**
Generates synthetic data based on the distance between the minority data and the closest minority data therefore the new synthetic data will be formed between the two minority data
Chawla et al. 2002

2. **Borderline-SMOTE**
Only the minority data near the borderline are over-sampled
He et al. 2005

3. **ADASYN**
Generate more synthetic data for observations that are harder to learn than those that are easier to learn for a given model
He et al. 2008

RESEARCH PROSES

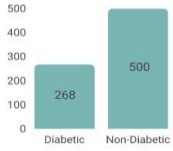


DATA

PIMA INDIANS DIABETES DATABASE

NATIONAL DIABETES AND DIGESTIVE AND KIDNEY DISEASES INSTITUTE

Dataset attributes:
Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, Age and Outcome



Sample: 768 females with at least 21 years old

CONCLUSION

Synthetic OS approaches has shown promise in addressing the challenges posed by imbalanced datasets in diabetes prediction.
ADASYN is the best synthetic OS approach in predicting diabetic cases since sensitivity is looking on the proportion of the total number of correctly classified under minority (positive) class.

RESULTS

CLASS DISTRIBUTION ON TRAINING SET

	Positive	Negative
Without SMOTE	213	401
SMOTE	401	401
ADASYN	380	401
Borderline- SMOTE	401	401

PERFORMANCE MEASURES ON TESTING SET

ACCURACY	SMOTE & ADASYN	74.7%
SPECIFICITY	Without SMOTE	77.8%
SENSITIVITY	ADASYN	76.4%
PRECISION	SMOTE	62.1%