Exploring the Relationship Between Feature Attribution Methods and Model Performance

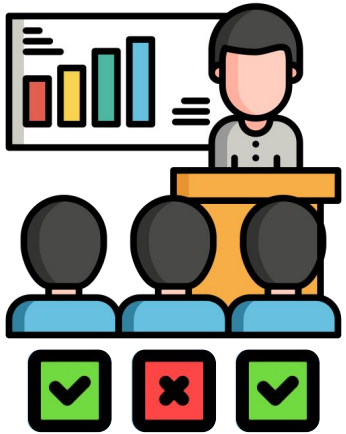
Priscylla Silva
Claudio Silva
Luis Gustavo Nonato

Motivation

Educational Context

ML models

Predictions

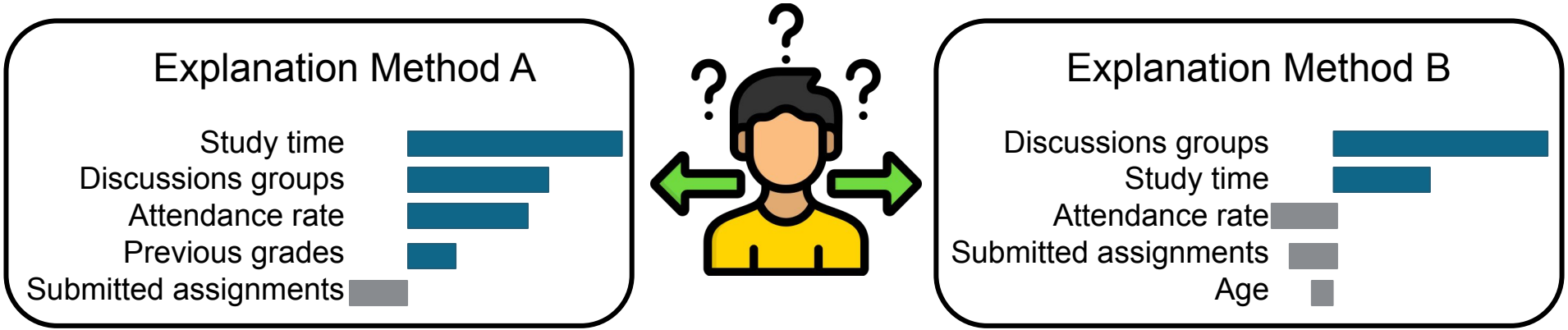


Why?

Image: Flaticon.com

Priscylla Silva (priscylla.silva@usp.br)

Feature Attribution Methods

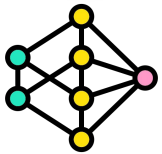
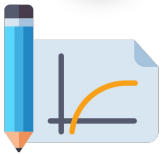
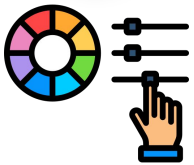


Research Question

Is there a correlation between the model's performance and the disagreement level observed among explanation methods?

Study methodology

1. Two datasets
2. Training a Neural Network
3. AUC Metric
4. 9 Feature attribution techniques
5. Measure (dis)agreement
6. Spearman's rank correlation

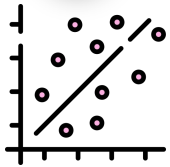
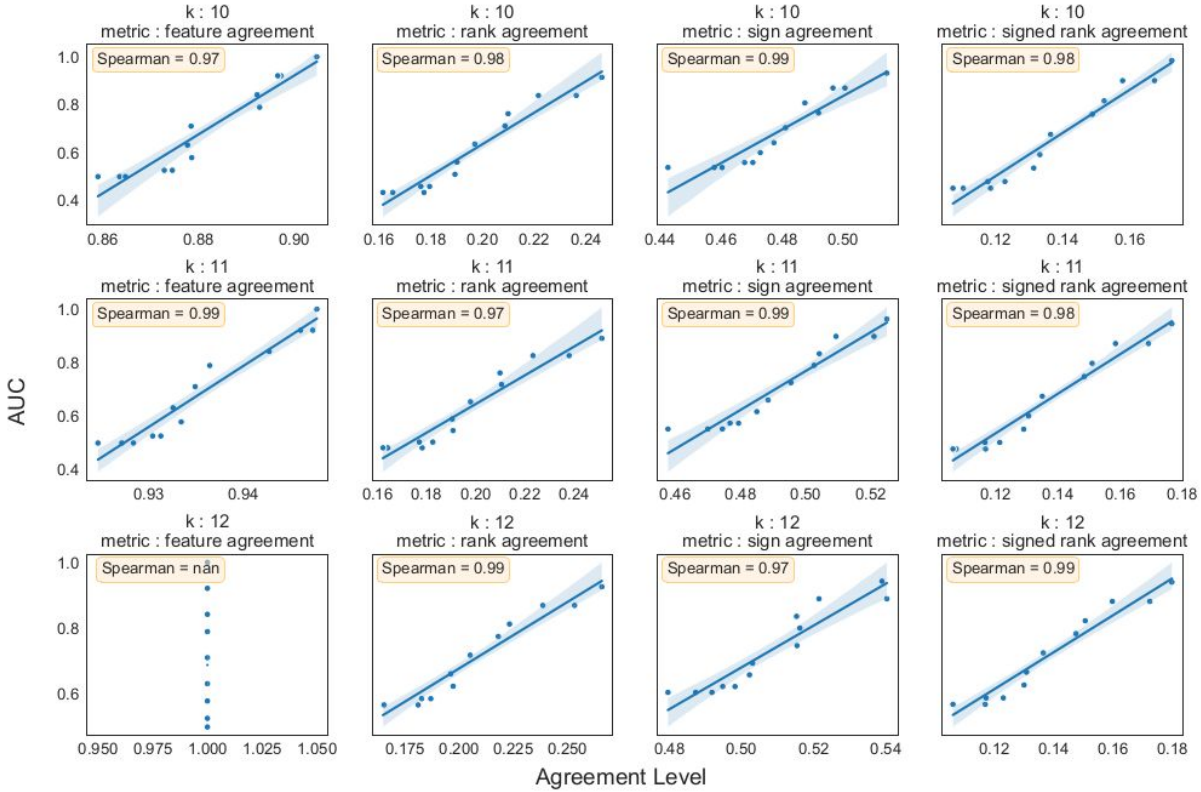


Image: Flaticon.com

Results

Our results show that models with an AUC greater than or equal to 0.8 consistently exhibit the highest levels of agreement among explanation methods.



Exploring the Relationship Between Feature Attribution Methods and Model Performance

Priscylla Silva
Claudio Silva
Luis Gustavo Nonato