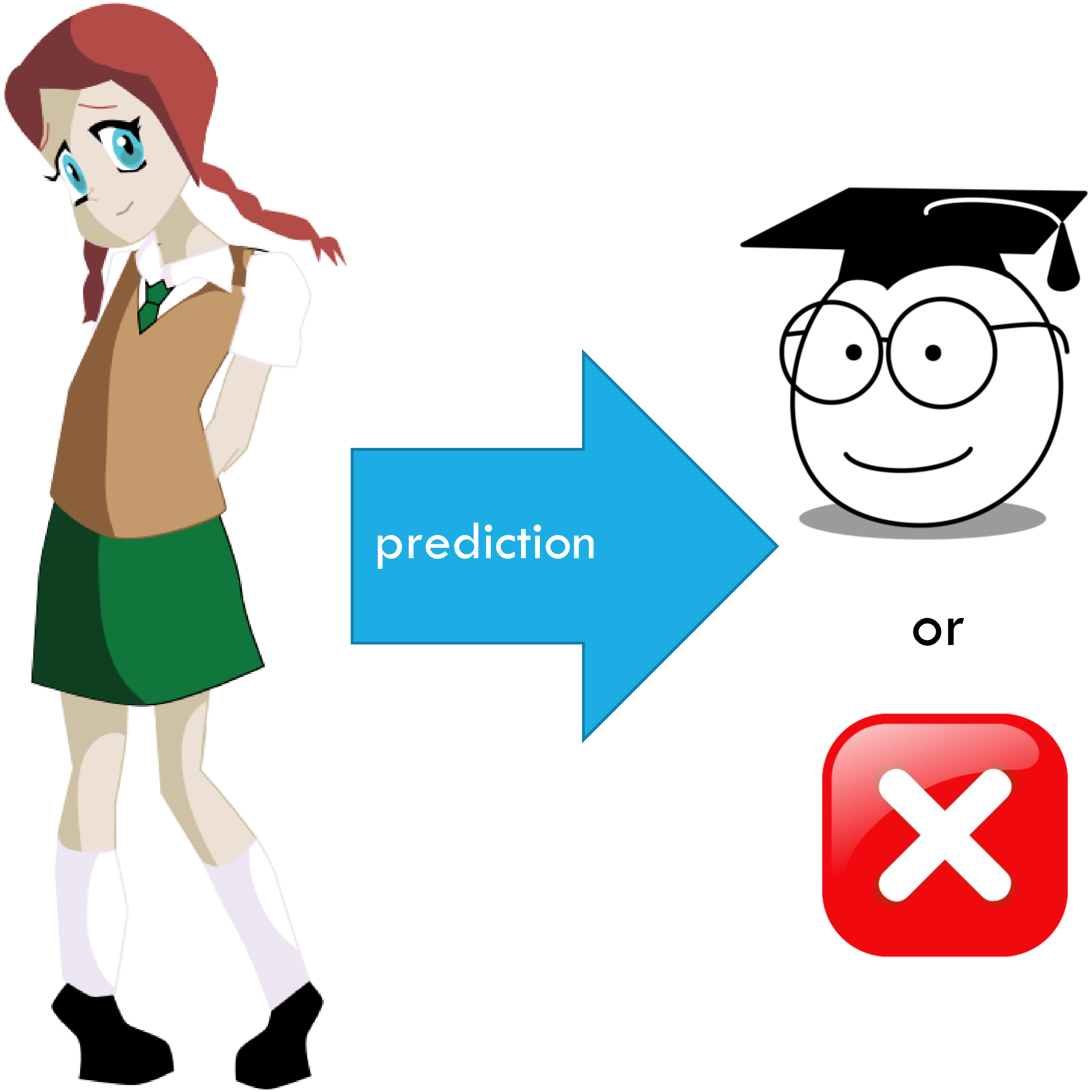
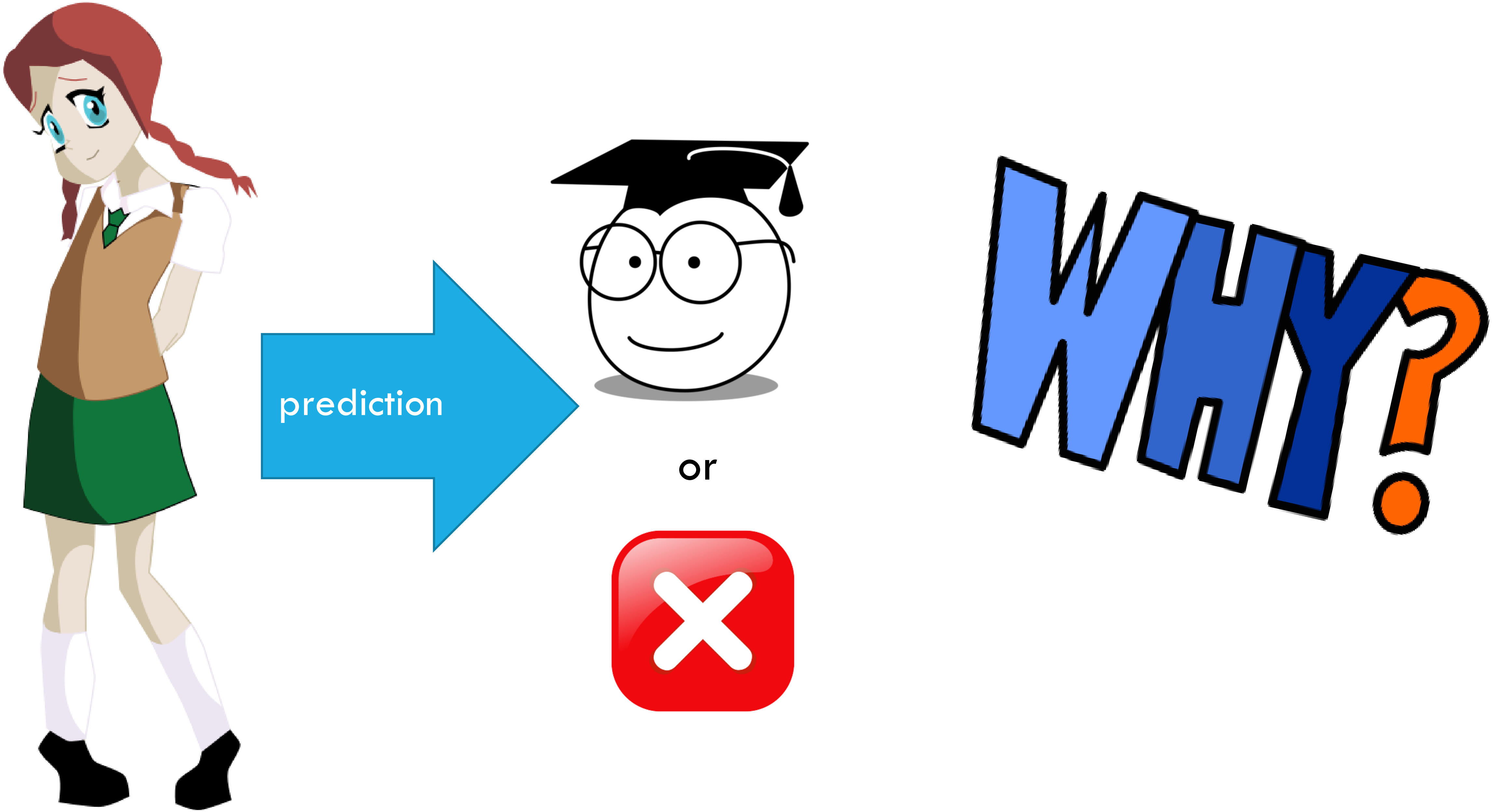
Predicting first-year engineering student success:

from traditional statistics to machine learning

Ramaravind Kommiya Mothilal
Tom Broos
Maarten Pinxten
Tinne De Laet

Tinne.DeLaet@kuleuven.be
@TinneDeLaet


prediction
or


prediction
or
WHY?

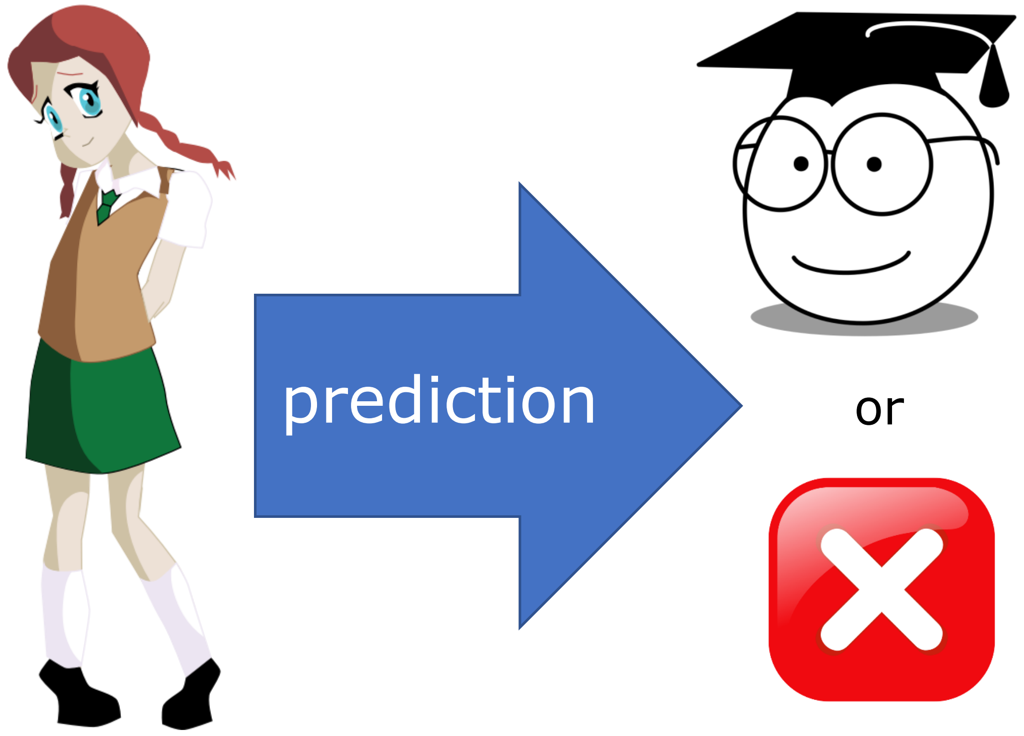
KU LEUVEN
LEUVEN ENGINEERING AND SCIENCE EDUCATION CENTER
STELA

prediction

or

population-wide insights

individual predictions


prediction
or

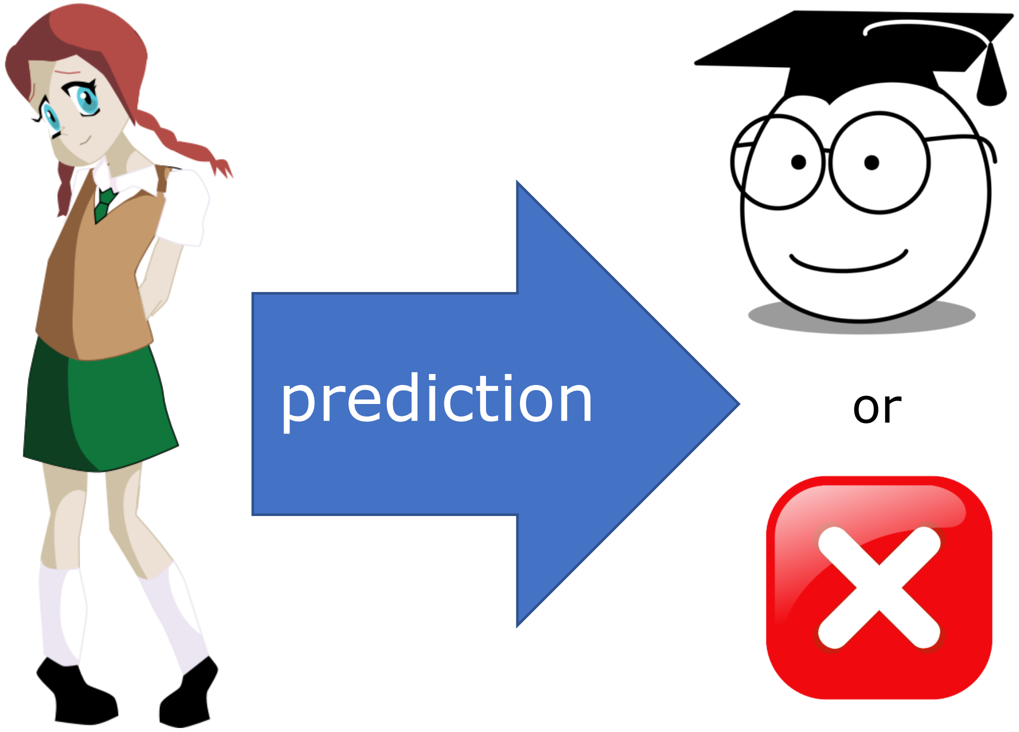
prior academic achievement (secondary education)

- grades math, physics, chemistry
- number of hours math
- effort level

learning and studying skills

- motivation
- time management
- concentration
- performance anxiety
- use of test strategies

preference for time pressure



prior academic achievement (secondary education)

- grades math, physics, chemistry
- number of hours math
- effort level

learning and studying skills

- motivation
- time management
- concentration
- performance anxiety
- use of test strategies

preference for time pressure

prediction

or

academic achievement (AA)
GPA of first semester (wavg)

prior academic achievement (secondary education)
- grades math, physics, chemistry
- number of hours math
- effort level

learning and studying skills
- motivation
- time management
- concentration
- performance anxiety
- use of test strategies

preference for time pressure

prediction

or

academic achievement (AA)
GPA of first semester (wavg)

explanatory modelling
- multiple linear regression

predictive modelling
- logistic regression
- boosted trees

Research questions

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?

Research questions

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?

Can boosted trees more accurately predict first-year student success than logistic regression?

Research questions

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?

Can boosted trees more accurately predict first-year student success than logistic regression?

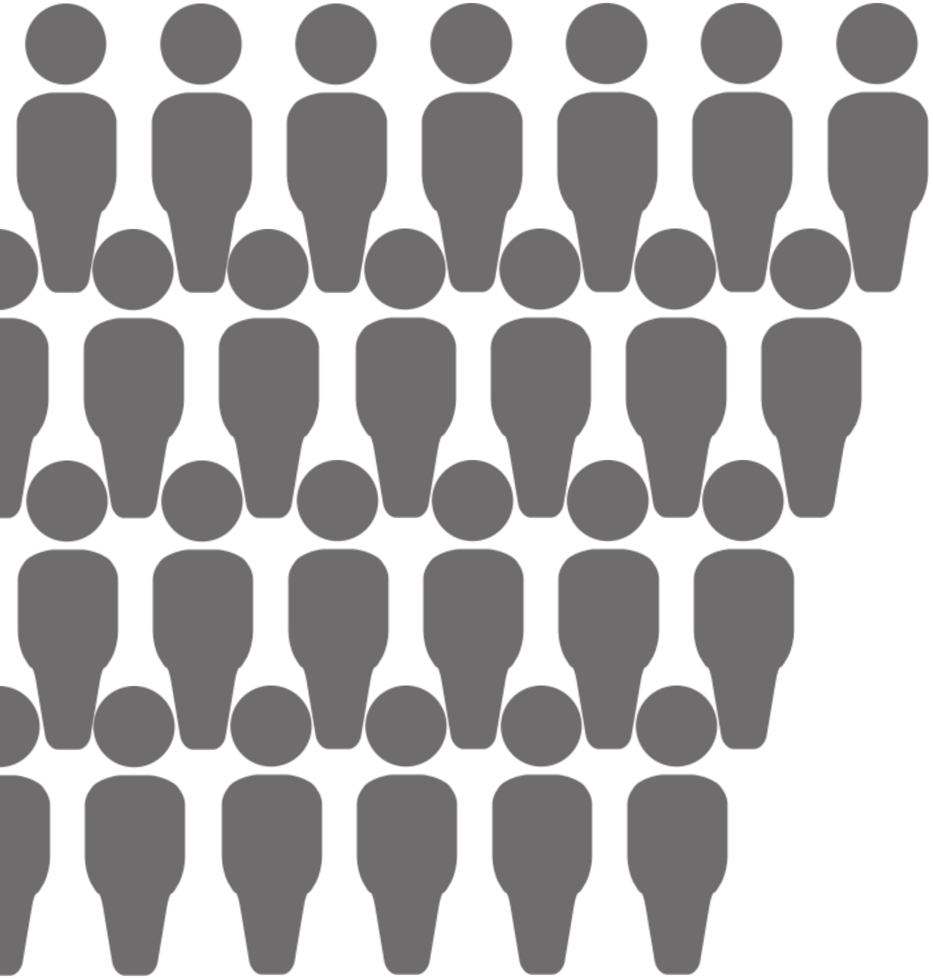
Can Local Interpretable Model-agnostic Explanations (LIME) generate interpretable insights in the factors important for predicting first-year student success?

KU LEUVEN

KU LEUVEN LESEC

LEUVEN ENGINEERING AND SCIENCE EDUCATION CENTER

STELA



first-year Bachelor of Engineering Science students
two academic years: 2015-2016 and 2016-2017
N=811

EXPLANATORY MODELLING
→ MULTIPLE LINEAR REGRESSION

Dependent Variable
Population Y intercept
Population Slope Coefficient
Independent Variable
Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Hypotheses

- Prior academic experience positively AA.
- Affective and goal strategies positively affect AA.
- Preference for time pressure does not affect AA.

EXPLANATORY MODELLING
→ MULTIPLE LINEAR REGRESSION

$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1} \overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Hypotheses

- Prior academic experience positively AA.
- Affective and goal strategies positively affect AA.
- Preference for time pressure does not affect AA.

	model	regression type	R^2
1	wavg ~ math+phy+chem+hrs	standard	0.37
2	wavg ~aff + goal + press	standard	0.06
3	wavg ~ math+phy+chem+hrs + aff + goal + press + eff	sequential	

EXPLANATORY MODELLING
→ MULTIPLE LINEAR REGRESSION

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

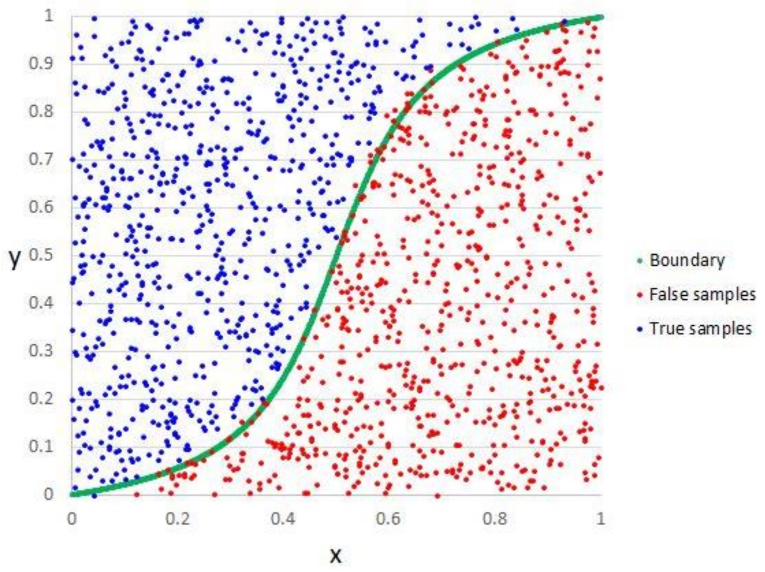
$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Hypotheses

- Prior academic experience positively AA.
- Affective and goal strategies positively affect AA.
- Preference for time pressure does not affect AA.

	model	regression type	R^2
1	wavg ~ math+phy+chem+hrs	standard	0.37
2	wavg ~aff + goal + press	standard	0.06
3	wavg ~ math+phy+chem+hrs + aff + goal + press + eff	sequential	

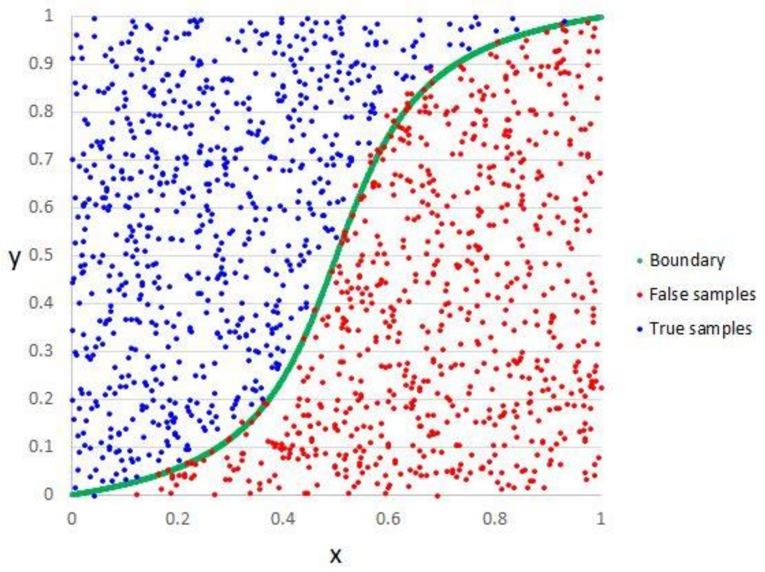
EXPLANATORY MODELLING WITH PREDICTIVE VALIDITY → LOGISTIC REGRESSION



prediction logistic regression

- no-risk (wavg > 11.5)
- moderate-risk
- at risk (wavg ≤ 8.5)

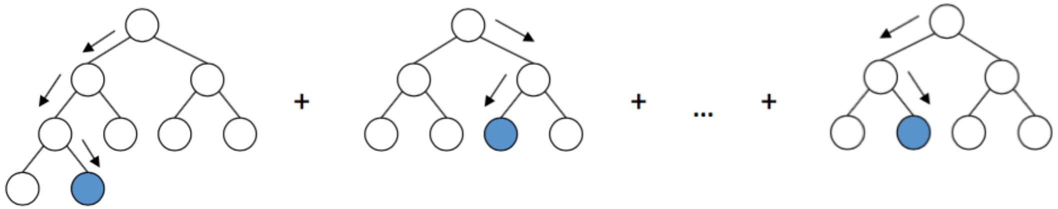
EXPLANATORY MODELLING WITH PREDICTIVE VALIDITY → LOGISTIC REGRESSION

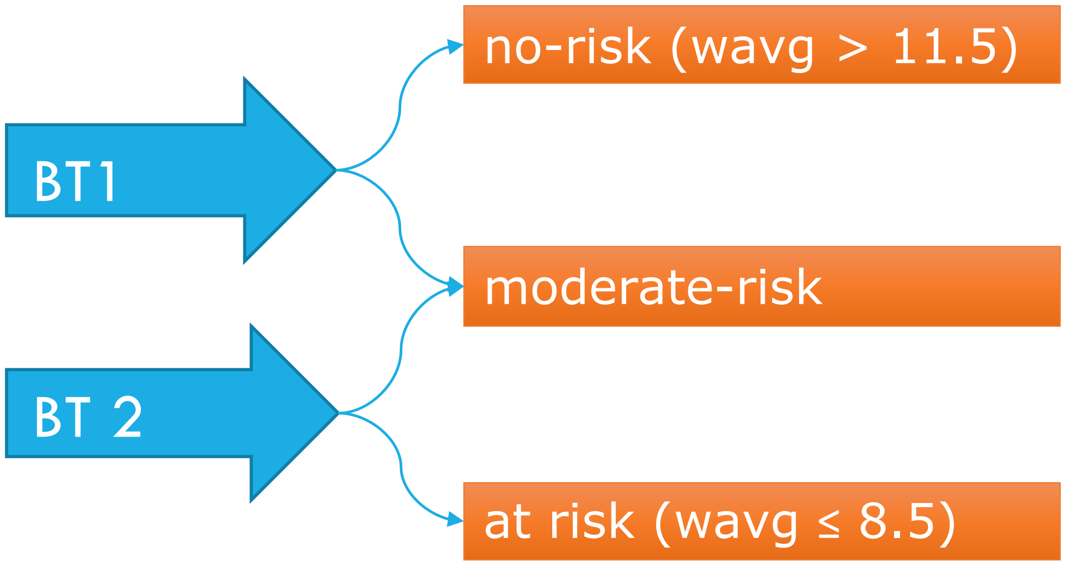


prediction logistic regression

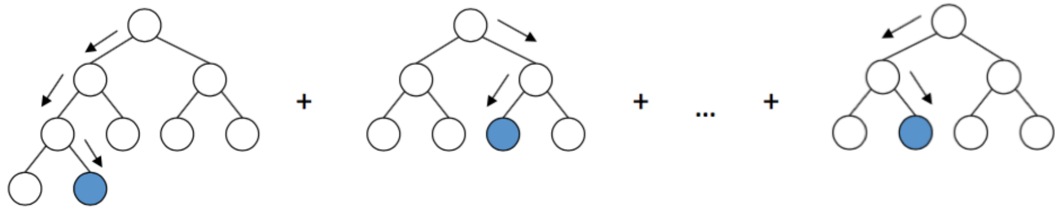
	precision	recall	F1-score
no-risk (wavg > 11.5)	0.41	0.45	0.43
moderate-risk	0.63	0.59	0.61
at risk (wavg ≤ 8.5)	0.63	0.60	0.62

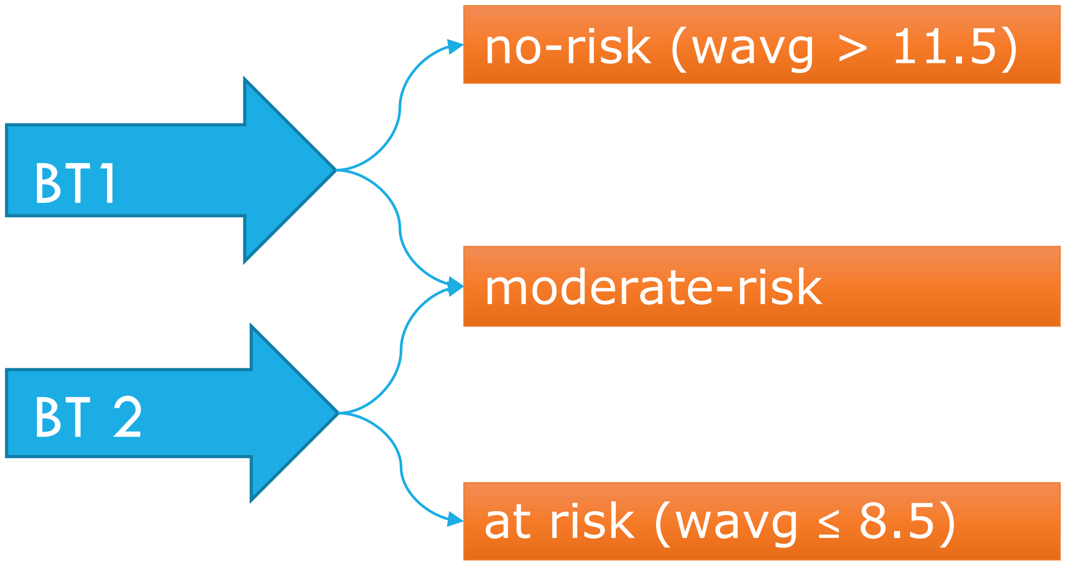
PREDICTIVE MODELLING
→ BOOSTED TREES



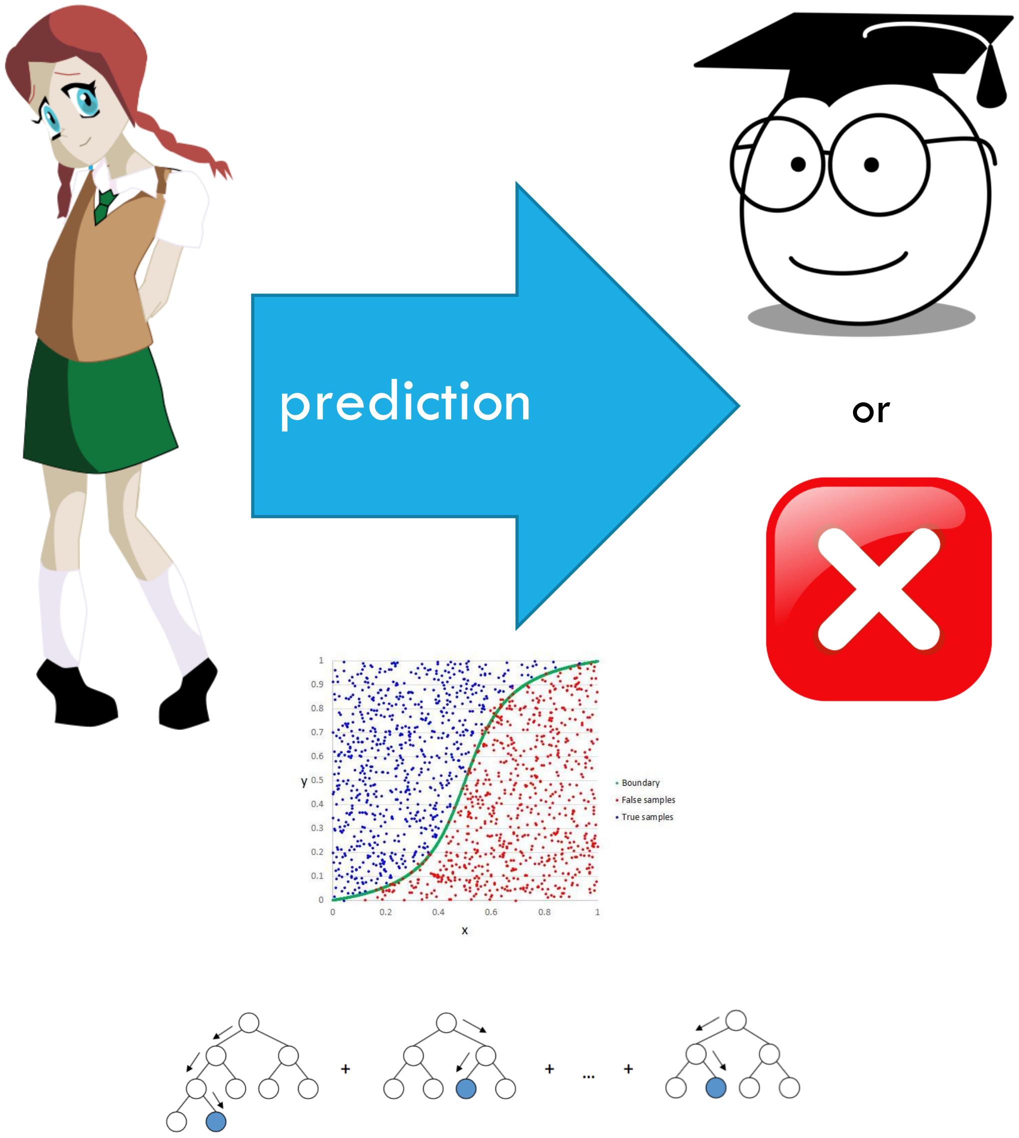


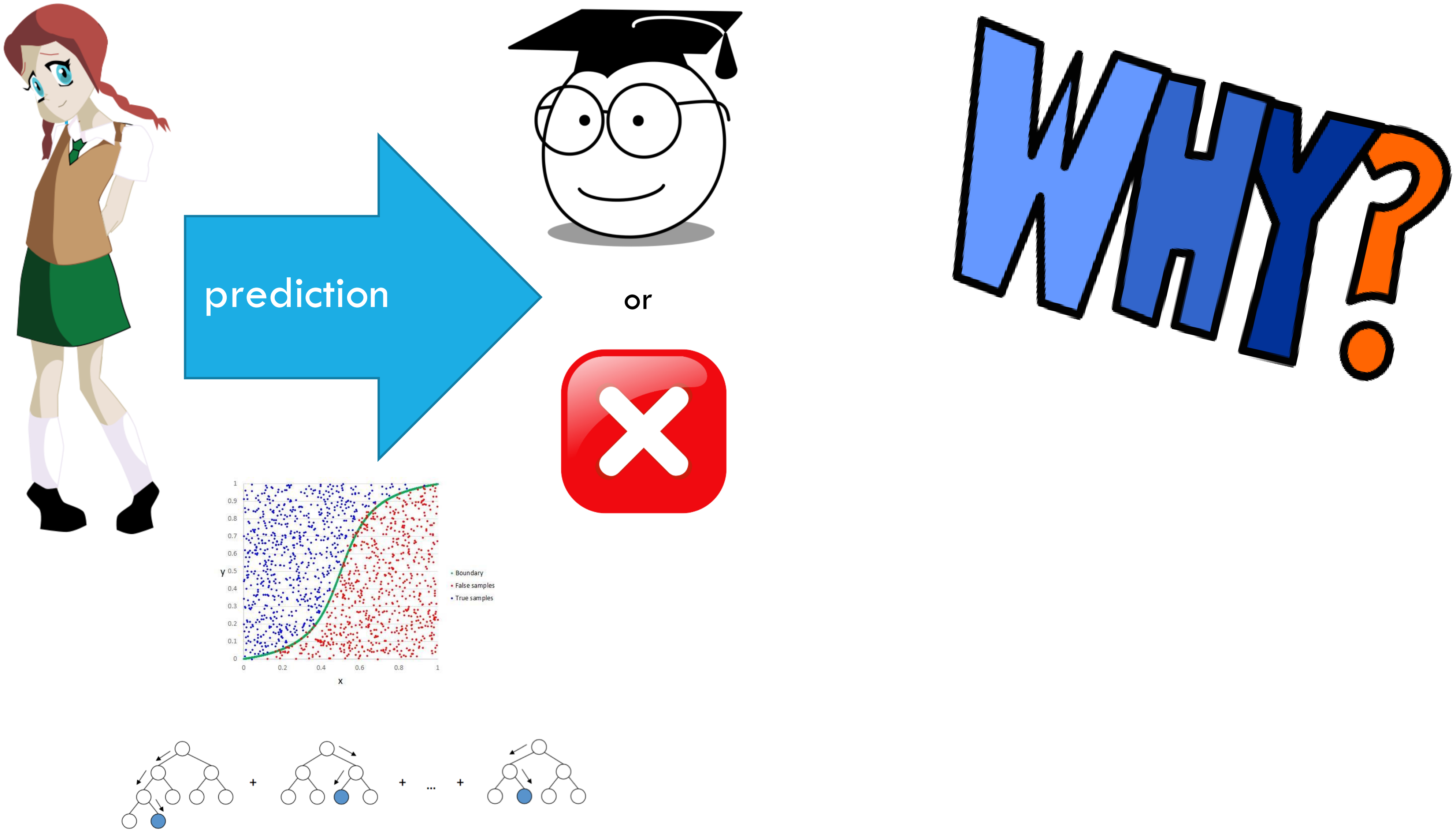
PREDICTIVE MODELLING
→ BOOSTED TREES

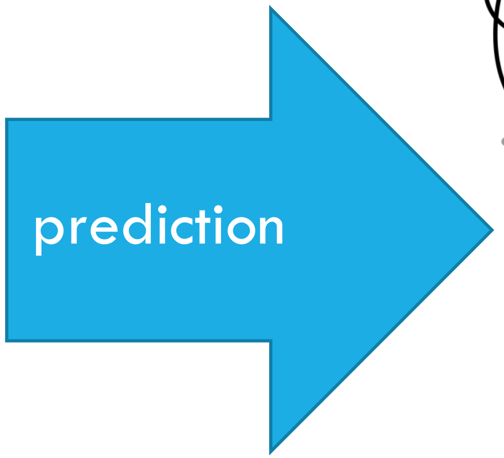




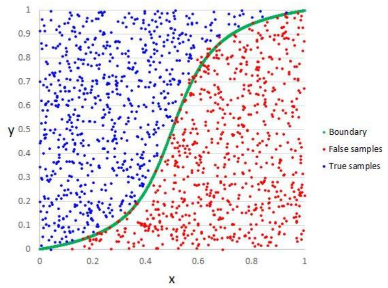
precision	recall	F1-score
0.64	*0.80*	0.71
0.88	0.77	0.82
0.87	*0.85*	0.86
0.68	0.70	0.69

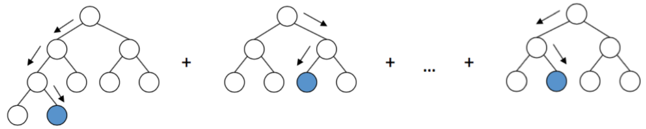

prediction
or
Boundary
False samples
True samples
x
y


prediction
or
Boundary
False samples
True samples
x
y
WHY?



or



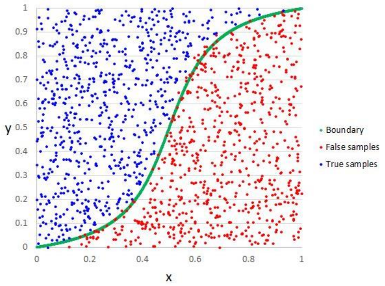


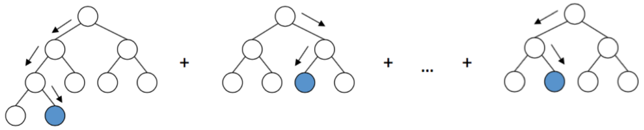
individual predictions

population-wide insights

prediction

or





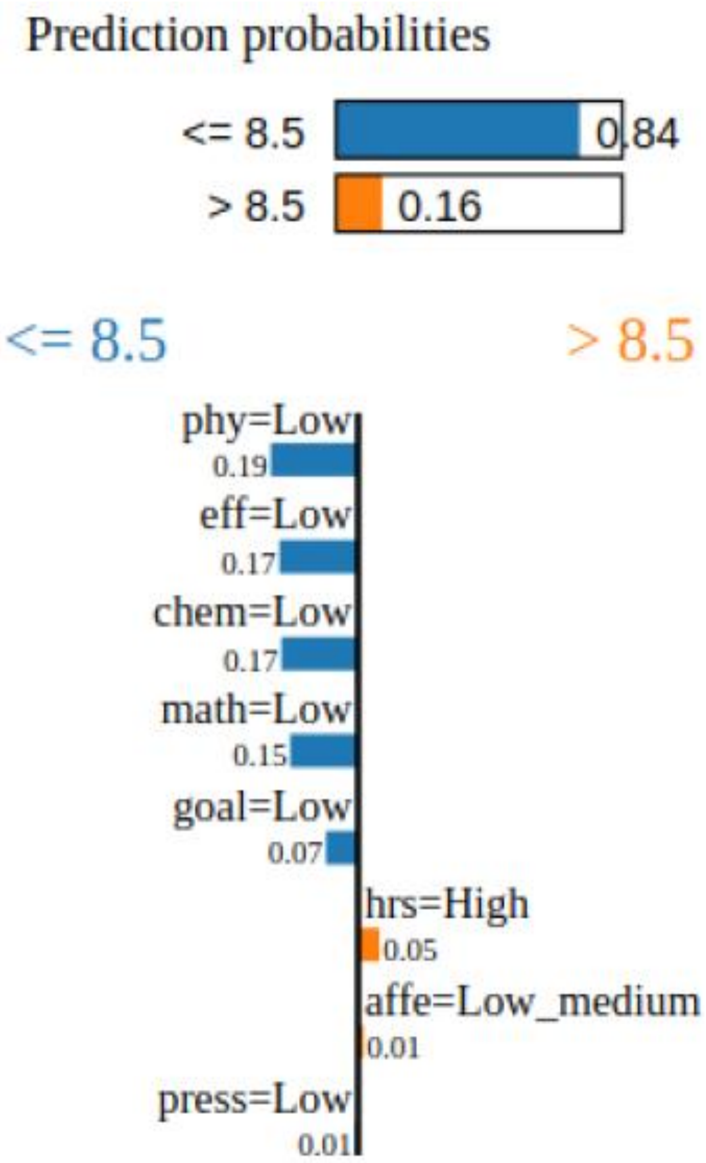
individual predictions

population-wide insights

Local Interpretable Model-agnostic Explanations (LIME)

EXPLANATORY MODELLING WITH PREDICTIVE VALIDITY → BOOSTED TREES + LIME

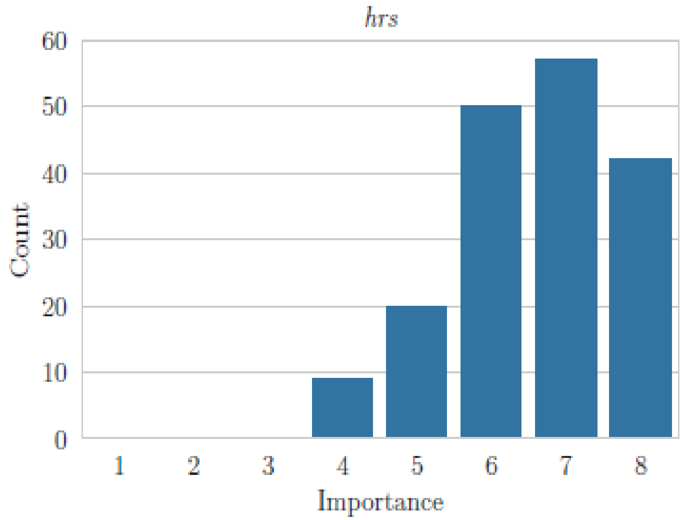
individual predictions

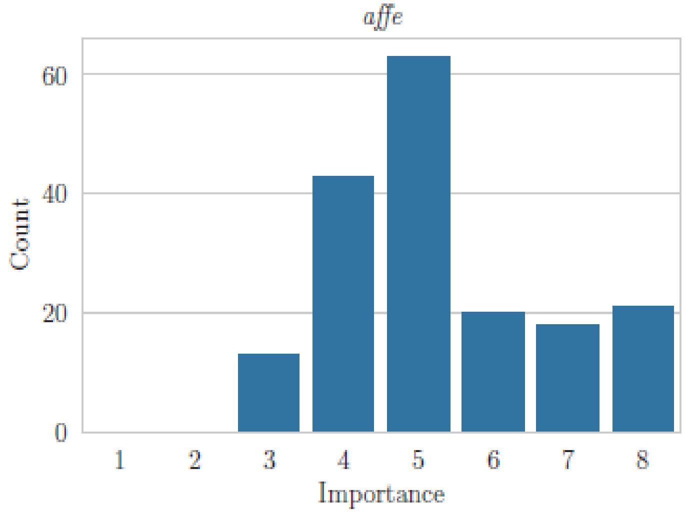


EXPLANATORY MODELLING WITH PREDICTIVE VALIDITY → BOOSTED TREES + LIME

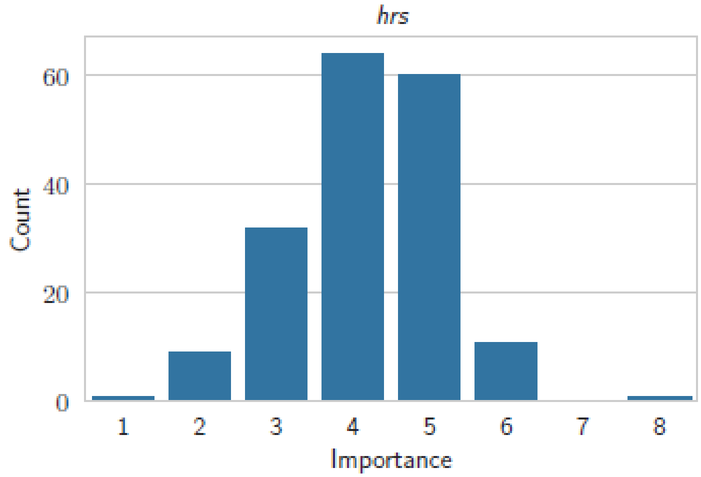
population-wide insights

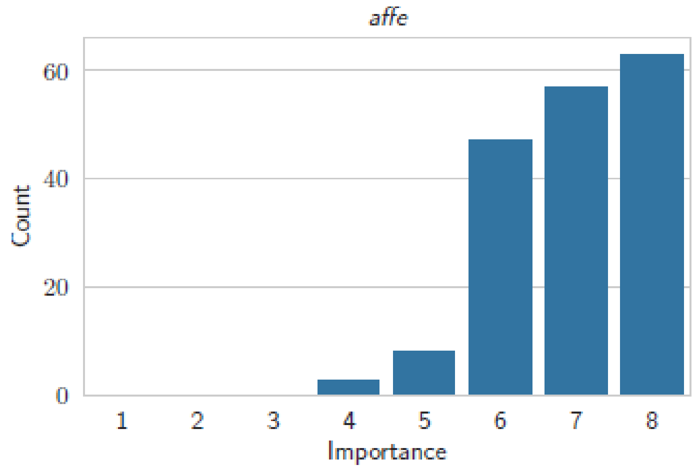
no-risk (wavg > 11.5)





at risk (wavg ≤ 8.5)





Conclusion

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?

Can boosted trees more accurately predict first-year student success than logistic regression?

Can Local Interpretable Model-agnostic Explanations (LIME) generate interpretable insights in the factors important for predicting first-year student success?

Conclusion

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success?

Hypotheses

- Prior academic experience positively AA. ✓
- Affective and goal strategies positively affect AA. ✓
- Preference for time pressure does not affect AA. ✓

Conclusion

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success? ✔

Can boosted trees more accurately predict first-year student success than logistic regression? ✔

precision & recall ↗ 20%

Conclusion

Do statistical modelling (multiple linear & logistic regression) and boosted trees identify the same factors for first-year engineering student success? ✔

Can boosted trees more accurately predict first-year student success than logistic regression? ✔

Can Local Interpretable Model-agnostic Explanations (LIME) generate interpretable insights in the factors important for predicting first-year student success? ✔

Questions for discussion during the conference

How would your university profit from research on first-year student success?

What is still required to transfer the research to practice?

Successful Transition from secondary to higher Education using Learning Analytics

enhance a **successful transition from secondary to higher education** by means of **learning analytics**

- ✓ design and build **analytics dashboards,**
- ✓ dashboards that go beyond identifying at-risk students, allowing **actionable feedback** for all students on a **large scale.**

www.stela-project.eu

@STELA_project

562167-EPP-1-2015-1-BE-EPPKA3-PI-FORWARD