NOVA IMS

Information Management School

Assessing the impact of AA determinants using ML methods:

Evidence from a European Country

Ana Beatriz Antunes Afonso

Advisor: Professor Doutor Frederico Cruz-Jesus

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Acreditações e Certificações

NOVA IMS Information Management School

Agenda

- Introduction
- Objectives
- Theoretical Background
- Contributions
- Methodology
- Findings
- Theoretical implications
- Practical implications

Introduction

- Work developed with DGEEC in the last months.

- Predict the grades on the Mathematics and Portuguese national exams using data science methods.

- Compare the efficiency of different Machine Learning methods.
- Understanding the determinants that lead to the different grades in these exams and quantify them.

- The same evaluation method and conditions for everyone.
- Exams to enter university for the great majority of degrees.

Introduction

Education leads to:

Better jobs

A more competitive country

Boost in the economy

Better health condition

Higher quality of life

Sustainability

Reduction of crime

Democratic world

Peace

Objectives

Understanding the underlying factors that lead do disparities in education.

Policymakers take more effective decisions, either at schools or governmental levels.

Improvement of the student's performance.
Reduction of educational gaps.

Theoretical Background

AA Drivers:

- Cognitive skills
- Past academic behaviour

- Socioeconomic and demographic characteristics

- Emotional intelligence
- Attitude toward school

- Usage of computers, internet and social media

- Time and quality of study

- Teachers' characteristics

- Schools' characteristics

- Classes' structure

Contributions

- Previous research is mostly based on samples.
- AI approaches only started to be used in the past decade.

Use of the **population** who performed national exams to choose the best model and **compare ML methods with a classical statistical approach.**

Comparison of **drivers for the success in mathematics vs. mother tongue** national exams.

New approach that allows the **measurability of the impact of the drivers** using Neural Networks.

Methodology

Data

Academic year 2018/2019

19 445 students of the 12th grade

35 780 exams

- 14 207 Mathematics national exams
- 21 573 Portuguese national exams

3 datasets, 3 targets:

- **Portuguese national exam grade**
- **Mathematics national exam grade**
- **Aggregate grade***

(*) – Portuguese and mathematics datasets together.

Data

Categories of the variables:

Students

Legal guardians

Professors

Schools

Data pre-processing

Data exploration

Data cleaning

Dummy encoding

Join with dataset of demographic information

+50 variables

Feature Selection

Selection of the number of features

Recursive Feature Elimination (RFE)

Which features to use

Recursive Feature Elimination (RFE) + Ridge Regression + Lasso Regression

Features Selected

Dataset ready for modelling

Modelling

Dataset → 75% training + 25% validation

Grid search for tunning the parameters

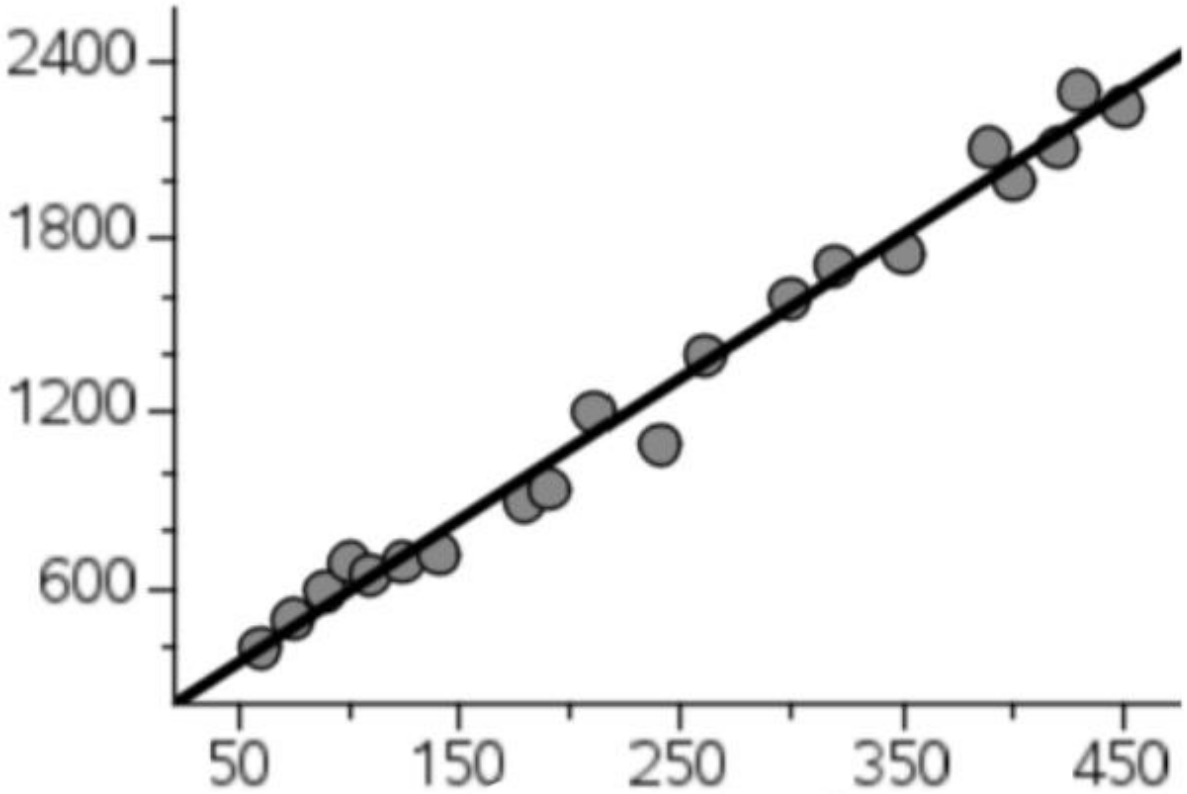
Modelling

Cross-validation

Modelling

Linear Regression

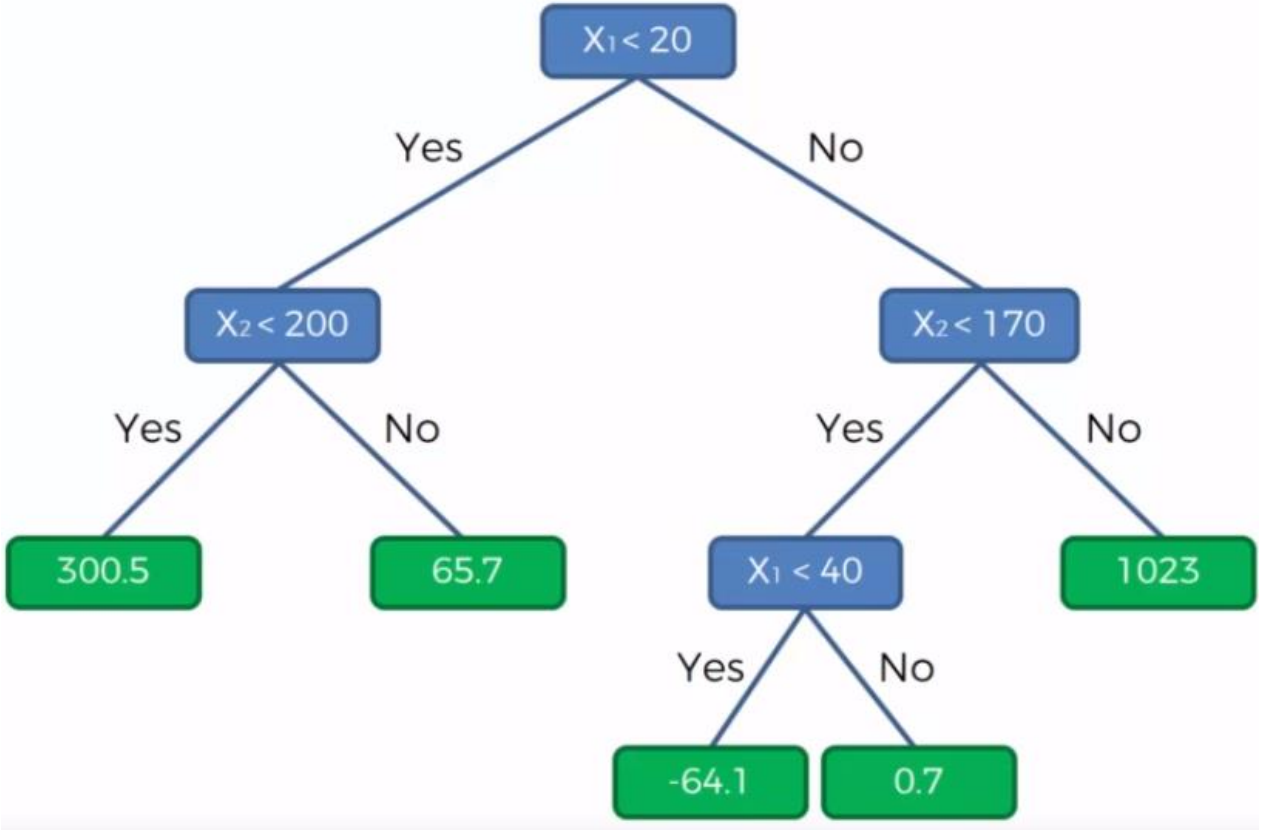
Finds the linear relationship between a dependent variable and one or more independent variables.



Modelling

Decision Tree

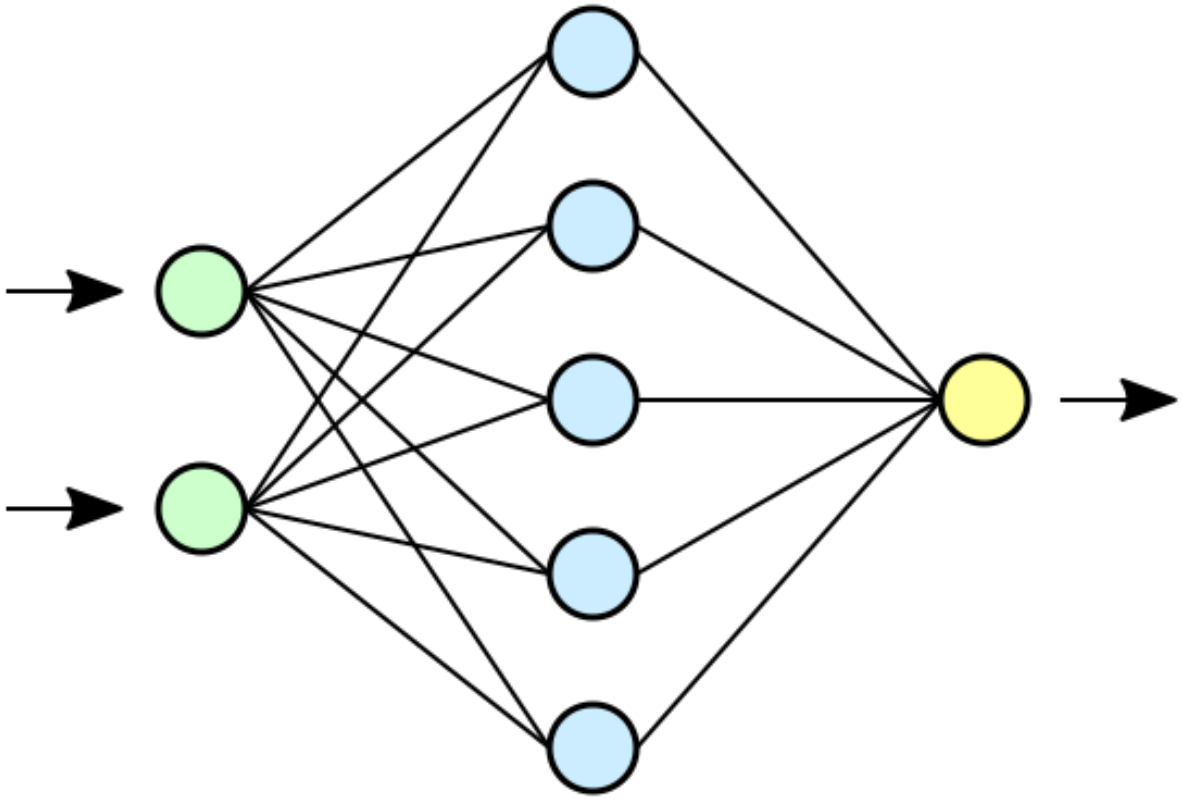
A graphical representation of all the possible solutions to a decision based on certain conditions.



Modelling

Neural Network

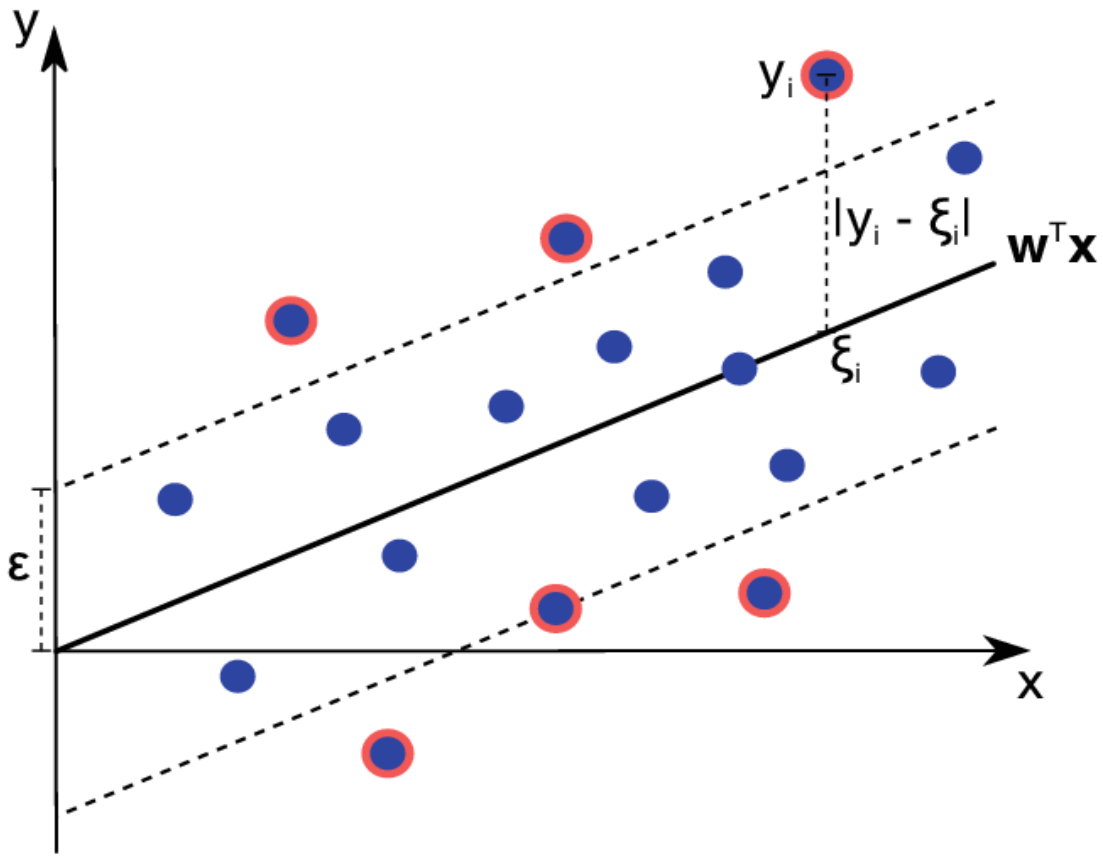
Mimics a human brain. It has an input layer, where data enters the network; a hidden layer, comprised of artificial neurons, each of which receives multiple inputs from the input layer. The artificial neurons summarize their inputs and pass the results to the output layer where they are combined again.



Modelling

Support Vector Regressor

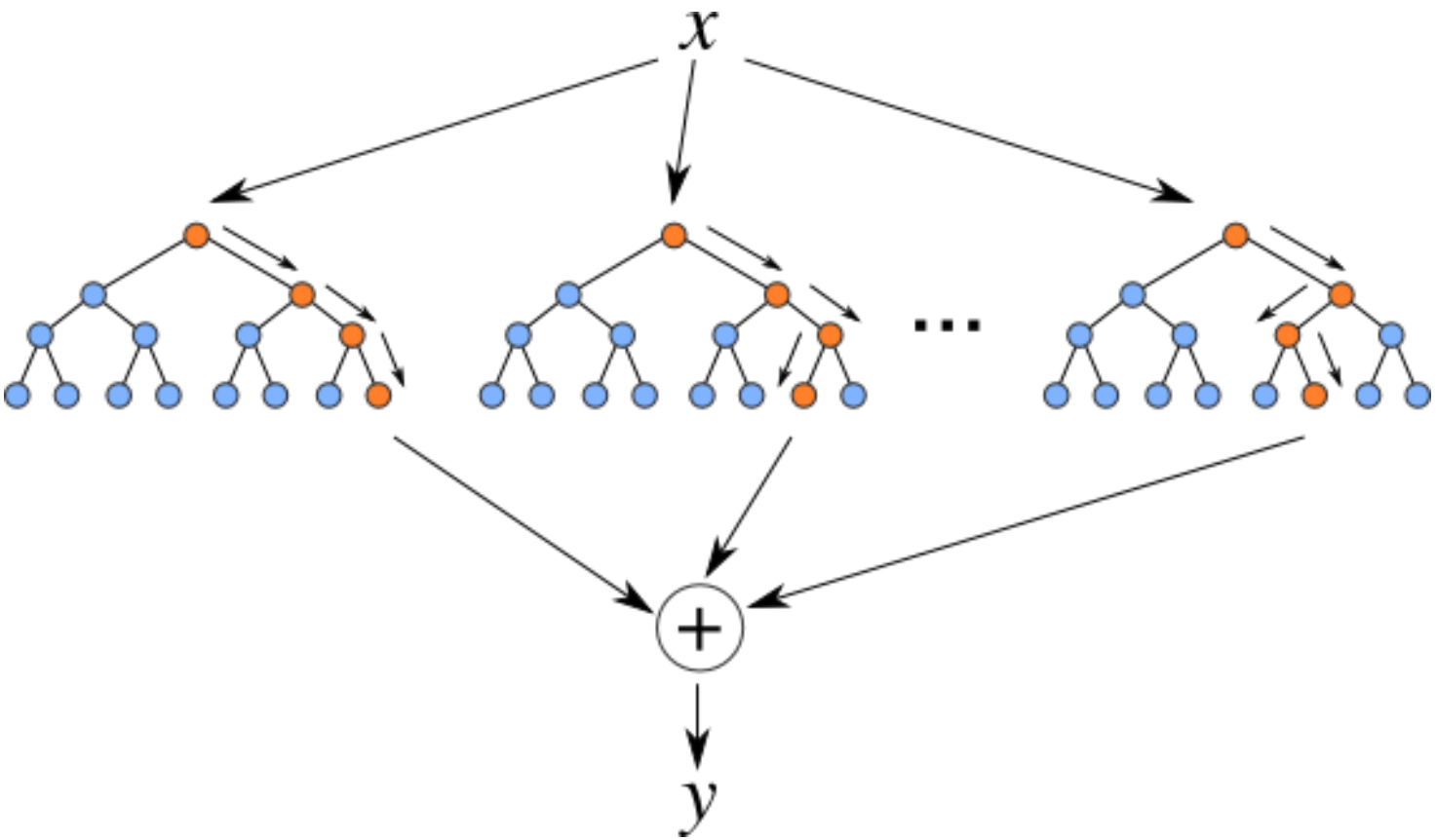
SVR tries to fit the best line within a threshold value which is the distance between the hyperplane and boundary line.



Modelling

Random Forests

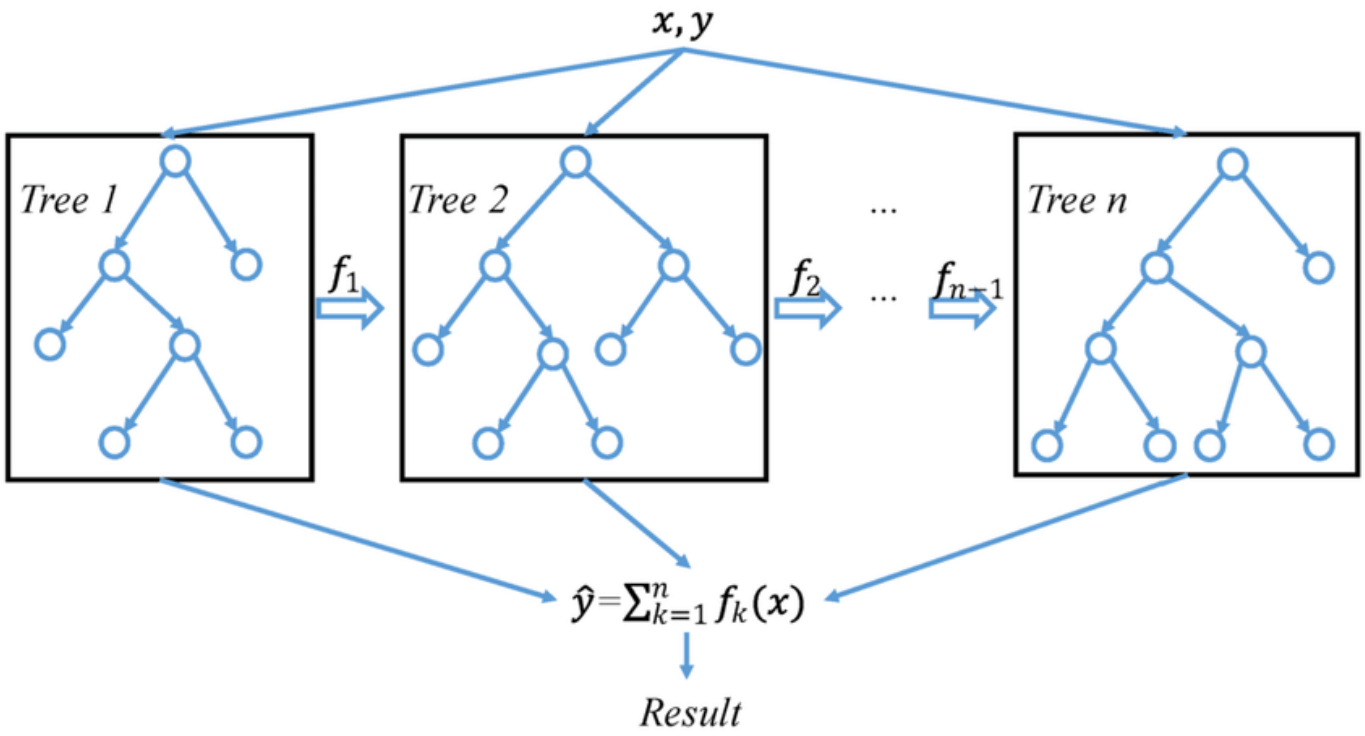
A Random Forest combines several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.



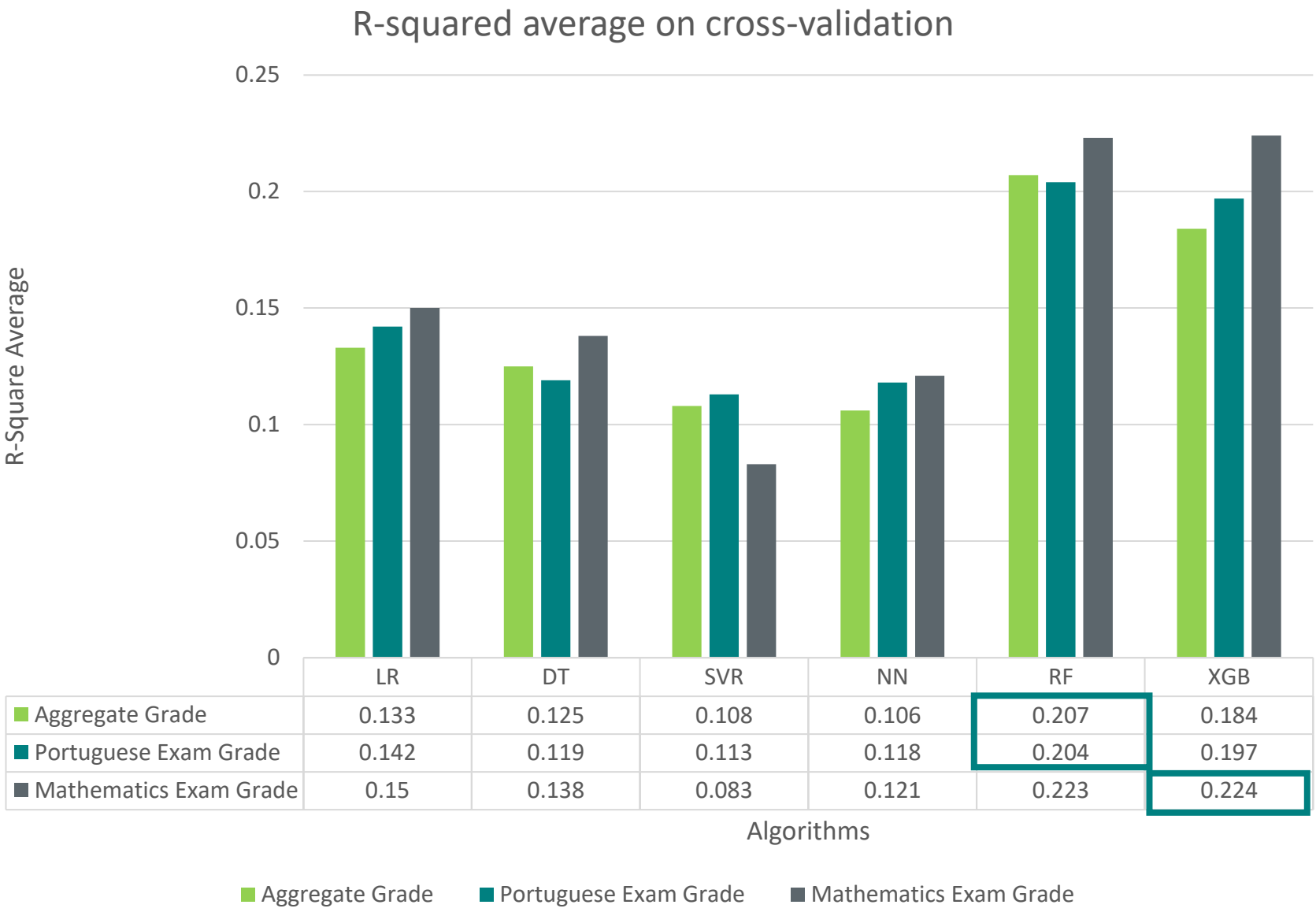
Modelling

Extreme Gradient Boosting

Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. XGBoost for short is an efficient open-source implementation of the gradient boosting algorithm



Results



	LR	DT	SVR	NN	RF	XGB
Aggregate Grade	0.133	0.125	0.108	0.106	0.207	0.184
Portuguese Exam Grade	0.142	0.119	0.113	0.118	0.204	0.197
Mathematics Exam Grade	0.15	0.138	0.083	0.121	0.223	0.224

Results

The most noticeable variables on feature importance:

Feature importance evaluated on the 2 best models: RF and XGB

- Student's age

- Female gender

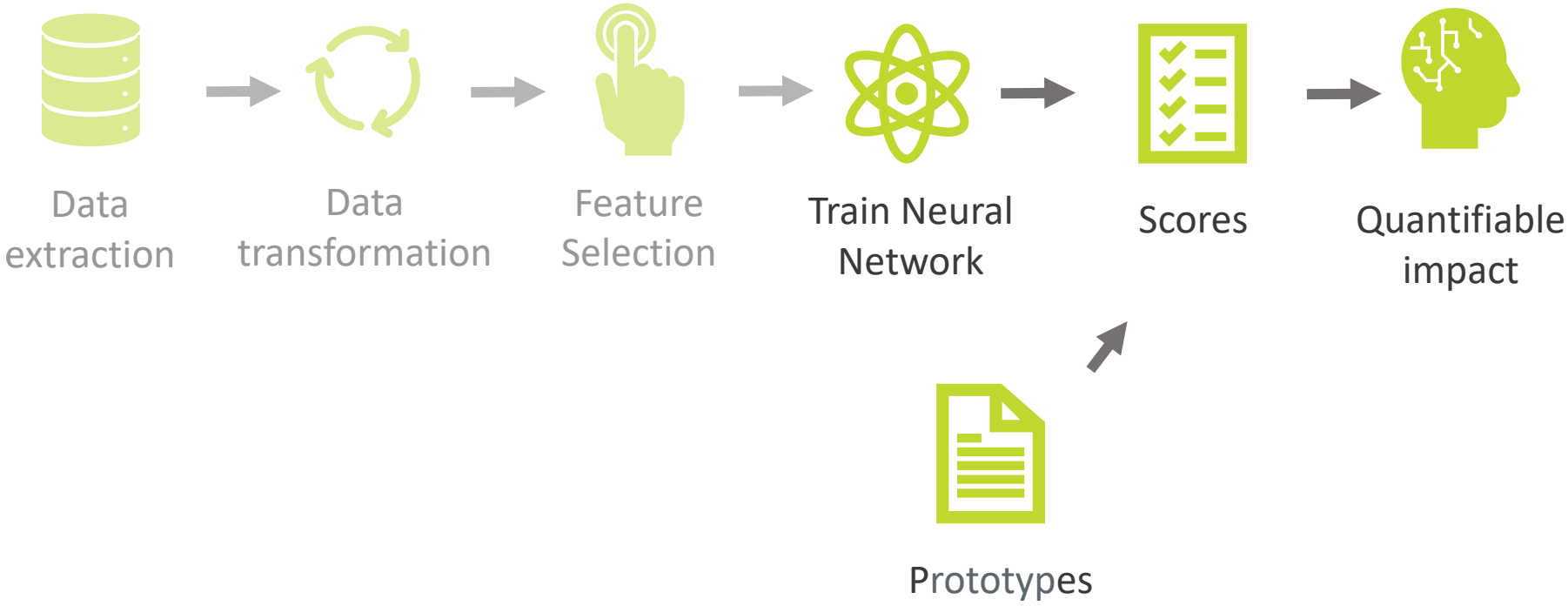
- Education of the legal guardian

- Rate of student's who have failed the year in that school

- Rate of students with social support of the school

- Rate of teachers with MSc or PhD on the school

Prototypes


Data extraction
Data transformation
Feature Selection
Train Neural Network
Scores
Quantifiable impact
Prototypes

Prototypes

	Variable A	Variable B	Variable C	Variable D	(...)
	Mean of A	Mean of B	Mean of C	Mean of D	
	Mean of A – STD of A	Mean of B	Mean of C	Mean of D	
	Mean of A + STD of A	Mean of B	Mean of C	Mean of D	
	Mean of A	Mean of B – STD of B	Mean of C	Mean of D	
(...)					

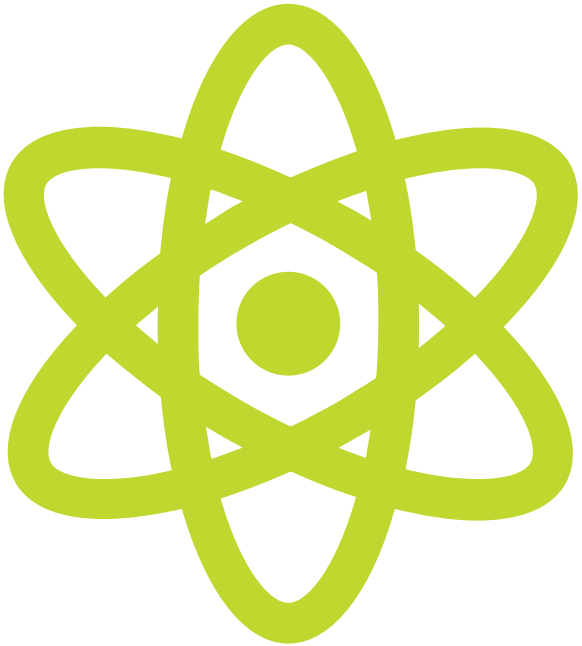
NOVA IMS Information Management School

Prototypes

FKENTIDADE_INSCR	Aluno_Idade	Aluno_EE_Hab	Escola_EscalaoA	Escola_Reprovados	Escola_MScPhD
media	17.25	11.41	0.09	0.17	0.18
Aluno_Idade - SD	16.71	11.41	0.09	0.17	0.18
Aluno_Idade + SD	17.79	11.41	0.09	0.17	0.18
Aluno_EE_Hab - SD	17.25	8.14	0.09	0.17	0.18
Aluno_EE_Hab + SD	17.25	14.68	0.09	0.17	0.18
Escola_EscalaoA - SD	17.25	11.41	0.03	0.17	0.18
Escola_EscalaoA + SD	17.25	11.41	0.15	0.17	0.18
Escola_Reprovados - SD	17.25	11.41	0.09	0.09	0.18
Escola_Reprovados + SD	17.25	11.41	0.09	0.25	0.18
Escola_MScPhD - SD	17.25	11.41	0.09	0.17	-0.02
Escola_MScPhD + SD	17.25	11.41	0.09	0.17	0.38

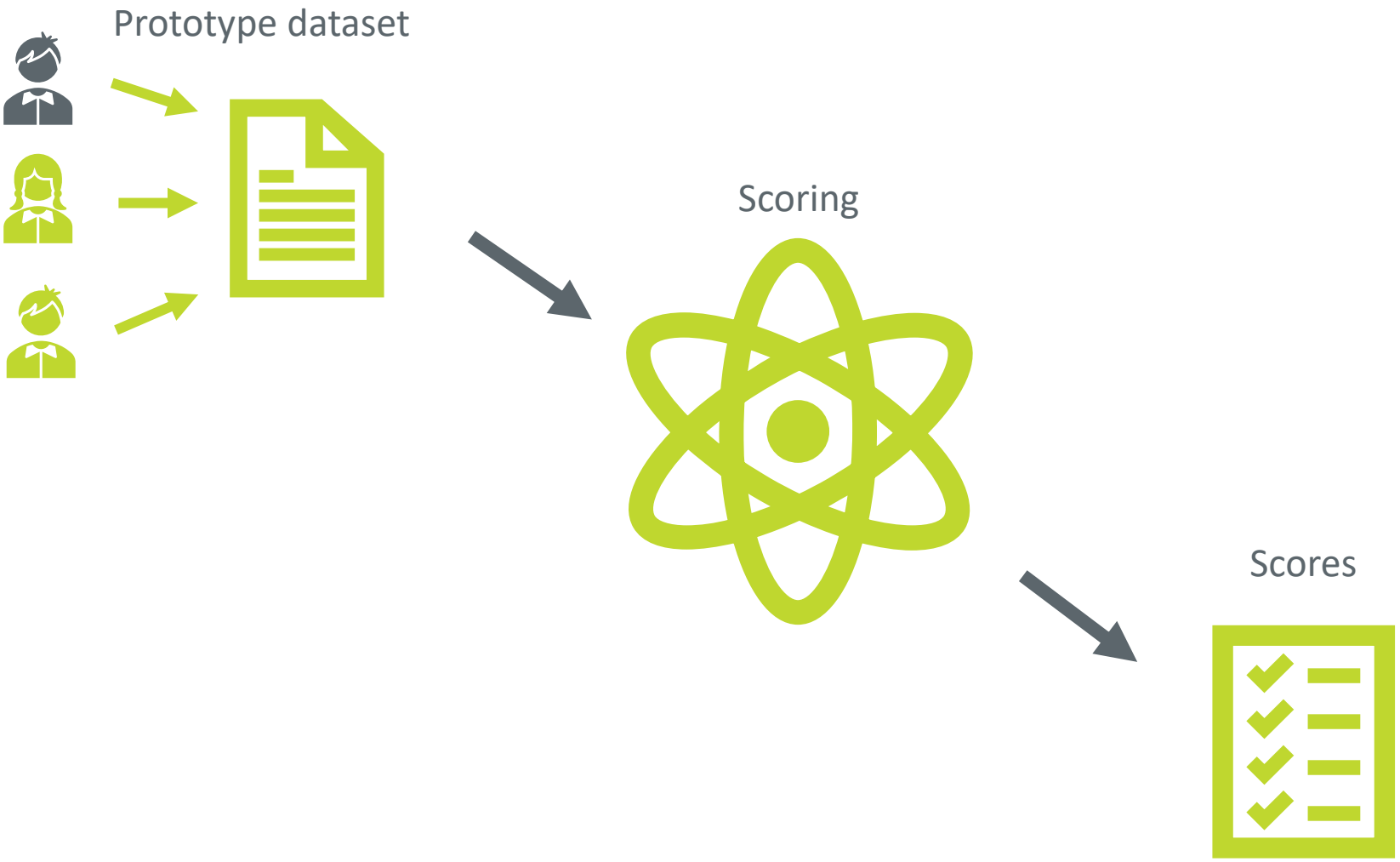
Prototypes

Train Neural Network with Original Dataset



Prototypes

Predict results using the prototype dataset



Prototypes

Predicted result of mean of A – STD of A prototype

Predicted result of the mean prototype

Quantifiable impact

Prototypes

Results of the prototypes	Aggregate Grade		Portuguese Grade		Mathematics Grade	
	Unit	$\hat{\beta}$	Unit	$\hat{\beta}$	Unit	$\hat{\beta}$
Student's age	0.58	-1.2	0.61	-1.1	0.54	-1.5
Education of the legal guardian	3.3	0.9	3.29	0.9	3.27	1.5
Feminine Gender	1	0.5	1	0.6	1	0.2
R. student's who have failed the year in that school	0.08	-0.7	0.08	-0.5	0.08	-0.8
R. students with social support of the school	0.06	-0.1	0.06	-0.2	0.06	-0.1
R. teachers with MSc or PhD on the school	0.2	-0.1	0.2	0.1	0.2	0

Findings

Students' age has a proxy for retention. Older students:
-1.5 points in math and -1.1 in Portuguese national exam.

Schools with higher rates of failing:
-0.8 point in math, -0.5 in Portuguese and -0.7 in the aggregate grade.

Legal guardians' with more than 12 years of education:
+1.5 point in math and +0.9 in Portuguese national exam.

School size, bigger schools with middle and high school:
-0.9 in math and -0.2 in mother tongue exams.

Findings

Schools with higher rate of economically disadvantage students:
-0.2 points in mother tongue national exam.

Teacher having a MSc or PhD:
+0.1 points in mother tongue national exam.

Girls outperform boys:
+0.6 points in the Portuguese national exam.

Internet has a positive but negligible impact.

Theoretical implications

Quantitative characterisation of the impact of each AA driver on final national exams.

Comparison between the importance of the different drivers related with students, legal guardians, teachers and schools.

While using virtually every student in Portugal and advanced data science techniques.

Practical implications

Retention criteria should be reviewed and **reinforce work at the student level** for the ones in risk of failing.

Considering legal guardians' education when assessing students to classes, to create heterogenous groups.

Given legal guardians' education, **flag students who might need extra help**.

Facilitation **of career opportunities for teachers** who decide to go post-graduate.

Digital reinforcement.

Minimalization of the effects of **living in socially and economically disadvantaged territories**.

Thank you!

Address: Campus de Campolide, 1070-312 Lisboa, Portugal
Phone: +351 213 828 610 Fax: +351 213 828 611

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa