

INVESTIGAÇÃO
ADMINISTRAÇÃO
PÚBLICA

Inteligência artificial e ciência de dados



Compreender os determinantes do desempenho académico: evidências do sistema de ensino secundário Português

Tiago Oliveira e Luísa Loura



DIREÇÃO-GERAL DE ESTATÍSTICAS
DA EDUCAÇÃO E CIÊNCIA

Equipa: Tiago Oliveira, Luísa Loura, Frederico Cruz Jesus, Mauro Castelli, Joana Duarte

O abandono escolar é um obstáculo ao crescimento económico e ao emprego.
O abandono escolar dificulta a produtividade e a competitividade, promovendo a pobreza e a exclusão social.

Compreender e potenciar o desempenho académico torna-se algo de crítica importância.

**Ambiente
Económico**

Saúde

Competitividade

Emprego

Educação

**Sustentabilidade
SS**

**Participação
Cívica**

Este projeto pretende analisar os antecedentes do desempenho académico, à escala nacional, usando micro dados dos alunos do ensino secundário público.

Dados do abandono escolar no âmbito da Estratégia Europa 2020



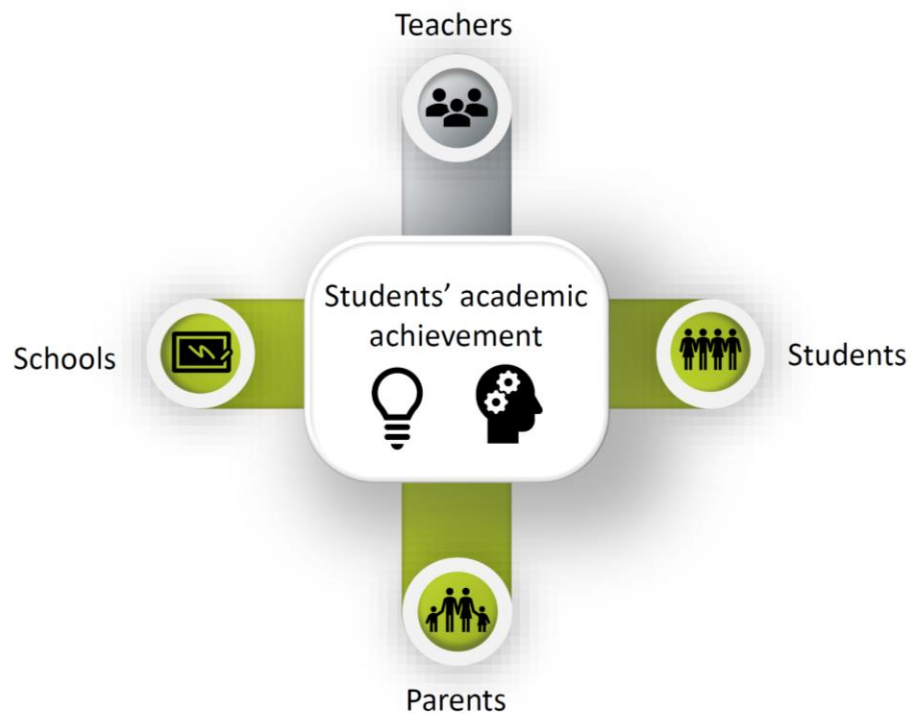
< 10%



≈ 12.6%



≈ 10.7%



- 1 Uma das primeiras iniciativas que usa métodos de *data science* no contexto de desempenho académico (DA) em grande Escala;
- 2 Compreender os antecedentes do DA;
- 3 Utilização de modelos logísticos baseados em árvores (*logistic model trees*) para resolver problemas de classificação e regressão;
- 4 Desenvolver e testar empiricamente um modelo conceptual abrangente para o DA.

- 1 Promover o DA providenciando aos decisores públicos, escolas, e professores, dados sobre os seus drivers, assim como previsões individuais dos alunos;
- 2 Ajudar Portugal e outros Estados-Membros a atingir os objetivos da Estratégia Europa 2020 em termos de DA;
- 3 Usar dados relevantes do contexto do DA que existem mas ainda não foram totalmente usados;
- 4 Desenvolver e disponibilizar um conjunto de modelos que providenciem as estimativas individuais do DA de cada aluno a cada disciplina no início de cada ano letivo;
- 5 Providenciar um conjunto abrangente de sugestões de melhoria para o sistema de bases de dados do Ministério da Educação Português.

Understanding the drivers of academic achievement: Evidence for Portugal's high school system

NOVA IMS Team:

Tiago Oliveira, Ph.D.

Frederico Cruz-Jesus, Ph.D.

Mauro Castelli, Ph.D.

Catarina Neves, Ph.D. Student

Ricardo Mendes, Ph.D. Student

References	Methods	St	Pa	Sc
Hanushek and Kimko (2000)	Regressions models	x		x
Hoxby (2000)	Regressions models	x		x
Fan and Chen (2001)	General linear model	x	x	
Barnett, Glass, Snowdon, and Stringer (2002)	Linear Programming techniques			x
Driessen, Smit, and Slegers (2005)	Frequency, Variance, and Structural models	x	x	x
Rivkin, Hanushek, and Kain (2005)	Regression models			x
Archibald (2006)	Hierarchical linear models	x		x
Jackson et al. (2006)	Internet recorded	x		
J.-S. Lee and Bowen (2006)	Hierarchical linear model	x	x	
Marks, Cresswell, and Ainley (2006)	Item Response Theory; Regressions models	x	x	x
Jeynes (2007)	Regression models		x	
Codjoe (2007)	Interviews	x		
Croninger, Rice, Rathbun, and Nishio (2007)	Hierarchical linear models	x		
H. Lee (2007)	Hierarchical linear models; Classic lineal regression model	x	x	x
Lei and Zhao (2007)	Hierarchical linear models; ANOVA tests	x		
Steinmayr and Spinath (2008)	Regression models	x		
Caro et al. (2009)	Hierarchical linear models; Panel data models	x		
Mensah and Kiernan (2010)	Tobit regression models; Univariate and Multivariate analyses	x	x	
Hartas (2011)	Univariate analyses of variance; Chi-square tests		x	
Patterson and Pahlke (2011)	Regression models	x	x	
Hanushek and Woessmann (2012)	Regression models	x		x
Brunner et al. (2013)	Multiple group factor analytic models; Full maximum likelihood	x		
Wally-Dima and Mbekomize (2013)	Descriptive statistics T tests	x		
Bosworth (2014)	Regression models	x		x
Krassel and Heinesen (2014)	Regression discontinuity design; Control for school fixed effects; Ordinary Least Squares	x	x	x
Vigdor, Ladd, and Martinez (2014)	Probit regression; Regression models	x		
Hodis et al. (2015)	Hierarchical linear models	x		
C. Lee and Walk (2015)	Ordinary Least Squares	x		

1st Scientific Paper

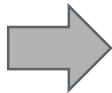
Variable	Description
x0	Year of the study cycle
x1	Binary variable indicating Portuguese citizenship
x2	Binary variable indicating whether a student was born in Portugal
x3	Binary variable indicating the sex of the student
x4	Age of the student
x5	Number of enrolled years in high school
x6	Number of failures in the educational career
x7	Scholarship
x8	Type of financial support received by the student according to the financial condition of his family (those receiving supported are economically disadvantaged)
x9	Binary variable indicating whether a student has a computer
x10	Binary variable indicating whether a student has access to the Internet
x11	Number of students in the class
x12	Number of students in the whole school
x13	Indicator of the economic condition of the family of the student
x14	Area of residence
x15	Binary variable indicating whether the school is in a rural area
x16	Number of unit courses attended in the present academic year

1st Scientific Paper

Random Forest

The main idea of ensemble learning methods is to combine, during the training phase, the prediction of different models (called base learners or weak learners) with the objective of producing a more accurate and reliable prediction. The base learners could be, for instance, decision trees and artificial neural networks or even a combination of different supervised learning techniques.

AI Methods



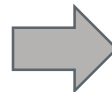
Artificial Neural Networks

Artificial Neural Networks (ANNs) are one of the best-known and widely used (AI) techniques. They are biologically inspired, and they mimic the structure of the human brain (Haykin, 1994).

Support Vector Machines

Support Vector Machines (SVMs) (Cortes & Vapnik, 1995) are a popular ML method for addressing classification and regression problems. Focusing on a classification problem where each training observation belongs to one of the possible two classes, the main idea of SVMs is to determine the best hyperplane that separates instances of one class from the instances of the second class.

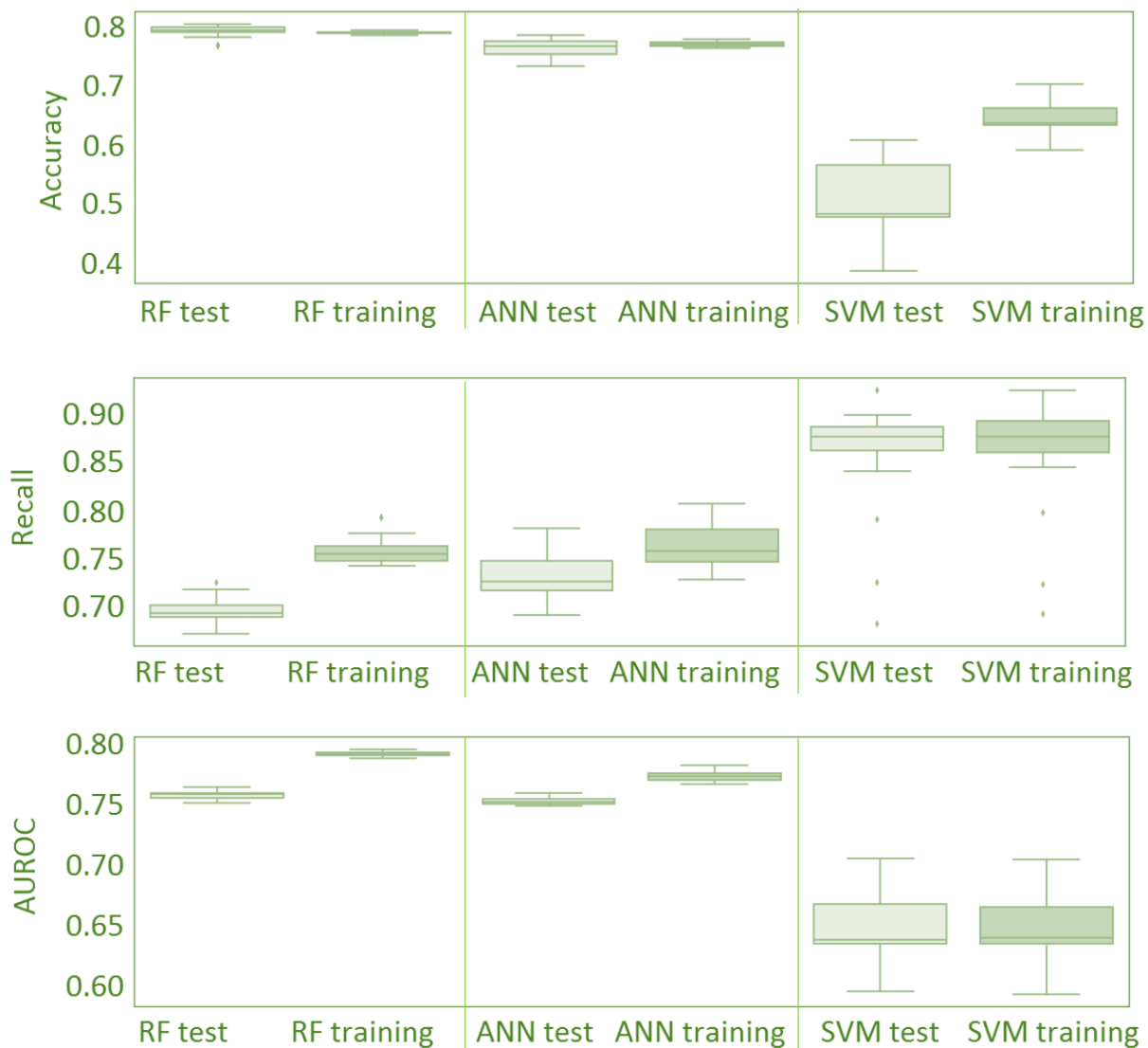
Classic Methods



Logistic Regression

Logistic regression (LR) is a well-known technique that is commonly applied in the field of statistics. LR is also used in machine learning as a baseline for testing the performance of more advanced techniques. The use of LR was proposed to overcome one of the limitation of linear regression, related to the fact that the output of a regression model is not constrained in $[0, 1]$.

1st Scientific Paper



1st Scientific Paper

Test Set	Random Forest	Neural Networks	Support Vector Machines	Logistic Regression
Accuracy	79.4%	76.5%	51.2%	81.1%
Recall	69.4%	73.0%	86.3%	48.7%
AUROC	76%	75%	65%	55%

Test Set	Random Forest	Neural Networks	Support Vector Machines	Logistic Regression
Cumulative Lift at 5%	4.65	4.26	2.45	2.59
Cumulative Lift at 15%	3.28	3.13	1.41	2.66
Cumulative Captured Response 5%	23%	21%	12%	13%
Cumulative Captured Response 15%	49%	47%	33%	40%
Threshold Score @ 15%	0.72	0.76	0.66	0.35

1st Scientific Paper

Acronym	Description	Worth x 100	Rank Worth
x0	Year of the study cycle	0.00	15
x1	Binary variable indicating Portuguese citizenship	0.00	15
X2	Binary variable indicating whether a student was born in Portugal	0.06	12
X3	Binary variable indicating the sex of the student	9.43	3
X4	Age of the student	6.26	4
X5	Number of enrolled years in high school	6.02	5
X6	Number of failures in the educational career	14.14	2
X7	Scholarship	3.83	7
x8	Type of financial support received by the student according to the financial condition of his family (those receiving supported are economically disadvantaged)	0.15	11
X9	Binary variable indicating whether a student has a computer	0.57	9
X10	Binary variable indicating whether a student has access to the Internet	0.00	15
X11	Number of students in the class	0.26	10
X12	Number of students in the whole school	5.74	6
X13	Indicator of the economic condition of the family of the student	3.38	8
X14	Area of residence	0.00	15
X15	Binary variable indicating whether the school is in a rural area	0.00	15
X16	Number of unit courses attended in the present academic year	50.17	1

Random Forest	From	To	Primeiros	Fails (n)	Fails (%)	Cum Fails (%)	Lift	Cum Lift	Capt
Ventile 1	1	2,212	Ventile 1	1929	87%	87%	4.65	4.65	23%
Ventile 2	2,213	4,425	Ventile 2	1229	56%	71%	2.96	3.80	38%
Ventile 3	4,426	6,639	Ventile 3	921	42%	61%	2.22	3.28	49%
Ventile 4	6,640	8,852	Ventile 4	778	35%	55%	1.87	2.93	58%
Ventile 5	8,853	11,066	Ventile 5	642	29%	50%	1.55	2.65	66%
Ventile 6	11,067	13,279	Ventile 6	416	19%	45%	1.00	2.38	71%
Ventile 7	13,280	15,493	Ventile 7	402	18%	41%	0.97	2.17	76%
Ventile 8	15,494	17,706	Ventile 8	369	17%	38%	0.89	2.01	81%
Ventile 9	17,707	19,920	Ventile 9	292	13%	35%	0.70	1.87	84%
Ventile 10	19,921	22,133	Ventile 10	292	13%	33%	0.70	1.75	88%
Ventile 11	22,134	24,347	Ventile 11	256	12%	31%	0.62	1.65	91%
Ventile 12	24,348	26,560	Ventile 12	219	10%	29%	0.53	1.56	93%
Ventile 13	26,561	28,774	Ventile 13	159	7%	27%	0.38	1.46	95%
Ventile 14	28,775	30,987	Ventile 14	111	5%	26%	0.27	1.38	97%
Ventile 15	30,988	33,201	Ventile 15	90	4%	24%	0.22	1.30	98%
Ventile 16	33,202	35,414	Ventile 16	69	3%	23%	0.17	1.23	98%
Ventile 17	35,415	37,628	Ventile 17	47	2%	22%	0.11	1.17	99%
Ventile 18	37,629	39,841	Ventile 18	32	1%	21%	0.08	1.10	99%
Ventile 19	39,842	42,055	Ventile 19	34	2%	20%	0.08	1.05	100%
Ventile 20	42,056	44,268	Ventile 20	18	1%	19%	0.04	1.00	100%

NAN	From	To	Primeiros	Fails (n)	Fails (%)	Cum Fails (%)	Lift	Cum Lift	Capt
Ventile 1	1	2,212	Ventile 1	1770	80%	80%	4.26	4.26	21%
Ventile 2	2,213	4,425	Ventile 2	1215	55%	67%	2.93	3.60	36%
Ventile 3	4,426	6,639	Ventile 3	911	41%	59%	2.19	3.13	47%
Ventile 4	6,640	8,852	Ventile 4	785	35%	53%	1.89	2.82	56%
Ventile 5	8,853	11,066	Ventile 5	642	29%	48%	1.55	2.56	64%
Ventile 6	11,067	13,279	Ventile 6	479	22%	44%	1.15	2.33	70%
Ventile 7	13,280	15,493	Ventile 7	414	18%	40%	1.00	2.14	75%
Ventile 8	15,494	17,706	Ventile 8	342	15%	37%	0.82	1.98	79%
Ventile 9	17,707	19,920	Ventile 9	375	17%	35%	0.90	1.86	83%
Ventile 10	19,921	22,133	Ventile 10	266	12%	33%	0.64	1.73	87%
Ventile 11	22,134	24,347	Ventile 11	272	12%	31%	0.66	1.64	90%
Ventile 12	24,348	26,560	Ventile 12	202	9%	29%	0.49	1.54	92%
Ventile 13	26,561	28,774	Ventile 13	173	8%	27%	0.42	1.45	94%
Ventile 14	28,775	30,987	Ventile 14	109	5%	26%	0.26	1.37	96%
Ventile 15	30,988	33,201	Ventile 15	102	5%	24%	0.25	1.29	97%
Ventile 16	33,202	35,414	Ventile 16	80	4%	23%	0.19	1.23	98%
Ventile 17	35,415	37,628	Ventile 17	47	2%	22%	0.11	1.16	99%
Ventile 18	37,629	39,841	Ventile 18	53	2%	21%	0.13	1.10	99%
Ventile 19	39,842	42,055	Ventile 19	39	2%	20%	0.09	1.05	100%
Ventile 20	42,056	44,268	Ventile 20	29	1%	19%	0.07	1.00	100%

SVM	From	To	Primeiros	Fails (n)	Fails (%)	Cum Fails (%)	Lift	Cum Lift	Capt
Ventile 1	1	2,212	Ventile 1	1015	46%	46%	2.45	2.45	12%
Ventile 2	2,213	4,425	Ventile 2	1134	51%	49%	2.73	2.59	26%
Ventile 3	4,426	6,639	Ventile 3	584	26%	41%	1.41	2.19	33%
Ventile 4	6,640	8,852	Ventile 4	513	23%	37%	1.24	1.96	39%
Ventile 5	8,853	11,066	Ventile 5	401	18%	33%	0.97	1.76	44%
Ventile 6	11,067	13,279	Ventile 6	483	22%	31%	1.16	1.66	50%
Ventile 7	13,280	15,493	Ventile 7	448	20%	30%	1.08	1.58	55%
Ventile 8	15,494	17,706	Ventile 8	686	31%	30%	1.65	1.59	63%
Ventile 9	17,707	19,920	Ventile 9	715	32%	30%	1.72	1.60	72%
Ventile 10	19,921	22,133	Ventile 10	405	18%	29%	0.98	1.54	77%
Ventile 11	22,134	24,347	Ventile 11	436	20%	28%	1.05	1.49	82%
Ventile 12	24,348	26,560	Ventile 12	117	5%	26%	0.28	1.39	84%
Ventile 13	26,561	28,774	Ventile 13	134	6%	25%	0.32	1.31	85%
Ventile 14	28,775	30,987	Ventile 14	353	16%	24%	0.85	1.28	89%
Ventile 15	30,988	33,201	Ventile 15	230	10%	23%	0.55	1.23	92%
Ventile 16	33,202	35,414	Ventile 16	112	5%	22%	0.27	1.17	94%
Ventile 17	35,415	37,628	Ventile 17	104	5%	21%	0.25	1.12	95%
Ventile 18	37,629	39,841	Ventile 18	180	8%	20%	0.43	1.08	97%
Ventile 19	39,842	42,055	Ventile 19	121	5%	19%	0.29	1.04	98%
Ventile 20	42,056	44,268	Ventile 20	134	6%	19%	0.32	1.00	100%

LogReg	From	To	Primeiros	Fails (n)	Fails (%)	Cum Fails (%)	Lift	Cum Lift	Capt
Ventile 1	1	2,212	Ventile 1	1076	49%	49%	2.59	2.59	13%
Ventile 2	2,213	4,425	Ventile 2	1159	52%	51%	2.79	2.69	27%
Ventile 3	4,426	6,639	Ventile 3	1075	49%	50%	2.59	2.66	40%
Ventile 4	6,640	8,852	Ventile 4	806	36%	47%	1.94	2.48	50%
Ventile 5	8,853	11,066	Ventile 5	644	29%	43%	1.55	2.29	57%
Ventile 6	11,067	13,279	Ventile 6	505	23%	40%	1.22	2.11	63%
Ventile 7	13,280	15,493	Ventile 7	473	21%	37%	1.14	1.97	69%
Ventile 8	15,494	17,706	Ventile 8	435	20%	35%	1.05	1.86	74%
Ventile 9	17,707	19,920	Ventile 9	376	17%	33%	0.91	1.75	79%
Ventile 10	19,921	22,133	Ventile 10	367	17%	31%	0.88	1.67	83%
Ventile 11	22,134	24,347	Ventile 11	306	14%	30%	0.74	1.58	87%
Ventile 12	24,348	26,560	Ventile 12	173	8%	28%	0.42	1.48	89%
Ventile 13	26,561	28,774	Ventile 13	144	7%	26%	0.35	1.40	91%
Ventile 14	28,775	30,987	Ventile 14	96	4%	25%	0.23	1.31	92%
Ventile 15	30,988	33,201	Ventile 15	104	5%	23%	0.25	1.24	93%
Ventile 16	33,202	35,414	Ventile 16	168	8%	22%	0.40	1.19	95%
Ventile 17	35,415	37,628	Ventile 17	147	7%	21%	0.35	1.14	97%
Ventile 18	37,629	39,841	Ventile 18	109	5%	21%	0.26	1.09	98%
Ventile 19	39,842	42,055	Ventile 19	96	4%	20%	0.23	1.05	99%
Ventile 20	42,056	44,268	Ventile 20	46	2%	19%	0.11	1.00	100%

1st Scientific Paper

- **Our results clearly demonstrate that an AI approach clearly outperform a more traditional one.**
- In RF, the first ventile, i.e., the 5% of students with higher likelihood estimated by the model to fail, effectively did so in 87% of the cases. This is 4.65 the average (lift), which corresponds to 23% of the total number of students failing that academic year (cumulative captured response).
- As marking 15% students with the higher estimated likelihood of failing. In the first three ventiles, the RF and ANN, identified students with an effective fail rate of 61% and 59%, respectively, corresponding to 3.28 and 3.13 times the average. Moreover, the percentage of captured fails are 49% and 47%.
- The most two important variables are related with the academic record of the student. **One might argue that this failing a course work almost has having a criminal record, to what AA is concerned.**

PhD in Inf Man

Catarina Neves Nunes

MSc in Psychology – Social Cognition

Faculdade de Psicologia da Universidade de Lisboa

Research Projects

- Applied Social Psychology
- Environment Psychology
- Judgement and Decision Making

PhD in Inf Man

Aim: Understanding the drivers of academic achievement

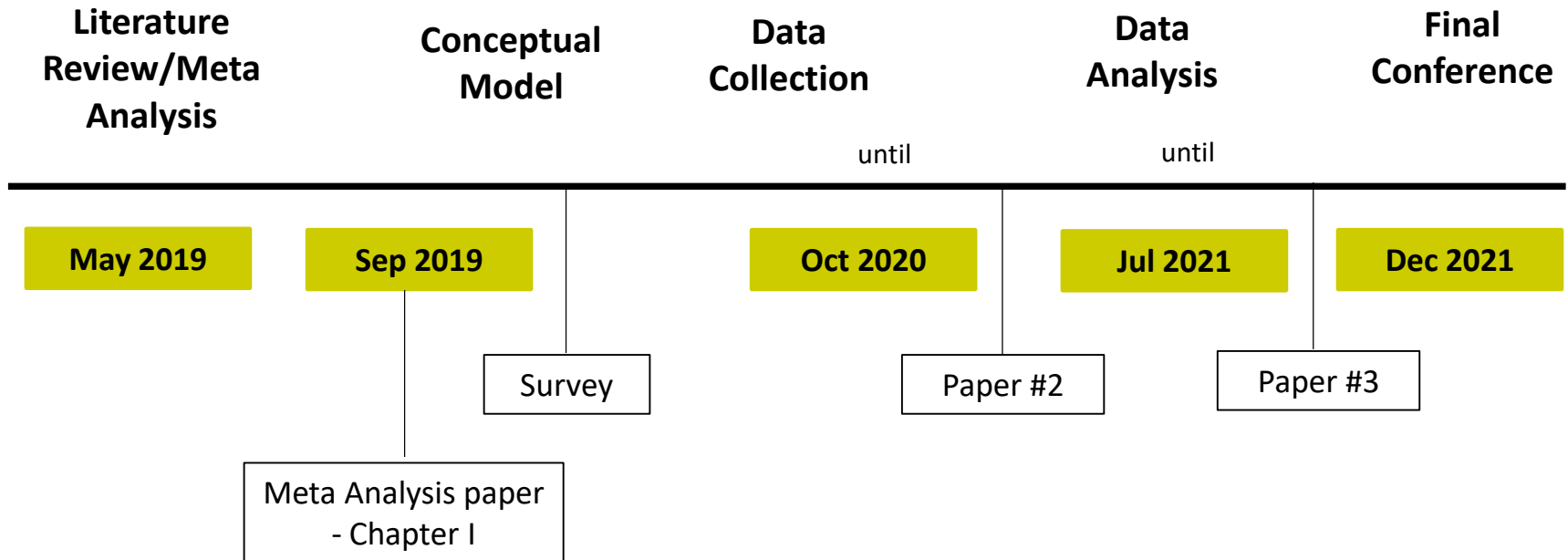
Sample: High school students from Portuguese schools

Factors: Students', parents' and schools' characteristics

Thesis Structure:

- Introduction
- Chapter I - Literature review (meta and weight analysis)
- Chapters II and III
 - Conceptual model
 - Data Collection (surveys)
 - Data Analysis
- Conclusions

PhD in Inf Man



Research Outputs:

- 3 papers in peer reviewed journals
- 1 national and 2 international conferences

PhD in Data Science

Ricardo Mendes

Bachelor degree in Economics Porto University

Master in Statistics and Information Management – NOVAims

Motivation: become a top professional/research in Data Science

PhD in Data Science

Aim

Develop high school students academic achievement predictive models

Understand and make explicit the main academic achievement antecedents

Database

DGEEC - Direção Geral de Estatística da Educação e Ciência: MISI and E360

Methods

Artificial Intelligence algorithms (logistic model trees, Decision Trees, Artificial Neural Networks and Support Vector Machines)

Scientific output

Participation in three international conferences

Two scientific articles to be submitted in peer reviewed journals

Deadline for first article submission: July 2019



NOVA

IMS

Information
Management
School

Obrigado!!!

Tiago Oliveira

<https://scholar.google.com/citations?user=RXwZPpoAAAAJ>
toliveira@novaims.unl.pt