



Universiteit  
Leiden

The Netherlands

## Reliable and fair machine learning for risk assessment

Pereira Barata, A.P.

### Citation

Pereira Barata, A. P. (2023, April 5). *Reliable and fair machine learning for risk assessment*. *SIKS Dissertation Series*. Retrieved from <https://hdl.handle.net/1887/3590289>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3590289>

**Note:** To cite this publication please use the final published version (if applicable).

**RELIABLE AND FAIR  
MACHINE LEARNING  
FOR RISK ASSESSMENT**

António Pereira Barata



Human Environment and Transport  
Inspectorate  
*Ministry of Infrastructure  
and Water Management*



**Universiteit  
Leiden**  
The Netherlands

The work in this book was funded by the Ministry of Infrastructure and Water Management through Leiden University. The author used facilities at the Leiden Centre of Data Science, the Leiden Institute of Advanced Computer Science, and the Innovation and Data Lab of the Human Environment and Transport Inspectorate.



SIKS Dissertation Series No. 2023-06

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Copyright ©2023 António Pereira Barata

Without written permission of the author, no part of this thesis may be reproduced, stored, or published in any form.

Typeset using  $\text{\LaTeX}$ , figures generated using `Matplotlib`

Printed by: Gildeprint, Enschede, the Netherlands, [gildeprint.nl](http://gildeprint.nl)

ISBN: 978-94-6419-719-8

# Reliable and Fair Machine Learning for Risk Assessment

Proefschrift

ter verkrijging van  
de graad van doctor aan de Universiteit Leiden,  
op gezag van rector magnificus prof.dr.ir. H. Bijl,  
volgens besluit van het college voor promoties  
te verdedigen op woensdag 5 april 2023  
klokke 13:45 uur

door

António Pereira Barata  
geboren te Lisboa, Portugal  
in 1989

*Promotor:* Prof.dr. H. J. van den Herik

*Co-promotores:* Dr. C.J. Veenman  
Dr. F.W. Takes

*Promotiecommissie:* Prof.dr. T.H.W. Bäck  
Prof.dr. C.M. Jonker (TU Delft)  
Prof.dr. M.V. Dignum (Umeå University)  
Prof.dr. J.N. Kok (Universiteit Twente)  
Prof.dr. H.H. Hoos (RWTH Aachen University)

*Dedicated to Alexa*



*"It is not your fault, but it is your responsibility."*

Cor J. Veenman





# Contents

<b>List of Abbreviations</b>	<b>xiii</b>
<b>List of Definitions</b>	<b>xv</b>
<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 The Inspectorate . . . . .	3
1.2 Machine Learning Preliminaries . . . . .	5
1.3 Data Quality . . . . .	7
1.3.1 Missingness . . . . .	8
1.3.2 Noise . . . . .	10
1.4 Fairness . . . . .	12
1.4.1 Performance-Fairness Trade-Off . . . . .	14
1.4.2 Addressing Fairness in Machine Learning . . . . .	15
1.5 Problem Statement and Research Questions . . . . .	16
1.5.1 Problem Statement . . . . .	16
1.5.2 Research Questions . . . . .	16
1.5.3 Research Methodology . . . . .	17
1.6 Research Contributions . . . . .	18
1.7 Thesis Overview . . . . .	20

<b>2</b>	<b>Imputation versus Missing-Indicator</b>	<b>23</b>
2.1	Minimising the Impact of Missing Data . . . . .	24
2.2	Missing Data . . . . .	26
2.2.1	Mechanisms of Missingness . . . . .	26
2.2.2	Synthesising Missing Values . . . . .	28
2.3	Related Work . . . . .	28
2.4	Method . . . . .	30
2.4.1	Missing-Indicator . . . . .	30
2.4.2	Data . . . . .	30
2.4.3	Value Removal . . . . .	32
2.4.4	Imputers and Classifiers . . . . .	32
2.5	Experiments . . . . .	34
2.5.1	Experimental Setup . . . . .	34
2.5.2	Results . . . . .	35
2.5.3	Discussion . . . . .	38
2.6	Chapter Conclusion . . . . .	39
<b>3</b>	<b>Crosslier Detection</b>	<b>43</b>
3.1	Crossliers and Miscategorisation . . . . .	44
3.2	Problem Description . . . . .	46
3.3	Related Work . . . . .	47
3.3.1	Supervised and Semi-supervised Learning . . . . .	47
3.3.2	Unsupervised Learning . . . . .	48
3.3.3	Data Quality Assurance . . . . .	49
3.4	The EXPOSE Method . . . . .	49
3.4.1	Classification Setup . . . . .	50
3.4.2	Model Selection . . . . .	50
3.4.3	Crosslier Score . . . . .	51
3.4.4	Crosslier Diagram . . . . .	52
3.5	Experiments . . . . .	53
3.5.1	Waste Transportation Setup . . . . .	53
3.5.2	Benchmark Setup . . . . .	54
3.6	Results . . . . .	56
3.6.1	Waste Transportation . . . . .	56
3.6.2	Benchmark . . . . .	58
3.7	Discussion . . . . .	60
3.8	Chapter Conclusion . . . . .	60

<b>4</b>	<b>Noise-Resilient Classifier</b>	<b>63</b>
4.1	Noise and Performance Degradation . . . . .	64
4.2	Problem Description . . . . .	67
4.2.1	Noise Interpretation . . . . .	67
4.2.2	Formal Problem Description . . . . .	67
4.3	Related Work . . . . .	68
4.3.1	Classifier Learning . . . . .	68
4.3.2	Label Noise Detection . . . . .	70
4.4	The DENOISE Method . . . . .	71
4.4.1	Learning with Sample Weights . . . . .	71
4.4.2	Posterior Estimation and Detection . . . . .	72
4.5	Experiments . . . . .	74
4.5.1	Data . . . . .	75
4.5.2	Synthetic Noise . . . . .	75
4.5.3	Evaluation . . . . .	76
4.6	Results . . . . .	77
4.6.1	Classifier Learning Task . . . . .	79
4.6.2	Label Noise Detection Task . . . . .	81
4.7	Chapter Conclusion . . . . .	82
<b>5</b>	<b>Fair Tree Classifier</b>	<b>85</b>
5.1	Algorithmic Fairness . . . . .	86
5.2	Problem Description . . . . .	87
5.3	Related Work . . . . .	88
5.3.1	Measures of Fairness . . . . .	88
5.3.2	Fair Tree Splitting Criteria . . . . .	90
5.4	The SCAFF Method . . . . .	91
5.4.1	AUC Computation . . . . .	91
5.4.2	Strong Demographic Parity . . . . .	92
5.4.3	Splitting Criterion AUC For Fairness . . . . .	93
5.4.4	Tree Construction . . . . .	95
5.5	Experiments . . . . .	96
5.5.1	Datasets . . . . .	96
5.5.2	Experimental Setup . . . . .	97
5.6	Results . . . . .	98
5.6.1	Binary Sensitive Attribute . . . . .	98
5.6.2	Multiple and Multicategorical Cases . . . . .	100
5.6.3	Relationship with Demographic Parity . . . . .	102
5.7	Chapter Conclusion . . . . .	104

<b>6 Conclusions</b>	<b>107</b>
6.1 Answers to the Research Questions . . . . .	107
6.2 Answer to the Problem Statement . . . . .	110
6.3 Future Research . . . . .	111
<b>References</b>	<b>115</b>
<b>Summary</b>	<b>135</b>
<b>Samenvatting</b>	<b>141</b>
<b>Curriculum Vitae</b>	<b>147</b>
<b>Publications</b>	<b>149</b>
<b>Acknowledgements</b>	<b>151</b>
<b>SIKS Dissertation Series</b>	<b>155</b>

# List of Abbreviations

AI	artificial intelligence
AP	average precision
AUC	area under the receiver operating characteristic curve
CV	cross-validation
EU	European Union
FAHT	fairness-aware Hoeffding tree
FPR	false positive rate
IF	isolation forest
ILT	Inspectie Leefomgeving en Transport
KNNC	k-nearest neighbours classifier
LOF	local outlier factor
LR	logistic regression
MAR	missing at random
MCAR	missing completely at random
MNAR	missing not at random
NAR	noise at random
NCAR	noise completely at random
NNAR	noise not at random
NWF	non-white female
NWM	non-white male
OvR	one-versus-rest
PS	problem statement
RQ	research question
SCAFF	splitting criterion AUC for fairness
SVMC	support vector machine classifier
TPR	true positive rate
WF	white female
WM	white male
XGBC	extreme gradient boosting classifier
XGBR	extreme gradient boosting regression



# List of Definitions

1.1	AI systems . . . . .	1
1.2	Trustworthy AI . . . . .	2
1.3	Risk assessment . . . . .	3
1.4	Machine learning . . . . .	5
1.5	Target function (model) . . . . .	5
1.6	AUC . . . . .	7
1.7	Fairness . . . . .	13
1.8	Demographic parity . . . . .	13
1.9	Strong demographic parity . . . . .	13
1.10	Data quality . . . . .	8
1.11	Missingness . . . . .	8
1.12	MCAR . . . . .	9
1.13	MAR . . . . .	9
1.14	MNAR . . . . .	9
1.15	Imputation . . . . .	9
1.16	Missing-indicator . . . . .	10
1.17	Noise . . . . .	10
1.18	Feature noise . . . . .	11
1.19	Label noise . . . . .	11
1.20	NCAR . . . . .	11
1.21	NAR . . . . .	11
1.22	NNAR . . . . .	12
1.23	CRISP-DM . . . . .	17
2.1	Class balance . . . . .	31
2.2	Warm start . . . . .	33



3.1	Crosslier . . . . .	43
3.2	Crosslier diagram . . . . .	43
3.3	Outlier . . . . .	45
3.4	Anomaly . . . . .	45
3.5	Crosslyingness . . . . .	46
4.1	Belongingness . . . . .	66
4.2	Noisy sample . . . . .	67
4.3	Sample weighting . . . . .	69
5.1	Threshold-dependent fairness measure . . . . .	86
5.2	Threshold-independent fairness measure . . . . .	87
5.3	Equal opportunity . . . . .	89
5.4	Equalised odds . . . . .	89
5.5	Sensitive AUC . . . . .	93
5.6	Orthogonality parameter $\Theta$ . . . . .	93

# List of Figures

1.1	Supervised classifier learning . . . . .	6
1.2	CRISP-DM . . . . .	18
2.1	Imputation through missing-indicator . . . . .	30
2.2	KNNC mean performances across missing proportions . . . . .	36
2.3	SVMC mean performances across missing proportions . . . . .	37
2.4	XGBC mean performances across missing proportions . . . . .	38
3.1	Anomaly, outlier, and crosslier . . . . .	45
3.2	Crosslier detection . . . . .	46
3.3	Distinction between outlier and crosslier . . . . .	48
3.4	Crosslier diagrams of four waste categories . . . . .	58
3.5	Crosslier detection performance across different methods . . . . .	59
4.1	Learning performances with noisy data (aggregated) . . . . .	78
4.2	Detection performances of noisy samples (aggregated) . . . . .	80
5.1	Effect of the orthogonality parameter $\Theta$ over SCAFF scores . . . . .	94
5.2	Computing AUC values for SCAFF . . . . .	96
5.3	Performance-fairness across methods per dataset . . . . .	99
5.4	Performance-fairness for the multiple and intersectional cases . . . . .	101
5.5	Effect of orthogonality $\Theta$ over demographic parity. . . . .	103



# List of Tables

2.1	Three different missing mechanisms . . . . .	27
2.2	Summary of dataset characteristics . . . . .	31
2.3	KNNC Wilcoxon p-values . . . . .	36
2.4	SVMC Wilcoxon p-values . . . . .	37
2.5	XGBC Wilcoxon p-values . . . . .	38
3.1	Datasets retrieved for crosslier simulations. . . . .	55
3.2	Model performances per waste category. . . . .	57
4.1	Datasets retrieved for noise simulations . . . . .	75
4.2	Learning performances with noisy data (per dataset) . . . . .	79
4.3	Detection performances of noisy samples (per dataset) . . . . .	81
5.1	Correlation between (strong) demographic parities . . . . .	104