

IDENTIFICATION OF THE RELATIONS BETWEEN SYNTHESIS
CONDITIONS AND COMPRESSIVE STRENGTH IN GEOPOLYMER
SYSTEMS VIA STATISTICAL LEARNING METHODS

by

Kadriye Deniz Köse

B.S., Chemical Engineering, Boğaziçi University, 2017

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University
2019

ACKNOWLEDGEMENTS

This study was supported by Boğaziçi University BAP projects: 12840 and 15201 .

I would firstly like to thank my thesis advisor Assoc. Prof. Burak Alakent who was always willing to teach me every new, for me, notion involved in the study. I am grateful for the endless hours he spent controlling my Thesis report and MATLAB codes, and explaining me every single line we needed to add. I would also like to thank my thesis co-advisor Assoc. Prof. Sezen Soyer Uzun who was always available to answer any question of mine. Both, supported me not only regarding this study but my life generally.

In addition, I would like to thank my thesis jury members Prof. Ramazan Yıldırım and Assoc.Prof. Erdem Günay who accepted to be my jury members and enlightened me with their valuable questions. I would also like to express my gratitude to Prof. Mehmet C. Çamurdan and Prof. Öner Hortaçsu who have supported me and Ekin in various aspects of our lives.

I feel very lucky for being the daughter of my great parents Yılşen Küçük-Hacı and Metin Köse who have supported me in everything, always encouraging me to discover new worlds. I would also love to thank my brother Mehmet Emin Köse who has taught me to seek for improvement through my decisions. I should mention that my grandmother Şefika has always been my second mother and a great assistant.

I would love to thank my dear Ekin Yurdakul who is beyond a sister to me. Thanks for the magnificent university years and for just being in my life which provides me with confidence every new day. Thank you for giving special meanings to clove, hedgehog, acibadem cookies and such trivial things which charge my life with brighter colors. God bless the day that we could not find KB 432(?)!

Dear Gülten Çelebi, thanks for your comfortable arms where lots of my spiritual wounds have been treated and for all the hours you spent formatting our messy reports. Its great to have met both the Black Sea's vicious wave part of you as well as the soft friendly persona that it covers.

And, dear Onur Can Boy I thank you because even though our foolish fights you still stand by me each and every time. Thanks for your older brother attitude and not kicking me out of the Uber during my argument with the police officers.

I should also acknowledge the substantial support of dear Sema Çömez who even during her own stressful days was all around me. I feel her helping hands from my washing machine to every call I currently make. Moreover, I appreciate the patience of Hilal Yildiz during my long complainant discussions and all the food she prepared.

My dear lab mates Aysun Urhan and Melike Dilara Göreke were under all occasions eager to help for anything. I appreciate your valuable support. Dear Semra Tuncay thanks for the pleasant converastions during design projects and dear Merve Yüce thanks for the litres of tea and hospitality in your lab! Dear Murat Düzgünoğlu and Yakup Bal your assistance is so valuable for us. The warm company of Hürrem, Korsan, Sefil and Onur's daughter is acknowledged, too.

Yet, I would like to express my appreciation to Dilara Aksoy and Merve Bavra. Without you I could not have survived in this jungle which you have transformed into home. I would love to thank H.M. Andriniaina R. Laisola for not allowing me to give up and return to my island and Theresa abla, Louange L. A L'Eternel and Hasan Yusuf for being my guardian angels.

Furthermore, I would like to express my gratitude to Arash M. Rezaeinzhad for the long Tarkovsky film analyses, and all Kandilli friends underlying Taylan Atakuru, Ulduz

Sobhiahshar, Massoud Pourmandi and Nesrin Conker who turned my data collection period into the most enjoyable dormitory nights.

I would reserve this last special paragraph to my lifetime teacher Aikaterini Karamitsiou and my lifetime friend Aikaterini Karamichali. The always innovative vision of the first and childish care of the latter are on my mind despite the passing years.

ABSTRACT

IDENTIFICATION OF THE RELATIONS BETWEEN SYNTHESIS CONDITIONS AND COMPRESSIVE STRENGTH IN GEOPOLYMER SYSTEMS VIA STATISTICAL LEARNING METHODS

Geopolymers are inorganic amorphous materials, which can be used in construction industry, and their production is characterized by low CO₂ emission. Compressive strength (CS) is a durability measure, which, in the case of geopolymers, depends on various manufacturing factors in a rather complex functional relation. The aim of this study is to develop a decision tree and extract reliable decision rules based on it. A database consisting of 879 data points and 19 predictors is constructed based on available publications. Forward Selection (FS), Backward Elimination (BE) and Rough Set theory are used as feature selection algorithms. The decision tree parameters (maximum number of splits, S_{max} and minimum parent size P_{min}) and number of relevant predictors for FS and BE and threshold parameter (eps), which is introduced for reduct set formation, are optimized via 5-fold CV with 20 repetitions. The reliability of the obtained decision rules is expressed via Stability index (SI) whereas the accuracy of the models via measures, such as RMSE. Minimum test RMSE is decreased up to 10.0 MPa for $S_{max} = 150$ BE based tree, but the interpretation of such large trees is demanding. Hence, a smaller sized ($S_{max} = 60$) reduct based tree is used for rule extraction. The CV RMSE and SI, test RMSE and R^2 of the resulting tree is 13.2 MPa, 0.79, 13.2 MPa and 0.64 respectively. Based on this tree, the most important predictor affecting geopolymer CS is found to be the type of aluminosilicate precursor material. Two material groups are identified corresponding to higher and lower CS. The data space is subsequently divided, and relations (linear regression, local models etc.) regarding smaller subspaces are developed. For higher CS group, water/Al, Na/Al and ambient curing time have a significant effect, whereas for lower CS compounds, purity and heat curing time are also found to be effective. Moreover, the effect of the predictors on CS can vary based on the precursor types, showing the presence of significant predictor interactions.

ÖZET

JEOPOLİMER SİSTEMLERİNDE SENTEZ KOŞULLARI VE BASINÇ DİRENCİ ARASINDAKİ İLİŞKİLERİN İSTATİSTİKSEL ÖĞRENME YÖNTEMLERİ ARACILIĞIYLA TANIMLANMASI

Üretiminde CO₂ salınımının düşük olması jeopolimerleri inşaat endüstrisinde kullanılabilecek uygun bir inorganik ve amorf malzeme yapmaktadır. Çeşitli üretim koşullarının karmaşık etkileriyle belirlenen basınç direnci (CS), jeopolimerlerin önemli bir dayanıklılık ölçüsüdür. Bu çalışmanın amacı, üretim koşulları ile basınç direnci arasında güvenilir kurallar elde etmek için bir karar ağacı oluşturmaktır. Bu nedenle, mevcut yayımları kullanarak, 879 veri noktası ve 19 değişkenden oluşan bir veri tabanı oluşturuldu. Gereksiz değişkenleri elemek için İleri Seçim (FS), Geri Eliminasyon (BE) ve Kaba Küme teorisi algoritmaları kullanıldı. Karar ağacı parametreleri (S_{max} ve P_{min}), indirgenmiş küme oluştururken bu çalışmada eklenen eşik parametresi (eps), ve FS/BE algoritmalarında uygun değişken sayısı 20 tekrar ile 5 kat çapraz doğrulama (CV) ile optimize edildi. Yapılan ağaçlardan elde edilen kuralların güvenilirliği kararlılık endeksi (SI) ile ölçüldü. BE tabanlı derin ağaçların ($S_{max} = 150$) test RMSE değeri 10.0 MPa'a kadar düşse de, FS/BE değişken seçmeli büyük ağaçlardan elde edilen kuralların kararlı olmadığı görüldü. Bu nedenle, kural oluşturmak için daha küçük boyutlu ($S_{max} = 60$) indirgenmiş küme bazlı bir ağaç kullanıldı. Oluşturulan bu ağacın CV RMSE ve SI, test RMSE ve R^2 sırasıyla 13.2 MPa, 0.79, 13.2 MPa ve 0.64'e eşit. Bu ağaca dayanarak düşük ve yüksek CS değerlerine karşılık gelen iki alüminosilikat prekürsor malzeme grubu tanımlandı. Çıkan ağacın yorumu için 2-boyutlu değişken uzayından doğrusal regresyon ile yerel modeller geliştirildi. Su/Al ve Na/Al oranları ve çevre koşullarında kürleme süresi yüksek CS grubunu, saflık ve yüksek sıcaklıkta kürleme süresi de düşük CS grubunu etkilediği görüldü. Ayrıca, değişkenlerin CS üzerindeki etkisinin prekürsor malzeme tipine göre değiştiği ve bunun değişkenler arası önemli etkileşimlerin varlığına işaret ettiği görüldü.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	vi
ÖZET	vii
LIST OF FIGURES	x
LIST OF SYMBOLS	xviii
LIST OF ACRONYMS/ ABBREVIATIONS	xxi
1. INTRODUCTION	1
2. THEORETICAL BACKGROUND	4
2.1. Geopolymers	4
2.1.1. Geopolymer Synthesis	5
2.1.2. Aluminosilicate Sources for Geopolymerization.....	6
2.1.3. Factors Affecting Compressive Strength of Geopolymers	7
2.2. Statistical Learning.....	9
2.2.1. Linear Regression	10
2.2.2. Classification and Regression Trees	10
2.2.3. Feature Selection Algorithms	14
2.2.4. Rough Set Theory	16
2.2.5. Stability in Feature Selection Algorithms.....	18
2.3. Application of Statistical Learning Methods in Materials Literature	20
3. METHODOLOGY	24
3.1. Exploratory Analysis of the Dataset	24
3.2. Construction of Regression Models	35
4. RESULTS AND DISCUSSION.....	38
4.1. Regression Trees with $S_{\max} = 30$	38
4.1.1. Constructing Regression Trees via FS and BE with $S_{\max} = 30$	38

4.1.2.	Constructing Regression Trees via Reduct Sets with $S_{\max} = 30$	50
4.2.	Regression Trees with $S_{\max} = 60$	56
4.2.1.	Constructing Regression Trees via FS and BE with $S_{\max} = 60$	56
4.2.1.	Construction Regression Trees via Reduct Sets with $S_{\max} = 60$	66
4.3.	Regression Trees with $S_{\max} = 150$	71
4.3.1.	Constructing Regression Trees via FS and BE with $S_{\max} = 150$	71
4.3.2.	Constructing Regression Trees via Reduct Sets with $S_{\max} = 150$	80
4.4.	A Summary of the Models Constructed	85
4.5.	Rule Extraction from the Regression Tree	87
5.	CONCLUSION AND RECOMMENDATIONS	100
5.1.	Conclusion	100
5.1.	Recommendations	101
	REFERENCES	102
	APPENDIX A: $N_{19}(60,14)$ FULL REGRESSION TREE	107

LIST OF FIGURES

Figure 2.1. Change in the number of publications on geopolymers since 1995.....	4
Figure 2.2. Two-part polymer synthesis from metakaolin	5
Figure 3.1. Per year distribution of the articles used	24
Figure 3.2. Material type distribution.....	29
Figure 3.3. (a) Si/Al, (b) Na/Al and (c) Water/ Al molar ratio distributions.....	30
Figure 3.4. Purity (SiO_2 and Al_2O_3 over raw material mass) distribution.....	31
Figure 3.5. Distributions of aggregate (a) presence (b) proportion, and (c) particle size.....	32
Figure 3.6. Distributions of additive (a) presence and (b) percentage.....	33
Figure 3.7. Geopolymer type distribution.....	33
Figure 3.8. (a) Maximum and (b) average particle size distributions.....	34
Figure 3.9. Distributions of the presence of (a) heat and (b) ambient curing.....	34
Figure 3.10. Distributions of (a) heat curing temperature, (b) ambient curing temperature, (c) heat curing time, and (d) ambient curing time.....	35
Figure 4.1. Surface plot of RMSE obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 30.....	39
Figure 4.2. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via FS; (d) RMSE, (e) MAE, and (f) stability index with respect to minimum parent size in full trees constructed for $S_{\max} = 30$	41
Figure 4.3. (a) Number of predictors included in the trees versus number of features selected by the FS search for FS(30,20) and FS(30,22), and (b) stability of the trees with	

respect to the levels for $FS_{12}(30,20)$, $FS_8(30,22)$, and full models $N_{19}(30,32)$ and $N_{19}(30,24)$.	42
Figure 4.4. Surface plot of RMSE obtained via CV with respect to minimum parent size parameter and number of features selected by BE for trees with maximum number of splits equal to 30.	43
Figure 4.5. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via BE for $S_{max} = 30$.	43
Figure 4.6 (a) Number of predictors included in the trees versus number of features selected by the FS search for $BE(30,32)$ and $FS(30,20)$ and (b) Tree level stability for $BE_{11}(30,32)$, $FS_{12}(30,20)$ and $N_{19}(30,32)$.	44
Figure 4.7. (a) Mean number of features included in the tree versus number of features included in the search for $S_{max} = 30$ and (b) Mean tree feature stability for $S_{max} = 30$.	45
Figure 4.8. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{max} = 30$.	46
Figure 4.9. Comparison of predictions via $BE_{11}(30,32)$ model with the experimental results on the test set.	48
Figure 4.10. Feature importance in the tree constructed via $BE_{11}(30,32)$ using the training set.	49
Figure 4.11. Feature importance in the tree constructed via $BE_{11}(30,32)$ using the whole dataset.	49
Figure 4.12. (a) RMSE, (b) MedAE and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $\epsilon = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{max} = 30$.	51
Figure 4.13. FS, BE and Reduct set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{max} = 30$.	52

Figure 4.14. Comparison of predictions via $R_{5e-4}(30,24)$ model with the experimental results on the test set.....	53
Figure 4.15. Feature importance in the tree constructed via $R_{5e-4}(30,24)$ using the training set.	54
Figure 4.16. Feature importance in the tree constructed via $R_{5e-4}(30,24)$ using the whole dataset.	54
Figure 4.17. Surface plot of RMSE obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 60.....	56
Figure 4.18. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via FS; (d) RMSE, (e) MAE, and (f) stability index with respect to minimum parent size in full trees constructed for $S_{\max} = 30$ and 60.	57
Figure 4.19. (a) Number of predictors included in the trees versus number of selected by the FS search FS(30,20) and FS(60,14), and (b) tree level stability for FS ₁₂ (30,20), FS ₁₃ (60,14), N ₁₉ (30,32) and N ₁₉ (60,14).	58
Figure 4.20. Surface plot of RMSE obtained via CV with respect to minimum parent size parameter and number of features selected by BE for trees with maximum number of splits equal to 60.....	59
Figure 4.21. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via BE for $S_{\max} = 30$ and 60.....	60
Figure 4.22 (a) Number of predictors included in the trees constructed during search versus number of features selected by final tree for FS and BE minimum RMSE decision trees for $S_{\max} = 30$ and 60 and (b) Stability of the tree levels for FS, BE and full models with minimum RMSE for $S_{\max} = 30$ and 60.....	61

Figure 4.23. (a) Mean number of features selected by the tree versus number of features included in the search for $S_{\max} = 30$ and (b) Mean tree feature stability for $S_{\max} = 30$ and 60.....	62
Figure 4.24. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 60$	63
Figure 4.25. Comparison of predictions via $FS_{13}(60,14)$ model with the experimental results on the test set.....	64
Figure 4.26. Feature importance in the tree constructed via $FS_{13}(60,14)$ using the training set. ...	65
Figure 4.27. Feature importance in the tree constructed via $FS_{13}(60,14)$ using the whole dataset.	65
Figure 4.28. (a) RMSE, (b) MedAE and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $\epsilon = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{\max} = 60$	66
Figure 4.29. FS, BE and Reduct set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 60$	67
Figure 4.30. Comparison of predictions via $R_{1e-4}(60,12)$ model with the experimental results on the test set.....	69
Figure 4.31. Feature importance in the tree constructed via $R_{1e-4}(60,12)$ using the training set.	70
Figure 4.32. Feature importance in the tree constructed via $R_{1e-4}(60,12)$ using the whole dataset.	70
Figure 4.33. Surface plot of RMSE obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 150.....	72

Figure 4.34. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via FS; (d) RMSE, (e) MAE, and (f) stability index with respect to minimum parent size in full trees constructed for $S_{\max} = 30, 60$ and 150	73
Figure 4.35. (a) Number of predictors included in the trees constructed during search versus number of features selected by final tree for FS(30,20), FS(60,14) and FS(150,4) and (b) Tree level stability for FS ₁₂ (30,20), FS ₁₃ (60,14), FS ₁₅ (150,4) full models N ₁₉ (30,32), N ₁₉ (60,14) and N ₁₉ (150,6).	74
Figure 4.36. CV RMSE 3D Surface Plot of Models with various Pmin and Number of Features Selected by BE for $S_{\max} = 150$	75
Figure 4.37. (a) RMSE, (b) MAE, and (c) stability index with respect to features selected via BE for $S_{\max} = 30, 60$ and 150	75
Figure 4.38 (a) Number of predictors in the trees versus number of features selected by FS/ BE search for $S_{\max} = 30, 60$ and 150 and (b) Tree level stability for $S_{\max} = 30, 60$ and 150	76
Figure 4.39. (a) Mean number of features selected by the tree versus number of features included in the search for and (b) Mean tree feature stability for $S_{\max} = 30, 60$ and 150	76
Figure 4.40. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 150$	77
Figure 4.41. Comparison of predictions via FS ₁₅ (150,4) model with the experimental results on the test set.	78
Figure 4.42. Feature importance in the tree constructed via FS ₁₅ (150,4) using the training data set.	79
Figure 4.43. Feature importance in the tree constructed via FS ₁₅ (150,4) using the whole dataset.	79

Figure 4.44. (a) RMSE, (b) MedAE and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $\text{eps} = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{\max} = 150$	80
Figure 4.45. FS, BE and rough set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 150$	81
Figure 4.46. Comparison of predictions via $R_{5e-5}(150,8)$ model with the experimental results on the test set.....	83
Figure 4.47. Feature importance in the tree constructed via $R_{5e-5}(150,8)$ using the training set.	83
Figure 4.48. Feature importance in the tree constructed via Feature importance in the tree constructed via $R_{5e-5}(150,8)$ using the whole dataset.....	84
Figure 4.49. A tabulated summary of the models and their performances in CV and the test set	86
Figure 4.50. The regression tree constructed via $R_{1e-4}(60,12)$	90
Figure 4.51. Local regression model surfaces constructed for left hand side of the $R_{1e-4}(60,12)$ regression tree.	92
Figure 4.52. Local regression model surfaces constructed for right hand side of the $R_{1e-4}(60,12)$ regression tree.....	99
Figure A.1. $N_{19}(60,14)$ Full Regression Tree	108

LIST OF TABLES

Table 3.1. Summary of the predictors considered.	26
Table 3.2. List of materials used as precursors and their abbreviations.	27
Table 4.1. Summary of CV and Test Results of Minimum RMSE and MedAE models	47
Table 4.2. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 30$	52
Table 4.3. CV and Test Result Comparison between most stable reduct based decision tree $R_{5e-4}(30,24)$ and best wrapper tree $BE_{11}(30,32)$ for $S_{\max} = 30$	55
Table 4.4. Summary of CV and Test Results of Minimum RMSE and MedAE models for $S_{\max} = 60$	63
Table 4.5. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 60$	68
Table 4.6. CV and Test Result Comparison between most stable Reduct based decision tree $R_{1e-5}(60,12)$ and minimum RMSE FS tree $FS_{13}(60,14)$ for $S_{\max} = 60$	71
Table 4.7. Summary of CV and Test Results of Minimum RMSE and MedAE models ($S_{\max} = 150$).	77
Table 4.8. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 150$	81
Table 4.9. CV and Test Result Comparison between minimum RMSE Reduct based decision tree $R_{5e-5}(150,8)$, minimum RMSE FS tree $FS_{15}(150,4)$ and full tree $N_{19}(150,6)$ for $S_{\max} = 150$	84
Table 4.10. Material groupings used in the regression tree.	91
Table 4.11. Predictor that the marker size in each subfigure represents.	91
Table 4.12. Statistics of the local models constructed for the left-hand side of the $R_{1e-4}(60,12)$ regression tree models	95

Table 4.13. Local regression model coefficients, their parameters estimates and p-values constructed from the left hand side of the $R_{1e-4}(60,12)$ regression tree.	96
Table 4.14. Local regression model coefficients, their parameters estimates and p-values constructed from the right hand side of the $R_{1e-4}(60,12)$ regression tree.....	97
Table 4.15. Statistics of the local models constructed for the right-hand side of the $R_{1e-4}(60,12)$ regression tree models	98
Table A.1. Denotation and Material Group Mapping for $N_{19}(60,14)$ Regression Tree.....	107

LIST OF SYMBOLS

A	Subset of X
$\mathbb{A}$	Feature Subset
B	Subset of A
$\underline{B}(X)$	Lower Limit of B
$\overline{B}(X)$	Upper Limit of B
$\underline{B'}(X)$	Reduct of B
C	Categories / Subset of A
c_{ij}	Inputs in $ U \times U $
$c\left(\frac{i}{j}\right)$	Probability of missclassifying a category j
$d(x)$	Function constructed based on $\mathcal{L}$
d_a	Decision attribute
d_s	Number of samples
$E(p)$	Entropy of the probability distribution of a feature p
$G(p, t)$	Gain index of the test set T and feature p
$g(t)$	Gini index at the node t
$IND(B)$	Indiscernability Relation for B
j_n	Measurement vector
k	Number of Features Selected
KI	Kuncheva Index
$\bar{K}$	Average number of predictors selected
$\mathcal{L}$	Training data set

$\mathcal{L}_1$	Training sample data set
$\mathcal{L}_2$	Testing sample data set
$\mathcal{L}_V$	Vth sample data set
M	Number of predictors
n	Number of different levels/ features
N	Number of Samples
N_1	Training Sample Size
N_2	Testing Sample Size
N_V	Vth Set Sample Size
p_i	Probability distribution of instances of feature p
$\hat{p}_j$	Number of predictors selected at each training set
$p\left(\frac{j}{t}\right)$	Probability of category j at the node t
$P_{\min}$	Minimum parent size
$POS_B(d_a)$	Positive Region of the decision attribute d_a based on subset B
p_s	Number of predictors selected in the search algorithm
p_T	Number of predictors in the final tree
$\bar{p}_t$	Average number of predictors in the cross validated trees
$q_{.90}$	0.90th quantile
r	Random hypergeometric variable
$r_{i,j}$	Number of common elements for set i and j
r_p	Correlation coefficient for the predicted data
R_p^2	Coefficient of determination for the predicted data
$R * (d)$	Number of missclassified instances

R^{CV}	CV-misclassification estimate
R^{TS}	Misclassified instances used for testing
S	Information System
S_j^2	Measure defined for stability index calculations
$S_{\max}$	Maximum number of splits
SI_{Tree}	Stability index of the regression tree
T	Test set
U	Universe of all objects (Rough Set Theory)
$w(a)$	Weight of attribute
X	Vector of independent features / Subset of U
x	Instance belonging to X
$\mathbf{X}$	Sample Vector
$\mathbf{x}$	Instance belonging to $\mathbf{X}$
Y	Dependent Variable (linear regression)
$\mathcal{Y}$	Expected value of the intersection of A and B
$\alpha(x)$	Attribute value of instance x in information system S
β_i	Regression Coefficient (linear regression)
γ	Interaction coefficient (linear regression)
ε	Error (linear regression)/ Tolerance value in DM
$\mu(C.B)$	Measure of dependency of B and C
$\sigma_{(A,d_a)}B$	Significance of B for the decision attribute d_a in the set A
$\chi(\cdot)$	Function which equals 1 if the expression enclosed is true or 0 otherwise

LIST OF ACRONYMS/ ABBREVIATIONS

Add	Presence of Additive (0/1)
Aggr	Presence of Aggregate (0/1)
AggrProp	Aggregate Proportion
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AvParSize	Average particle size
BE	Backward Elimination
CAC	Calcium Activated Concrete
CART	Classification and Regression Tree
CG	Coal Gangue
CS	Compressive Strength
CV	Cross-Validation
DM	Discernibility Matrix
EMDR	Electrolytic Manganese Dioxide
FA	Fly Ash
FS	Forward Selection
L	Lime
MAE	Mean Absolute Error
<i>maxAE</i>	Maximum absolute error
medAE	Median Absolute Error
MGAE	Mean Gross Absolute Error
<i>MGAE</i>	Mean gross absolute error
MK	Metakaolin
OPC	Ordinary Portland Cement
ParSize	Maximum particle size

POFA	Palm Oil Fly Ash
PSPP	Process, Structure and Property
RHA	Rice Husk Ash
RM	Red Mud
RMSE	Root Mean Squared Error
S	Slag
SD	Search direction (forward selection/backward elimination)
SE	Standard Error
SE	Standard Error
SF	Silica Fume
SMC	Soft Marine Clay
Type	Geopolymerization mechanism type

1. INTRODUCTION

A notable proportion of worldwide CO₂ emissions results from ordinary Portland cement (OPC) production plants. In contrast, geopolymers, characterized by affordability and low CO₂ emission during their production, are considered as future cement alternatives (Luukkonen *et al.*, 2018). Although other names such as low-temperature aluminosilicate glass and hydroceramics are also used, the term “geopolymer” invented by Davidovits (Davidovits, 1991) is the name most widely accepted to describe alkali aluminosilicate binders, produced mainly by alkali activation of an aluminosilicate source. In addition to alkali activation, geopolymers can be rarely formed by acid activation of aluminosilicate sources. (Han *et al.*, 2018). Geopolymers differ from significant amount of calcium-containing, alkali-activated cements, invented by Glukhovsky (1950s) (Duxson *et al.*, 2005). As construction materials, main advantages of geopolymers are their durability and outstanding mechanical properties, mainly derived from their X-ray amorphous structures at mild temperatures. Their environmentally friendly nature is strengthened from the fact that industrial wastes such as fly ash and rice husk ash can be utilized for their production. Furthermore, their inorganic nature implies fire resistance and thermal stability (Duxson *et al.*, 2005).

Geopolymer science is not only a research area, but geopolymer mortars are produced and used in construction and equivalent areas. Sea-shore structures, as bricks requiring lower production temperatures, thermal insulation and passive cooling systems by absorbing moisture and regulating the overall daytime temperature are examples of their use (Renca.org, n.d.). Their mechanical performance is illustrated in Toowoomba Wellcamp Airport and Global Change Institute Building constructions, where geopolymer mortars are

part of the building structure (Wagners, n.d.). As a result, extensive use of these materials in everyday life makes the need to understand their characteristics indispensable.

Statistical and machine learning methods nowadays are widely used to transform huge amount of data to valuable information. This information can vary from data in market basket analysis (Sagin and Ayvaz, 2018) to compositional and structural information on material science in case of geopolymer characteristics analysis (Yeh, 1998). One of the most popular machine learning methods is classification and regression tree algorithm. The primary advantage of classification and regression trees lies in ease of understanding the derived relations (Safavian & Landgrebe, 1991). Construction of trees involves partitioning the complex data space into simpler spaces, and flexibility in the construction procedure, i.e. freedom in choosing different predictors for various space partitionings, is another of their benefits. . In this way, a natural feature selection is enclosed in the decision tree algorithm. There are also explicit feature selection algorithms, such as Forward Selection (FS) and Backward Elimination (BE), which makes it possible to remove redundant information predictors, thus improve prediction accuracy (Karagiannopoulos *et al.*, 2007). Besides these conventional algorithms, logic based methods, such as reduct sets in rough set theory can be used for feature selection (Hu *et al.*, 2003). While it is possible to improve prediction accuracy using a combination of these methods, generalizability of the obtained rules is another issue, i.e. small changes in the training set of tree may yield significantly different set of decision rules. Hence, both accuracy and “stability” of the rules are important measures of the reliability of the derived rules.

In the current study, experimental data from the recently published literature are collected and relations, i.e. explanatory models, between processing conditions and compressive strength of geopolymers are determined in the form of regression trees. Hence, In the study, feature selection methods are implemented with decision trees to obtain acceptable prediction accuracy and stable models, robust to perturbations in the dataset. A detailed analysis of the resulting tree is presented to shed light onto how experimental processing may affect mechanical performance.

The structure of the thesis can be summarized as follows. Continued from the brief introduction in Section 1, theoretical backgrounds of geopolymer production, statistical/machine learning methods, including feature selection and decision tree construction methods, and application examples of the statistical methods in material science are presented in Section 2. A detailed description of the methodology followed in the current thesis is given in Section 3. In Section 4 results of various models are discussed, and a final model tree is used for deriving rules for different regions of the feature space. Conclusion and Recommendations will be given in the Section 5.

2. THEORETICAL BACKGROUND

In this section, a review of the theoretical background of the study is given. First, the materials analyzed in the current study are introduced, and the factors affecting their mechanical properties are discussed. Then, a short history of statistical learning with emphasis on regression/decision trees, feature selection methods, rough set theory, and measures of stability are presented. Lastly, a literature survey on the use of statistical and machine learning methods on material science is provided.

2.1. Geopolymers

The increased interest on geopolymers as a research field is evident, as the number of publications through the years has increased significantly. Figure 2.1 shows the number of geopolymer keyword related publications, which could be accessed through ScienceDirect. As seen, there are numerous studies regarding optimization of synthesis and curing procedures for obtaining enhanced geopolymer performance.

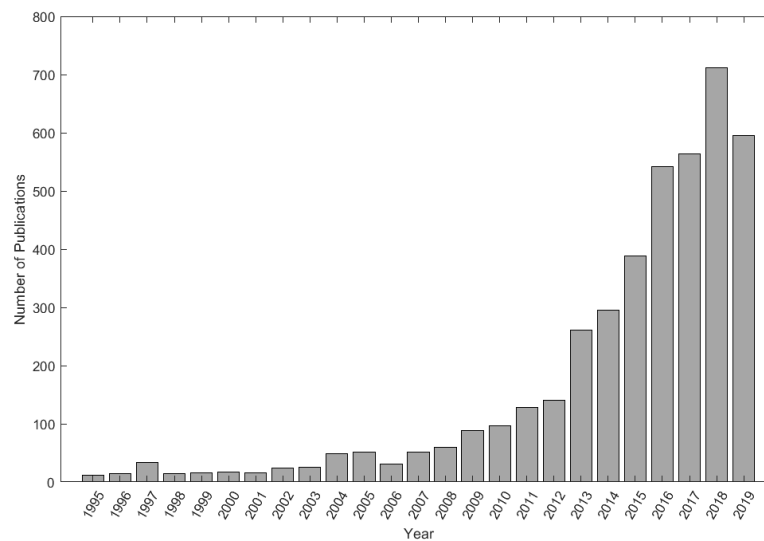


Figure 2.1. Change in the number of publications on geopolymers since 1995.

2.1.1. Geopolymer Synthesis

The traditional method of geopolymer production comprises a two-part synthesis, which includes dissolution and polycondensation steps (Figure 2.2): the raw material (aluminosilicate precursor) is activated by a high pH (usually viscous) activating solution, and a three-dimensional polymeric system is formed (Rovnaník, 2010).

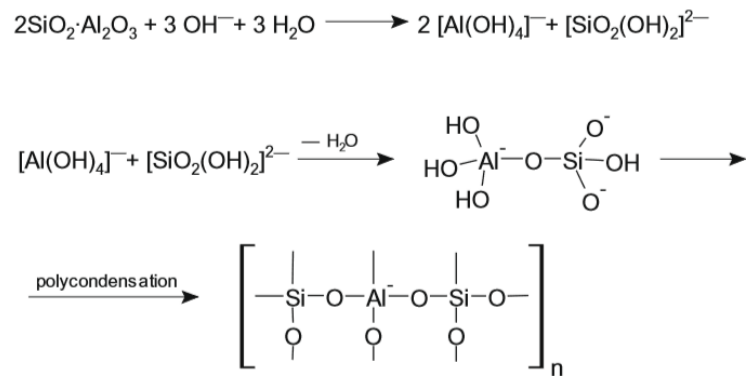


Figure 2.2. Two-part polymer synthesis from metakaolin (Rovnaník, 2010).

The requirement of a highly basic activating solution and the difficulty in controlling the viscosity of the preliminary geopolymer gel, during condensation, are the main drawbacks of two-part geopolymer synthesis which aids the development of modern one-part or just-add-water geopolymers. One-part manufacturing requires the steps of dry mixing of aluminosilicate source (e.g. metakaolin, fly ash, slag, etc.) as well as the alkali source (e.g. solid NaOH and/ or solid Na₂SiO₃) reacting with water to give the geopolymer gel. The regulation of gel viscosity is important as it gives an insight into the water molecules trapped in the pores of the system (Feng *et al.*, 2012). Although, one-part geopolymer production by this aspect seems more promising, the formation of crystalline tectosilicates (e.g. zeolites) is claimed to be disadvantageous because of leading to lower compressive strength (CS) values

which is indicator of the endurance of the polymer produced referring to the load that the system can withstand (Sturm *et al.*, 2016).

2.1.2. Aluminosilicate Sources for Geopolymerization

There are various of aluminosilicate sources for geopolymer production, one of which is metakaolin (MK) and is the purer form of aluminosilicate source, yielding a high proportion of the sum of SiO_2 and Al_2O_3 components over the whole MK mass. This measure is important in understanding the microstructure in geopolymeric systems (Duxson *et al.*, 2005). Fly ash (FA) is less pure than metakaolin and relation between microstructure and compressive strength is more difficult to determine experimentally. FA geopolymers are suggested to be characterized by high CS, low creep and shrinkage. (Temuujin *et al.*, 2010). Different types of fly ash can be observed: class C and class F. Due to its excessive calcium content and quick stabilization, the latter is not frequently utilized. Slag (ground granulated blast-furnace slag) obtained from coal combustion plants is rich in Al_2O_3 and SiO_2 is usually used in combination with fly ash aiming to improve mechanical performance (Luukkonen *et al.*, 2018).

Another $\text{SiO}_2/\text{Al}_2\text{O}_3$ source is rice husk ash (RHA). When calcined up to 800°C , RHA is an amorphous agriculture material. However, crystalline features (cristobalite and tridymite) are formed, when rice husk ash is cured at higher temperatures. As a geopolymer raw material, one-part RHA polymer samples with low crystalline components have produced when heat curing no more than one day is applied. As previously underlined, it is argued that crystallinity affects geopolymer durability negatively. Low dissolution of Al-Si precursor in liquid is a limitation of one-part synthesis, but the angular shape leading to rougher surface of RHA particles is suggested to yield higher CS (Sturm *et al.*, 2016).

Red mud (RM) is mainly a residue of alumina production and is also widely used for geopolymerization. Thus, red mud geopolymers do not generally exhibit high CS values (~ 10 MPa) so they are usually used in combination with Al-rich precursors, such as FA. Large RM containing geopolymers is asserted to show long term breakdown attributed to

depolymerization which may be limited by silica fume addition (Ye *et al.*, 2016). Microstructural analysis showed that addition of even small amounts of RM in MK geopolymer pastes increased the number of goethite and hematite crystals, microcracks and amount of unreacted MK (Kaya & Soyer-Uzun, 2016).

Silica Fume (SF) is a reactive pozzolan, which can be supplied by ferrosilicon and silicon metal plants. Its particle size is typically small ($< 1 \mu\text{m}$) and contains significant amounts of pure amorphous silicon dioxide. It is suggested that only silica fume containing geopolymers when properly cured, may reach CS values about 40 MPa (Rajerajeswari *et al.*, 2014).

2.1.3. Factors Affecting Compressive Strength of Geopolymers

Compressive strength (CS) is equal to the pressure that a sample can withstand without failing, and is a measure of the quality of the geopolymer produced. Curing time, curing temperature, ambient curing time, particle size of the reactive materials, geopolymer paste composition, Na content, aggregate addition/composition are some of the most important factors to affect the CS of geopolymer. Numerous experimental studies have been conducted to elucidate the effect of these factors on CS.

It is a common method to heat cure geopolymer mortars before letting them rest under ambient conditions, since curing time and temperature are known to affect CS. It should be noted that the effects of temperature and heat curing time on mechanical properties are nonlinear and depend on geopolymer material, while prolonged heat curing may lead to deterioration of the mortars. For instance, CS of electrolytic manganese residue-based geopolymer was found to increase as temperature was increased up to 80°C but further increase in temperature resulted crack formation on GP concrete (Han *et al.*, 2018). Similarly, heat curing time increase up to two days is hypothesized to lead to strength increase, whereas any longer treatment has the opposite effect (Han *et al.*, 2018).

Ambient curing time represents the setting time of the mortars. In a general manner, CS is enhanced with elongated setting time, however the rate in increase seems to vary for different time ranges. In Nagaraj and Babu, (2018) study, for example, a rapid rise in CS is observed until the ambient curing time is equal to 28 days, while mechanical strength of the majority of the samples is improved only marginally with further curing. On the contrary, when sodium silicate to sodium hydroxide ratio is the highest (6 M activator case) decrease is also observed (K & Babu, 2018).

Particle size of the reactive materials is related to the specific surface area, and may affect mortars' strength. On the other hand, the relationship between the two variables is not linear, e.g. raw materials with small particle sizes was found to increase strength, but in other circumstances a workable geopolymer gel was not formed from too fine materials (Duxson *et al.*, 2005).

Si-O-Si Bonds are stronger compared to Si-O-Al and Al-O-Al bonds, so it is important to regulate Si/Al ratio in order to form fully developed tetrahedrals for higher CS values. By means of compositional aspect only, a higher Si/Al ratio is expected to yield a higher strength. Accordingly, low Si/Al geopolymer mortars are characterized by large pores usually connected with each other and unreacted material, so low strength. In contrast, even small increases in the ratio have had SEM observable significant decreases in pore size (smaller pores). Thus, further increase is argued to have led to more brittle mortars and to Si density increase, attributed to possibly non-reacted aluminosilicate precursor (i.e. MK) (Duxson *et al.*, 2005).

Na content is also known to play an important role in determining CS. Theoretically, increase in alkalinity of the geopolymer gel results in an increase in compressive strength, as a result of the sufficient reaction between Na and Si-Al species (Yadollahi *et al.*, 2015). On the contrary, increasing sodium content ratio, expressed as Na/Si or Na/Al ratio, was shown to decrease the CS. This is attributed to poorly dissolved Na, or (in cases that water/solid is not constant) to excess water content (Zhang *et al.*, 2010). Water does not usually favour stable formation of either geopolymers, or conventional cement mortars.

However, addition of enough water is essential to ensure enough dissolution of reactive materials (Reddy *et al.*, 2018).

Aggregate addition is a common procedure in geopolymer production aiming to minimize pore and/or crack formation improving geopolymer stability. Yet, proper adjustment of aggregate proportion is substantial to inhibit bond formation between alkali activator and aggregate (or other inactive impurities and fillers), but not the aluminosilicate source – alkali bonds. The strength of the reactant bonds is asserted to impact geopolymer durability i.e. CS (Temuujin *et al.*, 2010).

In summary, there are various factors influencing the compressive strength of the geopolymer mortars, which have interacting and nonlinear effects. Hence examining the relation between CS and each factor individually is not adequate for an overall understanding of the polymeric systems. Furthermore, experimental determination of the relationships between all the factors (predictor variables) and CS is not an easy or even feasible task. As an alternative, statistical learning and data mining tools may help in rendering a holistic view via relating the physics and chemistry of geopolymerization with the resulting mechanical strength.

2.2. Statistical Learning

The desire to understand the distribution of data enabled the rise of the field of study named statistical learning, which are tools to understand and predict the data available. Statistical learning tools may be classified as supervised and unsupervised, based on whether the input data is used for predicting an output variable, or is just used to generate relation, assisting better understanding of the analysed class, respectively. The historical progression of statistical learning started with Legendre and Gauss (early 19th century) who introduced the method of least squares (early form of linear regression) and continued with Fisher (1936) who worked on linear discriminant analysis for the analysis of qualitative (categorical) values. In 1940s, logistic regression arose, and in 1970s, Nelder Wedderburn enlarged the field with generalized linear models. By the end of this decade, many linear

models have been developed. With the aid of computing technological improvements during 1980s, Breiman, Friedman, Olshen and Stone developed Classification and Regression Trees (CART) and became the pioneers of using cross-validation for model selection. In 1986, Hastie and Tibshirani broadened the learning methods using additive models. With the progression of machine learning, statistical learning, nowadays, is pronounced as a subfield under statistics, and its development is mainly based on the improvement of user-friendly software (e.g. R system) (James *et al.*, 2013).

2.2.1. Linear Regression

Linear regression is a widely accepted modelling method. Response variable is predicted based on the combination of independent variables and their interactions (Cosma, 2006). It should be noted that linear regression yields a global model i.e. spanning the whole space. Linear regression general equation is given in 2.1 where Y is the dependent variable, β_i is the regression coefficient vector, and X the vector of independent features, γXX^T represents the interaction of the variables and ε is the error term.

$$Y = \beta_0 + \beta^T X + \gamma XX^T + \varepsilon \quad (2.1)$$

Increase in the number of predicting variables increases model complexity and makes the interpretation of the resulting model difficult. Repeated partitioning of the data space is a methodology that eases model analysis, as a different pattern is assigned to each independent part of the space considered. The idea of data space partitioning is utilized by decision trees.

2.2.2. Classification and Regression Trees

Classification and regression decision trees are data based rule deriving statistical learning tools, and their development is attributed to Breiman *et al.* (Breiman *et al.*, 1984). The type of the decision tree is based on the type of response variable (the one predicted).

Variables may either bare a numeric value (qualitative-continuous) or can be numerically non-meaningful and discretize data into groups (categorical). A regression tree is constructed for numerical response data, while a classification/decision tree is built for categorical response data. Decision trees are advantageous as they can process any data type with little extrinsic intervention and lead to simple conditional rules. Each part of the data is partitioned in a parent node, and the resulting portions are named children nodes (Loh, 2014).. The parent node is split such as to gather similar objects at each node via minimizing impurity at each resulting child node. Splitting stops when a predefined criterion is attained.

THAID is regarded as the pioneer of classification trees, gathering similar response variables at children nodes, while C4.5 uses entropy and CART Gini index as impurity measurements (Loh, 2014). Entropy and gain formulas are given by Equations 2.2 and 2.3.

$$E(p) = - \sum_{i=1}^n p_i \times \log p_i \quad (2.2)$$

$$G(p, T) = E(p) - \sum_{j=1}^n p_j \times E(p_j) \quad (2.3)$$

Here, p_i is the probability distribution of the instances of a feature, n is the number of different levels that a feature can attain and T is the test set, i.e. feature tested. The partition resulting to minimum entropy and maximum gain is selected as the children nodes (HSSINA *et al.*, 2014). Gini Index impurity measurement used by CART is given by Equation 2.4.

$$g(t) = \sum_{j=1}^c c\left(\frac{i}{j}\right) p\left(\frac{j}{t}\right) p\left(\frac{i}{t}\right) \quad (2.4)$$

Here, $p\left(\frac{j}{t}\right)$ is the probability of category j at the node t and $C\left(\frac{i}{j}\right)$ is the probability of misclassifying a category j case as category i , for all categories k .

C4.5 and CART are eager to select features, comprising a larger set of levels. Aiming to overcome this limitation, other algorithms, such as CRUISE and GUIDE, are developed and take into account connection between each variable tested and the response variable specifically, carrying out chi-squared test (Loh, 2014). In that way, independence of the predictor from predicted feature is tested. The variable minimizing significance probability is chosen, and minimum Gini index split is preferred. ANOVA is also used by QUEST algorithm.

In regards to regression trees, although the response data refers to a numerical feature, the idea behind it, is similar to that of classification trees; minimizing impurity, which is expressed as the sum of squared deviations from the mean of the instances distributed at the node, which is the node prediction. (Loh, 2014). AID and CART are the preliminary regression tree algorithms. Another regression tree algorithm developed by Quinlan is $M5$, in which the nodes represent linear models. Moreover, GUIDE is used for regression problems based on classification tree philosophy but adapting a regression model and splitting parent nodes into children, based on whether the residuals are positive or negative. In this way, unbiased divisions are created while the models can vary being piecewise constant or quadratic. The degree in complexity depends on the complexity of each node and the size of it implying that simpler node trees are often larger. More complex trees tend to have greater accuracy, but their interpretation is much more difficult.

There are various accuracy measures. If $\mathcal{L}$ is a training data set with N samples, belonging to a space defined as $\mathbf{X} \times \mathbf{C}$ with, $\mathbf{x} \in \mathbf{X}$, $j \in \mathbf{C}$ and $A \subset \mathbf{X}$, then $R * (d)$ for instance gives the number of new instances misclassified, given a function $d(x)$ constructed based on the set $\mathcal{L}$ (Breiman *et al.*, 1984). In other words, it is the probability of new data misclassification which is assumed to be of identical distribution as $\mathcal{L}$. Symbolically, this may be summarized as follows:

$$P(\mathbf{x} \in A, Y = j) = p(A, j) \quad (2.5)$$

$$(\mathbf{x}, Y) \coprod \mathcal{L} \quad (2.6)$$

$$R * (d) = P(d(\mathbf{X}) \neq Y | \mathcal{L}) \quad (2.7)$$

When both d construction and accuracy estimate calculation are made based on $\mathcal{L}$ due to lack of new data, this $R *$ estimate is called an internal estimate. Resubstitution, sample estimation and V-fold cross-validation estimates are of interest. Resubstitution estimate gives the number of misclassified objects of $\mathcal{L}$ by d which is also constructed by $\mathcal{L}$. Based on this fact, it can lead to too optimistic R^* estimate conclusions. When sample estimation is considered training set $\mathcal{L}$ is divided into $\mathcal{L}_1$ and $\mathcal{L}_2$ with number of samples N_1 and N_2 respectively. The first is used to construct d and the latter to calculate accuracy.

$$R^{ts}(d) = 1/N_2 \sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_2} \chi(d(\mathbf{x}_n) \neq j_n) \quad (2.9)$$

Here, function $\chi(.)$ equals one when the expression within is true, and 0 otherwise for j_n measurement vectors in $\mathcal{L}$. Independency of the samples chosen in $\mathcal{L}_1$ and $\mathcal{L}_2$ is important and training sample size is significantly decreased. For that reason, V-fold cross-validation would rather be used. In this case, the data is divided in V parts, $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_V$ and d^V is the classifying function in each case constructed by the whole training sample excluding V -part at each step. The misclassification estimate for each V^{th} - fold is then computed as follows:

$$R^{ts}(d^V) = 1/N_V \sum_{(\mathbf{x}_n, j_n) \in \mathcal{L}_V} \chi(d^V(\mathbf{x}_n) \neq j_n) \quad (2.10)$$

Cross-validation (CV) misclassification estimate is finally calculated as the average of all individual folds.

$$R^{CV}(d) = 1/V \sum_{v=1}^V R^{ts}(d^v) \quad (2.11)$$

Since a large proportion of the training set is used, cross-validation estimate is thought to be more consistent.

As previously noted, accuracy of the tree is related with the tree complexity, hence the tree size. When optimization of classification/ regression at each step is solely considered, then excessively large trees are generally obtained. However, it should be emphasized that optimum size of trees exists, i.e. too many tree branches increase R^{ts} rate and worsen future predictions. Back pruning is preferable, since preliminary stopping may prevent significant impurity decrease, which might have been obtained with further branching. In CART and C4.5 algorithms, for example, maximum sized tree is constructed, and the best sized one is chosen. CV plays an important role in these methods and is mostly preferred for small sized data sets (Loh, 2014). As a maximum sized tree is pruned, CV misclassification decreases reaching minimum and increases again in lower levels of the tree, due to bias-variance trade-off (Breiman *et al.*, 1984). However, algorithms, in which tree size can be determined by a threshold to node impurity specification, also exist, e.g. THAID algorithm, for which CV is mostly preferred for small sized data sets.

2.2.3. Feature Selection Algorithms

Not all features that are apriori assumed to affect the dependent variable are certainly required to construct an empirical relation for the system analysed. Features can be characterized as (a) relevant; when their influence on the output is certain and necessary, (b) irrelevant; when they do not have any effect on the output so do not exhibit any relational pattern and (c) redundant; if the feature has an effect on the dependent variable but is correlated with another feature considered as relevant (Karagiannopoulos *et al.*, 2007). Consequently, algorithms for feature selection are developed with the scope to prevent

irrelevant feature use and overfitting and accelerate model building process (PK and V, 2013).

Feature selection algorithms are categorized as filter, wrapper and hybrid methods depending on whether learning algorithm used for final knowledge generation is utilized for optimal subset determination (Karagiannopoulos *et al.*, 2007). Filter algorithms do not use learning algorithm during feature selection and are generally computationally cheaper, but insufficient for redundant variable exclusion. Wrapper approaches use induction algorithm and are computationally expensive, but preferable since use learning algorithm for feature selection.

Two notable wrapper feature selection methods are Forward Selection (FS) and Backward Elimination (BE) (Karagiannopoulos *et al.*, 2007). In FS, features are added one by one each time to the set, which is initially empty. The feature selected for addition is the one that leads to a subset that gives the best result based on a criterion used (e.g. minimal *RMSE*). A disadvantage of FS is that it neglects any improvement derived from mutual contribution of more than one feature. Conversely, for backward elimination a feature is removed at each step. The feature selected is the one, the removal of which leads to the best result as mentioned for forward selection. Although backward elimination is costly, it also considers dependence of a variable to another.

There are other feature selection methods mainly depending on the methods mentioned, such as i) sequential forward floating selection where the number of features added or deleted varies, ii) best fit, for which returning to previously selected subset and continue selection from that point is possible, and iii) random generation, in which variable addition/ deletion is done randomly (Karagiannopoulos *et al.*, 2007).

2.2.4. Rough Set Theory

Rough Set theory is a rule deriving methodology based on data, Using rough set theory, it is possible to determine subsets of features, namely reducts, which consist of minimum number of features for classifying the whole dataset, similar to feature selection algorithms(Hu *et al.*, 2003). Reduct sets are usually determined by simplification of discernibility matrix, but its computation is expensive, hence it is usually employed on small data sets .

In what follows, some basic rough set terminology is introduced. A data set is assumed to consist of attributes (features), and objects (data points), and attribute value of each object is provided (Hu *et al.*, 2003). The feature to be predicted is called the decision on attribute (d_a), whereas the other features comprise the conditional feature set ($\mathbb{A}$) with $\mathbb{A} \cap \{d_a\} = \emptyset$. The ensemble of all objects is called the universe (U) and the information system ($S = (U, \mathbb{A} \cup \{d_a\})$) includes all the information about it. An equivalence relation comprises of all the equivalence classes, each of which approximates a subset of instances (concept). If B is a non-empty subset of A an indiscernibility relation is defined as;

$$IND(B) = \{(x, y) \in U \times U : \alpha(x) = \alpha(y) \forall \alpha \in B\} \quad (2.12)$$

Concepts are definable when determined by equivalence classes, or indefinable if described by upper and lower approximation subsets (Hu *et al.*, 2003). If $X \subseteq U$ and $B \subseteq \mathbb{A}$ a feature subset, lower and upper approximations of the subset, and the positive region of the decision d_a regarding subset X are described as follow:

$$\underline{B}(X) = \{x \in U : [x]_B \subseteq X\} \quad (2.13)$$

$$\overline{B}(X) = \{x \in U : [x]_B \cap X \neq \emptyset\} \quad (2.14)$$

$$POS_B(d_a) = \cup \{ \underline{B}(X) : X \in U/IND(d_a) \} \quad (2.15)$$

The minimum number of feature subset that contains the total information of the system is called a reduct, and is formally represented for a subset B as follows:

$$\underline{B'} \subseteq B, POS_B(d_a) = POS_{B'}(d_a) \quad (2.16)$$

noting that there is no other subset C satisfying;

$$POS_B(d_a) = POS_C(d_a) \quad (2.17)$$

However, there may be more than one reduct, and the set of all reducts is called the core of the system (Hu *et al.*, 2003). The measure of dependency of two subsets is given by Equation 2.18.

$$\mu(C, B) = |POS_C(B)|/|U| \quad (2.18)$$

As dependency approaches unity, B reliance to C is assumed to increase (Hu *et al.*, 2003). The significance of B , on the other side, is given by relation.

$$\sigma_{(A, d_a)} B = 1 - \mu(A - B, d_a) / \mu(A, d_a) \quad (2.19)$$

Discernibility matrix and discernibility function are also notions important for reduct calculation (Hu *et al.*, 2003). The first is the matrix $|U| \times |U|$ contains the inputs c_{ij} , for which $\{\alpha \in A | \alpha(x_i) \neq \alpha(x_j)\}$ if $d_a(x_i) \neq d_a(x_j)$ or else $\emptyset$. After simplification by absorption law, the intersection of all unities of attributes in each discernibility matrix entry leads to the system reducts. In other words, the discernibility function consists of the combination of the reducts. An attribute can appear in several reducts, along with different attributes. Number of reducts, in which a feature is present, gives insight on the importance of the feature on decision attribute, i.e. if a feature is present in a high number of reducts,

then that feature is likely to be important. Moreover, if the number of attributes in a reduct is low, then these attributes are likely to carry a higher importance. Within this frame, the importance -weight- of an attribute $\alpha_i \in C$ is given by Equation 2.20.

$$w(\alpha) = w(\alpha_i) + |A|/|C| \quad (2.20)$$

The weight is firstly zero and is recalculated every time that the feature appears in a new entry. Weights of the features calculated via this method may be used as a feature subset selection method (Hu *et al.*, 2003).

2.2.5. Stability in Feature Selection Algorithms

Feature selection methods contribute to faster classification/ regression, omitting non-essential features, and may yield higher accuracy, but may lead to different selected feature subsets for different subsets of data, hence the derived rules may easily be perturbed, i.e. unstable, in case of trivial changes in data set (Abeel *et al.*, 2009). Therefore, it is crucial to ensure the stability of the rules applied to different sections of data, hence it is important to use subset selection algorithm that leads to similar feature subsets with varying data sets. There are various methods to measure the stability- resemblance between selected set of features (Nogueira *et al.*, 2017). One of them is Kuncheva index (*KI*) and it is given by Equation 2.21. Here, $r_{i,j}$ is the number of common elements between sets i and j , d_s is the total number of samples and k the number of features selected. Kuncheva index varies between $[-1,1]$ with -1 implying total variability (Kuncheva, 2007).

$$KI = \frac{r_{i,j} - k^2/d_s}{k - k^2/d_s} \quad (2.21)$$

Here, k^2/d_s correction eliminates any similarity effect by chance. Let n be the feature number of a set X , A and B are two subsets of X , k the number of features of A and it is equal to B . The expected value of intersection of these two sets is given by $\mathcal{Y}$, a random hypergeometric variable and r the describing the elements chosen by A and B .

$$P(\mathcal{Y} = r) = \frac{\binom{k}{r} \binom{n-k}{k-r}}{\binom{n}{k}} = k^2/d \quad (2.22)$$

Another measure which converges to Kuncheva index, when the number of predictors selected do not change, is given by Nogueira *et al.* (2017), and used in the current study to determine the stability of the regression trees. $\bar{K}$ gives the average number of predictors selected in each training set, and is calculated as follows:

$$\bar{K} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M x_{ij} \quad (2.23)$$

Here, N and M denote the number of observations and predictors respectively and x_{ij} equals 1 if the j^{th} predictor is selected otherwise it is taken as zero with $x \in \mathbb{R}^{N \times M}$. The number of predictors selected at each training set is given by Equation 2.24 and S_j^2 is defined as in Equation 2.25.

$$\hat{p}_j = \frac{\sum_{i=1}^N x_{ij}}{N} \quad (2.24)$$

$$S_j^2 = \frac{N}{N-1} \hat{p}_j (1 - \hat{p}_j) \quad (2.25)$$

Hence, the stability is defined as follows:

$$Stability = 1 - \frac{\frac{1}{M} \sum_{j=1}^M S_j^2}{\frac{K}{M} \left(1 - \frac{K}{M}\right)} \quad (2.26)$$

2.3. Application of Statistical Learning Methods in Materials Literature

In chemical engineering literature, statistical learning methods have been mainly employed for prediction, knowledge extraction, and determining and visualizing relationships between variables. Statistical and machine learning methods have also been successfully employed in material science. The most frequently used machine learning methods employed on materials are i. decision/regression trees and random forests, ii. artificial neural network (ANN), and iii. combinations of various learners. The aim of these studies are accurate prediction of certain material characteristics via the statistical modelling of the systems (predictive modelling), and interpretation of the models to determine the partial effects of the factors (explanatory modelling). Most studies in the literature have leaned on the former type of modelling.

In one of the pioneering studies in this field, Sturrock and Bogaerts described the corrosion phenomenon for Austenitic Stainless Steels using parametric and non-parametric learning methods, such as Linear Discriminant, Nearest Neighbour, Polynomial Networks, Decision Trees and Fuzzy Logic (Sturrock & Bogaerts, 1997). By training and testing a data set consisting of 112 samples, prediction accuracy, assessed via substitution and leave-one-out errors, of the methods are compared. Nearest neighbour and decision tree were shown to be the most accurate methods. Moreover, decision trees were used to generate rules to determine whether corrosion will take place and had been shown to be accurate and easy to understand predicting tools even when applied to small data sets. Regression and classification decision trees have been used to describe the durability of Austenitic Stainless

Steel (Li, 2006). CART was applied to a creep rupture database for Austenitic Stainless Steel, consisting of around 1600 samples part of which is used for testing. The set includes 1 response variable (Rupture Life) and 20 predictors, one of which is discrete. The proportion of the missing values in the database was considerably low (<5%). A CART model was constructed importance of the predictors based on CART is given. CART model accuracy was also compared with other methods, including curvilinear regression, and regression tree is shown to have equal or better accuracy compared to them.

In a recent study, statistical analysis employing GUIDE regression tree algorithm is utilized, for a polylactic acid (PLA)-based biocomposite system aiming to extract knowledge in order to guide researchers in selection of experimental conditions for desired tensile strength values (Alakent and Soyer-Uzun, 2019). The data set was constructed using 26 publications between 1999 and 2018 and contained 135 experimental data points. 23 features related to composite synthesis were included; these features involved manufacturing method and temperature for compounding and testing, molecular weight of the PLA used, chemical composition of the composite, and type of the filler employed in synthesis. The response variable was selected as tensile strength, and unbiased cross validation error of the regression tree was found to be 12–17 MPa, with coefficient of determination for prediction (R^2_p) value equal to 0.5–0.7, showing moderate-to-high prediction accuracy. Results indicated that the method used for manufacturing/ testing, the molecular weight of the PLA and the temperature employed during test specimen preparation had a significant influence in predicting the final tensile strength of the composite.

Another method used for predicting CS and analysis of the factors that affect CS is M5P tree algorithm (Behnood *et al.*, 2017). The compressive strength of Normal and High Performance Concrete was predicted through the M5P model constructed based on 1912 samples collected from literature, and 15% the samples were used for testing the obtained model. Eight factors assumed to affect CS most significantly are used as the predictors in the model. Although no qualitative analysis of the factors was made, six models comprising the leaves of the M5P tree were used to predict the CS of the test set, yielding a high correlation coefficient (0.91). In another study, Random Forest, Naïve Bayes classifier and

k-nn classifier were employed for descriptive evaluation of four attributes (Si/Al, Na/Al, H₂O/Na and water/binder ratios) on CS: Metakaolin based geopolymers of different and different periods of time were synthesized, and models using compositional attributes from the obtained data were seen to yield accurate predictions (Lahoti *et al.*, 2017).

ANN is one of the most popular machine learners used in various fields, including material science. In one of the earliest attempts, an ANN model was constructed from a data set of ~1000 samples, comprising features such as water, superplasticizer and aggregate content and testing time to predict compressive strength of cement (Yeh, 1998). A multilayer ANN model, consisting of four intermediate and one output layer, was used in order to improve the experimental procedures for geopolymer design (Yadollahi *et al.*, 2015). Geopolymer mortars prepared experimentally are used as the input data of the model, consisting of 45 samples, 15 of which are used for testing. The geopolymers are prepared by pumice aluminosilicate precursor, NaOH/ silicate and super plasticizer and cured for different periods under varying temperature. The factors investigated during model building are silica modulus, sodium and water content as well as curing duration, but unfortunately only ground pumice based geopolymers were considered. The outcome of the study is an ANN model, successfully predicting CS of geopolymers synthesized experimentally. ANN was used again for compressive strength prediction of another concrete type -Rubberised Concrete- by Ashour *et al.* (2018). The training set consisted of 201 data used for the ANN construction, whereas the validating and testing ones of 43 samples each used to optimize the extend of generalization of the ANN and to test the final model, respectively. The predictors considered in the study are Coarse Aggregate, Fine Aggregate, water to cement ratio and Fine and Coarse Rubber. The number of neurons on the intermediate layer was chosen by trial and error based on which led to the greatest correlation coefficient between predicted and experimental data. Parametric studies were needed to generate rules based on the change of only three of the factors studied (water to cement ratio, fine and coarse rubber).

A more general approach is provided by Bui *et al.* who introduced ANN and modified firefly algorithm (MFA) for compressive and tensile strength of high performance concrete (ordinary concrete including additional materials such as slag) (Bui *et al.*, 2018). The ANN

model was constructed based on experimental data and the optimization of its parameters was achieved via MFA. 1133 data points were used, considering eight predictors such as cement, slag and superplasticizer amount to predict the CS of the concrete. The same dataset was also used in other studies such as that of Chou *et al.* (Chou *et al.*, 2016). Another data set consisting of 714 points is used for tensile strength prediction based on cubic compressive strength and curing time. The combined ANN-MFA method was shown to be a fast and accurate predictive tool. A network structure named Process, Structure Property and Performance (PSPP), in which factors affecting the performance of the material are represented as nodes, has been recently suggested to aid material design (Olson, 2013). Using PSPP, and knowledge extraction via natural language processing, a machine learning algorithm, a mapping chart (CAMaD) for systematization of material design is constructed (Onishi *et al.*, 2018). The modelling procedure in the study consist of identification of factors that affect the material properties via keyword searching, determining the relations among these factors using convolutional neural networks, and branching the PSPP chart.

3.METHODOLOGY

The procedure employed in the present study consist of i) preparation of the dataset, ii) construction of trees based on different feature selection methods at different tree sizes, iii) the selection of the optimum tree to be used as a rule generation tool, and iv) the interpretation of the tree obtained. Detailed information on how the data set is constructed is given in Section 3.1, followed by methods for selections of features and construction of tree in Section 3.2.

3.1. Exploratory Analysis of the Dataset

The data set used for model construction is created using 38 papers on geopolymer synthesis, leading to a total of 879 observations. Papers relevant to the keywords ‘‘geopolymer’’ and ‘‘Compressive Strength’’ are sorted, and papers, which are published after and including 2013 and can offer at least five data points, are selected for further utilization. The date distribution of papers is shown in Figure 3.1.

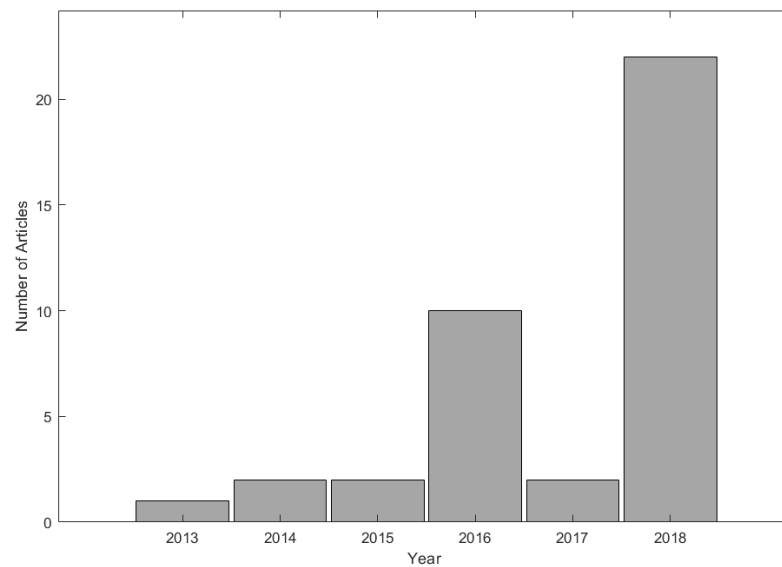


Figure 3.1. Per year distribution of the articles used

A brief description of geopolymer synthesis together with the features employed in the dataset can be noted as follows:

- (i) Geopolymers are synthesized mainly by alkali or rarely by acid-activation of raw materials (*Material*), such as metakaolin and fly ash and their combinations with other materials like slag and OPC. A full list of raw materials are given in Table 3.2.
- (ii) As previously stated, geopolymerization comprises mainly alkali activation of aluminosilicate source. In general the composition of geopolymer is expressed as $M_n[-(SiO_2)_z - AlO_2]_n \cdot wH_2O$ with n being the degree of polycondensation and M a cation such as Na⁺ (Yadollahi, Benli and Demirboğa, 2015). Therefore, the compositional parameters are selected as molar Na/Al, H₂O/Al and Si/Al ratios, based on the typical general formula to describe composition and are calculated for all the geopolymeric materials included in the database. It should be noted that the database contains a single study where acid-activation procedure is employed and in this context Na/Al for this set is rather low compared to most of the observations.
- (iii) Although geopolymers mainly refer to aluminosilicate-based materials, most of the aluminosilicate sources used in the studies contain impurities such as high Fe₂O₃ content in red mud and CaO in fly ash or slag. Therefore, in order to describe the potential aluminosilicate content in the final geopolymer, *Purity* is defined as a feature related to the total aluminosilicate content (Al₂O₃+SiO₂) in the raw material and employed as a numerical variable.
- (iv) As indicated before two main synthesis procedures are involved in geopolymerization: one part with “just add water” concept and two part synthesis method which includes a corrosive alkali activating solution. The synthesis procedure employed in geopolymer synthesis is described by a categorical variable (*Type*), which equals one for one-part and two for two-part synthesis, respectively.
- (v) Particle size of the raw material is also considered in the database. As particle size maximum 90th percentile in particle size distribution is considered. In case of a mixture of raw materials *ParSize* variable designates the maximum particle size among the components whereas *AvParSize* is the mass proportion based weighted average of the particle sizes of the distinctive species.
- (vi) A common practice in geopolymer field is adding fillers/ aggregate such as sand, quartz etc. to the geopolymer mix in order to promote mechanical properties. Both existence of aggregate (*Aggr*) and aggregate proportion in the final solid geopolymer

mix are considered (*AggrProp*). In addition, particle size of the used aggregate (*Aggr_ParSize*) is also noted.

- (vii) In some of the studies in the literature, additives such as steel fibers, super-plasticizers, PVA, sucrose and others are employed as a part of synthesis procedure for improving properties such as the elasticity of the geopolymer. Both existence of additive (*Add*) and percent of the additive in the mixture (*Add_per*) are considered in the model.
- (viii) When the paste is ready, it is put into moulds and cured. Usually, mortars are firstly subjected to heat treatment (referred as heat curing) around 40-80°C. The subjection to heat curing is designated as *Texist1*, the temperature of heat curing is noted as *T1* and its duration as *t1*.
- (ix) Curing at ambient conditions is also a common procedure that is usually followed after heat curing stage. The presence of ambient curing procedure is represented with a categorical variable *Texist2*, the temperature used for ambient curing is taken as *T2* whereas the time of ambient curing is *t2*.

The list of all predictor variables, their abbreviations, units and types are shown in Table 3.1. Only four of the predictors have missing values (NA) in the dataset, and the percent of missing values does not exceed 20%. Missing values are filled using median values of the existing data.

Table 3.1. Summary of the predictors considered.

Predictor Variable	Abbreviation	Unit	Type	% NA*
Material used as aluminosilicate precursor	<i>Material</i>	-	Categorical	-
Molar Si/Al ratio	<i>si_al</i>	-	Numerical	-
Molar Na/Al ratio	<i>Na_al</i>	-	Numerical	-
Molar H ₂ O Ratio	<i>w_al</i>	-	Numerical	6.0
(SiO ₂ + Al ₂ O ₃)/ total mass ratio	<i>Purity</i>	-	Numerical	-
Presence of Aggregate(0/1)	<i>Aggr</i>	-	Categorical	-

Table 3.1. Summary of the predictors considered (cont.).

Aggregate Proportion	<i>AggrProp</i>	-	Numerical	-
Aggregate Particle Size	<i>Aggr_ParSize</i>	μm	Numerical	1.9
Presence of Additive (0/1)	<i>Add</i>	-	Categorical	-
Additive percentage	<i>Add_per</i>	-	Numerical	-
Geopolymerization mechanism type	<i>Type</i>	-	Categorical	-
Maximum particle size	<i>ParSize</i>	μm	Numerical	19
Average particle size	<i>AvParSize</i>	μm	Numerical	19
Heat curing subjection	<i>Texist1</i>	-	Categorical	-
Ambient curing subjection	<i>Texist2</i>	-	Categorical	-
Heat curing temperature	<i>T1</i>	°C	Numerical	-
Heat curing time	<i>t1</i>	days	Numerical	-
Ambient curing temperature	<i>T2</i>	°C	Numerical	-
Ambient curing time	<i>t2</i>	days	Numerical	-

*: NA means “not available”, i.e. missing values in the database.

25 different material precursor types exist in the database; each unique combination of precursor types is taken as another level the material categorical variable can take. The complete list of the materials and their abbreviations are listed in Table 3.2. In the database, most of the mortars are prepared using Fly Ash +Slag (FA + S), followed by slag+fly ash +Palm Oil, and metakaolin only (Figure 3.2).

Table 3.2. List of materials used as precursors and their abbreviations.

Material Type	Abbreviation
Metakaolin	MK
Fly Ash	FA
Metakaolin + Fly Ash	MK+FA
Metakaolin + Red Mud	MK+RM
Rice Husk Ash	RHA

Table 3.2. List of materials used as precursors and their abbreviations (cont).

Red Mud	RM
Red Mud + Silica Fume	RM+SF
Fly Ash + Slag	FA+S
Red Mud + Fly Ash	RM+FA
Slag	S
Slag + Silica Fume + Lime	S+SF+L
Fly Ash + Lime	FA+L
Fly Ash + Slag+Lime	FA+S+L
Red Mud + Slag	RM+S
Fly Ash + Slag + Soft Marine Clay	FA+S+SMC
Electrolytic Manganese Dioxide	EMDR
Rice Husk Ash + Slag	RHA+S
Fly Ash + Silica Fume	FA+SF
Fly Ash + Silica Fume + Ordinary Portland Cement	FA+SF+OPC
Fly Ash + Calcium Activated Concrete	FA+CAC
Slag + Fly Ash + Palm Oil Fly Ash	S+FA+POFA
Metakaolin + Slag	MK+S
Fly Ash + Ordinary Portland Cement	FA+OPC
Fly Ash + Slag + Silica Fume	FA+S+SF
Coal Gangue	CG

Molar Si/Al, Na/Al and Water/Al ratios are calculated based on the synthesis procedure, available precursor and alkaline activator composition, and individual curing steps indicated in each paper. Generally, mass concentration of raw materials and mix proportions are available, hence molar ratios may be calculated using the molecular mass of the compounds. Mass of aggregates and additives are not taken into consideration during calculations, since they do not take part in geopolymerization. On the contrary, cement, due to being an aluminosilicate source, is included in calculations. Distribution of Si/Al ratio is seen to have two modes at ~2.0 and ~2.8, and also is spread between 3 and 6 at a very low frequency (Figure 3.3 top). Although other activators, such as phosphoric acid (Han *et al.*, 2018) and KOH (Luna-Galiano *et al.*, 2018), are also used, the majority of the studies involve

sodium-based alkali activators (NaOH and Na₂SiO₃ solution) with most of the sodium to alumina ratio values being smaller than unity (Figure 3.3 middle). In the prepared database, water to alumina ratio is seen to vary in a wide range between 0.46 – 18, with an observable mode at ~8.5 (Figure 3.3 bottom).

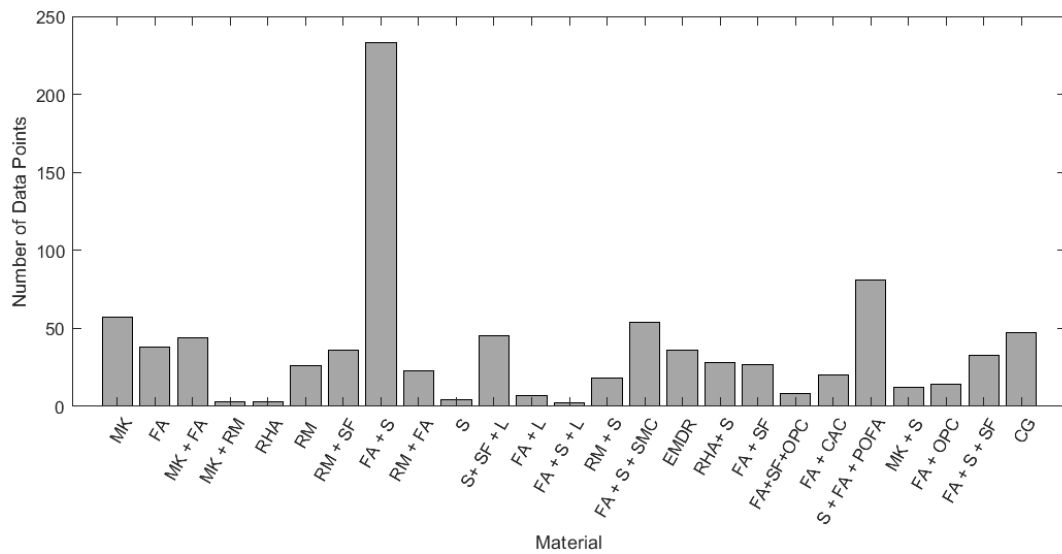


Figure 3.2. Material type distribution

Ratio of the total mass of SiO₂ and Al₂O₃ to the mass of the raw material, excluding the alkaline activator, aggregate and additives, is a measure of the available reactive Si and Al in the precursor composition. It is seen that ~250 observations in the dataset exhibit a ratio of ~80%, while the rest of the samples are approximately uniform distributed between 0.25 and 1.0 (Figure 3.4). The rest of the geopolymer paste consists of either undesirable impurities, or additional substances.

Coarse and fine aggregates, such as sand, limestone, dolomite, granite, quartz, expanded clay granule, are added to regulate CS and other characteristics, such as porosity of the prepared mortars. Presence of an aggregate exists is indicated with an aggregate variable (Aggr), taken to be equal to one or zero for the presence and absence of an aggregate, respectively. Almost half of the samples in the data contain aggregate as shown

in Figure 3.5a. The mass proportion of aggregate to the total mass of the dry material, excluding alkaline activator solid substances, such as NaOH, Na_2SiO_3 , is shown by the aggregate proportion variable (*AggrProp*). The distribution of the aggregate proportion of the aggregate containing mortars is shown in the middle of Figure 3.5b, in which two modes at 0.6 and ~ 0.75 are seen. As in the case of dry precursor materials, aggregate particle size (*Aggr_ParSize*) is also considered as a predictor. As seen in the bottom of Figure 3.5c, aggregate particle sizes are clustered into fine and coarse aggregates, corresponding to *Aggr_ParSize* smaller than 2000 μm and those larger than 7000 μm , respectively.

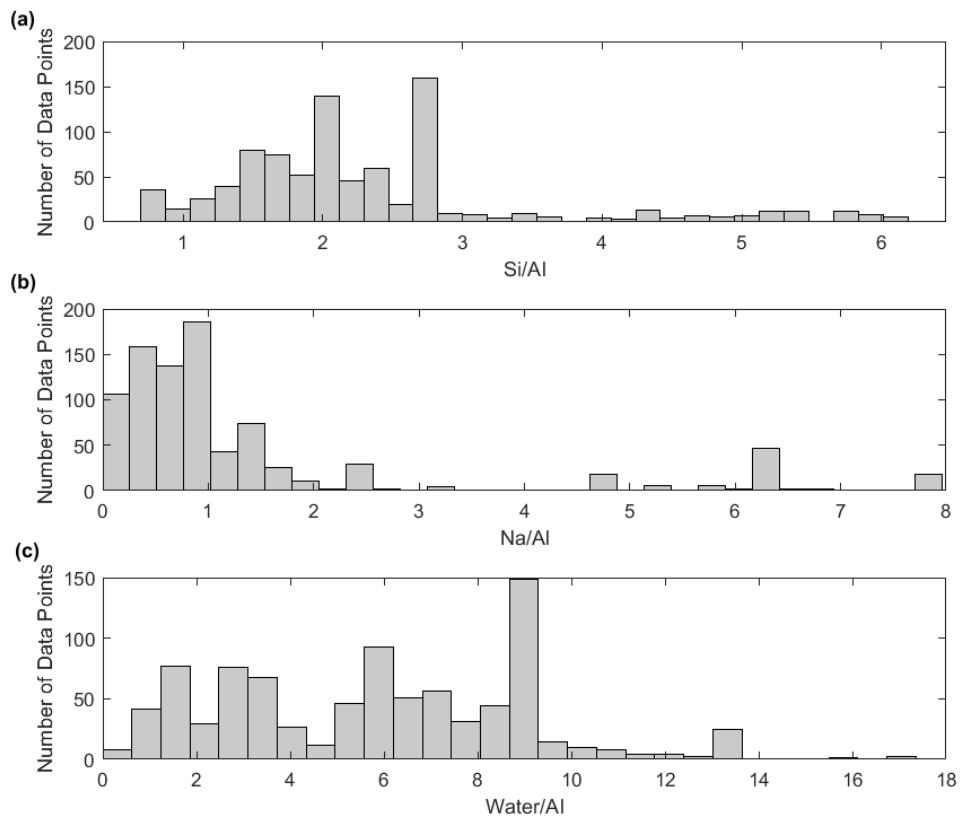


Figure 3.3. (a) Si/Al, (b) Na/Al and (c) Water/ Al molar ratio distributions.

Extra substances, such as super-plasticizer, steel fiber, PVA, sucrose, water reducer admixtures, HMGB (high mobility group proteins), MgO, Sodium lignosulfate, are added to regulate durability and compositional characteristics of the mortars. Presence of an additive is expressed as a categorical predictor (*Add*). Only one fifth of the data contains an additive

(Figure 3.6a). For additive containing samples, additive mass percent (*Add_per*) distribution is seen in Figure 3.6b. Most of the additive containing additive to dry mass ratio (not including alkaline activator) is below 2%, and two additional peaks at 5-6% and 13-14% are observed.

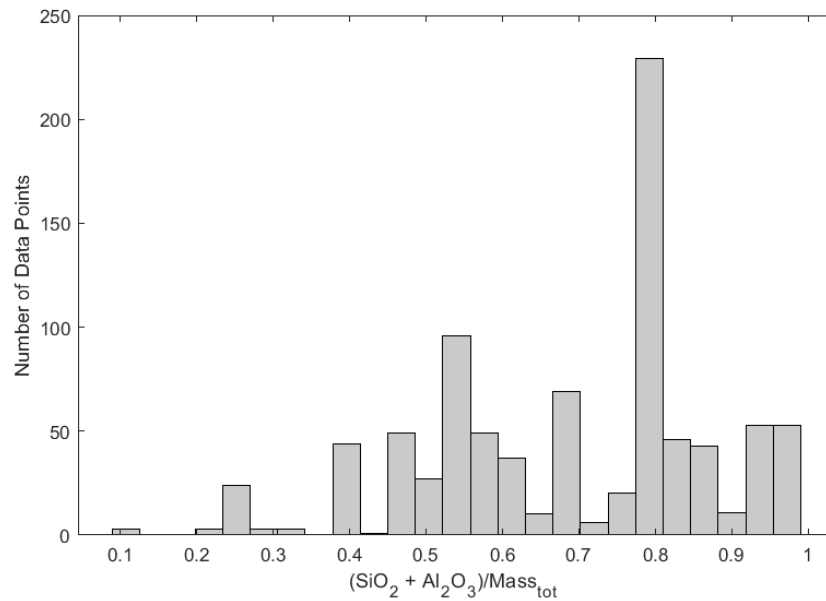


Figure 3.4. Purity (SiO₂ and Al₂O₃ over raw material mass) distribution.

Although the conventional geopolymer synthesis method is a two part method, consisting of separate preparation of the activating solution and mixing it with the precursor (raw material: Metakaolin, Red Mud, Fly Ash etc), a new synthesis method named just-add-water is also used in the literature. The latter is easier in preparation, since addition of certain amount of water makes the dry mixed materials (including solid alkali) ready to be used, similar to cement mortars. In the current study, the type of synthesis is represented as a categorical variable (*Type*). Most of the data is comprised of two-part geopolymers (Figure 3.7).

Degree of grinding is represented by particle size of solid materials before polymerization. Particle size variable (*ParSize*) is taken to be equal to 90th percentile of the particle size of a single or largest raw material In Figure 3.8a, four peaks at 40-50, ~75, 150

and 300 μm are seen. On the contrary, average particle size (*AvParSize*) refers to the weighted average of the mass proportion of the precursors. Average particle size is more evenly distributed, and mostly accumulated below 200 μm (Figure 3.8b).

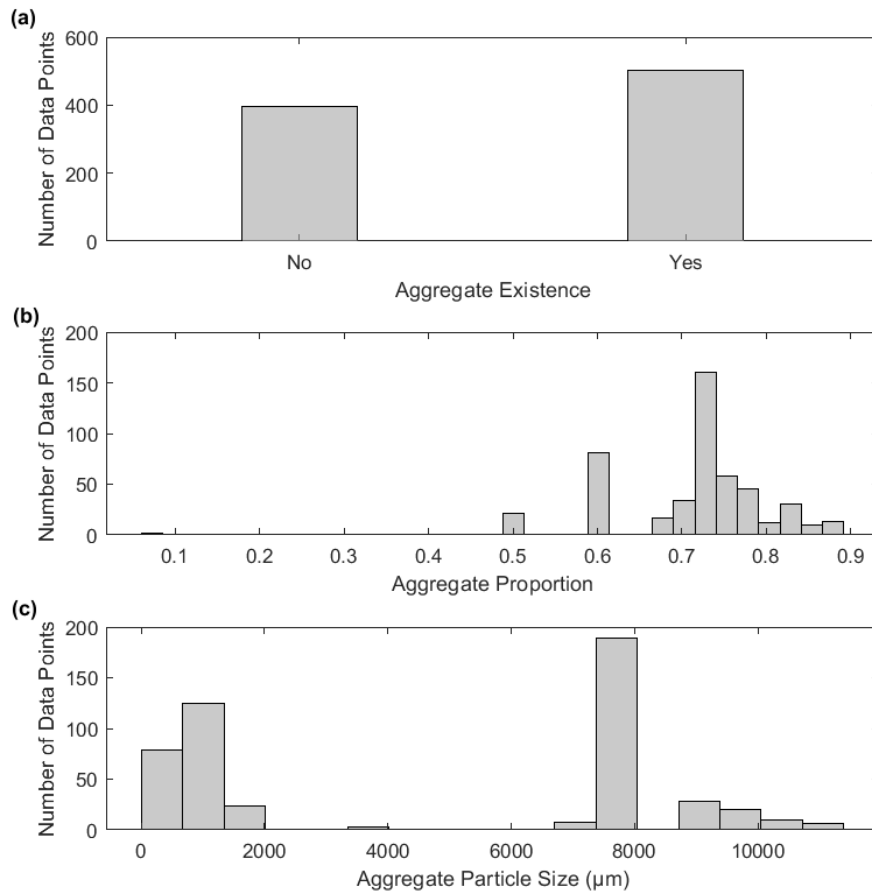


Figure 3.5. Distributions of aggregate (a) presence (b) proportion, and (c) particle size.

It is a common method to subject geopolymer mortars to high temperatures (above ambient) to enhance geopolymerization and enable hardening. In the dataset, number of heat (*Texist1*) and ambient (*Texist2*) cured specimens is shown in Figure 3.9 a and b respectively. It should be noted that the time elapsed before and after demoulding are both included in curing. Furthermore, ambient curing and heat curing correspond to curing below and above (and including) 40°C, respectively. The majority of specimens are not heat cured, but almost all samples in the dataset are ambient cured.

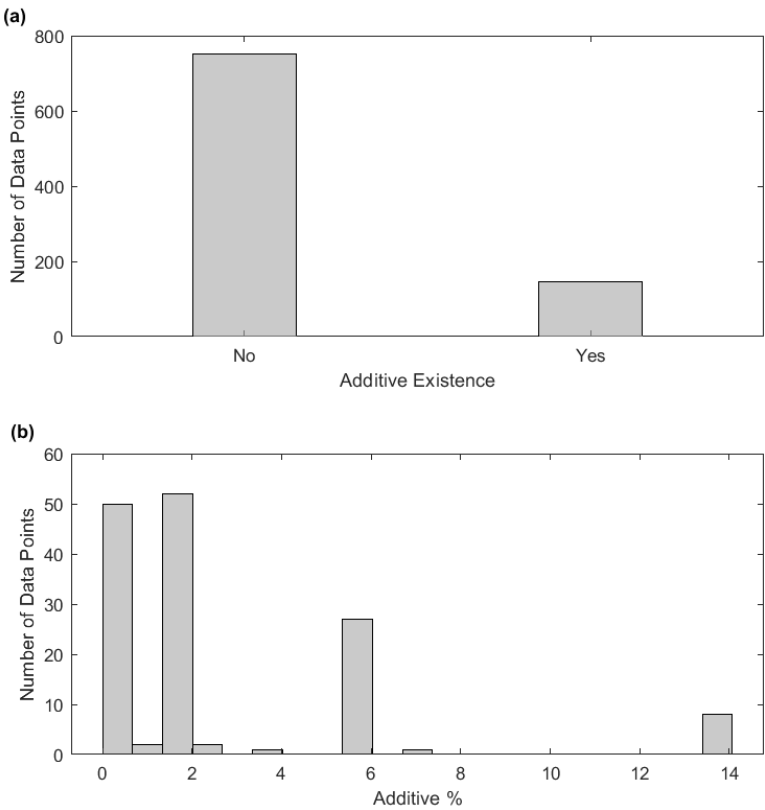


Figure 3.6. Distributions of additive (a) presence and (b) percentage.

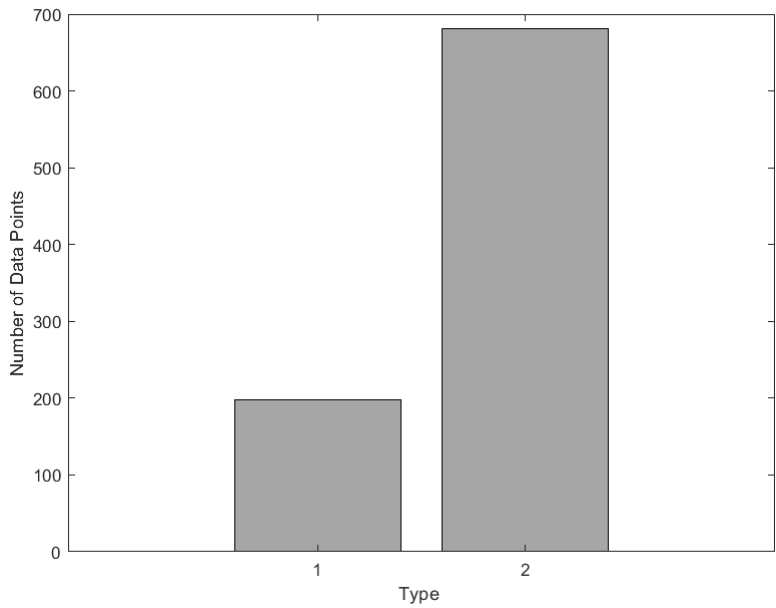


Figure 3.7. Geopolymer type distribution.

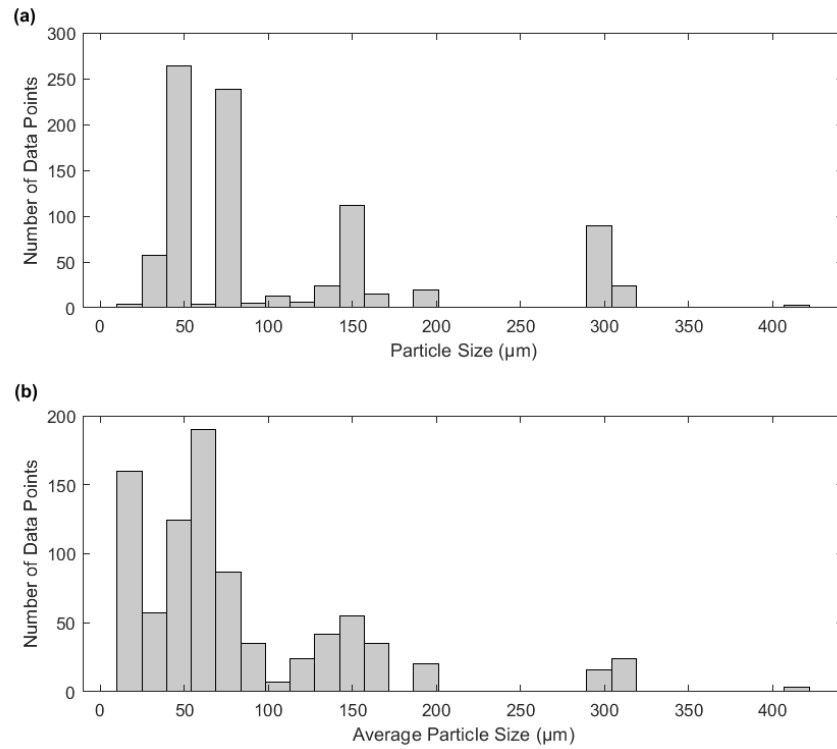


Figure 3.8. (a) Maximum and (b) average particle size distributions

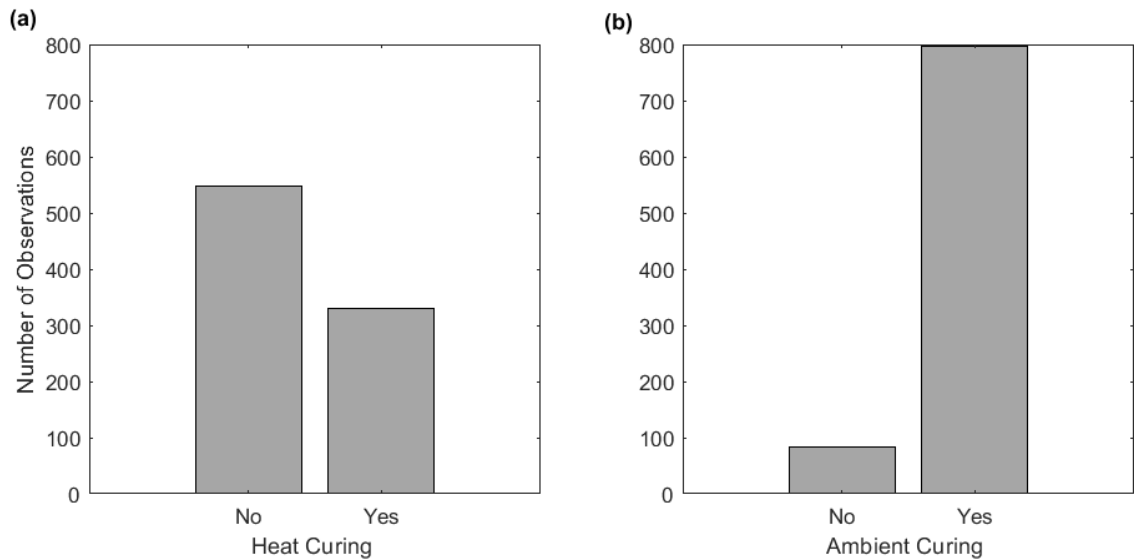


Figure 3.9. Distributions of the presence of (a) heat and (b) ambient curing

Heat curing and ambient curing temperature and time duration distributions of the data set are shown in Figure 3.10, in which water cured specimens are not included. For high

temperature curing cases (Figure 3.10a), there are two peaks at 40 °C and 60 °C, and accumulation around 80°C. High temperature curing usually lasts for < 5 days, but 7 and 28 day curing are also observed (Figure 3.10c). Ambient curing usually corresponds to temperature values around 25°C (Figure 3.10b) and time periods of mostly < 10 days, 29 days or 91 days (Figure 3.10d).

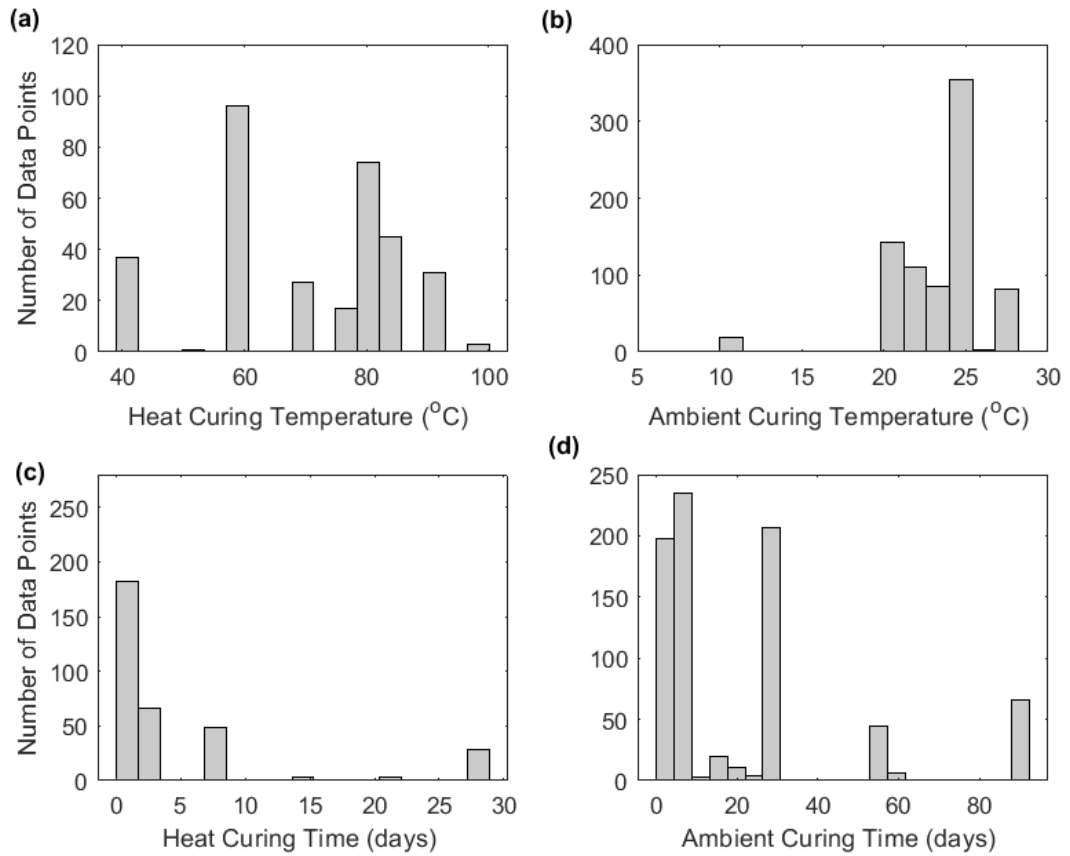


Figure 3.10. Distributions of (a) heat curing temperature, (b) ambient curing temperature, (c) heat curing time, and (d) ambient curing time.

3.2. Construction of Regression Models

The 879 samples in the dataset are shuffled and divided into a training set of 629 observations, and a test set of the 250 observations. Training was achieved in a three-dimensional parameter space: i) the maximum number of splits (divisions) representing the

size of the tree (S_{max}), ii) the minimum number of instances that exist in every node (P_{min}), iii) and the number of features used in a tree. Trees with three different maximum number of splits $S_{max} = \{30, 60 \text{ and } 150\}$ were constructed. Prediction accuracy of the constructed models were determined using 5-fold cross-validation (CV) with 20 repetitions; feature selection paths via forward selection and backward elimination are constructed using CV. The prediction accuracy in CV and hold-out test sets are assessed using the Root Mean Squared Error ($RMSE$), median and mean absolute errors ($medAE$ and MAE), between the experimental and predicted CS, Standard Error (SE) of the mean squared error and the sequence of the features selected by the algorithm and those used in the tree.

Decision trees are constructed using “fitrtree” routine in Statistics and Machine Learning Toolbox of MATLABTM, with S_{max} and P_{min} specified. The minimum leaf size (minimum number of members to allow leaf formation) is set to be equal to unity. For split predictors to be selected, “interaction-curvature” method is used, i.e. predictor selected at the split point minimizes chi-square p-value, testing the independence between individual predictors and predictor pairs with the response variable. The quadratic error tolerance is set at 1×10^{-6} . After a tree is constructed at its maximum size, it is pruned using CV (10-fold with 20 repetitions).

First, FS is used to select features. In this method, each predictor is subsequently added to the model and the regression tree is constructed as described previously. The calibration set consists of fourth fifth of the training set (5-fold CV), is used to construct for each predictor, and the set of predictors leading to minimum training $RMSE$ is selected. In each of FS algorithm, one additional predictor, leading to a maximum decrease in $RMSE$ of the training set is selected. In that way, all the predictors are selected along a FS path, and validation errors, using one fifth of the training set, are determined. Second, BE is used to select (or remove) predictors. Initially, a model is constructed using all the predictors. Then, the predictor, whose removal would lead to minimum increase in training $RMSE$ is removed; this step is iterated until all predictors are removed. Similar to FS path, predictions errors of validation set in terms of $RMSE$, $medAE$, MAE are calculated for the models constructed

along the BE path. Furthermore, stability indices of the $20 \times 5 = 100$ models during CV of each FS/BE step are calculated using Equation 2.23-2.26.

The third used algorithm in feature selection is a modified version of reduct set construction. First, all continuous variables, including the response variable, are discretised into 0 and 1, using the median values; for the categorical variable material, which consists of multiple levels, each of the different materials is taken as a separate binary variable. Then, a discernibility matrix (DM) is constructed via pairwise comparison. If the discretized values of the predictor variables of a pair of observations with different response variable values do not match, then the related entry of the DM is filled by one, otherwise, filled with zero. Hence, each entry of the discernibility matrix consists of the predictors, which take different values coinciding with the changes in the response variables. Entries of the discernibility matrix are sorted with respect to their cardinality, i.e. number of predictors discerning two samples (from one changed predictor till all of them). Weight of each predictor is calculated as shown on Equation 2.20 using the frequency of appearance of the predictors in the discernibility matrix and their cardinality in each entry. Predictor variables are included in the reduct set with respect to their weights, and inclusion of predictor stops when a specified threshold parameter (eps) is exceeded. After a reduct set is formed, a regression tree is constructed using the features in the reduct set only. It should be noted that if any of the material types is selected in the reduct set, then the Material predictor with all the levels, whether included in the reduct set or not, is used in constructing the regression tree. The computations stated above for feature selection methods are performed for various $\{S_{max}, P_{min}\}$ combinations.

4. RESULTS AND DISCUSSION

In this section, various models with different parameter values are constructed, assessed and a final interpretable model is used for explanatory purposes. Here, the following nomenclature for designating types of models is used: A letter abbreviation in the beginning refers to the feature selection method i.e. *FS* for Feature Selection via Forward Selection, *BE* for Backward Elimination, *R* via finding an optimum Reduct feature subset and *N* when none of the methods is used. The subscript indicates the number of features for the wrapper methods and the full tree and the threshold parameter for the reduct set based trees. The numbers in parentheses represent the maximum splits number S_{max} and minimum parent size P_{min} respectively. For instance, $FS_{14}(30,32)$ represents the regression tree model constructed using 14 features selected by Forward Selection with maximum split number and 32 minimum parent size equal to 30 and 32, respectively.

Trees with maximum number of splits $S_{max} = 30, 60$ and 150 are constructed and examined in Sections 4.1 to 4.3, respectively. For each distinct S_{max} , predictors are selected via Forward Selection (*FS*), Backward Elimination (*BE*), or Reduct Set methods within ranges of different values of P_{min} . The values of P_{min} , which yielded models with minimum *RMSE* and median absolute error (medAE) values and maximum stability metrics are used for model construction.

4.1. Regression Trees with $S_{max} = 30$

4.1.1. Constructing Regression Trees via FS and BE with $S_{max} = 30$

RMSE values obtained from CV on the regression trees constructed via FS using $S_{max} = 30$ are shown in Figure 4.1. *RMSE* gives an insight about the variation between CV folds indicating the prediction accuracy of the models, i.e. the one that will not lead to small prediction error using the whole data set. Here, trees with $P_{min} = 14$ to 38 are examined. Trees with $P_{min} = 20$ yield minimum error (with the lowest being $RMSE = 14.2$ MPa) and selected

for further analysis. Trees with higher $P_{\min}$ seem to lack details necessary to predict unseen data and those with lower $P_{\min}$ possibly yields overfitting

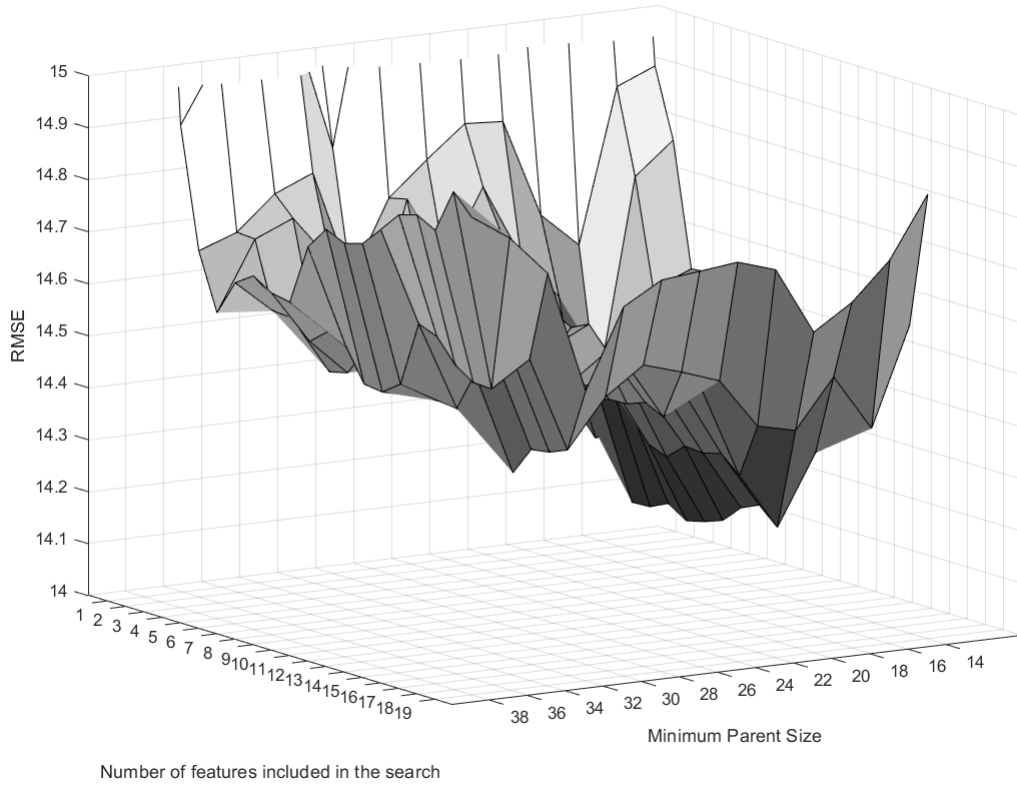


Figure 4.1. Surface plot of $RMSE$ obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 30.

$RMSE$ values obtained via employing CV for minimum $RMSE$ models (FS(30,20)) with respect to varying number of features selected by FS is shown in Figure 4.2a. Similar to that obtained for different values of $P_{\min}$, including a large number of predictors in the model increase prediction variance, and may lead to overfitting, while a set consisting of a small number of features may neglect important information about the process. However, trees with number of selected features $p_s = 9$ to 17 display similar $RMSE$ results. Along $RMSE$ Median Absolute Error ($medAE$) is another important measure of model quality, showing the error value separating 50% of the errors. Moreover, $medAE$ is a robust measure, and is not generally affected by gross errors. This might be both an advantage and

disadvantage in assessing predictive quality of models; *medAE* is not easily affected by outliers, however it may also disregard the systematic predictive shortcomings of the model upon certain input directions. Here, (Figure 4b) FS(30,22) model gives the smallest *medAE* with $p_s = 8$; is similar to that seen for *RMSE*, models with $p_s = 8$ to 17 yields the minimum *medAE* values. In other words, after including eight features, inclusion of additional features does not significantly improve the prediction accuracy. However, as shown in Figure 4c, number of predictors, which yield the minimum *RMSE* and *medAE* values, unfortunately, correspond to models with minimum stability. This means that trees, constructed via FS yielding 8 to 17 predictors, give highly variable rules (decision trees) for different calibration sets in CV. Stability is found to be minimum for only one selected feature, meaning that changes in the trained data for each CV fold leads to a variety of different predictors selected for one feature case.

The same CV analysis may be repeated for regression trees constructed without any feature selection algorithm, using all 19 predictors. Figures 4.2 d,e and f refer to the trees constructed with all the 19 variables. Using all predictors, two local minima for *RMSE* at $P_{min} = 20$ and $P_{min} = 32$ are seen. Minimum *medAE* is obtained for $P_{min} = 20$ to 32. It is seen that stability of the trees without feature selection is much higher than those on which FS is applied; the highest stability value is found to be equal to 0.68, much higher than 0.4-0.5 values obtained by the best performing FS method (Figure 4.2c).

The number of predictors included in the regression tree versus that selected by the feature selection method for FS(30,20) and FS(30,22) is plotted in Figure 4.3a. It can be observed that after the seventh feature the number of predictors used by the tree decreases drastically, implying that the predictive model either do not need additional features, or new features included in the tree is likely to replace the previously included ones. The same conclusion may be reached based on the decreasing stability after the eighth predictor. The stability of the models chosen with respect to the tree level is displayed in Figure 4.3b. As expected the full models have greater stability compared with less feature models, although, no clear comparison is possible as P_{min} also changes. Regarding full models, stability till the seventh level seems to be similar to FS models yielding minimum *RMSE* and *medAE*, i.e.

$FS_{19}(30,32)$ and $FS_{19}(30,24)$, respectively, but stability drops afterwards for the larger tree ($FS_{19}(30,24)$) so the priority of the features for the tree changes as the tree gets larger. It is, then, expected to obtain more stable rules from the upper nodes of the trees irrespective of P_{min} . It is also worth noting that all the curves, except the part of $FS_{12}(30,20)$ shown, show a minimum, so, although different predictors may have been included in upper levels of trees from fold to fold, a similar set of predictors is included after the fourth level. In summary, for $FS_8(30,22)$, $FS_{19}(30,24)$ and $FS_{19}(30,32)$, the same feature has been selected for almost all the CV-folds at the root node, while levels one to four were constructed with rather different predictors.

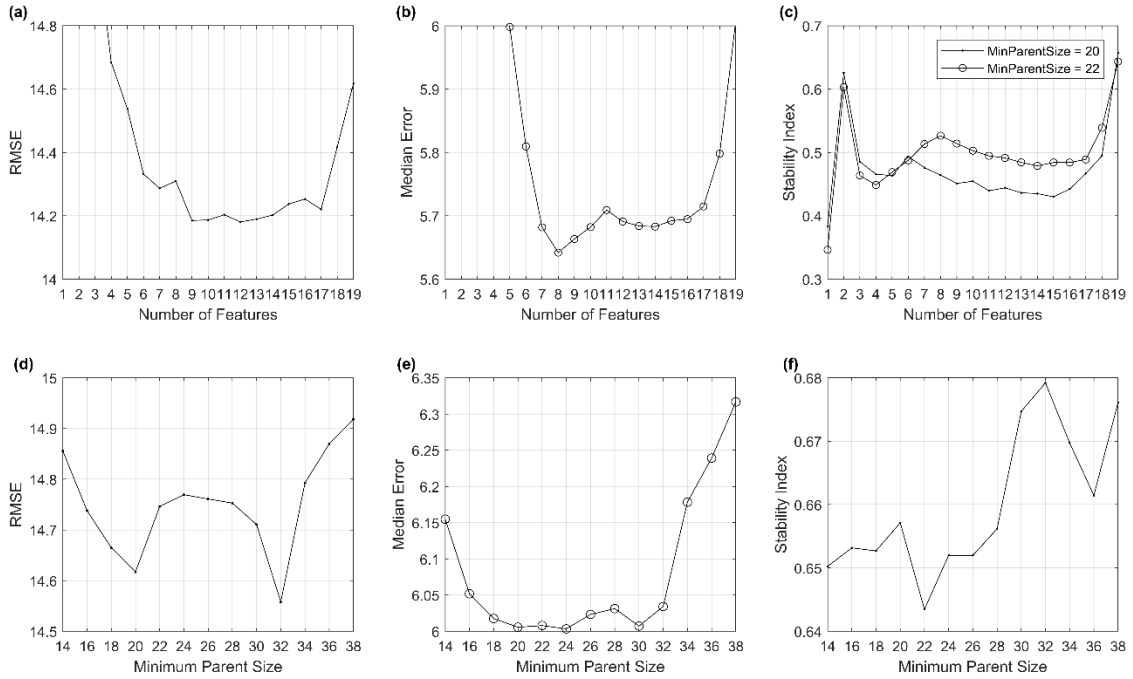


Figure 4.2. (a) *RMSE*, (b) *MAE*, and (c) stability index with respect to features selected via FS; (d) *RMSE*, (e) *MAE*, and (f) stability index with respect to minimum parent size in full trees constructed for $S_{max} = 30$.

Proceeding to BE feature selection method, *RMSE* of trees with $P_{min} = 30$ to 36 is shown in Figure 4.4. Minimum *RMSE* is observed for BE(30,32) for P_{min} value, larger than that seen for FS.

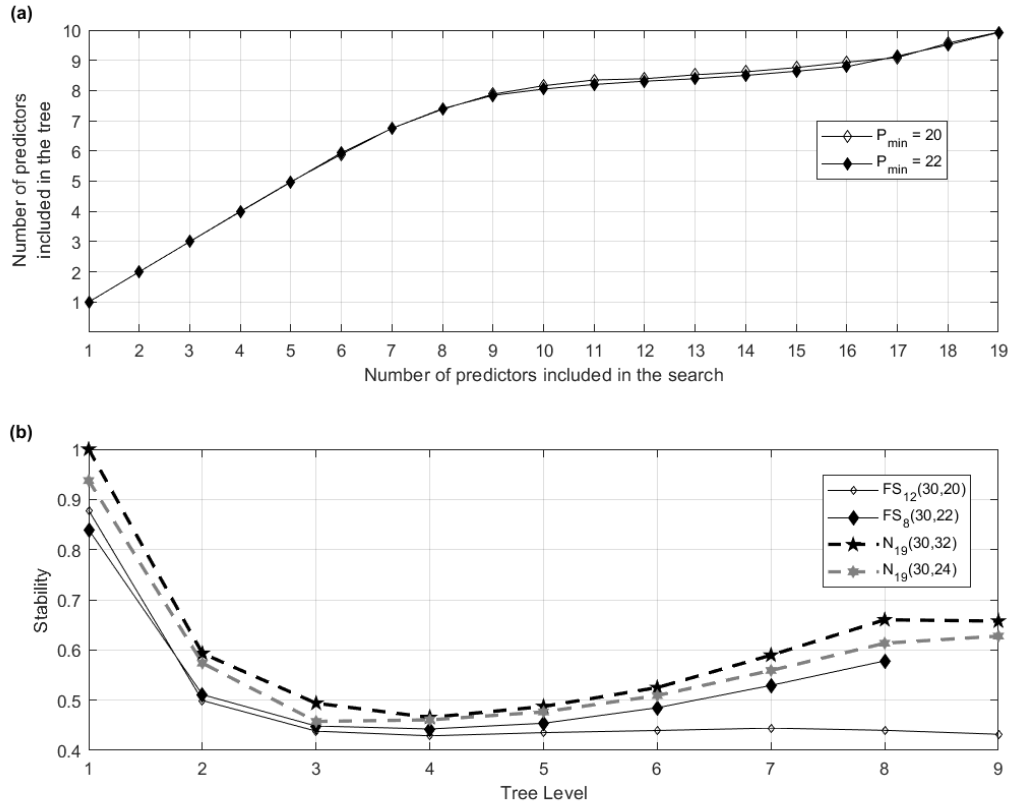


Figure 4.3. (a) Number of predictors included in the trees versus number of features selected by the FS search for $FS(30,20)$ and $FS(30,22)$, and (b) stability of the trees with respect to the levels for $FS_{12}(30,20)$, $FS_8(30,22)$, and full models $N_{19}(30,32)$ and $N_{19}(30,24)$.

$RMSE$, $medAE$ and stability values with respect to the number of features selected by BE on $BE(30,30)$ and $BE(30,32)$ are shown in Figures 4.5a,b and c respectively. $RMSE$ and $medAE$ decreases and then increases again showing a minimum for 11th and ninth features for $BE(30,32)$ and $BE(30,30)$ for $RMSE$ respectively. Since, $BE(30,32)$ is optimum based on $RMSE$ and $medAE$ values, these parameters seem to be reliable in model construction. The range of number of selected predictors with minimum error again corresponds to that with the lowest stabilities, similar to that seen in FS case (Figure 4.2c).

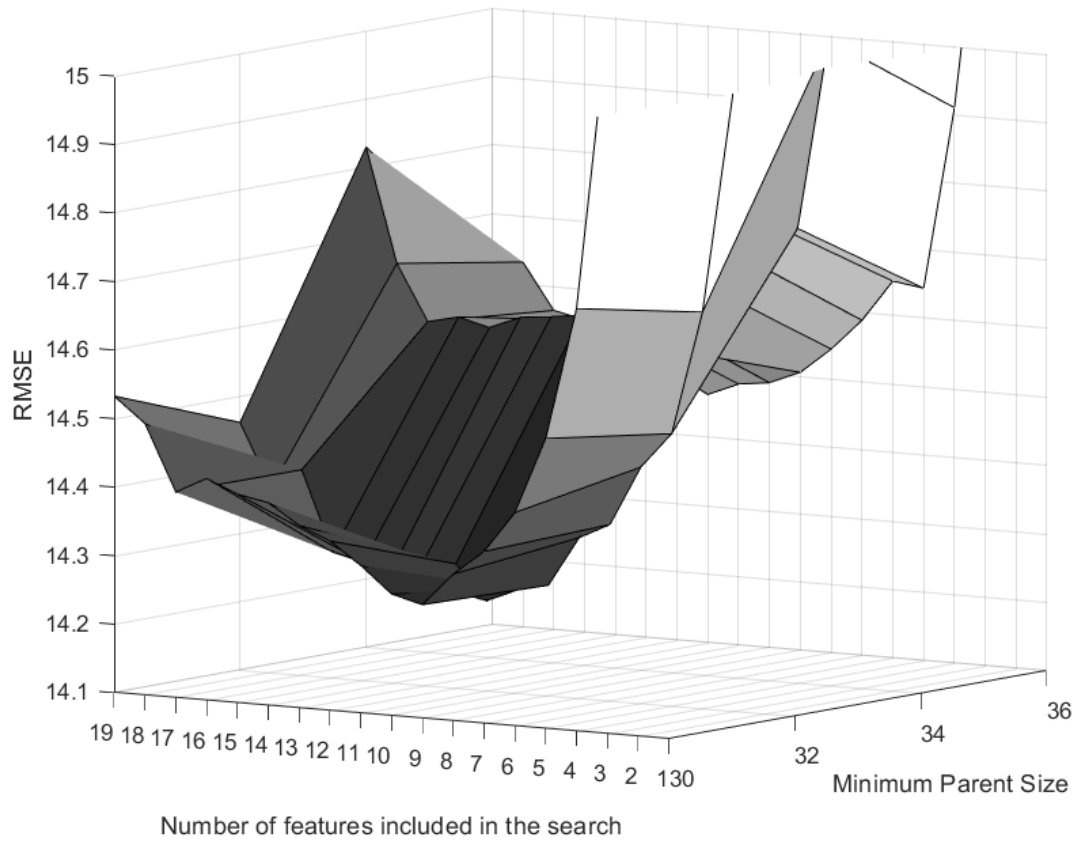


Figure 4.4. Surface plot of $RMSE$ obtained via CV with respect to minimum parent size parameter and number of features selected by BE for trees with maximum number of splits equal to 30.

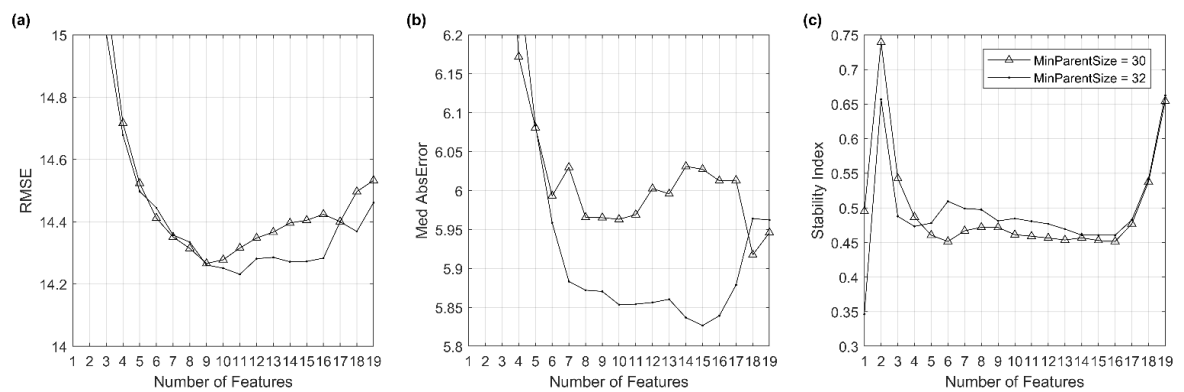


Figure 4.5. (a) $RMSE$, (b) MAE, and (c) stability index with respect to features selected via BE for $S_{max} = 30$.

The number of features used by the tree vs those selected by search mechanism for trees with smallest CV prediction errors obtained via BE and FS searches is shown in Figure 4.6a. In BE, approximately all predictors are included in the tree for small number of selected features, and the number of predictors in the tree almost stay constant as more features are selected. More features are chosen by BS(30,32) search for the same number of predictors included in the tree. Nevertheless, a premature conclusion should not be reached about the number of the predictors chosen by the tree based on feature selection algorithm, as same P_{min} trees should be compared. The greatest level stability is observed for the full tree, for which all predictors are “offered” to the regression tree. The similarity between the features of each level selected by each CV-fold is lower for $FS_{12}(30,20)$, showing That FS method is likely to produce more variable subsets of predictors.

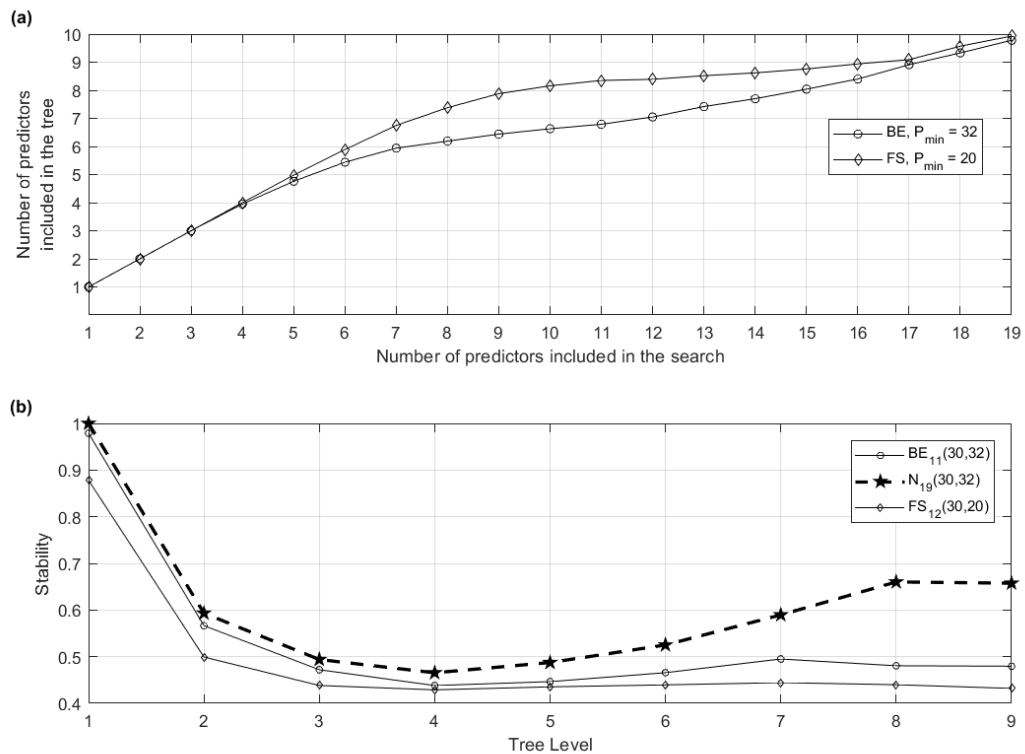


Figure 4.6 (a) Number of predictors included in the trees versus number of features selected by the FS search for BE(30,32) and FS(30,20) and (b) Tree level stability for BE₁₁(30,32), FS₁₂(30,20) and N₁₉(30,32).

Using trees with the same P_{min} for BE and FS would make it a more reliable comparison of these two search methods, hence Figures 4.7a and b is drawn using $P_{min} = 30, 32, 34$ and 36 for both BE and FS methods. Hereby, it is seen that for the same number of features included in the search, the tree includes more features selected by BE than those selected FS, between seven and 16 predictors. It is important to note that stability of the trees constructed by BE is also higher than those constructed by FS, in accordance with the greedy feature inclusion mechanism of FS method.

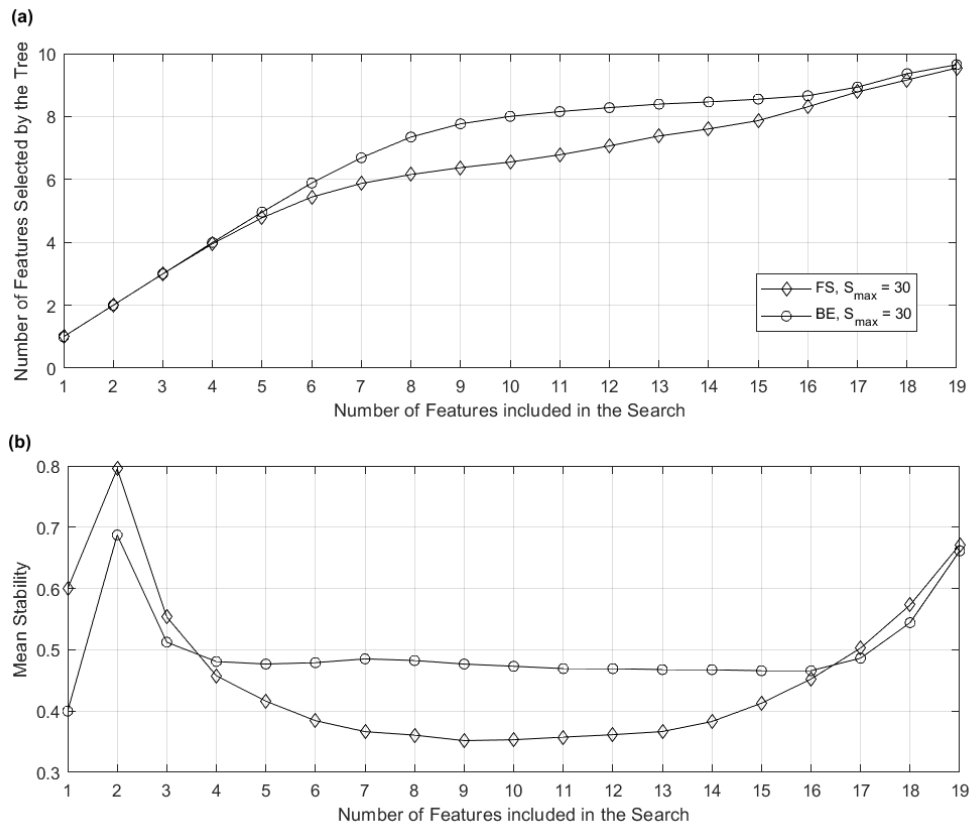


Figure 4.7. (a) Mean number of features included in the tree versus number of features included in the search for $S_{max} = 30$ and (b) Mean tree feature stability for $S_{max} = 30$.

The overall performance of the optimum BE and FS trees and the Full trees (19 predictors included for tree construction) is summarized in Figure 4.8, demonstrating the their $RMSE$ and stability values. Upper left corner of the figure represents the ideal tree models, yielding smalling $RMSE$ and having high stability. Moving to right, predictive

accuracy of the trees degrades, and moving down, the rules obtained from the trees are less stable. Here, it is seen that , with a small sacrifice in predictive accuracy, full models are more preferable; neglecting even one predictor decreases stability drastically and when six to 18 features are selected, the performance of the models is similar (slightly less erroneous but also less stable for FS).

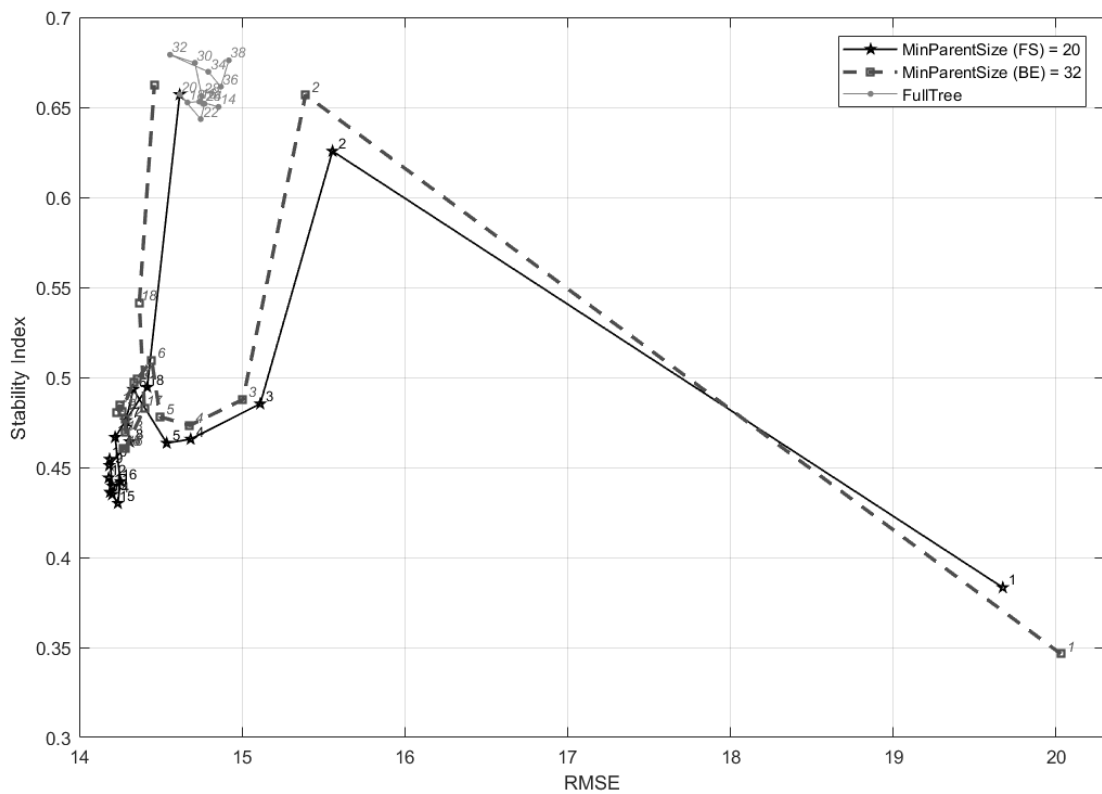


Figure 4.8. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 30$

CV and test set results of the minimum *RMSE* and *medAE* models for FS, BE and none feature selection method are tabulated in Table 4.1. *RMSE*, *MAE* and *medAE* of all models are close to each other and around 14, 10 and 6 MPa respectively. FS and BE models are marginally better in prediction accuracy, but full models outstand in terms of stability, which is above 0.65. It is interesting to note that full tree models use no more than ten variables, on the average. Hence, the lower stability of trees obtained via search methods is not derived from the limitation in variable number, but from which variables are selected. When it comes

to performances of the models on the test data, Trees of $BE_{11}(30,32)$ and $BE_{15}(30,32)$ trees give more accurate predictions than full trees. R^2_p are around 0.60 for all, with 0.62 for BE , and $RMSE$ is surprisingly lower than that predicted by CV for all except $FS_{12}(30,20)$. $BE_{11}(30,32)$ and $BE_{15}(30,32)$ outstand with respect to all the test error measures, except $medAE$ and q_{90} , for which $FS_{12}(30,20)$ and $N_{19}(30,32)$ prevail. It should be mentioned that 90% of the errors are below 24.5 MPa for all methods, but maximum absolute error is below 49 MPa for FS and BE, while reaches 69 MPa for the full tree model.

Table 4.1. Summary of CV and Test Results of Minimum $RMSE$ and $MedAE$ models

	$S_{\max}$	30					
Parameters	SD	FS		BE		None	
	$P_{\min}$	20	22	32	32	32	24
	p_s	12	8	11	15	19	19
CV results	$RMSE$	14.2	14.4	14.2	14.3	14.6	14.8
	MAE	9.45	9.51	9.56	9.58	9.82	9.92
	$medAE$	5.70	5.64	5.85	5.83	6.03	6.00
	SE	51.7	43.3	44.4	45.6	44.6	44.5
	SI_{Tree}	0.44	0.53	0.48	0.46	0.68	0.65
	$\bar{p}_T$	8.34	7.40	8.15	8.54	9.78	9.91
Test Results	p_T	7	6	8	9	9	9
	r_p	0.77	0.79	0.81	0.81	0.78	0.77
	R^2_p	0.57	0.59	0.62	0.62	0.57	0.55
	$RMSE$	14.5	14.0	13.6	13.6	14.5	14.8
	MAE	9.96	9.8	9.63	9.63	10.2	10.2
	$medAE$	6.03	6.45	6.33	6.33	6.58	6.24
	$q_{.90}$	24.5	23.1	24.0	24.1	22.6	24.1
	$MGAE$	34.9	33.3	31.9	31.9	33.9	34.4
	$maxAE$	48.9	48.5	47.8	47.8	68.8	68.8

Experimental CS data is drawn with respect to predicted CS from BE₁₁(30,32) for the test dataset in Figure 4.9. For the relatively small tree size, $S_{\max} = 30$, prediction accuracy is acceptable with $R_p^2 = 0.62$, and there seems to be no/negligible bias in predictions.

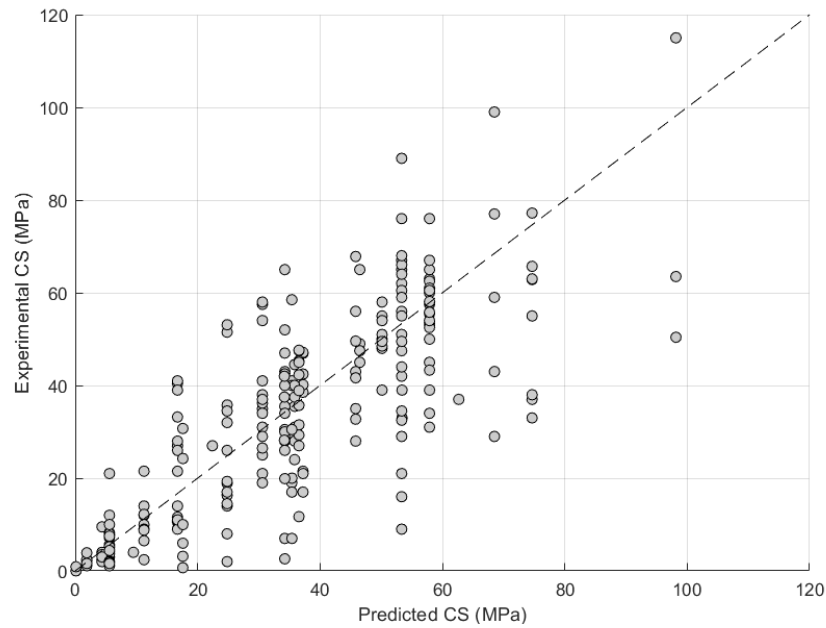


Figure 4.9. Comparison of predictions via BE₁₁(30,32) model with the experimental results on the test set.

To have a preliminary idea on the importance of predictors in the trees, features selected by BE₁₁(30,32) method and those used by the tree are shown in Figure 4.10 with decreasing feature importance computed based on the frequency and the level that the predictor appears on the tree. Material is the most important feature whereas Water/ Al ratio and Aggregate Proportion follow. Aggregate and additive existence and also additive percent features although selected by the algorithm are not used in the tree construction. The variable nature of the feature selection algorithm and the decision tree is noticeable when the feature importance of BE₁₁(30,32) is constructed based on the whole data (including the training and test set), e.g. Na/Al ratio does not appear for training set, but is on the third order of importance (Figure 4.11). Moreover, particle and average particle size, which highly depend on each other, both appear on the secondly constructed tree. It would be expected that a stable predictor selection algorithm would yield similar predictor importance profiles, so that the resulting rules would be reliable. This shows that while Material type and W/Al ratio

seem to be consistently important predictors in determining CS, contribution of other predictors varies significantly for different training sets.

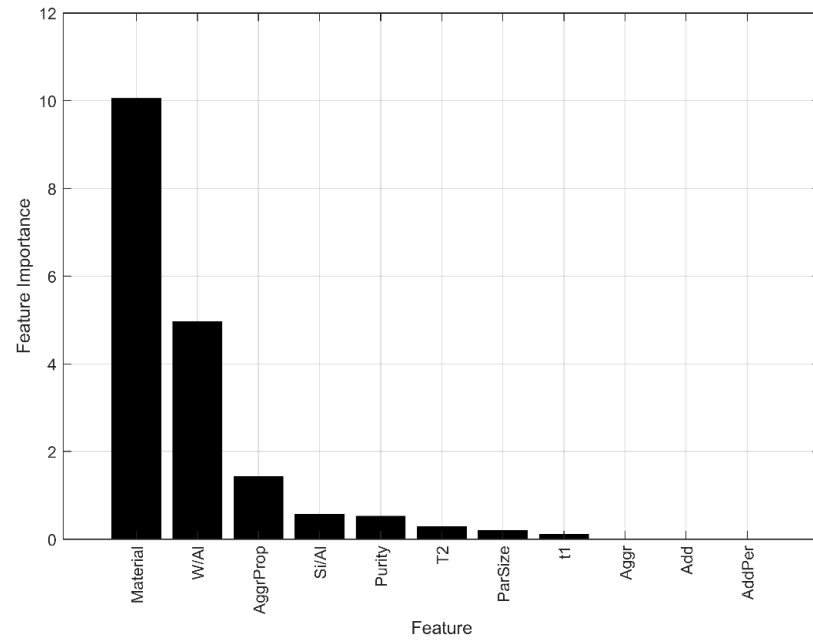


Figure 4.10. Feature importance in the tree constructed via $BE_{11}(30,32)$ using the training set.

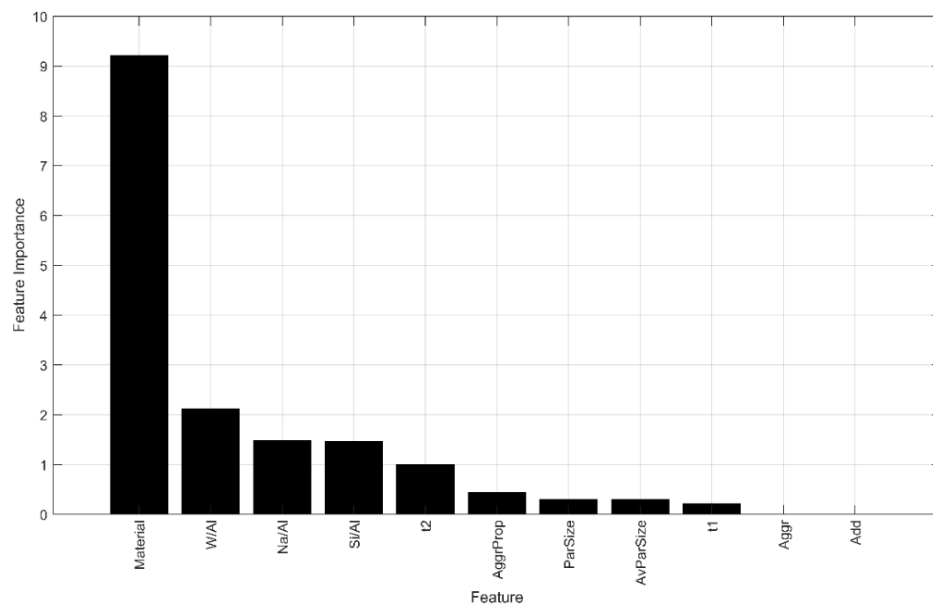


Figure 4.11. Feature importance in the tree constructed via $BE_{11}(30,32)$ using the whole dataset.

4.1.2. Constructing Regression Trees via Reduct Sets with $S_{max} = 30$

CV results for decision trees with $S_{max} = 30$ constructed using optimum reduct feature subset are shown in Figure 4.12. Compared to wrapper algorithms (FS and BE), RMSE is seen less likely to change with respect to P_{min} . The threshold value used in reduct set computations (eps) is related with the number of different classes obtained from the training set; as a larger eps is used, the classes obtained to explain the training set and the corresponding reduct set is smaller, via neglecting a larger portion of object pairs to be distinguished in the discernibility matrix. Hence, decreasing the eps value increases the size of the predictor subset, while increasing the eps value yields the opposite effect. Here, it is seen that $R_{5e-4}(30,24)$ and $R_{1e-5}(30,26)$ give the smallest *RMSE* values. Similarly, $R_{5e-4}(30,24)$ is the most ideal based on *medAE* plot as well (Figure 4.12b). Stability indices do not seem to depend on P_{min} but mostly on the number of classes used for reduct set formation (Figure 4.12c): larger eps values yield more stable trees. It is important to note that stability values are all above 0.7, much higher than those obtained by wrapper methods (Figure 4.5c). The stability of the tree levels is also higher for $R_{5e-4}(30,24)$ when the trees giving the lowest *RMSE* are plotted (Figure 4.12d).

It has been noted that full trees were more advantageous than trees constructed using FS and BE in terms of *RMSE* and stability values obtained by CV. Showing representative models obtained via FS and BE with reduct sets on the *RMSE* vs. stability plot shows that prediction accuracy of the methods are close, while rough set based methods have much higher stability, particularly $R_{5e-4}(30,24)$ is a promising model with stability above 0.85 and *RMSE* equal to 14.2 MPa (Figure 4.13).

CV and Test results for optimum reduct set based trees and full trees are tabulated in Table 4.2. It is seen that CV error results and stability values are equal or better for all reduct set trees. The number of features in the reduct set almost falls by half when threshold is increased from 1×10^{-5} to 5×10^{-4} , but no more than nine features are used for any tree constructed. test performance of $R_{1e-5}(30,30)$ is the best among those compared with high coefficient for determination and correlation between predicted and experimental CS. The

maximum absolute error is about 69 MPa for all the trees, so it is seen that trees with $S_{\max} = 30$ do not thoroughly exploit the relations in the training set to be able to predict all the test points.

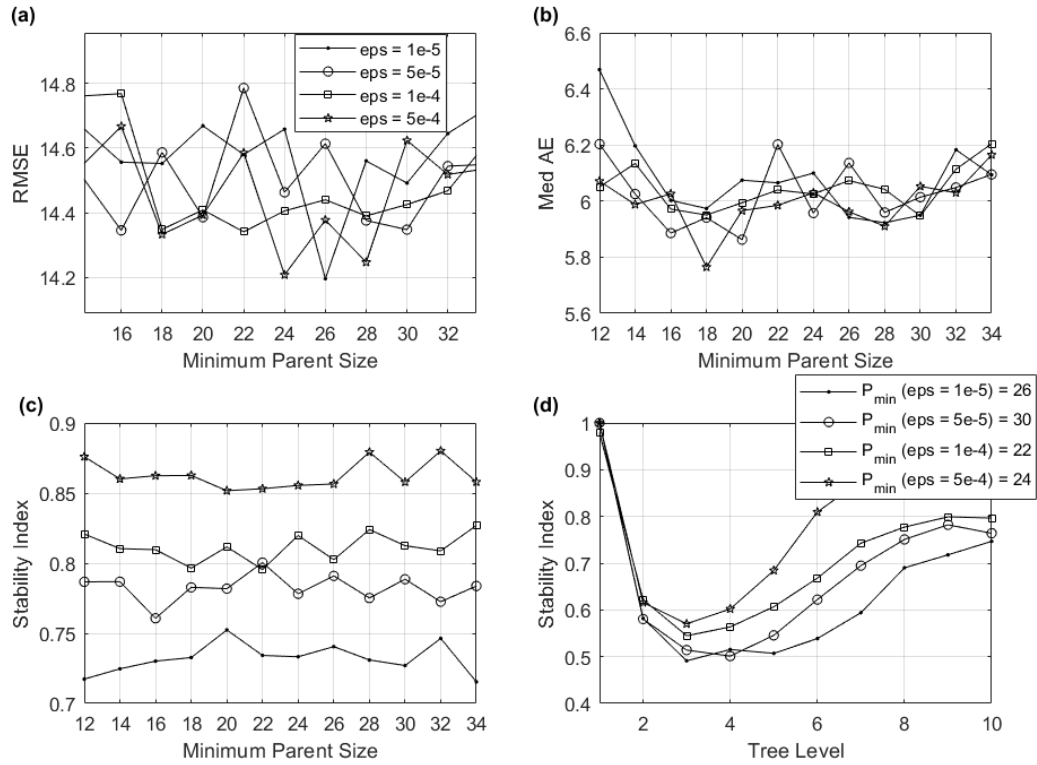


Figure 4.12. (a) *RMSE*, (b) *MedAE* and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $\epsilon = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{\max} = 30$.

The test performance of $R_{5e-4}(30,24)$ constructed based on the training set and the feature importance is shown in Figures 4.14 and 4.15, respectively. Reasonable correlation between the predicted and experimental data can be observed. Material is the most important feature (as for the wrapper cases) with aggregate proportion (which also appeared for $BE_{11}(30,32)$ based on training data in the third place) being the second most important one. Molar ratios and ambient curing time are also significant. When the same tree $-R_{5e-4}(30,24)$ - is constructed using the whole data set although the order of importance is slightly different than that described before, with aggregate proportion to be on the fourth place, none of the

lately appearing features is absent on the previous tree (Figure 4.16). Hence, the stability of the rules derived by reduct set trees is demonstrated here.

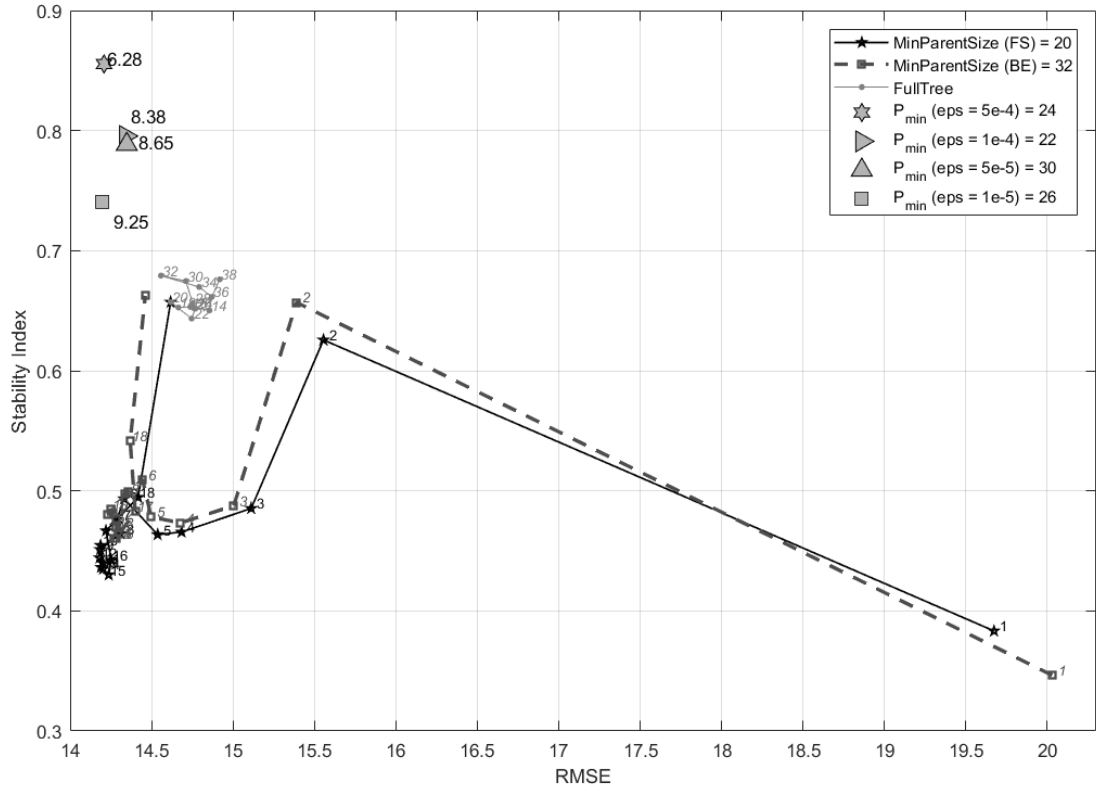


Figure 4.13. FS, BE and Reduct set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 30$.

Table 4.2. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 30$.

Parameters	$S_{\max}$	30					
	$P_{\min}$	26	30	22	24	32	24
	$\epsilon \times 10^{-5}/SA$	1	5	10	50	None	
CV results	RMSE	14.2	14.3	14.3	14.2	14.6	14.8
	MAE	9.60	9.70	9.70	9.66	9.82	9.92
	medAE	5.94	6.01	6.04	6.03	6.03	6.00
	SE	41.7	45.6	45.1	43.3	44.6	44.5
	SI_{Tree}	0.74	0.79	0.80	0.86	0.68	0.65

Table 4.2. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 30$
(cont.).

Test Results	$\bar{p}_A$	15.4	12.7	11.9	7.08	19.0	19.0
	$\bar{p}_T$	9.25	8.65	8.38	6.38	9.78	9.91
	p_A	16	13	12	6	19	19
	p_T	9	9	8	6	9	9
	r_p	0.77	0.80	0.78	0.76	0.78	0.77
	R^2_p	0.55	0.59	0.56	0.53	0.57	0.55
	$RMSE$	14.8	14.1	14.6	15.1	14.5	14.8
	MAE	10.2	9.62	10.1	10.6	10.2	10.2
	$medAE$	6.24	6.08	6.15	6.51	6.58	6.24
	$q_{.90}$	24.1	22.2	23	25.2	22.6	24.1
	$MGAE$	34.4	33.5	33.6	35.2	33.9	34.4
	$maxAE$	68.8	68.8	68.8	68.8	68.8	68.8

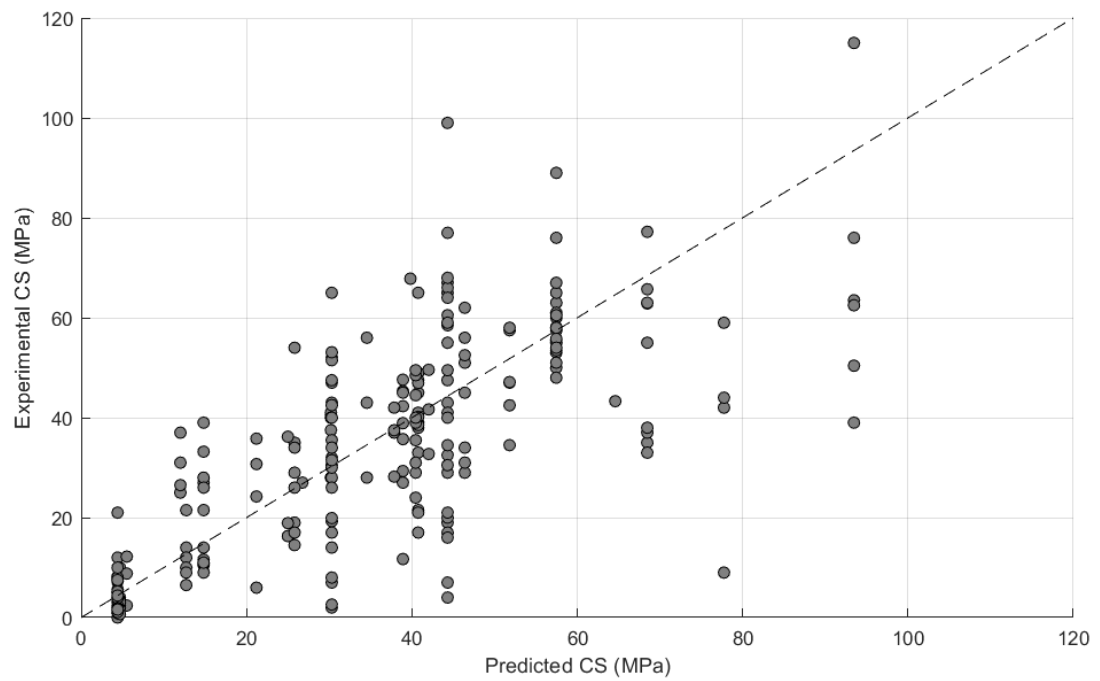


Figure 4.14. Comparison of predictions via $R_{5e-4}(30,24)$ model with the experimental results on the test set.

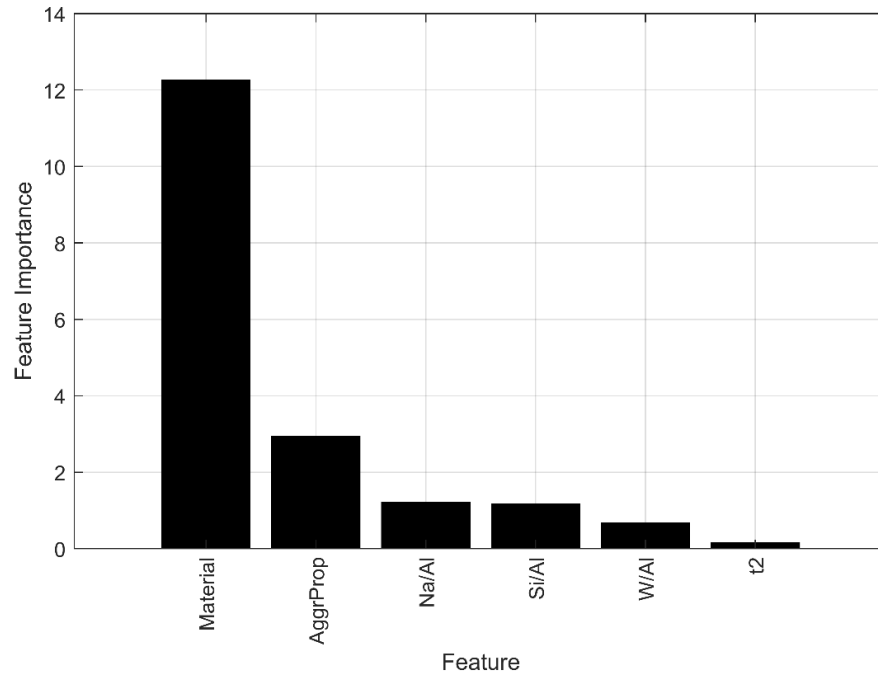


Figure 4.15. Feature importance in the tree constructed via $R_{5e-4}(30,24)$ using the training set.

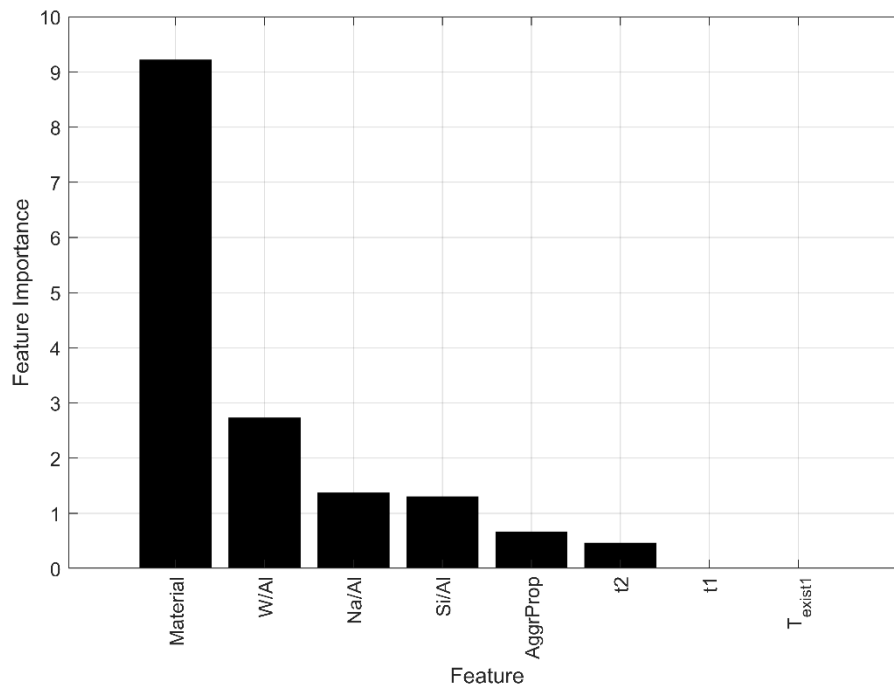


Figure 4.16. Feature importance in the tree constructed via $R_{5e-4}(30,24)$ using the whole dataset.

The CV and test set performance comparison between the best tree obtained by the filter and the wrapper algorithms is shown in Table 4.3. It is seen that $BE_{11}(30,32)$ has smaller CV errors compared to rough set method, but $R_{5e-4}(30,24)$ is preferable for its higher stability. It is also seen that BE tree has a significantly higher prediction accuracy for the test; it should be noted that $BE_{11}(30,32)$ is constructed by eight features, two more compared to $R_{5e-4}(30,24)$.

Table 4.3. CV and Test Result Comparison between most stable reduct based decision tree $R_{5e-4}(30,24)$ and best wrapper tree $BE_{11}(30,32)$ for $S_{\max} = 30$.

Parameters	$S_{\max}$	30	
	$P_{\min}$	24	32
	$\varepsilon \times 10^{-5}/SA$	5	BE
CV results	$RMSE$	14.2	14.2
	MAE	9.66	9.56
	$medAE$	6.03	5.85
	SE	43.3	44.4
	SI_{Tree}	0.86	0.48
	$\bar{p}_A$	7.08	11.0
	$\bar{p}_T$	6.28	8.15
Test Results	p_A	6	11
	p_T	6	8
	r_p	0.76	0.81
	R^2_p	0.53	0.62
	$RMSE$	15.1	13.6
	MAE	10.6	9.63
	$medAE$	6.51	6.33
	q.90	25.2	24.0
	$MGAE$	35.2	31.9
	$maxAE$	68.8	47.8

4.2. Regression Trees with $S_{\max} = 60$

4.2.1. Constructing Regression Trees via FS and BE with $S_{\max} = 60$

As the tree becomes deeper (increasing $S_{\max}$), it is expected that smaller $P_{\min}$ are required to capture all the information provided by the data set without leading to overfitting. So, compared to $S_{\max} = 30$, which yielded smallest RMSE around $P_{\min} = 20$, models with $P_{\min}$ values below 18 gave the smallest RSME, with $P_{\min} = 14$ to be the smallest (Figure 4.17).

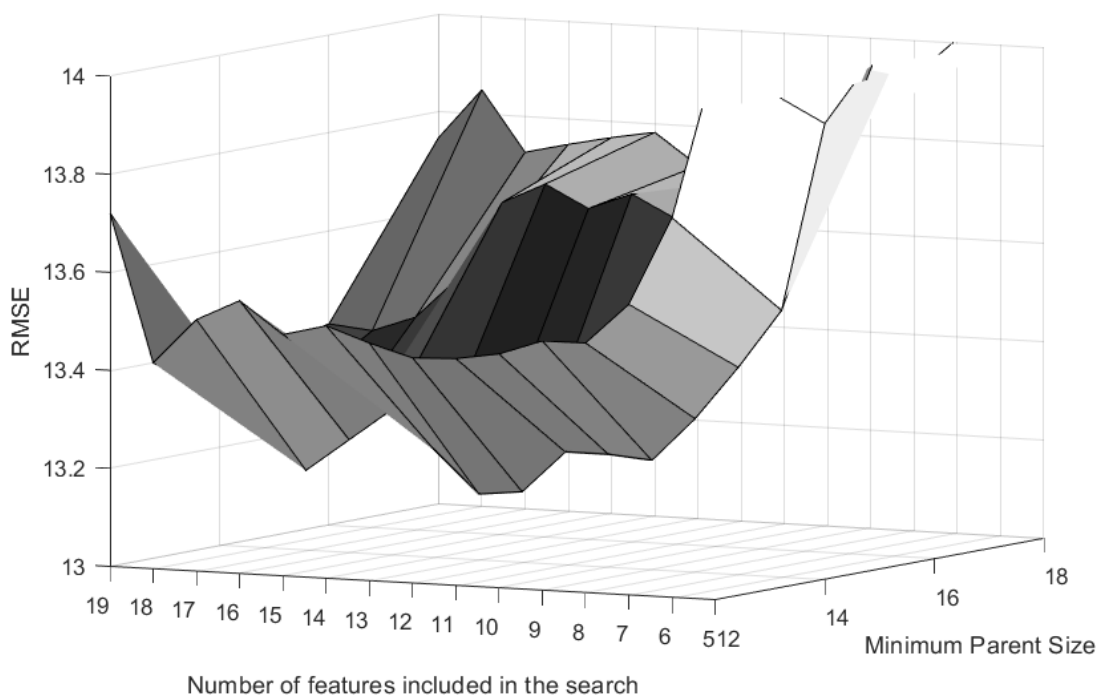


Figure 4.17. Surface plot of $RMSE$ obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 60.

Minimum $RMSE$ and $medAE$ trees for $S_{\max} = 30$ and 60 are compared in Figures 4.18 a,b and c. Similar to that seen for $S_{\max} = 30$, $RMSE$ and $medAE$ values show a convex behavior for and the number of predictors which yield the lowest tree stability again

coincides with those that yield the smallest errors for $S_{max} = 60$. Increasing the split size decreases both errors by ~ 1 MPa. When full models are compared (Figures 4.18 d,e and f) $RMSE$ and $medAE$ are decreased by > 1 MPa, and, in comparison with FS trees, full model stabilities are higher for $S_{max} = 60$. For further analysis, $FS_{13}(60,14)$ is chosen as the representative model.

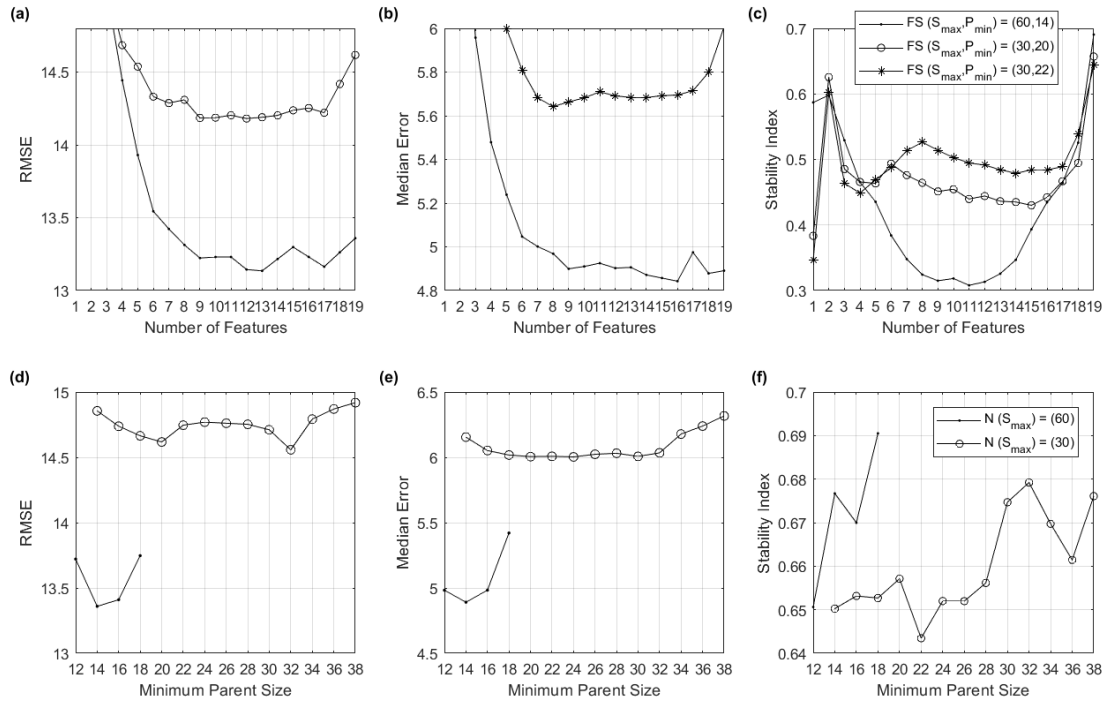


Figure 4.18. (a) $RMSE$, (b) MAE , and (c) stability index with respect to features selected via FS; (d) $RMSE$, (e) MAE , and (f) stability index with respect to minimum parent size in full trees constructed for $S_{max} = 30$ and 60.

For $S_{max} = 60$, the number of features included in the tree is almost identical to that obtained for $S_{max} = 30$ until the $\sim 13^{th}$ feature, as seen for $FS(60, 14)$ in Figure 4.19a. In this regard, there seems to be some essential predictors, which should be unavoidably used for the trees to grow, and additional predictors are required for the lowest branches. Compared to $FS_{12}(30,20)$, stability of the individual tree levels is generally higher for $FS_{13}(60,14)$, a clear conclusion cannot be reached since both P_{min} and p_s change (Figure 4.19b). In the case that full trees are compared, tree level stabilities are slightly higher for $S_{max} = 30$; though

overall tree stability increases when $S_{\max}$ is increased from 30 to 60, stability is decreased for the tree levels, showing that a greater variety of features are used in each CV fold.

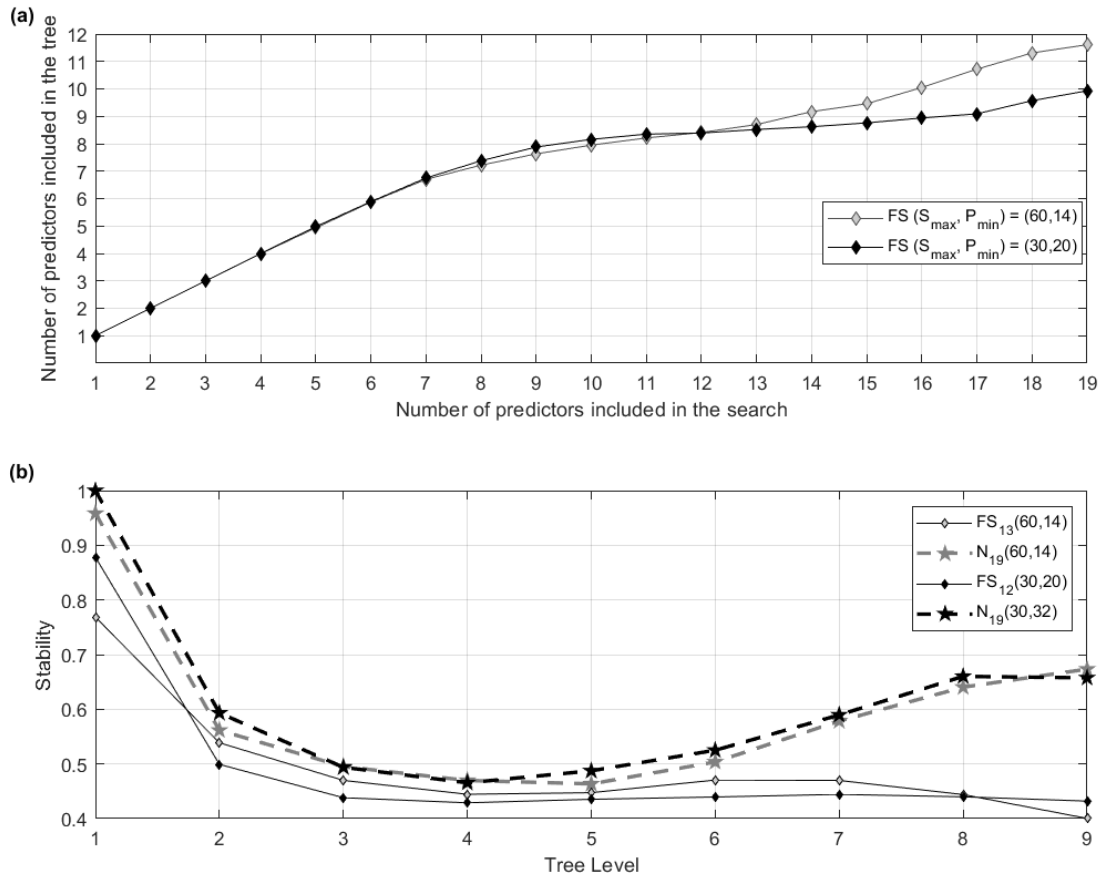


Figure 4.19. (a) Number of predictors included in the trees versus number of selected by the FS search $FS(30,20)$ and $FS(60,14)$, and (b) tree level stability for $FS_{12}(30,20)$, $FS_{13}(60,14)$, $N_{19}(30,32)$ and $N_{19}(60,14)$.

Figure 4.20 shows $RMSE$ values obtained from the models constructed using BE for $S_{\max} = 60$. In akin to FS results, smaller $P_{\min}$ values are likely to yield smaller prediction errors; $BE_{10}(60,12)$ is the model that gives the minimum $RMSE$.

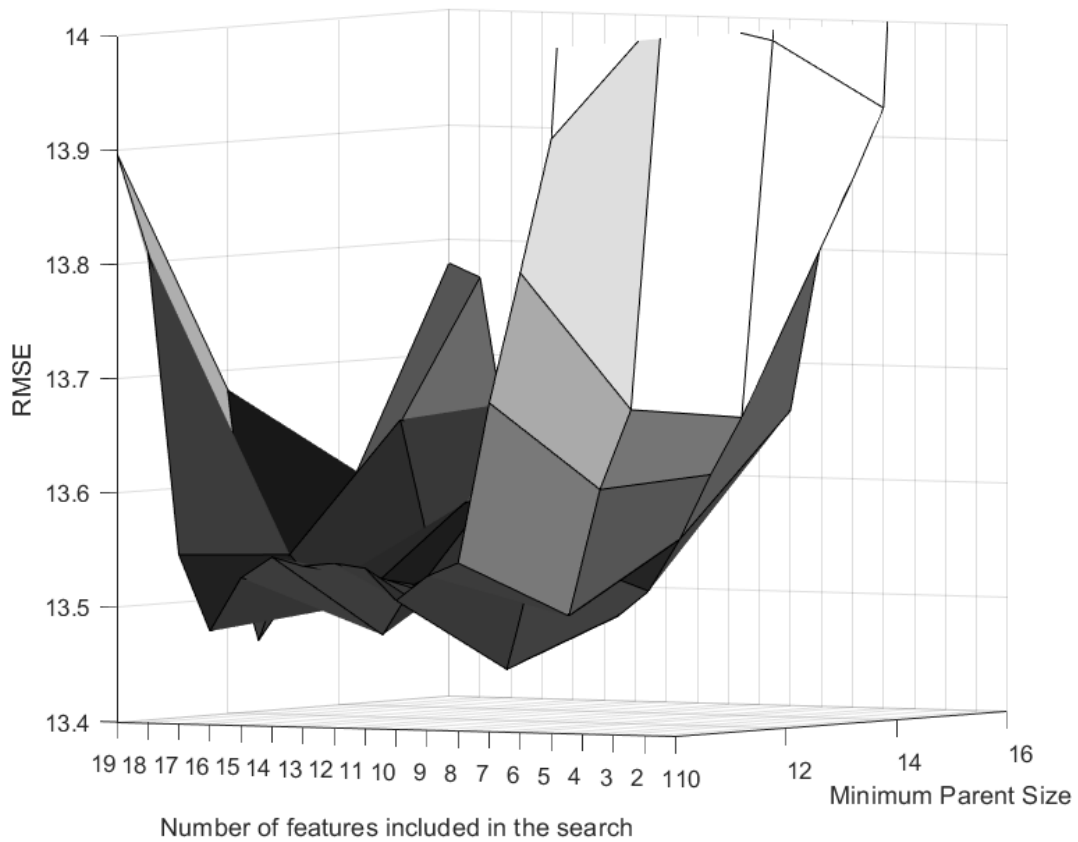


Figure 4.20. Surface plot of $RMSE$ obtained via CV with respect to minimum parent size parameter and number of features selected by BE for trees with maximum number of splits equal to 60.

BE(60,12) errors when considered p_s based, are lower than BE(30,32) as for the FS case with a difference about 1 MPa (Figure 4.21). Regarding stability, although, no certain comment cannot be made for features less than nine, BE(60,12) trees' stability is lower afterwards.

The number of features selected by the search algorithm with respect to those used for the tree construction are exhibited in Figure 4.22a. For $S_{max} = 60$, number of features included in the tree increases after the eighth feature selected by BE, in the region where $S_{max} = 30$ tree is saturated, i.e. the extra features selected by the search algorithm are not used during the tree construction. So, as the tree gets deeper, the relation between the number of features selected and those used approaches linearity (Figure 4.22a). In contrast, no more than 13

features are finally preferred. When tree level stability is analysed, there seems to be no significant difference between those constructed using $S_{max}=30$ and 60 (Figure 4.22b).

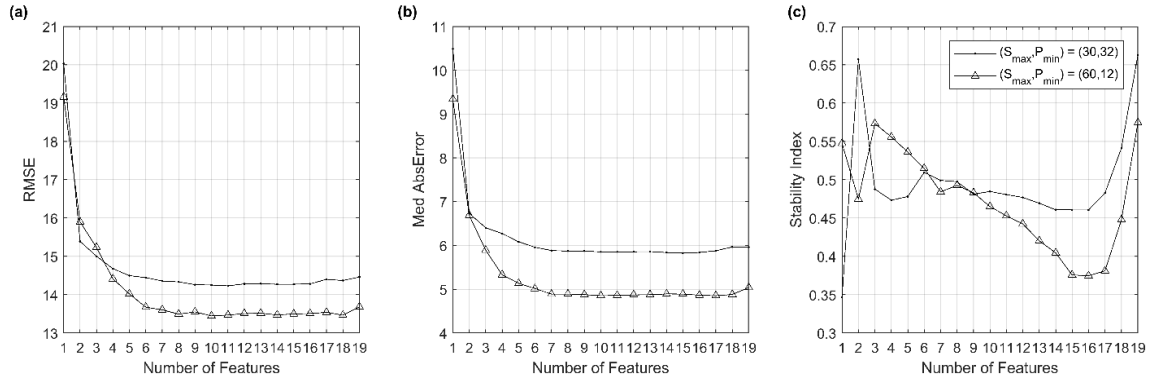


Figure 4.21. (a) *RMSE*, (b) *MAE*, and (c) stability index with respect to features selected via BE for $S_{max} = 30$ and 60.

It has been noted for $S_{max} = 30$ that the number of features included in the tree initially increases, followed by a stabilization with respect to the number of features used selected by the search algorithm (Figure 4.7a). For $S_{max} = 60$, however, the number of features needed for the trees to be constructed seems to increase constantly. For FS, on the other hand, the same S-Shape is still observable (Figure 4.23a). Stability of the trees constructed via FS is lower compared to those constructed by BE, for both split sizes. Nevertheless, stability of the trees constructed via BE(60,-) is even lower when $p_s > 15$ (Figure 4.23b).

Figure 4.24, shows the RMSE and stability values obtained from CV for trees constructed using $S_{max} = 60$, and full models are found to be more convenient with $RMSE < 14$ MPa and stability > 0.65 . FS(60,14) path is found to be on the left-top of BE(60,12) path for features above five, so may be preferable, contrary to that seen for the models constructed using $S_{max} = 30$, in which FS and BE paths almost coincide.

Performance of the models using $S_{max} = 60$ on the test set is tabulated in Table 4.4. As mentioned for Figure 4.24, FS trees are superior to BE and full trees in terms of *RMSE* and

medAE. Compared to result for $S_{max} = 30$ (Table 4.1), in which *RMSE* and *medAE* are ~ 14 MPa and ~ 6 MPa, respectively, the same statistics are < 12 MPa and ~ 5 MPa, respectively, for trees built using $S_{max} = 60$. . This shows that growing the tree increases the prediction accuracy significantly. Similar to that seen for $S_{max} = 30$, stability index of the full model is as much ~ 0.2 greater than those of the wrapper methods, indicating the superiority of the full model for explanatory purposes. On the other hand, maximum absolute error of the full model is still high.

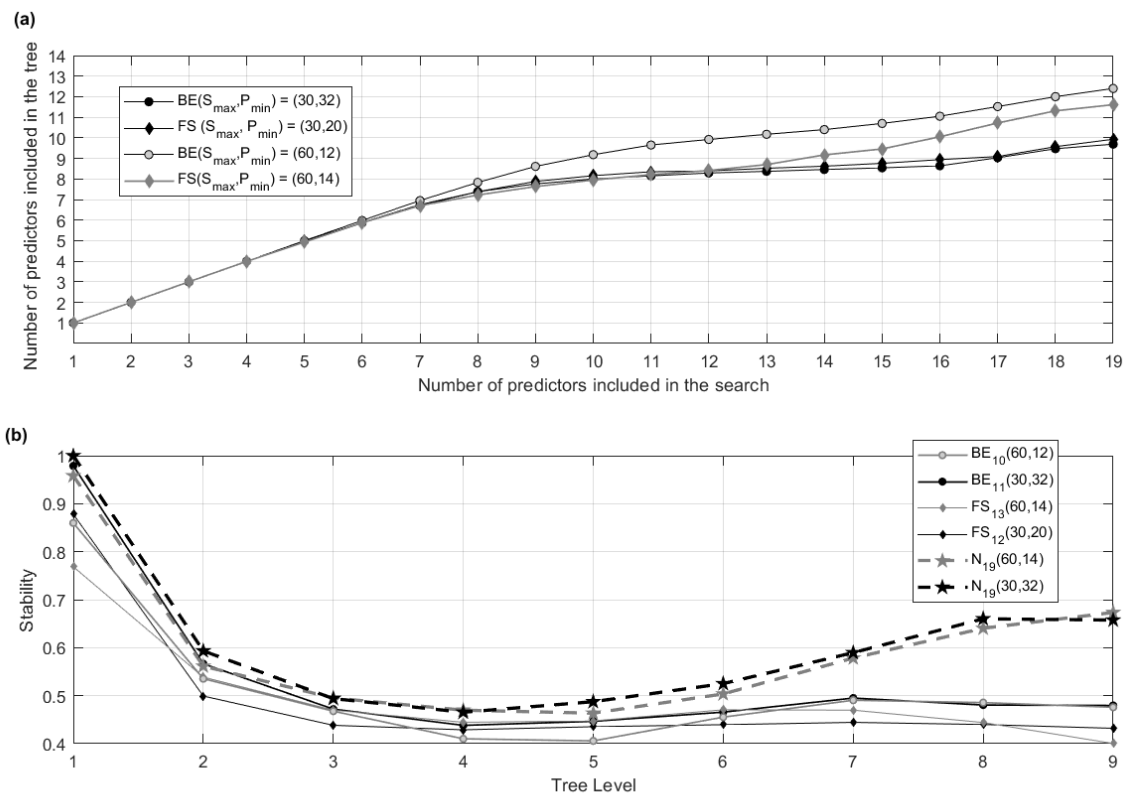


Figure 4.22 (a) Number of predictors included in the trees constructed during search versus number of features selected by final tree for FS and BE minimum *RMSE* decision trees for $S_{max} = 30$ and 60 and (b) Stability of the tree levels for FS, BE and full models with minimum *RMSE* for $S_{max} = 30$ and 60.

Predicted CS values are generally consistent with experimental CS for the test data (Figure 4.25). Examining the predictor importance in the tree constructed via FS₁₃(60, 14),

material and molar ratios, that also appear as the important predictors in the smaller optimum tree $BE_{11}(30,32)$, come out to be the most important features while there are also varying features, such as aggregate particle size and purity (Figure 4.26). When all the data set is used to construct the tree, predictor importance, hence, the potential decision rules significantly vary. For instance, Na/Al , which appears to be the second most important tree, is not included by the FS algorithm at all, once more indicating the sensitivity of the wrapper algorithm on the data set used (Figure 4.27).

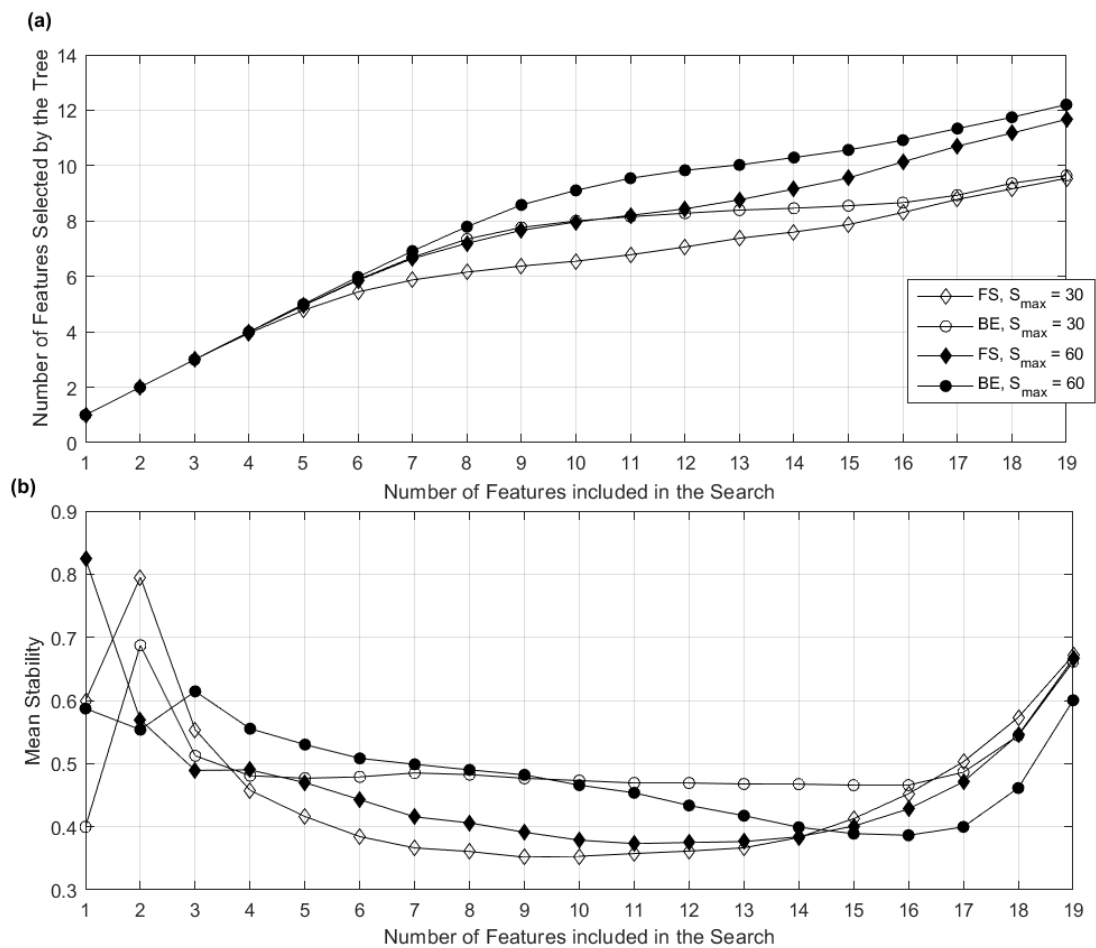


Figure 4.23. (a) Mean number of features selected by the tree versus number of features included in the search for $S_{max} = 30$ and (b) Mean tree feature stability for $S_{max} = 30$ and 60.

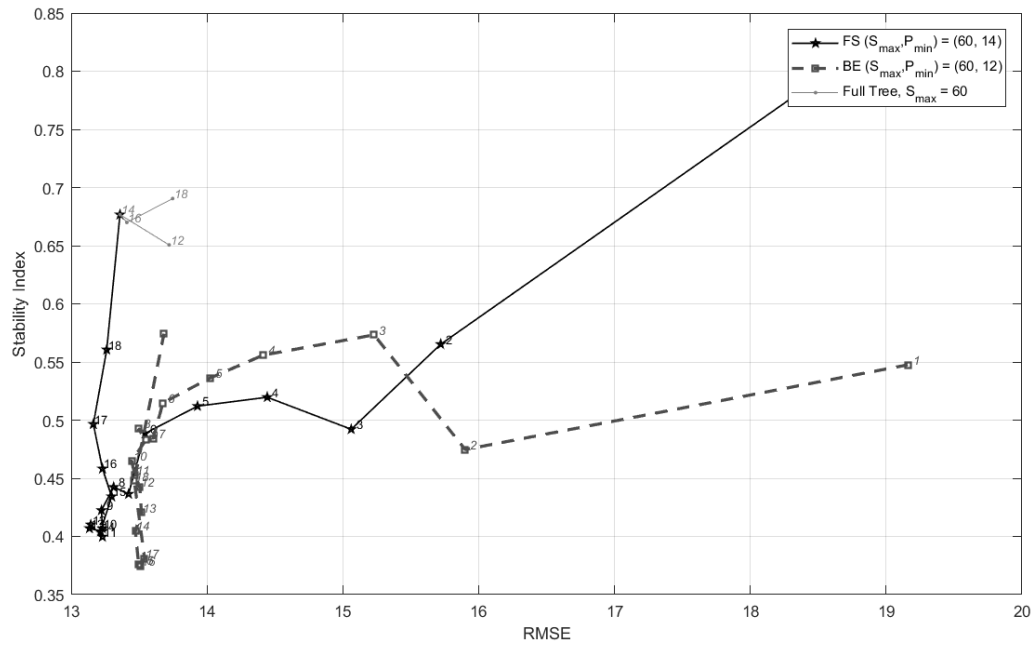


Figure 4.24. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 60$.

Table 4.4. Summary of CV and Test Results of Minimum $RMSE$ and $MedAE$ models for $S_{\max} = 60$.

	$S_{\max}$	60				
Parameters	SD	FS		BE		None
	$P_{\min}$	14	14	12	12	14
	p_s	13	16	10	17	19
CV results	$RMSE$	13.1	13.2	13.4	13.5	13.4
	MAE	8.31	8.32	8.39	8.43	8.46
	$medAE$	4.90	4.84	4.85	4.87	4.89
	SE	43.3	44.1	52.7	52.3	44.5
	SI_{Tree}	0.41	0.46	0.46	0.44	0.68
	$\bar{p}_T$	8.7	10.0	9.18	9.92	11.6
Test Results	p_T	7	9	9	11	12
	r_p	0.85	0.85	0.83	0.84	0.83
	R^2_p	0.71	0.71	0.65	0.67	0.65

Table 4.4. Summary of CV and Test Results of Minimum *RMSE* and *MedAE* models for $S_{\max} = 60$ (cont.).

	<i>RMSE</i>	11.9	11.8	13.1	12.7	13.0
	<i>MAE</i>	8.06	8.01	8.89	8.52	8.59
	<i>medAE</i>	5.00	5.00	5.44	5.27	5.22
	q.90	20.8	20.5	21.0	19.6	20.7
	<i>MGAE</i>	29.3	28.9	31.5	30.6	32.1
	<i>maxAE</i>	41.9	41.9	68.8	68.8	68.8

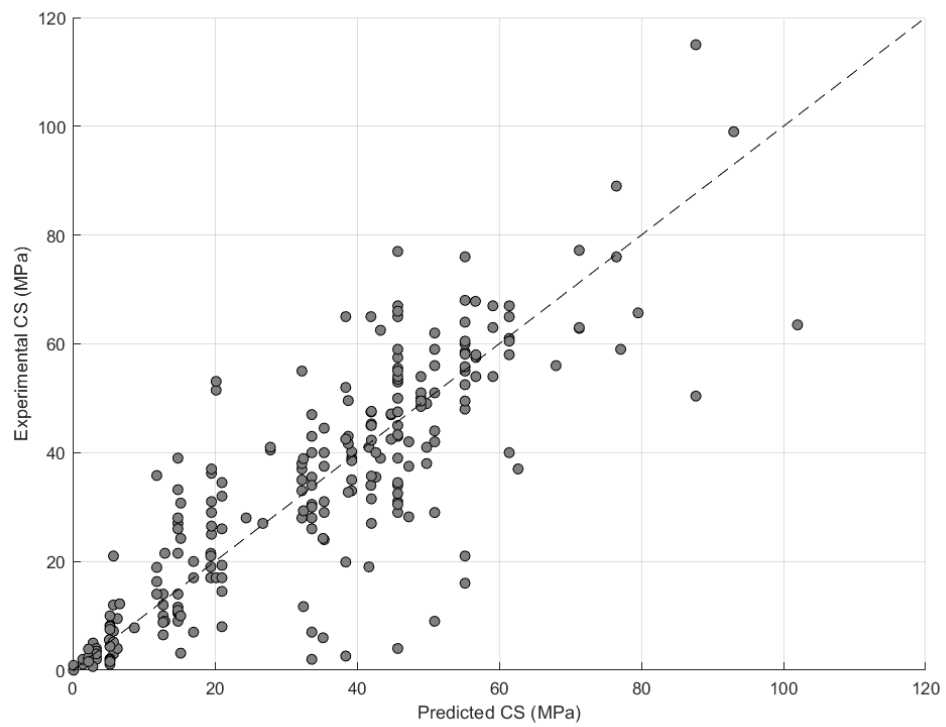


Figure 4.25. Comparison of predictions via $FS_{13}(60,14)$ model with the experimental results on the test set.

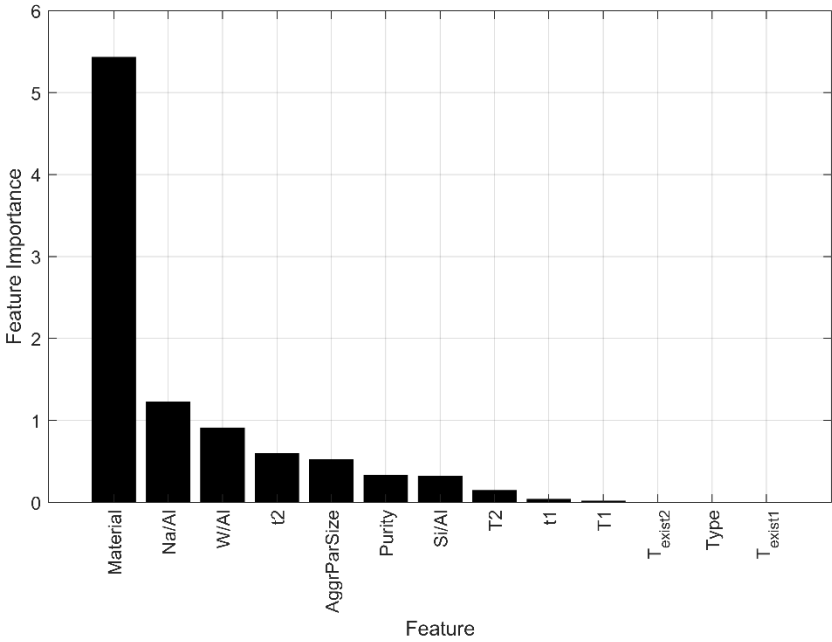


Figure 4.26. Feature importance in the tree constructed via $FS_{13}(60,14)$ using the training set.

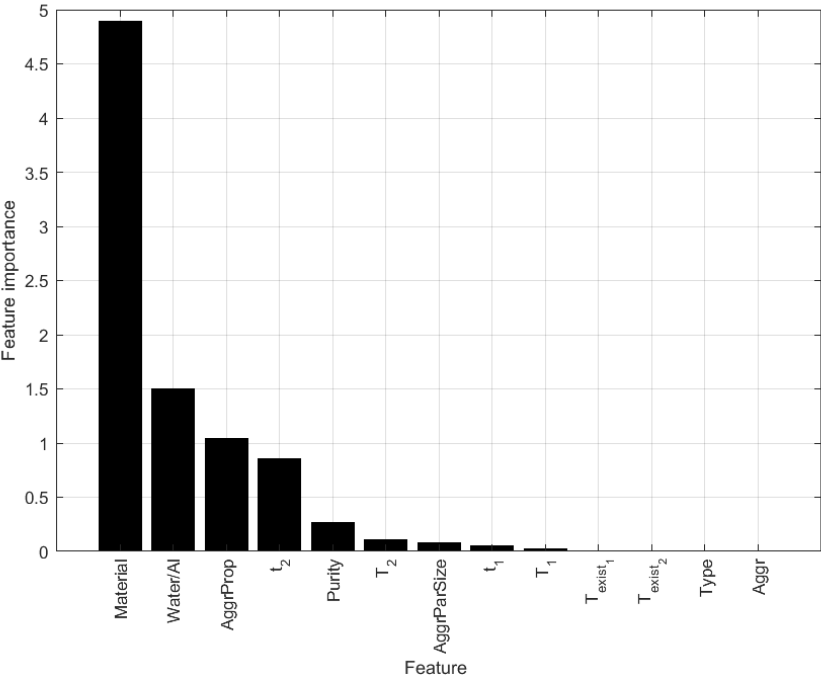


Figure 4.27. Feature importance in the tree constructed via $FS_{13}(60,14)$ using the whole dataset.

4.2.1. Construction Regression Trees via Reduct Sets with $S_{max} = 60$

CV results for decision trees with $S_{max} = 60$, constructed by optimum reduct feature subsets, are shown in Figure 4.28. It is observed that $R_{5e-5}(60,12)$ and $R_{1e-4}(60,12)$ yield the smallest RMSE, while $R_{1e-4}(60,10)$ yields the smallest *medAE* value (Figure 4.28b). Stability indices do not seem to depend on P_{min} .

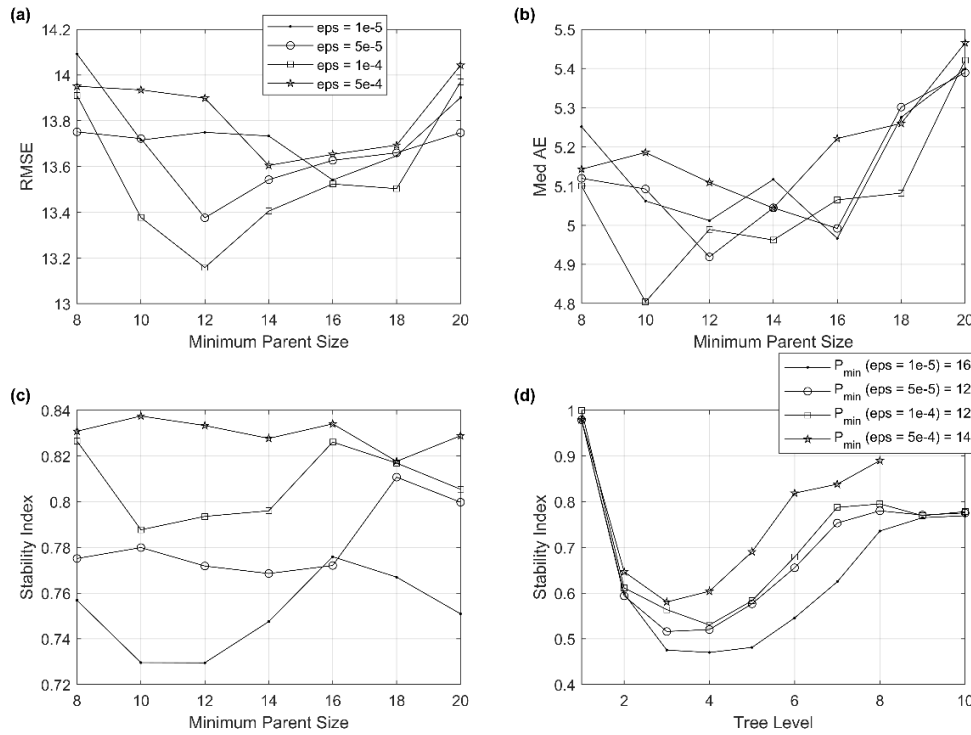


Figure 4.28. (a) *RMSE*, (b) *MedAE* and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $eps = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{max} = 60$.

It has been noted that full trees were more advantageous compared to those formed via FS and BE with respect to stability indices. Here, it is seen that all rough set based trees yield similar *RMSE* values, and are much more stable compared to full trees (Figure 4.29). As an example, model constructed via $R_{1e-4}(60, 12)$ has a stability of 0.78 and *RMSE* equal to 13.5 MPa.

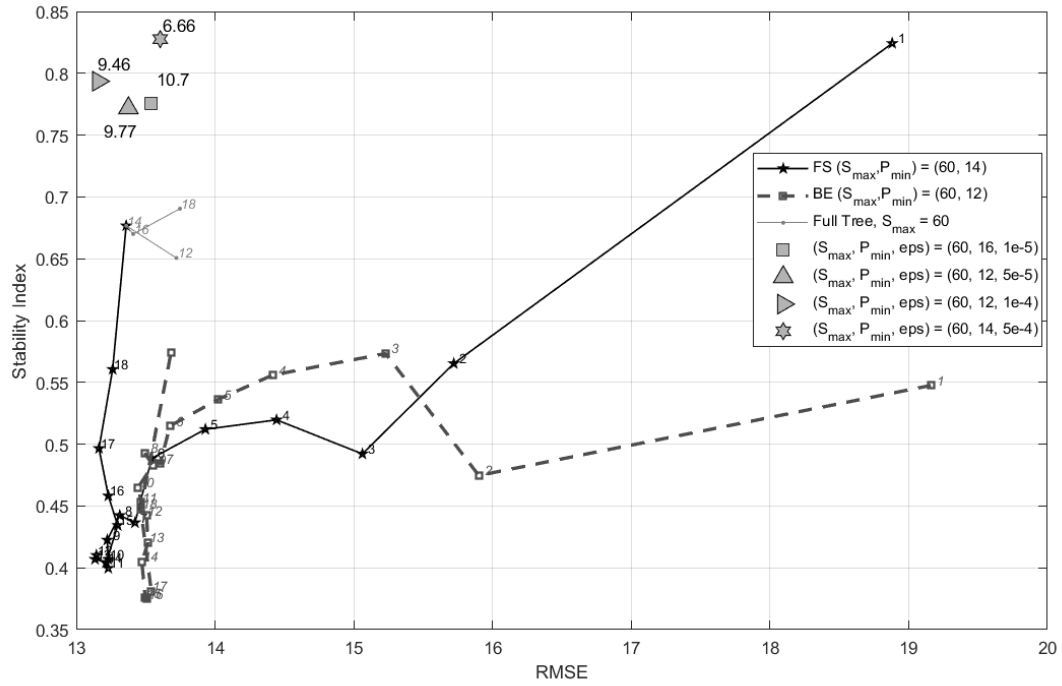


Figure 4.29. FS, BE and Reduct set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 60$.

CV and test results for best trees constructed based on reduct sets for $S_{\max} = 60$ are shown in Table 4.5. Even though tree constructed via $R_{1e-4}(60,12)$ gives the smallest *RMSE* value, model constructed via $R_{5e-5}(60,12)$ yields the smallest *medAE* value. is the largest number of features selected by the rough set methods is equal to 15.4, while 10.7 features are utilised on average in tree construction. In the full tree, however, 11.6 features out of for which 19 features are used on average. Hence, it may be stated that the source of instability for the trees constructed via wrapper methods is not due to inadequate number of features being provided by the feature selection algorithm, but uniqueness of the features selected for each specific fold. Reduct sets, on the other hand, seem to lead to more generalized feature groups applicable to a great majority of the folds. When test results are examined, $R_{1e-5}(60,16)$ is characterized with lower error measures (except for *medAE* which is lower for $R_{1e-4}(60,12)$) and with high coefficient for determination and correlation between predicted and experimental CS. The maximum absolute error is ~ 69 MPa, showing that the constructed model is unable to approximate the functional relation between CS and predictors over the whole predictor space. On the other hand, a R^2_p value equal to 0.64-0.67 and *medAE* equal

to ~5 MPa for the test set may be taken as a satisfactory result for an explanatory model of low-intermediate complexity, i.e. $S_{\max} = 60$.

Table 4.5. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 60$.

Parameters	$S_{\max}$	60				
	$P_{\min}$	16	12	12	14	14
	$\epsilon \times 10^{-5}/SA$	1	5	10	50	None
CV results	$RMSE$	13.5	13.4	13.2	13.6	13.4
	MAE	8.69	8.46	8.39	8.74	8.46
	$medAE$	4.97	4.92	4.99	5.04	4.89
	SE	45.1	48.0	46.5	45.5	44.5
	SI_{Tree}	0.78	0.77	0.79	0.83	0.68
	$\bar{p}_A$	15.4	12.6	11.9	7.17	19.0
	$\bar{p}_T$	10.7	9.77	9.46	6.66	11.6
Test Results	p_A	16	13	12	6	19
	p_T	11	11	9	6	12
	r_p	0.83	0.84	0.82	0.80	0.83
	R^2_p	0.65	0.67	0.64	0.61	0.65
	$RMSE$	13.1	12.7	13.2	13.7	13.0
	MAE	8.76	8.52	8.76	9.50	8.59
	$medAE$	5.38	5.27	4.82	6.26	5.22
	$q_{.90}$	21.8	19.6	22.1	25.0	20.7
	$MGAE$	32.0	30.6	32.0	32.4	32.1
	$maxAE$	68.8	68.8	68.8	68.8	68.8

The test performance of model constructed via R_{1e-4} (60, 12) using the training is shown in Figure 4.29, respectively. and a reasonable correlation between the predicted and experimental data can be observed. Predictor importance is shown in Figure 4.30, material is the most important feature (as for the wrapper cases) with aggregate proportion, which also appeared for the models constructed via $BE_{11}(30,32)$, as the second most important predictor. Furthermore, molar ratios and ambient curing time are also significant in

determining CS values. When the same tree, i.e. $R_{5e-4}(30,24)$, is constructed using the whole data set seven of the eight visible predictor contributions in both training sets overall (Figure 4.32), indicating the robustness of the rules derived by reduct set trees.

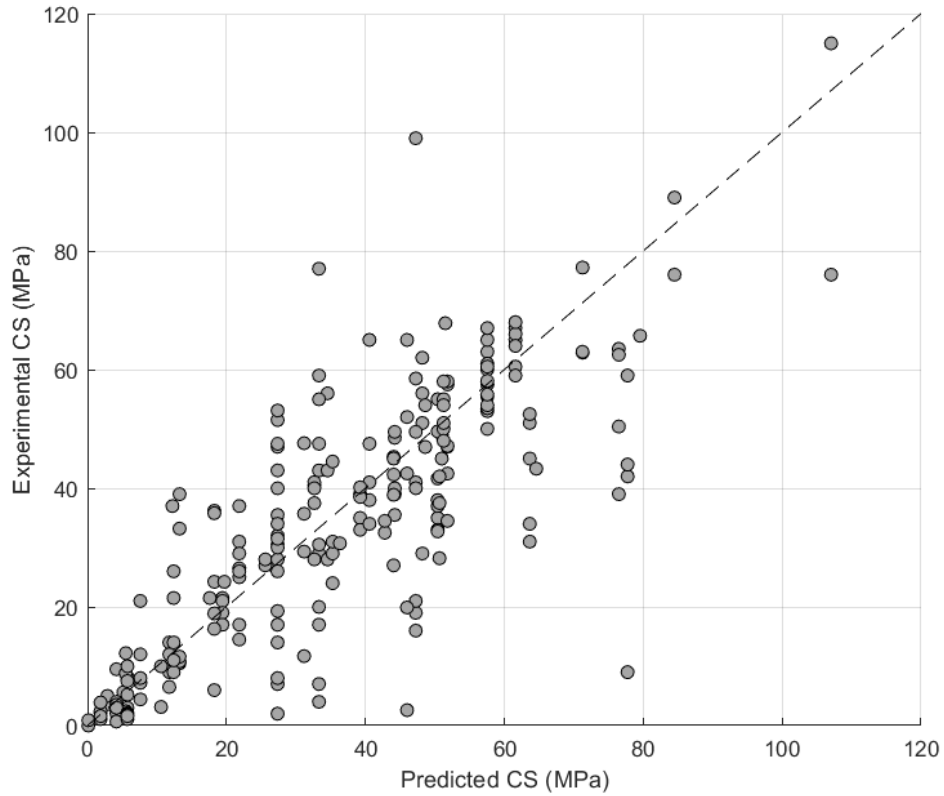


Figure 4.30. Comparison of predictions via $R_{1e-4}(60, 12)$ model with the experimental results on the test set.

Finally, the best reduct set, wrapper and full trees are in Table 4.6. Except for *medAE* which is the lowest for $R_{1e-4}(60,12)$, the test performance of $FS_{13}(60,14)$ is superior to those of other methods. It should also be noted that *maxAE* is lowered when FS tree is used. Nevertheless, the low stability of this model makes it susceptible for rule extraction.

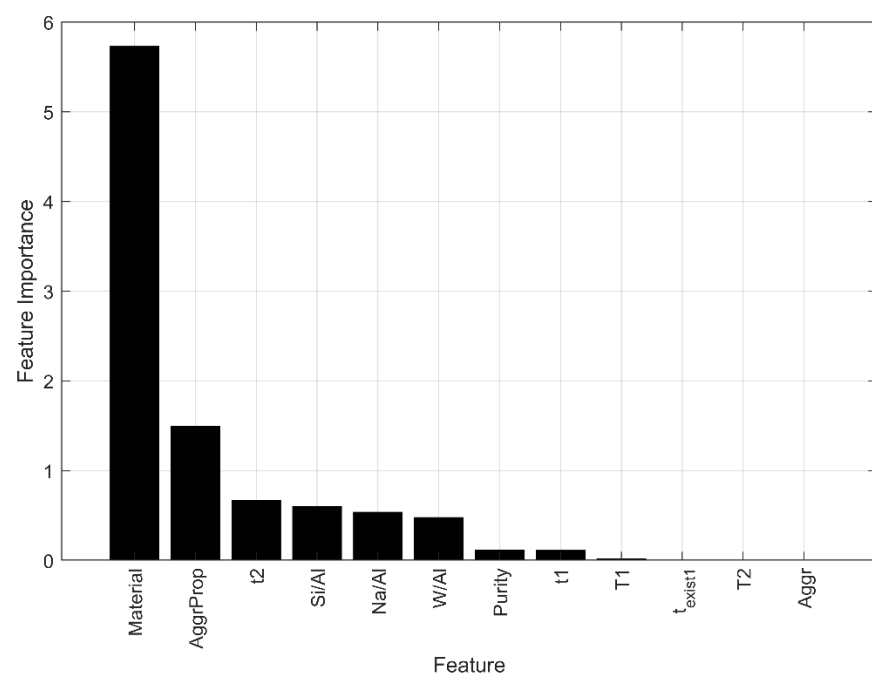


Figure 4.31. Feature importance in the tree constructed via $R_{1e-4}(60,12)$ using the training set.

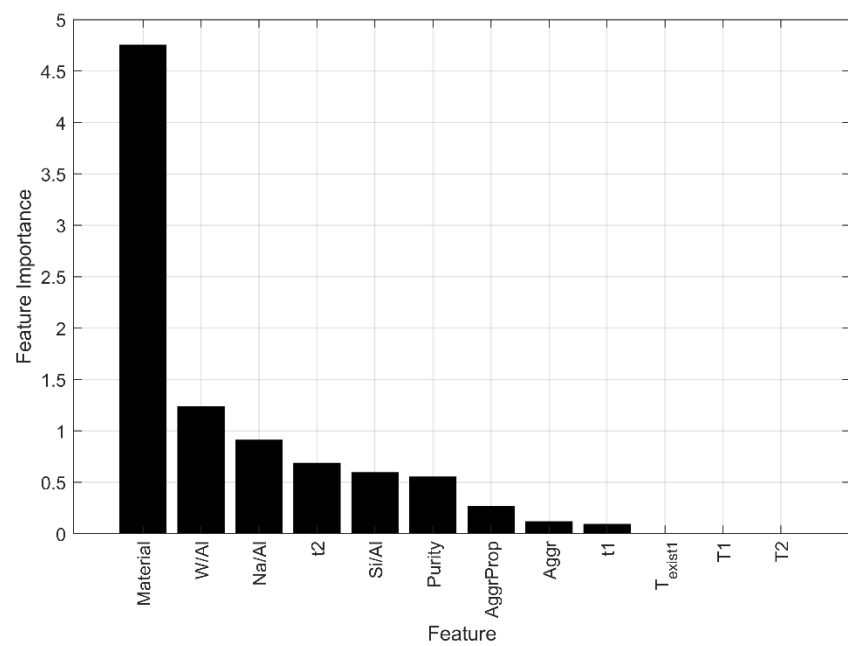


Figure 4.32. Feature importance in the tree constructed via $R_{1e-4}(60,12)$ using the whole dataset.

Table 4.6. CV and Test Result Comparison between most stable Reduct based decision tree $R_{1e-5}(60,12)$ and minimum $RMSE$ FS tree $FS_{13}(60,14)$ for $S_{max} = 60$.

Parameters	S_{max}	60		
	P_{min}	12	14	14
	$\epsilon \times 10^{-5}/SA$	10	FS	None
CV results	$RMSE$	13.2	13.1	13.4
	MAE	8.39	8.31	8.46
	$medAE$	4.99	4.90	4.89
	SE	46.5	43.3	44.5
	SI_{Tree}	0.79	0.41	0.68
	$\bar{p}_A$	11.9	13.0	19.0
	$\bar{p}_T$	9.46	8.70	11.6
Test Results	p_A	12	13	19
	p_T	9	7	12
	r_p	0.82	0.85	0.83
	R^2_p	0.64	0.71	0.65
	$RMSE$	13.2	11.9	13.0
	MAE	8.76	8.06	8.59
	$medAE$	4.82	5.00	5.22
	$q_{.90}$	22.1	20.8	20.7
	$MGAE$	32.0	29.3	32.1
	$maxAE$	68.8	41.9	68.8

4.3. Regression Trees with $S_{max} = 150$

4.3.1. Constructing Regression Trees via FS and BE with $S_{max} = 150$

As a result of tree depth increasing up to 150 splits, parent sizes much smaller than those examined for 30 and 60 splits are examined to determine the minimum prediction error. Hence, the minimum $RMSE$ by FS method is determined at $P_{min} = 4$ (Figure 4.33) .

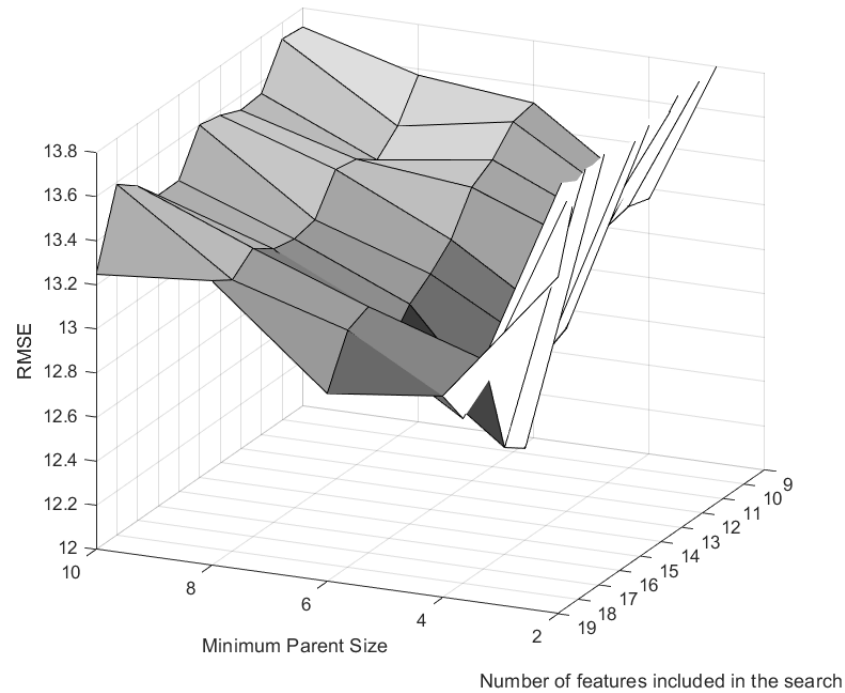


Figure 4.33. Surface plot of $RMSE$ obtained via CV with respect to minimum parent size parameter and number of features selected by FS for trees with maximum number of splits equal to 150.

Similar to those seen for $S_{max} = 30$ and 60, $RMSE$ and $medAE$ first decreases and then increases with the addition of new features. The region with the smallest prediction errors again coincide with that having lowest stability (Figure 4.34. a,b and c). In full models, it is possible to decrease $RMSE$ below 13 MPa and $medAE$ below 4 MPa using small P_{min} values. However, stability of the full trees barely exceeds 0.6 (Figures 4.34d,e and f).

Number of features included in the constructing tree is significantly inflated as tree size is increased: the first 10 features selected by FS are directly inserted in construction of the tree, while approximately half of the remaining features are used in the tree (Figure 4.35)

Selecting features via BE for $S_{max} = 150$ yields the minimum $RMSE$ at $P_{min} = 4$ are (Figure 4.36): Minimum $RMSE$ and $medAE$ values are found to be < 13 MPa and < 4 MPa, respectively. Unlike the results from models constructed via FS, the range of number of

features corresponding minimum $RMSE$ values do not coincide with those with minimum stability: model stability increases until the ninth feature, but stability decreases for larger number of features (Figures 4.37 a, b and c).

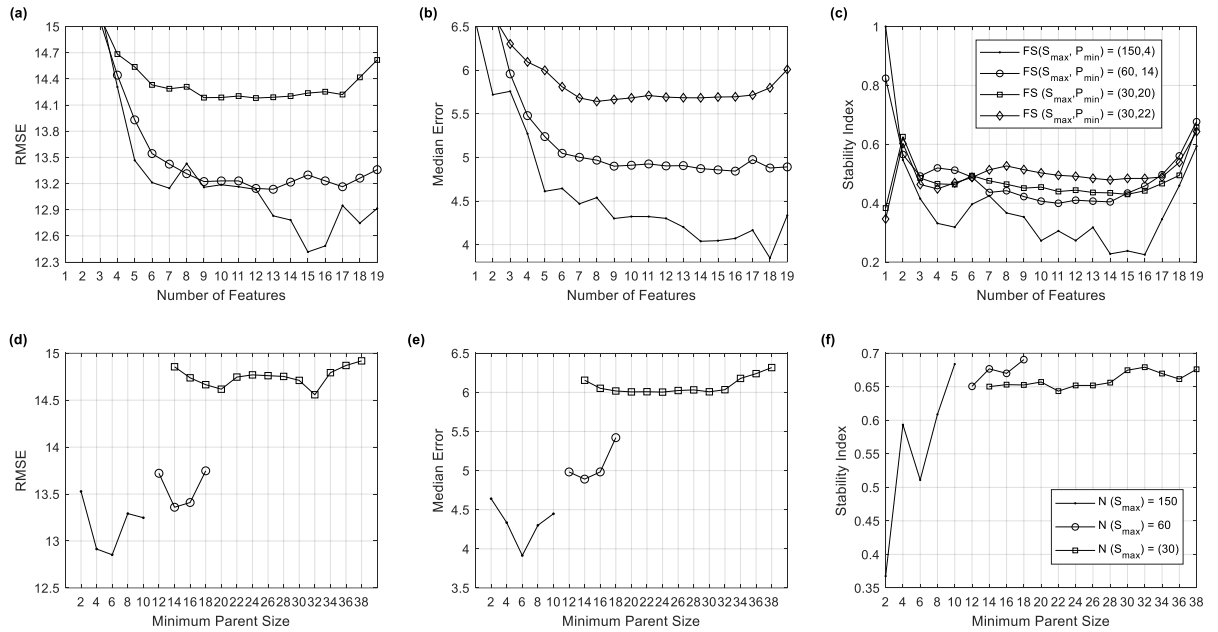


Figure 4.34. (a) $RMSE$, (b) MAE, and (c) stability index with respect to features selected via FS; (d) $RMSE$, (e) MAE, and (f) stability index with respect to minimum parent size in full trees constructed for $S_{max} = 30, 60$ and 150 .

In trees constructed via BE(150,4), most of the selected features via BE are included in the tree (Figure 4.38a), possibly explaining the extremely low stability of the models obtained for 9 to 18 features via BE (Figure 4.37c). It is also seen that more features are used in trees constructed via BE, compared to those constructed via BE. Predictors in each levels of tree is also likely to change for $S_{max} = 150$ (Figure 4.38b).

Comparing the number of predictors included in the tree as a result of search via FS and BE, it is seen that more predictors are included in the trees constructed using BE than those constructed via FS for the same number of features selected by the search method

(Figure 4.39a). Generally, the constructed tree is “saturated” with features after the eight and 10 feature for $S_{max} = 30$ and 60, respectively, but may still include more predictors for $S_{max} = 150$. Stability values of the models constructed via FS is smaller for number of features $< 13-14$, but higher for number of features < 3 and between 15-18 (Figure 4.39b).

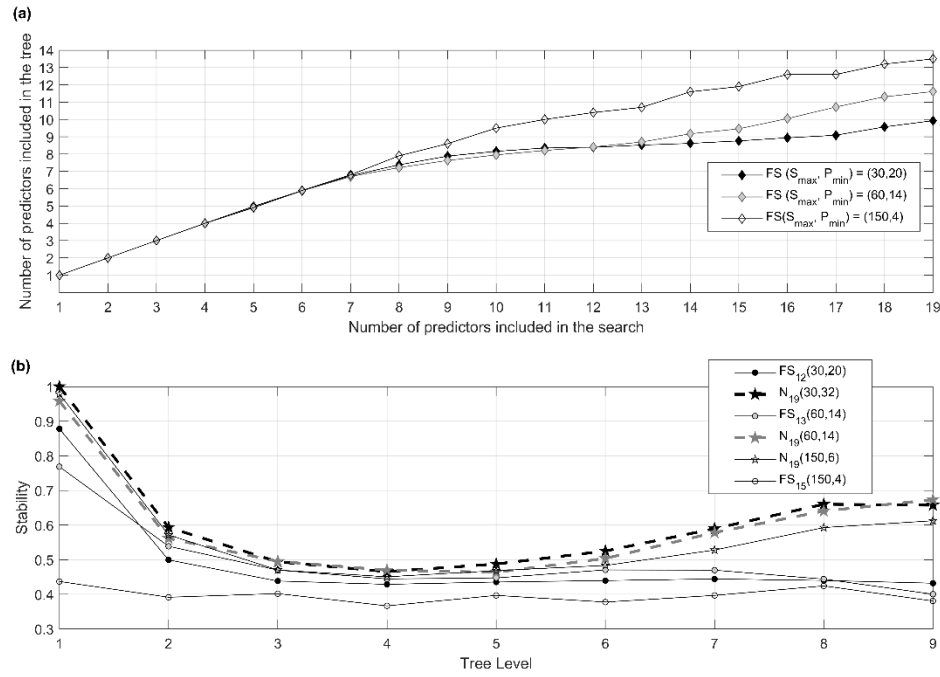


Figure 4.35. (a) Number of predictors included in the trees constructed during search versus number of features selected by final tree for FS(30,20), FS(60,14) and FS(150,4) and (b) Tree level stability for FS₁₂(30,20), FS₁₃(60,14), FS₁₅(150,4) full models N₁₉(30,32), N₁₉(60,14) and N₁₉(150,6).

Contrary to that seen for smaller trees, stability of the full trees N₁₉(150,10) is relatively high at ~ 0.7 with low CV errors (Figure 4.40).

CV and test results for $S_{max} = 150$ show that almost prediction accuracy of the trees are constructed via FS and BE methods exceed that of the full tree (Table 4.7). There is good correlation between predicted and experimental data, with the best performance by tree constructed via BE₉(60,6) yielding a R^2_p equal to 0.79, and $maxAE$ dropped to 46.9 MPa.

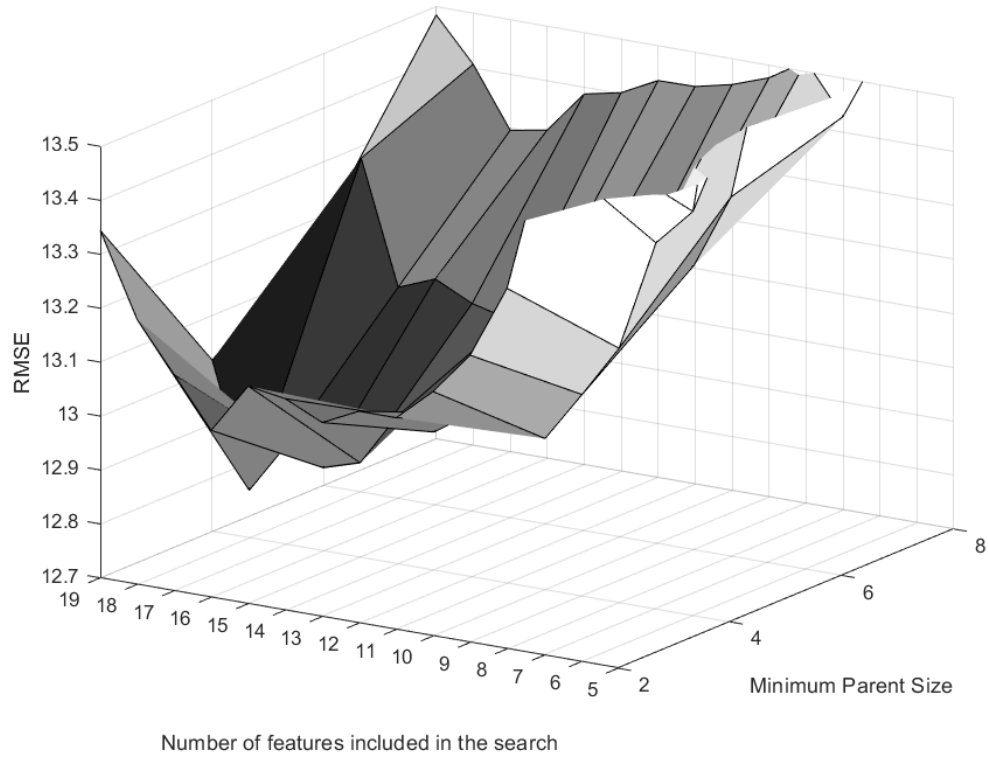


Figure 4.36. CV $RMSE$ 3D Surface Plot of Models with various P_{min} and Number of Features Selected by BE for $S_{max} = 150$.

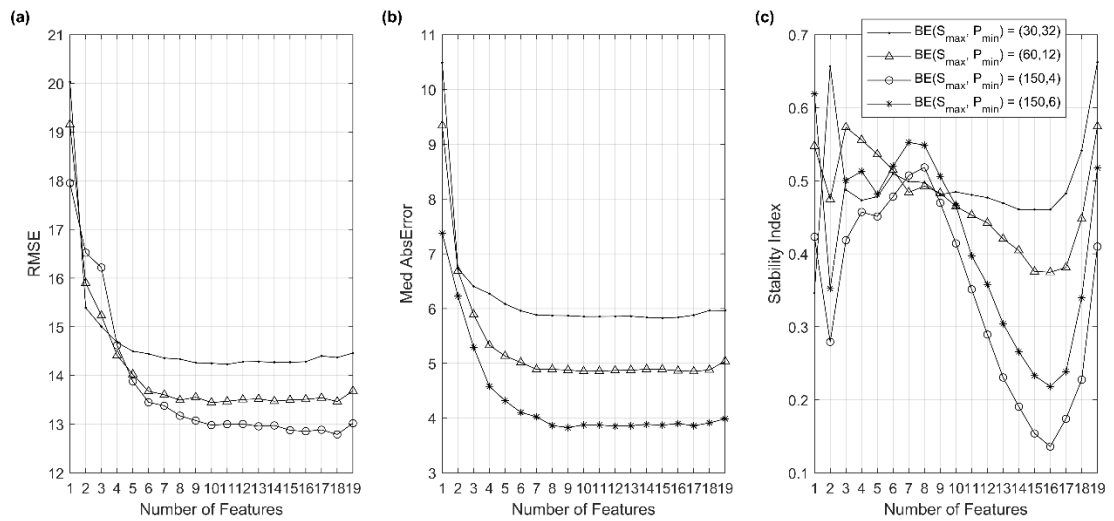


Figure 4.37. (a) $RMSE$, (b) MAE , and (c) stability index with respect to features selected via BE for $S_{max} = 30, 60$ and 150 .

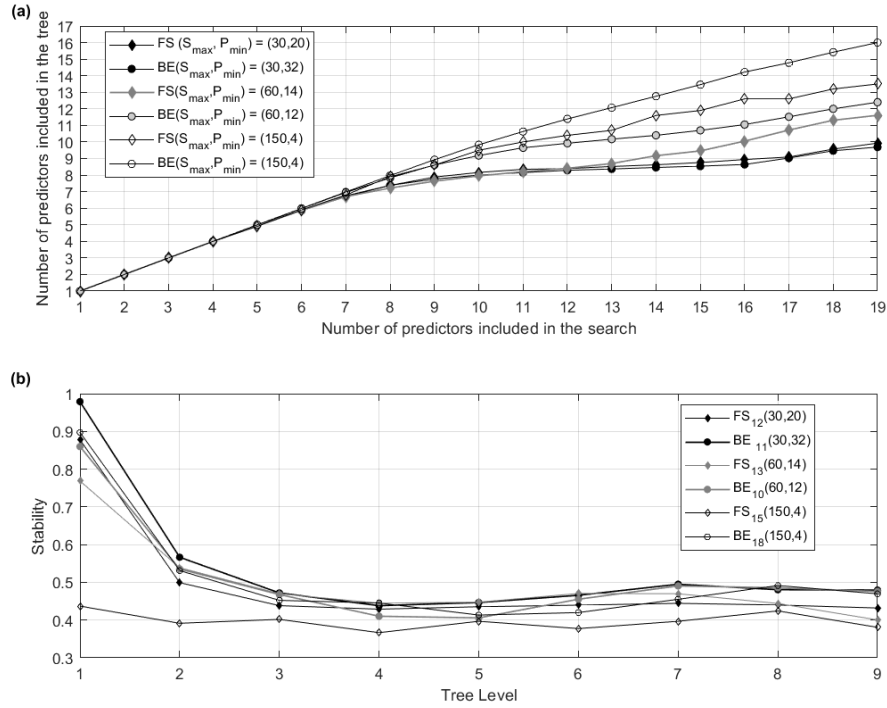


Figure 4.38 (a) Number of predictors in the trees versus number of features selected by FS/BE search for $S_{max} = 30, 60$ and 150 and (b) Tree level stability for $S_{max} = 30, 60$ and 150 .

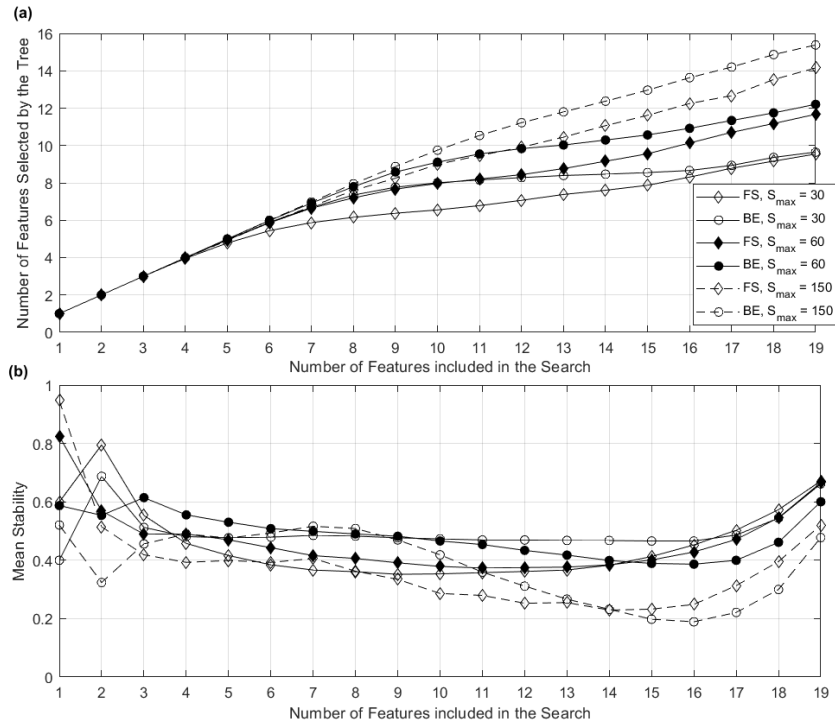


Figure 4.39. (a) Mean number of features selected by the tree versus number of features included in the search for and (b) Mean tree feature stability for $S_{max} = 30, 60$ and 150 .

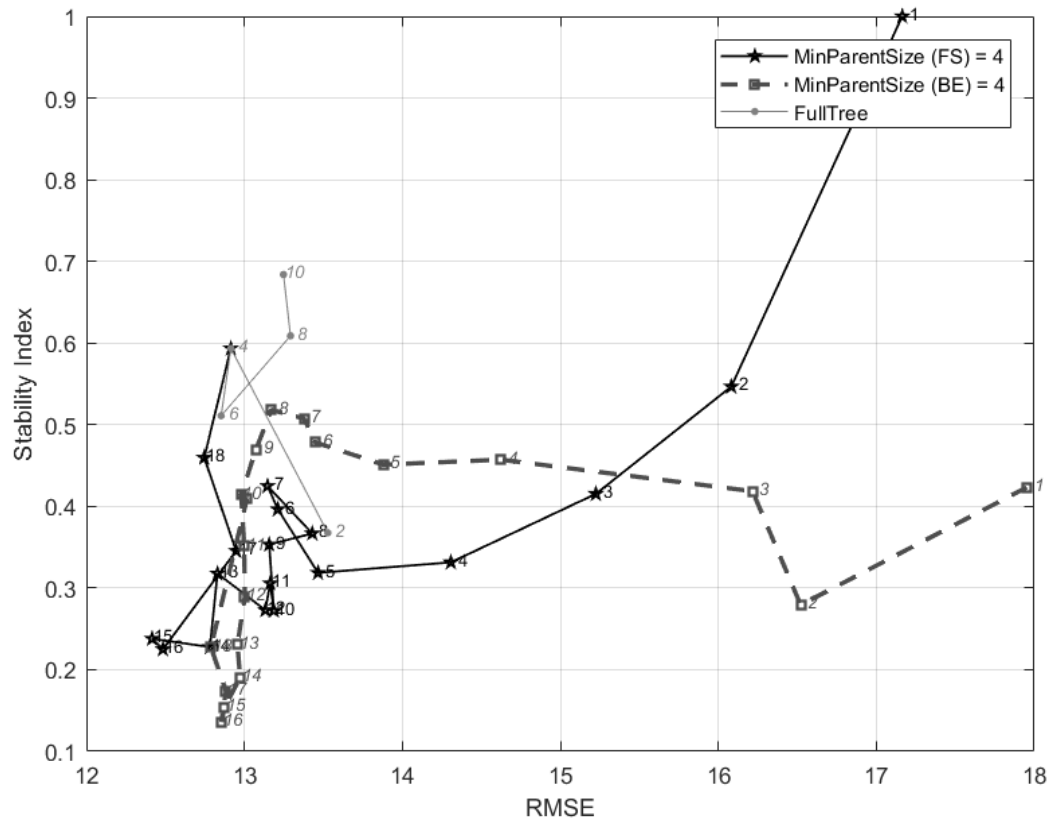


Figure 4.40. FS and BE trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 150$.

Table 4.7. Summary of CV and Test Results of Minimum *RMSE* and *MedAE* models ($S_{\max} = 150$)

	$S_{\max}$	150				
Parameters	SD	FS		BE		None
	$P_{\min}$	4	4	4	6	6
	p_s	15	18	18	9	19
CV results	<i>RMSE</i>	12.4	12.7	12.8	13.2	12.9
	<i>MAE</i>	7.57	7.51	7.57	7.69	7.68
	<i>medAE</i>	4.05	3.84	3.90	3.83	3.91
	SE	38.2	79.4	49.4	59.2	50.2
	SI_{Tree}	0.24	0.46	0.23	0.51	0.51
	$\bar{p}_T$	11.9	13.2	15.4	8.94	13.5

Table 4.7. Summary of CV and Test Results of Minimum *RMSE* and *MedAE* models for $S_{\max} = 150$ (cont.).

Test Results	p_T	13	13	17	9	14
	r_p	0.88	0.85	0.88	0.90	0.87
	R^2_p	0.75	0.71	0.76	0.79	0.72
	<i>RMSE</i>	11.0	11.8	10.8	10.0	11.6
	<i>MAE</i>	6.96	7.90	6.93	6.45	7.34
	<i>medAE</i>	4.44	5.29	3.98	3.54	3.9
	q _{.90}	16.5	18.3	17.7	17.5	19.7
	<i>MGAE</i>	27.9	29	27.6	25.2	28.6
	<i>maxAE</i>	53.9	53.9	53.9	46.9	68.8

When the test set is predicted via $FS_{15}(150,4)$, the experimental CS results to be in good accordance with predicted CS values (Figure 4.41). In this model, the most important predictor is found to be W/AI (Figure 4.42). Nevertheless, when a tree with identical parameters is constructed using the whole data set, an almost totally different feature importance profile is obtained with even the most important feature to be different than previously (Material versus W/AI). (Figure 4.43).

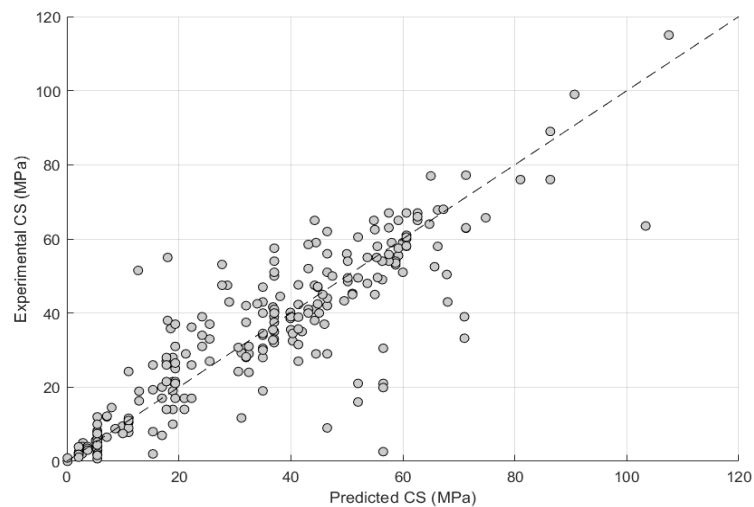


Figure 4.41. Comparison of predictions via $FS_{15}(150,4)$ model with the experimental results on the test set.

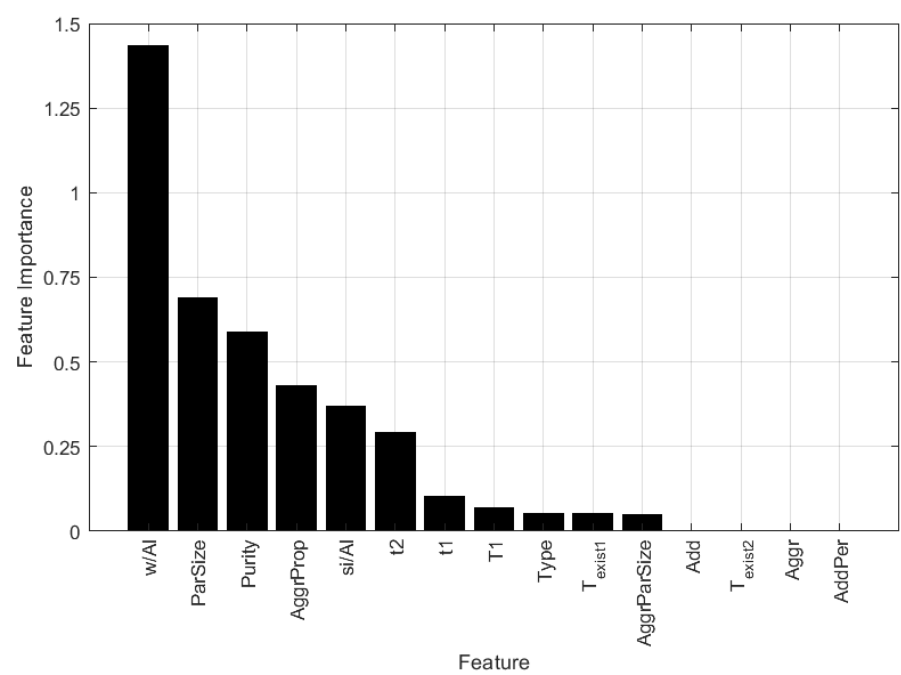


Figure 4.42. Feature importance in the tree constructed via $FS_{15}(150,4)$ using the training data set.

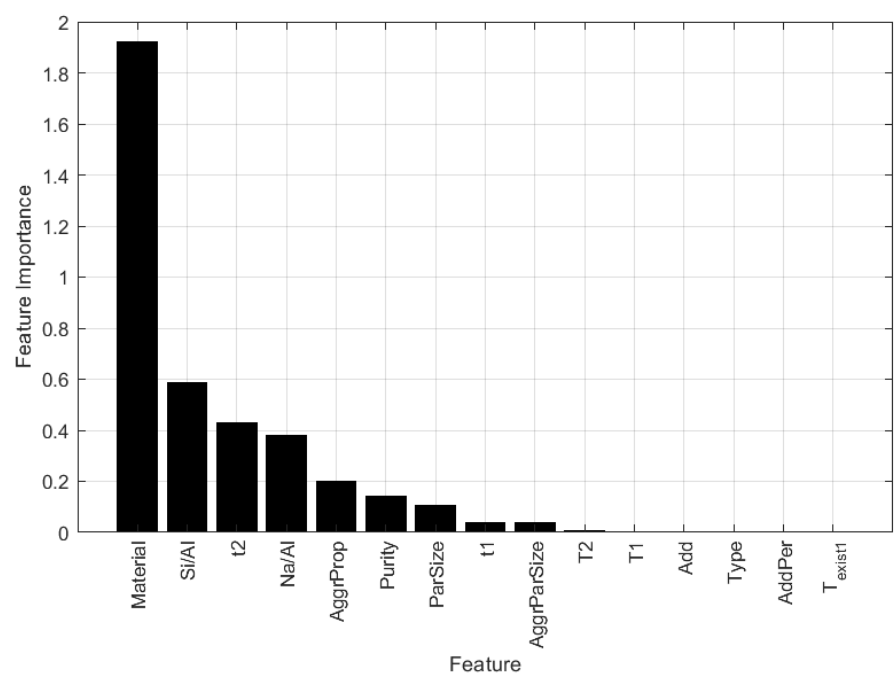


Figure 4.43. Feature importance in the tree constructed via $FS_{15}(150,4)$ using the whole dataset.

4.3.2. Constructing Regression Trees via Reduct Sets with $S_{\max} = 150$

It is seen that tree constructed via $R_{5e-5}(150,8)$ leads to minimum $RMSE$ and minimum $medAE$ (Figures 4.44a,b). Except the models constructed using $\varepsilon = 1.0 \times 10^{-5}$, stability values of the trees are at ~ 0.80 for varying $P_{\min}$ values, while stability values of tree levels are lower than those constructed for $S_{\max} = 30$ and 60. (Figures 4.44c and d). As seen in Figure 4.45, stability of the trees constructed with reduct sets exceeds those constructed via FS and BE methods.

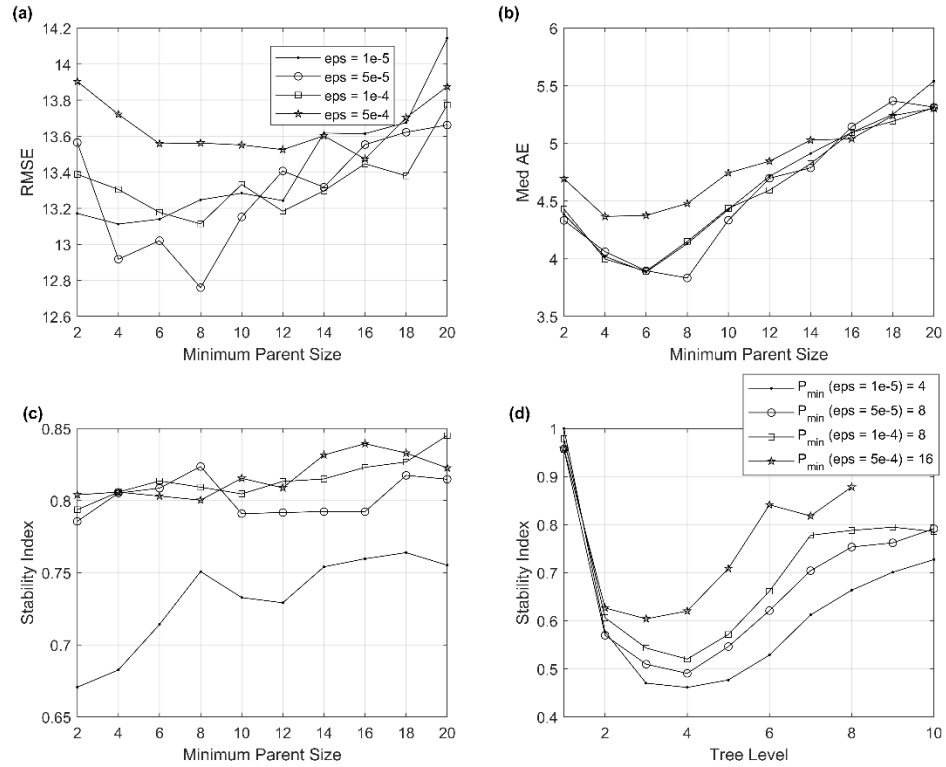


Figure 4.44. (a) $RMSE$, (b) $MedAE$ and (c) stability index with respect to minimum parent size of the trees, and (d) stability index versus vs tree levels for $\varepsilon = 1.0 \times 10^{-5}$, 5.0×10^{-5} , 1.0×10^{-4} and 1.0×10^{-4} and $S_{\max} = 150$.

The minimum $RMSE$ obtained by trees through reduct sets on test set is equal to 11.3 MPa, slightly more accurate than that of the full tree (Table 4.8). The stability of the same tree, $R_{5e-5}(150,16)$, is again marginally better than that of the full tree. The number of features

selected by the trees show variability in a range of 6 to 16 features, with higher stability but lower prediction accuracy than those obtained from wrapper methods.

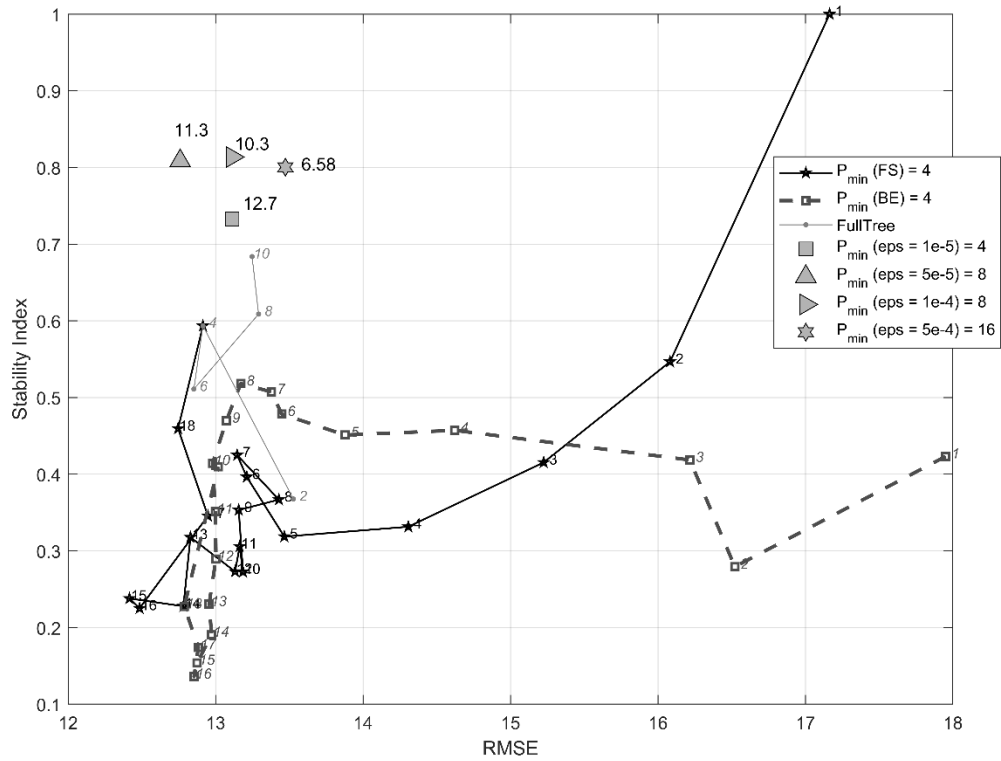


Figure 4.45. FS, BE and rough set trajectories on 2-D stability vs. RMSE plane determined using CV for $S_{\max} = 150$.

Table 4.8. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 150$.

Parameters	$S_{\max}$	150				
	$P_{\min}$	4	8	8	16	6
	$\epsilon \times 10^{-5}/SA$	1	5	10	50	None
CV results	RMSE	13.1	12.8	13.1	13.5	12.9
	MAE	7.78	7.56	7.82	8.72	7.68
	medAE	4.02	3.83	4.15	5.04	3.91
	SE	53.7	55.5	54.8	42.1	50.2
	SI_{Tree}	0.68	0.82	0.81	0.84	0.51
	$\bar{p}_A$	15.4	12.6	11.8	7.08	19.0

Table 4.8. CV and Test Results of Rough Set theory-based decision trees for $S_{\max} = 150$
(cont.).

	$\bar{p}_T$	12.7	11.3	10.3	6.58	13.5
Test Results	p_A	16	13	12	6	19
	p_T	16	12	11	6	14
	r_p	0.87	0.86	0.85	0.82	0.87
	R_p^2	0.74	0.69	0.68	0.64	0.72
	$RMSE$	11.3	12.3	12.5	13.3	11.6
	MAE	7.34	7.47	7.52	9.10	7.34
	$medAE$	4.16	3.48	3.58	6.02	3.90
	$q_{.90}$	19.6	19.7	20.4	21.7	19.7
	$MGAE$	28.1	31.3	32.2	32.0	28.6
	$maxAE$	49.1	68.8	68.8	68.8	68.8

The correlation between experimental and predicted data using $R_{5e-5}(150,8)$ tree is seen in Figure 4.46. Similar to most of the trees constructed before, the most important predictor found to be Material, followed by Aggregate Proportion, while the importance of the remaining predictors is close to each other (Figure 4.47). When the regression tree is constructed using the whole dataset, it is observed that most of the predictors have similar importance, showing the stability of the trees constructed using the rough set theory (Figure 4.48).

$RMSE$, MAE and $medAE$ values obtained from CV for FS, reduct set and full tree models are close to each other, but SE of $FS_{15}(150,4)$ model is significantly lower and stability of $R_{5e-5}(150,8)$ model is remarkably higher (Table 4.9). When test errors are considered, $FS_{15}(150,4)$ model has the highest correlation between experimental and predicted data, but its low stability index makes the rules derived from its decision tree unreliable. Full tree has a higher $RMSE$, but lower $medAE$ for test points; $R_{5e-5}(150,8)$ model has the largest $RMSE$, but lowest $medAE$ for test points. In general, one may see an inverse relation between stability and $RMSE$ values, and a rather interesting direct relation between stability and $medAE$ values for test points. Further research is required to determine whether the second suggested relation is coincidental or has a statistical basis.

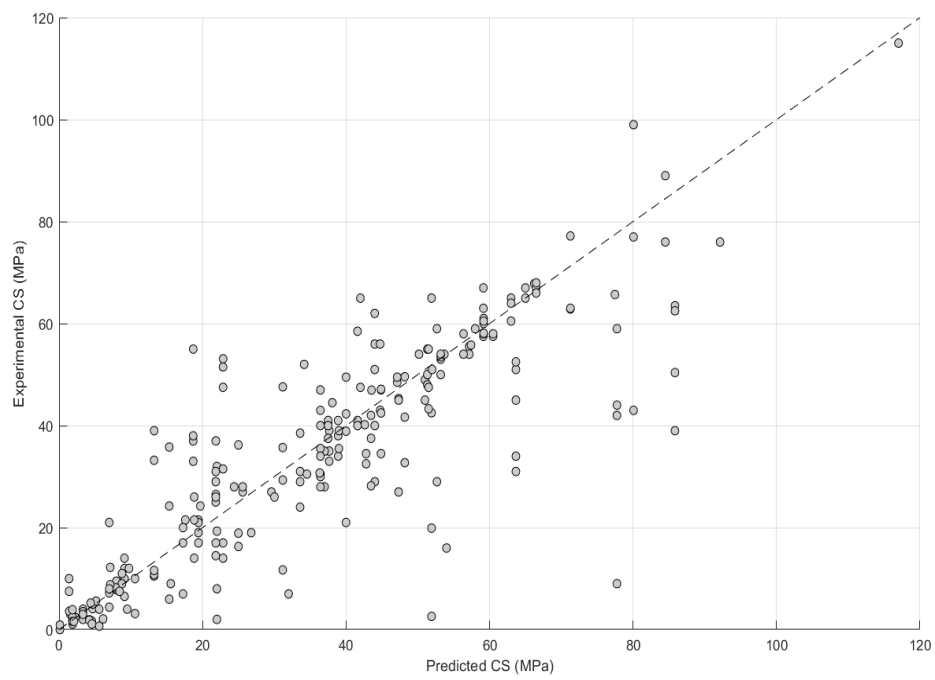


Figure 4.46. Comparison of predictions via $R_{5e-5}(150,8)$ model with the experimental results on the test set.

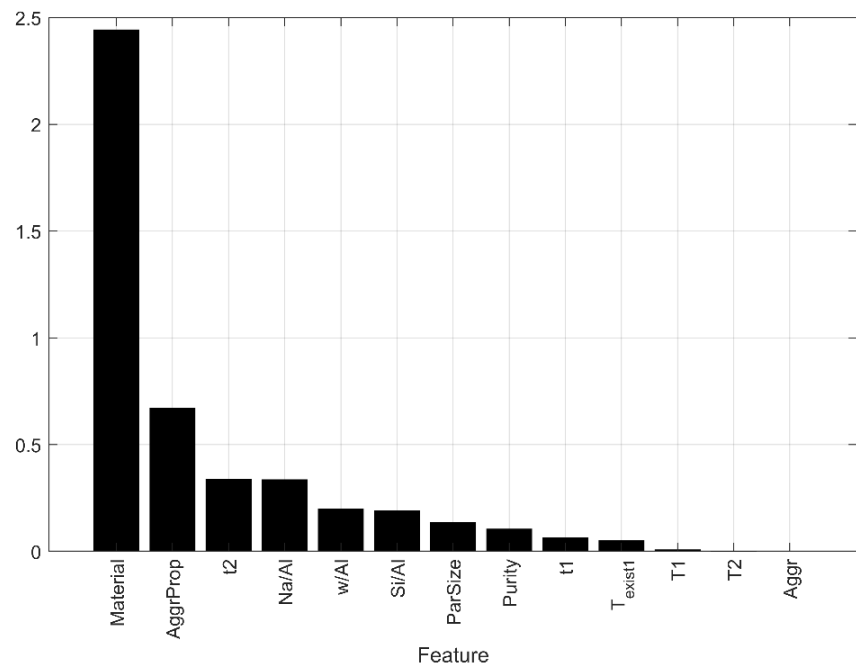


Figure 4.47. Feature importance in the tree constructed via $R_{5e-5}(150,8)$ using the training set.

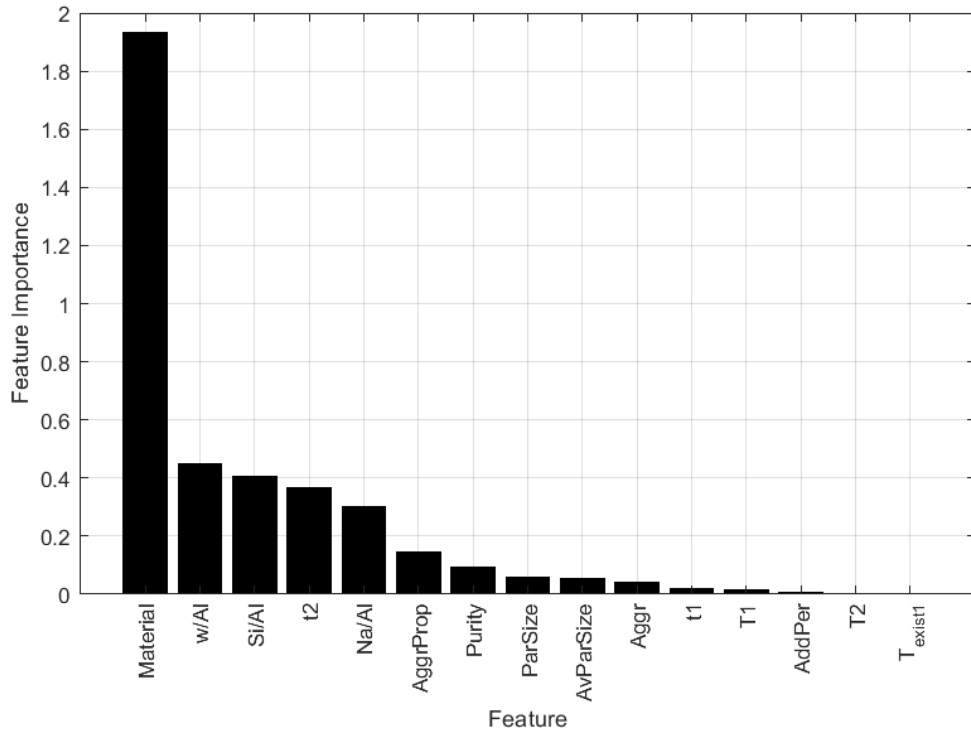


Figure 4.48. Feature importance in the tree constructed via Feature importance in the tree constructed via $R_{5e-5}(150,8)$ using the whole dataset.

Table 4.9. CV and Test Result Comparison between minimum $RMSE$ Reduct based decision tree $R_{5e-5}(150,8)$, minimum $RMSE$ FS tree $FS_{15}(150,4)$ and full tree $N_{19}(150,6)$ for $S_{max} = 150$.

Parameters	S_{max}	150		
	P_{min}	8	4	6
	$\epsilon \times 10^{-5}/SA$	5.00	FS	None
CV results	$RMSE$	12.8	12.4	12.9
	MAE	7.56	7.57	7.68
	$medAE$	3.83	4.05	3.91
	SE	55.5	38.2	50.2
	SI_{Tree}	0.82	0.24	0.51
	$\bar{p}_A$	12.6	15.0	19.0
	$\bar{p}_T$	11.3	11.9	13.5

Table 4.9. CV and Test Result Comparison between minimum *RMSE* Reduct based decision tree $R_{5e-5}(150,8)$, minimum *RMSE* FS tree $FS_{15}(150,4)$ and full tree $N_{19}(150,6)$ for $S_{max} = 150$ (cont.).

Test Results	p_A	13	15	19
	p_T	12	13	14
	r_p	0.86	0.88	0.87
	R^2_p	0.69	0.75	0.72
	<i>RMSE</i>	12.3	11.0	11.6
	<i>MAE</i>	7.47	6.96	7.34
	<i>medAE</i>	3.48	4.44	3.90
	q ₉₀	19.7	16.5	19.7
	<i>MGAE</i>	31.3	27.9	28.6
	<i>maxAE</i>	68.8	53.9	68.8

4.4. A Summary of the Models Constructed

In summary, regression trees with various S_{max} and P_{min} are constructed via FS, BE, reduct sets and using all the features. For each S_{max} value and feature selection algorithm, the optimum P_{min} and eps values are selected. It is seen that *RMSE* values from CV and for test sets decrease with increasing depth, however the tree becomes more complex and rule derivation becomes more demanding (Figure 4.49). Correlation coefficient between experimental and predicted data may reach 0.90, and the *RMSE* of test error falls down to 10 MPa by $BE_9(150,4)$ model, showing that a rather deep tree with a small parent size and 9 predictors can be rather successful in prediction. Although, models constructed via FS and BE show similar error profiles, models obtained via BE have generally higher stability around the optimum parameter values. It is seen that based on their stabilities full trees have moderate stability values for $S_{max} = 30$ and 60, and may also be used for rule derivation. Models constructed with reduct set has the highest stability, hence advantageous, particularly, for rule derivation.

What we looked at?

Feature Selection Method	Maximum Split Number	Minimum Parent Size	Selected Models	CV RMSE/ SI	Test RMSE/ R ²
FS	30	14-38	FS ₁₂ (30,20)	14.2/ 0.44	14.5/ 0.57
	60	12-18	FS ₁₃ (60,14)	13.1/ 0.41	11.9/ 0.71
	150	2-10	FS ₁₅ (150,4)	12.4/ 0.24	11.0/ 0.75
BE	30	30-36	BE ₁₁ (30,32)	14.2/ 0.48	13.6/ 0.62
	60	10-16	BE ₁₀ (60,12)	13.4/ 0.46	13.1/ 0.65
	150	2-8	BE ₉ (150,4)	13.2/ 0.51	10.0/ 0.79
Reduct	30	16-32	R _{5e-4} (30,24)	14.2/ 0.86	15.1/ 0.53
	60	8-20	R _{1e-4} (60,12)	13.2/ 0.79	13.2/ 0.64
	150	2-20	R _{5e-5} (150,8)	12.9/ 0.82	12.3/ 0.69
None	30	14-38	N ₁₉ (30,32)	14.6/ 0.68	14.5/ 0.57
	60	12-18	N ₁₉ (60,14)	13.4/ 0.68	13.0/ 0.65
	150	2-10	N ₁₉ (150,6)	12.9/ 0.51	11.6/ 0.72

Figure 4.49. A tabulated summary of the models and their performances in CV and the test set

4.5. Rule Extraction from the Regression Tree

Based on the analysis described on Sections 4.1 to 4.3, the model constructed via $R_{1e-4}(60,12)$ shows reasonable prediction accuracy with high stability and moderate tree depth, hence rules to predict CS are derived using this approach. In Figure 4.50, the resulting regression tree is shown. For the intermediate nodes the decision rules are noted inside the node boxes and the average CS values are shown on the top; if the condition in the rectangle is satisfied the right path is followed and vice versa. If the data are divided based on material type, the path to be followed is based on whether the material group is noted on the left or on the right. In case of terminal nodes, the numbers in the boxes show i) the mean CS and so the prediction of the tree, and ii) the number of observations at each node noted inside the parantheses. The number on the top of the terminal node boxes indicates the standard deviation of each terminal node. While a moderately deep tree with 60 splits is used, it is still difficult to derive rules, consistent with human understanding. Hence, a graphical/statistical approach is employed to make it easier to deduce important rules and place the “repeating themes” in the tree in a unifying perspective. To this end, rules are exhibited using locally constructed surfaces above 2-D predictor spaces, determined by the most important quantitative features revealed by the tree. In each subfigure representing local regression planes, the bold integer on the sub-figures show the model number, and the numbers in brackets indicate the number of data points and the number of different articles from which they are derived.

Material type used as aluminosilicate source resides on the top node of the tree: On the average, Group A materials (left) yield lower CS values (19 MPa), while Group B materials (right) have a higher CS (48 MPa). The material groupings seen in the regression tree are listed in Table 4.10 where the groups are listed based on the order of appearance in the regression tree (Figure 4.50), hence the alphabetic order of the group abbreviation. In each group the materials are mentioned based on the number of distinct materials combined; single and combinations of two or three. Groups B, M and N are noted with italic emphasizing their high performance. It should also be noted that there are material types among group A, that also yield high strength values, as discussed below. MK and RM containing geopolymers form most of the group A materials, which yield relatively lower CS

values. It is seen that most of the combinations of the aluminosilicate precursors include fly ash (Figure 3.2), and combination of FS with CAC, OPC, S, and SF+OPC are found to yield higher CS values. Addition of slag (S) in FA and RHA is presumed to strengthen the mortars, but this effect is not observed in the presence of MK and RM. The addition of Ca containing components in the form of Ordinary Portland Cement (OPC) and Calcium Activated Concrete (CAC) seem to increase CS, but this correlation is not observed when Lime (L), which is another Ca containing mineral, is combined with FA. Special components, such as EMDR and CG, yield high CS values. In this part of the tree, relatively the average CS is at a relatively low value equal to 19 MPa. The second level division yields Group C materials, which reside on the most left hand side of the tree, and include all Red Mud composites. Among the third level nodes, Group C yields the lowest average CS equal to 6.9 MPa. The other branch in the second level leads to Group D materials, in which MK based materials, except MK+RM in group C) yield a relatively higher average CS equal to 29 MPa. The analysis, so far, indicates that addition of Red Mud is correlated with obtaining composites with low CS.

The left part of the tree (including Group A materials) is visualized on a 2-D feature plane, constructed from Na/Al ratio and Purity (Figure 4.51), since these two predictors define splitting relations in the top levels numerous times. It should be noted that different markers are used describe the different material groups in Figure 4.51. First, Group E materials are focused on; these materials are split into Groups G and H in the following levels of the tree, and the FA+RM combination (Group G) is expected to yield a low CS value equal to 1.9 MPa. Group H is explained via the other relatively deeper branch, using purity and t_2 as split features. Hence, a local model using features purity and t_2 is constructed using 104 data points of Group H (shown with gray circles) from 20 different papers, that is the highest number of different sources analysed in a single local model in the current study (Model 1). Indeed, this large variety in sources is revealed in the complexity of the model; unlike all the models to be discussed, it is not possible to capture the overall distribution of the data using a low order linear model. Hence, nonparametric kernel ridge regression (Table 4.12) is employed, and the fitting RMSE value is found to be equal to 1.9 MPa (Table 4.13). The subfigure of model 1 in Figure 4.51 shows a rather complex behaviour, yet it is still possible to determine a number of local maxima of CS with respect purity and t_2 . For $t_2 < 5$ -

10 days, CS is increased by ~ 5 MPa at purity values equal to ~ 0.55 and ~ 0.85 ; a more significant increase of CS by ~ 10 MPa is obtained for $t_2 < \sim 25$ days at a purity value equal to ~ 0.65 . It should, however, be noted that another important factor may be the source of impurity, which is not included in the current study. CS decreases with increasing ambient curing time for purity < 0.60 and purity < 0.70 ; this may be attributed to crack formation during water evaporation in longer term curing, and the lack of aluminosilicate component to construct the tetrahedral structure of the polymers. These points show that significant interactions between purity and t_2 may be present. Projecting samples belonging to Group G on the surface (shown with stars in Figure 4.51) shows that RM+FA combinations lay below the surface defined by Group H samples over the low purity-low t_2 region, showing that Red Mud+Fly Ash combination is likely to yield a significantly low CS.

Next, samples in Group F, with an average CS of 14 MPa are examined (shown with black circles in Figure 4.51). In this side of the branches of the tree, Na/Al ratio and t_2 are used as splitting features, hence the local surface is constructed using these two features. 36 data points from 14 different papers reside in two separate locations of Purity-Na/Al ratio plane: high Purity-low Na/Al region, and low Purity-high Na/Al, but at the same time span t_2 dimension between 0 and 60 days. A model with second order terms of t_2 and first order interactions of Na/Al and t_2 is constructed (model 2 in Table 4.12), giving a RMSE value of 8.6 MPa with a low R^2 equal to 0.17 (Table 4.13). The high p-value of the model equal to 0.21 brings doubt on the reliability of the model, hence the derived rules should be met with caution. When Na content is low, CS seems to have a minimum at ambient curing time equal to 20-30 days, but t_2 seems to have a positive effect on CS at higher Na values. Moreover, Na content increase enhances the geopolymer strength development, particularly, for longer ambient curing times.

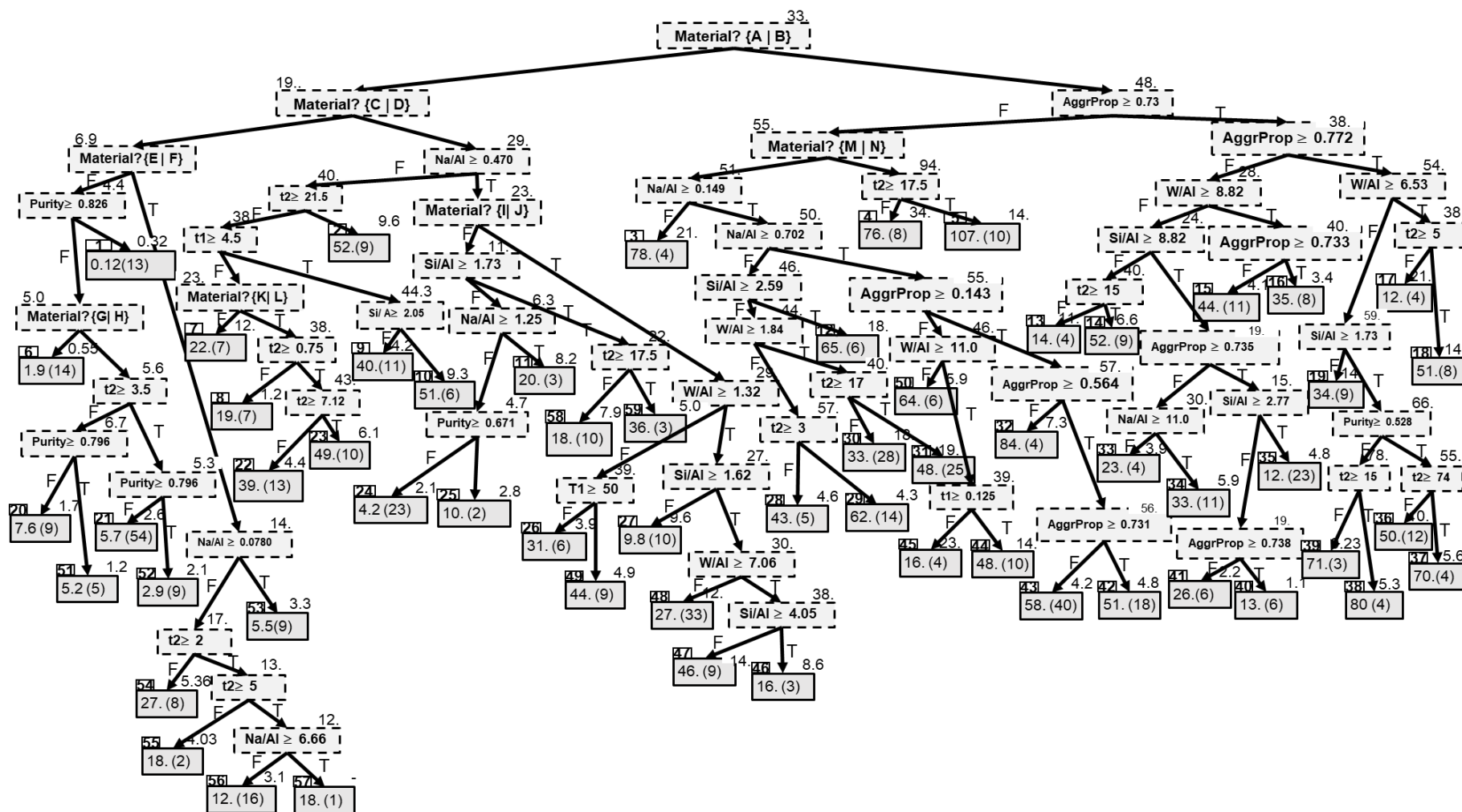


Figure 4.50. The regression tree constructed via $R_{1e-4}(60,12)$.

Table 4.10. Material groupings used in the regression tree.

Denotation	Material Group
A	{FA, MK, RM, RHA, FA+L, FA+MK, FA+RM, FA+SF, MK+RM, MK+S, RM+S, RM+SF, FA+S+L, FA+S+POFA, FA+S+SF, FA+S+SMC}
B	{CG, EMDR, S, FA+CAC, FA+OPC, FA+S, FA+SF+OPC, RHA+S, S+SF+L}
C	{RM, FA+L, FA+SF, FA+RM, MK+RM, RM+S, RM+SF, FA+S+L, FA+S+SMC}
D	{FA, MK, RHA, FA+MK, MK+S, FA+S+SF, FA+S+POFA}
E	{RM, FA+RM, RM+SF, FA+S+L, FA+S+SMC}
F	{FA+L, FA+SF, MK+RM, RM+S}
G	{RM+FA}
H	{RM, RM+SF, FA+S+L, FA+S+SMC}
I	{FA+MK, MK+S, FA+S+POFA}
J	{FA, MK, RHA}
K	{FA, FA+S+SF}
L	{FA+MK, FA+POFA+S}
M	{CG, EMDR, FA+S, FA+OPC, RHA+S, FA+OPC+SF }
N	{S, FA+CAC}

Table 4.11. Predictor that the marker size in each subfigure represents.

Model No*	Marker Size Predictor
3	Na/Al Ratio
7	t1
8	t2
11	Si/Al Ratio

*For the models that are not included in Table 4.11 no marker size representation has been used.

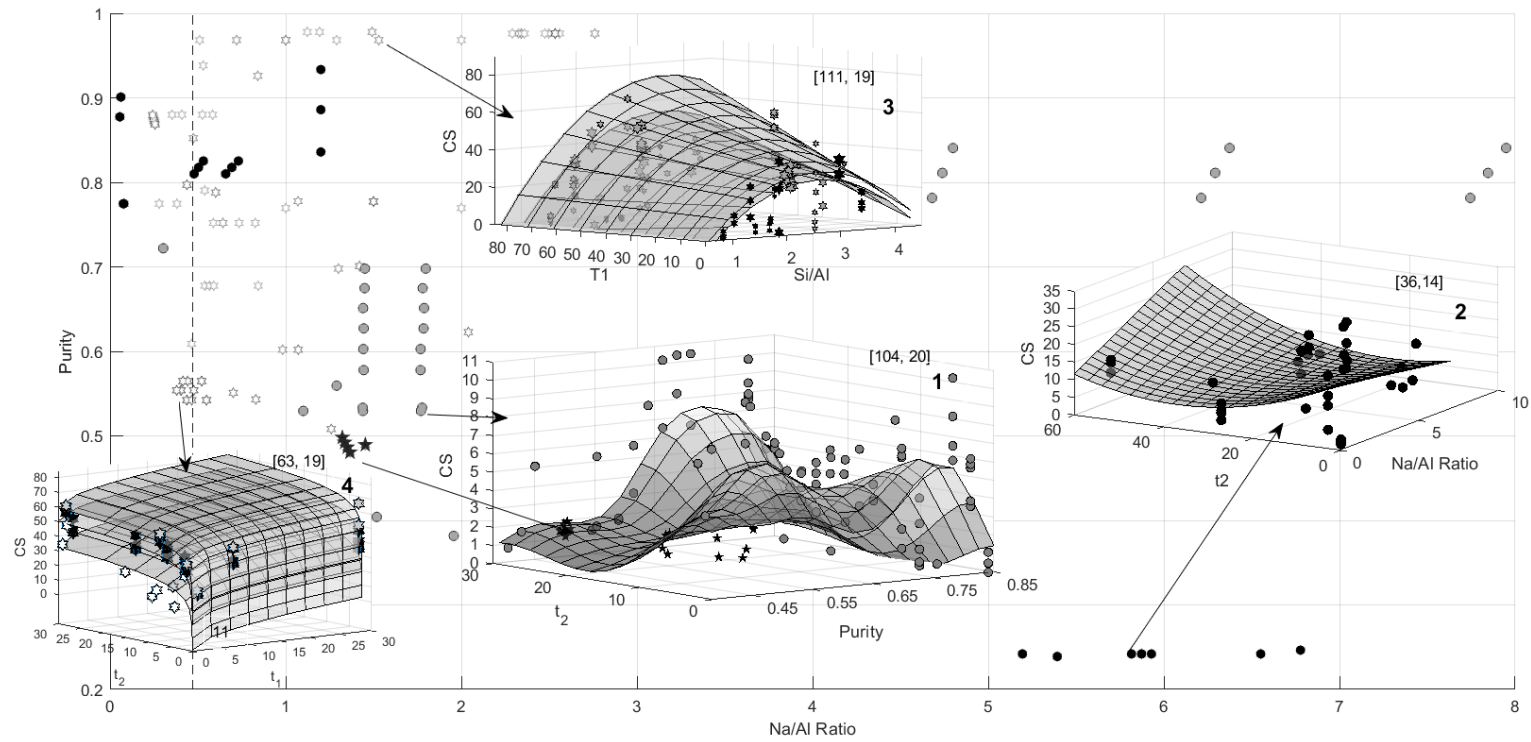


Figure 4.51. Local regression model surfaces constructed for left hand side of the $R_{1e-4}(60,12)$ regression tree. The marker size of subfigure 3 is proportional to the value of Na/Al ratio.

For samples synthesized using Group D materials satisfying $\text{Na}/\text{Al} < 0.475$, regression plane is constructed above t_1 and t_2 , involving two types of materials, groups K and L, and also an indicator variable, showing whether $\text{si_al} < 2.05$. (model 4 in Table 4.12). Model 4, using 63 samples from 19 papers, yields a satisfactory RMSE value equal to 8.6 MPa, with a RMSE value equal to 0.60 (Table 4.12). CS of Group K (solid hexagons), consisting of FA, FA+S+SF, is ~ 17 MPa lower than that of group L (empty hexagons), consisting of FA+MK, FA+S+POFA (similar to group I), showing (opposite to that in Model 3) that using solely FA yields lower CS compared to using FA+MK. This shows that both FA+MK and FA+S+POFA may be advantageous only under specific conditions, i.e. for $\text{Na}/\text{Al} < \sim 0.48$, as seen in the tree. Subfigure 4 of Figure 4.50 shows that CS increases rapidly with respect to t_2 in the first 5 days, followed by a slow increase, as a typical logarithmic response. A value of $\text{si_al} < 2.05$ decreases the CS by ~ 10 MPa, while t_1 has a positive partial effect, smaller compared to that of t_2 , on CS. The three surfaces (from top to bottom) in Subfigure 4 of Figure 4.50 correspond to group L with $\text{si_al} \geq 2.05$, group L with $\text{si_al} < 2.05$, and group K with $\text{si_al} < 2.05$, respectively.

In the right side of the tree (Group B), the predictors appearing more frequently on the upper nodes are aggregate proportion, Water/Al and Na/Al, hence the 2-D plane, over which local models are to be constructed, are formed using features aggregate proportion and Water/Al; additionally marker sizes of the samples are proportional to Na/Al ratio (Figure 4.52). This part of the tree consists of samples with higher CS values (48 MPa), compared to the left side of the tree. Group B materials include FA combined with calcium activated concrete. In the study of Temuujin *et al.*, it was noted that the addition of calcium in ambient cured geopolymers enhances CS, but deteriorates them if heat cured at 70°C (Temuujin *et al.*, 2009). It was also mentioned that higher porosities characterize heat cured mortars, possibly linked to higher rates of water evaporation. In contrast with the models derived for the left side of the tree (Table 4.12), the number of data points and number of various papers used to construct the Group B models are lower, at a maximum of 104 points and 6 papers (model 10, Table 4.14). Additionally, models 5-11 show relatively higher R^2 and lower P-values, showing the significance of local fits (Table 4.15).

Aggregate proportion values of the data used in Models 5 and 6, both of which have $R^2 \approx 0.75$, are > 0.77 , and both models contain highly significant $t_2/\log t_2$ terms (Table 4.14 and 4.15): CS sharply increases within $t_2 < \sim 5$ days, and followed by small/no changes (subfigures 5 and 6 Figure 4.52). This effect of t_2 on CS has also been seen in Model 4. The positive correlation of Si/Al with CS in Models 5 is again similar to that seen in Model 4. The predictors used in Model 6 are identical to those in Model 1, yet increasing purity has a simple positive effect on CS, unlike the complex interactions with t_2 seen in Model 1. This shows that relation of CS with purity and t_2 might show difference between {FA+L, FA+SF, MK+RM, RM+S} and {Fa+S}.

Model 7 is constructed from 45 points and 4 papers, covering a large range of aggregate proportion values, and yields a reasonable $R^2 = 0.62$ value (Table 4.15 and Figure 4.12). The size of the points on the Model 7 surface represent the heat curing time. The quadratic relation of AggrProp with CS reaches a theoretical maximum at ~ 0.35 , while small W/Al values and large t_1 seem to be correlated with large CS values (Table 4.14). It should be noted that the effect of t_1 has an identical positive sign to that seen in Model 7, showing that t_1 may have a positive contribution to CS under various conditions, such as small Na/Al, or high W/Al ratio values. Mortars without aggregates are generally processed for a longer time of heat curing than mortars containing high aggregate proportion. Additionally, mortars with large Water/Al proportion are longer heat cured than those with small water containing specimens. These observations remind one not to interpret the correlations as strict causal relations, but possible relations to be examined using controlled experiments.

Models 8 and 9 are separately constructed using data coming from single papers. RMSE value of Model 8 is very low at 4.72 MPa, but RMSE value of Model 9 is significantly high (24.6 MPa) with a low-moderate $R^2 = 0.36$ (Table 4.15). The common point in both models is the negative effect of Si/Al on CS. Aggregate proportion of these two data are somewhat close, at 0.5 and ~ 0.7 , hence it might be interesting to check whether the effect of Si/Al on CS might change its sign for this range of aggregate proportion values. As seen in models 2 and 3, sign of Na_Al parameter is positive in Model 8, and the positive logarithmic effect of t_2 is seen in both models 8 and 9. Model 10 is a simple model, constructed from 4

data points from 2 papers, and comprises only positive logarithmic effect of t_2 (Table 4.14). Model 11 is constructed from 78 data points from 6 papers, and yields a RMSE value equal to 17.9 MPa. Similar to that seen in model 7, W/Al ratio has a negative effect on CS. It should be noted that W/Al ratio of data used in these two models span a range from 0 to 14, hence one may say that increasing W/Al ratio is likely to decrease CS for group B materials, unless aggregate proportion is high ($>.7$). The positive effects of Si/Al ratio and t_2 on CS are again to be noted.

As an overall analysis of the effect of features, the ambient curing time effect seems to be the most significant predictor correlated with CS. For all the models, the effect of time is more obvious at early ages (before 5-10 days), and stabilizes afterwards (all 11 models except 3 and 7). The secondly more frequently appearing predictor is Si/Al. Si/Al ranges of Models 9 and 5 are similar ($Si/Al \in (1,3)$) and span different Aggregate proportion and W/Al ranges. It can be noted, then, that when W/Al and aggregate proportion increases an increase in Si/Al ratio has a positive effect on CS, compared for the cases where W/Al is considerably low (<1). In addition, the two different curves in Figure 6 indicate that when Si/Al is lower than 1.73, CS is lower (more than 10 MPa) and in cases that Si/Al is greater CS gets greater. When wider ranges are considered Si/Al ratio has a positive effect (Model 5 and 11) but for narrower ranges (Model 4) the effect is negative. The reason of the negative sign for Model 9 is only one point pulling the surface towards it; so, whether an increasing effect will be observed with increasing Si ratio depends on various parameters and not a general trend can be underlined.

Table 4.12. Statistics of the local models constructed for the left-hand side of the $R_{1e-4}(60,12)$ regression tree models

Model no.	N	ErrorDoF	P-value	RMSE	R^2	R_{adj}^2
1	104	-	-	1.87	0.64	-
2	36	31	0.209	8.23	0.17	0.06
3	111	103	$<10^{-10}$	9.87	0.65	0.63
4	63	58	$<10^{-10}$	8.58	0.60	0.58

Table 4.13. Local regression model coefficients, their parameters estimates and p-values constructed from the left hand side of the R1e-4(60,12) regression tree.

Model no.	Coefficient	Estimate	SE	t_{stat}	p_{value}
1(KRR with $\sigma = 6, \lambda = 3$)	<i>Purity</i>	-	-	-	-
	t_2	-	-	-	-
2	b_0	18.6	2.41	7.72	1.03×10^{-8}
	<i>na_al</i>	-0.729	0.956	-0.763	0.451
	t_2	-0.940	0.294	-2.18	3.71×10^{-2}
	$na_al \times t_2$	5.76×10^{-2}	4.80×10^{-2}	1.20	0.239
	t_2^2	8.70×10^{-3}	5.24×10^{-3}	1.66	0.106
3	b_0	-24.4	11.2	-2.17	3.26×10^{-2}
	<i>si_al</i>	36.8	6.34	5.80	7.32×10^{-8}
	<i>Group I</i>	-2.94	3.34	-0.880	0.381
	<i>na_al</i>	5.95	1.29	4.63	1.08×10^{-8}
	T_1	-0.208	0.155	-1.34	0.182
	$si_al \times T_1$	0.213	5.32×10^{-2}	4.00	1.17×10^{-4}
	$Group\ I \times T_1$	-0.130	6.10×10^{-2}	-2.13	3.51×10^{-2}
	si_al^2	-7.34	1.20	6.13	1.61×10^{-8}
4	b_0	37.3	2.84	13.1	$<10^{-10}$
	t_1	4.31	0.822	5.24	$<10^{-10}$
	$\log t_2$	6.49	0.849	7.65	2.26×10^{-6}
	$I\{Si_al < 2.05\}$	-9.12	2.73	-3.34	1.46×10^{-3}
	<i>Material</i>	-17.3	3.30	5.25	2.26×10^{-6}

Table 4.14. Local regression model coefficients, their parameters estimates and p-values constructed from the right hand side of the $R_{1e-4}(60,12)$ regression tree

Model no.	Coefficient	Estimate	SE	t_{stat}	$pValue$
5	b_0	-119	53.7	-2.21	0.038
	Si/Al	56.5	24.9	2.27	0.034
	t_2	15.9	3.345	4.80	1.0×10^{-4}
	$\log(t_2)$	-0.258	0.143	-1.80	0.087
6*	b_0	-16.1	17.2	-0.937	0.36
	$\log(t_2)$	15.2	4.74	3.20	3.5×10^{-3}
	$\log(Purity)$	-113	26.1	-4.43	1.8×10^{-4}
	$\log(t_2) \times \log(Purity)$	16.7	7.21	2.31	0.028
7	b_0	153	19.6	7.78	1.6×10^{-9}
	$AggrProp$	285	43.3	6.76	7.2×10^{-8}
	$\log(W/Al)$	-45.7	8.31	-5.60	2.4×10^{-6}
	t_1	1.50	0.462	3.28	2.4×10^{-3}
	$AggrProp^2$	-393	63.8	-6.22	2.3×10^{-7}
8	b_0	451	5090	8.00	$<10^{-10}$
	Si/Al	-227	20.4	-11.1	$<10^{-10}$
	Na/Al	226	6.32	35.8	$<10^{-10}$
	$\log(t_2)$	3.03	0.495	6.12	1.9×10^{-8}
9	b_0	73.9	27.4	2.69	0.017
	Si/Al	-30.6	16.8	-1.81	0.091
	$\log(t_2)$	23.7	8.42	2.82	0.013
10	b_0	69.0	2.65	26.1	1.5×10^{-3}
	$\log(t_2)$	8.45	1.14	7.41	0.018
11	b_0	39.1	6.09	6.42	1.2×10^{-8}
	$\log(W/Al)$	-24.8	13.3	-3.40	0.0011
	$\log(Si/Al)$	31.5	13.3	2.37	0.020
	$\log(t_2)$	5.98	2.06	2.91	4.8×10^{-3}

*: For $Si/Al < 1.73$, CS is 10-25 MPa lower than $Si/Al > 1.73$, with a $\log(t_2)$ coefficient equal to 11.0.

Table 4.15. Statistics of the local models constructed for the right-hand side of the $R_{le-4}(60,12)$ regression tree models

Model no.	N	ErrorDoF	P-value	RMSE	R^2	R^2_{adj}
5	25	21	9×10^{-7}	11.2	0.76	0.73
6	31	27	5×10^{-8}	8.01	0.74	0.71
7	45	40	5×10^{-8}	12.2	0.62	0.58
8	104	100	$< 10^{-10}$	4.72	0.93	0.93
9	18	15	0.034	24.6	0.36	0.28
10	4	2	0.02	4.74	0.97	0.95
11	78	74	8×10^{-4}	17.9	0.20	0.17

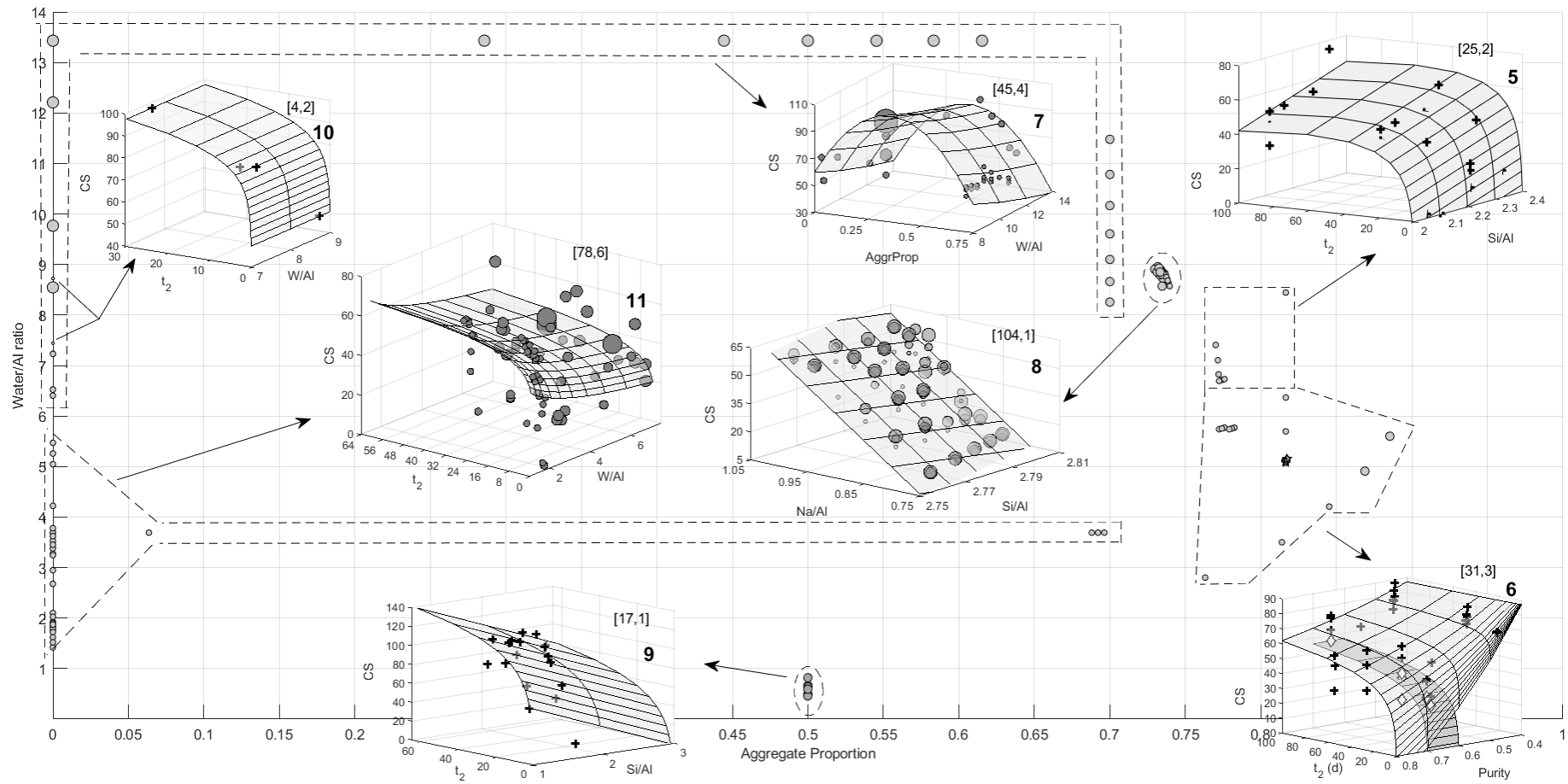


Figure 4.52. Local regression model surfaces constructed for right hand side of the $R_{1e-4}(60,12)$ regression tree. Sizes of the markers in subfigures 7,8 and 11 are proportional to the values of t_1 , t_2 and Si/Al, respectively

5. CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

Geopolymers are eco-friendly materials which are considered as future cement alternatives due to their low CO₂ emissions. As construction materials, producing geopolymers with high mechanical performance (e.g. high CS) is important. In this study, first, a data set comprising the measured compressive strength values of geopolymers under various experimental conditions are determined using available literature. Second, decision trees using all features and using selection algorithms, such as forward selection, backward elimination and reduct set, are constructed at three different tree sizes ($S_{max} = 30, 60$ and 150). The resulting trees are compared on the basis of prediction accuracy and stability, and finally a convenient regression tree with good stability and reasonable prediction accuracy is selected.

It is observed that increasing regression tree size leads to more accurate predictions with a RMSE equal to ~ 10 MPa, but interpretation of the resulting tree becomes difficult. Furthermore, wrapper algorithms, such as forward selection and backward elimination, are adequate for decreasing test error, however the resulting trees have very low stability. Hence, the regression tree using a reduct set feature algorithm ($R_{1e-4}(60,12)$) is selected with a stability index equal to 0.82 and a RMSE of 13.2 MPa on the test set.

Based on the regression tree, two group of materials characterized with lower (Group A) and higher (Group B) CS are determined. High performance leading precursors are mainly FA and Slag, whereas the lower ones are Red Mud containing compounds. For Group A, W/AI ratio aggregate proportion and ambient curing time are the most important predictors. The latter is found to have a logarithmic effect on the mortars. Generally lower W/AI values are to be used for high CS, while CS shows a maximum region around the aggregate proportion equal to 0.4. In contrast, ambient curing time effect may even have a defective effect in varying Purity cases. The effect of purity seems to be complicated and

highly depends on other experimental conditions. CS shows maximum around $\text{Si/Al} = 2.5$ for specific compounds. The effect of Si/Al seems to be relative to aggregate proportion, too. Synthesis type, i.e. one-part or two-part procedure, does not seem to influence CS.

5.1. Recommendations

As stated previously good mechanical performance is expected for geopolymers synthesized by the activation of *CG*, *EMDR*, *S*, *FA+CAC*, *FA+OPC*, *FA+S*, *FA+SF+OPC*, *RHA+S*, *S+SF+L* for aggregate proportion of 0.4. However, this region has not been spanned (based on the publications studied), so further experiments on this region are suggested. Moreover, it was noted that high CS is expected for EMDR activated geopolymers. The study considered during the data set formation is based on acid activation of EMDR. So, whether similar good performance is obtained for its alkali activation could also be experimentally checked. It is also seen that the type of the geopolymer (one or two-part) is not among the factors that affect the geopolymer CS the most. So, one-part geopolymer synthesis which is preferable for large scale production can be used for parameters optimized as discussed earlier.

REFERENCES

- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3), 392–398. <https://doi.org/10.1093/bioinformatics/btp630>
- Alakent, B., & Soyer-Uzun, S. (2019). Implementation of Statistical Learning Methods to Develop Guidelines for the Design of PLA-Based Composites with Tensile Strength Values. *Industrial and Engineering Chemistry Research*. <https://doi.org/10.1021/acs.iecr.8b05477>
- Ashour, A. F., Abdalhmid, J., Dai, X., & Khan, A. (2018). *Prediction of Rubberised Concrete Strength by Using Artificial Neural Networks*. (November).
- Behnood, A., Behnood, V., Modiri Gharehveran, M., & Alyamac, K. E. (2017). Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm. *Construction and Building Materials*. <https://doi.org/10.1016/j.conbuildmat.2017.03.061>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. <https://doi.org/10.1201/9781315139470>
- Bui, D. K., Nguyen, T., Chou, J. S., Nguyen-Xuan, H., & Ngo, T. D. (2018). A modified firefly algorithm-artificial neural network expert system for predicting compressive and tensile strength of high-performance concrete. *Construction and Building Materials*. <https://doi.org/10.1016/j.conbuildmat.2018.05.201>
- Chou, J.-S., Chong, W. K., & Bui, D.-K. (2016). Nature-Inspired Metaheuristic Regression System: Programming and Implementation for Civil Engineering Applications. *Journal of Computing in Civil Engineering*. [https://doi.org/10.1061/\(asce\)cp.1943-5487.0000561](https://doi.org/10.1061/(asce)cp.1943-5487.0000561)
- Cosma, S. (2006). Lecture 10 : Regression Trees. *October*.
- Davidovits, J. (1991). Geopolymers - Inorganic polymeric new materials. *Journal of*

Thermal Analysis. <https://doi.org/10.1007/BF01912193>

- Duxson, P., Provis, J. L., Lukey, G. C., Mallicoat, S. W., Kriven, W. M., & Van Deventer, J. S. J. (2005). Understanding the relationship between geopolymer composition, microstructure and mechanical properties. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 269 (1–3), 47–58. <https://doi.org/10.1016/j.colsurfa.2005.06.060>
- Feng, D., Provis, J. L., & Van Deventer, J. S. J. (2012). Thermal activation of albite for the synthesis of one-part mix geopolymers. *Journal of the American Ceramic Society*, 95(2), 565–572. <https://doi.org/10.1111/j.1551-2916.2011.04925.x>
- Han, Y., Cui, X., Lv, X., & Wang, K. (2018). Preparation and characterization of geopolymers based on a phosphoric-acid-activated electrolytic manganese dioxide residue. *Journal of Cleaner Production*, 205, 488–498. <https://doi.org/10.1016/j.jclepro.2018.09.141>
- Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*. <https://doi.org/10.14569/specialissue.2014.040203>
- Hu, K., Lu, Y., & Shi, C. (2003). Feature Ranking in Rough Sets. *AI Communications - Special Issue on Artificial Intelligence Advances in China*, 16(1), 41–50.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. <https://doi.org/10.1007/978-1-4614-7138-7>
- K, N. V., & Babu, D. L. V. (2018). Assessing the performance of molarity and alkaline activator ratio on engineering properties of self-compacting alkaline activated concrete at ambient temperature. *Journal of Building Engineering*, 20(June), 137–155. <https://doi.org/10.1016/j.jobbe.2018.07.005>
- Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). Feature Selection for Regression Problems. *The 8th Hellenic European Research on Computer Mathematics & Its Applications*, (i), 20–22.
- Kaya, K., & Soyer-Uzun, S. (2016). Evolution of structural characteristics and compressive strength in red mud-metakaolin based geopolymer systems. *Ceramics International*,

- 42(6), 7406–7413. <https://doi.org/10.1016/j.ceramint.2016.01.144>
- Kuncheva, L. I. (2007). A stability index for feature selection. *Proceedings of International Multi-Conference: Artificial Intelligence and Applications*, 390–395. Retrieved from papers2://publication/uuid/FC5277DF-B494-4316-8D02-E8CE794BAE37
- Lahoti, M., Narang, P., Tan, K. H., & Yang, E. H. (2017). Mix design factors and strength prediction of metakaolin-based geopolymer. *Ceramics International*. <https://doi.org/10.1016/j.ceramint.2017.06.006>
- Li, Y. (2006). Predicting materials properties and behavior using classification and regression trees. *Materials Science and Engineering A*. <https://doi.org/10.1016/j.msea.2006.06.100>
- Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*. <https://doi.org/10.1111/insr.12016>
- Luna-Galiano, Y., Leiva, C., Arenas, C., & Fernández-Pereira, C. (2018). Fly ash based geopolymeric foams using silica fume as pore generation agent. Physical, mechanical and acoustic properties. *Journal of Non-Crystalline Solids*. <https://doi.org/10.1016/j.jnoncrysol.2018.07.069>
- Luukkonen, T., Abdollahnejad, Z., Yliniemi, J., Kinnunen, P., & Illikainen, M. (2018). One-part alkali-activated materials: A review. *Cement and Concrete Research*, 103(October 2017), 21–34. <https://doi.org/10.1016/j.cemconres.2017.10.001>
- Nogueira, S., Sechidis, K., & Brown, G. (2017). On the Stability of Feature Selection Algorithms. *J. Mach. Learn. Res.*, 18(1), 6345–6398. <https://doi.org/10.1162/jmlr-2017-18-014>
- Olson, G. B. (2013). Genomic materials design: The ferrous frontier. *Acta Materialia*. <https://doi.org/10.1016/j.actamat.2012.10.045>
- Onishi, T., Kadohira, T., & Watanabe, I. (2018). Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *Science and Technology of Advanced Materials*. <https://doi.org/10.1080/14686996.2018.1500852>
- PK, A., & V, P. (2013). Review on Feature Selection Techniques of DNA Microarray Data. *International Journal of Computer Applications*, 61(12), 39–44. <https://doi.org/10.5120/9983-4814>

- Rajerajeswari, A., Dhinakaran, G., & Ershad, M. (2014). Compressive strength of silica fume based geopolymer concrete. *Asian Journal of Applied Sciences*, Vol. 7, pp. 240–247. <https://doi.org/10.3923/ajaps.2014.240.247>
- Renca.org. (n.d.). Product Range. <http://www.renca.org/products> , accessed in July 2019.
- Reddy, M. S., Dinakar, P., & Rao, B. H. (2018). Mix design development of fly ash and ground granulated blast furnace slag based geopolymer concrete. *Journal of Building Engineering*, 20(August), 712–722. <https://doi.org/10.1016/j.jobbe.2018.09.010>
- Rovnaník, P. (2010). Effect of curing temperature on the development of hard structure of metakaolin-based geopolymer. *Construction and Building Materials*, 24(7), 1176–1183. <https://doi.org/10.1016/j.conbuildmat.2009.12.023>
- Safavian, S. R., & Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man and Cybernetics*. <https://doi.org/10.1109/21.97458>
- Sagin, A. N., & Ayvaz, B. (2018). Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector. *Southeast Europe Journal of Soft Computing*. <https://doi.org/10.21533/scjournal.v7i1.149>
- Sturm, P., Gluth, G. J. G., Brouwers, H. J. H., & Kühne, H. C. (2016). Synthesizing one-part geopolymers from rice husk ash. *Construction and Building Materials*, 124, 961–966. <https://doi.org/10.1016/j.conbuildmat.2016.08.017>
- Sturrock, C. P., & Bogaerts, W. F. (1997). Empirical Learning Investigations of the Stress Corrosion Cracking of Austenitic Stainless Steels in High-Temperature Aqueous Environments. *Corrosion (Houston)*. <https://doi.org/10.5006/1.3280476>
- Temuujin, J., Van Riessen, A., & MacKenzie, K. J. D. (2010). Preparation and characterisation of fly ash based geopolymer mortars. *Construction and Building Materials*, 24(10), 1906–1910. <https://doi.org/10.1016/j.conbuildmat.2010.04.012>

- Temuujin, J., van Riessen, A., & Williams, R. (2009). Influence of calcium compounds on the mechanical properties of fly ash geopolymer pastes. *Journal of Hazardous Materials*. <https://doi.org/10.1016/j.jhazmat.2008.12.121>
- Wagners. (n.d.). Projects | Earth Friendly Concrete | What We Do | Main | Wagners. <https://www.wagner.com.au/main/what-we-do/earth-friendly-concrete/projects>, accessed in July 2019.
- Yadollahi, M. M., Benli, A., & Demirboğa, R. (2015). The effects of silica modulus and aging on compressive strength of pumice-based geopolymer composites. *Construction and Building Materials*, 94, 767–774. <https://doi.org/10.1016/j.conbuildmat.2015.07.052>
- Ye, N., Chen, Y., Yang, J., Liang, S., Hu, Y., Xiao, B., Huang, Q., Shi, Y., Hu, J., & Wu, X. (2016). Co-disposal of MSWI fly ash and Bayer red mud using an one-part geopolymeric system. *Journal of Hazardous Materials*, 318, 70–78. <https://doi.org/10.1016/j.jhazmat.2016.06.042>
- Yeh, I. C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*. [https://doi.org/10.1016/S0008-8846\(98\)00165-3](https://doi.org/10.1016/S0008-8846(98)00165-3)
- Zhang, G., He, J., & Gambrell, R. P. (2010). Synthesis, Characterization, and Mechanical Properties of Red Mud–Based Geopolymers. *Transportation Research Record: Journal of the Transportation Research Board*, 2167(1), 1–9. <https://doi.org/10.3141/2167-01>

APPENDIX A: N₁₉(60,14) FULL REGRESSION TREE

Table A.1. Denotation and Material Group Mapping for N₁₉(60,14) Regression Tree

Denotation	Material Group
A	{MK,FA,MK+FA, MK+RM, RHA, RM, RM+SF, RM+FA, FA+L, FA+S+L, RM+S, FA+S+SMC, FA+SF, S+FA+POFA, MK+S, FA+S+SF}
B	{FA+S, S, S+SF+L, EMDR, RHA+S, FA+SF+OPC, FA+CAC, FA+OPC, CG}
C	{MK+RM, RM, RM+SF, RM+FA, FA+L, FA+S+L, RM+S, FA+S+SMC, FA+SF}
D	{MK,FA,MK+FA, RHA, S+FA+POFA, MK+S, FA+S+SF}
E	{RM, RM + SF, RM + FA, FA +S+L, FA+S+SMC}
F	{MK+RM, FA+L, RM+S, FA+SF}
G	{MK+FA, S+FA+POFA, MK+S}
H	{MK, FA, RHA, FA+ SF+S}
I	{FA, FA+S+SF}
J	{MK+FA, S+FA+POFA}
K	{FA+S, EMDR, RHA+S, FA+SF+OPC, FA+OPC, CG}
L	{S, FA+CAC}
M	{FA+SF}
N	{S+SF+L}

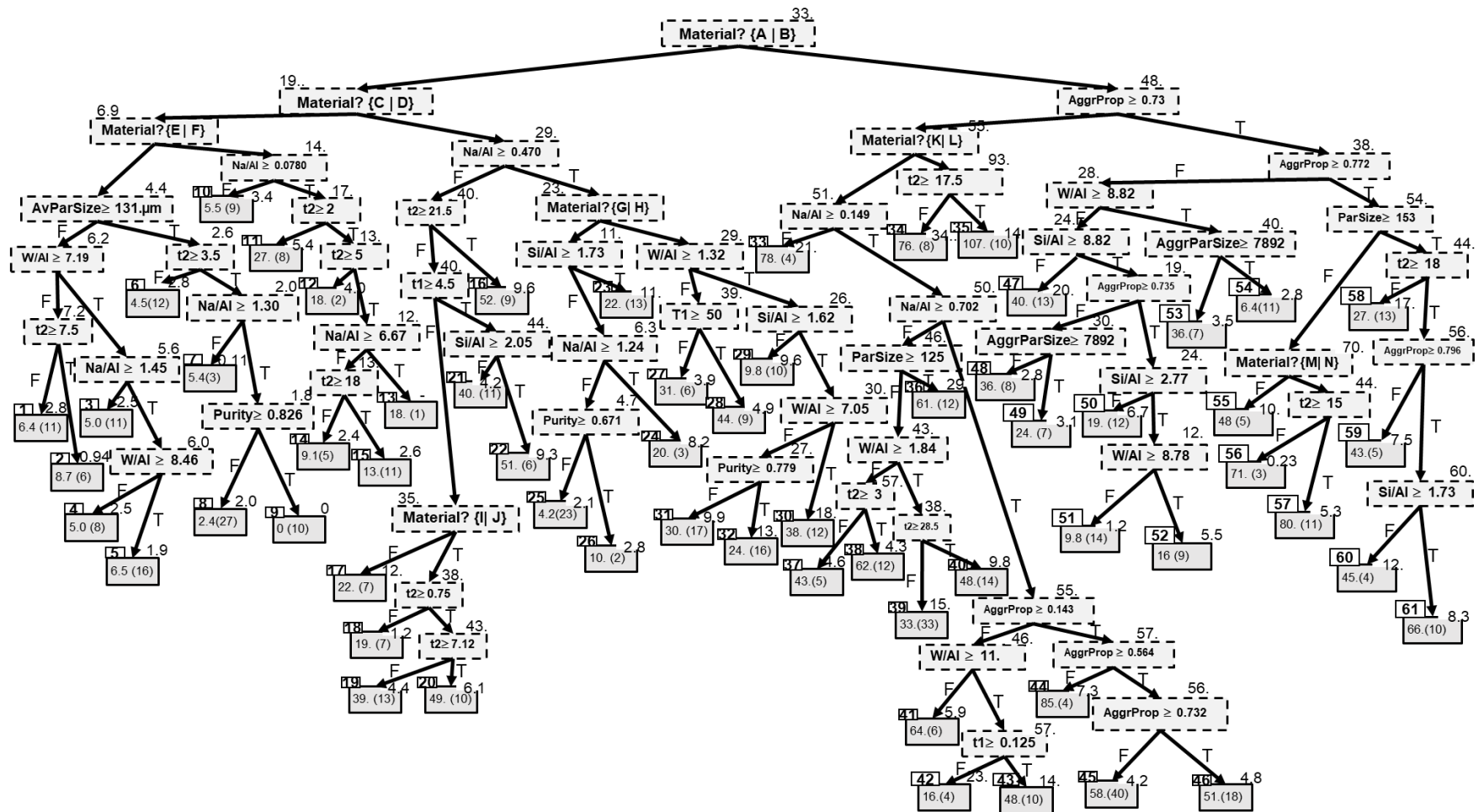


Figure A.1. N₁₉(60,14) Full Regression Tree

