

Interpretação e desempenho em modelos preditivos da resistência do concreto

¹ Sandro Martini 

² Lara Kühl Teles 

¹ Universidade São Judas Tadeu, Brasil, prof.martini@gmail.com

² Grupo de Materiais Semicondutores e Nanotecnologia (GMSN), Instituto Tecnológico de Aeronáutica (ITA), Brasil, lkteles@ita.br

JUTS – Journal of Urban Technology and Sustainability

ISSN: 2675-780X

DOI: <https://doi.org/10.47842/juts.v8i1.85>

Organização: Comitê Científico Interinstitucional

Editor Chefe: Guilherme Gaudereto

Avaliação: Double Blind Review pelo SEER/OJS

Recebido: 09/11/2025

Aceito: 10/06/2026

RESUMO

A modelagem preditiva da resistência do concreto enfrenta o equilíbrio entre a precisão de modelos não lineares complexos e a transparência das abordagens lineares. A contribuição original deste estudo é a quantificação explícita dessa relação de compromisso e a identificação, via análise de componentes principais, de fatores latentes estatísticos nos dados de dosagem, interpretados à luz da tecnologia do concreto. Analisou-se um banco de dados com 1.030 misturas, comparando quatro abordagens — modelo de referência, regressão por componentes principais, regressão linear múltipla e *random forest* — avaliadas no conjunto de teste e por validação cruzada de dez partições. A análise revelou três fatores latentes fisicamente interpretáveis: o contraste entre teor de água livre e uso de superplastificante e adições, o balanço entre escória de alto-forno e agregado graúdo e o balanço entre cimento e cinza volante, que juntos explicam 62,8% da variância total. O *random forest* atingiu a maior precisão ($R^2 = 0,88$; R^2 médio = $0,910 \pm 0,023$ na validação cruzada), enquanto a regressão por componentes principais ($R^2 = 0,48$; R^2 médio = $0,425 \pm 0,075$) demonstrou que a redução dimensional preserva a interpretabilidade ao custo de precisão, sendo recomendada como ferramenta exploratória, e não como modelo preditivo primário. A principal limitação é a sub-representação de concretos de alta resistência (acima de 60 MPa) no conjunto de dados utilizado, restringindo a generalização nessa faixa. Os resultados reforçam a necessidade de técnicas de interpretabilidade pós-hoc, como SHAP e LIME, para conciliar precisão e transparência em aplicações de engenharia.

Palavras-chave: resistência do concreto; análise de componentes principais; *trade-off* precisão-interpretabilidade; *random forest*; modelagem preditiva.

Interpretation and performance in predictive models of concrete strength

Predictive modeling of concrete compressive strength faces a fundamental trade-off between the accuracy of complex nonlinear models and the transparency of linear approaches. The original contribution of this study is the explicit quantification of this trade-off and the identification, through principal component analysis, of statistical latent factors in mix-design data, interpreted in light of concrete technology principles. A dataset of 1,030 mixtures was analyzed, comparing four approaches — a baseline model, principal component regression, multiple linear regression, and random forest — evaluated on a hold-out test set and through 10-fold cross-validation. The analysis revealed three physically interpretable latent factors: the contrast between free water content and the use of superplasticizer and mineral additions, the balance between blast furnace slag and coarse aggregate, and the balance between cement and fly ash, which together explain 62.8% of the total variance. Random forest achieved the highest accuracy ($R^2 = 0.88$; mean $R^2 = 0.910 \pm 0.023$ in cross-validation), whereas principal component regression ($R^2 = 0.48$; mean $R^2 = 0.425 \pm 0.075$) showed that dimensionality reduction preserves interpretability at the expense of accuracy, being recommended as an exploratory tool rather than a primary predictive model. The main limitation is the underrepresentation of high-strength concretes (above 60 MPa) in the dataset, restricting generalization in this range. The results reinforce the need for post-hoc interpretability techniques, such as SHAP and LIME, to reconcile accuracy and transparency in engineering applications.

Keywords: concrete strength; principal component analysis; accuracy–interpretability trade-off; random forest; predictive modeling.



Interpretación y desempeño en modelos predictivos de la resistencia del hormigón

La modelización predictiva de la resistencia a compresión del hormigón enfrenta el equilibrio entre la precisión de los modelos no lineales y la transparencia de los enfoques lineales. La contribución original de este estudio es la cuantificación explícita de esta relación de compromiso y la identificación, mediante análisis de componentes principales, de factores latentes en los datos de dosificación, interpretados según la tecnología del hormigón. Se analizó una base de 1.030 mezclas, comparando cuatro enfoques — modelo de referencia, regresión por componentes principales, regresión lineal múltiple y random forest — evaluados en conjunto de prueba y validación cruzada de diez particiones. El análisis reveló tres factores latentes físicamente interpretables: el contraste entre agua libre y el uso de superplastificante y adiciones, el balance entre escoria y agregado grueso y el balance entre cemento y ceniza volante (62,8% de la varianza total). El random forest alcanzó la mayor precisión ($R^2 = 0,88$; R^2 medio = $0,910 \pm 0,023$ en validación cruzada), mientras que la regresión por componentes principales ($R^2 = 0,48$; R^2 medio = $0,425 \pm 0,075$) evidenció que la reducción dimensional preserva la interpretabilidad a costa de la precisión, recomendándose como herramienta exploratoria y no como modelo predictivo primario. La principal limitación es la subrepresentación de hormigones de alta resistencia (>60 MPa) en la base de datos, que restringe la generalización en este rango. Los resultados refuerzan la necesidad de técnicas de interpretabilidad post-hoc, como SHAP y LIME, para conciliar precisión y transparencia en aplicaciones de ingeniería.

Palabras clave: resistencia del hormigón; análisis de componentes principales; equilibrio precisión–interpretabilidad; random forest; modelización predictiva.

1 INTRODUÇÃO

A tecnologia do concreto passa por uma transformação significativa impulsionada pela dupla exigência de sustentabilidade e aprimoramento de desempenho. A busca por concretos de baixo carbono, que utilizam materiais cimentícios suplementares e agregados reciclados, tornou a matriz do material substancialmente mais complexa e variável (Ahmad *et al.*, 2024; Juenger *et al.*, 2011). A identificação de novas fontes de materiais cimentícios suplementares, incluindo argilas calcinadas, pozolanas naturais e cinzas de biomassa, é crucial para sustentar a descarbonização da indústria do cimento (Snellings; Suraneni; Skibsted, 2023). Paralelamente, a incorporação de nanotecnologia e aditivos avançados objetiva desenvolver concretos com funcionalidades específicas, como autocicatrização e monitoramento de integridade estrutural (Monteiro; Moura; Soares, 2022).

Nesse contexto de crescente complexidade, a previsão precisa das propriedades mecânicas do concreto transcende o mero exercício de controle de qualidade, constituindo-se em desafio científico fundamental, quando a compreensão das interações entre múltiplos componentes torna-se imperativa para a inovação tecnológica. O desenvolvimento de modelos preditivos robustos é considerado desafiador devido à diversidade de aditivos e condições de cura que afetam o desempenho do concreto moderno (Jiang *et al.*, 2024).

A resistência à compressão configura-se como a propriedade de engenharia mais crítica do concreto (Mehta, 1999). A busca por modelos preditivos precisos para esta propriedade tem constituído campo

de intensa investigação científica, evoluindo de relações empíricas (Abrams, 1919) para modelos de aprendizado de máquina sofisticados (Asteris *et al.*, 2021). Historicamente, métodos tradicionais basearam-se em fórmulas empíricas, modelos analíticos e ensaios físicos, porém, essas abordagens empíricas têm demonstrado limitações para capturar as relações não-lineares complexas entre os diferentes constituintes do concreto (Elhishi; Elashry; El-Metwally, 2023). Algoritmos como redes neurais artificiais (Yeh, 1998), *gradient boosting* (Ahmad *et al.*, 2022) e *random forest* (RF) (Zhu *et al.*, 2023) consistentemente alcançam coeficientes de determinação (R^2) superiores a 0,90, superando significativamente os modelos estatísticos tradicionais. Técnicas de aprendizado de máquina podem prever a resistência à compressão de materiais cimentícios com boa acurácia e capacidade de aprendizado (Chi *et al.*, 2023).

Contudo, modelos de maior complexidade estrutural, como o RF, podem apresentar menor transparência direta na relação entre variáveis de entrada e saída, dificultando a interpretação imediata das relações físico-químicas subjacentes entre os componentes do concreto (Ribeiro; Singh; Guestrin, 2016). Essa característica tem sido amplamente discutida na literatura como parte da discussão sobre a relação entre precisão preditiva e interpretabilidade em modelos de aprendizado de máquina (Murdoch *et al.*, 2019). Na engenharia de materiais, em que a compreensão dos mecanismos subjacentes é tão importante quanto a quantificação do desempenho, a interpretabilidade dos modelos não constitui atributo secundário, mas assume papel central para o avanço científico e tecnológico.

Técnicas de redução de dimensionalidade, como a análise de componentes principais (*Principal Component Analysis*, PCA) (Jolliffe, 2002), oferecem abordagem promissora para a interpretabilidade. Mediante a transformação de variáveis correlacionadas em conjunto reduzido de componentes não correlacionados, a PCA possibilita a revelação da estrutura dos dados (Kobaka, 2021). A regressão por componentes principais (*Principal Component Regression*, PCR), que utiliza esses componentes como preditores, configura um método para construção de modelos lineares interpretáveis, cuja influência no desempenho preditivo da resistência do concreto tem sido investigada em estudos recentes (Wan; Xu; Šavija, 2021).

A necessidade de estruturas interpretativas para modelos de aprendizado de máquina na predição de propriedades do concreto tem sido recentemente enfatizada, dado que muitos estudos falham em explicar a natureza dos modelos de alta precisão (Zhang *et al.*, 2024). Embora estudos anteriores tenham comparado o desempenho preditivo de diferentes algoritmos de aprendizado de máquina para resistência do concreto, poucos quantificam explicitamente o custo de interpretabilidade associado à

adoção de modelos de alta precisão, nem avaliam sistematicamente a capacidade da PCA em extrair fatores latentes estatísticos nos dados de dosagem, cuja interpretação pode ser fundamentada em princípios da tecnologia do concreto (Nikoopayan Tak; Feng; Mahgoub, 2025; Wan; Xu; Šavija, 2021). Esse artigo parte da seguinte pergunta de pesquisa: em que medida a redução dimensional por PCA preserva a capacidade preditiva e, simultaneamente, oferece interpretabilidade física superior à regressão linear múltipla e ao RF? Para respondê-la, realiza-se uma análise crítica e comparativa entre um modelo de referência (baseline) e três abordagens de modelagem: (1) modelo PCR com foco na interpretabilidade; (2) regressão linear múltipla (*Multiple Linear Regression*, MLR) como referência linear; e (3) modelo RF como referência de alta precisão. É importante delimitar que este estudo não tem como objetivo a otimização de hiperparâmetros, a proposição de novos algoritmos ou a generalização para *datasets* distintos do utilizado; seu escopo é a análise estrutural comparativa das abordagens no contexto específico do *dataset* utilizado nesse estudo. O objetivo central é a análise das restrições inerentes a cada abordagem e de suas consequências para o avanço de modelos interpretáveis, generalizáveis e robustos no campo da engenharia de materiais.

2 MATERIAIS E MÉTODOS

A metodologia deste estudo foi implementada por meio de uma sequência integrada de etapas analíticas e de modelagem, utilizando como base o conjunto de dados de resistência à compressão do concreto disponibilizado por Yeh (1998) no repositório de aprendizagem de máquina da Universidade da Califórnia, Irvine (University of California, Irvine, 1987). Este conjunto contém 1.030 observações de misturas de concreto de alto desempenho, cada uma caracterizada por oito variáveis de entrada quantitativas: cimento (kg/m^3), escória de alto-forno (kg/m^3), cinza volante (kg/m^3), água (kg/m^3), superplastificante (kg/m^3), agregado graúdo (kg/m^3), agregado miúdo (kg/m^3) e idade de cura (dias), além da variável de resposta, resistência à compressão (MPa). Cabe destacar que este *dataset* apresenta limitações inerentes à sua origem experimental. As observações provêm de condições laboratoriais controladas, com concentração de amostras na faixa de resistência entre 20 e 50 MPa e sub-representação de concretos de alta resistência (acima de 60 MPa), o que pode restringir a capacidade de generalização dos modelos para misturas fora dessa faixa. Esse banco de dados foi inicialmente particionado em conjuntos de treinamento (80%) e teste (20%), de modo a assegurar a avaliação imparcial do desempenho dos modelos.

Para a análise exploratória, aplicou-se a PCA sobre as variáveis independentes previamente padronizadas mediante transformação Z-score, etapa essencial para eliminar a influência de diferentes escalas de medida e permitir comparação equitativa entre os constituintes, assegurando média zero e desvio padrão unitário. A análise de componentes principais foi implementada sobre o conjunto de treinamento padronizado, extraindo componentes ortogonais que maximizam a variância explicada dos dados originais (Jolliffe, 2002; Kobaka, 2021). A definição do número de componentes retidos foi guiada pela combinação de três critérios: o critério de Kaiser (autovalor $> 1,0$) (Jackson, 1993), que, dado que os componentes de quarta e quinta ordens explicam conjuntamente 24,6% da variância (autovalor médio inferior a 1,0), indica a retenção de até quatro componentes; a análise do ponto de inflexão do gráfico *scree*; e o desempenho preditivo em validação cruzada. Com base nesse conjunto de critérios, retiveram-se os cinco primeiros componentes que, em conjunto, explicaram aproximadamente 87,4% da variância total. A análise do coeficiente de determinação (R^2) médio obtido por validação cruzada em função do número de componentes revelou que a inclusão do quinto componente resulta em melhora substancial em relação a três componentes ($R^2 = 0,425$ frente a $R^2 = 0,318$), enquanto a inclusão do sexto e sétimo componentes, embora aumente o desempenho preditivo ($R^2 = 0,554$ e $0,587$, respectivamente), implica maior complexidade estrutural e reduz a clareza interpretativa dos fatores latentes. Dessa forma, cinco componentes foram considerados adequados para os objetivos deste estudo, por representarem um compromisso metodologicamente consistente entre parcimônia, interpretabilidade e desempenho preditivo, sem caracterizar uma solução ótima formal, mas uma escolha fundamentada nos critérios adotados.

Na etapa subsequente de modelagem preditiva, foram implementadas e comparadas quatro abordagens distintas. Como referência mínima de desempenho, foi incluído um modelo *baseline*, constituído pela predição da média aritmética da variável resposta no conjunto de treinamento (*DummyRegressor*), cujo desempenho esperado próximo a $R^2 = 0$ estabelece o limiar inferior de aceitabilidade para qualquer modelo preditivo. O primeiro modelo, PCR, utilizou os cinco primeiros componentes extraídos pela PCA como variáveis preditoras em modelo de regressão linear, mitigando problemas de multicolinearidade presentes nas variáveis originais (Wan; Xu; Šavija, 2021). O segundo modelo, MLR, empregado como referência linear, foi treinado com as oito variáveis originais padronizadas. O terceiro modelo, RF, considerado a referência de alta precisão, foi ajustado utilizando as variáveis originais com o intuito de capturar interações não lineares complexas (Khodaparasti; Alijamaat; Pouraminian, 2023). Os hiperparâmetros do modelo RF foram fixados a priori, utilizando a implementação da biblioteca *scikit-learn* (Pedregosa *et al.*, 2011): número de

árvores ($n_estimators$) = 100, número máximo de variáveis consideradas em cada divisão ($max_features$) = $\sqrt{p}$, definido explicitamente (o padrão do `RandomForestRegressor` para problemas de regressão é $max_features = 1.0$, isto é, todas as variáveis de entrada), número mínimo de amostras para divisão de nó ($min_samples_split$) = 2 e número mínimo de amostras por folha ($min_samples_leaf$) = 1. Esta configuração foi adotada deliberadamente, sem otimização de hiperparâmetros, para isolar o efeito da estrutura do modelo na comparação de desempenho entre as abordagens.

O desempenho preditivo de cada modelo foi quantificado mediante o R^2 e a raiz do erro quadrático médio (*Root Mean Square Error*, RMSE), métricas complementares que permitem a análise quantitativa da relação entre interpretabilidade e precisão preditiva (Zhang *et al.*, 2024). A avaliação foi conduzida em duas etapas complementares: (1) particionamento único em conjuntos de treinamento (80%) e teste (20%), preservando a comparabilidade direta com estudos anteriores que utilizaram o mesmo *dataset*; e (2) validação cruzada *k-fold* com $k = 10$, com embaralhamento aleatório controlado ($random_state = 42$), que fornece estimativas de desempenho médio e respectivo desvio padrão entre as partições, permitindo avaliar a estabilidade e a capacidade de generalização de cada modelo. Na validação cruzada, a padronização Z-score e a extração dos componentes principais foram reajustadas exclusivamente sobre os dados de treinamento de cada partição, mediante encadeamento em pipeline, de modo a evitar vazamento de informação do conjunto de avaliação. Esta abordagem dual permite tanto a comparação com a literatura quanto a quantificação da variabilidade das métricas.

3 RESULTADOS E DISCUSSÃO

A análise dos dados inicia-se com a caracterização estatística das variáveis de entrada e da variável de resposta. A Tabela 1 apresenta as estatísticas descritivas do conjunto de dados, evidenciando a ampla variabilidade nas dosagens dos constituintes do concreto. Observa-se que o cimento varia de 102,0 a 540,0 kg/m³, enquanto a resistência à compressão apresenta amplitude de 2,33 a 82,60 MPa, refletindo a diversidade de misturas presentes no *dataset*. Esta heterogeneidade é característica de bancos de dados experimentais de concreto, nos quais diferentes combinações de materiais cimentícios suplementares, aditivos e condições de cura são exploradas para otimização de propriedades mecânicas (Yeh, 1998).

Tabela 1 - Estatísticas descritivas das variáveis de entrada e da variável de resposta do conjunto de dados de concreto de alto desempenho ($n = 1.030$).

	Média	Desvio Padrão	Mínimo	Máximo
Cimento (kg/m^3)	281,17	104,51	102,0	540,0
Escória (kg/m^3)	73,9	86,28	0,0	359,4
Cinza Volante (kg/m^3)	54,19	64,0	0,0	200,1
Água (kg/m^3)	181,57	21,36	121,75	247,0
Superplastificante (kg/m^3)	6,2	5,97	0,0	32,2
Agregado Graúdo (kg/m^3)	972,92	77,75	801,0	1145,0
Agregado Miúdo (kg/m^3)	773,58	80,18	594,0	992,6
Idade (dias)	45,66	63,17	1,0	365,0
Resistência (MPa)	35,82	16,71	2,33	82,6

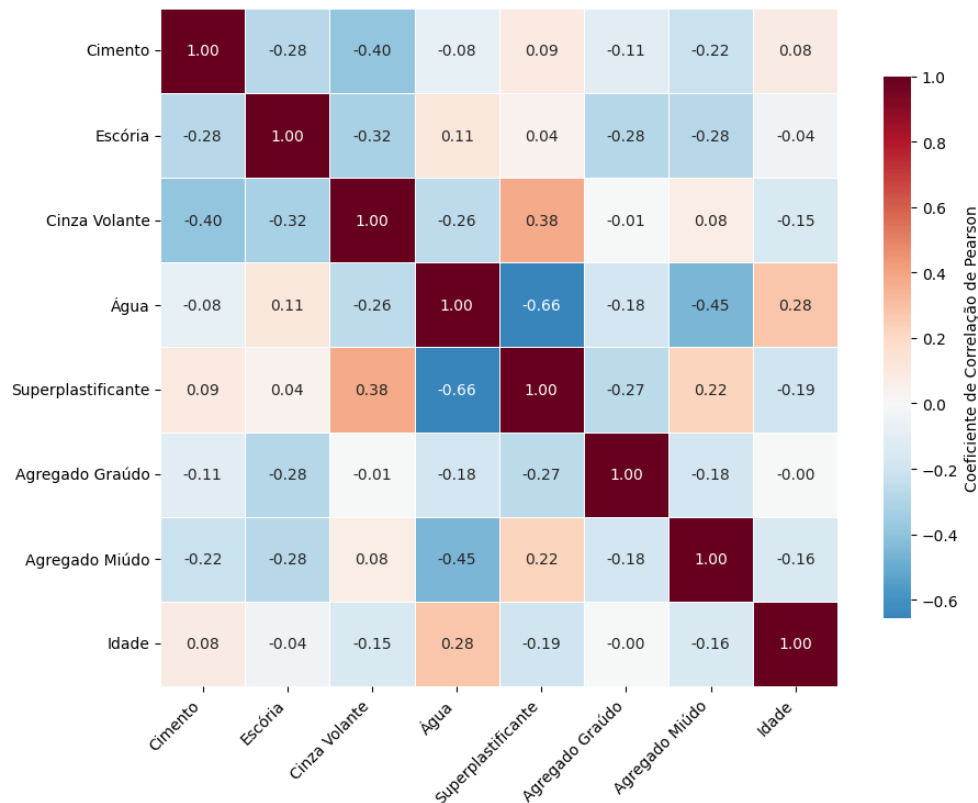
Fonte: Elaborado pelos autores.

A análise de correlação entre as variáveis de entrada, apresentada na Figura 1, revela padrões importantes que justificam a aplicação de técnicas de redução de dimensionalidade. Observa-se correlação negativa fraca entre cimento e escória de alto-forno ($r = -0,28$), refletindo a prática comum de substituição parcial do cimento por materiais cimentícios suplementares. Similarmente, a correlação negativa fraca a moderada entre cimento e cinza volante ($r = -0,40$) evidencia estratégias de dosagem voltadas à sustentabilidade e economia. A correlação negativa de magnitude moderada a forte entre água e superplastificante ($r = -0,66$) corrobora o princípio fundamental da tecnologia do concreto, no qual o aditivo reduz a demanda hídrica da mistura. A presença de multicolinearidade entre algumas variáveis, embora não severa, pode comprometer a estabilidade dos coeficientes em modelos de regressão linear múltipla, reforçando a necessidade de abordagens que mitiguem este problema, como a regressão por componentes principais (Wan; Xu; Šavija, 2021).

A aplicação da PCA sobre as oito variáveis de entrada padronizadas resultou na extração de componentes ortogonais que capturam a estrutura de variabilidade dos dados. A Figura 2a apresenta o gráfico scree, que ilustra a variância explicada por cada componente principal. O critério de Kaiser (autovalor $> 1,0$) indica a retenção de até quatro componentes, ao passo que a análise do ponto de inflexão no gráfico scree e o desempenho preditivo em validação cruzada sustentam a retenção de cinco componentes (Jackson, 1993). Esta redução dimensional de oito para cinco variáveis latentes representa uma compressão significativa da informação, mantendo 87,4% da variabilidade original,

o que é considerado adequado para análises exploratórias e modelagem preditiva em dados de concreto (Kobaka, 2021).

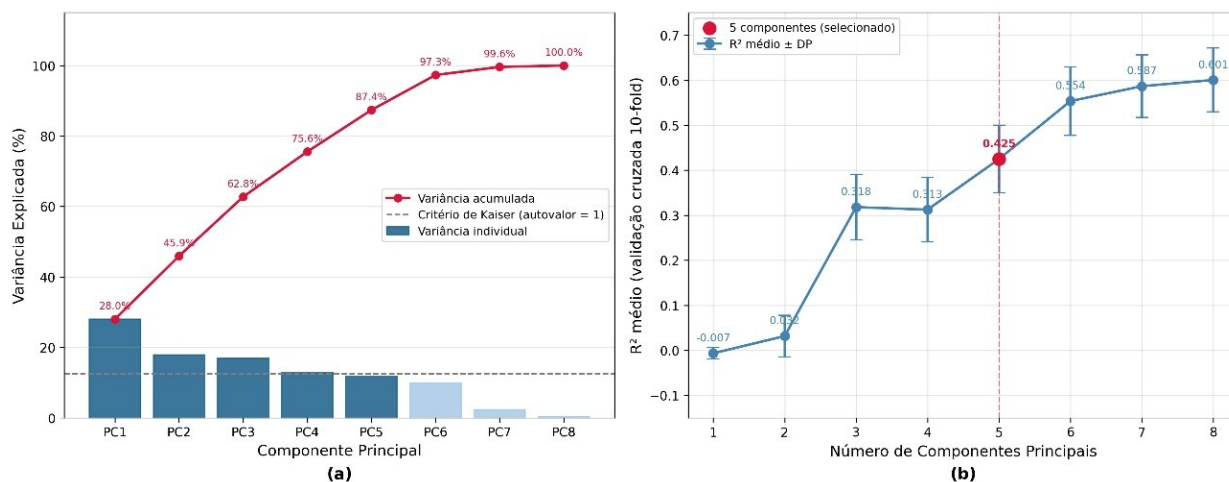
Figura 1 - Matriz de correlação de Pearson entre as variáveis de entrada do conjunto de dados de concreto de alto desempenho.



Nota: Valores positivos (vermelho) indicam correlação direta, enquanto valores negativos (azul) indicam correlação inversa. A intensidade da cor representa a magnitude da correlação.

Fonte: Elaborado pelos autores.

A análise de desempenho preditivo, apresentada na Figura 2b, contextualiza a escolha de cinco componentes como ponto de equilíbrio entre parcimônia e capacidade preditiva: o R^2 médio em validação cruzada aumenta de 0,318 com três componentes para 0,425 com cinco; embora componentes adicionais ainda elevem o R^2 (0,554 com seis), a retenção de cinco preserva o compromisso com a interpretabilidade física dos fatores latentes.

Figura 2 - Análise de componentes principais aplicada ao conjunto de dados de resistência à compressão do concreto.

Nota: (a) Gráfico scree com variância individual e acumulada explicada por cada componente principal; barras em azul escuro indicam os cinco componentes retidos com base no ponto de inflexão da curva acumulada e no desempenho preditivo em validação cruzada, responsáveis por 87,4% da variância total. (b) Desempenho preditivo do modelo PCR em função do número de componentes principais, avaliado por validação cruzada 10-fold; o ponto vermelho destaca a configuração selecionada (cinco componentes, R^2 médio = 0,425), que representa o equilíbrio entre parcimônia, interpretabilidade e capacidade preditiva; as barras de erro indicam o desvio padrão entre as partições.

Fonte: Elaborado pelos autores.

A interpretação física dos componentes principais é realizada mediante a análise das cargas (*loadings*), que quantificam a contribuição de cada variável original para cada componente. A Tabela 2 apresenta as cargas dos três primeiros componentes principais, que em conjunto explicam 62,8% da variância total. Embora cinco componentes tenham sido retidos para a modelagem PCR, estes três primeiros capturam os padrões mais relevantes para interpretação física das misturas. O PC1, responsável por 28,0% da variância total, é dominado por cargas positivas para água (0,54) e cargas negativas para superplastificante (-0,51), cinza volante (-0,40) e agregado miúdo (-0,40), sugerindo que este componente captura o contraste entre misturas com maior teor de água livre e aquelas com maior uso de aditivos redutores de água e materiais suplementares. O PC2, explicando 17,9% da variância, apresenta cargas positivas dominantes para escória (0,71) e cargas negativas para agregado graúdo (-0,59), indicando que este componente está associado à compensação volumétrica do traço: misturas com maior teor de escória, e portanto maior volume total de ligantes, acomodam menor volume de agregado graúdo por metro cúbico, padrão comum em concretos de alto desempenho que empregam materiais cimentícios suplementares em substituição parcial ao cimento. O PC3, com 16,8% da variância, é caracterizado por cargas elevadas para cimento (0,76) e cargas negativas para cinza volante (-0,45) e agregado graúdo (-0,29), refletindo o equilíbrio entre o uso de cimento Portland convencional, materiais pozolânicos e a composição granular da mistura. Esta estrutura

latente é consistente com os princípios fundamentais da tecnologia do concreto, nos quais a resistência mecânica é governada pela relação água/materiais cimentícios, pela estrutura do esqueleto granular e pela reatividade pozolânica dos materiais suplementares (Mehta, 1999; Snellings; Suraneni; Skibsted, 2023).

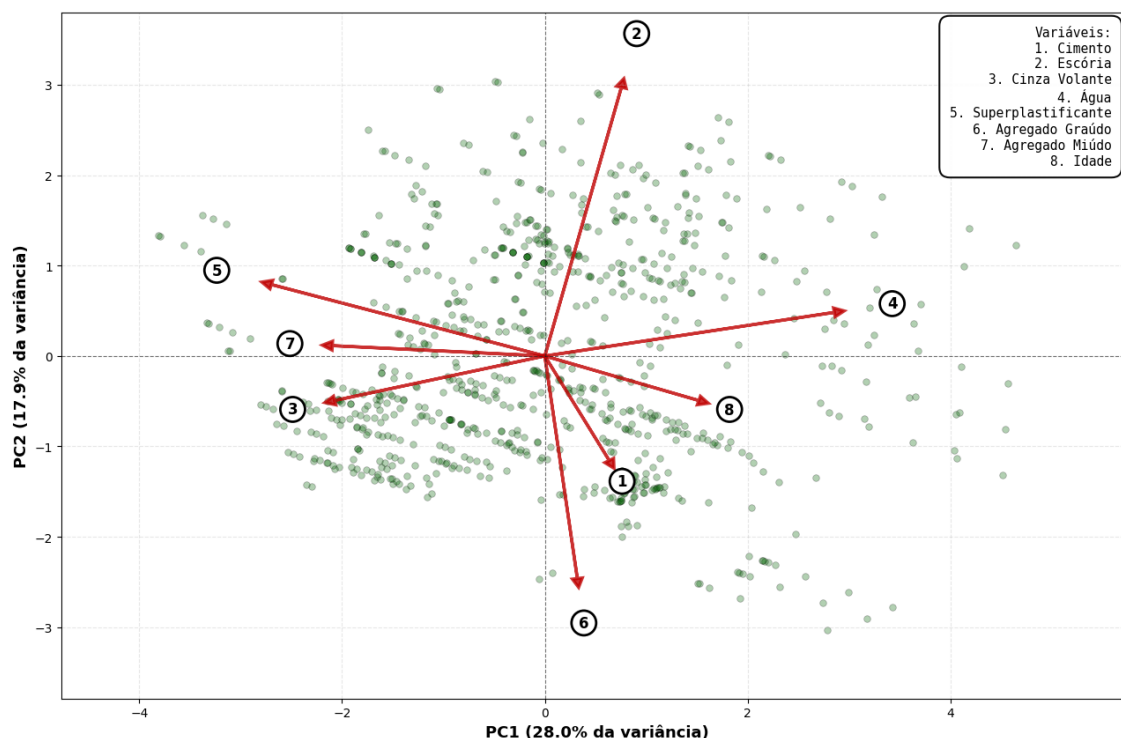
Tabela 2 - Cargas (*loadings*) dos três primeiros componentes principais (PC1, PC2, PC3) para cada variável de entrada.

	PC1	PC2	PC3
Cimento	0,12	-0,28	0,76
Escória	0,14	0,71	-0,04
Cinza Volante	-0,40	-0,12	-0,45
Água	0,54	0,12	-0,22
Superplastificante	-0,51	0,19	0,28
Agregado Graúdo	0,06	-0,59	-0,29
Agregado Miúdo	-0,40	0,03	0,03
Idade	0,29	-0,12	0,1
Variância Explicada (%)	28,02	17,91	16,83

Nota: Valores em negrito indicam as cargas mais relevantes ($|loading| \geq 0,40$) para a interpretação de cada componente. A última linha apresenta a variância explicada por cada componente.

Fonte: Elaborado pelos autores.

A visualização *biplot*, apresentada na Figura 3, permite a interpretação simultânea das observações (pontos) e das variáveis originais (vetores) no espaço reduzido dos dois primeiros componentes principais, que juntos explicam 45,9% da variância total. Os vetores representam as cargas das variáveis originais, cujo o comprimento indica a magnitude da contribuição e a direção aponta para a correlação com os eixos. Observa-se que os vetores de água e superplastificante apontam em direções opostas no PC1, confirmando a interpretação anterior de que este componente captura o contraste entre teor de água livre e uso de aditivos. A escória apresenta forte projeção no PC2, enquanto o cimento domina o PC3 (não visível nesta projeção bidimensional, mas evidenciado na Tabela 2). A dispersão das observações revela a heterogeneidade das misturas de concreto presentes no *dataset*, com alguns agrupamentos visíveis que podem corresponder a diferentes estratégias de dosagem. Esta representação visual facilita a identificação de padrões e a compreensão da estrutura multivariada dos dados, sendo particularmente útil para a interpretação de modelos de redução de dimensionalidade em engenharia de materiais (Kobaka, 2021).

Figura 3 -Biplot dos dois primeiros componentes principais (PC1 e PC2).

Nota: Os pontos verdes representam as observações individuais de misturas de concreto. Os vetores vermelhos representam as variáveis originais, cujo comprimento indica a magnitude da contribuição e a direção indica a correlação com os componentes principais. As porcentagens nos eixos indicam a variância explicada por cada componente.

Fonte: Elaborado pelos autores.

A Tabela 3 apresenta o desempenho comparativo dos quatro modelos preditivos avaliados no conjunto de teste e por validação cruzada 10-fold. O modelo baseline, baseado na predição da média da variável resposta, obteve $R^2 \approx 0$ e RMSE de 16,07 MPa no conjunto de teste (R^2 médio = $-0,006 \pm 0,005$; RMSE médio = $16,67 \pm 1,15$ MPa na validação cruzada), estabelecendo o limiar inferior de referência para avaliação dos demais modelos. O modelo PCR, utilizando cinco componentes principais que capturam 87,4% da variância dos dados originais, alcançou R^2 de 0,48 e RMSE de 11,58 MPa no conjunto de teste (R^2 médio = $0,425 \pm 0,075$; RMSE médio = $12,56 \pm 0,99$ MPa na validação cruzada), demonstrando capacidade moderada de predição com elevada interpretabilidade e redução dimensional significativa (de oito para cinco dimensões). O modelo MLR, treinado com as oito variáveis originais padronizadas, obteve desempenho superior ($R^2 = 0,63$, RMSE = 9,80 MPa no conjunto de teste; R^2 médio = $0,601 \pm 0,071$; RMSE médio = $10,44 \pm 0,91$ MPa na validação cruzada), evidenciando que a retenção de toda a informação original compensa a maior complexidade do modelo, embora a multicolinearidade entre as variáveis possa comprometer a estabilidade dos coeficientes estimados. O modelo RF apresentou o melhor desempenho ($R^2 = 0,88$, RMSE = 5,57

MPa no conjunto de teste; R^2 médio = $0,910 \pm 0,023$; RMSE médio = $4,95 \pm 0,71$ MPa na validação cruzada), capturando relações não-lineares complexas entre os constituintes do concreto e a resistência mecânica. O baixo desvio padrão do R^2 do RF ($\pm 0,023$) indica estabilidade de generalização consistente entre as partições, ao passo que o PCR apresenta maior variabilidade ($\pm 0,075$), refletindo maior sensibilidade à composição das partições de treinamento. Estes resultados são consistentes com a literatura, na qual modelos de conjuntos baseados em árvores de decisão frequentemente superam modelos lineares na predição de propriedades do concreto, com R^2 tipicamente superiores a 0,85 (Khodaparasti; Alijamaat; Pouraminian, 2023; Zhu *et al.*, 2023). Contudo, métricas elevadas de R^2 não constituem, por si só, evidência suficiente de superioridade científica: o RF, embora mais preciso, opera como modelo de caixa-preta, limitando a compreensão dos mecanismos físicos subjacentes e sua aplicação direta em contextos de dosagem racional e controle tecnológico do concreto (Murdoch *et al.*, 2019). A diferença de desempenho entre PCR e RF quantifica o *trade-off* entre interpretabilidade e precisão: um aumento absoluto de 0,40 no R^2 (de 0,48 para 0,88) vem ao custo de menor transparência direta na relação entre variáveis de entrada e saída, enquanto o PCR oferece uma representação parcimoniosa do espaço de variáveis, ao custo de redução na capacidade preditiva, observada na queda do coeficiente de determinação de $R^2 = 0,63$ no MLR para $R^2 = 0,48$ no PCR.

Tabela 3 - Desempenho comparativo dos modelos preditivos no conjunto de teste.

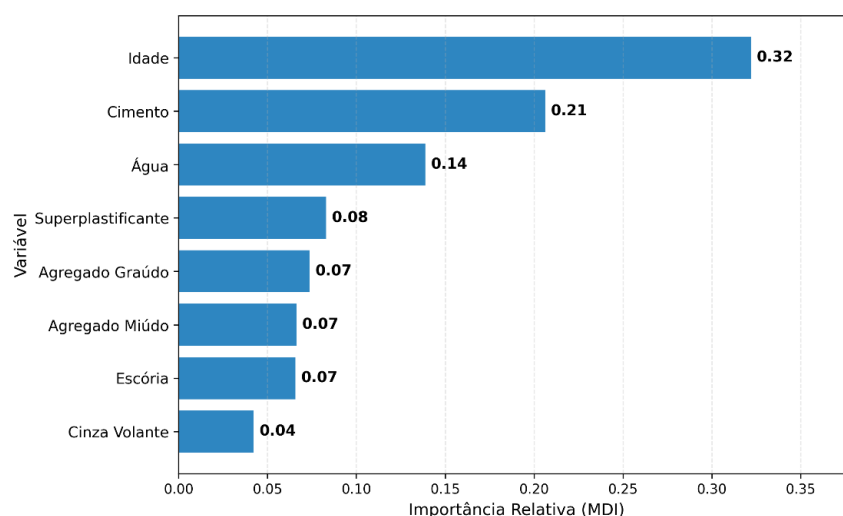
Modelo	Variáveis/Componentes	R^2	RMSE (MPa)
Baseline	Média da resposta	0,00	16,07
PCR	5 componentes (87,4%)	0,48	11,58
MLR	8 variáveis	0,63	9,8
RF	8 variáveis	0,88	5,57

Fonte: Elaborado pelos autores.

A Figura 4 apresenta a importância relativa das variáveis no modelo RF, quantificada pela redução média de impureza (*Mean Decrease in Impurity*, MDI) durante a construção das árvores de decisão. A idade do concreto emerge como a variável mais importante (importância relativa de 0,32), refletindo o papel fundamental do tempo de cura no desenvolvimento da resistência mecânica através da progressão das reações de hidratação do cimento (Mehta, 1999). O teor de cimento ocupa a segunda posição (0,21), confirmando sua função como principal ligante hidráulico responsável pela matriz cimentícia. A água (0,14) apresenta importância intermediária, evidenciando a influência da relação água/cimento na porosidade e microestrutura do concreto endurecido. O superplastificante

(0,08), o agregado graúdo (0,07), o agregado miúdo (0,07) e a escória (0,07) exibem importâncias intermediárias, enquanto a cinza volante (0,04) apresenta a menor importância relativa. Cabe destacar, contudo, que a métrica MDI apresenta vieses conhecidos: variáveis com maior número de categorias ou maior variância tendem a receber importâncias infladas, e variáveis correlacionadas entre si podem ter sua importância subestimada individualmente em favor de uma delas (Murdoch *et al.*, 2019; Strobl *et al.*, 2007). Neste *dataset*, a correlação entre cimento, água e superplastificante pode influenciar a distribuição de importâncias, de modo que os valores apresentados devem ser interpretados como indicadores relativos de relevância preditiva, e não como medidas absolutas de causalidade física. Ressalta-se que, no presente estudo, essa análise teve caráter exploratório, sendo que abordagens alternativas, como a importância baseada em permutação ou métodos pós-hoc de interpretabilidade, podem fornecer estimativas mais robustas e constituem direções metodológicas para trabalhos futuros. Estes resultados são consistentes com estudos prévios que identificam idade e teor de cimento como preditores dominantes em modelos de machine learning para resistência do concreto, com Nikoopayan Tak, Feng e Mahgoub (2025) reportando importâncias de 0,301 para idade e 0,425 para relação água/cimento em modelo de gradient boosting, e Wan, Xu e Šavija (2021) confirmando a idade de cura como feature crítica em múltiplos algoritmos de ML com R^2 superiores a 0,90. Do ponto de vista prático da engenharia, a predominância da idade e do teor de cimento como preditores dominantes reforça a importância do controle do tempo de cura e do traço de cimento em projetos de dosagem racional, indicando que intervenções nessas variáveis têm maior potencial de impacto na resistência final do que ajustes nos agregados ou nas adições minerais secundárias.

Figura 4 - Importância relativa das variáveis no modelo RF, medida pela MDI.

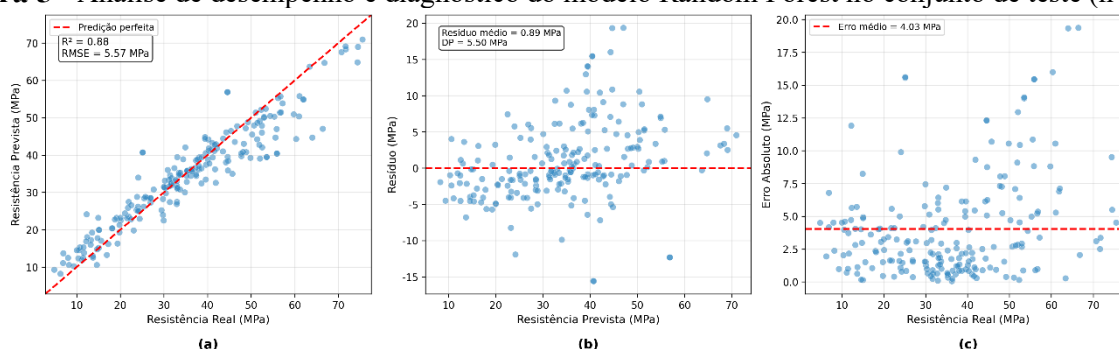


Nota: Valores normalizados para soma unitária. Barras ordenadas em ordem decrescente de importância.

Fonte: Elaborado pelos autores.

A Figura 5 apresenta a análise de desempenho e diagnóstico do modelo RF no conjunto de teste. A Figura 5a evidencia a proximidade dos pontos à linha de predição perfeita, confirmando a elevada capacidade preditiva do modelo ($R^2 = 0,88$, $RMSE = 5,57$ MPa). A Figura 5b apresenta o gráfico de resíduos em função dos valores preditos: os resíduos distribuem-se de forma aproximadamente simétrica em torno de zero (resíduo médio = 0,89 MPa; desvio padrão = 5,50 MPa), sem padrão sistemático evidente de subestimação ou superestimação em função dos valores preditos. Cabe registrar que o resíduo médio, embora pequeno em termos práticos frente à amplitude da variável resposta, é positivo e estatisticamente distinguível de zero (erro padrão de aproximadamente 0,38 MPa para $n = 206$), indicando leve tendência de subestimação que não compromete, contudo, a utilidade do modelo na faixa de resistências avaliada. A Figura 5c apresenta o erro absoluto em função da resistência real observada: o erro absoluto médio é de 4,03 MPa, com tendência de aumento para resistências superiores a 60 MPa. Esse comportamento é consistente com a menor representatividade dessa faixa no conjunto de treinamento, consequência direta das limitações do *dataset* UCI discutidas na seção de Métodos, sugerindo que a maior incerteza nessa região está associada principalmente à menor densidade amostral e não necessariamente a uma limitação estrutural do modelo. Para resistências inferiores a 50 MPa, que correspondem à maior parte das aplicações práticas em estruturas convencionais, o modelo demonstra excelente precisão, com erros absolutos tipicamente inferiores a 5 MPa, magnitude comparável à variabilidade natural da resistência do concreto produzido em obra e, portanto, compatível com aplicações de apoio ao controle tecnológico conduzido conforme a ABNT NBR 12655 (ABNT, 2022). Este desempenho é comparável aos melhores resultados reportados na literatura para modelos de ensemble aplicados à predição de propriedades mecânicas do concreto (Khodaparasti; Alijamaat; Pouraminian, 2023; Zhu *et al.*, 2023).

Figura 5 - Análise de desempenho e diagnóstico do modelo Random Forest no conjunto de teste ($n = 206$).



Nota: (a) Gráfico de dispersão entre resistência à compressão observada e predita; a linha tracejada vermelha representa a predição perfeita ($y = x$). (b) Gráfico de resíduos em função dos valores preditos; a linha tracejada vermelha indica o resíduo zero, referência para ausência de viés sistemático. (c) Erro absoluto em função da resistência real observada; a linha tracejada vermelha indica o erro absoluto médio (4,03 MPa).

Fonte: Elaborado pelos autores.

4 CONCLUSÃO

Este estudo demonstrou a aplicabilidade da PCA como técnica de redução dimensional e interpretação de relações multivariadas em dados de resistência à compressão do concreto, utilizando um *dataset* com 1.030 amostras e oito variáveis de composição. A análise revelou que os cinco primeiros componentes principais capturam 87,4% da variância total dos dados, com os três primeiros componentes interpretáveis capturando 62,8% da variância: PC1 (28,0%) associado principalmente ao contraste entre o teor de água livre e o uso de superplastificante e adições, PC2 (17,9%) refletindo o balanço entre escória e agregado graúdo, e PC3 (16,8%) dominado pelo teor de cimento e cinza volante. O *biplot* confirmou a estrutura de correlações entre as variáveis, evidenciando a oposição entre água e superplastificante, bem como a baixa correlação entre os grupos de variáveis ligantes e agregados, projetados em direções aproximadamente independentes no plano dos dois primeiros componentes.

A comparação de modelos preditivos demonstrou que o PCR com cinco componentes, capturando 87,4% da variância, alcançou R^2 de 0,48 e RMSE de 11,58 MPa no conjunto de teste (R^2 médio = $0,425 \pm 0,075$ na validação cruzada), oferecendo um compromisso razoável entre interpretabilidade e capacidade preditiva. A MLR com as oito variáveis originais obteve desempenho superior (R^2 = 0,63, RMSE = 9,80 MPa; R^2 médio = $0,601 \pm 0,071$ na validação cruzada), enquanto o modelo RF apresentou a melhor performance (R^2 = 0,88, RMSE = 5,57 MPa; R^2 médio = $0,910 \pm 0,023$ na validação cruzada), capturando relações não-lineares complexas entre os constituintes do concreto e a resistência mecânica. A análise de importância das variáveis no modelo RF identificou idade (0,32) e teor de cimento (0,21) como os preditores dominantes, seguidos por água (0,14) e superplastificante (0,08), resultados consistentes com o conhecimento estabelecido sobre os mecanismos de desenvolvimento de resistência no concreto.

A análise de diagnóstico do modelo RF revelou distribuição simétrica dos resíduos, sem viés sistemático (resíduo médio = 0,89 MPa), com erro absoluto médio de 4,03 MPa, embora com maior dispersão para resistências superiores a 60 MPa, indicando maior incerteza na predição de concretos de alta resistência, limitação diretamente associada à sub-representação desta faixa no *dataset* UCI. Para resistências inferiores a 50 MPa, o modelo demonstrou excelente precisão, com erros tipicamente inferiores a 5 MPa, magnitude comparável à variabilidade natural da resistência do concreto produzido em obra e compatível com aplicações de apoio ao controle tecnológico conduzido conforme a ABNT NBR 12655 (ABNT, 2022). Com base nos resultados obtidos, é possível

estabelecer diretrizes práticas de aplicação: o PCR é adequado quando o objetivo principal é a compreensão da estrutura multivariada dos dados, a identificação de fatores latentes fisicamente interpretáveis e a mitigação de multicolinearidade em contextos exploratórios ou de dosagem racional; o PCR não deve ser utilizado como modelo preditivo primário quando a precisão é requisito crítico, pois seu desempenho ($R^2 = 0,48$) é significativamente inferior ao do RF ($R^2 = 0,88$) e mesmo ao da MLR ($R^2 = 0,63$). Reconhece-se, ainda, que a contribuição deste estudo é de natureza comparativa e analítica, sem propor novos algoritmos ou avanços metodológicos originais; seu valor reside na sistematização do trade-off interpretabilidade-precisão e na validação empírica das estruturas latentes identificadas pela PCA no contexto específico do *dataset* UCI.

Conclui-se que o aprofundamento na investigação de técnicas de interpretabilidade aplicadas a modelos de aprendizado de máquina complexos é essencial para conciliar a elevada capacidade preditiva dos modelos de conjunto com o nível de transparência exigido em aplicações de engenharia. Como direções metodológicas concretas para estudos futuros, recomenda-se: (1) a aplicação de técnicas de interpretabilidade pós-hoc, como *SHapley Additive exPlanations* (SHAP) e *Local Interpretable Model-agnostic Explanations* (LIME), que permitem decompor as predições do RF em contribuições individuais de cada variável, conciliando alta precisão com transparência local; (2) a substituição da importância MDI por métricas de importância baseadas em permutação, que são menos suscetíveis aos vieses identificados neste estudo; e (3) a implementação de otimização de hiperparâmetros por busca em grade ou bayesiana, para avaliar o ganho de desempenho além da configuração fixa adotada. Além disso, a ampliação do conjunto de dados, com maior representatividade de concretos de alta resistência (> 60 MPa), e a incorporação de variáveis microestruturais, como porosidade e distribuição do tamanho de poros, podem aprimorar simultaneamente a capacidade preditiva dos modelos e a compreensão dos mecanismos físico-químicos que governam o desenvolvimento da resistência em matrizes cimentícias heterogêneas.

REFERÊNCIAS

- ABRAMS, D. A. **Design of concrete mixtures**. Chicago: Structural Materials Research Laboratory, Lewis Institute, 1919. Disponível em: <https://archive.org/details/designofconcrete00abrarich>. Acesso em: 6 nov. 2025.
- AHMAD, A.; AHMAD, W.; ASLAM, F.; JOYKLAD, P. Compressive strength prediction of fly ash-based geopolymer concrete via advanced machine learning techniques. **Case Studies in Construction Materials**, [s. l.], v. 16, e00840, 2022. DOI: <https://doi.org/10.1016/j.cscm.2021.e00840>.

AHMAD, M. R.; FERNÁNDEZ-JIMENEZ, A.; CHEN, B.; LENG, Z.; DAI, J.-G. Low-carbon cementitious materials: Scale-up potential, environmental impact and barriers. **Construction and Building Materials**, [s. l.], v. 455, art. 139087, 2024. DOI: <https://doi.org/10.1016/j.conbuildmat.2024.139087>.

ABNT. **NBR 12655**: Concreto de cimento Portland — Preparo, controle, recebimento e aceitação — Procedimento. Rio de Janeiro: ABNT, 2022. Disponível em: <https://www.abntcatalogo.com.br/pnm.aspx?Q=U1VhbkmrUi90T3YzSjQ1SDR1UHdkVUZBYIFUdll0Wjl5MXJ4dXRya3RQOD0=#hide1>. Acesso em: 28 fev. 2026.

ASTERIS, P. G.; SKENTOU, A. D.; BARDHAN, A.; SAMUI, P.; PILAKOUTAS, K. Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models. **Cement and Concrete Research**, [s. l.], v. 145, art. 106449, 2021. DOI: <https://doi.org/10.1016/j.cemconres.2021.106449>.

CHI, L.; WANG, M.; LIU, K.; LU, S.; KAN, L.; XIA, X.; HUANG, C. Machine learning prediction of compressive strength of concrete with resistivity modification. **Materials Today Communications**, [s. l.], v. 36, art. 106470, 2023. DOI: <https://doi.org/10.1016/j.mtcomm.2023.106470>.

ELHISHI, S.; ELASHRY, A. M.; EL-METWALLY, S. Unboxing machine learning models for concrete strength prediction using XAI. **Scientific Reports**, [s. l.], v. 13, art. 19892, 2023. DOI: <https://doi.org/10.1038/s41598-023-47169-7>.

JACKSON, D. A. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. **Ecology**, Washington, v. 74, n. 8, p. 2204-2214, 1993. DOI: <https://doi.org/10.2307/1939574>.

JIANG, Y.; ZHANG, J.; ZUO, W.; XU, G.; YUAN, C.; WANG, L.; DU, Z.; LU, Y.; SHE, W. Prediction of time-dependent concrete mechanical properties based on advanced deep learning models considering complex variables. **Case Studies in Construction Materials**, [s. l.], v. 21, art. e03629, 2024. DOI: <https://doi.org/10.1016/j.cscm.2024.e03629>.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer Science+Business Media, 2002. *E-book*. DOI: <https://doi.org/10.1007/b98835>.

JUENGER, M. C. G.; WINNEFELD, F.; PROVIS, J. L.; IDEKER, J. H. Advances in alternative cementitious binders. **Cement and Concrete Research**, [s. l.], v. 41, n. 12, p. 1232-1243, 2011. DOI: <https://doi.org/10.1016/j.cemconres.2010.11.012>.

KHODAPARASTI, M.; ALIJAMAAT, A.; POURAMINIAN, M. Prediction of the concrete compressive strength using improved random forest algorithm. **Journal of Building Pathology and Rehabilitation**, [s. l.], v. 8, art. 92, 2023. DOI: <https://doi.org/10.1007/s41024-023-00337-8>.

KOBAKA, J. Principal Component Analysis as a Statistical Tool for Concrete Mix Design. **Materials**, [s. l.], v. 14, n. 10, art. 2668, 2021. DOI: <https://doi.org/10.3390/ma14102668>.

- MEHTA, P. K. Concrete technology for sustainable development. **Concrete International**, Detroit, v. 21, n. 11, p. 47-53, nov. 1999.
- MONTEIRO, H.; MOURA, B.; SOARES, N. Advancements in nano-enabled cement and concrete: Innovative properties and environmental implications. **Journal of Building Engineering**, [s. l.], v. 56, art. 104736, 2022. DOI: <https://doi.org/10.1016/j.jobbe.2022.104736>.
- MURDOCH, W. J.; SINGH, C.; KUMBIER, K.; ABBASI-ASL, R.; YU, B. Definitions, methods, and applications in interpretable machine learning. **Proceedings of the National Academy of Sciences of the United States of America**, [s. l.], v. 116, n. 44, p. 22071-22080, 2019. DOI: <https://doi.org/10.1073/pnas.1900654116>.
- NIKOOPAYAN TAK, M. S.; FENG, Y.; MAHGOUB, M. Advanced Machine Learning Techniques for Predicting Concrete Compressive Strength. **Infrastructures**, [s. l.], v. 10, n. 2, art. 26, 2025. DOI: <https://doi.org/10.3390/infrastructures10020026>.
- PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COUNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, [s. l.], v. 12, p. 2825-2830, 2011. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>. Acesso em: 28 fev. 2026.
- RIBEIRO, M. T.; SINGH, S.; GUESTIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING (KDD '16), 2016, San Francisco. **Anais [...]** New York: ACM, 2016. p. 1135-1144. DOI: <https://doi.org/10.1145/2939672.2939778>.
- SNELLINGS, R.; SURANENI, P.; SKIBSTED, J. Future and emerging supplementary cementitious materials. **Cement and Concrete Research**, [s. l.], v. 171, art. 107199, 2023. DOI: <https://doi.org/10.1016/j.cemconres.2023.107199>.
- STROBL, C.; BOULESTEIX, A.-L.; ZEILEIS, A.; HOTHORN, T. Bias in random forest variable importance measures: illustrations, sources and a solution. **BMC Bioinformatics**, [s. l.], v. 8, art. 25, 2007. DOI: <https://doi.org/10.1186/1471-2105-8-25>.
- UNIVERSITY OF CALIFORNIA, IRVINE. **UCI Machine Learning Repository**. Irvine: School of Information and Computer Sciences, 1987-. Disponível em: <http://archive.ics.uci.edu/>. Acesso em: 6 nov. 2025.
- WAN, Z.; XU, Y.; ŠAVIJA, B. On the use of machine learning models for prediction of compressive strength of concrete: influence of dimensionality reduction on the model performance. **Materials**, [s. l.], v. 14, n. 4, art. 713, 2021. DOI: <https://doi.org/10.3390/ma14040713>.
- YEH, I. C. Modeling of strength of high-performance concrete using artificial neural networks. **Cement and Concrete Research**, [s. l.], v. 28, n. 12, p. 1797-1808, 1998. DOI: [https://doi.org/10.1016/S0008-8846\(98\)00165-3](https://doi.org/10.1016/S0008-8846(98)00165-3).

ZHANG, Y.; REN, W.; CHEN, Y.; MI, Y.; LEI, J.; SUN, L. Predicting the compressive strength of high-performance concrete using an interpretable machine learning model. **Scientific Reports**, [s. l.], v. 14, art. 28346, 2024. DOI: <https://doi.org/10.1038/s41598-024-79502-z>.

ZHU, J.; FANG, S.; YANG, Z.; QIN, Y.; CHEN, H. Prediction of concrete strength based on random forest and gradient boosting machine. *In*: IEEE 3RD INTERNATIONAL CONFERENCE ON POWER, ELECTRONICS AND COMPUTER APPLICATIONS (ICPECA), 2023, Shenyang, China. **Anais [...]**. [S. l.]: IEEE, 2023. p. 306-312. DOI: <https://doi.org/10.1109/ICPECA56706.2023.10075839>.