XLI CILAMCE
XLI Ibero-Latin American Congress on Computational Methods in Engineering
16-19 November, 2020 FOZ DO IGUAÇU PARANÁ - BRAZIL

Previsão da resistência à compressão do concreto com técnicas de aprendizado de máquina

Pedro B. A. Moreira[1], Victor M. Silva[2]

[1]*Estudante, Engenharia Civil, Universidade Veiga de Almeida*
Avenida das Américas, 22631-004, Rio de Janeiro, Brazil
pedrobermoreira@gmail.com
[2]*Professor, Engenharias, IBMEC/RJ*
Avenida Armando Lombardi 940, 22640-000, Rio de Janeiro, Brazil
victor.silva@professores.ibmec.edu.br

Resumo. A resistência à compressão é a principal característica do concreto. A previsão correta deste parâmetro significa redução de custo e tempo. Este trabalho construiu modelos preditivos para 6 diferentes idades de amostras de concreto. Foi utilizado um conjunto de 1.030 amostras de estudos anteriores, com 9 variáveis. Outras 6 variáveis foram adicionadas para representar as proporções dos ingredientes principais em cada amostra. Os modelos preditivos foram desenvolvidos em linguagem *R*, utilizando o algoritmo *Parallel Random Forest* e a técnica de validação cruzada repetida para otimizar os parâmetros. Os resultados foram compatíveis com outros estudos usando o mesmo conjunto de dados. O modelo mais importante, de 28 dias, obteve um erro quadrático médio (*RMSE*) de 4,717. O modelo de 3 dias obteve o melhor resultado, *RMSE* de 3,310. O trabalho mostrou que a resistência à compressão do concreto pode ser prevista. Criar um modelo para cada idade permitiu obter resultados compatíveis com os dados disponíveis em cada idade. Foi uma alternativa promissora, pois bons resultados foram alcançados treinando com apenas um algoritmo. Este trabalho facilita a exploração e novos esforços para prever a resistência à compressão do concreto, pode ser usado como linha de base para diferentes algoritmos ou a combinação de vários.

Palavras-chave: Concreto, Resistência à compressão, Aprendizado de Máquina, Previsão, *Parallel Random Forest*

1 Introdução

A resistência à compressão é a principal característica do concreto, medida por ensaios de padrões internacionais que consistem na quebra de corpos de prova [1]. A medição aos 28 dias é obrigatória e representa a classe do concreto. Saber com antecedência o resultado em uma determinada idade, com base nas proporções de seus ingredientes, é de grande interesse para fabricantes de concreto, construtoras e engenheiros civis.

A resistência à compressão é uma função não linear de seus ingredientes e idade, dificultando o estabelecimento de uma fórmula analítica, embora algumas já tenham sido propostas por Hasan [2] e Kabir et al. [3]. Porém, a maioria dos estudos construiu modelos incluindo a idade como característica junto com os ingredientes, mas devido à não linearidade entre a resistência à compressão e a idade, constatamos a necessidade de maiores investigações de modelos que separem a idade e analisem apenas os ingredientes.

Portanto, o presente estudo visa a construção de modelos preditivos da resistência à compressão do concreto em diferentes idades utilizando apenas seus ingredientes como características.

2 Trabalhos Relacionados

Yeh [4] demonstrou a possibilidade de usar Redes Neurais Artificiais para prever a resistência à compressão do concreto, concluindo que é um método mais preciso do que os modelos de regressão. Neste estudo, mais de 1.000 amostras de concreto foram coletadas de 17 fontes diferentes. Este conjunto de dados foi posteriormente utilizado em vários estudos sobre concreto, alguns dos quais são mencionados a seguir.

Alshamiri et al. [5] propôs um novo método, *Regularized Extreme Learning Machine* (*RELM*) para treinar modelos de Redes Neurais Artificiais para predizer a resistência à compressão. Os resultados foram comparados

com vários algoritmos conhecidos rodando no mesmo conjunto de dados, incluindo individual e *ensembles*, e o modelo proposto teve o melhor resultado de longe.

Hameed and Khalid [6] compara modelos de Rede Neural Artificial com Regressão Linear Múltipla para prever a força de resistência à compressão e descobriu que os modelos de Rede Neural Artificial obtêm muito mais precisão do que a Regressão Linear Múltipla.

Além desses estudos publicados, agora é muito comum publicar projetos pessoais em páginas da web. Por fácil acesso a este banco de dados e o crescente interesse em ciência de dados e aprendizado de máquina, alguns estudos não publicados usando esse mesmo banco de dados incluem Modukuru [7], Raj [8], Abban [9] e Pierobon [10]. No geral, todos seguiram as etapas padrão no desenvolvimento de modelos de aprendizado de máquina, os dois primeiros usando o pacote *scikit-learn* em linguagem *python* desenvolvido por Pedregosa et al. [11] e os dois últimos usaram o pacote *caret* desenvolvido por Kuhn [12] em linguagem *R* [13].

Ao final deste trabalho, na seção de discussão e conclusão, os resultados encontrados neste trabalho são comparados com todos esses estudos relacionados.

3 Metodologia

3.1 Materiais e reprodutibilidade

A metodologia foi realizada no software *RStudio* [14], um ambiente virtual integrado para desenvolvimento de código em linguagem *R* [13]. Ao longo do processo, todo o código executado foi documentado na mesma ordem de sua execução e enviado para o repositório online hospedado no *github* [15]. O repositório contém uma versão estendida deste trabalho, incluindo todo o código, pacotes necessários e versões. Para garantir a reprodutibilidade, sempre que um código utilizou operações probabilísticas, um *seed* foi definido antes de sua execução, garantindo a consistência dos resultados ao rodar em outra máquina.

3.2 Conjunto de dados

Os dados foram baixados do site da Universidade da Califórnia em Irvine [16]. No total, são 1.030 linhas com 9 colunas. Cada linha representa uma amostra com as variáveis: resistência à compressão, idade e 7 ingredientes (água, cimento, agregado miúdo, agregado graúdo, cinza volante, escória de alto-forno e superplastificantes).

3.3 Preparação dos dados

Os trabalhos relacionados utilizaram o conjunto de dados em sua totalidade ou realizaram um mínimo de preparação. De forma diferente, neste trabalho uma etapa específica foi dedicada apenas a limpar as amostras e prepará-las para as próximas etapas. As principais etapas executadas nesta seção estão listadas abaixo:

1. Amostras duplicadas foram removidas;
2. As amostras foram agregadas e identificadas (com um nova coluna de *ID*) pelas configurações de proporção dos ingredientes, independente de sua idade;
3. Idades de 90, 91 e 100 dias foram fundidas em uma única categoria de 100 dias. Essa etapa foi realizada com o seguinte desenvolvimento: primeiro, traçando e analisando o *boxplot* agrupado por idade na Fig. 1 mostrou que as amostras de 90, 91 e 100 dias têm concentrações distintas de valores de resistência à compressão. Em seguida, uma análise de componentes principais (*PCA*) dos ingredientes foi feita na Fig. 2, mostrando que as combinações de ingredientes dessas idades são muito distintas. Como são idades muito próximas, é razoável que possamos juntar essas idades sem qualquer prejuízo nas previsões. Por fim, a marca dos 100 dias foi escolhida porque a análise de amostras que continham dados em pelo menos cinco idades diferentes mostrou que a resistência à compressão tende a aumentar com o tempo. Portanto, uma idade mais avançada tende a fornecer resultados mais conservadores.
4. Após juntar as idades de 90, 91 e 100 dias, foram retiradas as idades com frequência menor que 50, restando apenas as idades de 3, 7, 14, 28, 56 e 100 dias;
5. Para amostras com o mesmo *ID* e a mesma idade, mas com diferentes valores de resistência à compressão, os dados foram combinados em uma única linha para cada combinação de *ID* e idade, contendo uma média das resistências à compressão;
6. Foram mantidos apenas os *ID*s que possuem dados na marca de 28 dias;
7. Adição de 6 novas variáveis contínuas que foram usados nos modelos de predição, representando as proporções entre os ingredientes principais (água / cimento, agregado miúdo / cimento, agregado graúdo / cimento, agre-

gado miúdo / agregado graúdo, água / agregado graúdo e água / agregado miúdo);

8. Acréscimo de 2 novas variáveis categóricas utilizadas para visualizar a distribuição das amostras (classe do concreto e traço aproximado).

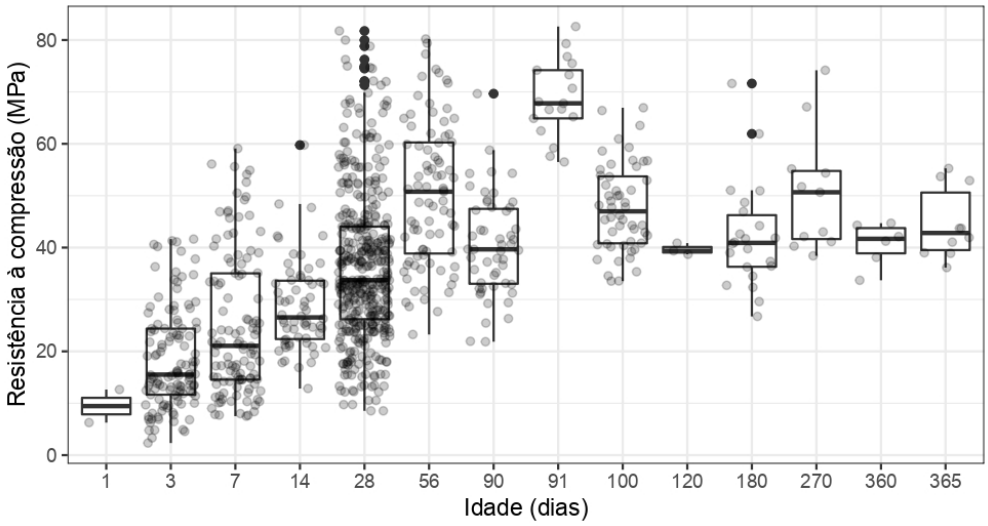

Figure 1. *Boxplot* - resistência à compressão agrupada por idade

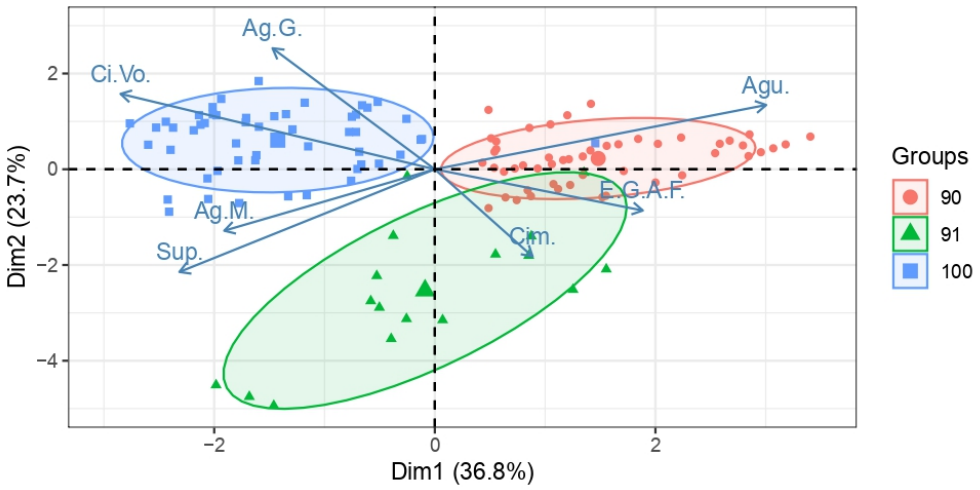

Figure 2. Análise de componente principal - 90, 91 e 100 dias

Após essas manipulações, o número final de amostras foi reduzido de 1.030 linhas para 916 linhas, com 416 configurações de ingredientes diferentes (*IDs*). Um arquivo *xls* dos dados neste momento está disponível para download no repositório no github [15].

3.4 Visualização dos dados

Além dos gráficos já apresentados, vários outros gráficos e tabelas foram construídos para realizar a exploração e visualização das amostras preparadas na etapa anterior, dentre elas:

- Estatística descritiva das variáveis contínuas e categóricas;
- Distribuição das variáveis em relação à resistência à compressão;
- Correlação entre as variáveis agrupadas por idade;

- Relação entre o traço aproximado e a resistência à compressão;
- Relação entre os principais ingredientes do concreto e a resistência à compressão;
- Análise de componentes principais (*PCA*) dos ingredientes.

Os gráficos para da estatística descritiva das variáveis categóricas são apresentados, na Fig. 3.

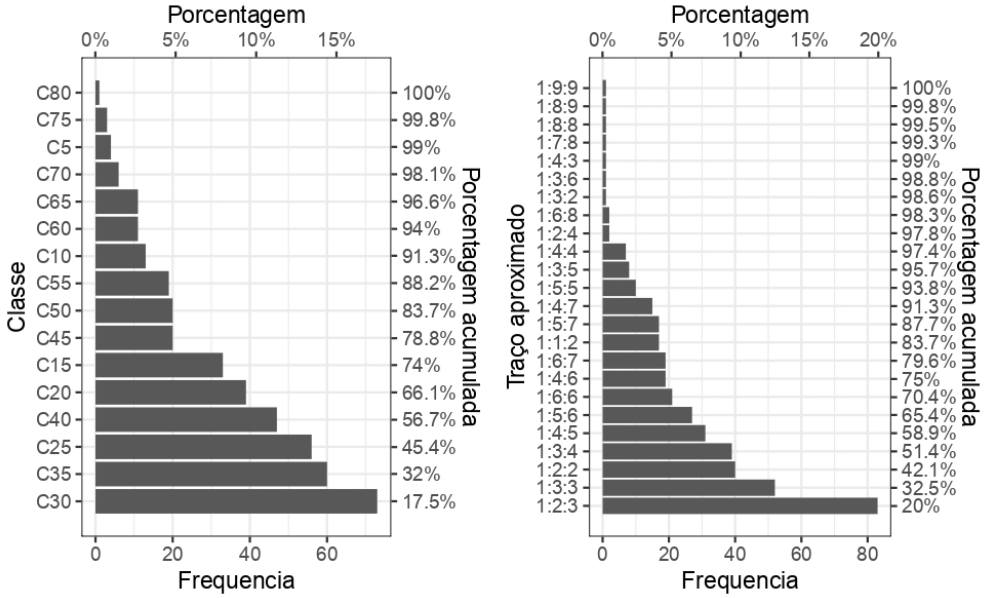

Figure 3. Estatística descritiva das variáveis categóricas

3.5 Pré-processamento e separação dos dados

O principal pacote usado para construir os modelos de aprendizado de máquina foi o pacote *Caret* [12]. Ele fornece todas as funcionalidades e utilitários para construir modelos de previsão para qualquer conjunto de dados, tem uma documentação direta e clara que orienta o processo e fornece cerca de 200 algoritmos diferentes para construir modelos. Neste trabalho, foram realizados os principais passos descritos por Irizarry [17] e Kuhn [18]. Começando por algumas etapas de pré-processamento descritas abaixo:

1. Remoção das variáveis categóricas;
2. Separação do conjunto de dados por idade, resultando em 6 conjuntos de dados menores;
3. Remoção de variáveis com variância próxima de zero (apenas a cinza volante do conjunto de 7 dias foi removida);
4. Verificação de variáveis com correlação acima de 0,999, o que não ocorreu;
5. Cada conjunto de dados foi dividido em conjuntos de teste e treino, 20% e 80% respectivamente, mostrado na Tabela 1 e a distribuição na Fig. 4.

Table 1. Separação das amostras

Modelo	Amostras	Treino (80%)	Teste (20%)
3 dias	121	97	24
7 dias	114	94	20
14 dias	62	50	12
28 dias	416	335	81
56 dias	83	67	16
100 dias	120	96	24

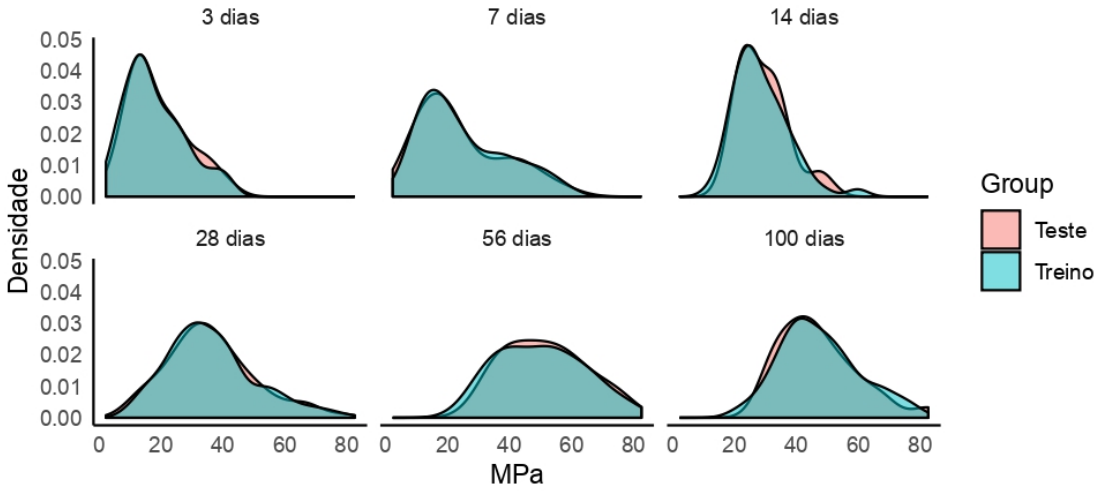

Figure 4. Distribuição dos conjuntos de teste e treino em relação à resistência à compressão

3.6 Modelos ingênuos

Antes de construir os modelos reais, para fins de comparação, foram criados modelos ingênuos. Eles simplesmente prevêem que a resistência à compressão do conjunto de teste é a resistência à compressão média do conjunto de treino. Em outras palavras, os modelos ingênuos são simplesmente a melhor estimativa possível para avaliar o quão próximo / distante o modelo real está de um palpite aleatório.

3.7 Modelos de aprendizado de máquina

Utilizamos apenas um algoritmo neste trabalho, escolhido pela maior probabilidade para atingir o melhor resultado possível, de acordo com Fernandez-Delgado et al. [19], que comparou 179 algoritmos em 121 bancos de dados diferentes, e concluiu que o mais provável de alcançar o melhor resultado possível é o *Parallel Random Forest*, chamado *prRF* no pacote *Caret* [12].

Seis modelos diferentes foram construídos, um para cada faixa de idade, e as seguintes etapas foram feitas para cada modelo:

1. Definir o esquema de reamostragem, com método de validação cruzada repetida;
2. Definir a *tuning grid* para o parâmetro *mtry*, que é uma sequência de 1 ao número de colunas de cada conjunto de dados. Todos, exceto o de 7 dias, são iguais, visto que apenas o conjunto de 7 dias teve uma coluna removida no pré-processamento;
3. Definir o *seed* igual a 1, escolhido arbitrariamente para garantir reprodutibilidade. Esse *seed* pode ser manipulado para obter resultados mais satisfatórios, mas foi optado por não fazê-lo.
4. Fazer a transformação do pré-processamento dos dados com os métodos *center* e *scale*;
5. Executar a função *train* do *Caret* com as configurações acima e modelo *parRF*;

4 Resultados

A avaliação de desempenho dos modelos foi realizada pelo erro de raiz quadrada média (*RMSE*). O *RMSE* é a medida utilizada em todos os trabalhos citados na introdução permitindo a comparação dos modelos na discussão.

O *RMSE* de teste para cada modelo em ordem crescente de idade foi 3,31, 4,36, 4,62, 4,72, 5,94 e 5,85, respectivamente. A tabela 2 apresenta os detalhes e resultados de cada modelo, inclusive o ingênuo. A Fig. 5 compara os valores reais e previstos para os modelos finais.

Table 2. Resultados dos modelos finais

Modelo	*mtry*	*CV*	Repetições	*RMSE* ingênuo (teste)	*RMSE* final (treino)	*RMSE* final (teste)
3 dias	6	30	10	9.303229	3.905196	3.310370
7 dias	2	10	10	13.443646	4.475981	4.361987
14 dias	13	30	10	7.593319	5.136687	4.620515
28 dias	11	30	10	14.283824	5.847334	4.716698
56 dias	8	30	10	12.702112	6.702565	5.939163
100 dias	8	10	10	12.614652	6.381940	5.851088

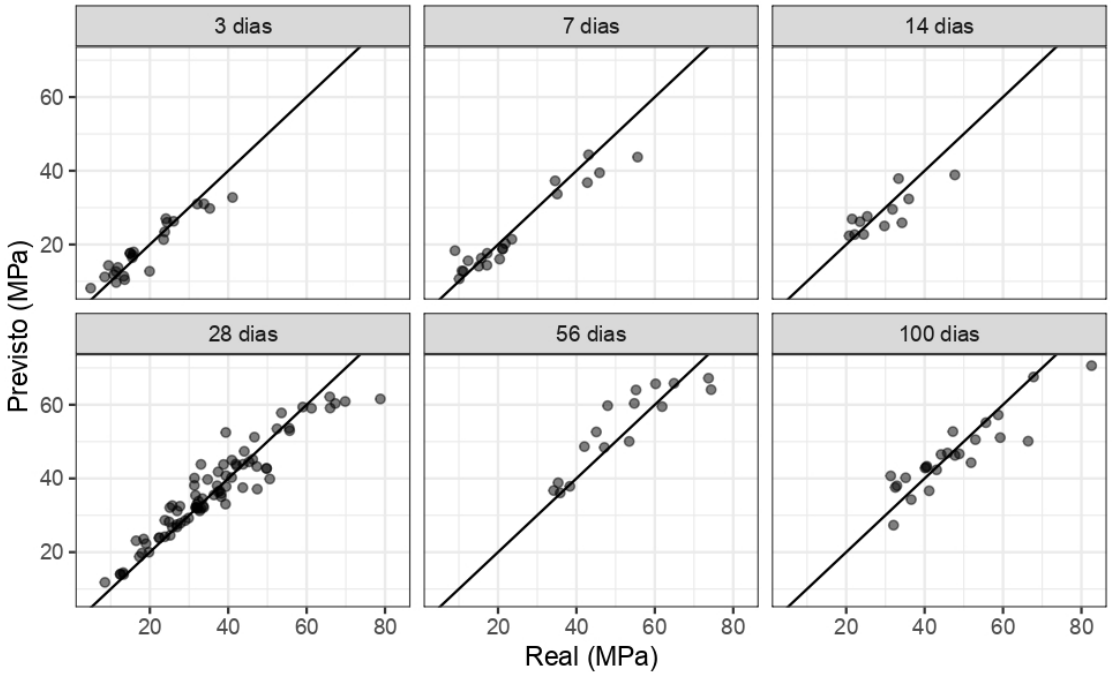

Figure 5. Real *vs* previsto para cada modelo

5 Discussão e conclusão

Os modelos construídos apresentam resultados satisfatórios e comprovam que a resistência à compressão do concreto pode ser prevista com relativa facilidade. A alternativa adotada de criar um modelo para cada conjunto de idades mostrou-se um método válido, ao invés de usar a idade como preditor junto com os ingredientes, como realizado nos estudos relacionados com o mesmo conjunto de dados. A adoção dessa estratificação obteve resultados diferenciados para cada faixa de idade. O *RMSE* calculado em nosso trabalho e o obtido nos trabalhos relacionados foram próximos. A tabela 3 mostra a comparação entre esses estudos e o modelo de 28 dias aqui desenvolvido.

Seguindo a linha de raciocínio deste trabalho, a mesma hipótese pode ser avaliada utilizando outros algoritmos além do aqui utilizado (*Parallel Random Forest*), pois podem apresentar resultados melhores. Outra opção é criar um *ensemble* de vários algoritmos, como realizado por Pierobon [10], mas com a separação dos conjuntos de idades aqui proposta. Além disso, este estudo pode ser reproduzido com um conjunto de dados maior, idealmente com um número semelhante de amostras em cada faixa de idade e uma distribuição mais homogênea de resistência à compressão e classe de concreto, visto que na Fig. 3 mostra que esse conjunto de dados é muito inclinado para a classe de concreto entre C25 e C35.

Table 3. Comparação com outros trabalhos com o mesmo conjunto de dados

Autor	Ano	Algoritmo	*RMSE*	Diferença (%)
Pierobon [10]	2018	*Ensemble* de 5 algoritmos	4.150	-12
Esse trabalho (28 dias)	**2020**	*Parallel Random Forest*	**4.717**	**0**
Hameed and Khalid [6]	2020	Redes neurais artificiais	4.736	0
Raj [8]	2018	*Gradient Boosting Regressor*	4.957	+5
Modukuru [7]	2020	*Random Forest Regressor*	5.080	+8
Alshamiri et al. [5]	2020	*Regularized Extreme Learning Machine*	5.508	+17
Abban [9]	2016	*SVM* com *Radial Basis Function Kernel*	6.105	+29

Declaração de autoria. Os autores confirmam que são os únicos responsáveis pela autoria deste trabalho, e que todo o material aqui incluído como parte do presente trabalho é propriedade (e autoria) dos autores, ou tem a permissão dos proprietários a serem incluídos aqui.

Referêcias

[1] Gambhir, M., 1990. *Concrete Technology: Theory and Practice*. Tata McGraw-Hill Education, 5e edition.
[2] Hasan, Md e Kabir, A., 2011. Prediction of compressive strength of concrete from early age test result. In *Conference: 4th Annual Paper Meet and 1st Civil Engineering Congress, At Dhaka, Bangladesh*.
[3] Kabir, A., Hasan, M., & Miah, M., 2012. Predicting 28 days compressive strength of concrete from 7 days test result. In *International Conference on Advances in Design and Construction of Structures (ADCS 2012)*, pp. 18–22.
[4] Yeh, I.-C., 1998. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, vol. 28, pp. 1797–1808.
[5] Alshamiri, A., Yuan, T.-F., & Kim, a., 2020. Non-tuned machine learning approach for predicting the compressive strength of high-performance concrete. *Materials*, vol. 13, pp. 1023.
[6] Hameed, M. & Khalid, M., 2020. *Prediction of Compressive Strength of High-Performance Concrete: Hybrid Artificial Intelligence Technique*, pp. 323–335. Spriger.
[7] Modukuru, P., 2020. Concrete compressive strength prediction. https://towardsdatascience.com/concrete-compressive-strength-prediction-using-machine-learning-4a531b3c43f3.
[8] Raj, P., 2018. Predicting compressive strength of concrete. https://www.kaggle.com/pavanraj159/predicting-compressive-strength-of-concrete.
[9] Abban, D., 2016. Concrete compresive strength. https://rpubs.com/brother_abban/220101.
[10] Pierobon, G., 2018. A comprehensive machine learning workflow with multiple modelling using caret and caretensemble in r. https://towardsdatascience.com/a-comprehensive-machine-learning-workflow-with-multiple-modelling-using-caret-and-caretensemble-in-fcbf6d80b5f2.
[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, vol. 12, n. Oct, pp. 2825–2830.
[12] Kuhn, M. e. a., 2020. *caret: Classification and Regression Training*. R package version 6.0-86.
[13] R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
[14] RStudio Team, 2020. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.
[15] Moreira, P. B. A., 2020. Repository. https://github.com/PedroBern/concrete-compressive-strength-prediction.
[16] Yeh, I.-C., 2008. Concrete compressive strength data set archive to download. https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength.
[17] Irizarry, R., 2019. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. Chapman & Hall/CRC Data Science Series. CRC Press.
[18] Kuhn, M., 2008. Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, vol. 28, n. 5.
[19] Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181.