Uso de Inteligência artificial para determinação da resistência à compressão no concreto simples

A. I. de Campos

Reusmo--O presente trabalho tem como objetivo aplicar algoritmos de Mineração de Dados em um conjunto de dados conhecido como *Concrete Compressive Strength Dataset*, dessa forma é possível prever a resistência à compressão do concreto a partir da sua idade e dos demais ingredientes que compõem a mistura. Trabalhos anteriores analisaram uma parcela desse conjunto de dados utilizando técnicas de Redes Neurais. Portanto a proposta deste trabalho engloba fazer análises com diversos algoritmos de Mineração de Dados baseadas em aprendizado supervisionado de forma a ampliar a base de algoritmos utilizadas para esse tipo de problema além de que modelos desse tipo podem contribuir no controle de qualidade do concreto produzido. O conjunto de dados selecionado possui uma 1030 registros com oito variáveis de entrada relativas à dosagem do material e a saída é dada como à resistência a compressão do concreto simples (f_{ck}). O conjunto de dados será dividido com a técnica de validação cruzada do tipo *KFold* e diferentes modelos de regressão foram empregados como por exemplo: Regressão Linear, Regressão LASSO, Regressão de Ridge, K Vizinhos mais próximos (KNN), *AdaBoost*, Árvores de Decisão, Florestas aleatórias, *GradienteBoosting*. Os métodos se mostraram satisfatórios para o problema de previsão da resistência do concreto sendo que o modelo que melhor se adequo foi o *GradienteBoosting* com um valor de raiz do erro médio quadrático igual a 5,294.

***Palavras-chave*—Concreto, Inteligência Artificial, Regressão.**

I. INTRODUÇÃO

A RESISTÊNCIA, à compressão do concreto simples (f_{ck}) é uma das variáveis mais importantes para o projeto de estruturas de concreto, sendo que o conhecimento desse atributo é essencial no controle de qualidade e segurança dos sistemas estruturais produzidos. Por norma o f_{ck} é definido como a medida de resistência à compressão que representa 95% de grau de confiança, sendo exemplos de utilização dessa medida: (a) A determinação do tempo de retirada de escoras no planejamento da construção de uma estrutura; e (b) Principal indicador a nível de projeto para dimensionamento das peças estruturais.

A resistência à compressão (f_{ck}) é uma medida determinada experimentalmente através do ensaio de compressão de corpos-de-prova cilíndricos descrito na NBR 5739 [1]. Porém, essa medida será influenciada por diversos fatores como relação água/cimento, tipo de cimento, corpo de prova, velocidade de ensaio, etc. Devido a essa grande quantidade de parâmetros o controle de qualidade desse atributo é uma tarefa complexa visto que em outras variações do concreto convencional uma maior quantidade de parâmetros é adicionada ao problema, como por exemplo no Concreto com Fibras (CRF) e o Concreto de Alto Desempenho (CAD), sendo o último a temática foco deste artigo.

O Concreto de Alto Desempenho possui uma versatilidade de aplicações e permite uma série vantagens a nível de projeto e execução, como: (a) Redução da taxa de armadura; (b) Redução das dimensões das seções dos elementos estruturais e consequentemente redução do peso da estrutura; (c) Facilidade de colocação e compactação do concreto nas formas devido ao emprego de aditivos plastificantes; (d) Aumento da impermeabilidade e resistência mecânica do concreto devido a introdução de aditivos minerais reativos [2]–[4].

Portanto a utilização de técnicas que possibilitem a predição de valores de resistência do concreto do tipo CAD, especialmente na situação de compressão, são técnicas fundamentais para redução de erros e gastos. Logo a utilização de ferramentas de Mineração de Dados pode ser útil para esse tipo de situação, de forma que predições assertivas resultam em um controle de qualidade mais adequado impedindo situações graves de reparo estrutural em estruturas de f_{ck} inadequado.

Neste trabalho será utilizado as técnicas de predição para estimar o valor de uma variável numérica a partir da de técnicas de Mineração de Dados. A proposta deste trabalho engloba propor análises com algoritmos de inteligência computacional baseados em regressões.

O trabalho está dividido em seis seções onde nas quatro primeiras são apresentadas introduções sobre a determinação da resistência em concretos de alto desempenho como também nos algoritmos com modelos de regressão.

II. UMA BREVE DESCRIÇÃO DO CONCRETO DE ALTO DESEMPENHO

Para aplicação de qualquer tipo de concreto, seja ele CAD ou não, a fase de proporcionamento é essencial. A fase de proporcionamento ou dosagem consiste na fase de mistura dos materiais que compõem o compósito concreto simples, sendo que o produto final após a execução de um método de dosagem é o traço.

O estudo da dosagem é essencial para obras de pequeno e grande porte. Tutikian e Helene [5] afirmam que um estudo de dosagem visa obter a mistura ideal e mais econômica, numa determinada região e com os materiais ali disponíveis, de forma que essa mistura atenda aos requisitos relacionados em projeto.

Em termos de dosagem os procedimentos são os mais variados, desde formulações teóricas universais como os métodos do ACI (*American Concrete Institute*) e ABCP (Associação Brasileira de Cimento Portland), até procedimentos experimentais mais complexos como o método do empacotamento utilizados para o CAD.

O CAD normalmente é composto dos mesmos materiais que o Concreto Convencional, porém alguns materiais são adicionados a essa mistura de forma a caracterizar o CAD. No caso essas adições consistem na colocação das adições minerais e aditivos químicos plastificantes.

Nas adições minerais Tutikian *et al.* [6] afirma que é comum aplicações de materiais como a pozolanas naturais, cinza volante, escória de alto forno e sílica ativa. Já os aditivos químicos utilizados são os de plastificação da pasta, que permitem uma maior fluidez no CAD sem aumento do consumo de água e cimento.

Além das introduções do aditivo plastificante e adição mineral reativa Guimarães [2] afirma que cabe uma importante observação a respeito do emprego dos agregados no estudo do CAD. Nessa tipologia de concreto a pasta é o elemento de maior resistência diferentemente do concreto convencional onde os agregados têm resistência superior a pasta. Portanto devido a essa situação a escolha da qualidade do agregado é fundamental para utilização do CAD.

III. TÉCNICAS PARA EXTRAÇÃO DE CONHECIMENTO DE UMA BASE DE DADOS

Mineração de Dados (ou *Data Mining*) é o processo pelo qual programas de computadores são usados para pesquisar repetidamente padrões utilizáveis em uma base de dados que é normalmente extensa [7].

O processo de obtenção de conhecimento a partir de uma base de dados é muitas vezes denominado como KDD (*Knowledge Discovery in Databases*). A Fig. 1 detalha o processo de KDD que é a fundamentação básica para a Mineração de Dados.

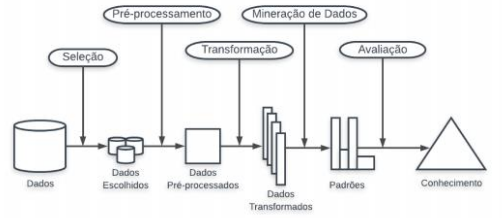

Figura 1. Passos para obtenção de conhecimento a partir de uma base de dados [8], [9].

Problemas de Predição e Descrição são exemplos de técnicas de Mineração de Dados [10]. Nesse trabalho o foco é dado a um problema de predição da resistência à compressão do concreto.

O problema de predição através de uma regressão consiste na descoberta de uma função preditiva de forma similar à feita em classificação, mas com o objetivo de calcular um valor numérico real ao invés de obter uma classe discreta [9]–[11].

Dessa forma o problema preditivo pode ser escrito no formato de um problema inverso dada uma saída quais os parâmetros do modelo que melhor se ajustam aquela curva. Portanto dessa forma os problemas de Mineração de dados que utilizam o processo de regressão vão se utilizar de um problema inverso ao longo do processo iterativo conforme equação (1) onde b_{obs} são as medidas observadas e b_{num} as medidas numéricas obtidas via modelo de predição. Essa função de ajuste também é chamada de função objetivo.

$$min \, \|b_{num} - b_{obs}\| \tag{1}$$

Regressão Linear Múltipla

A eq. (2) apresenta o modelo de predição do conjunto de dados via aproximação linear. Já a eq. (3) apresenta a função de problema inverso utilizada para determinação dos pesos w_i que permitem a melhor aproximação entre conjunto de dados e modelo.

$$y_{num,i}(w_i, x_i) = w_0 + w_1.x_1 + \cdots + w_{n-1}.x_{n-1} + w_n.x_n \tag{1}$$

$$min \, = \frac{1}{m}\sum_{i=1}^{m}(y_{num,i} - y_{obs,i})^2 \tag{2}$$

Regressão Lasso e Ridge

Para evitar que um algoritmo de regressão tenha sobreajuste em relação ao conjunto de dados utilizasse uma técnica de regularização para correção da função de ajuste (nesse trabalho dada pela eq. 2). Para isso introduz-se um coeficiente de penalização na função objetivo do problema de ajuste. As eq's. (3) e (4) apresentam o modelo da função objetivo do problema de mineração de dados para regressão Ridge e regressão Lasso respectivamente [12]–[14].

$$min \, = \frac{1}{m}\sum_{i=1}^{m}(y_{num,i} - y_{obs,i})^2 + \alpha.\sum_{j=1}^{p} {w_j}^2 \tag{3}$$

$$min \, = \frac{1}{m}\sum_{i=1}^{m}(y_{num,i} - y_{obs,i})^2 + \alpha.\sum_{j=1}^{p} |w_j| \tag{4}$$

Onde α representa um parâmetro de penalização que deve ser R^+.

Regressão KNN

Originalmente o método KNN (*K-Nearest Neighbor*) é utilizado como um algoritmo de classificação. Este algoritmo utiliza-se de modelos de agrupamento para determinação de uma classe discreta. Nesse tipo de algoritmo a ideia geral é encontrar um valor discreto (ou classe) com base nos valores rotulados mais próximos.

A métrica de distância entre as partículas é dada por uma função de distância. Essa função de distância pode ser dada de várias formas, sendo a forma geral denominada como distância Minkowski de ordem p.

$$d(x, y) = \left(\sum_{i=1}^{k}|x_i - y_i|^p\right)^{1/p} \tag{5}$$

Para $p = 1$ temos a distância de Manhattan e para $p = 2$ a distância Euclidiana. Para o caso geral de geração de um modelo numérico para a regressão do tipo KNN o valor de numérico gerado será dado pela e. (6).

$$y_{num,i}(w_i, x_i) = \frac{\sum_{i=1}^{k} w_i.y_{obs}(x \in \mathcal{N}_k(x_i))}{\sum_{i=1}^{k} w_i} \tag{6}$$

Onde k indica o número de vizinhos selecionados para comparação e w_i os pesos de contribuição de cada parâmetro ou atributo do problema.

Árvores de regressão

Os modelos de árvore de regressão foram introduzidos por Breiman *et al.* [15]. Em linhas gerais o procedimento de árvores de regressão consiste na divisão do espaço i-dimensional, criado pelas i variáveis preditoras [16].

Segundo Garcia [17], as Árvores de Decisão são constituídas de nós, que representam os atributos, e de ramos, provenientes desses nós e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível

valor desse atributo). Nas árvores existem nós-folha (folha da árvore), que representam os diferentes valores de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de regressão. As Árvores de Decisão podem ser representadas como conjuntos de regras do tipo "se-então". As regras são escritas considerando o trajeto do nó raiz até uma folha da árvore.

Já as Florestas Aleatórias são um conjunto de Árvores de Decisão. Em geral, as Florestas Aleatórias atingem boa acurácia preditiva quando comparadas a outros métodos de aprendizado de máquina supervisionados [18].

Regressão AdaBoost

O algoritmo de regressão *AdaBoost* inicia-se com um método ou algoritmo para encontrar as previsões ruins. O algoritmo de *Boosting* chama esse algoritmo de aprendizado repetidamente, cada vez alimentando-o com um diferente subconjunto dos exemplos de treinamento. Cada vez que é chamado, o algoritmo de aprendizado gera uma nova hipótese de predição fraca, e depois de muitas execuções, no caso da tarefa de classificação, o algoritmo combina essas hipóteses fracas dentro de uma única hipótese de predição que, de acordo com o esperado, será mais precisa que qualquer outra gerada anteriormente [19].

A medida da discrepância é que vai permitir dizer se o valor previsto é ou não aceitável, define-se para isto uma função de perda. De acordo com Hastie *et al*. [20] utilizando a função de perda exponencial eq. (7), o algoritmo *Adaboost* apresenta um melhor desempenho.

$$L_t(x_i) = 1 - exp\left(-\frac{|y_{num,i} - y_{obs,i}|}{max|y_{num,i} - y_{obs,i}|}\right) \quad (7)$$

Regressão Gradient Boosting

O método de regressão *Gradient Boosting* também é uma combinação de métodos de aprendizado "fracos" que resulta em um método forte, treinando vários modelos de árvores de decisão de forma gradual, aditiva e sequencial. A função de perda é uma medida que indica quão bons são os coeficientes do modelo na adequação dos dados subjacentes. Dessa forma, depois de cada execução os algoritmos que obtiveram uma acurácia maior recebem um maior peso enquanto os que não conseguiram prever bem os valores recebem pesos menores, depois de várias iterações é obtido um algoritmo adaptado para prever a variável de saída de determinado modelo.

IV. MATERIAIS E MÉTODOS

O desenvolvimento desse artigo se deu inicialmente pela escolha de um *dataset* de referência. O *dataset* foi obtido via plataforma UCI e se trata da base de dados de Yeh [21], [22]. Esse banco de dados contém 1030 registros relativos a Concretos de Alto Desempenho (CAD).

O *dataset* contempla oito variáveis (unidade padrão *kg/m*^{3} e *dias*) de entrada relativas à dosagem e uma variável de saída relativa à resistência característica do concreto (f_{ck}), totalizando então um *dataset* 1030 linhas com 9 colunas.

Para aplicação da modelagem numérica de Mineração de Dados foi empregada a linguagem Python 3.0 com o uso das seguintes bibliotecas: (a) Scikit-learn; (b) Pandas; (c) Matplotlib; (d) Seaborn e (e) Numpy.

O problema estudado consiste em um caso de predição onde serão empregados métodos de aprendizado de máquina supervisionado para determinação de uma variável contínua no caso um parâmetro do material.

Pré-processamento

A fase de pré-processamento tem como objetivo preparar a base de dados para eventuais simulações. Nessa fase verificou-se a quantidade de espaços vazios do conjunto de dados, distribuição das variáveis ou atributos de entrada, como também a correlação entre as variáveis. No caso desse *dataset* o conjunto de dados não tinham valores vazios. Como o conjunto de dados possui um intervalo diferente para cada atributo (ou variável) de entrada foi aplicado o procedimento de normalização *Z-Score*.

Predição

O conjunto de dados foi separado em para treino e teste, sendo 70% dos registros para treino e 30% para teste e os algoritmos empregados foram: Regressão Linear, Regressão LASSO, Regressão de Ridge, K Vizinhos mais próximos (KNN), AdaBoost, Árvores de Decisão, Florestas aleatórias, GradienteBoosting. Os resultados foram comparados com a separação do conjunto de dados com a validação cruzada. Todos os modelos já são implementados na biblioteca *Scikit-learn*.

Validação Cruzada e Ajuste dos parâmetros

O método de validação cruzada divide aleatoriamente o conjunto de treinamento em *K-folds* (neste caso K = 10), ou seja, divide em 10 subconjuntos distintos e então treina e avalia o modelo 10 vezes escolhendo uma parte diferente de cada uma delas para avaliação e treinando nas outras 9 partes [23] [24].

Esta técnica é amplamente empregada em problemas em que o objetivo da modelagem é a predição. Busca-se então estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados [25].

Após a execução de todos os modelos descritos anteriormente é necessário realizar um ajuste do modelo final, então os hiper parâmetros do melhor algoritmo encontrado para este conjunto de dados serão variados visando melhorar seu desempenho, todas as combinações de valores serão testadas, para isso será utilizado o *GridSearchCV* do Scikit-Learn para realizar esta busca automaticamente.

Estatísticas de avaliação

A avaliação do desempenho de cada um dos modelos descritos anteriormente será realizada com estatísticas de validação como a raiz quadrada do erro médio quadrático *REMQ* (em inglês RMSE) que corresponde à raiz quadrada da média da diferença entre os valores estimados e os valores reais ao quadrado, como apresentado na eq. (8), ou seja, quanto

menor o valor do *REMQ* melhor é o modelo. Os modelos também serão avaliados pelo coeficiente de determinação R^2, que indica o quão próximos os dados estão do modelo ajustados, é calculado com a eq. (9), sendo, portanto, um valor entre 0 e 1, onde valores próximos de 1 indicam que o modelo melhor se ajusta à amostra.

$$REMQ = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y_{num,t} - y_{obs,t}\right)^2} \quad (8)$$

$$R^2 = \frac{\sum_{t=1\ldots N}(y_{num,t} - y_{obs,t})^2}{\sum_{t=1\ldots N}(y_{obs,t} - \overline{y_{obs}})^2} \quad (9)$$

V. RESULTADOS E DISCUSSÃO

Nessa seção são apresentados os resultados e discussões a respeito das simulações descritas.

Pré-processamento

A Fig. 1 indica as estatísticas básicas do conjunto de dados como: número de registros (*count*), média (*mean*), desvio padrão (*std*), valor mínimo (min), o percentil 25% (primeiro quartil), percentil 50% ou mediana (segundo quartil), o percentil superior a 75% (terceiro quartil), e o valor máximo (max), estes dados são importantes para descrever e entender a forma da distribuição dos conjuntos de dados e possíveis outliers.

	cimento	escória	cinzas	água	superplastificante	ag_grosso	ag_fino	idade	resistência
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.167864	73.895825	54.188350	181.567282	6.204660	972.918932	773.580485	45.662136	35.817961
std	104.506364	86.279342	63.997004	21.354219	5.973841	77.753954	80.175980	63.169912	16.705742
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.330000
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.710000
50%	272.900000	22.000000	0.000000	185.000000	6.400000	968.000000	779.500000	28.000000	34.445000
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.135000
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.600000

Figura 2. Estatísticas básica do conjunto de dados.

Uma forma de analisar a relação entre as variáveis em um conjunto de dados é construir a chamada matriz de correlação, onde nos eixos verticais e horizontais estão dispostas as variáveis e em cada ponto de plano cartesiano é a relação entre as variáveis do eixo horizontal com a variável do eixo vertical, portanto, é uma matriz quadrada com diagonal principal toda igual a 1 e simétrica. Na Fig. 3 está indicado a matriz de correlação para o conjunto de dados em estudo, onde cada número indica coeficiente de correlação entre duas variáveis, neste caso o coeficiente de correlação foi transformado para visualização em um mapa de cores para melhor visualizar os resultados. Observa-se que o maior valor de correlação está entre as variáveis água e superplastificante com o coeficiente igual a -0,60, o que indica que estas variáveis apresentam uma correlação inversa. Existem também uma correlação de 0,50 entre resistência e cimento.

Ambos os casos informados acima são de fato confirmados experimentalmente pois à medida que aumenta a quantidade de água na mistura a necessidade por superplastificante é reduzida. Outro fator comprovado experimentalmente é também a forte correlação entre o cimento e a resistência, neste último caso uma correlação positiva.

Observa-se que a maioria dos valores estão proximos de cores claras o que indica um coeficiente de correlação próximo a zero, ou seja, as variáveis nestes casos não estão relacionadas ou têm uma correlação fraca. O que quer dizer que as informações não são redundantes tornando isso um fator que torna interessante a aplicação de uma técnica de Mineração de Dados.

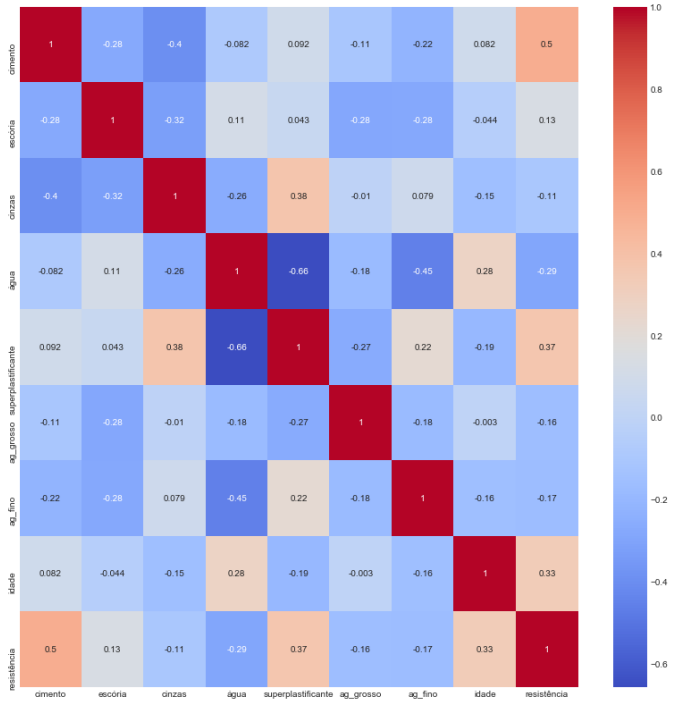

Figura 3. Matriz de correlação dos atributos com mapa de cores.

A Fig. 4 indica os histogramas de cada atributo do conjunto de dados. Observa-se que os atributos 'cinzas', 'escória', 'idade' e 'superplastificante' apresentam uma distribuição sem algum padrão enquanto os atributos possuem distribuições com padrões distorcidos à direita, isso pode ser explicado pelo o padrão de criação do conjunto de dados que analisou com mais frequência uma faixa de valores do que outra. Todas os outros atributos apresentaram uma distribuição com um padrão normal ou gaussiana.

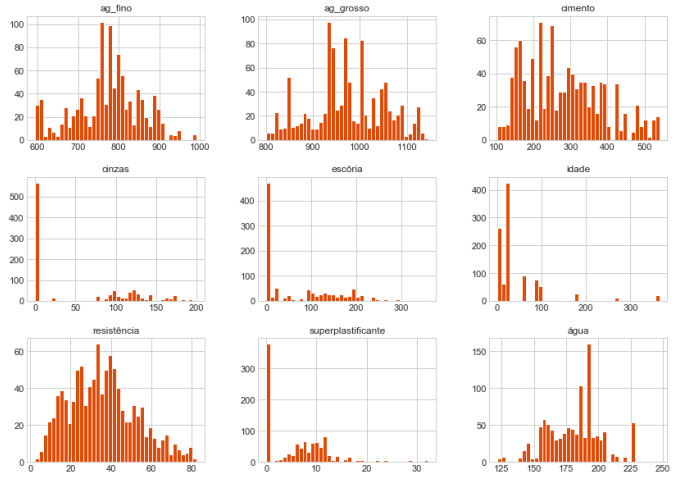

Figura 4. Histogramas dos atributos.

O chamado *boxplot* é um tipo de gráfico usado para avaliar a distribuição dos dados, com base nas estatísticas básicas

anteriormente calculadas na Fig.2. Segundo Evsukoff [X] a construção do *boxplot* segue o seguinte procedimento, de cima para baixo, o traço superior representa o limite superior, o primeiro traço da caixa é o valor do terceiro quartil (75%), seguido pelo segundo quartil ou mediana (50%) e o último traço da caixa é o valor do primeiro quartil (25%), o último traço é o valor do limite inferior. A Fig. 5 indica os *boxplots* para os dados numéricos do *dataset* analisado neste trabalho.

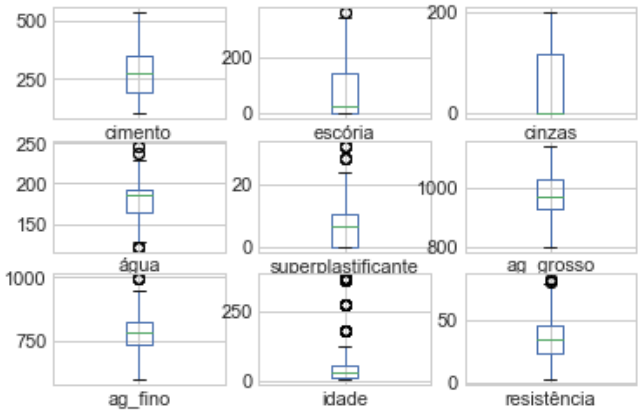

Figura 5. *Boxplot* do conjunto de dados.

Modelos de Regressão

Os modelos de regressão apresentados no item III foram aplicados no conjunto de dados em estudo, aplicando a técnica de validação cruzada com *K=10*, os resultados das estatísticas de validação *REQM* e R^2 foram comparados com os *REQM* e R^2 de execuções dos modelos simplesmente dividindo os dados em treino e teste, como indicado na Tab. 1.

Tabela 1. Resultados do modelos de regressão

Modelo	Treino e Teste		Validação Cruzada	
	REQM	R^2	REQM	R^2
Regressão Linear	10,944	0,601	10,254	0,621
Regressão LASSO	11,765	0,539	10,831	0,579
Regressão de Rigde	10,944	0,601	10,253	0,621
K vizinhos mais próximos	9,124	0,723	9,029	0,706
AdaBoost	8,274	0,772	7,589	0,787
Árvores de Decisão	9,045	0,728	7,298	0,813
Florestas Aleatórias	6,258	0,870	5,341	0,896
Gradiente *Boosting*	6,436	0,862	5,164	0,904
Gradiente *Boosting* (gridsearchcv)	4,627	0,917	5,294	0,907

A Fig. 6 apresenta para todos os modelos aplicados a correlação entre os resultados reais em função dos resultados preditos, nesse tipo de representação observa-se que quanto mais próximos os pontos estiverem da reta tracejada na diagonal melhor o modelo se ajustou aos dados, como por exemplo no modelo Gradiente *Boosting* que tem o menor coeficiente R^2.

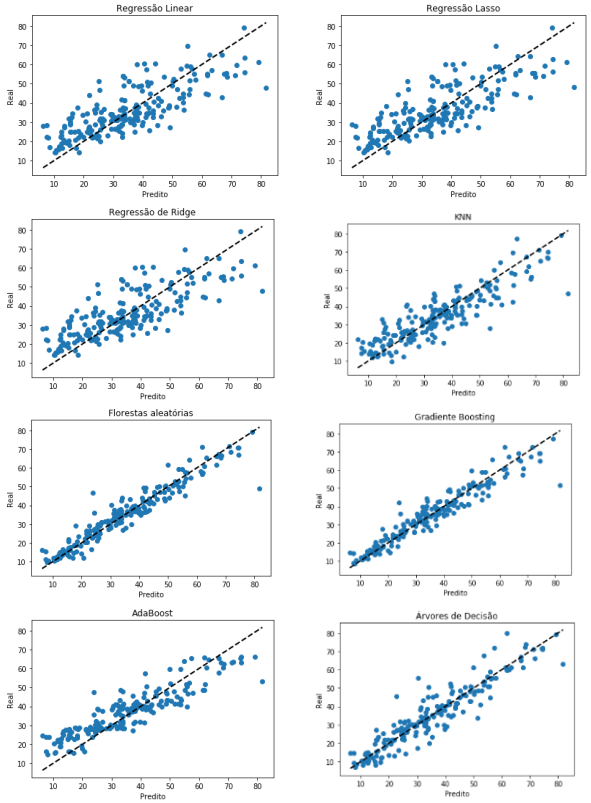

Figura 6. *Correlação real vs. predito para os diferentes modelos.*

A Fig. 7 apresenta a importância de cada atributo no modelo para o modelo de floresta aleatória, isto porque sabe-se que quanto mais características, mais provavelmente o modelo irá sofrer superajuste (*overfitting*). Observa-se que o atributo "cinzas" apresentou a menor influência no modelo e provavelmente poderia ser retirado das análises sem perdas significativas de resultados, essa observação é importante para os laboratórios ao analisar qual será a mistura adotada porque a quantidade de cinzas não tem grande influência na previsão da resistência, usando o modelo de florestas aleatórias. O contrário ocorre com a quantidade de cimento e a idade, que tem elevada influência no modelo e não podem ser excluídas das análises. Nesse ponto é interessante salientar que não só quantidade de cimento é importante como também o tipo do cimento utilizado pois estes podem influenciar o comportamento da curva de resistência em diferentes idades.

	Importância
cimento	0.330079
escória	0.075474
cinzas	0.021808
água	0.103928
superplastificante	0.066722
ag_grosso	0.027455
ag_fino	0.046049
idade	0.328486

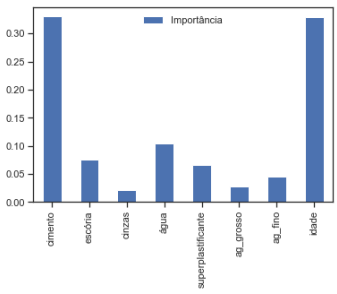

Figura 7. Importância dos atributos no modelo de Floresta Aleatória.

A Fig. 8 indica a árvore de decisão para o conjunto de dados, a previsão do valor final funciona da seguinte forma: suponha

que deseja-se prever o valor da resistência de um determinado concreto. Começando pelo nó da raiz (profundidade 0, na parte superior): este nó pergunta se a idade do concreto é menor do que 21. Se for, então o processo é deslocado para o nó filho esquerdo da raiz (profundidade 1, esquerda). Caso contrário, é deslocado para a direita, essa sequência é repetida até encontrar-se um nó da folha (que não tem nenhum nó filho), este irá indicar o valor de resistência do concreto predito pelo modelo.

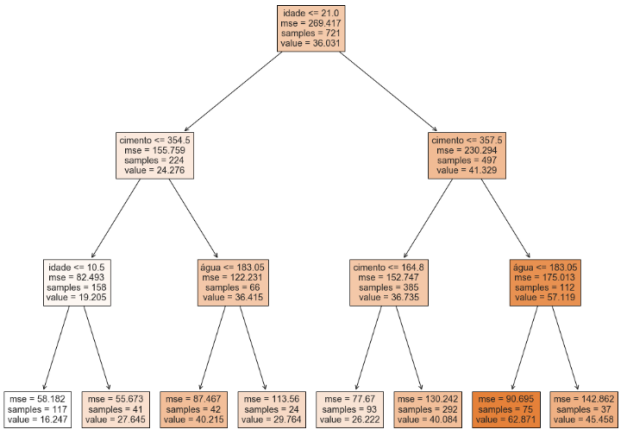

Figura 8. Árvore de decisão do conjunto de dados.

A árvore de decisão (Fig.8) obtida com o modelo condiz com o esperado, a partir da opinião de um especialista, uma vez que quanto menor a idade do concreto menor a resistência e vice e versa. O mesmo pode ser afirmado para o atributo cimento. Para a análise dos outros atributos seria necessário observar uma arvores com maior profundidade.

VI. CONCLUSÕES E SUGESTÕES

Portanto esse trabalho teve como objetivo a avaliação de métodos de Mineração de Dados para estudo da resistência do concreto. Foi possível verificar que tais métodos podem ser úteis para extração de conhecimento em uma base de dados voltadas a dosagem do concreto.

A técnica de validação cruzada foi aplicada para garantir uma amostragem mais representativa do conjunto de dados e garantir maior veracidade dos resultados.

É válido salientar que nos trabalhos consultados se utilizam de técnicas de aprendizado complexas como redes neurais e nesse trabalho foram aplicados modelos mais simples que uma rede neural e os resultados foram satisfatórios. Logicamente uma testagem maior é necessária, porém podemos afirmar que para um primeiro momento os métodos se mostraram promissores

Tal pesquisa contribui no sentido de que empresas possam ter um maior controle de qualidade dos concretos produzidos no pátio de obras. Pois prever a resistência do mesmo antes de executar ensaios exaustivos e demorados seria interessante do ponto de vista financeiro.

REFERÊNCIAS

[1] Associação Brasileira de Normas Técnicas, *ABNT NBR 5739 - Concreto - Ensaios de compressão de corpos-de-prova cilíndricos*. Rio de Janeiro: ABNT, 2018.

[2] J. P. Z. Guimarães, "Estudo experimental das propriedades do concreto de alto desempenho", Mestrado em Engenharia Civil, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2002.

[3] F. T. F. do Nascimento, "Concreto de alto desempenho e sua aplicação em vigas de edifícios", Mestrado em Engenharia de Estruturas, Universidade de São Paulo, São Carlos, 1997.

[4] A. G. C. Baccin, "Fundamentos do concreto de alto desempenho e sua aplicação no projeto de pilares", Mestrado em Engenharia de Estruturas, Universidade de São Paulo, São Carlos, 1998.

[5] B. F. Tutikian e P. Helene, "Dosagem dos Concretos de Cimento Portland", in *Concreto: Ciência e Tecnologia*, Ibracon, 2011.

[6] B. F. Tutikian, G. C. Isaia, e P. Helene, "Concreto de Alto e Ultra-Alto Desempenho", in *Concreto: Ciência e Tecnologia*, Ibracon, 2011, p. 44.

[7] S. I. McClean, "Data Mining and Knowledge Discovery", *Encyclopedia of Physical Science and Technology*, p. 18, 2001.

[8] U. Fayyad, G. Piatetsky-Shapiro, e P. Smyth, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, vol. 17, nº 3, Art. nº 3, doi: https://doi.org/10.1609/aimag.v17i3.1230.

[9] D. A. Silva, "Aplicação de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do Ministério Público do Trabalho", Bacharelado em Ciência da Computação, Universidade Federal de Uberlândia (UFU), Uberlândia, 2018.

[10] M. M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. Wiley-IEEE Press, 2019.

[11] R. Santos, "Conceitos de Mineração de Dados na Web", in *XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos*, Belo Horizonte (BH), 2009, p. 81–124.

[12] M. H. Casagrande, "Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade (p > n)", Mestrado em Estatística, Universidade de São Paulo, São Carlos, 2019.

[13] M. A. P. Marques, "Análise e comparação de alguns métodos alternativos de seleção de variáveis preditoras no modelo de regressão linear", Mestrado em Estatística, Universidade de São Paulo, São Paulo, 2018.

[14] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow", p. 564.

[15] L. Breiman, J. H. Friedman, R. A. Olshen, e C. J. Stone, *Classification and Regression Trees*, 1ª Edição. Wadsworth: Routledge, 1984.

[16] R. L. de Morais, "Uso de Árvores Aleatórias para Classificação Sensorial de Arroz Cozido", Bacharelado em Estatística, Universidade de Brasília (UnB), Brasília, 2017.

[17] S. C. Garcia, "O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde", Porto Alegre, Universidade Federal do Rio Grande do Sul (UFRGS), 2003.

[18] M. Fernandez-Delgado, E. Cernadas, S. Barro, e D. Amorim, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?", p. 49.

[19] Y. Freund e R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, vol. 55, nº 1, p. 119–139, ago. 1997, doi: 10.1006/jcss.1997.1504.

[20] T. Hastie, R. Tibshirani, e J. Friedman, "Additive Models, Trees, and Related Methods", in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T. Hastie, R. Tibshirani, e J. Friedman, Orgs. New York, NY: Springer, 2009, p. 295–336.

[21] I.-C. Yeh, "Concrete Compressive Strength Data Set", *UCI Machine Learning Repository*, 2007. https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

[22] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks", *Cement and Concrete Research*, vol. 28, nº 12, p. 1797–1808, 1998, doi: 10.1016/S0008-8846(98)00165-3.

[23] S. Borra e A. Di Ciaccio, "Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods", *Computational Statistics & Data Analysis*, vol. 54, nº 12, p. 2976–2989, dez. 2010, doi: 10.1016/j.csda.2010.03.004.

[24] J. P. Z. Cunha, "Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos", Mestrado em Estatística, Universidade de São Paulo (USP), São Paulo, 2019.

[25] A. P. dos Santos Júnior, "Análise das características de jogabilidade no PUGB usando árvore de decisão", bacharelado em Estatística, Universidade Federal de Uberlândia (UFU), Uberlândia, 2019.

Amanda Isabela de Campos Atualmente é aluna de doutorado do Programa de Engenharia Civil - COPPE/UFRJ na linha de pesquisa Estruturas e Sistemas Offshore. Possui mestrado em estruturas e materiais pelo Programa de Engenharia Civil - COPPE/UFRJ. Possui graduação em Engenharia Civil pela Universidade Federal de Juiz de Fora.