Proposta de aplicação de aprendizado de máquina na previsão da resistência do concreto

Josilon Campos Souza

(josilonsouza7@gmail.com)

Professora orientadora: Andrea dos Reis Fontes

Coordenação de curso de Engenharia Civil

Resumo

O presente estudo buscou demonstrar a aplicabilidade prática de um algoritmo de aprendizado de máquina na previsão de resistência do concreto e colaboração dessa tecnologia na construção civil. Foram usados métodos de análise exploratória de dados através de um banco de dados com amostras de misturas de concreto e suas respectivas resistências. A construção do modelo de previsão se deu a partir do Google colaboratory, ambiente gratuito para construção do código do aprendizado de máquina. Os resultados demonstraram que o modelo construído consegue prever com alta precisão a resistência do concreto depois de determinado período de cura, além de determinar os pesos de cada componente para a mistura de concreto. Dessa forma concluiu-se que a tecnologia aplicada neste estudo tem grande impacto na construção civil na previsão da performance de um dos principais componentes dessa indústria.

Palavras-chave: Aprendizado de Máquina. Concreto. Random Forest Regression. Análise de Dados.

Abstract

The present study sought to demonstrate the practical applicability of machine learning algorithm in predicting the strength of concrete, as also the support provided by this technology in civil construction. Exploratory data analysis methods were used through a database with samples of concrete mixtures and their respective strengths. The construction of the prediction model was produced on google colaboratory, a free environment for building the machine learning code. The results showed that the developed model is able to predict with high accuracy the strength of the concrete after a certain curing period, in addition to determining the weights of each component for the concrete mix. Thus, the conclusion is that the technology applied in this study has a great impact on civil construction in the performance prediction of one of the main components of this industry.

Keywords: Machine Learning. Concrete. Random Forest Regression. Data Analysis.

1. Introdução

Considerada a quarta revolução industrial, a indústria 4.0 traz impactos econômicos, culturais e sociais em todo o planeta. O uso da internet para o fluxo de comunicação entre o mundo real e o cibernético, sejam máquinas, sensores, robôs ou

computadores aumentou exponencialmente e alterou a indústria evitando desperdícios, melhorando o desempenho e aumentando a segurança (CAVALCANTI, et al. 2018).

Dentro desse termo, indústria 4.0, se encaixam outros que compõem todo esse progresso. *Big data e data analysis*, são ferramentas que viabilizam o processo de implementação dessa nova revolução. Esse processo, em partes, só se torna possível pelo grande poder de processamento computacional disposto atualmente (CAVALCANTE; ALMEIDA, 2017).

Os dados gerados diariamente dão a base para se usar essas ferramentas. A quantidade, velocidade, variedade, variabilidade e complexidade dão forma à *Big Data* que junto à Data *analysis* proporcionam um conhecimento que pode ser aplicado à indústria, processando dados, gerando testes e elaborando modelos das mais diversas formas. Entretanto, sem o apoio de técnicas específicas esse processo não oferece benefício algum (FAVERO, 2022).

Data analysis ou "análise de dados" é a responsável por tratar esse grande volume de dados, com o propósito de extrair informações úteis para a tomada de decisão. Dentro desse processo podemos identificar abordagens que são usadas em vários tipos de negócios e ciências. Uma das técnicas usadas para extrair essas informações é a análise preditiva, que consiste em usar padrões estatísticos para tentar prever determinadas situações definidas antes pelo usuário (KUDYBA et.al, 2014).

Junto a essa revolução, a economia globalizada desperta uma carência na diminuição da probabilidade de falhas em qualquer linha de produção, resultando assim num destaque crescente na confiabilidade. Dessa forma o conhecimento provido pela análise dessas falhas e da minimização de sua ocorrência nos revela uma grande variedade de cenários (FOGLIATTO; RIBEIRO, 2009; RANGEL et al., 2012).

Na indústria da construção civil, um dos processos mais significativos é a fabricação do concreto. Segundo a Federación Iberoamericana de Hormigón Premesclado (FIHP) o consumo mundial desse material chega a 11 bilhões de toneladas por ano, destacando sua importância no setor.

Na obra o concreto deve atingir níveis de performance rígidos para garantir a segurança do que está sendo construído. Essa qualidade é garantida por normas, que usam cálculos matemáticos para encontrar valores de resistência seguros, transgredir esses cálculos pode ocasionar não apenas um risco de segurança como também econômico. A demolição da estrutura total ou parcial eleva os prejuízos e pode inviabilizar o projeto (FILHO; HELENE, 2011).

Uma das técnicas que auxiliam o processo de segurança do concreto, é a retirada de amostras dessa mistura para produzir corpos de prova que serão ensaiados para identificar se a fabricação alcançou os níveis de resistência previstos em cálculo. Esses ensaios demandam tempo e máquinas pesadas para gerar o relatório final que atestam ou não a resistência da mistura (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 1994).

Na produção do concreto, seja em usinas ou in loco, existem fatores que definem o valor final da resistência do concreto e devem ser considerados na elaboração da mistura. A proporção dos materiais impacta diretamente no resultado final, além do tempo

que a mistura leva para atingir o seu máximo de desempenho (MEHTA; MONTEIRO, 2008).

1.1 Delimitação do assunto

Na indústria, como foi dito, a produção do concreto deve seguir padrões de qualidade rígidos. O tempo necessário para constatar esses padrões também é elevado trazendo custos adicionais para o orçamento. A elaboração do concreto para alcançar os resultados previstos em projeto dependem das proporções dos materiais usados na mistura, dessa forma o presente trabalho tem o objetivo de utilizar a análise preditiva através de algoritmos de aprendizado de máquina como método para prever a resistência final do concreto.

O trabalho foca na utilização de um modelo para treinar um algoritmo de máquina preditiva e na realização de testes para validar os resultados com amostras reais. Esse algoritmo se baseia em um outro chamado *árvore de decisão*, onde pontos de decisão são criados e o modelo deve escolher por quais pontos seguir para tomar uma decisão que tenha mais probabilidade de acontecer. O *random forest regression*, usado na pesquisa, deriva da árvore de decisão, porém traz algumas diferenças para calcular as probabilidades dos resultados (KAMINSKI; JAKUBCZYK; SZUFEL, 2017).

1.2 Objetivos

O objetivo geral é aplicar o aprendizado de máquina através da análise de dados, para propor um modelo preditivo com um percentual de acerto que consiga contribuir no processo de obtenção do fck do concreto ou substituir o modelo atual para calcular a resistência à compressão do concreto.

Objetivos específicos:

- Analisar a relevância de cada componente na mistura do concreto;
- Utilizar a análise exploratória de dados com o objetivo de entender o banco de dados;
- Demonstrar a aplicabilidade prática de um algoritmo de aprendizado de máquina.

1.3 Justificativa

A construção civil é um dos setores da economia com maior relevância para o país, contribuindo no desenvolvimento social e econômico. A geração de emprego deste setor e sua contribuição no consumo de bens e serviços é uma das maiores evidenciando seu destaque (ABIKO; GONÇALVES, 2003).

Entretanto há uma deficiência nessa indústria. No Brasil a Confederação Nacional da Indústria (CNI), destaca que apenas 48% das indústrias brasileiras implementam alguma tecnologia, empresas grandes, no entanto tem um percentual maior de 63%, enquanto empresas de pequeno porte seguem com apenas 25% (SILVA, 2018).

De acordo com Nascimento e Santos apud Silva et.al.:

> Nas últimas décadas, a necessidade de se implementar novas Tecnologias de Informação e Comunicação (TICs) cresceu gradualmente dentro da indústria da construção civil, visando a otimização de etapas, o acesso a informações em tempo real e, consequentemente, ganhos na produtividade tanto dos escritórios quanto dos canteiros de obra. A utilização de TICs passa a ser ainda mais relevante dentro do cenário atual da pandemia, uma vez que atividades que dependem exclusivamente de ações humanas são realizadas constantemente em regime reduzido, o que impulsiona ainda mais as discussões sobre as inovações tecnológicas no setor. (SILVA et.al. 2021, p. 2).

O presente trabalho tenta tratar essa deficiência aplicando técnicas de análise de dados na construção de um algoritmo de aprendizado de máquina para diminuir custos da indústria da construção civil.

2 Referencial teórico

2.1 Concreto

Segundo Pinheiro (2007), o concreto pode ser definido por um material de construção formado a partir de uma mistura que tenha o equilíbrio adequado entre: aglomerantes, agregados e água. Cada um desses componentes tem uma função específica na pasta. O aglomerante mais usual é o cimento Portland que reage com a água e endurece a mistura com o tempo, o agregado são minerais com tamanho em torno de 0,075 mm e 4,8 mm (miúdo) e maiores que 4,8 mm (graúdo) que tem a função de dar volume ao concreto para reduzir o custo. O concreto simples é formado pela junção da mistura argamassa (cimento, areia e água) com o agregado graúdo, gerando uma alta resistência ao esforço de compressão, sendo essa sua principal função na construção.

As vantagens do concreto na construção civil são imensas, sendo possível moldá-lo em várias formas, tendo baixo custo nos materiais, a estrutura é monolítica e trabalha como uma peça só quando solicitada, facilidade e rapidez na execução e baixo custo de manutenção. Porém ainda existem algumas desvantagens, a principal delas é sua baixa resistência a tração e algumas outras também estão presentes como fragilidade, fissuração, peso próprio elevado e o custo das fôrmas para moldagem (PINHEIRO, 2007).

Pinheiro (2007), explica também como contornar essas desvantagens usando o concreto de alto desempenho. A obtenção desse tipo de mistura vem da adição de sílica ativa e aditivos plastificantes, que podem ser substituídos por cinza volante ou resíduo de alto-forno apresentando um resultado melhor do concreto.

Para Parizotto (2017), os aditivos mencionados acima conseguem melhorar de forma significativa essa mistura. A escória ou resíduo de alto-forno eleva a trabalhabilidade, além de aumentar a resistência, reduzir a permeabilidade e o aumento da temperatura durante a hidratação e elevar a resistência a sulfatos. A adição das cinzas volante tem o objetivo de reduzir a porosidade e diminuir a retração por secagem do concreto, essa adição também consegue melhorar a trabalhabilidade e resistência a sulfatos. Os superplastificantes para o concreto de alto desempenho também são adicionados e tem o objetivo de reduzir a água usada, mas obter a mesma trabalhabilidade, esse aditivo tem maior demanda em obras com condições de concretagem complexas.

As proporções na fabricação do concreto os fatores de mais impacto, dosar essas quantidades de material para garantir a resistência e a trabalhabilidade evita prejuízos

posteriores, como por exemplo, aumentar a quantidade de cimento com a finalidade de dar mais resistência sem ter essa necessidade (PARIZOTTO, 2017).

De acordo com Allen e Iano apud Parizotto (2017), uma mistura de concreto com agregados bem graduados a relação água-cimento é o principal fator para determinar a resistência. No entanto, no processo de lançamento e acabamento do concreto ainda se adiciona mais água para proporcionar uma fluidez e plasticidades necessária para trabalhar com a mistura. Essa adição de água acaba por ocasionar vazios microscópicos no concreto endurecido por conta da retração do concreto ao evaporar essa água.

Os agregados são outros elementos que influenciam no concreto. A forma, porosidade e distribuição granulométrica trazem características distintas para a mistura. Agregados esféricos trazem uma trabalhabilidade maior facilitando o movimento e diminuindo a quantidade de massa para envolvê-los. A porosidade influencia na quantidade de água usada na mistura, agregados com maior porosidade absorvem parte da água usada reduzindo a fluidez da mistura. a distribuição granulométrica também tem impacto na trabalhabilidade, pois quanto maior a área superficial mais pasta será necessária para envolvê-los (PARIZOTTO, 2017).

Neville apud Parizotto (2017), afirma que além de todas essas características as condições climáticas e o tempo de lançamento também irão influenciar na trabalhabilidade. Nos dias mais quentes a evaporação se dá de forma mais rápida, já a demora no lançamento da mistura faz com que o cimento presente ali já tenha iniciado a sua reação e o agregado absorvido mais água. As perdas nesse sentido podem ser evitadas com a implementação dos aditivos.

Essas propriedades do concreto são medidas em alguns ensaios normatizados pela ABNT. As normas técnicas NBR NM 67:1998 e NBR 5739:2007 traz as regras de como executar esses ensaios, o primeiro é o ensaio de abatimento de tronco ou *slump* test, quanto ao segundo ensaio de resistência à compressão do concreto (PARIZOTTO, 2017).

O ensaio de abatimento de tronco consiste em um tronco de cone com altura de 300 mm, que é posicionado sobre uma superfície lisa com a menor altura para cima. O concreto preenche esse tronco em três camadas e cada camada recebe 25 golpes de uma haste metálica padronizada com 16 mm de diâmetro e ponta arredondada. A camada final é nivelada com uma desempenadeira e pelo movimento de rolagem da haste, em todo esse processo esse cone deve estar firmemente parado contra sua base. Ao final o molde é levantado lentamente e o concreto sofre um abatimento que deve ser analisado pela tabela a seguir (NEVILLE, 2016):

Tabela 1 – Slump test

Tipo de Trabalhabilidade	Abatimento (mm)
Abatimento zero	0
Muito baixa	5-10
Baixa	15-30
Média	35-75
Alta	80-155
Muito alta	160 ao colapso

Fonte: Neville (2016).

Allen e Iano apud Parizotto (2017), afirma que o ensaio de resistência à compressão é realizado ao final do período de cura do concreto e se dá a partir do seguinte processo. Ao receber o concreto usinado ou fabricado in loco, retira-se parte da mistura para moldar corpos de prova cilíndricos e normalizados que serão levados ao laboratório, lá essas amostras esperam o período de cura de 28 dias e são ensaiados através de prensa hidráulica para determinar a resistência à compressão. O calcula é feito a partir da divisão do valor da carga de ruptura da amostra pela área de aplicação de força, na hipótese dessa resistência não alcançar os valores de projeto são retiradas mais amostras do concreto já executado e ensaiado novamente, caso também não atinjam a resistência deve-se descartar tudo o que foi construído com aquela mistura.

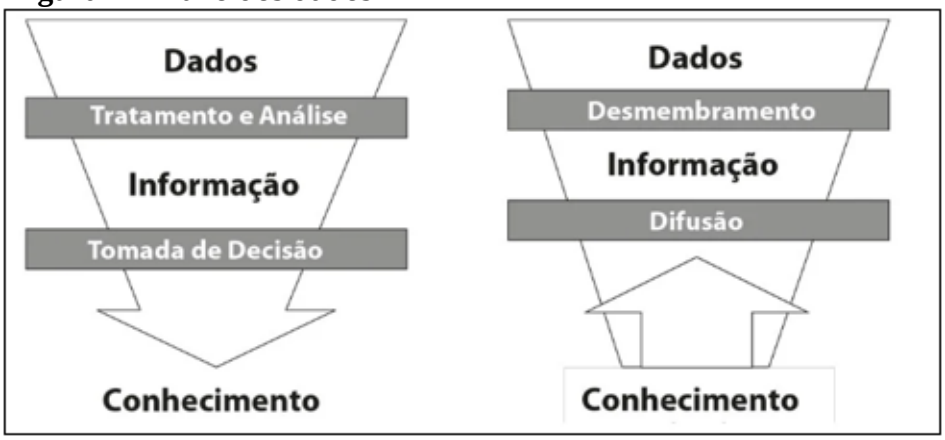
2.2 Análise de dados

No atual momento, no qual o poder computacional chegou a níveis gigantescos e que o processo de digitalização de dados, anteriormente físicos, se tornou igualmente grande, a análise de dados ganha um espaço de destaque para melhorar a tomada de decisão de qualquer atividade (FAVERO, 2022).

Sobre o termo análise de dados, uma abordagem atual é que a análise está presente em processos, frameworks, tecnologias e algoritmos sendo utilizada para colher percepções de valor. Apenas o volume de dados não tem relevância até serem processados em informações significativas. Então dá-se a análise a função de extrair e criar informações úteis de um conjunto de dados brutos, e que se divide nas seguintes etapas: filtragem, processamento, categorização, condensação e contextualização. Após esses passos pode-se então produzir conhecimento sobre o sistema, ambiente, usuários e suas operações tornando a tomada de decisão mais inteligente (ARSHDEEP; MADISETTI, 2019, p. 22, tradução nossa).

Entende-se então que há uma hierarquia entre dados, informações e conhecimento. O primeiro quando tratado converte-se em informações, o conhecimento por sua vez nasce quando essas informações são percebidas e aplicadas no processo de tomada de decisão. O processo pode também ser revertido como demonstrado na figura 1 (FAVERO, 2022).

Figura 1 - Fluxo dos dados



Fonte: Favero (2022).

Para dar início a esse processo Kudyba (2014), define que deve haver uma pré-análise dos dados, também chamada de análise exploratória de dados, permitindo ao analista uma compreensão melhor do que está disposto. Dito isto, antes de aplicar

qualquer modelo analítico, o pesquisador deve investigar as variáveis desse banco de dados com objetivo de encontrar vieses ou distorções dos dados, discrepâncias e erros. Aplicar métodos de cálculo como: máximo, mínimo, variância e desvio padrão pode ser uma das formas de encontrar esses problemas.

Na figura 2 é demonstrado um exemplo de banco de dados, no qual foi aplicado uma pré-análise superficial, revelando a quantidade de valores preenchidos nesta base de dados (count), desvio padrão (std), máximo (max) e mínimo (min). Ao final desse processo ou algum outro definido pelo analista pode-se então começar o processo de análise do banco de dados.

Figura 2 - Exemplo pré-análise

	count	mean	std	min	25%	50%	75%	max
cimento	1030.0	281.167864	104.506364	102.00	192.375	272.900	350.000	540.0
escória	1030.0	73.895825	86.279342	0.00	0.000	22.000	142.950	359.4
cinzas	1030.0	54.188350	63.997004	0.00	0.000	0.000	118.300	200.1
água	1030.0	181.567282	21.354219	121.80	164.900	185.000	192.000	247.0
superplastificante	1030.0	6.204660	5.973841	0.00	0.000	6.400	10.200	32.2
agreg.graúdo	1030.0	972.918932	77.753954	801.00	932.000	968.000	1029.400	1145.0
agrer.miúdo	1030.0	773.580485	80.175980	594.00	730.950	779.500	824.000	992.6
idade	1030.0	45.662136	63.169912	1.00	7.000	28.000	56.000	365.0
resistência	1030.0	35.817961	16.705742	2.33	23.710	34.445	46.135	82.6

Fonte: Elaborado pelo autor.

Há atualmente três metodologias que se destacam no universo analítico, a descritiva, a comparativa e a preditiva. A escolha de uma dessas depende do objetivo que o analista tem para sua pesquisa, independente da escolha o resultado final dessa análise é fornecer um conhecimento para melhorar a tomada de decisão (KUDYBA, 2014).

A análise preditiva usada nessa pesquisa é responsável por grandes inovações no mercado como: identificação de fraudes, análise de risco, otimização de operações e aumento da competitividade. O primeiro é utilizado bastante em bancos, com algoritmos de *machine learning* que agem na identificação e redução de fraudes tentando prever e evitar que operações fraudulentas sejam concluídas. Já a análise de risco é difundida no setor de seguros de carros, no qual é identificado os perfis de motoristas com menos chances de se acidentarem para oferecer valores menores (VELOSO, 2021).

Kudyba (2014, p. 224) afirma que "O termo análise preditiva, cunhado e popularizado no final da década de 1990, é uma maneira relativamente nova de descrever a prática relativamente antiga de usar a análise estatística e outras técnicas matemáticas para prever o comportamento".

Sobre esse tipo de análise, Arshdeep e Madisetti (2019) definem o conceito e a abordagem mais adotada. Verifica-se então que, a análise preditiva pode ser utilizada na previsão da ocorrência de um evento, provável resultado de um evento ou a previsão de valores futuros. Essa análise tenta responder à pergunta: O que *é mais provável acontecer?* Para isso é usado algum dado existente para treinar o modelo de análise, no qual se percebe padrões e tendências na base de dados. O recurso adotado com maior frequência para esse tipo de treinamento é a divisão da base de dados em dados de

treinamento e dados de teste, exemplo, 75% dados de treinamento e 25% dados de teste. Dessa forma perguntas como *Qual a chance de ocorrer uma falha em uma máquina ou de ocorrer um desastre natural*? Podem ser respondidas.

2.2 Aprendizado de Máquina

O termo aprendizado de máquina é relativamente velho comparado a outras tecnologias, atualmente no cenário tecnológico o aprendizado de máquina vem ganhando grande força. Impulsionado em grande parte pela indústria 4.0 que prioriza tecnologias emergentes e o avanço tecnológico recente, esse recurso acaba destacando-se e sendo o responsável por vários resultados positivos (MAISUECHE, 2019).

No passado, para se resolver um problema os programas computacionais precisavam ser codificados através de uma série de etapas definidas pelas práticas necessárias para a resolução do problema. Na década de 1970, na qual o uso da inteligência artificial estava difundido, começou a se usar essa tecnologia como um novo artifício para esclarecer novos problemas. A obtenção de conhecimento através de especialistas de uma determinada área, era usado como base das inteligências artificiais que por fim tentavam resolver obstáculos reais. Entretanto, esse processo era complexo e se limitava pela subjetividade do especialista ou por seu receio em ser substituído pela máquina. Recentemente o volume, complexidade e velocidade dos dados gerados trouxe uma nova perspectiva com o objetivo de criar novas tecnologias baseadas em IA que sejam independentes desse processo complexo (FACELI, LORENA, GAMA, ALMEIDA, & Carvalho, 2021).

Segundo Faceli et.al. (2021), o aprendizado de máquina deriva da inteligência artificial e é hoje responsável por vários avanços tecnológicos em setores como a economia, segurança pública e automação industrial. Empresas como Netflix e Google usam AM em seus negócios e sua aplicação vai desde recomendações de filmes baseadas em padrões do usuário a legendas automáticas em vídeos.

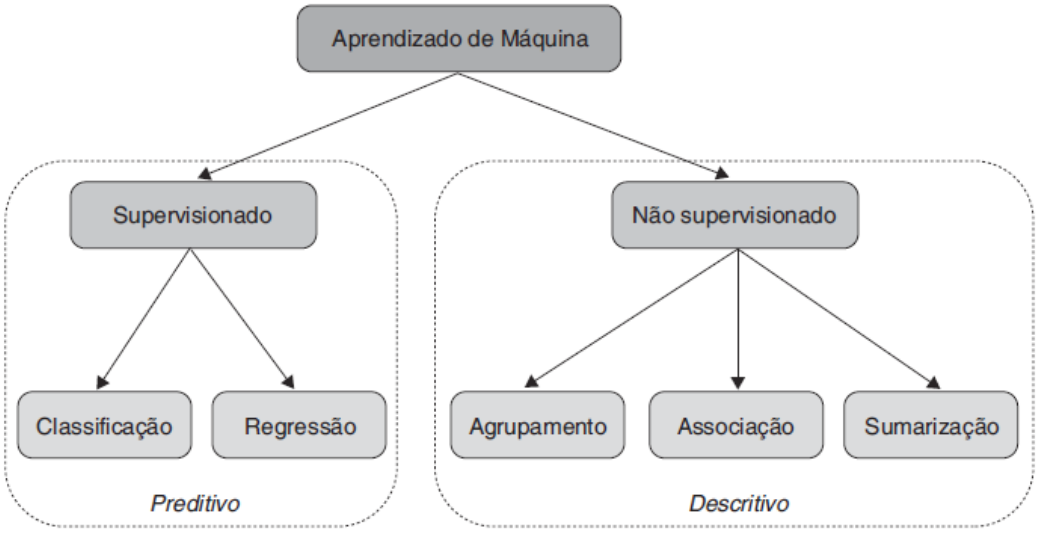
Sobre as aplicações de AM, pode-se afirmar que:

> Aplicações baseadas em AM utilizam heurísticas que buscam por modelos capazes de representar o conhecimento presente em um conjunto de dados. Em geral, os conjuntos de dados são estruturados em formato tabular, uma matriz atributo-valor, em que cada linha representa um objeto (instância ou exemplo) e cada coluna representa um atributo (característica ou variável). Os atributos podem ser divididos em atributos preditivos, cujos valores descrevem características dos objetos, que formam um vetor de entrada, e atributo alvo, cujo valor rotula o objeto, com uma classe ou valor numérico. Essas denominações têm por origem o frequente uso dos valores dos atributos preditivos de um objeto para predizer o valor de seu atributo alvo. Nem todos os conjuntos de dados possuem atributo alvo. Quando possuem, são chamados de conjuntos de dados rotulados. (FACELI et.al. 2021, p. 2).

Há ainda uma divisão dos tipos de tarefas que o AM pode atuar, são elas tarefas preditivas e descritivas. A primeira como dito acima busca predizer o valor do atributo alvo através de atributos preditivos, já a segunda, busca extrair padrões dos valores dos dados e uma das tarefas principais é encontrar regras de associação de valores de um subconjunto de atributos preditivos com outro (FACELI et.al. 2021).

Na figura 3 é demonstrado a divisão dessas tarefas e os resultados que elas entregam como, classificar, agrupar ou associar os dados.

Figura 3 - Fluxograma aprendizado de máquina



Fonte: FACELI et.al (2021)

No centro do aprendizado de máquina, a ferramenta mais importante é o algoritmo usado para manipular os dados. O sucesso da AM depende da escolha correta do algoritmo e os dados usados nessa análise devem ter uma preparação cuidadosa do cientista (MUELLER; MASSARON, 2019).

No campo científico é necessário que haja uma padronização das regras usadas na construção do AM. Esse processo tem grande significado, pois padroniza o processo deixando-o mais eficiente (MUELLER; MASSARON, 2019).

Sobre esse processo Mueller e Massaron afirmam que:

> À medida que os cientistas continuam a trabalhar com uma tecnologia e a transformar hipóteses em teorias, ela se torna mais relacionada à engenharia (onde as teorias são implementadas) do que à ciência (onde são criadas). À medida que as regras que governam uma tecnologia se tornam mais claras, grupos de especialistas trabalham em conjunto para defini-las por escrito. O resultado são as especificações (um grupo de regras com que todos concordam).
> Finalmente, as implementações das especificações se tornam padrões que uma organização regulatória, como o IEEE (Institute of Electrical and Electronics Engineers) ou uma combinação do ISO/IEC (International Organization for Standardization/International Electrotechnical Commission), gerencia. A IA e o aprendizado de máquina existem há tempo suficiente para criar especificações, mas atualmente você não encontrará quaisquer padrões para nenhuma delas. (MUELLER; MASSARON, 2019, p. 20).

O aprendizado de máquina é programado a partir de experiências passadas, para isso o princípio de inferência denominado indução extrai conclusões genéricas de um conjunto particular de dados. Existe uma característica de AM que deve estar presente nessa inferência que é a capacidade de lidar com dados imperfeitos. A base de dados normalmente apresenta ruídos, dados inconsistentes, ausentes ou redundantes. Os

algoritmos de AM devem ser capazes de contornar esses problemas e minimizar sua influência no processo, no entanto dependendo da quantidade de problemas o aprendizado pode ser prejudicado ou inviabilizado, para isso as técnicas de pré-processamento são aplicadas para reduzir ou eliminar esses obstáculos (FACELI et.al. 2021).

Atualmente o desenvolvimento do aprendizado de máquina é aplicado como uma disciplina da ciência e não da engenharia. Nesse processo ainda é necessária uma certa subjetividade que não se aplica a uma matemática exata. Ao analisar os dados, por exemplo, deve-se considerar como esses são usados, alguns dados são a base para o algoritmo atingir resultados específicos, enquanto outros fornecem a saída para entender os padrões subjacentes. O equilíbrio desse processo depende do cientista para que descubra como produzir a saída mais adequada, por exemplo, na limpeza dos dados a subjetividade de cada analista tem impacto, o tratamento que cada um deles dá ao banco de dados é distinta, claro que a remoção de valores duplicados é feita de forma regular, porém a análise de subconjuntos dessa base pode ser feita de forma única por cada um deles (MUELLER; MASSARON, 2019).

Há, no entanto, duas características ao desenvolver um algoritmo, que devem ser evitadas. O aprendizado de máquina busca entender os padrões da base de dados para conseguir responder questionamentos sobre a mesma, esse aprendizado deve ser capaz de ser aplicado também a outros objetos do mesmo domínio do problema, mas que não façam parte do mesmo conjunto de dados de treinamento. Essa capacidade é denominada generalização e permite ao algoritmo ser aplicado a outros conjuntos de dados. A baixa capacidade de generalização produz uma característica de super ajuste aos dados (overfitting), fazendo com que o modelo decore o conjunto de dados de treinamento não tendo realmente aprendido. O contrário é o sobre ajuste (underfitting) que consiste no algoritmo ter uma baixa capacidade preditiva, isso acontece muitas vezes quando os dados de treinamento são poucos representativos ou o algoritmo é simples e não consegue entender os padrões dos dados (FACELI et.al. 2021).

2.2.1 Árvore de decisão

Os algoritmos de árvore de decisão se baseiam em regras para tomar cada decisão. A estrutura de uma árvore assemelha-se a um fluxograma, a criação de nós é feita e em cada um deles uma condição é verificada direcionando a decisão do algoritmo entre nós até o final da árvore.

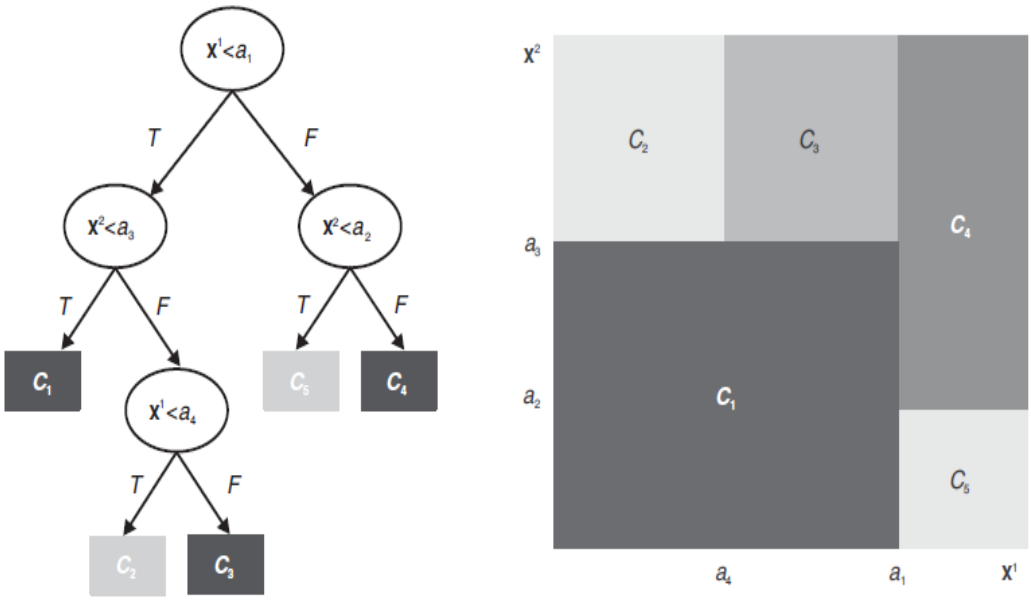
Segundo Faceli et.al. (2021), uma árvore de decisão é um grafo direcionado e acíclico. Os nós podem ter um (nó folha) ou mais sucessores (nó de divisão). Um nó folha é identificado como uma função que consiste em uma constante que minimiza uma função de custo, normalmente essa constante é a moda e em árvores de regressão, como o *random forest regression*, minimiza o erro médio quadrático e o desvio absoluto através da média e mediana respectivamente. Nos nós de divisão um teste condicional é aplicado aos atributos direcionando a árvore por um caminho ou outro.

É possível analisar essa estrutura a partir do exemplo dado por Faceli et.al. através da figura a seguir, onde ela diz que:

> A figura representa uma árvore de decisão e a divisão correspondente no espaço definido pelos atributos x1 e x2. Cada nó da árvore corresponde a uma região nesse espaço. As regiões definidas pelas folhas da árvore são mutuamente excludentes, e a reunião dessas regiões cobre todo o

> espaço definido pelos atributos. A interseção das regiões abrangidas por quaisquer duas folhas é vazia. A união de todas as regiões (todas as folhas) é U. Uma árvore de decisão abrange todo o espaço de instâncias. Esse fato implica que uma árvore de decisão pode fazer predições para qualquer exemplo de entrada (FACELI et.al. 2021, p. 78).

Figura 4 – Estrutura da árvore de decisão



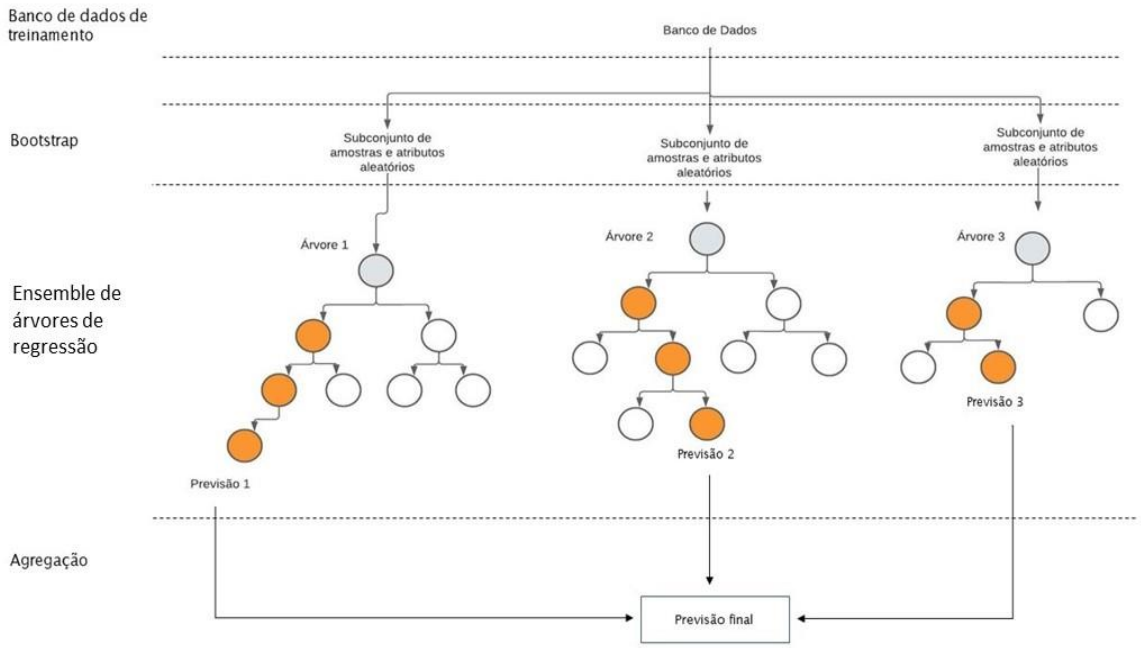
Fonte: Faceli et.al (2021).

2.2.2 Random forest regression

Segundo Dietterich et. al. apud Ponte (2020), o *random forest* é um algoritmo de aprendizado de máquina que se baseia em árvores de decisão, criando florestas compostas cada uma por várias árvores, dessa forma a previsão será melhor do que feita apenas por uma árvore de decisão. Em uma *random forest* há também a aleatoriedade de atributos para cada árvore, isso traz para a floresta uma redução do *overfitting*, muito presente em uma árvore de decisão.

Esse tipo de algoritmo tem origem de um método conhecido como *ensemble*, que consiste em um conjunto de vários algoritmos, do mesmo tipo ou diferentes, contribuindo para a previsão final. O *random forest* utiliza uma técnica denominada *bootstrap* que busca, a partir dos dados originais, criar subconjuntos de amostras com reposição, o objetivo é diminuir a variância do estimador. Outra característica de destaque é que cada árvore produzida traz atributos preditores aleatórios, esse processo ajuda a reduzir o *overfitting*, dificultando que o algoritmo decore os dados de treinamento. Ao final do processo é construída a média previsão do conjunto dos algoritmos e pode-se fazer uma estimativa de performance através do erro médio quadrático (MSE), erro médio absoluto (MAE) e coeficiente de determinação (R^2) (PONTE et. al, 2020).

Figura 5 – Exemplo de estrutura da random forest



Fonte: Elaborado pelo autor.

No aprendizado de máquina a avaliação da performance é um ponto de extremo cuidado e as métricas de desempenho trazem uma boa noção da acurácia do modelo. O erro médio absoluto calculado a partir da média absoluta dos erros, no qual se está preocupado apenas com o módulo da diferença do real para o que foi previsto, é usado para medir a distância do valor real. No erro médio quadrático costuma penalizar mais o erro, pois cada valor é elevado ao quadrado tornando-o mais distante do real. O uso de cada uma das métricas depende se o usuário está preocupado em punir valores que se afastem muito da média.

Abaixo estão as equações que descrevem os erros:

Equação 1 – Erro médio absoluto (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|Yi - \hat{Y}i\right| \quad (1)$$

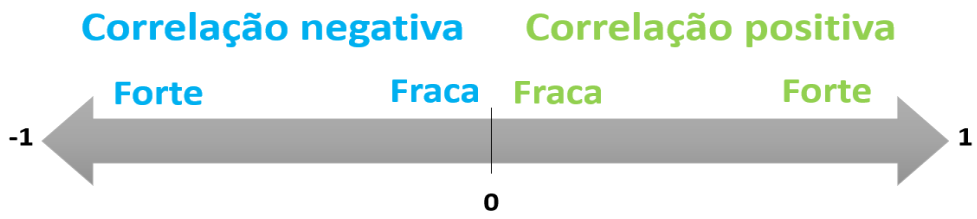
Equação 2 – Erro médio quadrático (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Yi - \hat{Y}i)^2 \quad (2)$$

Outra métrica útil na avaliação do modelo de aprendizado de máquina é o coeficiente de determinação (R^2), que nada mais é do que o quadrado do coeficiente de

correlação de Pearson. Essa correlação é um grau de relação entre duas variáveis quantitativas com valores entre -1 e 1. Valores que tendem a se aproximar de 1 define uma correlação positiva, informando que quando o valor de uma variável aumenta a outra também aumentará. Valores abaixo de zero implicam em uma correlação negativa, definindo uma relação inversamente proporcional. Uma correlação de valor 0 define que não há relação entre as variáveis.

Figura 6 – Coeficiente de Correlação



Fonte: Elaborado pelo autor.

3. Metodologia

O presente trabalho é de caráter exploratório e está baseado em fontes secundárias dentre as quais: livros, artigos, teses e publicações relacionadas ao assunto escritos em língua portuguesa e inglesa.

Foi utilizado como objeto de estudo prático uma base de dados, na qual constam várias amostras de misturas de concreto, idade e resistência final. Por conta da carência de dados reais online referente ao assunto, foi usado uma base de dados do site Kaggle, que é uma plataforma onde é disposto dados de diferentes naturezas e que pessoas que estão aprendendo sobre esse assunto podem analisar esses dados e completar tarefas indicadas pelo próprio site ou por outros contribuintes da plataforma. Os dados encontrados no site na maioria das vezes são dados verídicos e alguns outros gerados artificialmente, neste estudo não se tem certeza da veracidade dos dados, entretanto eles servirão para treinar o algoritmo.

Para realizar essa análise é necessário um ambiente para a construção de todo o código. O google collaboratory ou google colab é uma plataforma na nuvem, desenvolvida para incentivar a pesquisa de aprendizado de máquina e inteligência artificial. Nesse ambiente o usuário consegue trabalhar de forma didática e escrever anotações sobre o código como em um caderno. A plataforma é ainda colaborativa e permite que o usuário compartilhe seu caderno de códigos com outros usuários que podem executar, analisar e modificar o que foi escrito. O google colab usa como código fonte o python, uma linguagem de programação relativamente nova, no entanto é possível utilizar outras, a escolha do usuário.

Na análise de dados a escolha da linguagem de programação é de extrema relevância, pois a quantidade de dados envolvida é enorme e é necessária uma interação veloz entre dados, linguagem e máquina. O python atualmente é considerado uma dessas linguagens e consegue fazer essa ponte com bastante eficiência. Seu desenvolvimento se deu a partir da ideia de ser uma linguagem de aprendizado fácil e que conseguisse resolver problemas com menos linhas de códigos comparada a outras concorrentes. Pode-se dizer que a comunidade é bastante ativa e a criação de bibliotecas que buscam resolver

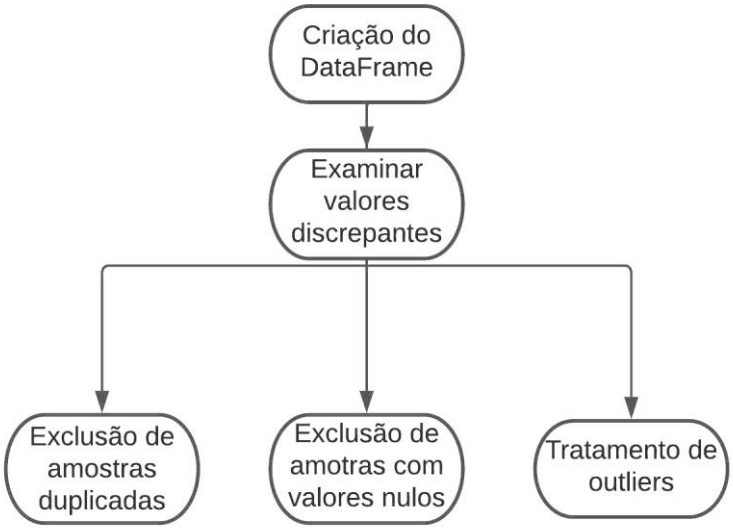
problemas corriqueiros na programação é em razão disso. Uma biblioteca em python é um pacote de várias funções que já resolvem um determinado problema sem que haja a necessidade de criar novamente as mesmas linhas de código.

Nessa pesquisa o ambiente usado para a construção do código foi o google colab com a base em python. Algumas bibliotecas como pandas, numpy, matplotlib e scikit learn foram utilizadas na construção do algoritmo de predição.

3.1 Análise exploratória de dados

O primeiro passo foi criar um dataframe através da biblioteca pandas denominado *concreto*, dentro desse foi alocado o banco de dados tabulado. A biblioteca pandas é capaz de criar a estrutura, manipular e limpar dados do data frame. Após ter criado a estrutura foi feita uma análise exploratória dos dados, como mostrado no seguinte fluxograma:

Figura 4 - Fluxograma análise exploratória de dados



Fonte: Elaborado pelo autor.

Para encontrar os valores duplicados do banco de dados foi utilizado algumas funções da biblioteca pandas. A função *duplicated* retorna os valores booleanos de cada linha dos dados, demonstrando a existência de valores duplicados (True) ou não duplicados (False), porém apenas a visualização desses valores é complicada e não dá uma análise completa da quantidade de duplicados. Atribuir essa função em uma variável permite usar outra função para somar esses valores, havendo algum valor duplicado podemos descartá-lo através da função *drop_duplicated*, também da biblioteca pandas.

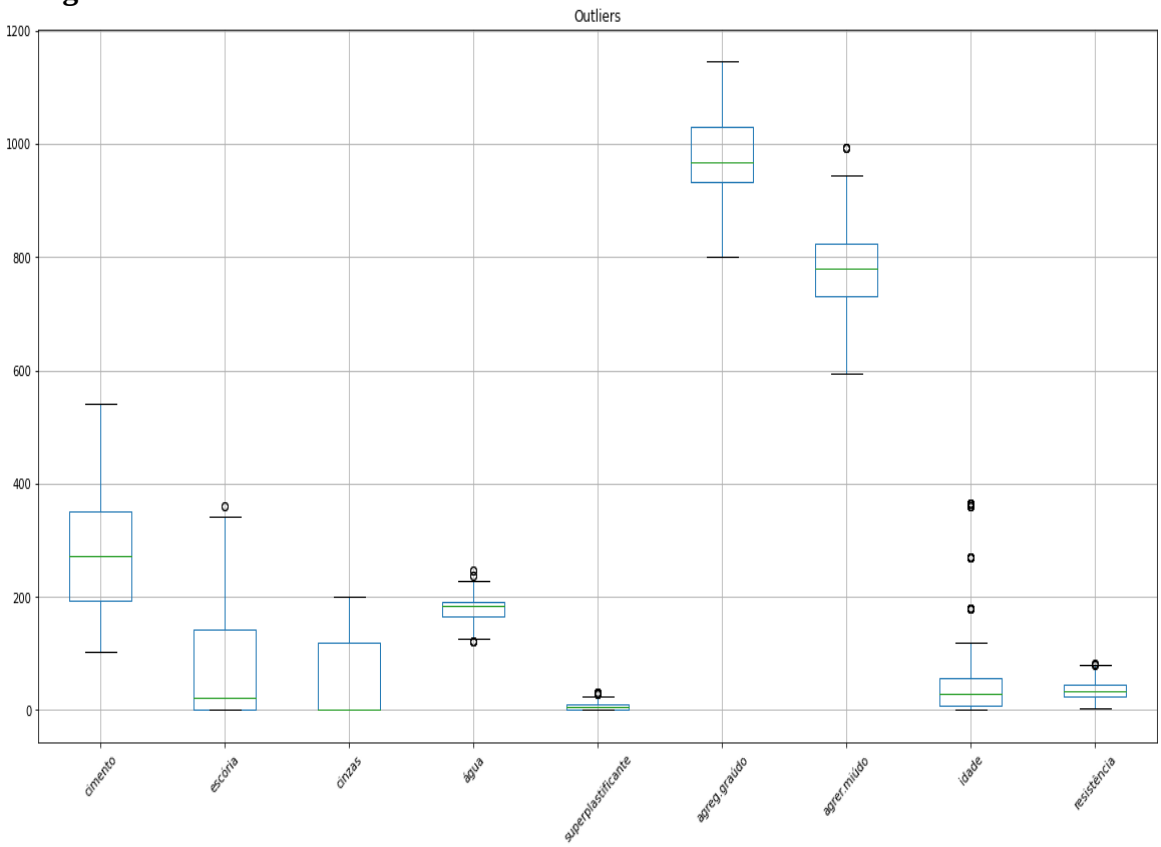
Valores nulos podem seguir a mesma ideia, porém com a função isnull que retorna valores nulos, atribuir ela a uma outra variável também permite fazer a contagem e descartar através da função *dropna*.

Para a extração de valores outliers a complexidade é um pouco maior. Foi utilizado o método de Tukey, que consiste em encontrar os limites inferior e superior usando o intervalo interquartil (IQR) e o primeiro (Q1) e terceiro (Q3) quartis. Esses

quartis dividem os dados de cada coluna do dataframe e demonstram a dispersão dos dados.

Utilizando a biblioteca seaborn para criar um gráfico boxplot é possível ver os valores outliers de suas respectivas colunas do dataframe.

Figura 5 – Antes do tratamento de outliers



Fonte: Elaborado pelo autor.

Para encontrar o intervalo interquartil foi utilizado uma função do pandas chamada *quantile*, que retorna os quantis de acordo com os parâmetros que dividem os quartis 0.25 e 0.75 respectivamente. Esses valores foram atribuídos às variáveis Q1 e Q3.

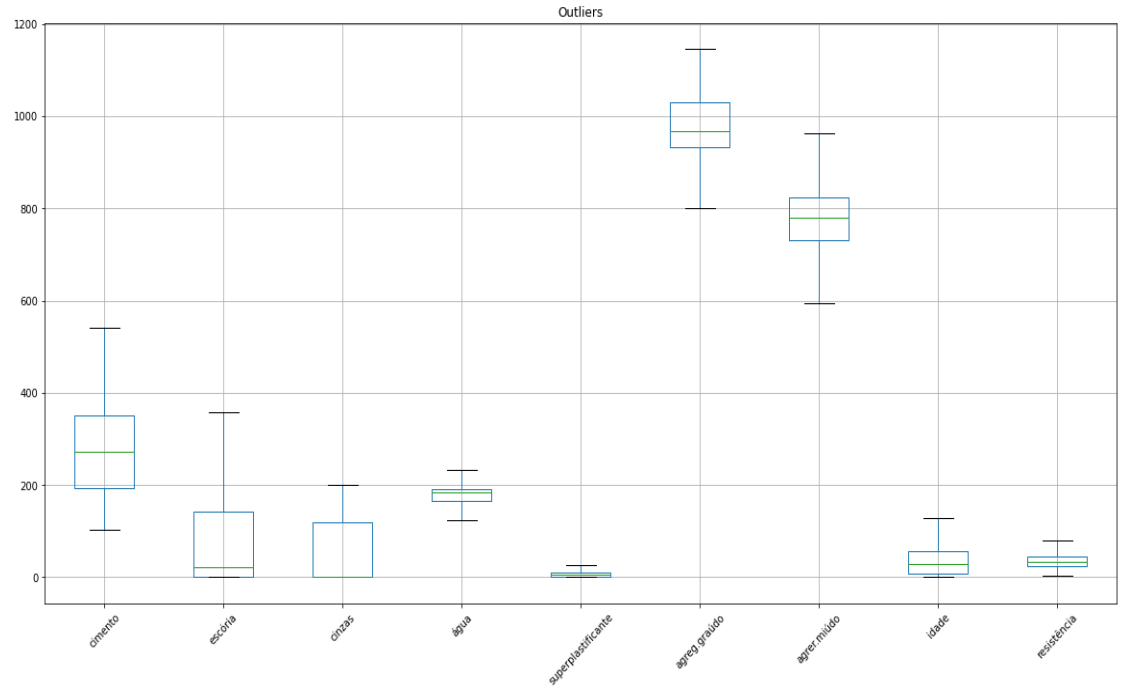
Para os limites inferior e superior foi realizado o seguinte cálculo:

$$\text{Limite inferior} = Q1 - (1{,}5 * IQR)$$

$$\text{Limite superior} = Q3 + (1{,}5 * IQR)$$

Após essa observação pode-se então fazer o tratamento dos dados discrepantes, substituindo esses valores pelos limites calculados anteriormente através de uma estrutura de repetição do python, dada pela função *for*, que varre todas as linhas da base de dados e substitui os valores quando eles estão abaixo ou acima dos limites.

Figura 6 – Depois do tratamento de outliers



Fonte: Elaborado pelo autor.

Após o tratamento de outliers foi construído um mapa de calor para entender a correlação entre as variáveis, através da biblioteca seaborn com a função *heatmap*.

Figura 7 – Mapa de calor

Fonte: Elaborado pelo autor.

3.2 Algoritmo de aprendizado de máquina

Nessa fase da pesquisa foi criado e treinado o algoritmo de aprendizado de máquina. Para tanto a biblioteca scikit learn, especializada em vários modelos desse método, foi utilizada. Dentre os inúmeros algoritmos de AM, o *random forest regression* foi escolhido pelo motivo de minimizar o *overfitting* do modelo.

O primeiro passo na construção do algoritmo é fazer a divisão da base de dados entre atributos previsores e o atributo alvo. Para isso, cria-se duas variáveis x e y e atribuímos à primeira todas as colunas da base com exceção da coluna contendo o atributo alvo resistência, essa atribuição é possível pelo comando drop da biblioteca pandas que descarta apenas a coluna escolhida. A atribuição do atributo alvo à variável y se deu apenas pelo comando de "chamar" uma coluna específica do dataframe *concreto*. As duas linhas de código são escritas da seguinte forma:

```
X = concreto.drop('resistência', axis=1)

y = concreto['resistência']
```

Após essa separação dos dados aplicamos a divisão entre dados de treinamento e teste através do comando *train_test_split* da biblioteca scikit learn. Com esse comando é possível fazer a divisão dos dados em qualquer proporção, outro parâmetro é o de aleatoriedade denominado *random_state*, com ele podemos duplicar a mesma aleatoriedade da divisão em outros cadernos. A proporção de divisão dos dados foi de 70% treinamento e 30% para teste do algoritmo, o código foi escrito da seguinte forma:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

A próxima etapa foi atribuir o algoritmo *random forest regression* à uma variável denominada *modelo*. Os parâmetros usados na criação do modelo foram *n_estimators e random_state*. O primeiro determina quantas árvores serão criadas e o segundo a aleatoriedade como explicado anteriormente. Ao final do processo pode-se alocar os dados de treinamento dentro do modelo como mostra o código abaixo:

```
modelo = RandomForestRegressor(n_estimators=20000, random_state=0)

modelo.fit(X_train, y_train)
```

Na última etapa foi usada a função *predict* do algoritmo para prever os dados de teste separados anteriormente, o parâmetro dessa função é apenas o conjunto de dados teste. Essa função foi atribuída a uma variável denominada *previsao* e retorna valores que o algoritmo tentou prever. Para isso o código abaixo foi escrito:

```
previsao = modelo.predict(X_test)
```

3.2.3 Avaliação do algoritmo

Nessa seção foi realizada a avaliação do modelo de aprendizado de máquina a partir das métricas citadas anteriormente. A raiz do erro médio quadrático (RMSE) e o coeficiente de determinação foram escolhidos por se encaixarem bem na estimativa de performance desse algoritmo de regressão. A RMSE é apenas uma forma para comparar modelos sem se preocupar com a escala entre eles, sua base é o MSE. A função *sqrt* (raiz

quadrada) da biblioteca numpy foi usada e seu parâmetro foi apenas o MSE da própria biblioteca scikit learn junto aos dados de teste e os previstos. O coeficiente de determinação foi usado através da função *r2_score*, também do scikit learn, passando dados de teste e previsão para a estimativa. O código abaixo foi utilizado para obter a performance do modelo:

```
np.sqrt(mean_squared_error(y_test, previsao))

r2_score(y_test, previsao)
```

4. Resultados e Discussões

O banco de dados usado no trabalho possuía 25 amostras duplicadas e nenhum valor nulo, essas amostras foram excluídas proporcionando uma base de dados mais limpa. Foi notado também alguns valores outliers, que foram substituídos pelos limites superior e inferior. Todo esse tratamento dos dados traz ao modelo de aprendizado de máquina melhor entendimento e menos ruído de dados discrepantes.

Através da máquina preditiva construída, o objetivo proposto de identificar a resistência futura do concreto teve resultados satisfatórios. As métricas usadas para a avaliação do modelo atingiram valores altos, o RMSE alcançou um valor de apenas 5.13, lembrando que quanto menor esse valor melhor será o modelo. O coeficiente de determinação R^2 foi de 90%, ou seja, apenas 10% dos dados não conseguem ser explicados pelo modelo.

É possível entender também a correlação entre as variáveis e sua relação com o atributo alvo. O mapa de calor construído anteriormente demonstra uma correlação de 0.5 da variável cimento com a resistência, dando ao cimento um alto impacto na mistura do concreto. Outra informação importante que pode ser tirada desse mapa é a relação entre a idade do concreto e a resistência em um valor de 0.47. A água também é relevante na mistura com um grau de correlação de 0.29. Um caso de destaque nesse mapa é visto entre a variável superplastificantes e a água com o maior valor de correlação 0.67, isso é explicado pelo fato desse aditivo ser usado justamente para diminuir o volume de água na mistura de concreto.

Houve também a análise da importância de cada variável para o modelo e os seguintes percentuais foram encontrados:

Tabela 2 – Rank de importância dos atributos

Atributo	Peso
Idade	35%
Cimento	29%
Água	11%
Superplastificante	9%
Escória	7%
Agregado miúdo	4%
Agregado graúdo	3%
Cinzas	2%

Fonte: Elaborado pelo autor.

A tabela acima obtida através de função da biblioteca scikit learn, consegue explicar o peso de cada variável para o modelo. Os três maiores percentuais estão relacionados com a idade, cimento e água. Essa tabela remete as principais características dessas variáveis, citadas anteriormente neste trabalho.

5. Conclusão

Conforme foi analisado durante todo o trabalho, é possível afirmar que esse tipo de tecnologia tem grandes impactos na construção civil. O modelo de aprendizado de máquina conseguiu com grande acurácia explicar 90% dos valores no banco de dados, no entanto, as medidas de segurança na construção são rígidas para evitar ao máximo acidentes. A substituição do método atual de constatação do *fck* do concreto pelo modelo estudado aqui é improvável, porém esse tipo de proposta pode auxiliar todo o processo de antemão reduzindo custos no controle tecnológico do concreto, afinal sabendo dos possíveis resultados de misturas variadas o responsável pode se antecipar a situações adversas.

O modelo conseguiu também entender a relevância de cada um dos componentes da mistura de concreto, trazendo o peso percentual do componente na identificação da resistência e percepções da relação dos materiais entre si.

É necessário dizer que os algoritmos de aprendizado de máquina podem ser aperfeiçoados através de técnicas de estatística e com um tratamento da base de dados melhorado, isso poderia trazer resultados mais assertivos.

Esse estudo consegue também demonstrar o poder da análise de dados aplicado a engenharia civil. Entender uma base de dados com milhares de valores tornou-se fácil e os ambientes para construir esse conhecimento, acessíveis.

Como citado anteriormente, no Brasil esse tipo de abordagem na indústria da construção civil é ínfimo. Entretanto, os estudos sobre o assunto se tornam cada vez maiores e isso pode alavancar no futuro uma integração maior do aprendizado de máquina nessa indústria.

Agradecimentos

Agradeço aos meus pais por todo o apoio durante a minha graduação e a minha mulher e meu filho por suportarem meu humor durante a construção desse trabalho. Agradeço aos professores que colaboraram na construção do conhecimento necessário para chegar aqui. Aos meus amigos, colegas de curso e república, muito obrigado.

REFERÊNCIA

ABIKO, A. K.; GONÇALVES, O. M. O futuro da construção civil no Brasil. Resultados de um estudo de prospecção tecnológica da cadeia produtiva da construção habitacional. Escola Politécnica da Universidade de São Paulo- 2003.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. (1994). *NBR 5738: Moldagem e cura de corpos-de-prova.* Rio de Janeiro.

ARSHDEEP, B. MADISETTI, V. Big Data Analytics: A Hands-On Approach. 2019. ISBN: 978-1-949978-00-1

CAVALCANTE, C. G. Sá; DOMINGUES, De A. T. Os benefícios da Indústria 4.0 no gerenciamento das empresas. 2017

CAVALCANTI, V. Y. S. de Lima. et a. Indústria 4.0: Desafios e Perspectivas na Construção Civil.

DA SILVA, A. D; Dos Santos Simão-Orientadora, Alessandra; Gabriel Menezes-Co-Orientador, Carlos Augusto. Impactos da Indústria 4.0 na Construção Civil brasileira. 2018

FACELI, Katti; Lorena, Ana C.; Gama, João; et.al. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: Grupo GEN, 2021. E-book. ISBN 9788521637509. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788521637509/. Acesso em: 10 nov. 2022.

FAVERO, Luiz P. Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Grupo GEN, 2017. E-book. ISBN 9788595155602. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595155602/. Acesso em: 25 set. 2022.

FOGLIATTO, F. S.; RIBEIRO, J. L. D. Confiabilidade e Manutenção Industrial. Rio de Janeiro: Elsevier, 2009.

HELENE, Paulo; SILVA FILHO, L. C. Análise de estruturas de concreto com problemas de resistência e fissuração. Concreto: Ciência e Tecnologia, v. 2, 2011.

KAMIŃSKI, B. Jakubczyk, M.; Szufel, P. (2017). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*.

KUDYBA, *S* et.al. Big Data, Mining, and Analytics: Components of strategic decision making. 1º Edição. Nova York: Auerbach Publications, 2014.

Luiz Carlos Pinto da Silva Filho, Paulo Helene. Análise de Estruturas de Concreto com Problemas de Resistência e Fissuração. 2011.

MAISUECHE, Cuadrado Alberto. UTILIZACIÓN DEL MACHINE LEARNING EN LA INDUSTRIA 4.0. Valladolid, septiembre, 2019

MEHTA, Povindar Kumar; MONTEIRO, Paulo J. M. Concreto: Microestrutura, Propriedades e Materiais. 3. ed. São Paulo: Ibracon - Instituto Brasileiro de Concreto, 2008.

MUELLER, John P.; MASSARON, Luca. Aprendizado de Máquina Para Leigos. Rio de Janeiro: Editora Alta Books, 2019. E-book. ISBN 9788550809250. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788550809250/. Acesso em: 17 nov. 2022.

NEVILLE, A.M. Propriedades do Concreto. Porto Alegre: Grupo A, 2016. E-book. ISBN 9788582603666. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788582603666/. Acesso em: 16 nov. 2022.

PARIZOTTO, Liana. Concreto Armado. Porto Alegre: Grupo A, 2017. E-book. ISBN 9788595020917. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595020917/. Acesso em: 16 nov. 2022.

PINHEIRO, Libânio M. Fundamentos do concreto e projeto de edifícios. 2007.

PONTE, Caio; CAMINHA, Carlos; FURTADO, Vasco. Otimização de florestas aleatórias através de ponderação de folhas em árvore de regressão. In: Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2020. p. 698-708.

SILVA, Larissa Eterna Taveira et al. Tecnologias digitais utilizadas pela indústria da construção civil. SIMPÓSIO BRASILEIRO DE GESTÃO E ECONOMIA DA CONSTRUÇÃO, v. 12, p. 1-8, 2021.

VELOSO, Lee. Análise Preditiva: benefícios, exemplos e como implementá-la. Moki, 2021. Disponível em: https://site.moki.com.br/analise-preditiva/. Acesso em: 24 de novembro de 2022.