UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Otimização do Processo de Moagem de Cimento

Enrique Guevara Salas

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Otimização do Processo de Moagem de Cimento

Enrique Guevara Salas

USP - São Carlos

2024

Enrique Guevara Salas

Otimização do Processo de Moagem de Cimento

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof. Dr. Odemir Martinez Bruno

USP - São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G939o

Guevara Salas, Enrique
Otimização do Processo de Moagem de Cimento / Enrique Guevara Salas; orientador Odemir Martinez Bruno. -- São Carlos, 2024.
72 p.

Tese (Doutorado - MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. MBA em Inteligência Artificial e Big Data. 2. Trabalho de Conclusão de Curso. 3. Otimização do Processo de Moagem de Cimento. 4. Enrique Guevara Salas. 5. Prof. Dr. Odemir Martinez Bruno. I. Martinez Bruno, Odemir , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

A minha esposa e aos meus filhos pela compreensão, carinho e apoio incansável.

AGRADECIMENTOS

Gostaria de agradecer ao Professor Odemir Martinez pelo valioso orientação e apoio, essenciais para o desenvolvimento e conclusão deste trabalho.

Aos professores, funcionários e monitores do Instituto de Matemática e Computação da Universidade de São Paulo, ICMC-USP.

Sou imensamente grato à minha querida esposa Lizeth e aos meus filhos Maria Eduarda e Ayrton por toda sua paciência, compreensão, carinho e amor.

E por fim, gostaria de agradecer à minha mãe Clara pelos ensinamentos, pela dedicação e paciência, em casa e na escola, que me tornaram o homem que sou hoje.

Nada na vida deve ser temido, apenas compreendido. Agora é a hora de entender mais, para que possamos temer menos.

Marie Curie (1903)

RESUMO

GUEVARA, E. **Otimização do Processo de Moagem de Cimento**. 2024. 73 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

A produção de cimento é uma das atividades mais intensivas em consumo de recursos e energia, resultando em elevados impactos ambientais, como emissões de CO2 e a extração de matérias-primas não renováveis. O concreto, por sua vez, é um dos materiais de construção mais utilizados globalmente, e sua produção segue crescendo em resposta à demanda populacional. O desafio enfrentado pelas indústrias de cimento é a busca por processos mais eficientes e sustentáveis, especialmente na etapa de moagem, onde o índice de Blaine (finura do cimento) é um fator crítico para garantir a qualidade do produto final. O presente estudo visa desenvolver um modelo prescritivo de otimização que permita estimar com maior precisão o valor de Blaine em tempo real, minimizando o tempo de resposta e os desvios de qualidade. A metodologia adotada foca na aplicação de aprendizado de máquina, especialmente modelos de regressão e aprendizado por reforço, para identificar e prever combinações ideais de variáveis operacionais que garantam a estabilidade do processo de produção de cimento. Na fase inicial de modelagem, foram construídos diferentes modelos de regressão, incluindo regressão linear, Ridge, Lasso, e modelos de árvores de decisão e de boosting. Cada modelo foi avaliado com base em métricas como R^2 e RMSE, sendo o modelo de regressão linear escolhido por sua consistência e explicabilidade. O modelo apresentou uma margem de erro de 4,4% na previsão do valor de Blaine. Em seguida, foi implementado o modelo de aprendizado por reforço deep deterministic policy gradient (DDPG), que se mostrou eficiente para ajustar variáveis contínuas em um ambiente de controle industrial. O DDPG foi utilizado para recomendar as melhores ações operacionais a fim de manter o valor de Blaine dentro da faixa ideal (4700-4750). Esse modelo, ao interagir com o modelo de regressão, permitiu uma redução da margem de erro para 3,6% durante o processo de validação, mostrando potencial para otimizar o controle da qualidade em tempo real. A combinação de modelos de regressão e aprendizado por reforço oferece uma solução robusta para otimizar o processo de moagem de cimento, reduzindo erros operacionais e melhorando a eficiência energética.

Palavras-chave: Modelos de regressão; aprendizagem por reforço; cimento; Blaine.

ABSTRACT

GUEVARA, E. **Optimization of Cement Grinding Process**: subtitle. 2024. 73 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

The production of cement is one of the most intensive activities in the consumption of resources and energy, resulting in high environmental impacts, such as CO2 emissions and the extraction of non-renewable raw materials. Specifically, in turn, it is one of the two construction materials most used globally, and its production continues to grow in response to population demand. The challenge faced by the cement industries is the search for more efficient and sustainable reprocesses, especially in the molding stage, where the Blaine index (fineness of the cement) is a critical factor to guarantee the quality of the final product. This study aims to develop a prescriptive optimization model that allows to estimate with greater precision the value of Blaine in real time, minimizing the response time and quality deviations. The adopted methodology focuses on the application of reinforcement learning, especially regression models and reinforcement learning, to identify and predict ideal combinations of various operations that guarantee the stability of the cement production process. In the initial modeling phase, different regression models were built, including linear regression, Ridge, Lasso, and decision tree and boosting models. Each model was evaluated based on metrics such as R^2 and RMSE, with the linear regression model chosen for its consistency and explainability. The model presents a margin of error of 4,4% in the forecast of Blaine's value. Next, the learning model was implemented by reinforcing deep deterministic policy gradient (DDPG), which was shown to be efficient in adjusting continuous variables in an industrial control environment. The DDPG was used to recommend the best operational actions to maintain the Blaine value within the ideal range (4700-4750). This model, which interacts with the regression model, allows a reduction of the margin of error to 3,6% during the validation process, showing potential to optimize or control quality in real time. The combination of regression and reinforcement learning models offers a robust solution to optimize the foundation mowing process, reducing operational errors and improving energy efficiency.

Keywords: Regression models; reinforcement learning; cement; Blaine.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Elemento opcional, elaborada seguindo a mesma ordem apresentada no texto com cada item designado por seu nome e respectivo número de página.

LISTA DE ABREVIATURAS E SIGLAS

RL	–	Reinforcement Learning (Aprendizagem por Reforço)
Train	–	Dados de treinamento
Test	–	Data do teste
Val	–	Dados de validação
DDPG	–	Deep Deterministic Policy Gradient
SVM	–	Suport Vector Machine
$R2$	–	Coeficiente de Determinação
RMSE	–	Raiz do Erro Quadrático Médio
DQN	–	Aprendizagem por reforço profundo
GCP	–	Google Cloud Platform
RDV	–	Raw Data Vault
UDV	–	Universal Data Vault
DDV	–	Description Data Vault
IQR	–	Intervalo interquartil

SUMÁRIO

1 **INTRODUÇÃO**

1.1 Contextualização do problema

Depois da água, o concreto é a substância mais utilizada no mundo, por exemplo em edifícios, estradas, barragens, etc. A sua produção quadruplicou nos últimos 30 anos e a sua procura continuará a crescer dado o crescimento populacional em todo o mundo (Bamigboye et al., 2018).

O concreto é o material de construção mais utilizado no mundo, pelas vantagens que apresenta em resistência, durabilidade, entre outras, em comparação a outros materiais. Estima-se que o concreto convencional seja produzido cerca de 6 bilhões de toneladas por ano no mundo. À medida que a demanda pela oferta de cimento cresce continuamente, isso provoca um aumento na utilização de agregados, principalmente calcário, por ser importante na produção de cimento Portland (Soltannzadeh et al., 2018).

Nele são utilizados diferentes elementos como água, areia, brita e cimento, este último representa aproximadamente 10% do concreto, e é o que mais utiliza recursos para sua produção. A indústria do cimento é uma das indústrias do mundo que mais consome energia e recursos, bem como um dos principais geradores de CO2 e outros efeitos ambientais.

Tendo em conta que esta atividade produtiva implica uma extração crescente de matérias-primas e, portanto, uma redução destes recursos não renováveis, além desta extração implica um impacto no ambiente e exposição ao risco de esgotamento no futuro (Mohamad et al., 2021). Embora esses impactos não possam ser eliminados, eles podem ser reduzidos tornando o processo produtivo mais eficiente.

Nas diferentes fases do processo produtivo na fábrica, são liberados poeira, ruído e gases de efeito estufa, principalmente dióxido de carbono. Este tipo de problema ambiental pode agravar a qualidade de vida e os desequilíbrios nas zonas envolventes. Portanto, é necessário aprimorar as tecnologias utilizadas no processo produtivo da planta, para garantir uma produção mais limpa e otimização de recursos (Mohamad et al., 2021).

No nível regulatório, esta indústria deve atender à qualidade necessária de resistência e durabilidade para garantir a integridade de seus clientes e evitar desastres e perdas econômicas (Mohamad et al., 2021).

1.2 Justificativa

Tendo em conta o exposto, este é um processo produtivo tradicional que, embora sejam feitos esforços para reduzir a pegada de carbono, isso também implica maiores investimentos e custos para fazer as implementações necessárias, mas a empresa também procura fazer otimizações e fazer o processo mais eficiente, o que provocaria reduções no impacto da pegada de carbono, otimizando o uso de matérias-primas e energia como carvão e petróleo. Este projeto visa utilizar ferramentas de inteligência artificial que permitam a empresa encontrar essas lacunas no processo produtivo para manter controlado o processo selecionado neste caso: moagem de cimento.

Portanto, o foco está em como contribuir para a otimização do processo e reduzir o consumo de energia e matérias-primas de forma que gere maiores lucros para a empresa, além de reduzir o impacto ao meio ambiente.

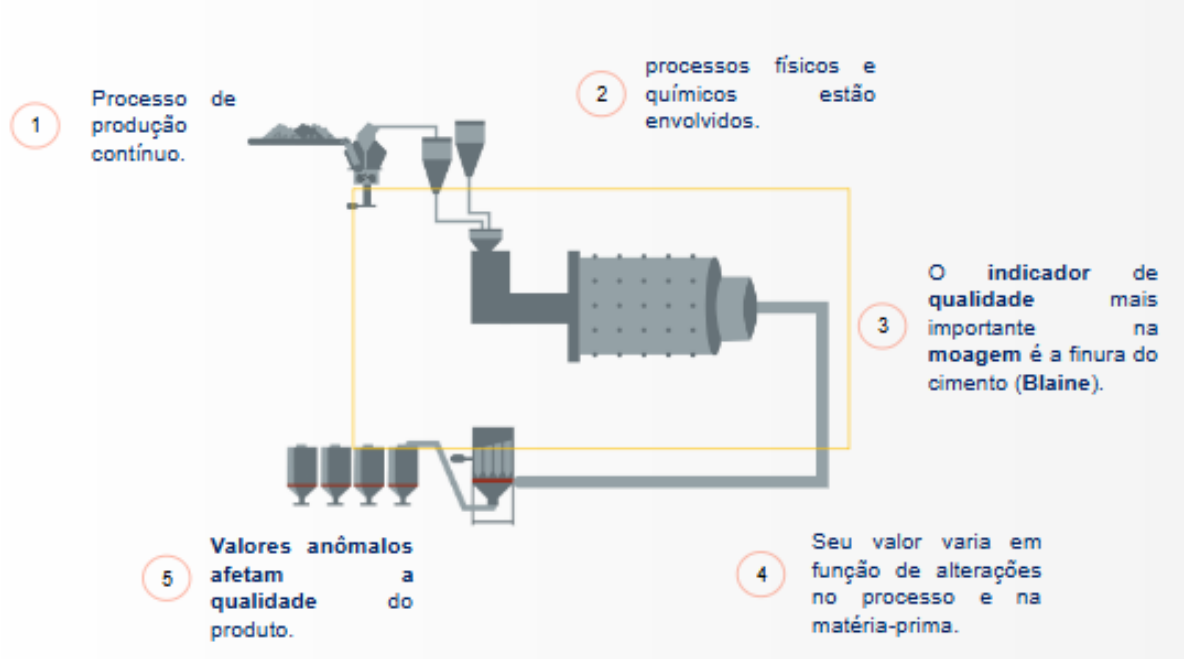
Entrando em detalhes, o ciclo de produção de cimento possui 5 macroprocessos como 1) Extração de matéria-prima, 2) Moagem e homogeneização das matérias-primas, 3) Clinquerização, 4) Moagem do clínquer, 5) Expedir.

Figura 1 – Processo de produção em fábrica de cimento.

Fonte: Elaborador pelo autor (2024).

O processo de moagem de cimento é um fluxo de produção contínuo no qual estão envolvidos diversos equipamentos mecânicos e elétricos, matérias-primas como calcário e gesso e materiais intermediários como o clínquer. Neste processo, um indicador chave da qualidade do cimento produzido é a finura do cimento, que é chamado de “Blaine” ou “superfície específica”.

Figura 2 – Processo de moagem de cimento no moinho 3.



Fonte: Elaborador pelo autor (2024).

1.3 Problema de pesquisa

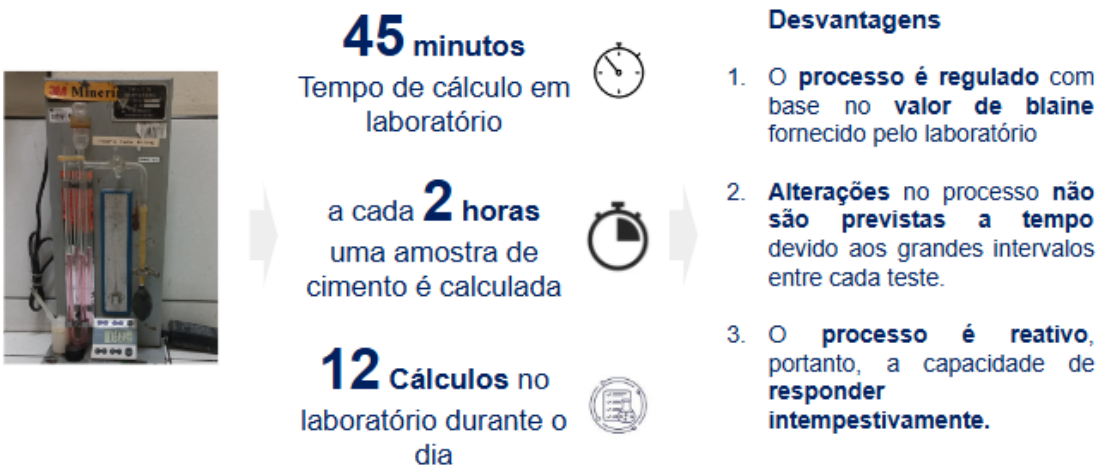
O controle da moagem de cimento não é totalmente eficiente porque ainda existem processos que dependem da supervisão humana. Por exemplo, o indicador Blaine (que regula a qualidade do cimento) não é obtido como resultado de um processo automatizado em tempo real, pois existem restrições operacionais como tempos que envolvem a retirada da amostra e que sejam realizados sequencialmente outros testes laboratoriais que regulam a qualidade dos processos intermediários na produção de cimento. Por isso, cada exame laboratorial possui tempos de avaliação próprios. Conforme indicado, o valor de Blaine é obtido a cada 2 horas, horário em que os resultados poderiam ter sido alterados e saído das faixas de qualidade. Este processo é detalhado a seguir:

- Atualmente, a equipe de controle de qualidade realiza o cálculo de Blaine através de testes laboratoriais, retirando uma amostra de cimento na saída da produção, e esse teste dura em média 45 minutos.

- Esses testes são realizados a cada 2 horas, dando a Blaine 12 resultados por dia. O fluxo completo da avaliação de Blaine, além de contar com uma parte do processo operacional (coleta de amostra) e laboratorial, implica um desfasamento de tempo entre a coleta da

amostra e o resultado de Blaine. Portanto, caso ocorra alguma irregularidade no processo, é só poderá ser regularizado após 45 min (tempo de laboratório).

Figura 3 – Processo de controle e medição com equipamento Blaine.



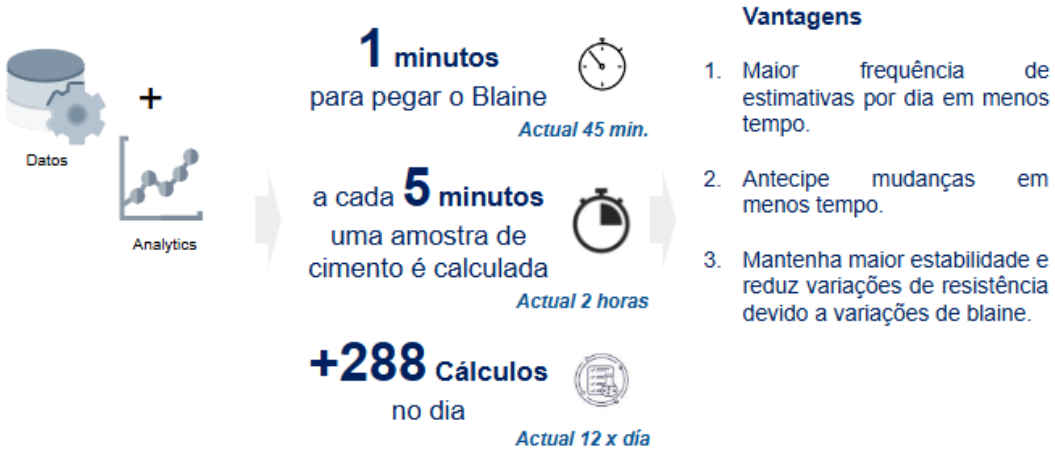
Fonte: Elaborador pelo autor (2024).

Este atraso no tempo de resposta implica que quando o resultado é obtido no Blaine (laboratório), as condições das variáveis envolvidas na moagem podem ter mudado, pelo que as decisões tomadas podem já não responder às condições atuais.

1.4 Objetivo

O objetivo desta pesquisa é o desenvolvimento de um modelo de otimização prescritivo em que com um primeiro modelo o valor de Blaine é estimado muito próximo daquele calculado em laboratório com margem de erro mínima aceitável e no menor tempo possível (em minutos). Com base nesta estimativa, um segundo modelo pode gerar cenários ótimos em que a fábrica deveria estar operando. Estas informações chegariam aos operadores do processo como recomendações para que possam fazer as alterações relevantes para garantir a qualidade do cimento.

Figura 4 – Variação do teor de gordura com tratamento térmico.



Fonte: Elaborador pelo autor (2024).

O valor do Blaine é fundamental porque está ligado às resistências iniciais do cimento e da pega. A resistência relacionada com questões regulatórias e comerciais e a configuração relacionada com o consumo de água na construção.

Garantir o valor do Blaine dentro de uma faixa ótima contribui para garantir uma elevada resistência inicial do concreto, o que gera benefícios para as construtoras além de cumprirem as normas regulamentadoras, que possam avançar mais rapidamente em sua atividade produtiva.

Em termos de produção, conseguir garantir a qualidade do cimento com um bom estimador Blaine evitaria o reprocessamento na produção, além de reduzir custos com consumo de matéria-prima, combustível e energia elétrica.

2 **REFERENCIAL TEÓRICO**

Sendo a produção de cimento uma atividade económica muito importante e de capital intensivo, a empresa deve garantir que os seus processos são rentáveis, pelo que qualquer meio que lhe permita melhorar o seu lucro será muito importante, por exemplo otimizando processos e reduzindo custos sem que isso afete negativamente a qualidade do produto (Condor et al., 2001).

2.1 Conceitos Básicos

Tendo em consideração as metodologias discutidas na seção anterior, e as vantagens e desvantagens de passar de um processo manual para um semiautomático e automático em alguns processos de produção, especialmente no processo de moagem de cimento, considera-se que ainda há espaço para melhoria

Portanto dentro da literatura sobre inteligência artificial podemos encontrar diferentes opções que nos permitiriam superar as dificuldades das lacunas de informação entre teste e teste (a cada 2 horas) e adicionalmente o tempo envolvido nos testes de cálculo do Blaine, é que opções como regressão.

Os modelos aparecem dentro do mundo dos modelos paramétricos e não paramétricos como árvores de decisão, boosting, redes neurais entre outros, mas que até certo ponto são modelos preditivos.

Este trabalho busca explorar opções que permitam a esta indústria tornar o processo muito mais automatizado. Embora existam ferramentas automáticas como os PLC que são os controladores dos equipamentos, eles não possuem inteligência para decidir mudanças, mas são baseados em cronogramas, além disso, eles não podem fazer estimativas (Condor et al., 2001).

Portanto, a proposta é baseada em sistemas que tenham a capacidade de aprender e tomar decisões automaticamente sobre alterações nos principais componentes envolvidos no processo de moagem, a fim de garantir que o nível de finura do cimento ou Blaine seja mantido nos padrões de qualidade e economizado. recursos como matérias-primas e energia.

No mercado existem soluções que visam a aplicação de automação industrial. Tal caso pode ser visto com ferramentas como o INTOUCH, que é um software de automação industrial

que permite trabalhar com bancos de dados, realizar processamento em lotes, além de aplicações de internet. Isto permite simulações dos diferentes subprocessos e equipamentos envolvidos na moagem de cimento (Condor et al., 2001).

A simulação do processo é baseada na programação de scripts, em que os comandos estão vinculados aos botões de controle. Mas dentro do software devem ser marcadas as especificações de cada um dos subprocessos para que o programa possa realizar o cálculo de ajuste do equipamento (Condor et al., 2001).

Para resolver este problema de otimização e controle de qualidade no processo de moagem de cimento, diversas metodologias podem ser consideradas, incluindo matemática, estatística, aprendizado de máquina e IA. Algumas alternativas exploradas:

a. Controle estatístico de processo (CEP):

- Utiliza técnicas de controle de processo, como gráficos de controle, para monitorar e manter o Blaine dentro da faixa alvo. Isso envolve coletar dados do processo de moagem e analisá-los para detectar desvios do alvo.

b. Modelagem preditiva:

- Você pode desenvolver modelos preditivos para prever o valor de Blaine com base nas configurações do equipamento de processo.
- Algoritmos como regressão linear, regressão logística ou modelos de árvore de decisão podem ser úteis aqui.

c. Otimização:

- Emprega técnicas de otimização para encontrar configurações ideais de equipamento que maximizem Blaine dentro da faixa alvo.
- Algoritmos como programação linear, programação quadrática ou algoritmos genéticos podem ser úteis dependendo da complexidade do problema.

d. Aprendizagem por reforço:

- Se o processo de fresagem for dinâmico e for necessário um controle adaptativo, o aprendizado por reforço pode ser uma opção.
- O sistema aprende de forma dinâmica e autônoma por meio da interação com o ambiente. Esses algoritmos podem ajudar a encontrar políticas ideais para ajustar os valores dos equipamentos com base nas observações de Blaine.
- Algoritmos como Q-learning ou DQN podem ser úteis para aprender uma política de controle ideal para ajustar equipamentos em tempo real.

e. Inteligência artificial e robótica:

- Embora possa ser uma abordagem mais complexa e não necessariamente aplicável em todos os casos, você pode considerar técnicas avançadas de inteligência artificial e robótica para automatizar o processo de ajuste de equipamentos online.

2.2 Fundamentos teóricos

No mundo dos modelos de aprendizagem supervisionada temos métodos tradicionais como árvores de decisão, Boosting, redes neurais, etc. e em cada um destes casos é necessário fornecer dados de treinamento e definir o que se espera que o modelo calcule. Então é uma pessoa quem define o que o modelo deve aprender a identificar.

No caso de modelos não supervisionados, os dados são fornecidos e é o algoritmo o responsável por encontrar padrões nos dados que lhe permitam gerar grupos diferentes entre si, mas com semelhança interna dos grupos.

Mas na aprendizagem por reforço, o agente aprende por tentativa e erro, através da interação com o seu ambiente. Aqui o ser humano não coleta esses dados, o próprio agente decide de forma autônoma e decide quais dados levar para autoaprendizagem, portanto isso não requer intervenção humana, e este tipo de algoritmo é o mais próximo do que pode ser observado em filmes de ficção científica de IA, onde o computador desenvolve a sua própria inteligência através da sua interação com o seu ambiente.

A aprendizagem por reforço apresenta relação com outras áreas do conhecimento como:

a. Teoria de controle: veículos autônomos, desde a captura de sinais do ambiente até a tomada de decisões sobre como dirigir.

b. Aprendizagem por reforço: ajudam a interagir com o meio ambiente (Sutton e Barto, 2018).

- Matemática: soluções de pesquisa: buscam interação por meio de tentativa e erro para tomar decisões ideais para um ambiente específico.
- Economia: agentes económicos dinâmicos, com e sem informação perfeita.
- Neurociências: relacionado às experiências do algoritmo e do sistema de recompensa.

2.3 Modelos de Aprendizagem por Reforço

Como conceito inicial, um sistema de aprendizagem por reforço tem a capacidade de aprender como tomar decisões ótimas para atingir um objetivo. A entrada é a interação de tentativa e erro com o sistema de interesse, na interação dentro de seu ambiente, estado, ações e resposta recebe o feedback necessário para autorregular seu comportamento para alcançar uma recompensa (Sutton e Barto, 2018).

2.3.1 Estrutura de um sistema de aprendizagem por reforço

Dentro do algoritmo do processo de aprendizagem por reforço temos uma estrutura para tomar decisões sequenciais - processo de decisão Markov (MDP) no qual podemos encontrar 5 componentes:

1. **Agente:** corresponde ao indivíduo que aprende e toma decisões.
2. **Ambiente:** conjunto de coisas com as quais o agente interage.
3. **Estado:** representação do ambiente observável pelo agente.
4. **Ação**
5. **Recompensa**

Figura 5 – Fluxo de interação entre o agente e o ambiente onde atua baseado em recompensas.

Fonte: Pröllochs, N. y Feuerriegel, S, Reinforcement Learning, Business Analytics Practice (2015).

Interação: ocorre como uma sequência de etapas em cada uma das quais o agente:

a. Observe o estado s_t do ambiente.
b. Execute uma ação a_t, usando a função de política.

c. Como consequência da ação, um passo depois o agente observa o novo estado s_t+1 e recebe uma recompensa (valor numérico) r_t+1
d. Ajustar a função política: adaptar o comportamento, aprender.

Inicialmente o agente não sabe nada sobre o ambiente, por isso tomará decisões aleatórias. Neste caso, se uma ação traz uma recompensa positiva, o agente deve aprender a escolher essa ação com mais frequência, enquanto uma ação que não proporciona uma recompensa, ou mesmo gera uma recompensa negativa, o agente deve aprender a evitar essa ação. Ou, na pior das hipóteses, não escolhê-lo com frequência. Nesse sentido, o agente, por meio de suas ações, atinge seu objetivo ao maximizar o valor total da recompensa que recebe, também conhecida como retorno. (Sutton e Barto, 2018).

Abaixo estão algumas áreas de aplicação da aprendizagem por reforço:

- Comunicações;
- Controle industrial;
- Logística e administração de processos;
- Robótica;
- Sistemas elétricos de potência em decisão e controle;
- Sistemas de recomendação em comércio eletrônico;
- Transporte.

2.3.2 Tipos de algoritmos de aprendizado por reforço

Os algoritmos de aprendizagem por reforço podem ser classificados em algoritmos baseados em valor (baseados em valor) e algoritmos baseados em políticas (baseados em políticas).

2.3.2.1 Algoritmos Baseados em Políticas

Este grupo inclui aqueles algoritmos que implementam uma política que condiciona diretamente a ação a ser tomada em cada momento. Ou seja, este tipo de algoritmo define a política que decide a probabilidade de tomar uma ação com base em cada estado.

Esta função não estima quanta recompensa o agente receberá de cada estado, pois não aprende com base no valor $V(s)$ nem em um valor de ação $Q(s,a)$, mas sim esses algoritmos são definidos através de um função de política (policy function) $\pi(a|s)$ que, conforme dito, estima a probabilidade de execução de cada ação com base em cada estado (Sutton e Barto, 2018).

Se avaliadas as vantagens presentes entre cada tipologia, verifica-se o seguinte:

Vantagens dos algoritmos baseados em valor:

i. Eles não precisam de política ou restrições para treinar.
ii. Como não há restrições de política dentro das funções, é mais eficiente em relação ao número de variáveis (eficiência amostral).
iii. Há menos variação.

Vantagens dos algoritmos baseados em valor:

i. Eles podem ser mais eficientes quando são apresentadas ações contínuas, portanto, para ambientes estocásticos, reconhecem melhor os padrões nos dados.
ii. Centra-se na otimização da função do nosso interesse, que neste caso é a política definida no início.

Como existe um objetivo mais claro, permite uma convergência mais rápida, ao contrário da função de valor.

2.3.2.2 Algoritmos baseados em valor (Value-based)

Aqui temos algoritmos que usam apenas um valor ou função de valor de ação sem implementar explicitamente uma política, eles se enquadram no grupo de algoritmos baseados em valor. Esses algoritmos não informam explicitamente qual ação o agente deve realizar, mas sim quanta recompensa receberá de cada estado ou estado de ação. Portanto, o agente deve decidir que ação tomar depois de ver esses valores. Um exemplo pode ser sempre realizar a ação com o valor Q mais alto. Outro exemplo poderia ser seguir uma política ε-ganancioso. Entre os algoritmos baseados em valor estão Q-Learning, DQN entre outros (Sutton e Barto, 2018).

Conforme mencionado, trata-se de uma política que busca orientar as ações do agente com base na maximização do valor da recompensa. Neste campo podemos encontrar um conceito muito interessante.

a) Política ε-Voraz (ε-greedy policy):

Esta será uma política que condicionará as ações que o agente irá realizar. Neste caso, a política ε -greedy consiste em o agente escolher quase sempre a melhor opção possível com base na informação que já recolheu. Porém, com probabilidade "ε" o agente realizará uma ação

aleatória. Esse conceito permite que o agente não fique preso apenas em ações que já tiveram resultados positivos ou recompensas positivas, pois isso não permitirá que o agente continue aprendendo sobre novas possibilidades que também proporcionem recompensas positivas.

Este valor "ε", é decidido pelo executor, e será a forma como o problema de "exploração" e "explotação" será equilibrado. A exploração consiste em o agente poder explorar e conhecer todas as ações possíveis quantas vezes forem necessárias para poder avaliar qual ou quais são as melhores, independentemente de serem obtidas ou não recompensas positivas durante essa exploração.

Já a explotação consiste em o agente maximizar suas recompensas, portanto o agente decidirá sobre as ações que proporcionam as maiores recompensas positivas. Nesse sentido, é importante equilibrar essas duas situações, pois se o agente apenas explorar, nunca obterá benefícios, e se apenas explotar, o agente não saberá que outras ações poderiam tê-lo ajudado a obter recompensas maiores (Sutton e Barto, 2018).

No problema de aprendizagem por reforço, o estado muda continuamente quando uma ação é executada. O agente recebe o estado do ambiente, que é representado pela letra *s* (estado), o agente executa a ação que escolher (inicialmente aleatória) representada por a (ação). Quando a ação for executada, o ambiente responderá fornecendo uma recompensa *r* (recompensa), e o ambiente transferirá o agente para um novo estado *s'* (próximo estado) (Sutton e Barto, 2018).

b) Função de valor:

Para calcular o nível de recompensa que será obtido a longo prazo em cada estado, deve ser introduzida uma função de valor V(*s*). Esta função produz uma estimativa da recompensa que o agente obterá até o final do jogo, partindo de um estado inicial (*s*). Se conseguirmos estimar o valor corretamente, o agente pode decidir executar a ação que maximiza o benefício da soma das ações (Sutton e Barto, 2018).

2.3.3 Algoritmos

a) Q-Learning

Este algoritmo tenta saber quanta recompensa obterá no longo prazo para cada par de estados e ações (*s*,*a*). Esta função é chamada de função valor-ação. Este algoritmo a representa com a função Q(*s*,*a*), que retorna a recompensa que o agente receberia ao executar a ação a do estado s, e assumindo que seguirá a mesma política. definido pela ação Q até o final do jogo.

Portanto, se a partir do estado inicial tivermos duas ações disponíveis, *a1* e *a2*, a função Q nos fornecerá os valores Q de cada uma das ações. Por exemplo, se Q(*s*,*a1*) = 1 e Q(*s*,*a2*) = 4, então o agente decidirá pela ação 2, que é a que trará maiores recompensas (Sutton e Barto, 2018).

b) Equação de Bellman

A explicação é baseada no fato de que o valor Q do estado s e da ação a Q(*s*,*a*) deve ser igual à recompensa r obtida pela execução daquela ação, mais o valor Q da execução da melhor ação possível *a'* do próximo estado *s'*, multiplicado por um fator de desconto γ, que é um valor com classificação $\gamma \in (0, 1)$. Este valor γ é usado para decidir quanto peso queremos dar ao curto prazo e recompensas no longo prazo, e é um hiperparâmetro que devemos decidir.

c) Deep Q-Network

Devido às limitações do volume de dados e variáveis, o que implica um desafio para o Q-Learning, o que Mnih (2015) propõe em equipe é o algoritmo Deep Q-Network. Este algoritmo combina o algoritmo Q-learning com redes neurais profundas. Já no campo da IA, as redes neurais são uma alternativa muito eficiente para aproximar funções não lineares. Portanto, este algoritmo utiliza uma rede neural para aproximar a função Q, evitando assim utilizar uma tabela para representá-la.

Na verdade, ele usa duas redes neurais para estabilizar o processo de aprendizagem. A primeira, a Rede Neural principal, representada pelos parâmetros θ, é utilizada para estimar os valores Q do estado atual *s* e da ação *a*: Q(*s,a;*θ). A segunda, a Rede Neural alvo, parametrizada por θ', terá a mesma arquitetura da rede principal, mas será utilizada para aproximar os valores Q do próximo estado *s´* e da próxima ação *a´* (Mnih et al., 2015).

A aprendizagem ocorre na rede principal e não na rede alvo. A rede alvo é congelada (seus parâmetros não são alterados) por diversas iterações (geralmente em torno de 10.000), e então os parâmetros da rede principal são copiados para a rede alvo, transmitindo assim o aprendizado de uma para a outra, fazendo com que as estimativas sejam calculadas pela rede alvo são mais precisos (Mnih et al., 2015).

d) Equação de Bellman em DQN

De acordo com Wang (2019), no algoritmo original de atualização de vantagem de Baird, a equação de atualização residual compartilhada de Bellman é dividida em duas

atualizações: uma para a função de valor de estado e outra para a função de vantagem associada. Demonstrou-se que a atualização de vantagem convergiu mais rapidamente do que o Q-learning em domínios de tempo contínuo simples (Harmon et al., 1995). Seu sucessor, o algoritmo de aprendizagem de borda, representa apenas uma função de borda (Harmon e Baird, 1996).

A arquitetura de duelo representa as funções de valor V(*s*) e vantagem A(*s*, *a*) com um único modelo profundo cuja saída combina os dois para produzir um valor de ação de estado Q(*s*, *a*). Ao contrário da atualização de vantagens, a representação e o algoritmo são dissociados por construção. Consequentemente, a arquitetura de duelo pode ser usada em combinação com uma ampla variedade de algoritmos aprendizagem por reforço livres de modelo (Wang et al., 2019).

Há uma longa história de funções de vantagem em gradientes de política, começando por (Sutton et al., 2000). Como exemplo recente desta linha de trabalho, Schulman et al. (2015) estimam valores de vantagem online para reduzir a variância de algoritmos de gradiente de política. Consideramos uma configuração de tomada de decisão sequencial, na qual um agente interage com um ambiente E em intervalos de tempo discretos. No exemplo do domínio Atari, por exemplo, o agente percebe um vídeo st composto por M quadros de imagem:

$$S_t = (x_t - M + 1, \ldots, x_t) \in;$$

S_t *: S* na etapa de tempo *t*.

O agente então escolhe uma ação de um conjunto discreto em ∈ A = {1, . . ., |A|} e observe um sinal de recompensa r_t produzido pelo emulador de jogo. O agente busca maximizar o retorno descontado esperado, onde definimos o retorno descontado como:

$$R_t = \sum_{T=t}^{\infty} (\boldsymbol{\gamma}^{T-t} r_t)$$

Nesta formulação, $\gamma \in [0, 1]$ é um fator de desconto que compensa a importância das recompensas imediatas e futuras. Para um agente que se comporta de acordo com uma política estocástica π, os valores do par estado-ação (*s*, *a*) e do estado *s* são definidos como segue:

$$Q^{\pi}(s,a) = E[\,[R_t] s_t = s, a_t = a, \pi\,], and$$

$$V^{\pi}(s) = E_{a\sim\pi(s)}[Q^{\pi}(s,a)]$$

A função de valor de ação de estado acima (função Q abreviada) pode ser calculada recursivamente com programação dinâmica (Janner et al., 2019).

$$Q^{\pi}(s,a) = E_{s\prime}\left[r + \ \gamma E_{a\sim\pi(s)}[Q^{\pi}(s,a)]|\ s,a,\pi\right]$$

Nós definimos o ideal:

$$Q^{*}(s,a) = max_{\pi}Q^{\pi}(s,a)$$

Sob a política determinista:

$$a = \arg.max_{a\prime}Q^{*}(s,a')$$

segue que:

$$V^{*}(s) = max_{a}Q^{*}(s,a)$$

Disto segue-se também que a função Q ótima satisfaz a equação de Bellman:

$$Q^{\pi}(s,a;\theta) = \gamma\, max_{a\prime}Q^{\cdot}(s',a';\theta')$$

Definimos outra grandeza importante, a função vantagem, relacionando o valor e as funções Q:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - \ V^{\pi}(s)$$

tenha em conta que: $E_{a\sim\pi(s)}[A^{\pi}(s,a)] = 0$

Intuitivamente, a função de valor V mede quão bom é estar num determinado local estado s. A função Q, entretanto, mede o valor da escolha de uma ação específica neste estado. A função vantagem subtrai o valor do estado da função Q para obter uma medida relativa da importância de cada ação (Wang et al., 2019).

Para gerar uma previsão a partir do conjunto, simplesmente selecionamos um modelo uniformemente ao acaso, permitindo que diferentes transições ao longo de uma única implementação de modelo sejam amostradas a partir de diferentes modelos dinâmicos.

3 METODOLOGIA

3.1 Proposta de desenvolvimento

Produto da atividade industrial, os dados operacionais provenientes do moinho e dos diversos equipamentos que interagem no processo de moagem de cimento, que são transferidos de uma base de dados industrial para a base de dados para o Data Lake operacional especificamente em tabelas bigquery.

Da mesma forma, os resultados dos testes de laboratório chegam ao Data Lake. Cada equipamento e/ou sensor que participará desta pesquisa serão as variáveis determinantes nos resultados da nossa variável a prever, ou seja, o Blaine (indicador químico a prever e alterar). Esses dados operacionais são registrados com frequência a cada 5 segundos.

Figura 6 – Fluxo de geração de dados entre ambiente industrial e banco de dados corporativo.

Fonte: Elaborador pelo autor (2024).

Com relação à variável “Blaine” (valor a ser manipulado), são registrados a cada 2 horas, por ser produto de um teste laboratorial (testes químicos através de uma amostra), não podem ser obtidos com a mesma frequência com que são obtidos são dados operacionais de registro.

Agora a integração destes dados implica também um desafio quanto ao esquema que definirá como tanto os dados do preditor quanto os dados a serem previstos estarão relacionados com uma ferramenta analítica capaz de antecipar e sugerir mudanças para manter o desempenho do resultado dentro de um desejável padrão.

Além disso, para ter uma ferramenta analítica que lhes permita antecipar o que vai acontecer, é de interesse por parte da equipe de operações e qualidade poder ter os dados no mesmo ambiente de onde a equipe de supervisão monitora esses dados operacionais para manter o controle sobre os equipamentos e/ou componentes envolvidos.

Portanto, antes da proposta de construção de um modelo analítico, pode considerar 3 blocos que devem ser considerados necessários antes, durante e depois da construção desta ferramenta analítica:

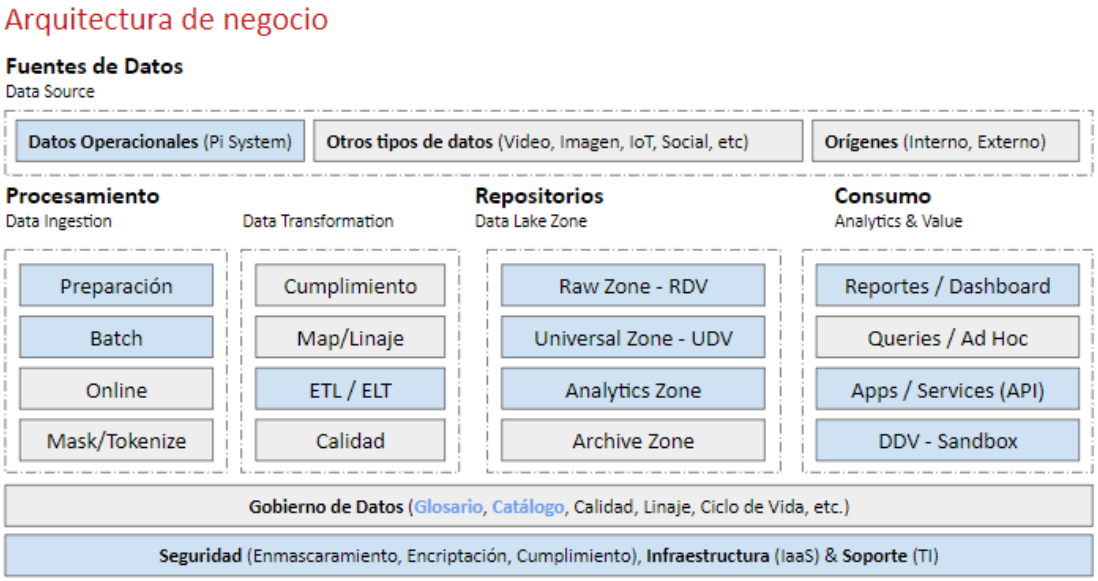
i. Conexão com Data Lake e ingestão de dados.
ii. Construção do modelo analítico.
iii. Pipeline de implantação de modelo analítico.

3.2 Coleta de Dados

3.2.1 Conexão com Data Lake e ingestão de dados

Os dados gerados pelos diferentes componentes e sensores são registrados em um ambiente industrial denominado Scada, deste ambiente são direcionados através de uma arquitetura em nuvem para o Data Lake corporativo. O lago corporativo possui 3 camadas de dados, que vão desde RDV, UDV, DDV.

Figura 7 – Fluxo de geração de dados entre ambiente industrial e banco de dados corporativo.



Fonte: Elaborador pelo autor (2024).

Uma vez que os dados estão na camada analítica, eles são armazenados em tabelas, dependendo da origem dos dados, neste caso teremos “dados operacionais” vindos dos sensores do equipamento, e teremos “dados químicos” que são os produtos de testes e ensaios laboratoriais. Considerando que ambos os tipos de dados devem manter uma ligação temporal, um modelo de dados é executado para integrar esses dois tipos de dados.

3.2.2 População e amostra

Os dados foram obtidos de uma empresa produtora de cimento, especificamente de um ambiente bigquery em um Data Lake integrado aos sistemas industriais da empresa, tanto operacionais quanto laboratoriais, que possuem acesso restrito. Esses dados são atualizados a cada 5 minutos e tiveram que passar por um pré-processamento para serem integrados.

i. População

   Considera-se um total de 2361 dados gerados pelos diferentes componentes que fazem parte da análise no período de julho de 2022 a março de 2023. Todos os dados referem-se a 109 sinais dos quais, como se verá, serão reduzidos no processo limpeza.

ii. Amostra

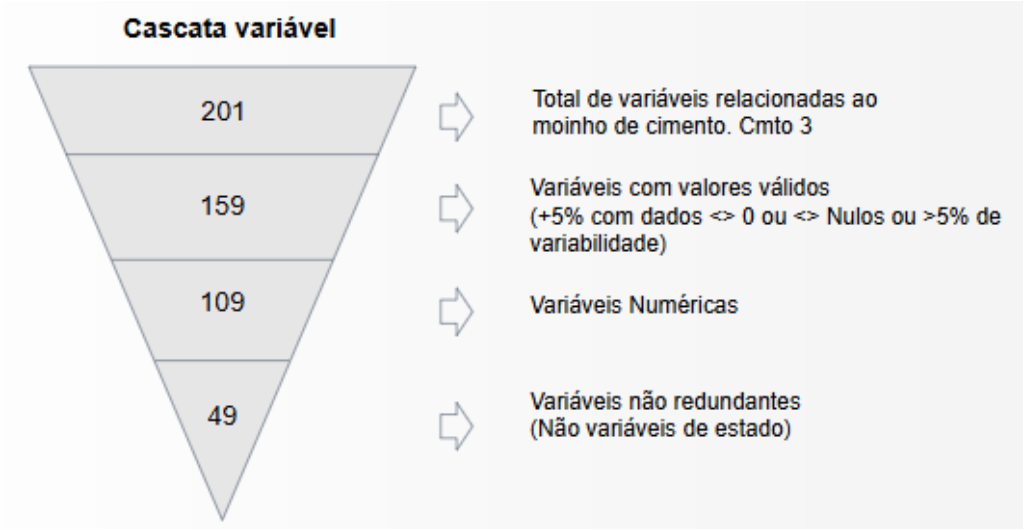
   Não foi estimada uma amostra porque trabalhamos com os dados totais da população-alvo.

3.3 Pré-processamento

3.3.1 Análise exploratória do banco de dados de sinais

O gráfico seguinte (Figura 8) apresenta o total de variáveis identificadas que correspondem aos dados fornecidos pelos diferentes equipamentos do ciclo de moagem de cimento. Porém, nem todos apresentam dados válidos, por isso a cascata mostra as limpezas iniciais que foram realizadas nessas variáveis também como parte do processo de exploração.

Figura 8 – cascata variável



Fonte: Elaborador pelo autor (2024).

Para revisar a relação direta ou indireta entre os componentes do moinho com o resultado Blaine, de um total de 109 variáveis operacionais inicialmente identificadas fizemos um top das mais relacionadas (Figura 9).

Figura 9 – Correlação superior das variáveis operacionais com o valor de Blaine.

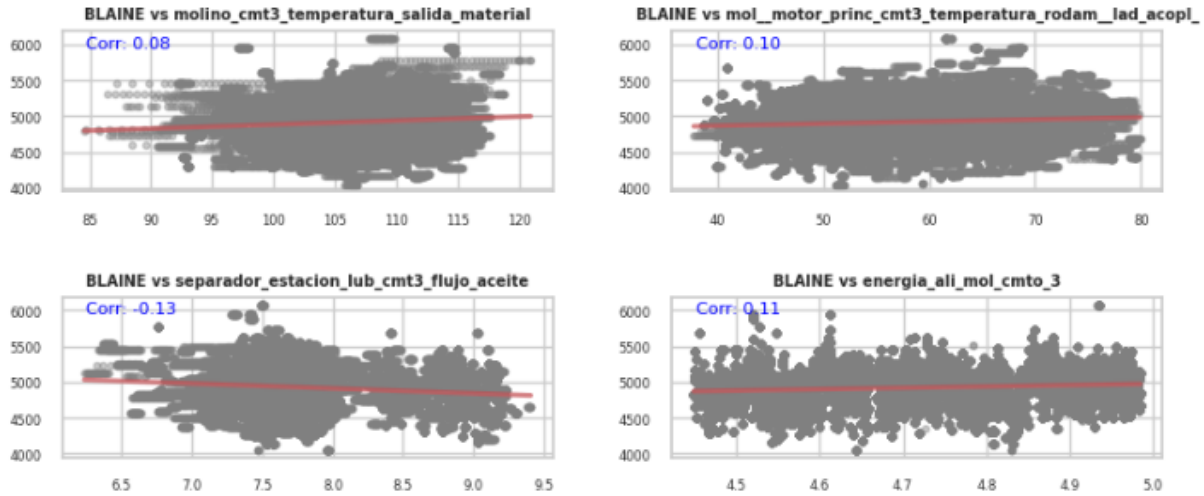
Variables	Correlación
separador_succion_cmt3_presion_gases	0.36
sistema_spray_cmt3_peso_total_grasa	0.18
mol__motor_princ_cmt3_temperatura_rodam__lad_acopl_	0.16
energia_aux_mol_cmto_3	0.15
energia_ali_mol_cmto_3	0.15
biza_clinker_d_cmt3_flujo_material	**0.15**
separador_succion_cmt3_temperatura_gases	0.14
mol__motor_princ_cmt3_temperatura_devanado_v2	0.12
mol__motor_princ_cmt3_temperatura_devanado_w1	0.12
mol__motor_princ_cmt3_temperatura_devanado_w2	0.12
mol__motor_princ_cmt3_temperatura_devanado_u1	0.12
mol__motor_princ_cmt3_temperatura_devanado_u2	0.12
mol__motor_princ_cmt3_temperatura_devanado_v1	0.12
mol__cmt3_corriente_motor_principal	**0.09**
potencia_activa_total_ali_mol_cmto_3	**0.09**
mol__pinon_lad_acopl_axial_cmt3_envolvente_vibracion	0.08
molino_cmt3_temperatura_salida_material	0.08
faja_alim__mol__cmt3_corriente_motor	0.08
corriente_promedio_ali_mol_cmto_3	0.08
mol__pinon_lad_acopl_axial_cmt3_aceleracion_vibracion	0.08

Variables	Correlación
canal__alim__seprador_cmt3_corriente_motor_soplador	-0.09
mol_CMT3_corriente_motor_bomb_1_2__tot_lub_motor_princ	-0.09
molino_estacion_lub_reductor_cmt3_temperatura	-0.09
estacion_lub_motor_princ_mol__cmt3_corriente_motor	-0.1
elev_silos_cmt3_temperatura_material	-0.11
mol__chumacera_entrada_cmt3_temperatura	-0.11
biza_clinker_a_cmt3_flujo_material	**-0.12**
mol__est__lub__chumacera_entrada_cmt3_temperatura_entrada	-0.13
biza_caliza_cmt3_flujo_material	**-0.14**
filtro_princ__cmt3_temperatura_salida	-0.14
elev_silos_cmt3_corriente_motor	-0.15
filtro_princ__cmt3_temperatura_entrada	-0.16
filtro_princ__cmt3_corriente_motor_ventilador	-0.17
separador_estacion_lub_cmt3_flujo_aceite	-0.19
elev_alim_separador_cmt3_corriente_motor	-0.27
separador_din__mico_cmt3_corriente_motor_	-0.28
filtro_princ__cmt3_frecuencia_motor_ventilador	-0.29
corriente_promedio_aux_mol_cmto_3	-0.32
potencia_activa_total_aux_mol_cmto_3	-0.34
separador_cmt3_presion_gases	-0.37

Fonte: Elaborador pelo autor (2024).

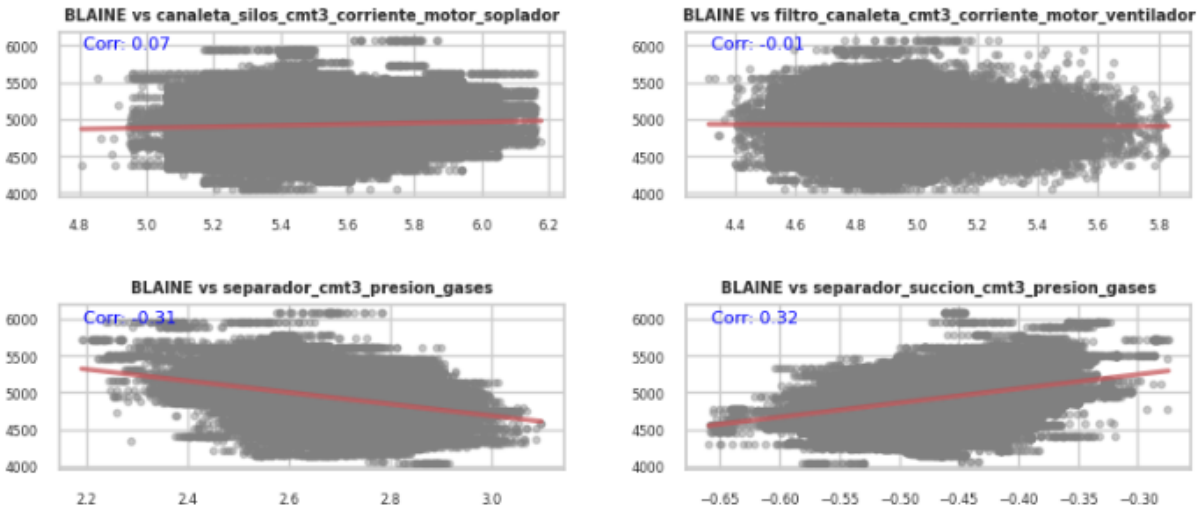
A nível gráfico, observa-se nos dois gráficos seguintes (Figura 10 e Figura 11) que o nível de correlação entre as variáveis em relação ao valor de Blaine não é muito elevado, embora possam ser diferenciadas relações positivas e negativas, também há casos de elevada concentração em alguns pontos que também não facilitam a identificação de um nível significativo de correlação, porém no processo de construção serão validadas essas relações, que não precisam necessariamente ter um padrão linear.

Figura 10 – Distribuição e correlação entre dados operacionais e Blaine – Parte 1.



Fonte: Elaborador pelo autor (2024).

Figura 11 – Distribuição e correlação entre dados operacionais e Blaine – Parte 2.

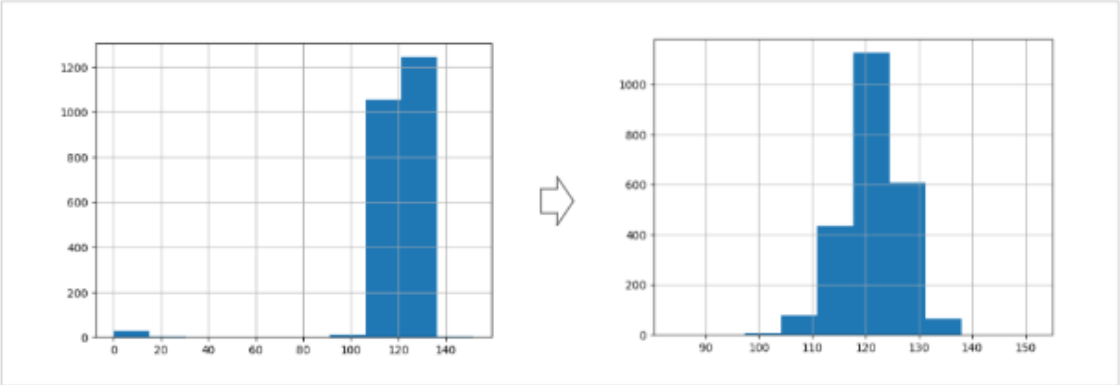


Fonte: Elaborador pelo autor (2024).

3.3.2 Detecção e remoção de outliers

No processo de exploração focado na variável Blaine podemos ver na Figura 12 que ela apresenta valores ilógicos mesmo esse dado vindo do laboratório, portanto tendo os valores de referência máximo e mínimo da equipe de operações, um menor o corte é feito para valores inferiores a 4000.

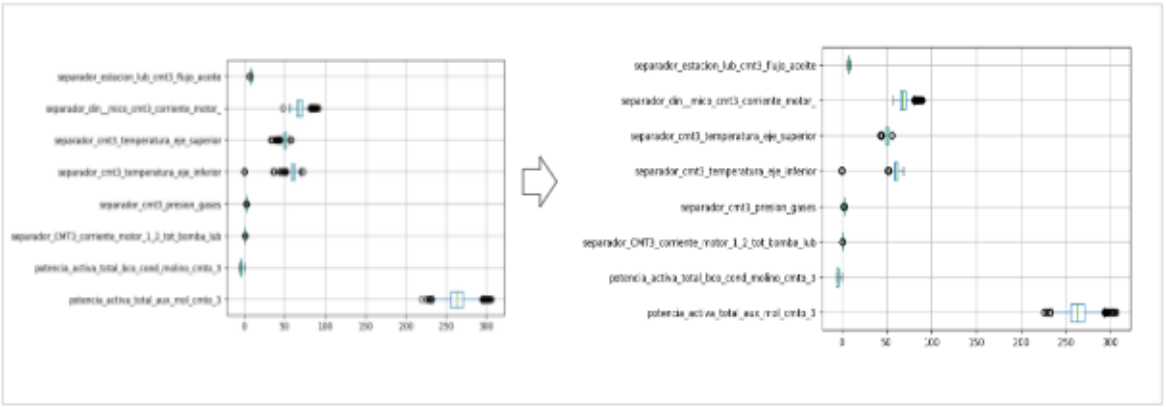
Figura 12 – Distribuição Blaine sem e com limpeza de outliers.



Fonte: Elaborador pelo autor (2024).

Da mesma forma, é realizada uma primeira visualização dos dados provenientes dos diferentes sensores ou componentes da moagem (variáveis preditoras e de controle), aos quais se aplica a mesma metodologia de limpeza, redução nas extremidades das caixas (Figura 13).

Figura 13 – Distribuição de dados operacionais com e sem outliers



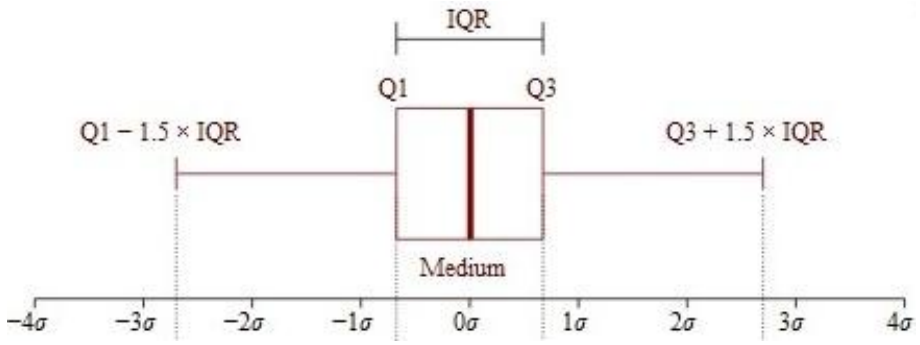
Fonte: Elaborador pelo autor (2024).

No ponto anterior foi mencionado que foi feita uma delimitação no valor de Blaine considerando valores acima de 4000 cm²/g, que se refere à superfície de partículas de cimento por grama.

Mas é necessário aplicar um método de limpeza de dados incongruentes nas restantes variáveis antes de passar a uma análise mais aprofundada, pelo que o método aplicado foi o corte com o intervalo interquartil.

- Q1: primeiro quartil, relativo aos primeiros 25% dos dados observados;
- Q2: segundo quartil (mediana), relativo aos primeiros 50% dos dados observados;
- Q3: terceiro quartil, relativo aos primeiros 75% dos dados observados;

Figura 14 – Distribuição de dados operacionais com e sem outliers



Fonte: Oracle EPM Cloud Planning Translated Books (2024).

Tomando como referência o método IQR (Figura 14), os dados extremos nos limites foram anulados ao considerar todas as variáveis exógenas juntamente com a variável endógena.

3.3.3 Eliminação de variáveis com baixa variabilidade

Foi realizado um processo de limpeza variável devido aos valores nulos, além deles foi realizado um processo de limpeza variável com baixa variabilidade em seus valores (<5% de variabilidade).

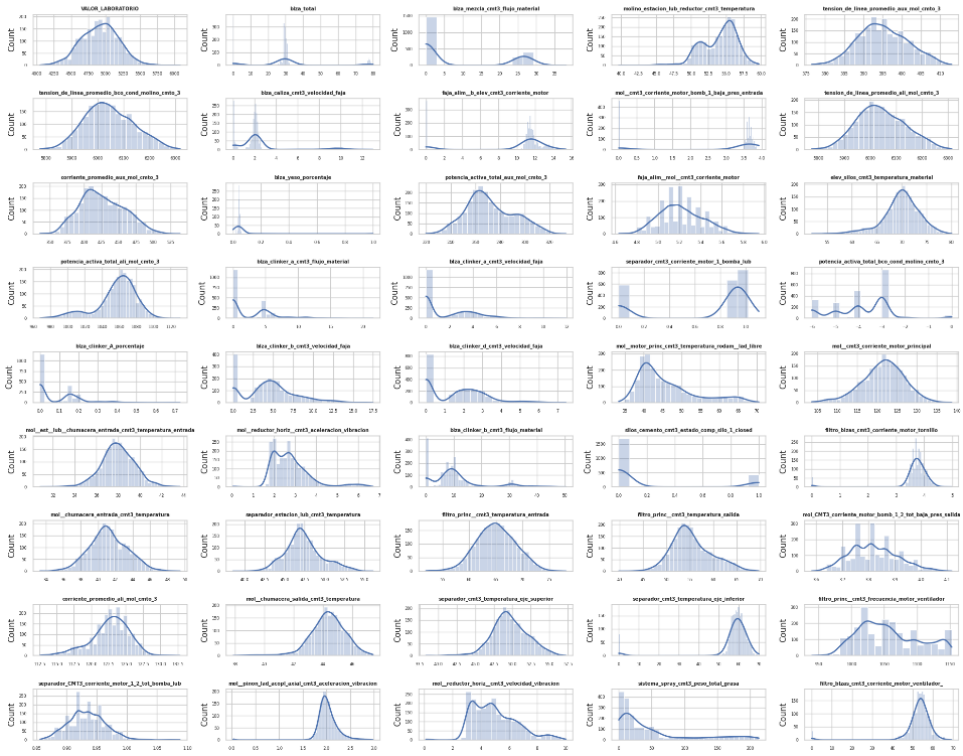
3.3.4 Detecção e eliminação de multilinearidade e autocorrelação com VIF

Foi realizado um processo de detecção de multicolinearidade e autocorrelação entre variáveis explicativas, utilizou-se a metodologia VIF (variance inflation factor). O limite de corte foi VIF > 5.

3.3.5 Análise exploratória do banco de dados de sinais componentes da área da fábrica de cimento

Após o processo de tratamento e limpeza dos dados, observa-se (Figura 15) que as variáveis também apresentam picos ou comportamentos multimodais, algumas apresentam distribuições com tendência à normalidade, no processo de modelagem do modelo que valida a importância destas e quais variáveis finalmente possuem maior explicabilidade no Blaine.

Figura 15 – Distribuição de variáveis explicativas.



Fonte: Elaborador pelo autor (2024).

3.4 Modelagem

3.4.1. Dividindo conjunto de dados em conjuntos de treinamento, teste e validação

Para desenvolver os modelos de regressão, o conjunto de dados foi separado em 3 partes, treinamento, teste e validação. Com esta separação, obtém-se um conjunto de dados para treinar o modelo (data train), um conjunto de dados para o modelo validar e otimizar seus parâmetros (data test) e um conjunto de dados de validação para validar que o modelo pode continuar estimando fora do tempo ou contexto dos dados com os quais foi treinado (Figura 16).

Figura 16 – Distribuição do conjunto de dados em treinamento, teste e validação.

Fonte: Elaborador pelo autor (2024).

Antes de treinar o modelo, o conjunto de dados foi separado em três, dados de treinamento, dados de teste, dados de validação. Abaixo está a porcentagem dos dados que foram obtidos em cada subconjunto e a utilização que tiveram no processo de construção e validação:

- Conjunto de treinamento (treinamento):
  - Aproximadamente 60% - 80% do conjunto de dados total.
  - Este conjunto é usado para ajustar o modelo e aprender padrões a partir dos dados.
- Conjunto de teste (teste):
  - Aproximadamente 10% - 20% do conjunto de dados.
  - Este conjunto é usado para avaliar o desempenho do modelo depois de treinado. É importante não usar este conjunto durante o treinamento para evitar overfitting.
- Conjunto de validação (validação):
  - Aproximadamente 10% - 20% do conjunto de dados.
  - É usado para selecionar os melhores hiperparâmetros do modelo (como regularização em modelos como Ridge ou Lasso) e para ajustar o modelo, sem tocar no conjunto de teste. Isso ajuda a evitar overfitting.

3.4.2 Construção do modelo analítico

Para executar o processo de construção do modelo analítico, foi utilizada a plataforma em nuvem Vertex IA, que possui notebooks Jupyter próprios e bibliotecas Python próprias. A partir desta plataforma é invocado o conjunto de dados preparado em Bigquery.

Com base nos objetivos deste TCC, foi procurado otimizar o processo de moagem de cimento que é medido com o Blaine indicado, dado que este indicador varia em valor, dependendo das alterações presentes em cada um dos componentes que interagem com o

moinho, portanto esta ferramenta analítica proposta deve ser capaz de prever cada combinação de movimentos nas variáveis operacionais (dentro de uma faixa de operabilidade do equipamento) de tal forma que mantenha os valores de Blaine dentro de uma faixa ótima, que seria a mesma neste caso como o modelo de aprendizagem por reforço entre as políticas que devem seguir.

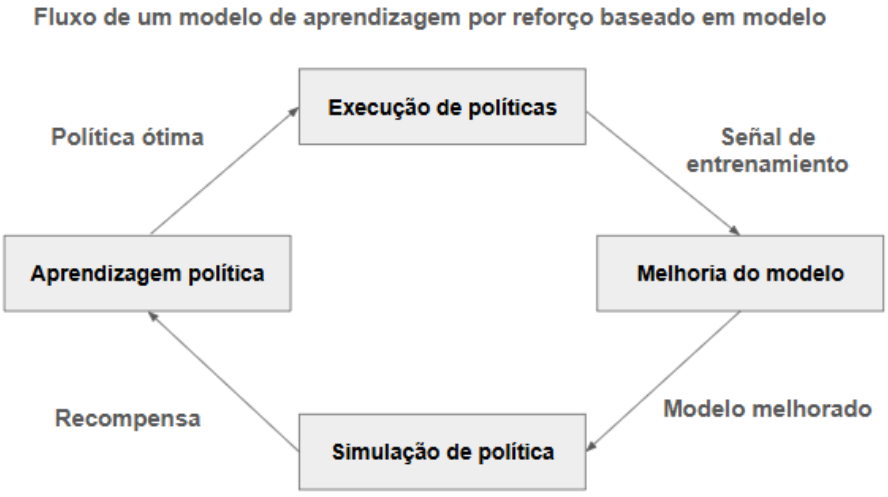
Tendo em conta os modelos de aprendizagem por reforço, as decisões podem ser tomadas a partir de duas abordagens.

- **Aprendizagem por reforço baseada em modelo.**
- **Aprendizagem por reforço sem modelos (Aprendizagem por reforço).**

Para tanto, a proposta centra-se na utilização de um algoritmo de Modelo Baseado em Aprendizagem por Reforço, ou seja, um modelo de aprendizagem por reforço que é suportado por um modelo preditivo que lhe permite ter simulações das possíveis ações que poderia realizar, analisando a melhor decisão antes de ser aplicada fisicamente aos componentes do moinho.

Isso evitaria que o modelo de aprendizagem por reforço fizesse movimentos aleatórios e cometesse erros físicos, portanto este algoritmo teria mais inteligência e seria capaz de antecipar os múltiplos erros que poderia cometer (Figura 17).

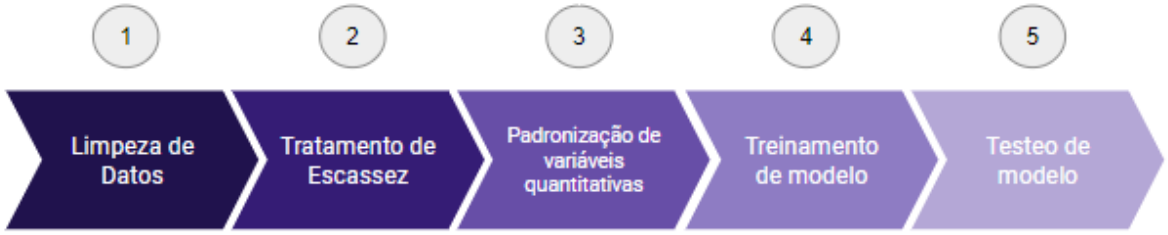
Figura 17 – Fluxo de um modelo de aprendizagem por reforço baseado em modelo.



Fonte: Elaborador pelo autor (2024).

O diagrama a seguir especifica o processo a ser seguido inicialmente como parte da metodologia de construção do modelo analítico (Figura 18).

Figura 18 – Metodologia de construção de modelo.



Fonte: Elaborador pelo autor (2024).

Embora o modelo desenvolvido seja um modelo de aprendizagem por reforço, é um modelo de aprendizagem por reforço baseado em modelos que irão guiar o agente (o agente é considerado o modelo treinado, mas que por sua vez pode aprender a tomar decisões com base nos dados de treino).

Neste caso, como o que se pretende é tornar mais eficiente o tempo de aprendizagem do agente e para que este possa antecipar possíveis erros que possa cometer no seu processo de aprendizagem, utiliza-se um modelo preditivo, neste caso um modelo de regressão para que o agente pode prever se as decisões que toma estão corretas com base nos seus objetivos de tal forma que cada decisão que toma já tem a capacidade de antecipar a resposta através deste modelo de regressão.

Portanto, a construção é separada em 2 momentos:

- *Primeiro:* construção de um modelo de regressão que preveja os valores de Blaine.
- *Segundo:* Utilizando um modelo de aprendizado por reforço que toma o modelo de regressão como entrada, treinar um agente para que ele aprenda quais valores das variáveis dependentes (5 variáveis) ele deve recomendar para que o valor de Blaine fique dentro de uma faixa ótima do processo de produção de cimento (entre 4700 a 4750) e isso é conseguido fazendo previsões e reajustando os valores das 5 variáveis dependentes que podem ser movidas (nem todas são ajustáveis).

3.4.2.1 Primeiro: Modelos de regressão

No processo de construção do modelo de regressão foram utilizados diferentes algoritmos e bibliotecas. Um processo de otimização relativa foi aplicado entre os modelos para reduzir o overfitting no momento do treinamento. Entre os modelos utilizados:

- Regressão Linear;
- Ridge (regressão de Ridge);
- Lasso (Operador de Contração e Seleção Mínima Absoluta);
- ElasticNet;
- Decision Tree (Árvore de Decisão);
- Random Forest;
- Gradient Boosting;
- XGBoost;
- LightGBM;
- KNeighborsRegressor (regressão dos vizinhos mais próximos);
- SVR (regressão de vetor de suporte);
- Rede Neural.

3.4.2.2 Segundo: aprendizado por reforço

O algoritmo deep deterministic policy gradient (DDPG) foi selecionado devido à sua capacidade de lidar com problemas de controle contínuo no contexto de aprendizagem por reforço (RL), como encontrar o valor ótimo de Blaine em um processo industrial. Vantagens do DDPG para atingir esse objetivo:

i. Problema de controle contínuo:

O objetivo principal é otimizar o valor de Blaine ajustando variáveis modificáveis dentro de uma faixa contínua. Este tipo de problema é o controle contínuo, ou seja, as ações que o agente realiza (modificações nas variáveis do processo) são contínuas e não valores discretos.

O DDPG foi projetado especificamente para lidar com esses tipos de ambientes, onde as ações a serem tomadas são sobre variáveis contínuas e não discretas (como "0" ou "1"). No caso do Blaine, é necessário ajustar continuamente algumas variáveis para atingir o valor ideal.

ii. Exploração eficiente em espaços de ação contínua:

O DDPG combina o melhor de dois tipos de abordagens em RL:

Ator-Crítico: O ator seleciona as ações (ajusta variáveis modificáveis), enquanto o crítico avalia o quão boa a ação foi com base na recompensa (quão próxima está do valor ideal de Blaine).

Gradientes de Política: Aprende políticas determinísticas (ações específicas em situações específicas) e pode ajustar variáveis de forma eficiente. Isso permite que o agente aprenda ações ideais em um ambiente contínuo, testando diferentes configurações das variáveis e melhorando com o tempo.

iii. Determinismo nas ações:

No DDPG, o agente aprende uma política determinística, o que significa que para cada estado do processo (cada combinação de variáveis), o agente seleciona exatamente uma ação ótima. Isto é crucial para otimizar o valor de Blaine, uma vez que é necessária uma solução precisa e reprodutível: o DDPG garante que, uma vez aprendido, o modelo dará a mesma resposta (ajuste variável) para a mesma entrada. Isto é importante em processos industriais onde os ajustes devem ser consistentes e confiáveis.

iv. Algoritmo fora da política:

DDPG é um algoritmo fora da política, o que significa que pode usar experiências passadas (armazenadas em um buffer de repetição) para aprender com elas. Isso torna o treinamento mais eficiente, já que não é necessário gerar novas interações com o ambiente o tempo todo. Isso reutiliza experiências passadas para melhorar seu desempenho.

Isto é útil no caso de encontrar o valor ideal de Blaine, pois o modelo pode aprender com uma quantidade limitada de dados ou simulações sem ter que gerar continuamente novas amostras, o que pode ser caro em termos de tempo e recursos.

v. Otimização em espaços de ação de alta dimensão:

O DDPG é eficaz ao trabalhar com problemas onde o espaço de ação (as variáveis que o modelo pode modificar) é grande ou multidimensional. Neste caso, o ajuste do valor de Blaine provavelmente envolve a modificação de algumas variáveis independentes (como temperatura, fluxo de ar, velocidade do moinho, etc.).

vi. Uso de Redes Neurais:

O DDPG usa redes neurais profundas para aprender a prever ações (ator) e recompensas esperadas (crítico). As redes neurais permitem que o modelo aprenda relações complexas e não lineares entre as variáveis do processo e o resultado (valor de Blaine).

Isto é crucial para problemas industriais complexos onde a relação entre as variáveis modificáveis e o objetivo final não é trivial nem linear. A capacidade de aprender representações não lineares permite que o modelo se ajuste melhor e, com o tempo, descubra configurações ideais para o processo.

vii. Exploração com Ruído:

Exploração contínua: para evitar que o agente fique preso em soluções abaixo do ideal, o ruído é introduzido nas ações via Normal action noise. Este ruído permite ao agente explorar diferentes configurações das variáveis para melhorar continuamente a previsão e ajuste do Blaine. Este processo de exploração e aproveitamento é crucial para melhorar o modelo no longo prazo.

3.4.3 Escolha do Melhor Modelo

Após os modelos de regressão terem sido treinados para avaliar seu desempenho, seus resultados foram avaliados em dados de treinamento, teste e validação (dados fora do tempo), com o objetivo de ver seu desempenho e como ele se mantinha trabalhando com diferentes conjuntos de dados. As métricas analisadas foram:

- Raiz do Erro Quadrático Médio (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum (y_{real} - y_{predicho})^2}$$

- R^2 (Coeficiente de Determinação):

$$R^2 = 1 - \frac{\sum(y_{real} - y_{predicho})^2}{\sum(y_{real} - \gamma_{real})^2}$$

Com base nessas métricas, foi escolhido o melhor modelo de regressão que apresentasse explicabilidade para o negócio ou operação. Isso também foi utilizado para o modelo de aprendizagem por reforço, pois também busca reduzir o erro estimado pelo agente.

3.4.4 Modelo de Referência

Deep deterministic policy gradient (DDPG) é um algoritmo de aprendizado por reforço projetado para lidar com espaços de ação contínua, tornando-o ideal para problemas como ajuste de variáveis em um processo industrial para otimizar o valor de Blaine. Ao contrário de algoritmos mais simples como Q-Learning e SARSA, que funcionam apenas com ações discretas, o DDPG permite o controle preciso de variáveis contínuas. Isto é crucial para otimizar processos em que as decisões não são binárias, mas requerem ajustes graduais (Wang et al., 2016).

Modelos como DDQN e REINFORCE também estão limitados a ambientes discretos ou apresentam problemas de ineficiência ao exigirem muitas interações com o ambiente, o que não é ideal para ambientes industriais onde os recursos são limitados. Por outro lado, o PPO e o A3C conseguem lidar com ações contínuas, mas não alcançam a mesma precisão e eficiência que o DDPG, principalmente no ajuste fino de variáveis.

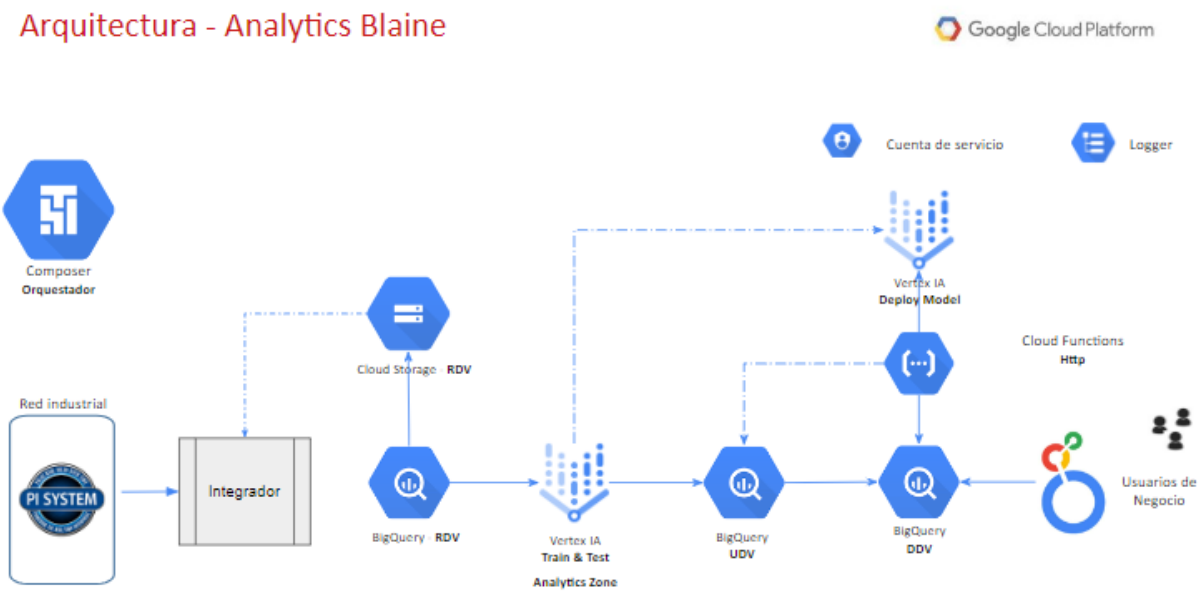
O DDPG utiliza uma arquitetura Ator-Crítico, que permite aprender uma política determinística, ou seja, uma ação específica para cada estado. Isto é essencial para processos em que são necessárias decisões consistentes e reproduzíveis. Além disso, por estar fora da política, o DDPG reutiliza experiências passadas de forma eficiente, reduzindo a necessidade de geração constante de dados. O DDPG é mais adequado para este tipo de problemas de controle contínuo em comparação com outros algoritmos RL, que são mais limitados em termos de ações contínuas e eficiência.

3.5 Aplicação prática do modelo preditivo

3.5.1 Implementação

A plataforma GCP possui um pipeline de implantação de modelo denominado Kuberflow, por meio do qual é possível monitorar o modelo de simulação dos valores de Blaine com o qual o modelo de aprendizagem por reforço tomará decisões sobre os valores que devem ser alterados. Como ainda temos os valores reais dos resultados de Blaine dos testes de laboratório, o modelo poderá continuar validando os resultados obtidos e fornecer feedback e manter este modelo calibrado.

Figura 19 – Proposta de arquitetura de extração de dados, construção de modelo e implementação.



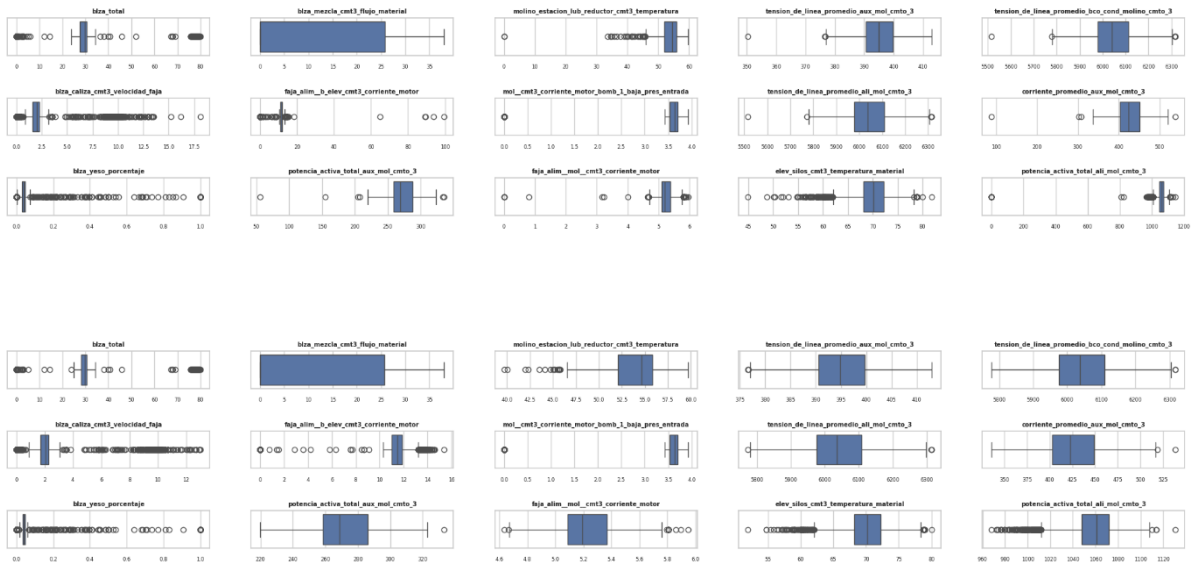
Fonte: Elaborador pelo autor (2024).

4 RESULTADOS

4.1 Coleta e Pré-processamento

Após a limpeza dos dados outliers, as variáveis que permaneceram como variáveis preditoras do melhor modelo apresentaram redução na sua dispersão.

Figura 20 – Caixas de Distribuição Variáveis.



Fonte: Elaborador pelo autor (2024).

4.2 Análise Exploratória e Descritiva

Da mesma forma, com as variáveis preditoras, é feita uma observação do seu nível de relacionamento (direto ou indireto) com os valores de Blaine, bem como do seu nível de correlação.

4.3 Modelagem: modelo de regressão

Foram construídos 12 modelos com técnicas diferentes conforme mostrado na Tabela 1, dos quais os primeiros quatro modelos pertencem a modelos de regressão linear, os próximos dois a árvores, os próximos três a boosting, o próximo a redes neurais e uma variante de máquina de vetores de suporte (SVM).

Tabela 1 - Desempenho dos modelos treinados de Regressão.

Modelos	R^2 Train	R^2 Test	RMSE Train	RMSE Test	CV RMSE
Linear Regression	0,3036	0,2508	225,72	241,20	235,39
Ridge	0,3036	0,2511	225,73	241,16	235,02
Lasso	0,3030	0,2511	225,81	241,17	234,58
ElasticNet	0,2943	0,2488	227,23	241,53	233,41
Decision Tree	0,3375	0,0827	220,17	266,89	261,81
Random Forest (Optimizado)	0,6918	0,2202	150,16	246,09	237,40
Gradient Boosting (Optimizado)	0,4632	0,2144	198,18	247,00	236,53
XGBoost	0,5716	0,2264	177,04	245,11	235,86
LightGBM	0,5287	0,2168	185,70	246,61	235,75
KNeighborsRegressor	0,4064	0,1466	208,40	257,44	257,83
SVR	0,0163	0,0147	268,27	276,61	268,88
Neural Network	-0,2038	-1,1211	296,78	405,85	428,68

Fonte: Elaborador pelo autor (2024).

4.3.1 Escolha do melhor modelo

Como se pode verificar nos resultados, os modelos não paramétricos apresentaram um valor superior no R^2 no treino, no entanto apresentam uma queda bastante forte no teste R^2. O mesmo acontece quando se observa o RMSE, que ao nível do treino o valor é entre os mais baixos, mas quando observamos os resultados dos testes, ele sobe bem acima dos valores de treinamento. No processo de treinamento do modelo, foram aplicadas otimizações ao nível de seus hiperparâmetros para estabilizar seus resultados.

Ao observar as métricas resultantes dos modelos paramétricos (Regressão Linear, Lasso, Ridge, ElasticNet), mantêm resultados mais estáveis ao nível de R^2 e RMSE, ambos com dados de treino, teste e validação.

Os resultados entre esses modelos de regressão paramétrica são muito semelhantes, portanto, não geram muita diferença. Razão pela qual o modelo de regressão linear foi selecionado adicionalmente por sua vantagem de explicabilidade.

Figura 21 – Resumo do modelo de regressão linear

```
Resumen del modelo OLS final:
                            OLS Regression Results
==============================================================================
Dep. Variable:      VALOR_LABORATORIO   R-squared:                       0.285
Model:                            OLS   Adj. R-squared:                  0.275
Method:                 Least Squares   F-statistic:                     28.08
Date:                Mon, 30 Sep 2024   Prob (F-statistic):          6.53e-84
Time:                        13:51:56   Log-Likelihood:                -9339.5
No. Observations:                1360   AIC:                         1.872e+04
Df Residuals:                    1340   BIC:                         1.882e+04
Df Model:                          19
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
const                                          4907.4316      7.886    622.308      0.000    4891.962    4922.902
corriente_promedio_aux_mol_cmto_3               -31.7717      9.838     -3.230      0.001     -51.071     -12.473
faja_alim__a_cmt3_corriente_motor                55.8548     15.004      3.723      0.000      26.421      85.289
potencia_activa_total_ali_mol_cmto_3             32.3219      7.929      4.076      0.000      16.767      47.877
blza_clinker_a_cmt3_velocidad_faja              -22.9244      6.918     -3.314      0.001     -36.497      -9.352
blza_yeso_cmt3_velocidad_faja                   -23.5243     10.769     -2.184      0.029     -44.651      -2.398
separador_estacion_lub_cmt3_temperatura          25.9376      8.123      3.193      0.001      10.002      41.873
mol__cmt3_corriente_motor_bomb_1_baja_pres_salida    18.0476      9.010      2.003      0.045       0.373      35.722
mol__pinon_lad_acopl_axial_cmt3_velocidad_vibracion    15.8200      7.502      2.109      0.035       1.104      30.536
separador_cmt3_presion_gases                    -78.0061      7.852     -9.935      0.000     -93.410     -62.603
filtro_princ__cmt3_temperatura_entrada          -42.8223      8.289     -5.166      0.000     -59.084     -26.561
mol__estacion_lub_motor_cmt3_temperatura        -30.5005      9.956     -3.063      0.002     -50.032     -10.969
separador_cmt3_corriente_motor_2_bomba_lub      -19.3550      7.510     -2.577      0.010     -34.088      -4.622
blza_clinker_D_porcentaje                        35.5939      8.290      4.294      0.000      19.332      51.856
elev_alim_separador_cmt3_corriente_motor        -27.4416      7.878     -3.483      0.001     -42.897     -11.987
filtro_blzas_cmt3_corriente_motor_tornillo       22.2219      7.616      2.918      0.004       7.281      37.163
estacion_lub_motor_princ_mol__cmt3_corriente_motor    15.6493      7.619      2.054      0.040       0.702      30.597
separador_estacion_lub_cmt3_flujo_aceite        -18.2021      6.684     -2.723      0.007     -31.315      -5.089
mol__cmt3_corriente_motor_principal              23.9241      7.328      3.265      0.001       9.549      38.299
blza_caliza_porcentaje                           31.8262     12.280      2.592      0.010       7.735      55.917
==============================================================================
Omnibus:                       10.080   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               10.223
Skew:                          -0.204   Prob(JB):                      0.00603
Kurtosis:                       2.879   Cond. No.                         4.65
==============================================================================
```

Fonte: Elaborador pelo autor (2024).

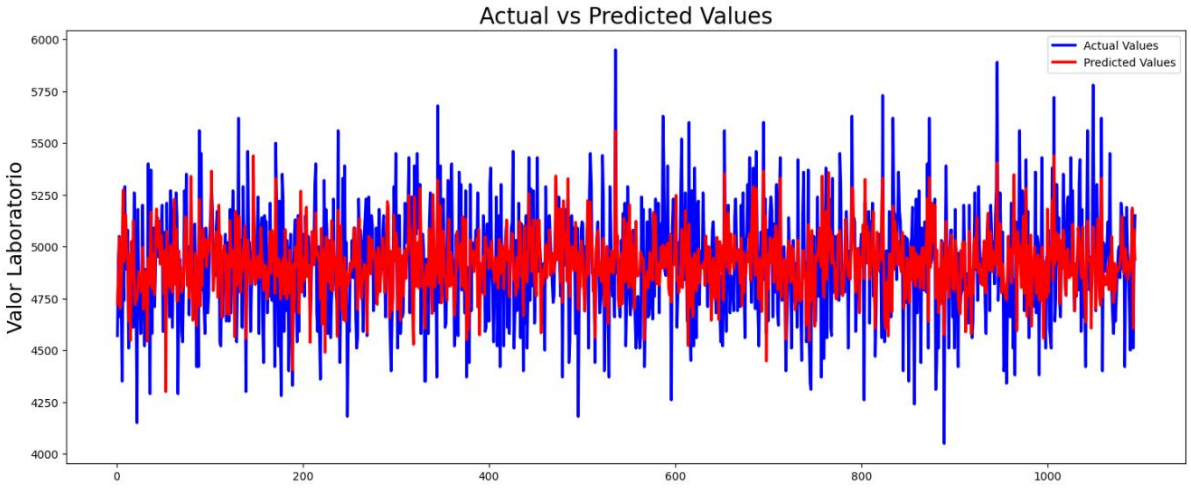
Os resultados apresentados na figura 21 refletem diversas métricas comuns na análise de um modelo de regressão linear:

- Teste Omnibus (Prob(Omnibus): 0,002): Indica que os resíduos não seguem uma distribuição normal, o que sugere possíveis problemas no ajuste do modelo.
- Durbin-Watson (2,041): Está próximo de 2, sugerindo que não há autocorrelação significativa nos resíduos, um bom indicador de independência.
- Teste Jarque-Bera (Prob(JB): 0,00266): Reafirma que os resíduos não são normais, o que pode impactar na validade das inferências.
- Skew (-0,200): Mostra uma ligeira assimetria negativa nos resíduos, mas não é preocupante.
- Curtose (2,779): Próximo de 3, o que indica que os resíduos possuem formato semelhante a uma distribuição normal.

- Cond. No. (2.43): Sugere que não há problemas de multicolinearidade.
- Também como pode ser observado, as variáveis selecionadas mantêm um nível de significância de 0,05%, portanto há evidência estatística suficiente para rejeitar a hipótese nula.

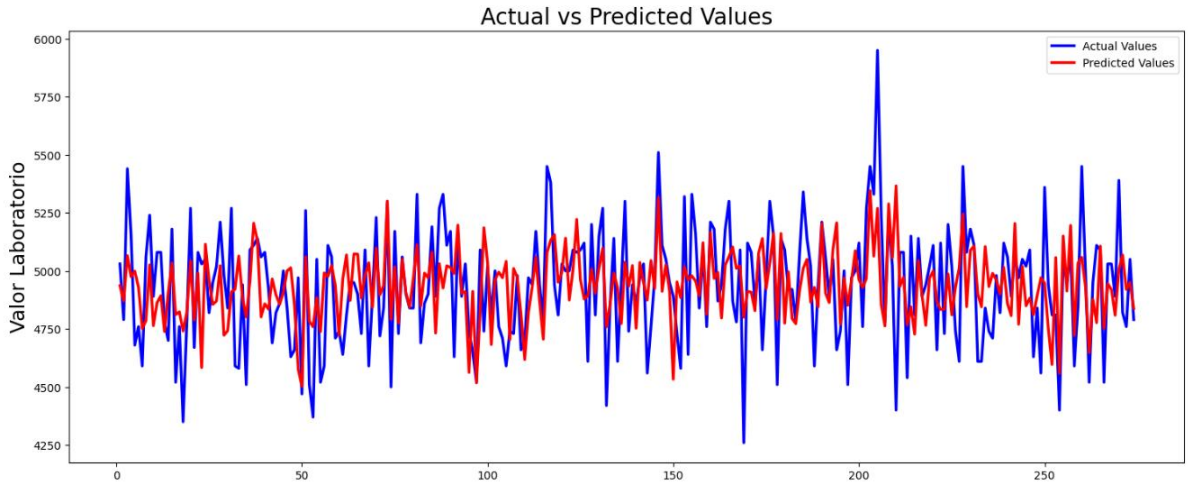
Embora o modelo possa melhorar, contaremos com os resultados do RMSE e sua estabilidade ao fazer previsões.

Figura 22– Resultado das predições para o conjunto de treinamento.



Fonte: Elaborador pelo autor (2024).

Figura 23– Resultado das predições para o conjunto de teste.



Fonte: Elaborador pelo autor (2024).

O erro médio do modelo de regressão é de 4,4%. Mas no nível de visualização, são observados picos que o modelo não consegue prever (Figura 23).

4.4 Modelagem: modelo de aprendizagem por reforço

O processo de treinamento do modelo de aprendizagem por reforço (DDPG) também incluiu um conjunto de dados de treinamento e teste, os mesmos dados que foram utilizados para o processo de treinamento do modelo de regressão.

Nesse sentido, o modelo de aprendizagem por reforço (DDPG) passou por um treinamento onde, com o apoio do modelo de regressão, aprendeu quais valores atribuir às 5 variáveis dependentes que poderiam ser modificadas de forma a permitir o previsto valor para se aproximar do valor médio da faixa ideal em que o valor de Blaine deve ser mantido. Nesse sentido, os resultados dos testes no processo de treinamento, teste e validação são os seguintes

Tabela 2 – Desempenho dos modelos treinados Reinforcement Learning.

Data	RMSE Global	R^2 Global	RMSE Global	Dif Blaine Real vs Predicho	Target Mean	% Error RL	% Error Real
Train	174,25	0,00	246,34	90,97	4725	3,69%	5,21%
Test	241,89	0,00	228,31	8,40	4725	5,12%	4,83%
Validación	208,79	0,00	283,82	66,74	4725	4,42%	6,01%

Fonte: Elaborador pelo autor (2024).

- O campo RMSE global mostra o RMSE entre o valor previsto e o valor médio ideal.
- O campo RMSE previsto mostra o RMSE entre o valor previsto e o valor médio ideal do Blaine.
- R^2 Global, mostra o R^2 entre o valor previsto e o valor médio ótimo, pois neste caso o valor médio ótimo era constante, o valor de R^2 era zero.
- O campo target média mostra a média do valor de Blaine ideal que se espera que seja alcançado.
- O campo % erro mostra a % de erro médio entre o valor estimado pelo modelo de aprendizagem por reforço e o valor médio ideal de Blaine.

Já que o objetivo do modelo de aprendizagem por reforço é treinar um agente que, utilizando o modelo de regressão como ferramenta, teste diferentes combinações ou valores para cada uma das cinco variáveis, estimando valores do Blaine até obter valores próximos para um valor médio de 4725.

Tabela 3 – Demonstração de aplicação do agente treinado com aprendizagem por reforço.

Variável Modificável	Valor Atual	Valor sugerido	Blaine Real	Blaine previsto	RMSE to Target
corriente_promedio_aux_mol_cmto_3	1,430	2,430	4640	4723	1,53
faja_alim__a_cmt3_corriente_motor	0,629	1,629	4640	4723	1,53
separador_estacion_lub_cmt3_temperatura	0,944	-0,056	4640	4723	1,53
separador_cmt3_corriente_motor_2_bomba_lub	0,117	-0,883	4640	4723	1,53
mol__cmt3_corriente_motor_principal	-0,127	-0,169	4640	4723	1,53

Fonte: Elaborador pelo autor (2024).

No exemplo apresentado na Tabela 3, observa-se no campo “Valor Atual” que cada um dos sinais das cinco variáveis possui um valor inicial. Após o processo de estimação pelo agente, no campo “Valor sugerido”, ele estima os valores que cada uma das cinco variáveis deve apresentar para que o valor do Blaine fique dentro de uma faixa de otimalidade cujo resultado é refletido no campo “Blaine previu” e no campo “RMSE to Target” observa-se que o modelo de aprendizagem por reforço consegue se aproximar de um valor dentro da faixa de otimalidade.

5 **CONCLUSÃO**

Ao longo do desenvolvimento desta análise foram observadas uma série de variáveis operacionais relacionadas ao processo de produção de cimento da empresa, e no processo de compreensão foi identificada a presença de variáveis que também influenciam o processo de moagem que não são captadas por sensores como medir umidade do material, temperatura ambiente, exposição, entre outros.

Também não são incluídas variáveis referentes à composição química de matérias-primas e materiais intermediários como gesso, calcário, clínquer etc. Dada a sua elevada contribuição de informação para a variabilidade do Blaine, sugere-se para futuros treinamentos avaliar sua viabilidade e inclusão.

Estas variáveis não foram consideradas inicialmente porque não foram integradas ao processo. Sugere-se também incorporar mais registros históricos ao modelo à medida que são gerados, a fim de alimentá-lo e identificar novos casos.

Apesar de possuir apenas informações sobre variáveis operacionais, foi alcançado um modelo de regressão que mantém uma margem de erro de 4,4% na previsão do valor de Blaine (variável prevista). Embora haja dificuldades na previsão de valores de pico, o modelo mantém estabilidade ao fazer previsões com teste e dados de validação.

Como segundo ponto, o modelo de aprendizagem por reforço mostrou a capacidade deste agente em se ajustar com base no modelo de regressão a um valor da metade ótima do Blaine, conseguindo mesmo uma margem de erro de 3,4% ao valor real quando usando validação de dados ou fora do tempo.

Concluindo que a combinação dos dois modelos resultou na manutenção de uma baixa margem de erro na tentativa de estimar o valor de Blaine e que este agente pode auxiliar na tomada de decisão das equipes de operações e qualidade da planta.

Visto que seguindo suas recomendações reduziria esta margem de erro de uma média de 5,20% para 6% para 3,6% para 4,4% com a oportunidade de reduzir o erro com recomendações sobre a incorporação de mais informações na construção dos modelos.

Com relação aos resultados observados com o modelo de aprendizagem por reforço na tabela 3, pode-se inferir o potencial desta ferramenta não só para fornecer suporte e recomendações às equipes de operações e qualidade, mas também permitir o controle de determinadas variáveis que implica controle.

O custo de produção, bem como evitar desperdícios e perdas de materiais de produção, uma vez que é produzido dentro de uma faixa de aceitabilidade do produto, o que implica economia para a empresa e evita sanções a nível regulatório por parte de entidades estatais.

REFERÊNCIAS

CONDOR, E; PAUTA, F. **Simulación de la dosificación, molienda y separación de cemento.** 2001. Disponível em chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://bibdigital.epn.edu.ec/bitstream/15000/11500/1/T1851.pdf. Acesso em: 19 Jan. 2024.

HARMON, M; HARMON, S. **Reinforcement Learning: A Tutorial** - Wright State University. 1996. Disponível em: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cs.toronto.edu/~zemel/documents/411/rltutorial.pdf. Acesso em: 10 Abr. 2024.

HARMON, M; LEEMON, C. **Advantage Updating Applied to a Differential Game**. 1996. Disponível em: https://www.researchgate.net/publication/2790049_Residual_Advantage_Learning_Applied_to_a_Differential_Game. Acesso em: 21 Jul. 2024.

JANNER, M; FU, J; ZHANG, M; LEVINE, S. **When to Trust Your Model: Model-Based Policy Optimization**. 2019. Disponivel em: https://arxiv.org/pdf/1906.08253. Acesso em: 20 Sep. 2024.

MNIH, V; KAVUKCUOGLU, K; SILVER, D; RUSU, A; VENESS, J; BELLEMARE, M; GRAVES, A; RIEDMILLER, M; FIDJELAND, A; OSTROVSKI, G; PETERSEN, S; BEATTIE, CH; SADIK, A; ANTONOGLOU, I; KING, H; KUMARAN, D; WIERSTRA, D; LEGG, SH; HASSABIS, D. **Human-level control through deep reinforcement learning** - Macmillan Publishers Limited. 2015. Disponivel em: https://web.stanford.edu/class/psych209/Readings/MnihEtAlHassibis15NatureControlDeepRL.pdf. Acesso em: 15 Jun. 2024.

MOHAMADA, N; MUTHUSAMYA, K; EMBONG, R; KUSBIANTORO, A: HASHIM, M. **Environmental impact of cement production and Solutions**: - A review Faculty of Civil Engineering Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak. 2018. Disponível em: https://www.researchgate.net/publication/349877298_Environmental_impact_of_cement_production_and_Solutions_A_review. Acesso em: 15 Maio. 2024.

OJEWUMI, E; BAMIGBOYE, G; OLUKANNI, D. **Experimental and modelling of flexural strength produced from granite-gravel combination in selfcompacting concrete**. pp. 437–447. Set 2018.

PRÖLLOCHS, N; FEUERRIEGEL, S**. Reinforcement Learning, Business Analytics Practice**, 2015. Disponible en http://www.is.uni-freiburg.de/ressourcen/ business-analytics/13_reinforcementlearning.pdf. Acesso em: 30 Sep. 2024.

SOLTANNZADEH, F; EMAM-JOMEH, M; EDALAT, A. **Development and characterization of blended cements containing seashell poder**. 2018. Disponivel em: https://www.researchgate.net/publication/349877298_Environmental_impact_of_cement_production_and_Solutions_A_review. Acesso em: 10 Jan. 2024.

SUTTON, R. S; BARTO, A. **Reinforcement learning: An introduction. MIT press. - The MIT Press Cambridge**, Massachusetts London, England. 2018. Disponível em: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf. Acesso em: 20 Fev. 2024.

SCHULMAN, J; M; LEVINE, S; MORITZ, P; JORDAN, M; ABBEEL, P. **Trust Region Policy Optimization** - University of California, Berkeley, Department of Electrical Engineering and Computer Sciences**.** 1996. Disponível em: https://proceedings.mlr.press/v37/schulman15.pdf. Acesso em: 03 Jun. 2024.

WANG, T; BAO, X; CLAVERA, I; HOANG, J; WEN, Y; LANGLOIS, E; ZHANG, S; ZHANG, G; ABBEEL, P; BA, J. **Benchmarking Model-Based Reinforcement Learning**. 2019. Disponivel em: https://arxiv.org/pdf/1907.02057. Acesso em: 21 Ago. 2024.

WANG, Z; SCHAUL, T; HESSEL, M; VAN HASSELT, H; LANCTOT, M; FREITAS, N. **Dueling Network Architectures for Deep Reinforcement Learning** - Google DeepMind, London, UK. 2016. Disponível em: https://arxiv.org/pdf/1511.06581.pdf. Acesso em: 05 Abr. 2024.