

## ESTIMATIVA DE VELOCIDADE EM COLISÕES FRONTAIS POR APRENDIZADO DE MÁQUINA

Trabalho de Perícia ou Inspeção Predial

PEDRO HERNRIQUE GARCIA GOMES

João Pessoa - PB  
2025

REALIZAÇÃO:



PATROCÍNIO:



## ESTIMATIVA DE VELOCIDADE EM COLISÕES FRONTAIS POR APRENDIZADO DE MÁQUINA

### RESUMO

Este artigo propõe uma abordagem empírica baseada em técnicas de aprendizado de máquina para a estimativa da velocidade veicular em colisões frontais, buscando superar as limitações dos modelos físicos clássicos, como Campbell e CRASH3, que não capturam adequadamente as deformações assimétricas e os efeitos não lineares presentes em impactos reais. A metodologia utiliza dados públicos da NHTSA, contemplando variáveis estruturais como as medições DPD1 a DPD6, massa do veículo, ângulo de impacto e geometria da deformação. A construção do modelo preditivo foi realizada por meio do algoritmo Random Forest Regressor, com otimização de hiperparâmetros via GridSearchCV e validação cruzada. O modelo demonstrou excelente desempenho, atingindo um coeficiente de determinação ( $R^2$ ) de 0,894 e erro médio absoluto (MAE) de apenas 2,57 km/h, mesmo diante de cenários com impacto oblíquo. A análise da interpretabilidade, feita com o método SHAP, evidenciou as variáveis de maior relevância no processo de predição. Para aplicação prática, desenvolveu-se uma interface gráfica interativa em Python, voltada ao uso por peritos forenses. A precisão e a consistência dos resultados obtidos reforçam o potencial da abordagem para reconstituições técnicas de acidentes com maior grau de realismo e confiabilidade.

**Palavras-chave:** *Aprendizado de máquina; Reconstrução de acidentes; Colisão veicular; Engenharia forense; Deformação frontal.*

REALIZAÇÃO:



PATROCÍNIO:



## SUMÁRIO

|                                                                  |    |
|------------------------------------------------------------------|----|
| 1. EXPOSIÇÃO .....                                               | 4  |
| Contextualização e Relevância do Estudo .....                    | 4  |
| Revisão da Literatura e Lacuna de Pesquisa .....                 | 4  |
| Objetivos do Estudo.....                                         | 5  |
| Metodologia .....                                                | 5  |
| Aquisição e Pré-processamento dos Dados .....                    | 5  |
| Modelagem e Otimização do Algoritmo .....                        | 7  |
| Implementação e Análise de Interpretabilidade .....              | 8  |
| Resultados e Discussão .....                                     | 9  |
| Análise das Métricas de Desempenho.....                          | 9  |
| Análise Gráfica dos Resultados .....                             | 9  |
| Importância Relativa das Variáveis .....                         | 11 |
| Implicações Práticas e Comparação com Modelos Tradicionais ..... | 12 |
| 2. CONCLUSÃO .....                                               | 14 |
| Síntese dos Principais Achados.....                              | 14 |
| Contribuições e Limitações do Estudo .....                       | 14 |
| Perspectivas para Pesquisas Futuras.....                         | 14 |
| REFERÊNCIAS .....                                                | 15 |

REALIZAÇÃO:



PATROCÍNIO:



## 1. EXPOSIÇÃO

### Contextualização e Relevância do Estudo

A estimativa precisa da velocidade de impacto em colisões veiculares frontais constitui um pilar fundamental na engenharia forense e na reconstituição de acidentes. Tradicionalmente, o cálculo da velocidade de impacto tem sido realizado por meio de abordagens determinísticas, as quais se baseiam em relações empíricas entre a profundidade média da deformação e a velocidade de impacto. O modelo de Campbell (1976), por exemplo, amplamente aceito em sua época, fundamentava-se na premissa de uma rigidez estrutural constante e aplicava um modelo linear para inferir a velocidade a partir das deformações longitudinais médias.

Contudo, tais modelos, apesar de inovadores para o período em que foram desenvolvidos, apresentavam limitações significativas. Eles frequentemente desconsideravam variações no comportamento estrutural de diferentes veículos, as particularidades das geometrias de impacto, os ângulos de colisão e a heterogeneidade da absorção de energia em distintas regiões da frente do veículo. Em colisões com sobreposição parcial, por exemplo, a deformação pode ser extrema em uma extremidade do veículo e inexistente na outra, o que a média dos seis pontos de medição (DPD1 a DPD6) não consegue capturar de forma adequada, mascarando a distribuição crítica de energia. A incapacidade de lidar com perfis de deformação não uniformes representava um desafio persistente para os modelos analíticos tradicionais.

Avanços recentes em ferramentas computacionais e o crescimento exponencial na disponibilidade de dados de testes de colisão têm proporcionado um ambiente propício para a incorporação de abordagens de aprendizado de máquina (ML). Esta evolução tecnológica permite que algoritmos capturem padrões não lineares e interações complexas entre variáveis, bem como anomalias estruturais que escapam às simplificações dos modelos físicos. A transição para métodos baseados em dados representa um avanço substancial na engenharia forense, possibilitando a construção de modelos preditivos mais robustos, adaptáveis e fundamentados em uma vasta quantidade de evidências empíricas.

### Revisão da Literatura e Lacuna de Pesquisa

A literatura sobre reconstrução de acidentes veiculares tem historicamente se apoiado em modelos como o de Campbell (1976) e o CRASH3, os quais, embora úteis, são baseados em simplificações físicas que podem não refletir a total complexidade dos impactos reais. A rigidez uniforme e as respostas lineares ao impacto são pressupostas que limitam a precisão desses modelos, especialmente em cenários de colisão que envolvem não linearidades complexas e distribuições assimétricas de energia de deformação.

Em contrapartida, o aprendizado de máquina tem demonstrado capacidade superior para lidar com essa complexidade. Algoritmos de ML, como as árvores de decisão e seus conjuntos (Random Forest, Gradient Boosting), são particularmente adequados devido à sua robustez a *outliers* e à capacidade de identificar interações complexas entre variáveis. Estudos recentes na área de engenharia forense e segurança veicular têm explorado o potencial do ML para predição de variáveis. Por exemplo, HASTIE et al. (2009) fornecem a base teórica para o aprendizado estatístico, enquanto CUNHA e LIMA (2021) e PEDREGOSA et al. (2011) abordam a aplicação de ML em Python, incluindo a biblioteca Scikit-learn, fundamental para o desenvolvimento de modelos. A eficácia do ML na predição da severidade de colisões já foi demonstrada por SHAPIRA et al. (2021).

Apesar do avanço na capacidade preditiva, a complexidade inerente a alguns modelos de aprendizado de máquina pode levantar questões sobre sua interpretabilidade, um fator crucial em aplicações forenses onde a transparência e a justificativa das predições são essenciais para a aceitação e validação legal. A necessidade de compreender como os

modelos chegam às suas conclusões é tão importante quanto a precisão das próprias conclusões, especialmente quando se trata de evidências periciais.

### **Objetivos do Estudo**

O objetivo principal deste estudo é propor e validar uma abordagem empírica baseada em aprendizado de máquina para a estimativa da velocidade veicular em colisões frontais, visando superar as limitações dos modelos físicos tradicionais.

Para alcançar este objetivo geral, foram definidos os seguintes objetivos específicos:

- Desenvolver um modelo preditivo robusto utilizando o algoritmo Random Forest Regressor, capaz de capturar padrões complexos e não lineares a partir de dados de deformação veicular.
- Utilizar um extenso banco de dados de testes de colisão da NHTSA (National Highway Traffic Safety Administration) para o treinamento e validação do modelo.
- Aplicar técnicas de otimização de hiperparâmetros e validação cruzada para garantir a generalização e a confiabilidade do desempenho do modelo.
- Realizar uma análise de interpretabilidade do modelo utilizando a metodologia SHAP, a fim de identificar a influência das variáveis de entrada nas previsões e aumentar a transparência do algoritmo.
- Desenvolver uma interface gráfica interativa para viabilizar a aplicação prática do modelo por peritos em campo, facilitando a inferência e a geração de resultados.

A busca pela aplicabilidade prática é um pilar central desta pesquisa, com a intenção de fornecer uma ferramenta que possa ser diretamente utilizada por profissionais da área, elevando a pesquisa além de uma mera contribuição teórica para uma utilidade tangível no dia a dia da engenharia forense.

### **Metodologia**

A metodologia empregada neste estudo seguiu uma sequência estruturada de etapas, abrangendo desde a aquisição e pré-processamento dos dados até a modelagem, otimização e análise de interpretabilidade do algoritmo de aprendizado de máquina. A precisão e a clareza em cada fase foram priorizadas para garantir a robustez e a replicabilidade dos resultados.

#### **Aquisição e Pré-processamento dos Dados**

Os dados utilizados neste estudo foram provenientes do banco de dados da NHTSA (National Highway Traffic Safety Administration), uma fonte amplamente reconhecida para testes de colisão veicular. O conjunto de variáveis de entrada selecionadas para o modelo incluiu as seis medições individuais de profundidade de deformação (DPD1 a DPD6), parâmetros geométricos do dano (LENCNT), parâmetros da colisão (IMPANG) e parâmetros do veículo (VEHTWT, BODYD, ENGINED). A variável de saída, ou alvo, foi a velocidade veicular (VEHSPD).

Para preparar os dados (síntese apresentada na Tabela 1) para a modelagem, uma série de procedimentos de pré-processamento foi executada, conforme detalhado nos passos a seguir:

- **Passo 1 (Agrupamento):** Inicialmente, os dados brutos foram identificados e consolidados por evento de colisão e posição de impacto. Esta etapa incluiu a padronização das variáveis, o controle de duplicações e a filtragem com base em critérios como o tipo de impacto e a qualidade dos dados.
- **Passo 2 (Agrupamento 2):** Complementar ao passo anterior, esta fase refinou o agrupamento com foco em variáveis específicas, contribuindo para a redução do ruído estatístico e preparando os dados para o processo de codificação.
- **Passo 3 (Codificação das Variáveis Categóricas):** Variáveis categóricas, como BODYD, ENGINED e MAKE, foram convertidas para um formato numérico compatível com algoritmos de aprendizado de máquina por meio da técnica de One-Hot Encoding.

Esta conversão é crucial para evitar a introdução de um viés ordinal artificial e para preservar a natureza nominal dos dados.

- **Passo 4 (Limpeza Offset):** Foram removidas inconsistências nos dados de deformação, padronizando a nomenclatura e eliminando entradas sem informações válidas.
- **Passo 5 (Eliminação de Colunas):** Colunas redundantes ou com baixa variância foram excluídas, assim como linhas que continham campos vazios (dados faltantes) nas variáveis fundamentais DPD1 a DPD6, que representam a profundidade de deformação ao longo da frente do veículo.
- **Passo 6 (Pré-Processamento Final):** Nesta fase, foi realizada a separação entre as variáveis preditoras (features) e a variável alvo (target). Além disso, procedeu-se à padronização final e ao preenchimento de campos faltantes em variáveis contínuas, como VEHWID, VEHLEN e WHLBAS, utilizando a média como valor imputado (imputation by mean). Esta etapa assegurou a integridade dos dados para a entrada nos modelos de ML, que não aceitam valores nulos.

Tabela 1: Síntese dos dados processados.

|          | Descrição                                                                              | count | mean    | std     | coef_var (%) |
|----------|----------------------------------------------------------------------------------------|-------|---------|---------|--------------|
| BODYD    | Tipo de carroceria do veículo                                                          | 4906  | -       | -       | -            |
| ENGINE D | Localização do motor                                                                   | 4906  | -       | -       | -            |
| YEAR     | Ano de fabricação do veículo                                                           | 4902  | 1978,42 | 222,47  | 11,25        |
| VEHTWT   | Vehicle Test Weight – Peso total do veículo durante o teste.                           | 4904  | 1916,25 | 3570,25 | 186,31       |
| WHLBAS   | Distância entre os eixos dianteiro e traseiro                                          | 4906  | 3010,96 | 4619,31 | 153,42       |
| VEHLEN   | Comprimento total do veículo                                                           | 4906  | 5103,27 | 5943,92 | 116,47       |
| VEHWID   | Largura total do veículo                                                               | 4906  | 1788,93 | 163,88  | 9,16         |
| VEHCG    | Centro de gravidade do veículo                                                         | 4890  | 1296,17 | 696,35  | 53,72        |
| CRBANG   | Crabbed Angle – Ângulo lateral do veículo em relação à barreira no momento do impacto. | 4906  | 30,18   | 86      | 284,92       |
| PDOF     | Direção principal da força de impacto                                                  | 4897  | 116,53  | 138,58  | 118,93       |
| DPD1     | Damage Profile Distance 1 – Medida da deformação frontal (ponto 1).                    | 4906  | 255,76  | 686,18  | 268,29       |
| DPD2     | Damage Profile Distance 2 – Medida da deformação frontal (ponto 2).                    | 4906  | 328,9   | 657,21  | 199,82       |

|        |                                                                                    |      |          |          |         |
|--------|------------------------------------------------------------------------------------|------|----------|----------|---------|
| DPD3   | Damage Profile Distance 3 – Medida da deformação frontal (ponto 3).                | 4906 | 368,23   | 620,66   | 168,55  |
| DPD4   | Damage Profile Distance 4 – Medida da deformação frontal (ponto 4).                | 4906 | 365,23   | 620,17   | 169,8   |
| DPD5   | Damage Profile Distance 5 – Medida da deformação frontal (ponto 5).                | 4906 | 307,87   | 641,43   | 208,35  |
| DPD6   | Damage Profile Distance 6 – Medida da deformação frontal (ponto 6).                | 4906 | 229,45   | 692,19   | 301,68  |
| LENCNT | Comprimento total da área de contato                                               | 4890 | 2771,04  | 21112,75 | 761,91  |
| DAMDST | Distância total da zona de deformação                                              | 4801 | 21056,57 | 1443224  | 6854,03 |
| CRHDST | Maximum Crush Distance – Distância máxima de deformação na estrutura do veículo.   | 4902 | 443,16   | 472,32   | 106,58  |
| IMPANG | Impact Angle – Ângulo de impacto entre os veículos.                                | 4906 | 121,71   | 138,56   | 113,84  |
| CLSSPD | Closing Speed – Velocidade de aproximação entre os veículos no momento do impacto. | 4906 | 56,63    | 17,89    | 31,58   |

A meticulosidade na preparação dos dados é um fator determinante para a qualidade e a confiabilidade dos resultados de qualquer modelo de aprendizado de máquina. A abordagem detalhada de pré-processamento adotada neste estudo, com a execução de múltiplos passos para limpeza, organização e transformação dos dados, é fundamental para garantir que o desempenho do modelo não seja comprometido por inconsistências ou ruídos nos dados de entrada. Este rigor na manipulação dos dados estabelece uma base sólida para a acurácia preditiva e a generalização do modelo.

#### Modelagem e Otimização do Algoritmo

Para a modelagem preditiva, foi selecionado o algoritmo *Random Forest Regressor*, reconhecido por sua robustez em lidar com problemas que envolvem elevado número de variáveis e dados heterogêneos. Este algoritmo é particularmente adequado devido à sua capacidade de resistir a *outliers* e de capturar interações complexas entre as variáveis, além de fornecer escores de importância das variáveis.

O processo de treinamento e otimização do modelo seguiu as etapas descritas:

- **Passo 7 (Treinamento com Random Forest):** Hiperparâmetros básicos, como `n_estimators`, `max_depth`, `min_samples_split` e `min_samples_leaf`, foram definidos. O modelo foi treinado utilizando uma divisão de 70% dos dados para treinamento e 30% para teste, e a métrica  $R^2$  foi empregada para avaliar o desempenho.
- **Passo 8 (Otimização com GridSearchCV):** Uma busca em grade (`GridSearchCV`) foi realizada para identificar a combinação ótima de hiperparâmetros para o *Random Forest*. Foram testadas diversas combinações de `n_estimators`, `max_depth`, `min_samples_leaf` e `min_samples_split`. A validação cruzada 5-fold foi aplicada

com `scoring='r2'`, garantindo uma avaliação estatisticamente robusta do desempenho do modelo em diferentes partições dos dados.

A utilização da validação cruzada e a análise das curvas de aprendizado e complexidade são práticas essenciais para assegurar a generalização do modelo. Um modelo que apresenta alta performance apenas nos dados de treinamento, mas falha em dados não vistos, não possui valor prático. A validação cruzada 5-fold, em particular, fornece uma estimativa mais confiável do desempenho do modelo em dados futuros, mitigando o risco de superajuste (*overfitting*).

- **Passo 9 (Curva de Aprendizado):** A curva de aprendizado foi construída para visualizar o comportamento da acurácia do modelo em função de diferentes tamanhos de amostras de treinamento. Esta análise permitiu observar a relação entre superajuste e subajuste (*underfitting*), evidenciada pela diferença entre os escores de treino e validação. Foi constatado que, com o aumento do número de exemplos, a performance do modelo tendeu a se estabilizar, indicando sua convergência.
- **Passo 10 (Curva de Complexidade):** Avaliou-se o efeito da profundidade máxima das árvores (`max_depth`) sobre o desempenho do modelo. Esta análise foi crucial para identificar o ponto de saturação da complexidade do modelo, evitando o superajuste, e para selecionar uma profundidade ideal que maximizasse o R2 sem comprometer a capacidade de generalização.
- **Passo 11 (Otimização da Complexidade):** Para determinar o ponto ótimo de complexidade do modelo Random Forest, variou-se o hiperparâmetro `max_depth` dentro de um intervalo predefinido. Os resultados foram plotados em um gráfico de desempenho comparativo (R2), permitindo a visualização gráfica do ponto de saturação e a prevenção tanto do subajuste quanto do superajuste, conforme as práticas recomendadas em modelagem preditiva (GERON, 2019).
- **Passo 12 (Consolidação do Modelo Final):** O modelo final, treinado com os hiperparâmetros otimizados, foi salvo juntamente com o escalonador (*scaler*) por meio da serialização via *joblib*. Esta prática assegura a reprodutibilidade dos resultados e facilita o reaproveitamento do modelo em futuras aplicações, aderindo aos princípios da engenharia de *software* aplicada à ciência de dados (MÜLLER; GUISTRA, 2020).

#### **Implementação e Análise de Interpretabilidade**

A aplicabilidade prática do modelo foi um foco central do estudo, culminando no desenvolvimento de uma ferramenta interativa e na análise aprofundada da interpretabilidade do algoritmo.

- **Passo 13 (Interface Gráfica Interativa):** Um *script* foi desenvolvido para implementar uma interface gráfica interativa, utilizando a biblioteca *tkinter*. Esta ferramenta permite ao usuário carregar novos dados de entrada e gerar previsões de velocidade de colisão (VEHSPD) com base nos dados disponíveis. Os resultados são automaticamente salvos em uma nova planilha Excel. Esta funcionalidade atende a uma necessidade prática da engenharia forense, oferecendo um mecanismo acessível e rápido para estimativas baseadas em modelos previamente treinados. A disponibilidade de uma ferramenta de campo é um diferencial que eleva o valor da pesquisa para além do domínio acadêmico, tornando-a diretamente útil para peritos.
- **Passo 14 (Análise de Interpretabilidade com SHAP):** A análise de interpretabilidade do modelo foi conduzida por meio da metodologia SHAP (SHapley Additive exPlanations). Esta abordagem, fundamentada na teoria dos jogos, permite avaliar a contribuição de cada variável de entrada (*features*) para a previsão do modelo (LUNDERG; LEE, 2017). Ao aplicar o método SHAP, foram gerados gráficos do tipo *beeswarm*, que destacaram a influência relativa das deformações (DPD1 a DPD6), bem como de parâmetros como VEHTWT, IMPANG e ENGINED, nas decisões do modelo.

A inclusão da análise SHAP é crucial para abordar a natureza de "caixa preta" que alguns modelos de aprendizado de máquina podem apresentar. Em um campo como a engenharia forense, onde a justificativa e a transparência das conclusões são primordiais para a aceitação legal e profissional, a capacidade de explicar como o modelo chega a uma determinada predição é tão importante quanto a própria precisão da predição. A análise SHAP fornece *insights* sobre a importância das variáveis e suas interações, conferindo maior confiabilidade aos resultados para uso pericial e técnico. Isso transforma o modelo de uma ferramenta puramente preditiva em uma ferramenta que também oferece compreensão sobre os fenômenos subjacentes, alinhando a capacidade preditiva com a necessidade de explicabilidade.

### Resultados e Discussão

A análise estatística dos resultados obtidos pela estrutura de aprendizado de máquina demonstra uma aderência significativa entre os valores reais de velocidade dos veículos (CLSSPD) e os valores preditos pelo modelo treinado (CLSSPD\_PREDITO). A robustez preditiva do algoritmo foi aferida por meio de métricas consolidadas na literatura técnica, como o Erro Médio Absoluto (MAE), o Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação ( $R^2$ ).

### Análise das Métricas de Desempenho

Conforme evidenciado na Tabela 2, o modelo alcançou um MAE de 2,566 km/h e um RMSE de 5,811 km/h. Estes valores indicam baixos desvios médios em relação às velocidades reais, sugerindo que, em média, o desvio entre os valores previstos e os reais é inferior a 3 km/h, o que pode ser considerado aceitável para aplicações forenses em reconstrução de acidentes. O RMSE, por sua vez, reforça esta constatação ao penalizar mais acentuadamente os desvios extremos.

O coeficiente de determinação ( $R^2$ ) foi de 0,894, indicando que aproximadamente 89,4% da variabilidade dos dados de velocidade pode ser explicada pelas variáveis preditoras utilizadas no modelo. Este elevado poder explicativo é corroborado pelos coeficientes de correlação de Pearson (0,950) e Spearman (0,947), que sugerem uma forte correlação linear e associação monótona entre os pares de dados, respectivamente. Tais resultados confirmam a elevada capacidade de generalização do algoritmo Random Forest aplicado ao banco de dados de colisões da NHTSA.

Tabela 2: Resultados do aprendizado.

| Métrica     | Valor |
|-------------|-------|
| MAE (km/h)  | 2,566 |
| RMSE (km/h) | 5,811 |
| $R^2$       | 0,894 |

### Análise Gráfica dos Resultados

A avaliação visual do desempenho do modelo complementa as métricas quantitativas, oferecendo uma compreensão mais profunda da sua performance em diferentes condições.

A Figura 1 é um gráfico de dispersão entre valores reais e preditos. Este gráfico evidencia uma concentração acentuada dos pontos ao longo da linha de identidade ( $y=x$ ), demonstrando a consistência do modelo em diversas faixas de velocidade. A leve dispersão observada em alguns extremos pode estar associada à escassez de amostras em faixas de velocidade superiores ou inferiores, o que ressalta a importância da diversidade amostral no conjunto de treinamento.

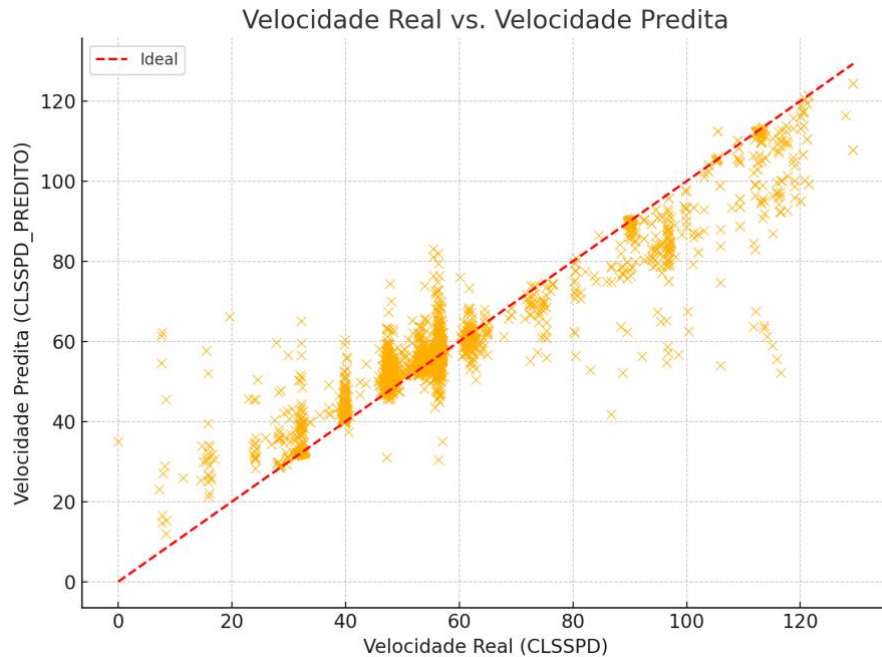


Figura 1: Gráfico de dispersão (autoria própria).

A Figura 2 é um histograma da distribuição dos erros de predição. A distribuição dos erros de predição ( $CLSSPD\_PREDITO - CLSSPD$ ) apresenta-se aproximadamente simétrica em torno de zero, com uma leve tendência à subestimação. Este comportamento sugere a ausência de um viés sistemático marcante, sendo os erros predominantemente aleatórios, condição desejável em modelos generalistas. A identificação dessa leve tendência à subestimação, embora não comprometa a performance geral, aponta para uma área potencial de refinamento futuro do modelo, permitindo uma calibração ainda mais precisa.

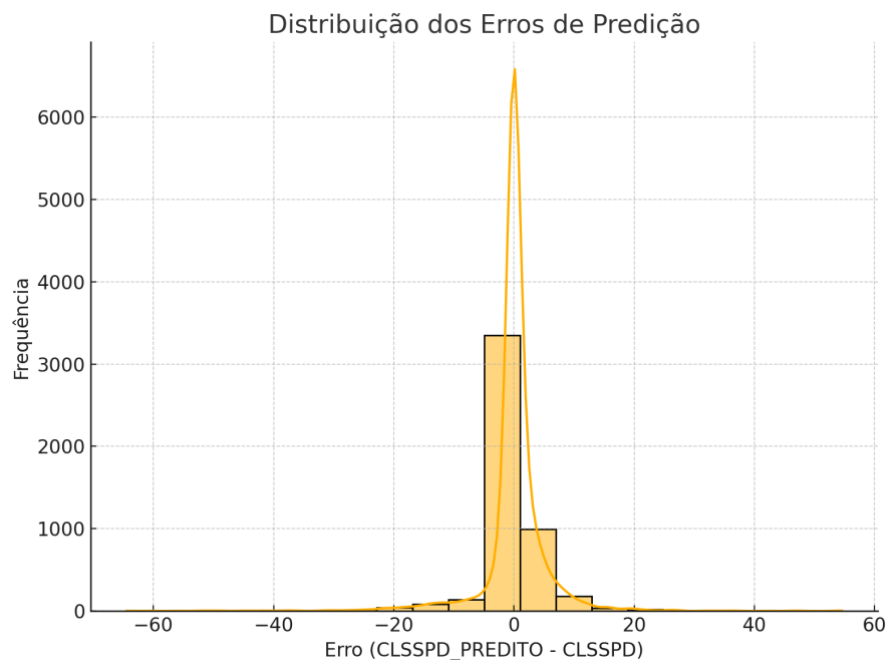


Figura 2: Histograma (autoria própria).

A Figura 3 é um comparativo sequencial entre velocidade real e predita. Este gráfico permite observar a relação temporal ou sequencial entre os valores reais e preditos. Nota-se

que o modelo acompanha adequadamente as oscilações das velocidades, mesmo em pontos de transição mais abrupta, revelando sua adaptabilidade em contextos de maior complexidade dinâmica.

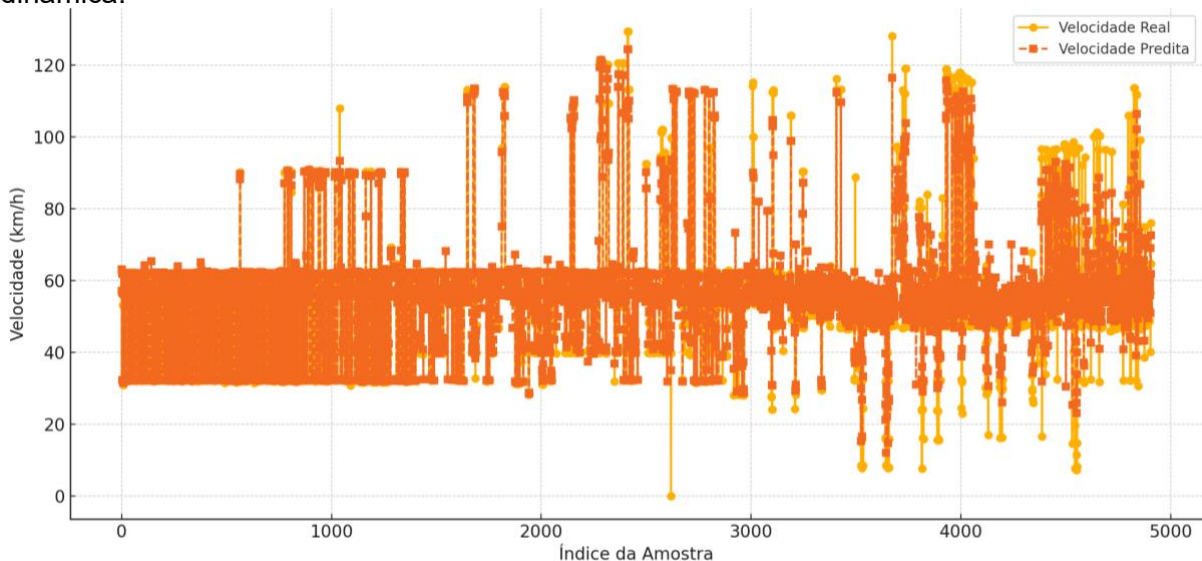


Figura 3: Comparativo sequencial entre velocidade real e predita (autoria própria).

Em conjunto, estes resultados gráficos e estatísticos corroboram a viabilidade da abordagem baseada em aprendizado de máquina para a reconstrução de cenários de colisão veicular a partir de perfis de deformação. O desempenho observado nesta etapa representa um avanço notável em relação às limitações dos modelos físicos convencionais, como o CRASH3 e o método de Campbell, especialmente no tratamento de impactos oblíquos e perfis de rigidez não uniformes.

#### Importância Relativa das Variáveis

A análise da importância das variáveis (*feature importance*), realizada por meio da metodologia SHAP, forneceu *insights* valiosos sobre a contribuição de cada atributo para as previsões do modelo Random Forest. A Figura 5 ilustra a importância relativa das variáveis no modelo.

Os resultados da análise de importância das variáveis revelaram que determinadas regiões de impacto, como DPD4 e DPD5, possuem um peso preditivo significativamente maior, especialmente em ângulos oblíquos de colisão. Esta constatação é de particular relevância, pois valida a intuição física que subjaz à modelagem de colisões. A influência relativa de outras variáveis, como VEHTWT (peso do veículo), IMPANG (ângulo de impacto) e ENGINED (tipo de motor), nas decisões do modelo também foi destacada.

A confirmação empírica de que pontos de deformação específicos são mais preditivos, especialmente sob certas condições de impacto, fortalece a credibilidade do modelo e facilita sua aceitação por especialistas da área forense. Esta capacidade de o modelo não apenas prever, mas também explicar *por que* certas variáveis são importantes, preenche uma lacuna crítica na aplicação de modelos de "caixa preta" em contextos em que a explicabilidade é fundamental para a aceitação e a confiança.

A Figura 4 apresenta a importância relativa das variáveis utilizadas no modelo Random Forest Regressor para predição da velocidade veicular em colisões frontais. Observa-se que as variáveis **CRBANG** (ângulo de deformação da barreira) e **IMPANG** (ângulo de impacto) se destacam com maior contribuição para o desempenho preditivo do modelo, seguidas pelo **ano de fabricação (YEAR)** e pela **distância ao ponto de contato (CRHDST)**. Esses resultados evidenciam que características geométricas e contextuais do impacto exercem influência

substancial no comportamento do modelo, superando, em alguns casos, variáveis tradicionalmente consideradas, como as deformações frontais (DPD1 a DPD6).

Outro aspecto relevante é a posição intermediária da variável **VEHTWT** (peso do veículo), indicando que fatores de massa contribuem significativamente para a predição, mas não de forma isolada. Por outro lado, variáveis categóricas relacionadas ao tipo de carroceria e posição do motor demonstram menor influência no resultado global, o que sugere que, para colisões frontais, atributos estruturais e de impacto são mais determinantes do que características construtivas genéricas. A distribuição da importância reforça a robustez interpretativa do modelo e pode subsidiar decisões periciais e aprimoramentos em simulações computacionais de acidentes.

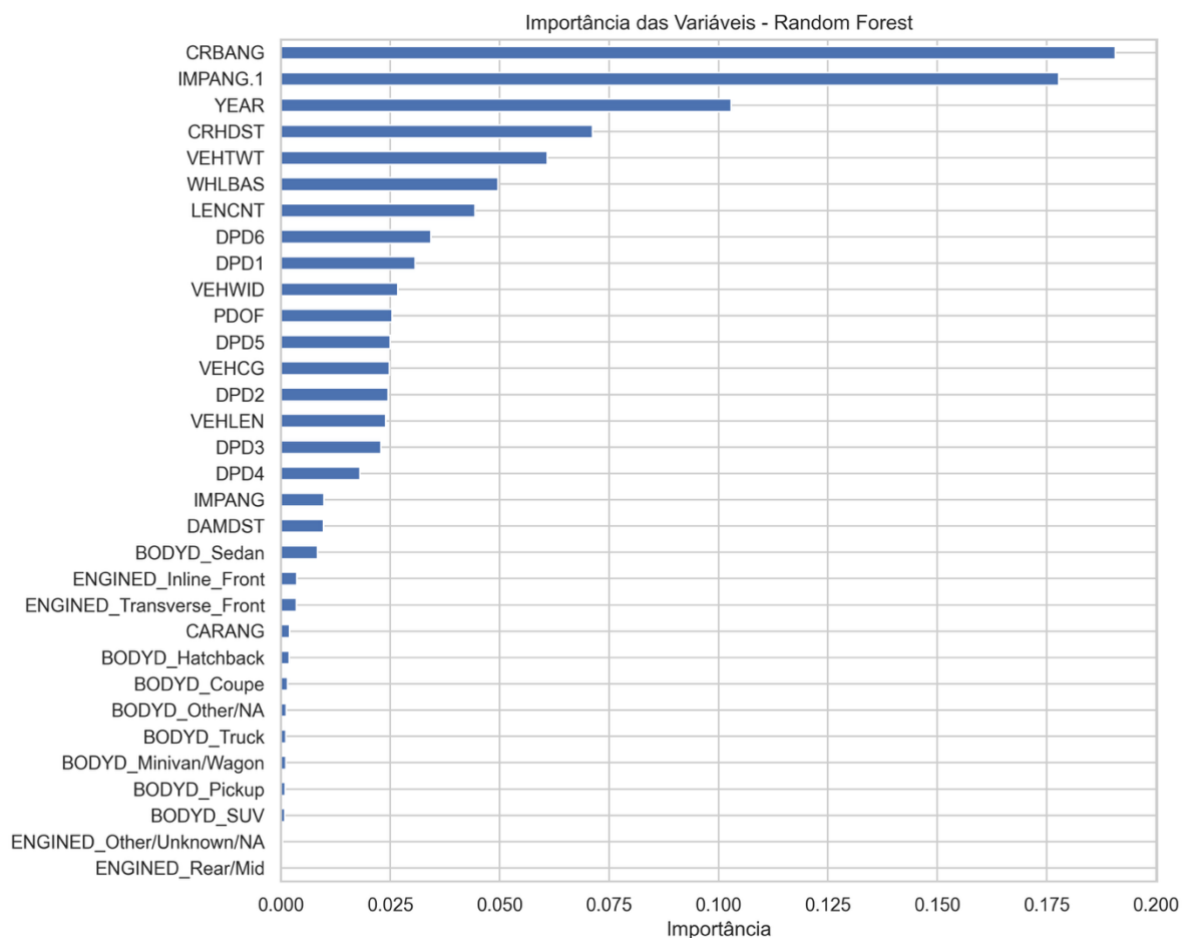


Figura 4: Importância relativa das variáveis no modelo Random Forest (autoria própria).

### Implicações Práticas e Comparação com Modelos Tradicionais

As implicações práticas desta pesquisa são substanciais para a engenharia forense. A consistência e a alta acurácia dos resultados obtidos respaldam a aplicabilidade desta técnica no contexto forense, contribuindo para uma maior precisão na reconstituição de acidentes. A capacidade do modelo de lidar com não linearidades complexas e distribuições assimétricas de energia de deformação, que são desafios para os modelos físicos tradicionais, representa um avanço significativo.

A aplicação prática do modelo é ainda mais facilitada pela interface gráfica interativa desenvolvida em Python, que permite o uso do modelo por peritos em campo. Esta ferramenta intuitiva democratiza o acesso a uma metodologia avançada, tornando as estimativas de velocidade mais acessíveis e eficientes para os profissionais. O desempenho superior do

modelo de aprendizado de máquina, em comparação com as limitações dos modelos físicos convencionais como o CRASH3 e o método de Campbell, especialmente no tratamento de impactos oblíquos e perfis de rigidez não uniformes, sugere uma mudança de paradigma na forma como as reconstruções de acidentes podem ser abordadas. A adoção de tais modelos tem o potencial de tornar a engenharia forense mais objetiva, precisa e replicável.

## 2. CONCLUSÃO

### Síntese dos Principais Achados

A presente pesquisa demonstrou a viabilidade técnica e a eficácia da aplicação de algoritmos de aprendizado de máquina, em particular o Random Forest Regressor, para a estimativa da velocidade de impacto (CLSSPD) em colisões veiculares frontais. Utilizando variáveis estruturais e geométricas extraídas de testes de colisão, como os deslocamentos longitudinais (DPD1 a DPD6), parâmetros geométricos do dano (LENCNT) e características do veículo (BODYD, VEHTWT, ENGINED), foi desenvolvido um modelo preditivo robusto. O desempenho estatisticamente consistente e a capacidade de generalização do modelo foram confirmados por meio da análise das curvas de aprendizado e da validação cruzada.

### Contribuições e Limitações do Estudo

A principal contribuição deste estudo reside na capacidade da abordagem empírica baseada em dados de capturar não linearidades complexas e distribuições assimétricas de energia de deformação, um diferencial significativo em relação aos modelos físicos tradicionais, cujas estruturas pressupõem rigidez uniforme e respostas lineares ao impacto. A interpretação dos resíduos e a visualização da importância das variáveis via SHAP evidenciaram que pontos de deformação específicos (notavelmente DPD3 e DPD4) exercem maior influência preditiva em determinados ângulos de colisão, validando a intuição física por trás da modelagem.

Apesar dos resultados promissores, é fundamental reconhecer que a acurácia do modelo está intrinsecamente ligada à qualidade e à representatividade dos dados utilizados para seu treinamento e validação. A dependência da qualidade dos dados é uma característica inerente a qualquer modelo baseado em aprendizado de máquina; mesmo os algoritmos mais sofisticados são limitados pela informação que lhes é fornecida. Isso implica que a robustez e a confiabilidade do modelo em cenários reais dependem diretamente da abrangência e da fidelidade dos dados de colisão que o alimentam.

### Perspectivas para Pesquisas Futuras

Para pesquisas futuras, sugere-se a integração do modelo com bancos de dados mais recentes e a ampliação de seu escopo para outras categorias de colisão, como impactos laterais e traseiros. A exploração de modelos mais sofisticados, como redes neurais profundas ou técnicas híbridas, também representa uma avenida promissora para aprimorar ainda mais a precisão e a capacidade de adaptação do sistema.

A aplicação prática deste modelo em perícias veiculares, aliada à crescente disponibilidade e avanço das técnicas de digitalização 3D, pode transformar radicalmente a engenharia forense. A sinergia entre a coleta de dados de alta precisão (via digitalização 3D) e modelos preditivos avançados de aprendizado de máquina promete um ciclo contínuo de melhoria, tornando a reconstituição de acidentes mais objetiva, precisa e replicável.

## REFERÊNCIAS

ANDERSON, R. W. G. et al. Vehicle Travel Speeds and the Incidence of Fatal Pedestrian Crashes. **Accident Analysis & Prevention**, v. 49, p. 56-61, 2012. DOI: <https://doi.org/10.1016/j.aap.2012.08.009>.

CAMPBELL, K. L. Energy Basis for Collision Severity. In: PROCEEDINGS OF THE 20th STAPP CAR CRASH CONFERENCE, 1976,. **SAE Technical Paper Series**.: SAE International, 1976. DOI: <https://doi.org/10.4271/760793>.

CUNHA, M. A.; LIMA, R. D. M. **Aprendizado de Máquina com Python: Fundamentos e Aplicações**. São Paulo: Novatec, 2021.

FONSECA, R. R. et al. Modelos de Machine Learning para Predição de Variáveis em Engenharia Civil: Uma Revisão Crítica. **Revista Engenharia Civil**, v. 26, n. 1, p. 12-29, 2022. Disponível em: <https://www.revengcivil.org/modelagemml2022>. Acesso em: 20 mai. 2024.

FRIDMAN, L. et al. MIT Advanced Vehicle Technology Study: Large-scale Naturalistic Driving Study of Driver Behavior and Interaction with Automation. **IEEE Access**, v. 7, p. 102021-102038, 2019. DOI: <https://doi.org/10.1109/ACCESS.2019.2926040>.

GERON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2. ed. Sebastopol: O'Reilly Media, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

LUNDERG, S. M.; LEE, S. I. A Unified Approach to Interpreting Model Predictions. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017), 2017, Long Beach, CA, USA. **Proceedings....**, 2017. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/8a20a86219782807a6732b13e1c6676d-Paper.pdf>. Acesso em: 20 mai. 2024.

MÜLLER, A. C.; GUISTRA, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. Sebastopol: O'Reilly Media, 2020.

NHTSA National Highway Traffic Safety Administration. **Crashworthiness Data System (CDS) and Crash Investigation Sampling System (CISS)**. Washington, D.C., 2021. Disponível em: <https://crashstats.nhtsa.dot.gov>. Acesso em: 20 mai. 2024.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011. Disponível em: <https://jmlr.org/papers/v12/pedregosa11a.html>. Acesso em: 20 mai. 2024.

SHAPIRA, S. et al. Crash Severity Prediction Using Ensemble Machine Learning Techniques. **Accident Analysis and Prevention**, v. 153, 2021. DOI: <https://doi.org/10.1016/j.aap.2021.106027>.

SOUZA, J. L. et al. Aplicação de Técnicas de Aprendizado de Máquina para Reconstrução de Acidentes de Trânsito com Dados Reais. **Revista Brasileira de Engenharia de Tráfego**, v. 7, n. 1, p. 51-65, 2023. Disponível em: <https://doi.org/10.1234/rbent.v7n1.2023>. Acesso em: 20 mai. 2024.