# Previsão de Fluxo de Tráfego Rodoviário com *Machine Learning*

Pedro Reis
*Escola de Engenharia*
*Universidade do Minho*
Braga, Portugal
PG59908@alunos.uminho.pt

Guilherme Pinto
*Escola de Engenharia*
*Universidade do Minho*
Braga, Portugal
PG60225@alunos.uminho.pt

João Azevedo
*Escola de Engenharia*
*Universidade do Minho*
Braga, Portugal
PG61693@alunos.uminho.pt

Luís Silva
*Escola de Engenharia*
*Universidade do Minho*
Braga, Portugal
PG60390@alunos.uminho.pt

***Abstract*—O presente relatório documenta o desenvolvimento de modelos de *Machine Learning* para a previsão multiclasse de fluxo de tráfego rodoviário na cidade do Porto, estruturado com a metodologia CRISP-DM. O estudo incidiu sobre um conjunto de dados históricos (2018-2019), abordando desafios críticos de preparação de dados, nomeadamente o preenchimento de valores em falta em variáveis meteorológicas, a transformação cíclica de atributos temporais e o severo desequilíbrio de classes, mitigado através da técnica de *Oversampling* SMOTE combinada com validação cruzada estratificada.**

**Foram implementadas e comparadas diversas famílias de algoritmos para aferir o equilíbrio entre viés e variância: desde modelos de referência lineares e não-lineares (*Logistic Regression*, SVM), passando por métodos de *Deep Learning* (Redes Neuronais Artificiais), arquiteturas de especialistas (*One-vs-Rest*, *Mixture of Experts*), até métodos de *ensemble* avançados (*Random Forest*, *XGBoost*, *Gradient Boosting*).**

**Os resultados demonstraram que a arquitetura de *Stacking* — integrando a heterogeneidade de cinco estimadores — obteve o melhor desempenho global, atingindo uma *accuracy* de 82% nos dados de validação. A robustez desta solução foi validada externamente através de uma competição na plataforma Kaggle, onde a consistência entre a métrica local, o *Public Score* e o *Private Score* confirmou a elevada capacidade de generalização do modelo com pouco *overfitting*. A análise de interpretabilidade (SHAP) validou a coerência física do sistema, identificando o diferencial de tempo acumulado como o fator preditor dominante, modulado contextualmente pela luminosidade e hora do dia.**

***Index Terms*—*Machine Learning*, Previsão de Tráfego, CRISP--DM, Stacking, SMOTE, Classificação, Ensemble, *Multiclass***

## I. Introdução

### A. Contextualização

Este relatório apresenta o desenvolvimento de modelos de *Machine Learning* capazes de prever o fluxo de tráfego rodoviário, no âmbito da UC de Dados e Aprendizagem Automática.

O trabalho recorreu a um histórico de dados fornecido superior a um ano (entre 24 de julho de 2018 e 2 de outubro de 2019) para treinar e explorar diversas abordagens. O projeto integrou a participação numa competição na plataforma Kaggle, utilizada estrategicamente para a validação dos modelos, permitindo aferir a sua capacidade de generalização.

### B. Metodologia

Entre as diferentes metodologias analisadas, a selecionada para a realização deste trabalho foi o CRISP-DM (*Cross Industry Standard Process for Data Mining*). Esta decisão fundamenta-se, primordialmente, no seu caráter iterativo, característica que se revela ideal para lidar com a complexidade e as constantes adaptações exigidas no desenvolvimento de modelos de *Machine Learning*.

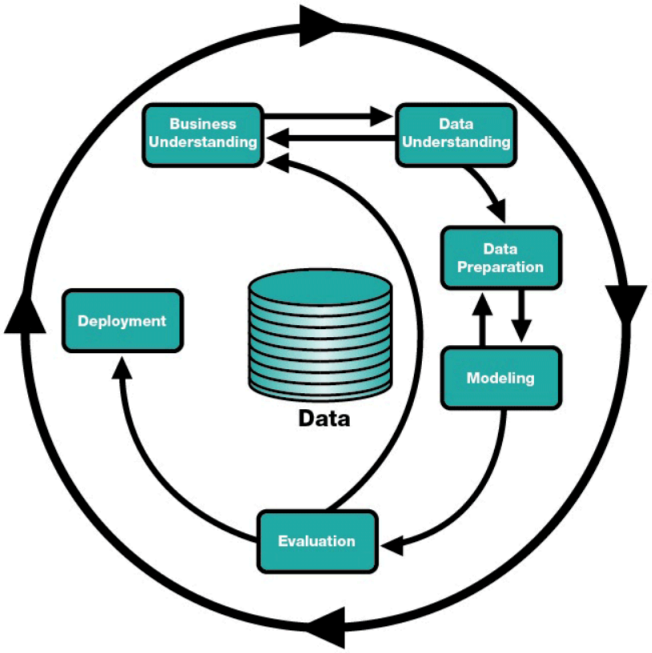


Fig. 1. Metodologia CRISP-DM

Seguindo a estrutura do CRISP-DM, o processo desenvolveu-se da seguinte forma:

- **Compreensão do negócio:** Estabelecimento dos objetivos a atingir com os modelos de previsão.
- **Compreensão dos dados:** Análise exploratória para entender a natureza das variáveis.
- **Preparação dos dados:** Limpeza e transformação dos dados para alimentar os algoritmos.
- **Modelação:** Criação e treino dos modelos.
- **Avaliação:** Análise das métricas de desempenho.

A última fase, correspondente à **Implementação**, não será abordada no âmbito estrito de colocação em produção, dado o contexto acadêmico do trabalho.

### C. Estrutura do documento

O restante deste relatório organiza-se da seguinte forma, a Secção II detalha as especificidades do problema, enquanto a Secção III foca-se na caracterização dos dados através de análises uni-variada e multivariada.

Posteriormente, a Secção IV documenta as técnicas de tratamento de dados e *feature engineering* aplicadas. As arquiteturas dos algoritmos selecionados, bem como as respetivas configurações de treino, são descritas na Secção V. O documento encerra com a discussão comparativa dos desempenhos obtidos na Secção VI e a síntese das conclusões na Secção VII, seguindo-se a listagem das referências bibliográficas utilizadas no trabalho.

## II. Compreensão do Negócio

### A. Enquadramento do Problema

O desafio central consiste na previsão precisa do fluxo de tráfego num intervalo temporal específico. A capacidade de antecipar congestionamentos não possui apenas valor técnico, mas também social, permitindo otimizar rotas de emergência, reduzir emissões poluentes e melhorar a qualidade de vida urbana através de uma gestão de tráfego proativa.

### B. Objetivos Estratégicos

Para garantir o sucesso da solução, definiram-se os seguintes objetivos:

- **Extração de conhecimento:** Identificar os atributos que mais influenciam o fluxo rodoviário.
- **Excelência preditiva:** Desenvolver modelos que maximizem a *accuracy*.
- **Robustez e generalização:** Garantir que o modelo mantém a performance em cenários de tráfego atípicos evitando *overfitting*.
- **Validação competitiva:** Obter uma classificação de destaque no *ranking* da competição Kaggle como prova de eficácia técnica.

### C. Entregáveis e Avaliação

Para cumprir os requisitos académicos e garantir a transparência do processo de desenvolvimento, o projeto estrutura-se nos seguintes componentes:

- **Relatório:** Documento que descreve a metodologia, a exploração de dados, as arquiteturas dos modelos e a análise crítica dos resultados obtidos.
- **Participação na competição:** Submissão de previsões na plataforma para aferição da métrica de *accuracy* e posicionamento no *ranking*.
- **Artefactos de software:** Pasta comprimida contendo o código-fonte, dados processados e instruções que permitam a reprodutibilidade integral de todos os passos e resultados apresentados.

## III. Compreensão dos Dados

### A. Descrição do conjunto de dados

O conjunto de dados utilizado neste projeto provém de registos reais de tráfego rodoviário na cidade do Porto. A estrutura de dados está organizada em dois componentes principais:

- **`training_data.csv`:** O conjunto de treino que inclui as variáveis independentes e a variável dependente.
- **`test_data.csv`:** O conjunto de teste que contém os mesmos atributos do ficheiro de treino, com a omissão da variável alvo.

### B. Análise uni-variada

Nesta etapa, procedeu-se à caracterização isolada de cada atributo e da variável alvo, com o intuito de compreender a sua distribuição estatística e validar a integridade dos dados. Esta exploração é fundamental para identificar medidas de tendência central, avaliar a dispersão e detetar a presença de valores atípicos que possam comprometer a robustez dos modelos de aprendizagem.

**`AVERAGE_SPEED_DIFF`**

**Descrição:** Esta variável representa a diferença entre a velocidade média em condições de fluxo livre e a velocidade real observada. É o atributo que os modelos pretendem prever.
**Tipo:** Categórica ordinal.
**Valores únicos:** 5 (cinco).
**Valores nulos:** 0 (zero).

TABLE I
Quantificação AVERAGE_SPEED_DIFF

| Categoria | Frequência | Percentagem |
|---|---|---|
| None | 2200 | 32.3% |
| Low | 1419 | 20.8% |
| Medium | 1651 | 24.2% |
| High | 1063 | 15.6% |
| Very_High | 479 | 7.1% |

**Análise**: A análise da Tabela I revela um desbalanceamento de classes significativo. Este cenário exige cautela, pois o modelo poderá apresentar uma tendência para subestimar situações de tráfego extremo em favor de classes maioritárias.

**Nota:** Durante a ingestão dos dados, a string "None" foi erroneamente interpretada como um valor nulo (NaN). Então, foi aplicada uma correção sistemática de mapeamento para restaurar a categoria original, garantindo que a classe fosse corretamente contabilizada.

### city_name

**Descrição:** Nome da cidade.
**Tipo:** Categórica nominal.
**Valores nulos:** 0 (zero).
**Valores únicos:** 1 (um).
**Análise:** Considerando a existência apenas de um valor neste atributo, esta variável não tem poder preditivo e pode ser removida.

### record_date

**Descrição:** A marca temporal associada ao registo.
**Tipo:** Temporal.
**Formato:** 'MM/DD/YYYY HH:MM'.
**Valores nulos:** 0 (zero).

TABLE II
RECORD_DATE EM BRUTO

| ID | record_date |
|---|---|
| 0 | 8/29/2019 7:00 |
| 1 | 8/10/2018 14:00 |
| 2 | 9/1/2019 16:00 |
| 3 | 2/26/2019 11:00 |
| 4 | 6/6/2019 12:00 |

**Análise:** Embora a variável forneça o contexto temporal exato de cada observação, o seu formato bruto não é diretamente interpretável pela maioria dos algoritmos de aprendizagem. No entanto, possui uma elevada relevância preditiva latente.

### AVERAGE_FREE_FLOW_SPEED

**Descrição:** O valor médio da velocidade máxima que os carros podem atingir em cenários sem trânsito.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

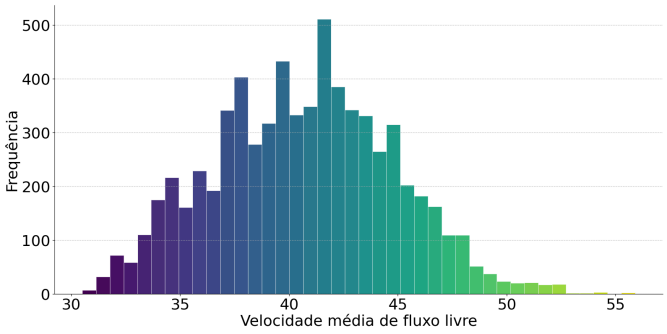


Fig. 2. Distribuição da velocidade de fluxo livre

**Análise:** A Fig. 2 mostra uma distribuição saudável da velocidade de fluxo livre, com média em torno de 40,7 km/h. A presença de valores acima de 50 km/h é minoritária, sugerindo que a amostra inclui algumas vias de maior capacidade onde os limites de velocidade são superiores ao padrão urbano ou a velocidade praticada pelos condutores exceda o limite legal estabelecido para a via.

### AVERAGE_FREE_FLOW_TIME

**Descrição:** O valor médio do tempo que demora a percorrer um determinado conjunto de ruas quando não há trânsito.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

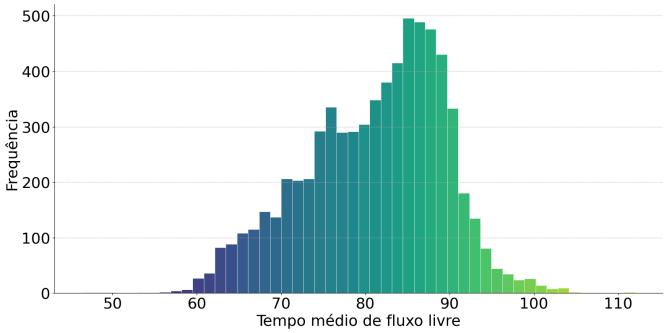


Fig. 3. Distribuição do tempo de fluxo livre

**Análise:** Na Fig. 3 os dados apresentam uma média de 81,1s e uma ligeira assimetria negativa, refletindo uma maior densidade de segmentos com tempos de percurso próximos do limite superior.

### AVERAGE_TIME_DIFF

**Descrição:** O valor médio da diferença do tempo que se demora a percorrer um determinado conjunto de ruas sem trânsito e com o trânsito atual.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

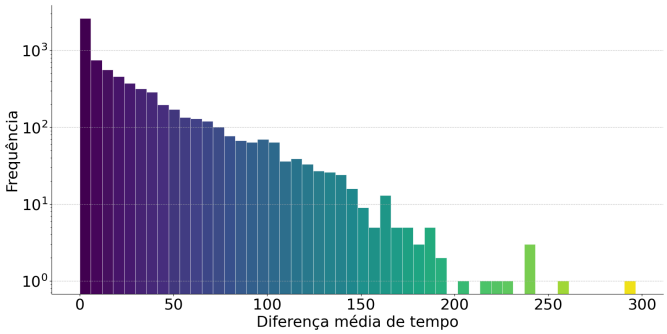


Fig. 4. Distribuição da diferença média de tempo

**Análise:** A Fig. 4 revela uma distribuição com forte assimetria positiva, onde a grande maioria das observações se concentra próximo de zero, indicando condições de fluxo livre. A utilização de uma escala logarítmica no eixo da frequência permite visualizar a "cauda longa" que se estende até aos 300 segundos, representando episódios de congestionamento severo. Pela sua natureza, esta variável é o principal indicador físico dos atrasos que definem as categorias da variável dependente.

### LUMINOSITY

**Descrição:** O nível de luminosidade.
**Tipo:** Categórica ordinal.
**Valores nulos:** 0 (zero).
**Valores únicos:** 3 (três).

TABLE III
QUANTIFICAÇÃO LUMINOSITY

| Categoria | Frequência | Percentagem |
|---|---|---|
| LIGHT | 3293 | 48,3% |
| DARK | 3253 | 47,8% |
| LOW_LIGHT | 266 | 3,9% |

**Análise:** Observando a distribuição da variável, é visível que o valor único 'LOW_LIGHT' tem baixa representação, o que poderá levar a uma dificuldade na previsão onde este caso está presente.

**AVERAGE_TEMPERATURE**

**Descrição:** Valor médio da temperatura.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

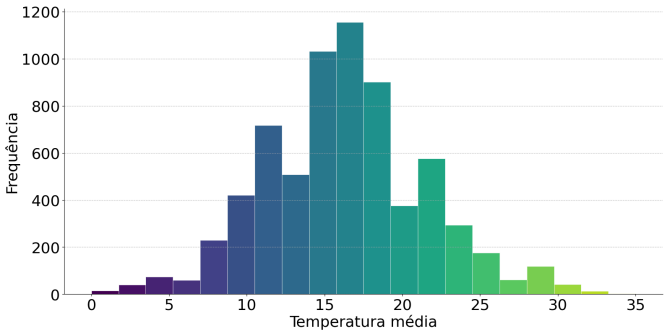


Fig. 5. Distribuição da temperatura média

**Análise:** A Fig. 5 apresenta uma distribuição aproximadamente normal, com maior concentração de registos por volta dos 16°C. A amplitude térmica observada é plenamente consistente com o clima temperado da cidade do Porto ao longo do ano.

**AVERAGE_ATMOSP_PRESSURE**

**Descrição:** Valor médio da pressão atmosférica.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

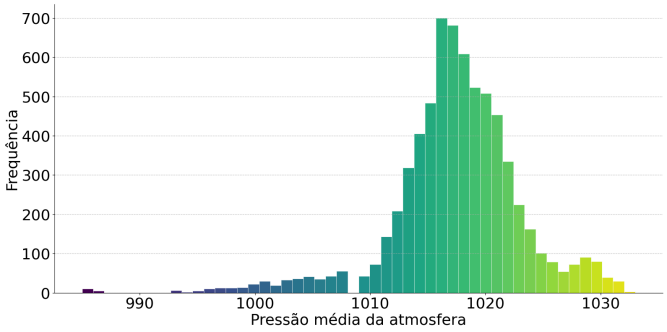


Fig. 6. Distribuição da pressão média da atmosfera

**Análise:** A Fig. 6 caracteriza-se como uma distribuição *leptocurtic* e média de 1017,4 hPa. A elevada concentração em torno da mediana sugere que a pressão atua como um indicador de estabilidade ambiental, onde apenas as quedas bruscas podem sinalizar eventos climatéricos com impacto indireto na condução.

**AVERAGE_HUMIDITY**

**Descrição:** Valor médio de humidade.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

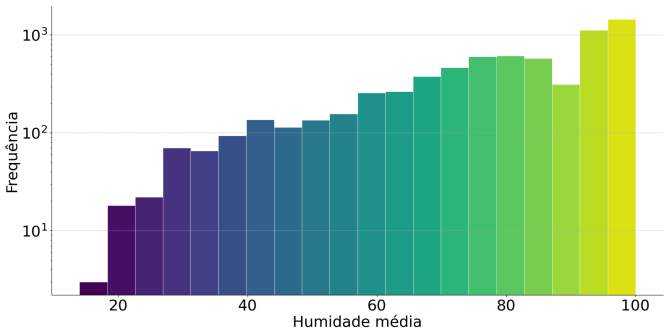


Fig. 7. Distribuição da humidade média

**Análise:** A Fig. 7 revela um acentuado enviesamento para altas concentrações. Para a modelação, esta variável poderá atuar como um indicador de condições de saturação, embora a sua reduzida variância na maioria dos registos sugira um papel de suporte em vez de um preditor linear primário.

**AVERAGE_WIND_SPEED**

**Descrição:** Valor médio da velocidade do vento.
**Tipo:** Numérica contínua.
**Valores nulos:** 0 (zero).

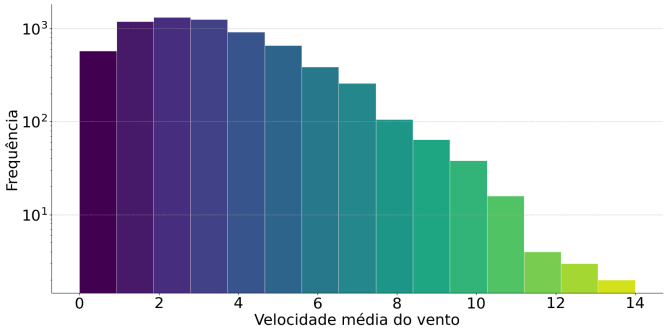


Fig. 8. Distribuição da velocidade média do vento

**Análise:** A Fig. 8 exibe uma distribuição com forte assimetria positiva, concentrando a massa de dados em velocidades reduzidas. Estatisticamente, esta variável atua como um indicador de condições climatéricas adversas esporádicas.

**AVERAGE_CLOUDINESS**

**Descrição:** O valor médio da percentagem de nuvens.
**Tipo:** Categórica ordinal.
**Valores únicos:** 9 (nove).

TABLE IV
QUANTIFICAÇÃO AVERAGE_CLOUDINESS

| Categoria | Frequência | Percentagem |
|---|---|---|
| céu limpo | 153 | 3.70% |
| céu claro | 1582 | 38.31% |
| algumas nuvens | 422 | 10.22% |
| nuvens dispersas | 459 | 11.11% |
| nuvens quebradas | 416 | 10.07% |
| nuvens quebrados | 448 | 10.85% |
| céu pouco nublado | 516 | 12.49% |
| tempo nublado | 67 | 1.62% |
| nublado | 67 | 1.62% |

**Valores nulos:** 2682 (dois mil seiscentos e oitenta e dois).

TABLE V
AVERAGE_CLOUDINESS EM BRUTO

| ID | AVERAGE_CLOUDINESS |
|---|---|
| 15 | nuvens quebrados |
| 16 | |
| 17 | NULL |
| 18 | nuvens quebrados |
| 19 | algumas nuvens |

**Análise:** A tabela de dados em bruto revela que estes valores em falta não são uniformes, manifestando-se de duas formas distintas:

- String vazia
- NULL

Além deste problema de dados em falta, identifica-se a necessidade de limpeza nas categorias, que por vezes são redundantes.

**`AVERAGE_PRECIPITATION`**

**Descrição:** Valor médio de precipitação.
**Tipo:** Numérica contínua.
**Valores únicos:** 1 (um).
**Valores nulos:** 0 (zero).
**Análise:** Sendo um valor constante, esta *feature* não tem poder preditivo e pode ser removida.

**`AVERAGE_RAIN`**

**Descrição:** Avaliação qualitativa do nível de precipitação.
**Tipo:** Categórica ordinal.
**Valores únicos:** 13 (treze)

TABLE VI
QUANTIFICAÇÃO AVERAGE_RAIN

| Categoria | Frequência | Percentagem |
|---|---|---|
| chuva fraca | 261 | 46.36% |
| chuva moderada | 153 | 27.17% |
| chuva leve | 45 | 7.99% |
| aguaceiros fracos | 38 | 6.74% |
| chuva | 30 | 5.32% |
| aguaceiros | 11 | 1.95% |
| chuva forte | 8 | 1.42% |
| trovoada com chuva leve | 7 | 1.24% |
| chuvisco fraco | 5 | 0.88% |
| chuva de intensidade pesado | 2 | 0.35% |
| chuva de intensidade pesada | 1 | 0.17% |
| trovoada com chuva | 1 | 0.17% |
| chuvisco e chuva fraca | 1 | 0.17% |

**Valores nulos:** 6249 (seis mil duzentos e quarenta e nove)

TABLE VII
AVERAGE_RAIN EM BRUTO

| ID | AVERAGE_RAIN |
|---|---|
| 11 | NULL |
| 12 | |
| 13 | NULL |
| 14 | chuva fraca |
| 15 | |

**Análise:** A tabela de dados em bruto revela que estes valores em falta não são uniformes, manifestando-se novamente de duas formas distintas:

- String vazia
- NULL

Além deste problema de dados em falta, não temos uma categoria com ausência de chuva e também se identifica categorias redundantes.

### *C. Análise Multivariada*

Nesta etapa, a análise foca-se na interação entre as múltiplas variáveis do conjunto de dados, procurando identificar redundâncias, dependências não lineares e, fundamentalmente, determinar o poder preditivo dos atributos em relação à variável alvo. Esta análise é crucial para validar se as variáveis independentes oferecem separabilidade suficiente para os algoritmos de classificação.

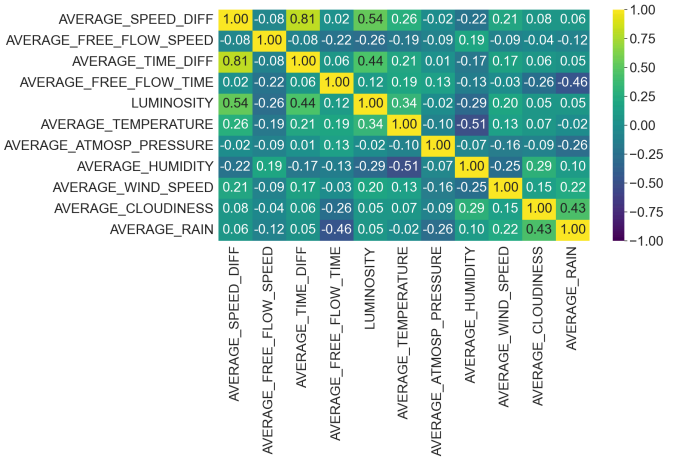


Fig. 9. Matriz de correlação linear entre variáveis numéricas

A matriz de correlação permite quantificar a força das relações lineares. O destaque imediato é a correlação positiva muito forte 0.81 entre **AVERAGE_TIME_DIFF** e **AVERAGE_SPEED_DIFF**. Do ponto de vista físico, este resultado é intuitivo: o atraso temporal acumulado num segmento rodoviário é o reflexo direto da redução da velocidade média. No entanto, o facto de a correlação não ser perfeita indica que a classificação final não depende apenas de um limiar fixo de tempo, mas sofre influência da interação com outras variáveis.

Observa-se ainda uma multicolinearidade moderada entre variáveis ambientais, como a relação inversa entre **AVERAGE_TEMPERATURE** e **AVERAGE_HUMIDITY**, −0.51. Este fenómeno sugere que, para modelos sensíveis à correlação entre preditores, a inclusão de ambos os atributos pode ser redundante. Por outro lado, a correlação de 0.54 entre a **LUMINOSITY** e **AVERAGE_SPEED_DIFF** reforça que a visibilidade e o ciclo horário são determinantes no comportamento dos condutores.

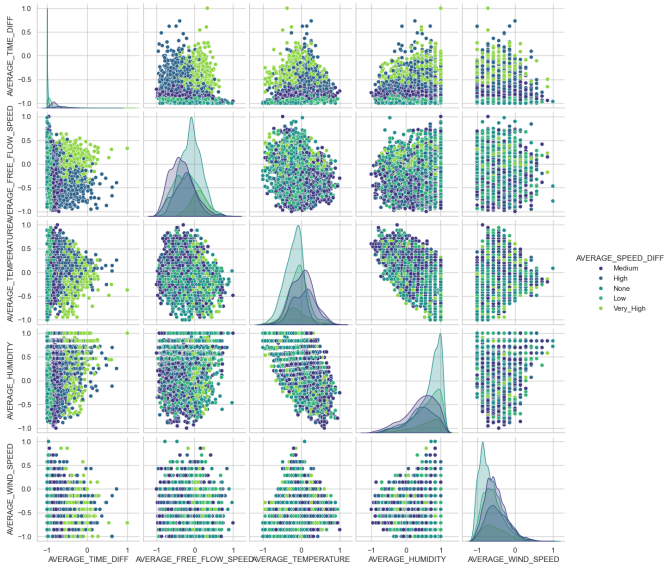


Fig. 10. Distribuição conjunta e separação de classes

A Fig. 10 confirma visualmente a supremacia da **AVERAGE_TIME_DIFF**. Enquanto os atributos meteorológicos resultam em "nuvens" de pontos sobrepostas, a diferença de tempo permite uma separação clara das aglomerações por classe.

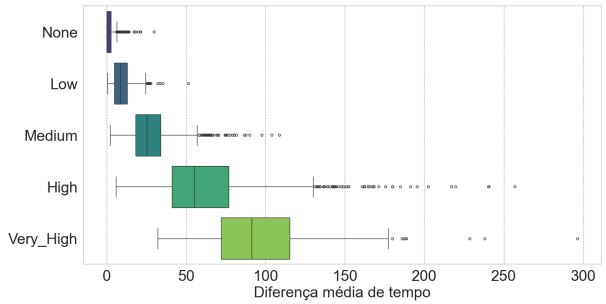


Fig. 11. Relação ordinal entre o atraso temporal e a gravidade do tráfego

O rigor desta relação é detalhado na Fig. 11. Observa-se uma progressão estritamente crescente das medianas e da amplitude interquartil à medida que a classe de tráfego se agrava.

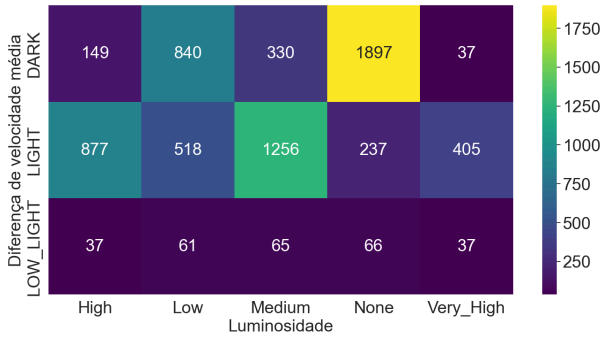


Fig. 12. Influência da luminosidade no estado do tráfego

Por fim, a Fig. 12 permite inferir padrões de comportamento humano. A predominância da classe "None" durante o período "DARK" é consistente com o fluxo livre esperado durante a noite e madrugada. Inversamente, as condições de "Medium" e "High" tráfego concentram-se esmagadoramente no período "LIGHT", coincidindo com o horário laboral e escolar.

## IV. Preparação dos Dados

O objetivo primordial desta fase é maximizar o potencial preditivo dos dados disponíveis, garantindo a sua consistência e relevância. Mais do que uma etapa técnica de limpeza e correção de erros, trata-se de um processo de refinamento que impacta diretamente a fiabilidade dos resultados finais. Através de técnicas de *feature engineering* e seleção criteriosa de variáveis, procura-se entregar aos algoritmos um conjunto de dados otimizado, reduzindo o risco de *overfitting* e assegurando uma utilização eficiente dos recursos do projeto.

### A. Remoção de atributos

Numa fase inicial, procedeu-se à redução de dimensionalidade e na garantia da integridade da variável alvo. Na

etapa anterior, verificou-se que os atributos **city_name** e **AVERAGE_PRECIPITATION** possuíam variância nula, contendo um valor constante para todas as observações. A ausência de variabilidade implica ausência de informação discriminativa para o modelo, estas variáveis foram removidas do conjunto de dados.

### *B. Codificação de variáveis categóricas*

A maioria dos algoritmos de *Machine Learning* requer que os dados de entrada sejam estritamente numéricos. Como identificado na análise exploratória, diversas variáveis, incluindo a variável alvo, apresentam-se em formato de texto.

Dada a natureza ordinal destas variáveis — onde existe uma hierarquia clara entre as classes — optou-se pela técnica de *Label Encoding* em detrimento do *One-Hot Encoding*. Esta abordagem preserva a relação de ordem intrínseca dos dados e evita o aumento desnecessário da dimensionalidade do *dataset*.

Procedeu-se aos seguintes mapeamentos numéricos:

- **AVERAGE_SPEED_DIFF:** A variável foi codificada numa escala de 0 a 4, refletindo a gravidade do congestionamento.
  - None → 0
  - Low → 1
  - Medium → 2
  - High → 3
  - Very_High → 4
- **LUMINOSITY:** Codificada de forma a refletir a intensidade da luz, facilitando a correlação positiva com a visibilidade.
  - DARK → 0
  - LOW_LIGHT → 1
  - LIGHT → 2

As restantes variáveis categóricas, nomeadamente **AVERAGE_CLOUDINESS** e **AVERAGE_RAIN**, exigiram tratamentos mais complexos de limpeza e imputação antes da sua codificação, processos que são detalhados na subsecção de tratamento de valores em falta.

### *C. Feature engineering*

Reconhecendo a forte componente temporal do tráfego, a variável **record_date** foi decomposta para capturar padrões sazonais e diários em múltiplas escalas. Foram extraídos os seguintes atributos temporais:

- **Ano (year) e mês (month):** Para capturar tendências de longo prazo e sazonalidade anual.
- **Semana do ano (week) e dia do ano (dayofyear):** Para identificar padrões infra-anuais mais granulares.
- **Dia da semana (dayofweek):** Para distinguir padrões entre dias úteis e fins de semana.
- **Hora (hour):** Para capturar a variação diária do tráfego.

Contudo, a representação linear da hora é problemática para algoritmos de regressão e redes neuronais, porque ignora a continuidade temporal entre as 23h00 e as 00h00. Para resolver isto, a hora foi projetada num espaço circular através de transformações trigonométricas [1], criando duas novas *features*:

$$\text{hour\_sin} = \sin\left(\frac{2\pi \cdot \text{hour}}{24}\right) \tag{1}$$

$$\text{hour\_cos} = \cos\left(\frac{2\pi \cdot \text{hour}}{24}\right) \tag{2}$$

Esta transformação garante que a proximidade temporal é preservada matematicamente. Após a extração destas características, a variável original **record_date** e **hour** foram removidas do conjunto de dados.

### *D. Tratamento de valores em falta*

A gestão de valores em falta é uma etapa crítica no pré-processamento, uma vez que a simples remoção de registos incompletos pode introduzir viés de seleção e reduzir significativamente o tamanho da amostra. A análise exploratória identificou três cenários distintos que exigiram abordagens específicas:

a) Correção de *parsing:* O **AVERAGE_SPEED_DIFF** apresentava valores nulos que, após inspeção cruzada com os dados originais, foram identificados como erros de leitura. A categoria textual "None", indicativa de fluxo normal, foi incorretamente convertida para valor nulo durante a ingestão dos dados. Procedeu-se à reversão destes valores para a categoria "None". A exclusão destes registos teria sido prejudicial ao modelo, pois eliminaria 32.3% dos dados (Tabela I), removendo especificamente os exemplos que representam condições normais de trânsito.

b) Tratamento do **AVERAGE_RAIN***:* Esta variável apresentava três desafios, a elevada taxa de valores nulos, a ausência de uma categoria explícita para "sem chuva" e redundância semântica nas suas 13 categorias. A estratégia adotada consistiu em:

1) **Inferência:** A análise de padrões nos ficheiros em bruto (**.csv**) permitiu deduzir que os valores `NULL` correspondiam, na sua maioria, a ausência de registo pluviométrico. Assim, estes valores foram convertidos para a categoria "sem chuva".
2) ***Binning*:** Devido ao severo desbalanceamento de classes e à redundância entre termos, optou-se por simplificar a granularidade da variável. As categorias foram agrupadas numa variável binária: "Sem Chuva" (0) e "Com Chuva" (1). Esta redução de dimensionalidade foca-se no impacto macroscópico da precipitação, ignorando nuances de intensidade que, dada a elevada taxa de valores em falta, poderiam introduzir ruído e incerteza adicionais no modelo.
3) **Imputação:** Para o preenchimento dos valores em falta, foi realizado um estudo comparativo envolvendo a imputação pela moda estatística, previsão através de algoritmos de *Machine Learning* e propagação temporal. Embora tecnicamente viável, a utilização de modelos de *Machine Learning* para a imputação foi descartada após análise dos resultados preliminares, pois verificou-se que estes introduziam um viés (bias) considerável nos dados, distorcendo a distribuição original. A moda estatística, por sua vez, revelou-se demasiado simplista. Consequentemente, optou-se pelo método de *Forward*

*Fill*. Esta técnica, ao propagar o último estado conhecido, respeita a continuidade temporal dos fenómenos meteorológicos sem impor o viés artificial dos modelos preditivos.

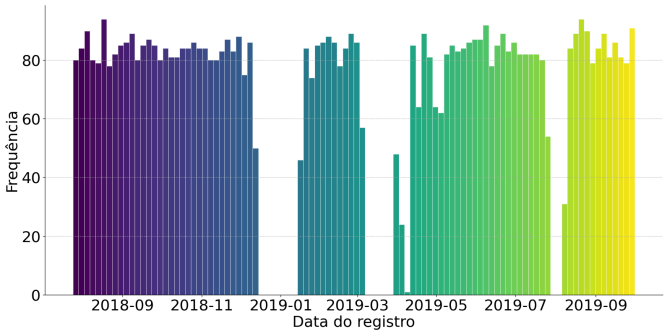


Fig. 13. Frequência temporal

Embora a descontinuidade temporal evidenciada na Fig. 13 comprometa a integridade sequencial, a adoção do *Forward Fill* justifica-se como uma aproximação conservadora, pressupondo a persistência do estado meteorológico durante os hiatos de registo.

c) Tratamento do **`AVERAGE_CLOUDINESS`***:* Diferente da precipitação, a ausência de dados na nebulosidade representa uma lacuna real de informação.

1) ***Binning*:** As 9 categorias originais apresentavam redundância semântica e inconsistências de formatação. Para corrigir a dispersão, as categorias foram agrupadas em 3 classes ordinais representativas da densidade da cobertura, "Céu limpo" (0), "Céu parcialmente limpo" (1) e "Céu nublado" (2). Esta simplificação permitiu eliminar o ruído provocado por distinções subtis entre termos sinônimos.
2) **Imputação:** Seguindo a metodologia aplicada no **`AVERAGE_RAIN`**, foram testadas abordagens baseadas em *Machine Learning* e na moda. Novamente, a imputação via ML foi rejeitada devido ao viés introduzido nas previsões, que comprometia a integridade estatística da variável. A escolha recaiu sobre o *Forward Fill*, validada pela natureza gradual da evolução da cobertura de nuvens. Esta abordagem provou ser a mais robusta, preservando a coerência sequencial dos dados meteorológicos e evitando as distorções observadas com os outros métodos.

### *E. Normalização e tratamento de outliers*

Relativamente aos *outliers* detetados na etapa anterior, nomeadamente em **`AVERAGE_TIME_DIFF`**, optou-se pela sua manutenção. Em contexto de tráfego, estes valores extremos representam frequentemente eventos reais de congestionamento severo ou acidentes, constituindo informação crítica que o modelo deve ser capaz de prever, e não ruído a eliminar.

A análise multivariada revelou que as variáveis possuem escalas muito distintas. Para garantir a convergência eficiente de algoritmos baseados em gradiente e distâncias, aplicou-se o *Min-Max Scaling*, normalizando todas as variáveis numéricas para o intervalo $[0, 1]$.

### *F. Balanceamento dos dados*

A análise da distribuição do **`AVERAGE_SPEED_DIFF`** (Tabela I) evidenciou um desequilíbrio severo, onde a classe maioritária possui cerca de 4,5 vezes mais registos que a classe minoritária. Sem intervenção, os modelos de aprendizagem tendem a enviesar as previsões para as classes frequentes.

Para mitigar este problema, adotou-se uma estratégia combinada de validação e *oversampling*:

- **Validação *Stratified K-Fold*:** Esta técnica assegura que cada subconjunto de dados mantém a proporção original das classes. Isto foi fundamental para garantir que as métricas de avaliação fossem fiáveis e representativas de todas as categorias de tráfego [2].
- ***Oversampling* com SMOTE:** O SMOTE (*Synthetic Minority Over-sampling Technique*), ao contrário da simples duplicação de registos, gera novos exemplos sintéticos nos *folds* de treino através da interpolação linear entre vizinhos próximos da classe minoritária. Deste modo, é possível lidar com o desequilíbrio e aumentar a representatividade destas classes, conduzindo a uma melhoria no desempenho dos modelos preditivos e à redução do viés [3].

### *G. Seleção de atributos*

A elevada dimensionalidade impõe desafios significativos, nomeadamente o aumento do custo computacional e o risco de *overfitting*. Para identificar o subconjunto de variáveis que maximiza a performance preditiva sem introduzir ruído, realizou-se um estudo comparativo utilizando três abordagens distintas de seleção e redução de dimensionalidade:

a) Análise de componentes principais*: Principal Component Analysis* (PCA) é uma técnica não-supervisionada para reduzir a dimensionalidade através da projeção dos dados em componentes ortogonais. Embora tenha sido eficaz na compressão da variância e na eliminação de multicolinearidade, resultou na perda de interpretabilidade semântica das variáveis, dificultando a compreensão física dos fatores que causam o tráfego.

b) Otimização via *Particle Swarm Optimization:* Implementou-se uma abordagem meta-heurística utilizando o algoritmo PSO adaptado para seleção de *features*. Neste método, cada "partícula" representou um subconjunto candidato de atributos, e a *fitness function* foi o desempenho do modelo. O PSO permitiu explorar o espaço de combinações de forma eficiente, procurando ótimos globais [4].

Contudo, a viabilidade desta abordagem foi severamente comprometida pelo custo computacional, particularmente quando o algoritmo foi acoplado a uma estratégia de *Grid Search* para a otimização de hiperparâmetros. Como cada avaliação da função de *fitness* de uma partícula exigia o treino e validação cruzada de um modelo, o processo resultou numa explosão combinatória do tempo de execução, tornando a pesquisa inviável para iterações rápidas de desenvolvimento.

c) Eliminação Iterativa com *Feature Importance:*

Por fim, implementou-se um procedimento manual de eliminação iterativa de variáveis (*Manual Backward Elimination*). Esta abordagem consistiu nos seguintes passos:

1) Treinar o modelo com o conjunto de variáveis atual.
2) Calcular a importância global de cada atributo através da *Feature Importance*, utilizando *Mean Decrease in Impurity* (para modelos baseados em árvores) e *Permutation Importance* (para modelos agnósticos ou lineares).
3) Identificar e remover a variável com a menor contribuição relativa.
4) Repetir o ciclo, reavaliando a performance do modelo a cada iteração, até que a remoção de variáveis começasse a degradar significativamente o desempenho.

Esta última abordagem revelou-se a mais robusta, permitindo não só reduzir a complexidade do modelo, mas também validar as hipóteses de negócio, mostrando quais são os preditores dominantes.

## V. Modelação

Nesta secção, descrevem-se as arquiteturas dos algoritmos selecionados e as estratégias de treino adotadas. A aplicação de *Grid Search* para a otimização de hiperparâmetros foi utilizada nos modelos de referência e métodos de *ensemble*, assegurando a escolha da melhor configuração dentro do espaço de pesquisa definido; contudo, esta técnica não foi aplicada às arquiteturas de *Deep Learning* e aos modelos especialistas. A abordagem seguiu uma complexidade incremental: iniciou-se com modelos de referência interpretáveis, progrediu-se para métodos de *ensemble* e as referidas arquiteturas avançadas, e culminou-se numa estratégia de *Stacking* para a integração final das previsões.

### A. Decision Tree

Como ponto de partida, utilizou-se uma árvore de decisão que, pela sua natureza, permite uma interpretação direta das regras inferidas a partir das *features* de tráfego. Embora propensas a *overfitting*, as árvores servem como *baseline* para aferir o ganho de performance obtido com modelos mais complexos.

Na fase de configuração, e em consonância com o método iterativo de seleção de atributos, o treino foi restringido às variáveis com maior poder discriminante. Consequentemente, foram removidas do conjunto de dados as colunas **`year`**, **`LUMINOSITY`**, **`AVERAGE_RAIN`** e **`AVERAGE_CLOUDINESS`**. A configuração final obtida foi:

- `criterion`: Gini;
- `min_samples_split`: 3;
- `min_samples_leaf`: 1;
- `max_depth`: 7.

TABLE VIII
Classification Report Decision Tree

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.91 | 0.87 | 0.89 |
| Low | 0.62 | 0.71 | 0.66 |
| Medium | 0.79 | 0.70 | 0.74 |
| High | 0.73 | 0.78 | 0.76 |
| Very_High | 0.77 | 0.82 | 0.79 |
| *accuracy* | | | 0.77 |
| *macro avg* | 0.76 | 0.77 | 0.77 |
| *weighted avg* | 0.78 | 0.77 | 0.78 |

### B. Logistic Regression

Como complemento à abordagem baseada em árvores, implementou-se a *Logistic Regression* para estabelecer uma *baseline* linear. O objetivo principal foi verificar o comportamento de um modelo paramétrico mais simples e interpretável perante a complexidade destes dados.

No que concerne à configuração, e em linha com a estratégia de redução de dimensionalidade, procedeu-se à exclusão dos atributos com menor relevância preditiva ou redundância temporal, nomeadamente: **`AVERAGE_CLOUDINESS`**, **`week`**, **`year`**, **`AVERAGE_RAIN`**, **`AVERAGE_ATMOSP_PRESSURE`** e **`dayofyear`**.

A configuração final selecionada privilegiou a penalização L1 (Lasso) para promover a esparsidade do modelo, exigindo a utilização do otimizador `saga` para garantir a convergência:

- `C`: 1;
- `penalty`: L1;
- `solver`: saga;
- `max_iter`: 2000.

TABLE IX
Classification Report Logistic Regression

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.88 | 0.90 | 0.89 |
| Low | 0.62 | 0.70 | 0.66 |
| Medium | 0.81 | 0.72 | 0.76 |
| High | 0.77 | 0.70 | 0.73 |
| Very_High | 0.68 | 0.82 | 0.74 |
| *accuracy* | | | 0.77 |
| *macro avg* | 0.75 | 0.76 | 0.76 |
| *weighted avg* | 0.78 | 0.77 | 0.77 |

### C. Support Vector Machine

Dando continuidade à procura de fronteiras de decisão mais complexas, implementou-se o algoritmo *Support Vector Machine* (SVM). A utilização deste modelo visou superar as limitações lineares da Regressão Logística, projetando os dados num espaço dimensional superior através do *Kernel Trick*, permitindo a separação de classes não-linearmente separáveis no espaço original.

Na fase de pré-processamento específica, identificou-se que a variável **`AVERAGE_WIND_SPEED`** introduzia ruído sem

contribuir significativamente para a discriminação das classes, tendo sido consequentemente removida. A otimização de hiperparâmetros revelou que um *kernel* polinomial se ajusta melhor à distribuição dos dados do que as abordagens linear. A configuração final obtida foi:

- `C`: 10;
- `kernel`: poly;
- `gamma`: scale.

TABLE X
CLASSIFICATION REPORT SVM

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.88 | 0.89 | 0.89 |
| Low | 0.66 | 0.70 | 0.68 |
| Medium | 0.79 | 0.76 | 0.78 |
| High | 0.75 | 0.71 | 0.73 |
| Very_High | 0.76 | 0.80 | 0.78 |
| *accuracy* | | | 0.79 |
| *macro avg* | 0.77 | 0.77 | 0.77 |
| *weighted avg* | 0.79 | 0.79 | 0.79 |

### D. Random Forest

Para mitigar a tendência de *overfitting* e a elevada variância características de uma árvore de decisão isolada, avançou-se para a implementação do *Random Forest*. Este método de *ensemble*, fundamentado na técnica de *Bagging* (*Bootstrap Aggregating*), constrói múltiplas árvores de decisão independentes durante a fase de treino e agrega as suas previsões, resultando num modelo globalmente mais robusto e estável face ao ruído nos dados.

No que respeita à configuração, e mantendo a coerência com a análise de importância de variáveis, o treino foi restringido aos atributos com maior poder discriminante, excluindo-se **`AVERAGE_RAIN`** e **`AVERAGE_CLOUDINESS`**. A otimização de hiperparâmetros conduziu à seguinte configuração final:

- `criterion`: entropy;
- `n_estimators`: 120;
- `max_depth`: 18;
- `min_samples_split`: 3;
- `min_samples_leaf`: 1.

TABLE XI
CLASSIFICATION REPORT RANDOM FOREST

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.90 | 0.89 | 0.90 |
| Low | 0.68 | 0.77 | 0.72 |
| Medium | 0.83 | 0.78 | 0.81 |
| High | 0.80 | 0.78 | 0.79 |
| Very_High | 0.85 | 0.79 | 0.82 |
| *accuracy* | | | 0.81 |
| *macro avg* | 0.81 | 0.80 | 0.81 |
| *weighted avg* | 0.82 | 0.81 | 0.81 |

### E. Gradient Boosting

Explorando uma abordagem sequencial, aplicou-se o algoritmo *Gradient Boosting*. Diferenciando-se do crescimento paralelo e independente do *Random Forest*, este método adota uma estratégia aditiva onde cada novo estimador é treinado para corrigir os erros residuais do modelo anterior. Esta técnica permite minimizar progressivamente a função de perda através do método *gradient descent*, refinando a capacidade preditiva nos casos mais complexos.

Na fase de configuração, optou-se por simplificar o espaço de características através da remoção das variáveis **`AVERAGE_RAIN`**, **`AVERAGE_CLOUDINESS`**, **`year`** e **`LUMINOSITY`**. Relativamente à otimização de hiperparâmetros, a pesquisa destacou a importância da densidade do *ensemble*, resultando na seguinte configuração final:

- `n_estimators`: 150

TABLE XII
CLASSIFICATION REPORT GRADIENT BOOSTING

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.91 | 0.90 | 0.90 |
| Low | 0.68 | 0.77 | 0.72 |
| Medium | 0.84 | 0.77 | 0.80 |
| High | 0.78 | 0.77 | 0.78 |
| Very_High | 0.78 | 0.80 | 0.79 |
| *accuracy* | | | 0.81 |
| *macro avg* | 0.80 | 0.80 | 0.80 |
| *weighted avg* | 0.82 | 0.81 | 0.81 |

### F. XGBoost

Como evolução algorítmica, implementou-se o *XGBoost* (*Extreme Gradient Boosting*). A principal inovação deste modelo reside na introdução de termos de regularização (L1 e L2) diretamente na função objetivo, penalizando a complexidade dos estimadores para evitar *overfitting*. Esta característica, aliada a uma execução otimizada para eficiência computacional e dados esparsos, permite um controlo superior do equilíbrio entre viés e variância em comparação com o *Gradient Boosting* tradicional.

Na fase de configuração, optou-se por uma estratégia conservadora na seleção de atributos, removendo apenas a variável **`AVERAGE_WIND_SPEED`** por demonstrar um contributo marginal para a redução da função de perda. Relativamente aos hiperparâmetros, a otimização identificou a seguinte configuração ideal:

- `max_depth`: 6;
- `n_estimators`: 120;
- `colsample_bytree`: 0.8.

TABLE XIII
CLASSIFICATION REPORT XGBOOST

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.91 | 0.89 | 0.90 |
| Low | 0.68 | 0.77 | 0.72 |
| Medium | 0.84 | 0.78 | 0.81 |
| High | 0.78 | 0.76 | 0.78 |
| Very_High | 0.80 | 0.80 | 0.80 |
| *accuracy* | | | 0.81 |
| *macro avg* | 0.80 | 0.80 | 0.80 |
| *weighted avg* | 0.82 | 0.81 | 0.81 |

TABLE XIV
CLASSIFICATION REPORT STACKING

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.91 | 0.90 | 0.91 |
| Low | 0.70 | 0.75 | 0.73 |
| Medium | 0.82 | 0.81 | 0.82 |
| High | 0.80 | 0.77 | 0.78 |
| Very_High | 0.83 | 0.79 | 0.81 |
| *accuracy* | | | 0.82 |
| *macro avg* | 0.81 | 0.81 | 0.81 |
| *weighted avg* | 0.82 | 0.82 | 0.82 |

### G. Stacking

Numa ótica de meta-aprendizagem, implementou-se a técnica de *Stacking*. A motivação para esta escolha reside na capacidade desta arquitetura de aprender os padrões de erro dos modelos individuais. Ao contrário de uma simples votação, o "meta-modelo" identifica em que cenários cada algoritmo base é mais fiável, corrigindo as suas fraquezas através das forças dos outros estimadores.

Para maximizar a diversidade de opiniões, selecionaram-se cinco algoritmos distintos para o nível base. As suas configurações foram ajustadas especificamente para este *ensemble*, permitindo uma maior complexidade individual, delegando a regularização final para o meta-modelo:

- ***Random Forest*:**
  - `n_estimators`: 200;
  - `max_depth`: 30.
- ***Logistic Regression*:**
  - `solver`: saga;
  - `penalty`: L1.
- **SVM:**
  - `kernel`: poly.
- ***XGBoost*:**
  - `n_estimators`: 200;
  - `max_depth`: 30;
- ***Gradient Boosting*:**
  - `n_estimators`: 200;
  - `max_depth`: 30.

Como "meta-modelo", utilizou-se uma *Logistic Regression*, responsável por ponderar linearmente as probabilidades geradas pelos modelos base.

Na construção deste *ensemble*, procedeu-se à remoção das variáveis **`AVERAGE_CLOUDINESS`**, **`AVERAGE_RAIN`**, **`month`** e **`AVERAGE_WIND_SPEED`** para reduzir o ruído na agregação.

### H. Artificial Neural Networks

Para modelar as relações não-lineares complexas inerentes aos dados de tráfego, implementou-se uma rede neuronal profunda com uma arquitetura *Feedforward*.

A arquitetura da rede foi desenhada para maximizar a capacidade de generalização e mitigar o *overfitting*, integrando múltiplas camadas de regularização:

- **Camada de Entrada:** Ajustada à dimensionalidade dos dados após o pré-processamento.
- **Camadas Ocultas:** A rede é composta por três camadas densas sequenciais:
  - **Primeira camada:** 128 neurónios e ativação *ReLU*, com regularização L2.
  - **Segunda camada:** 64 neurónios e ativação *ReLU*, com regularização L2.
  - **Terceira camada:** 32 neurónios e ativação *ReLU*.

Entre as camadas densas, foram introduzidas camadas de *BatchNormalization* para estabilizar a aprendizagem, seguidas por camadas de *Dropout* para prevenir a coadaptação dos neurónios.

- **Camada de Saída:** Composta por 5 neurónios com ativação *Softmax*, gerando a distribuição de probabilidades para as classes.

O treino foi realizado com o otimizador *Adam* (taxa de aprendizagem inicial de 0.001) e função de perda *Categorical Crossentropy*. Para otimizar o processo de aprendizagem, implementaram-se dois *callbacks* essenciais, *EarlyStopping*, com uma paciência de 20 épocas para interromper o treino caso a perda de validação estagne, e *ReduceLROnPlateau*, que reduz a taxa de aprendizagem em 50% após 3 épocas sem melhoria, permitindo um ajuste dos pesos nos mínimos locais.

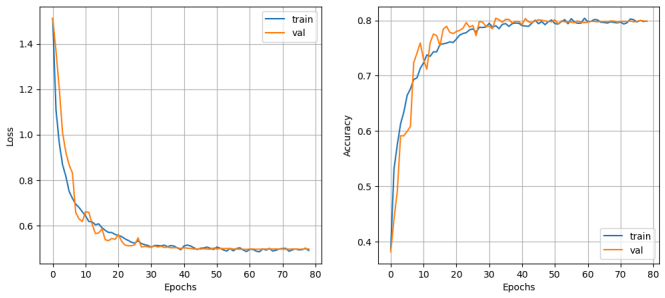


Fig. 14. Convergência das curvas de treino e validação

A Fig. 14 ilustra a evolução das métricas durante o treino. Observa-se a convergência das curvas de treino e validação, indicando que as técnicas de regularização foram eficazes em controlar o *overfitting*, mantendo a capacidade de generalização do modelo.

TABLE XV
CLASSIFICATION REPORT ARTIFICIAL NEURAL NETWORKS

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.87 | 0.93 | 0.90 |
| Low | 0.69 | 0.68 | 0.69 |
| Medium | 0.84 | 0.74 | 0.79 |
| High | 0.75 | 0.75 | 0.75 |
| Very_High | 0.72 | 0.86 | 0.79 |
| *accuracy* | | | 0.80 |
| *macro avg* | 0.77 | 0.79 | 0.78 |
| *weighted avg* | 0.80 | 0.80 | 0.80 |

### *I. Agentes Especialistas e Mixture of Experts*

Como alternativa às abordagens de ensemble apresentadas anteriormente (como o *Random Forest* ou *XGBoost*), exploramos uma estratégia de modularização do problema recorrendo a 'especialistas'. O objetivo destas arquiteturas foi mitigar o enviesamento provocado pelo desequilíbrio de classes já abordado, forçando o sistema a se dedicar a cada categoria de tráfego, em vez de tentar otimizar uma única função de perda global.

a) Estratégia *One-vs-Rest* com Reamostragem*:* Nesta primeira abordagem, o problema *multiclass* foi decomposto em cinco subproblemas binários (estratégia *One-vs-Rest*). Para cada classe de tráfego, foi treinado um "Agente Especialista" — um classificador *Random Forest* independente — com a tarefa exclusiva de distinguir se uma observação pertence à sua classe específica ou a qualquer outra. Foram utilizados os hiperparâmetros obtidos anteriormente para a *Random Forest* e as previsões foram realizadas através do grau de certeza dos modelos especialistas (a sua percentagem).

Nesta implementação foi aplicada também a técnica de *SMOTE*, assim como testadas alternativas como o *SMOTE-ENN* e *ADASYN*. Ao contrário do treino global, onde o *oversampling* tenta equilibrar todas as classes simultaneamente (o que pode gerar ruído excessivo nas fronteiras de decisão), nesta arquitetura o *SMOTE* foi aplicado individualmente dentro de cada caso binário. Isto permitiu que o especialista da classe "Very_High", por exemplo, fosse treinado com um conjunto de dados onde essa classe minoritária tinha um peso estatístico igual à soma de todas as outras classes.

A decisão final é obtida por competição direta: as probabilidades geradas por todos os especialistas são comparadas, selecionando-se a classe cujo agente demonstra o maior grau de confiança (*argmax*).

TABLE XVI
CLASSIFICATION REPORT ONE VS REST

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.87 | 0.91 | 0.89 |
| Low | 0.70 | 0.70 | 0.70 |
| Medium | 0.79 | 0.76 | 0.77 |
| High | 0.76 | 0.78 | 0.77 |
| Very_High | 0.86 | 0.80 | 0.83 |
| *accuracy* | | | 0.80 |
| *macro avg* | 0.80 | 0.79 | 0.79 |
| *weighted avg* | 0.80 | 0.80 | 0.80 |

b) Arquitetura *Mixture of Experts* (MoE)*:* Evoluindo a estratégia de especialistas independentes, implementou-se uma arquitetura de *Mixture of Experts* (MoE). Enquanto o *Stacking* utiliza um meta-modelo estático para combinar previsões, o MoE baseia-se no conceito de computação condicional. A arquitetura desenvolvida compõe-se por dois elementos fundamentais:

- ***Experts*:** são um conjunto de classificadores treinados em diferentes subconjuntos de dados rebalanceados, que desenvolvem competências locais em regiões específicas do espaço de características.
- ***Gating Network*:** é um modelo supervisor que não realiza a classificação final, mas sim avalia o contexto de entrada e decide qual o perito mais competente para lidar com aquele cenário específico.

Deste modo, a previsão final não é uma média democrática, como no caso anterior, mas uma soma ponderada dinâmica. Para cada nova observação, a Rede de Controlo atribui pesos diferentes a cada perito. Isto permite que o modelo confie mais num perito treinado para detetar tráfego intenso quando os dados sugerem horas de ponta, e noutro perito para tráfego leve durante a madrugada, entre outros possíveis casos.

TABLE XVII
CLASSIFICATION REPORT MOE

| | *precision* | *recall* | *f1-score* |
|---|---|---|---|
| None | 0.88 | 0.90 | 0.89 |
| Low | 0.67 | 0.72 | 0.69 |
| Medium | 0.78 | 0.72 | 0.75 |
| High | 0.76 | 0.75 | 0.75 |
| Very_High | 0.87 | 0.82 | 0.84 |
| *accuracy* | | | 0.79 |
| *macro avg* | 0.79 | 0.78 | 0.78 |
| *weighted avg* | 0.79 | 0.79 | 0.79 |

## VI. RESULTADOS E AVALIAÇÃO

Nesta secção, detalha-se o desempenho dos modelos desenvolvidos. A análise não se restringe apenas à métrica global de *accuracy*, aprofundando-se na interpretabilidade das decisões, na capacidade de generalização e no impacto do desequilíbrio das classes nos erros de previsão.

### A. *Performance Global*

A Tabela XVIII sumariza o desempenho dos modelos no conjunto de teste local. A ordenação decrescente pela *accuracy* evidencia a evolução da capacidade preditiva à medida que a complexidade da arquitetura aumenta, complementada pela análise do *f1-score* para ponderar o desequilíbrio de classes.

TABLE XVIII
COMPARATIVO DE PERFORMANCE GLOBAL

| Modelo | *accuracy* | *f1-score* |
|---|---|---|
| Stacking | 0.82 | 0.82 |
| XGBoost | 0.81 | 0.81 |
| Random Forest | 0.81 | 0.81 |
| Gradient Boosting | 0.81 | 0.81 |
| ANN | 0.80 | 0.80 |
| 1vRest | 0.80 | 0.80 |
| SVM | 0.79 | 0.79 |
| MoE | 0.79 | 0.79 |
| Logistic Regression | 0.77 | 0.77 |
| Decision Tree | 0.77 | 0.78 |

### B. *Análise de ensembles e arquiteturas*

A análise comparativa dos resultados evidencia uma clara hierarquia de desempenho, validando a premissa de que a combinação de múltiplos estimadores supera modelos isolados. Enquanto arquiteturas singulares, como a *Decision Tree* e a *Logistic Regression*, estagnaram num patamar de 77% de *accuracy*, as estratégias de agregação permitiram extrair padrões mais subtis, ultrapassando consistentemente a barreira dos 80%.

Dentro do universo dos ensembles, observa-se uma distinção qualitativa entre abordagens. Os métodos de *Bagging* e *Boosting*, embora robustos, convergiram para um teto de desempenho de 81%. Isto sugere que, apesar das suas diferenças algorítmicas, estes modelos baseados em árvores partilham enviesamentos semelhantes ao lidar com este conjunto de dados específico.

O modelo de *Stacking* destacou-se como a abordagem mais eficaz, atingindo o máximo global de 82%. A superioridade desta arquitetura reside na sua natureza heterogénea. Ao integrar as previsões de modelos com fronteiras de decisão geometricamente distintas — a linearidade da Regressão Logística, os hiperplanos multidimensionais do SVM e os cortes ortogonais dos algoritmos baseados em árvores — o *Stacking* maximizou a diversidade de informação.

O meta-modelo foi capaz de aprender a "especialização" de cada algoritmo base, ponderando qual o modelo mais fiável para cada tipo de registo. Esta capacidade de corrigir os erros residuais de um modelo com as forças de outro resultou numa generalização superior. Consequentemente, e alinhando com o objetivo estratégico de maximizar a *accuracy* para a competição, o *Stacking* foi selecionado como o modelo final deste projeto.

### C. *Validação e robustez*

A fiabilidade da solução foi validada externamente através da competição na plataforma Kaggle. A comparação entre as métricas de validação local e os *scores* da competição revela uma elevada consistência do modelo:

- **Estimativa Local:** 0.820
- ***Public Score* (30%):** 0.835
- ***Private Score* (70%):** 0.813

Observa-se um alinhamento rigoroso entre a estimativa local e o *Private Score*. A variação mínima confirma que a estratégia de validação cruzada foi eficaz na prevenção de estimativas otimistas e que o modelo não sofreu de muito *overfitting*. O facto de o *Public Score* ser ligeiramente superior deve-se à variância estatística natural em subamostras menores, não refletindo uma discrepância estrutural.

### D. *Interpretabilidade do modelo*

Embora o modelo de *Stacking* tenha sido selecionado como a solução final pela sua *accuracy* superior, a sua arquitetura de meta-aprendizagem introduz camadas de abstração que dificultam a extração direta da importância intrínseca de cada variável.

Para contornar esta limitação de "caixa negra", a análise de interpretabilidade foi conduzida utilizando o *Random Forest*. Esta opção metodológica é validada assumindo-se que os padrões dominantes são partilhados pela elevada proximidade de desempenho entre os dois modelos (0.81 vs 0.82), permitindo assumir que as variáveis determinantes para o *Random Forest* são igualmente críticas para o *Stacking*.

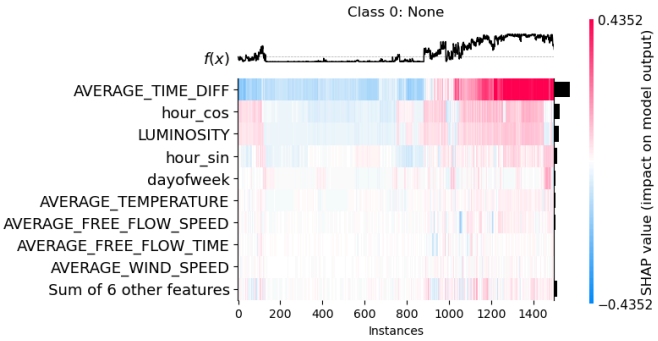


Fig. 15. *Heatmaps* SHAP, classe None

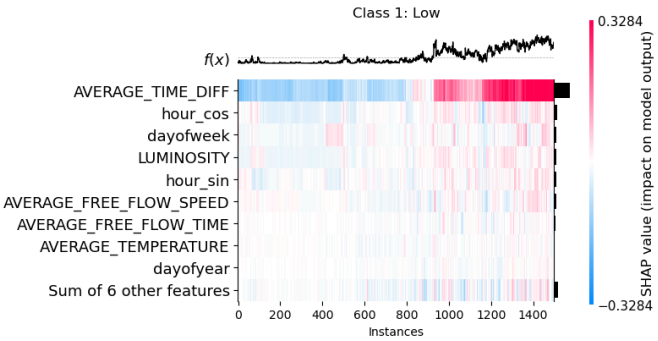


Fig. 16. *Heatmaps* SHAP, classe Low

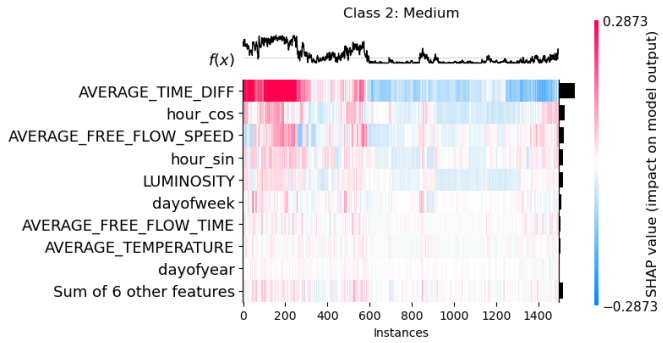


Fig. 17. *Heatmaps* SHAP, classe Medium

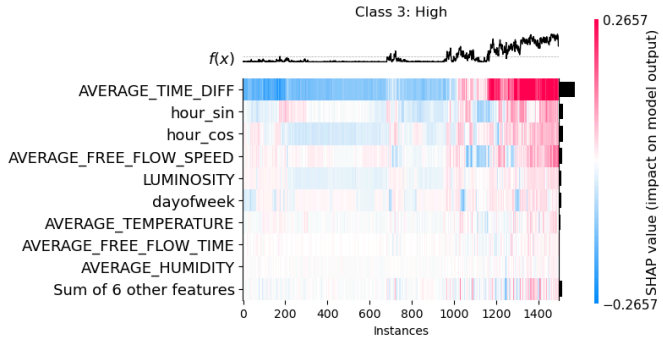


Fig. 18. *Heatmaps* SHAP, classe High

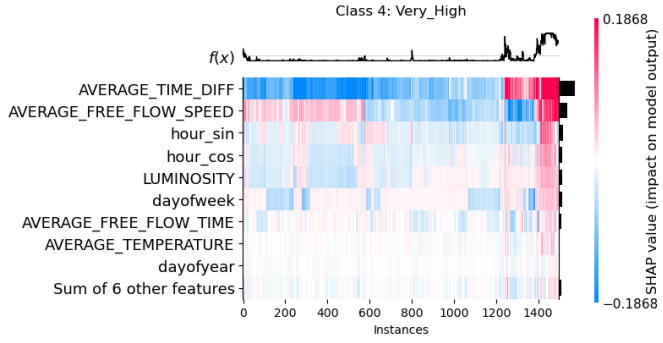


Fig. 19. *Heatmaps* SHAP, classe Very_High

A análise global dos valores SHAP [4] permite validar a coerência "física" do modelo através de dois pilares fundamentais:

- **Dominância Causal:** A variável **`AVERAGE_TIME_DIFF`** afirma-se transversalmente como o preditor dominante em todas as classes. O modelo aprendeu corretamente que a magnitude do atraso temporal é o indicador direto e quantificável da severidade do tráfego, evitando a dependência de correlações espúrias.
- **Modulação Contextual:** Para refinar as previsões, especialmente nas classes intermédias onde a fronteira de decisão é ténue, o modelo recorre a variáveis de contexto. Atributos como a hora do dia, a **`LUMINOSITY`** e a **`AVERAGE_FREE_FLOW_SPEED`** funcionam como fatores de ajuste, permitindo ao algoritmo distinguir entre congestionamentos estruturais e incidentais.

### *E. Análise de erros e desequilíbrio de classes*

Apesar da *accuracy* de 82% indicar um modelo robusto, uma análise detalhada da matriz de confusão revela nuances críticas impostas pelo desequilíbrio das classes.

O modelo demonstra um comportamento tendencialmente conservador. Observa-se que a maioria dos erros de classificação ocorre entre classes adjacentes. Dada a natureza ordinal do problema, este tipo de erro é menos penalizador do ponto de vista prático do que erros dispersos, pois indica que o modelo consegue captar a tendência de agravamento do tráfego, falhando apenas na definição exata do limiar de intensidade.

## VII. Conclusão

O presente trabalho cumpriu com sucesso o objetivo de desenvolver e validar modelos de *Machine Learning* capazes de prever o fluxo de tráfego rodoviário na cidade do Porto. Adotando a metodologia CRISP-DM, foi possível estruturar um *pipeline* robusto que abrangeu desde a limpeza complexa de dados meteorológicos até à implementação de arquiteturas de *ensemble* avançadas.

A análise comparativa permitiu identificar o *Stacking* como a abordagem mais eficaz, atingindo uma *accuracy* de 82% e demonstrando uma superioridade clara sobre os modelos individuais. A capacidade em integrar as fronteiras de decisão heterogéneas da *Logistic Regression*, SVM e algoritmos baseados em árvores revelou-se decisiva para capturar a complexidade não-linear do tráfego urbano. A consistência observada entre a validação cruzada local e os resultados obtidos na competição Kaggle corrobora a robustez da solução e a ausência de *overfitting* significativo.

Do ponto de vista da prospeção de dados, confirma-se a supremacia da variável **`AVERAGE_TIME_DIFF`** como o indicador determinante do estado do tráfego. A análise de interpretabilidade (SHAP) validou que o modelo alinha a sua "aprendizagem" com a realidade física: o atraso temporal define a existência do congestionamento, enquanto variáveis contextuais como a hora e a luminosidade refinam a probabilidade e a intensidade do cenário.

Não obstante os resultados positivos, o projeto enfrentou desafios, nomeadamente o severo desequilíbrio de classes. Embora mitigado através de técnicas de *Oversampling* (SMOTE) e validação estratificada, a análise de erros evidenciou uma tendência conservadora do modelo, com falhas concentradas predominantemente entre classes adjacentes.

## Referências

[1] E. Lewinson, «Three Approaches to Encoding Time Information as Features for ML Models». Acedido: 6 de janeiro de 2026. [Em linha]. Disponível em: https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/

[2] M. Bhagat e B. Bakariya, «Implementation of Logistic Regression on Diabetic Dataset using Train-Test-Split, K-Fold and Stratified K-Fold Approach», *National Academy Science Letters*, vol. 45, n.º 5, pp. 401–404, 2022, doi: 10.1007/s40009-022-01131-9.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, e W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[4] M. Raquib *et al.*, «PSO-XAI: A PSO-Enhanced Explainable AI Framework for Reliable Breast Cancer Detection», *arXiv preprint arXiv:2510.20611*, 2025, [Em linha]. Disponível em: https://arxiv.org/abs/2510.20611