UNIVERSIDADE BEIRA INTERIOR

Estudo da previsão da gravidade dos acidentes rodoviários e dos fatores contribuintes utilizando *machine learning*

(Versão final após defesa pública)

Victoria Carolina Pita Rodrigues

Dissertação para obtenção do Grau de Mestre em
Engenharia Civil
(2º ciclo de estudos)

Orientadora: Prof. Doutora Bertha Maria Batista dos Santos

Agosto de 2025

Declaração de Integridade

Eu, Victoria Carlina Pita Rodrigues, que abaixo assino, estudante com o número de inscrição M14339 de/o Engenharia Civil da Faculdade de Enagenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridades da Universidade da Beira Interior**.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 13/08/2025

gov.br
Documento assinado digitalmente
VICTORIA CAROLINA PITA RODRIGUES
Data: 13/08/2025 06:31:00-0300
Verifique em https://validar.iti.gov.br

(assinatura conforme Cartão de Cidadão ou preferencialmente
assinatura digital no documento original se naquele mesmo formato)

Dedicatória

Dedico esta conquista a Deus, pela força e sabedoria em cada passo da jornada; àqueles que, mesmo distantes, sempre se fazem presentes; à minha família de sangue e à família que escolhi — meus amigos —, pelo apoio, carinho e companheirismo ao longo do caminho.

Agradecimentos

Gostaria de agradecer à minha família por todo o apoio durante a minha jornada, à minha orientadora por me mostrar o caminho certo a seguir, aos meus amigos e ao Departamento de Engenharia Civil e Arquitetura (DECA) da UBI por fornecerem todo o suporte necessário para a realização desta dissertação. Agradeço também à Autoridade Nacional de Segurança Rodoviária (ANSR), por ter disponibilizado os dados utilizados no caso de estudo, fundamentais para o desenvolvimento deste trabalho de dissertação.

Resumo

Os acidentes rodoviários representam uma das principais causas de morte e ferimentos em Portugal. Em 2023, mais da metade dos acidentes com vítimas (52,7%) foi provocada por colisões, os despistes representaram 33,6% dos casos, enquanto os atropelamentos corresponderam a 13,4%, resultando em 479 mortes, 2.646 feridos graves e 42.890 feridos leves. Diante desse cenário, este estudo tem como objetivo desenvolver modelos preditivos para a gravidade dos acidentes rodoviários do tipo colisão e despiste, utilizando técnicas de aprendizado de máquina (*Machine Learning*), com base em dados fornecidos pela Autoridade Nacional de Segurança Rodoviária (ANSR) referentes ao período de 2019 a 2023.

Para tratar o desequilíbrio entre as classes de gravidade da variável de resposta (feridos leves vs. feridos graves e vítimas fatais), foram aplicadas as técnicas de balanceamento SMOTE (*Synthetic Minority Over-sampling Technique*) e RUS (*Random Under-Sampling*). Os algoritmos analisados foram: *Naive Bayes* (NB), *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), *Decision Tree* (DT), *Random Forest* (RF) e *K-Nearest Neighbors* (KNN). Os modelos foram treinados utilizando a validação cruzada com 10 subconjuntos e a divisão percentual 80/20 (80% para treinamento e 20% para teste).

Os resultados demonstraram que os melhores desempenhos foram obtidos com o algoritmo *Random Forest*, especialmente quando aplicado à base de dados balanceada com a técnica *SMOTE*, apresentando para os acidentes do tipo colisão uma área sob a curva ROC de 0,886 e um coeficiente *Kappa* igual a 0,607, já para os acidentes do tipo despiste a área sob a curva ROC apresentou um valor de 0,871 e um coeficiente *Kappa* de 0,573. A análise dos modelos também permitiu a identificação das variáveis mais influentes nos acidentes (fatores de risco), possibilitando a definição de recomendações para intervenções na infraestrutura viária e comportamento dos condutores, com vista à redução dos acidentes com consequências mais graves. Permitiu ainda a realização de estudos de previsão da gravidade dos acidentes mediante a modificação desses fatores na base de dados original. Os resultados desta pesquisa evidenciam o potencial do uso de técnicas de *Machine Learning* como ferramenta de apoio à tomada de decisão na área da segurança viária.

Palavras-chave

Segurança Rodoviária; Acidentes Rodoviários; Gravidade dos Acidentes; *Machine Learning*; Previsão; Fatores de Risco; Simulação**.**

Abstract

Road accidents are one of the main causes of death and injury in Portugal. In 2023, more than half of accidents with victims (52.7%) were caused by collisions, run-off-road accidents represented 33.6% of cases, while run-overs accounted for 13.4%, resulting in 479 deaths, 2,646 serious injuries and 42,890 minor injuries. In view of this scenario, this study aims to develop predictive models for the severity of collision and run-off-road accidents, using machine learning techniques, based on data provided by the Portuguese Road Safety Authority (ANSR) for the period from 2019 to 2023.

To address the imbalance between the severity classes of the response variable (minor injuries vs. serious injuries and fatalities), the SMOTE (Synthetic Minority Over-sampling Technique) and RUS (Random Under-Sampling) balancing techniques were applied. The algorithms analysed were: Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and K-Nearest Neighbors (KNN). The models were trained using 10-fold cross-validation and a 80/20 percentage split (80% for training and 20% for testing).

The results showed that the best performances were obtained with the Random Forest algorithm, especially when applied to the balanced database with the SMOTE technique, presenting for collision-type accidents an area under the ROC curve of 0.886 and a Kappa coefficient of 0.607, while for run-off-road accidents the results were an area under the ROC curve of 0.871 and a Kappa coefficient of 0.573. The analysis of the models also allowed the identification of the most influential variables in accidents (risk factors), enabling the definition of recommendations for interventions in road infrastructure and driver behavior, with a view to reducing accidents with more serious consequences. It also enabled studies to predict accident severity by modifying these factors in the original database. The results of this research highlight the potential of using Machine Learning techniques as a tool to support decision-making in the area of road safety.

Keywords

Road Safety; Road Accidents; Accident Severity; Machine Learning; Prediction; Risk Factors; Simulation.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Acrónimos

AE	Autoestrada
ANN	Artificial Neural Netwok
ANSR	Autoridade Nacional de Segurança Rodoviária
AUC	área sob a curva ROC
DT	Decision Tree
EM	Estrada Nacional
IA	Inteligência Artificial
IC	Itinerário Complementar
IP	Itinerário Principal
KNN	K-Nearest Neighbors
ML	Machine Learning
NB	Naive Bayes
OMS	Organização mundial da saude
RF	Random Forests
ROC	Receiver Operating Characteristic
ROS	Random Over Sampling
RUS	Random Under Sampling
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
VIF	Variance Inflation Factor

1. Introdução

A segurança rodoviária é um tema de crescente preocupação em escala global, especialmente em virtude do elevado número de mortes e lesões causadas por acidentes de trânsito. Em 2021, por exemplo, mais de 1,19 milhões de pessoas morreram em acidentes rodoviários, segundo a Organização Mundial da Saúde (OMS). Esse cenário reforça a necessidade de se adotar medidas mais eficazes para a prevenção de sinistros e para a mitigação dos seus impactos.

Em Portugal, apesar dos avanços nas políticas públicas e no investimento em infraestrutura viária, os números ainda são preocupantes. De acordo com dados da Autoridade Nacional de Segurança Rodoviária (ANSR), entre janeiro e outubro de 2024 foram registadas mais de 30 mil ocorrências com vítimas, resultando em 392 mortes e mais de 2 mil feridos graves. A maioria desses acidentes ocorreu por colisões ou despistes

Diante dessa realidade, torna-se evidente a necessidade de aprofundar o entendimento sobre o que contribui para a gravidade dos acidentes e como é possível, de forma prática e eficiente, reduzir os seus impactos. A tradicional análise estatística, embora útil, nem sempre consegue captar toda a complexidade por trás dos acidentes. Nesse contexto, os métodos baseados em *Machine Learning* (ML) vêm ganhando destaque por sua capacidade de processar grandes volumes de dados, extrair relações ocultas e gerar previsões com elevado grau de precisão.

Com a disponibilidade de bases de dados cada vez mais completas, o uso de *Machine Learning* na área da segurança rodoviária tem se mostrado promissor. Essa tecnologia pode ajudar a modelar e prever a gravidade dos acidentes com base em diversos fatores, tais como: o tipo de estrada e as suas características, as condições atmosféricas, o tipo de veículos envolvidos, entre outros. Além disso, oferece a possibilidade de simular cenários e testar medidas de intervenção, o que representa um apoio importante à tomada de decisão por parte de gestores públicos e profissionais da área.

A importância deste estudo pode ser compreendida a partir de três pontos centrais, o primeiro diz respeito à dimensão social e económica dos acidentes rodoviários, que continuam a provocar perdas humanas irreparáveis e a gerar impactos profundos na vida das pessoas e nos sistemas públicos. O segundo ponto está no potencial das técnicas de *Machine Learning*, que permitem analisar grandes volumes de dados e identificar padrões que ajudam a antecipar situações de risco. E, por fim, destaca-se a utilidade prática dos resultados, que podem apoiar a formulação de políticas públicas mais eficientes, baseadas em evidências concretas e orientadas para a prevenção.

1.1 Objetivos

Diante desse cenário, esta dissertação tem como principal objetivo desenvolver e avaliar modelos preditivos capazes de estimar a gravidade dos acidentes rodoviários, com especial

atenção aos dois tipos mais recorrentes: colisões e despistes. Para isso, será utilizado o banco de dados de sinistralidade portuguesa, dos anos de 2018 a 2023, com diferentes algoritmos de *Machine Learning*, aliados a técnicas de balanceamento de dados, com o intuito de identificar os fatores que mais influenciam a gravidade dos acidentes. Com base nos resultados obtidos, serão propostas medidas de intervenção que serão testadas por meio de simulações nos modelos que melhor traduzem este fenómeno, de modo a apoiar a tomada de decisões mais eficazes na área da segurança rodoviária.

1.2 Estrutura do documento

Esta dissertação está organizada em cinco capítulos, sendo que o Capítulo 1 – Introdução, contextualiza o tema, apresenta a motivação para o desenvolvimento da pesquisa e destaca a relevância da utilização de técnicas de *Machine Learning* no campo da segurança rodoviária. Também são definidos os objetivos do estudo e a forma como o trabalho está estruturado.

O Capítulo 2 – Sinistralidade Rodoviária apresenta o panorama global e nacional da sinistralidade rodoviária, destacando os principais tipos de acidentes, além das estratégias adotadas por entidades internacionais e por Portugal no combate à insegurança viária.

O Capítulo 3 – *Machine Learning* apresenta o conceito e explica de forma clara os principais algoritmos utilizados nesta área, suas vantagens e limitações, bem como as abordagens específicas aplicadas ao problema de desequilíbrio dos dados (classes) da variável resposta — uma característica comum em bases relacionadas a acidentes rodoviários.

O Capítulo 4 – Estudo de caso apresenta o estudo desenvolvido com base nos dados fornecidos pela ANSR. São descritas detalhadamente a metodologia adotada, o tratamento e preparação dos dados, os testes realizados com diferentes algoritmos e técnicas de balanceamento, e os resultados obtidos com cada modelo. A partir dos modelos desenvolvidos, são identificados fatores de risco e propostas e avaliadas medidas de intervenção através da realização de simulações. O objetivo destas operações é o de verificar o impacto potencial de alterações em variáveis relevantes na redução da gravidade dos acidentes.

Por fim, o Capítulo 5 – Conclusões e trabalhos futuros apresenta as conclusões do estudo, destacando os principais resultados, as contribuições práticas e científicas da pesquisa e as possibilidades de trabalhos futuros que possam dar continuidade e aprofundamento ao estudo realizado.

Além dos capítulos principais, a dissertação inclui também um conjunto de anexos, que trazem informações complementares e ajudam a entender a problemática estudada.

2. Sinistralidade Rodoviária

2.1 Enquadramento

Os acidentes rodoviários continuam sendo uma das principais causas de mortes no mundo. De acordo com o Informe sobre a Situação Mundial da Segurança Vial 2023, da Organização Mundial da Saúde (OMS), cerca de 1,19 milhões de pessoas morreram em acidentes de trânsito em 2021, o que corresponde a 15 mortes por 100.000 habitantes. Essas fatalidades atingem desproporcionalmente os países de baixa e média renda, onde ocorrem 92% das mortes, embora esses países possuam apenas cerca de 1% da frota mundial de veículos.

Os motociclistas e condutores de veículos de duas ou três rodas são os mais afetados, representando 30% das mortes, seguidos pelos ocupantes de veículos de quatro rodas (25%), pedestres (21%) e ciclistas (5%) (WHO, 2023).

Os acidentes rodoviários, além dos danos físicos, muitas vezes fatais, geram consequências sociais e econômicas significativas. Estima-se que os custos associados aos sinistros de trânsito absorvam entre 1% e 3% do produto interno bruto (PIB) da maioria dos países, podendo chegar até 6% em alguns casos (WHO, 2023). Esses impactos incluem não apenas os custos diretos com serviços de emergência e saúde, mas também perdas de produtividade, danos à infraestrutura e consequências emocionais para as vítimas e suas famílias.

O problema envolvendo os sinistros de trânsito não é recente, assim como os estudos voltados à sua compreensão. Entre as décadas de 1960 e 1970, William Haddon dedicou-se a identificar os principais fatores contribuintes para os acidentes de trânsito, bem como os elementos que devem ser analisados para a redução das suas consequências. Como apresentado na Tabela 1, que apresenta a “Matriz de Haddon”, é possível observar que os acidentes resultam da interação de três grandes grupos de fatores: fatores humanos, fatores relacionados aos veículos e equipamentos, e fatores ambientais.

Tabela 1 -Matriz de Haddon (Haddon, 1968, apud Goniewicz et al., 2015).

Fase	Fator Humano	Fator Veículo/Equipamento	Fator Ambiental
Pré-evento	Treinamento e comportamento do condutor, fadiga, álcool, distrações, velocidade excessiva	Condições dos freios, pneus, luzes, dispositivos de segurança	Sinalização, visibilidade, clima, desenho viário
Evento (impacto)	Reação do condutor, uso de cinto ou capacete	Capacidade de absorção de impacto, airbags, integridade da estrutura	Condições da via no local do impacto (curvas, obstáculos)
Pós-evento	Capacidade de pedir ajuda, primeiros socorros	Facilidade de resgate (portas travadas, estabilidade)	Acesso de emergência, tempo de resposta dos serviços

Entre os fatores identificados, o fator humano é o mais relevante, sendo responsável por cerca de 90 a 95% dos acidentes. Esse fator está relacionado a comportamentos de risco por parte dos condutores e demais usuários da via, como o excesso de velocidade, a condução sob efeito de álcool ou outras substâncias psicoativas, bem como o desrespeito pelas normas de trânsito. Além disso, a falta de habilidades adequadas para lidar com situações de risco na condução também contribui significativamente para o aumento do número de acidentes.

O segundo fator de influência é o fator veículo/equipamento, que responde por aproximadamente 8 a 10% dos casos de sinistro. Ele abrange falhas técnicas ou mecânicas dos veículos, como problemas nos freios, pneus ou sistemas de iluminação, além da ausência ou ineficiência de dispositivos de segurança ativa e passiva, como cintos de segurança, *airbags* e sistemas de freios ABS. A má conservação dos veículos e a não realização de manutenções periódicas também elevam o risco de acidentes graves.

Por fim, o fator ambiental e da infraestrutura representa cerca de 28 a 35% das causas, estando associado às condições das vias e do entorno. Situações como falta de sinalização adequada, iluminação deficiente, condições climáticas adversas, ausência de calçadas e ciclovias, bem como falhas no desenho viário, podem contribuir decisivamente para a ocorrência e agravamento de sinistros.

A matriz apresentada serve como ponto de partida para compreender os fatores que contribuem para a gravidade dos acidentes, possibilitando, assim, a formulação de estratégias integradas voltadas à segurança rodoviária. Estas estratégias devem abranger todos os grupos interessados, incluindo a administração pública, os institutos de investigação científica, as instituições não governamentais e os serviços médicos de emergência (Goniewicz et al., 2015). De entre as estratégias existentes, destaca-se a instituída pela ONU, intitulada de Década de Ação pela Segurança no Trânsito 2021–2030, com a meta de reduzir em pelo menos 50% o número de mortes e lesões no trânsito até 2030. O Plano Global para a Década de Ação propõe uma abordagem de "Sistemas Seguros".

A abordagem do Sistema Seguro difere significativamente das estratégias tradicionais de segurança rodoviária, que frequentemente colocavam o foco principal no comportamento humano. Embora não desresponsabilize os condutores, o Sistema Seguro enfatiza a importância de veículos e infraestruturas seguras, autoexplicativas e tolerantes aos erros humanos. A meta é criar um ambiente rodoviário que absorva as limitações físicas dos seres humanos e minimize as consequências dos erros inevitáveis (ANRS, 2021).

O Sistema Seguro baseia-se nos seguintes princípios fundamentais:

- As pessoas cometem erros que podem levar a acidentes;
- O corpo humano tem uma capacidade limitada de resistência a colisões.

- A segurança é uma responsabilidade partilhada por todos os intervenientes no sistema de transporte, incluindo projetistas, construtores, gestores, utilizadores das vias e veículos, autoridades de fiscalização e serviços de emergência.

- Todos os elementos do sistema devem ser reforçados em combinação, de modo que, se um falhar, os outros possam compensar, garantindo a proteção dos utilizadores da estrada

Entre as medidas recomendadas no plano estão: o desenvolvimento de infraestruturas viárias seguras (com calçadas, ciclovias, cruzamentos protegidos, entre outros), normas de segurança veicular mais rigorosas (como cintos de segurança obrigatórios, freios ABS e sistemas de assistência eletrônica), além de políticas urbanas voltadas ao transporte público e mobilidade ativa (caminhada e bicicleta) (ONU, 2021).

A Comissão Europeia adota quatro estratégias fundamentais para reduzir o número de acidentes com vítimas e mortes nas estradas. A primeira delas é a campanha "Visão Zero", que estabelece como meta de médio prazo a redução em 50% do número de mortes e feridos graves até 2030, e como objetivo de longo prazo a eliminação total de vítimas mortais e feridos graves nas estradas europeias até 2050 (P. E., 2021).

A segunda estratégia foca em infraestruturas seguras, priorizando investimentos com maior retorno em termos de segurança rodoviária, sobretudo nas áreas com maior concentração de acidentes. A manutenção da infraestrutura existente é considerada prioridade, complementada pela construção de novas vias, quando necessário (P. E., 2021).

A terceira estratégia diz respeito aos veículos seguros, incentivando a incorporação de tecnologias avançadas de segurança, como sistemas de adaptação inteligente da velocidade e assistência à permanência na faixa de rodagem, que aumentam a proteção dos ocupantes e usuários da via (P. E., 2021).

A quarta estratégia envolve a utilização segura das estradas, com medidas como a tolerância zero ao consumo de álcool ou substâncias psicoativas na condução, além da promoção de limites de velocidade seguros, como os 30 km/h em zonas residenciais, visando proteger especialmente os usuários mais vulneráveis (P. E., 2021).

2.2 Estratégias e sinistralidade em Portugal

Em atendimento às políticas de segurança viária da Comissão Europeia para o período de 2021–2030 e aos objetivos estabelecidos pela Organização das Nações Unidas na Década de Ação pela Segurança no Trânsito 2021–2030, Portugal adotou o programa Visão Zero 2030. Esta iniciativa estabelece o horizonte de médio prazo para a política de segurança rodoviária no país,

definindo objetivos estratégicos e operacionais a serem concretizados por meio de planos de ação bienais.

Para atender às metas propostas, o programa Visão Zero 2030 está sendo desenvolvido em três fases distintas. A primeira fase refere-se aos Princípios Balizadores da Estratégia Nacional de Segurança Rodoviária, que indicam a abordagem da estratégia nacional para a década, considerando as necessidades específicas de Portugal, a experiência adquirida com estratégias anteriores e os compromissos assumidos no plano internacional.

A segunda fase consiste na elaboração de relatórios técnico-científicos que apresentam os elementos-chave da nova estratégia, descrevem os caminhos para sua implementação e fornecem recomendações sobre metodologias para a formulação dos planos de ação bienais, bem como os processos de monitorização e avaliação necessárias.

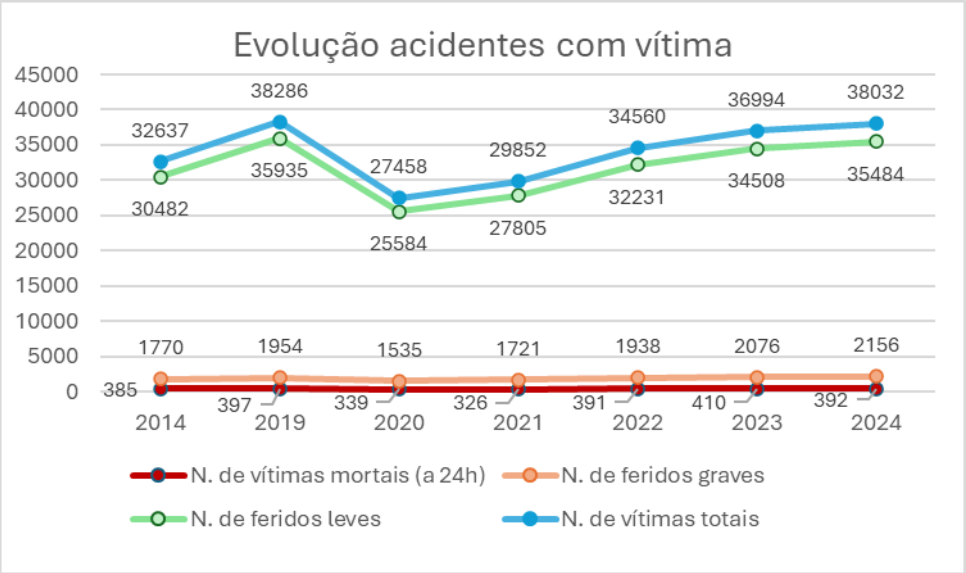
Atualmente, o programa encontra-se na terceira fase, dedicada à elaboração da Estratégia Visão Zero 2030 e do respetivo Plano de Ação, com base nos relatórios técnico-científicos previamente produzidos, nas análises diagnósticas da situação atual e na definição das bases que orientarão as ações futuras.

Com o objetivo de acompanhar o alcance das metas estipuladas, a Autoridade Nacional de Segurança Rodoviária (ANSR) publica anualmente um relatório anual de sinistralidade a 30 dias com os dados de acidentes com vítimas em Portugal. A Tabelas 2 e ao Gráfico 1 apresentam a evolução do número de acidentes em Portugal.

Tabela 2 - Acidentes com vítimas em Portugal (adaptado ANSR, 2025).

Ano	N. de vítimas mortais (a 24h)	N. de feridos graves	N. de feridos leves	N. de vítimas totais
2014	385	1770	30482	32637
2019	397	1954	35935	38286
2020	339	1535	25584	27458
2021	326	1721	27805	29852
2022	391	1938	32231	34560
2023	410	2076	34508	36994
2024	392	2156	35484	38032

Gráfico 1 - Evolução dos acidentes com vitímas em Portugal.



Com base nas informações disponibilizadas, observa-se que os anos de 2020 e 2021 — período marcado pelas restrições impostas pela pandemia de COVID-19 — apresentaram os menores índices de sinistralidade do período analisado. A partir de 2022, no entanto, verifica-se uma tendência de crescimento no número de acidentes, o que se relaciona diretamente com a retoma gradual da circulação e da atividade normal em todo o país.

Quanto à natureza dos acidentes, a Tabela 3 apresenta os casos registados nos anos de 2019, 2023 e 2024.

Tabela 3 - Acidentes por natureza (ANSR, 2025).

Janeiro-outubro	Acidentes com vítimas			Vítimas Mortais			Feridos Graves			Feridos Leves		
	2019	2023	2024	2019	2023	2024	2019	2023	2024	2019	2023	2024
Atropela-mento	4195	3720	3875	56	48	56	367	285	335	4166	3688	3847
Colisão	15790	15558	16184	159	158	158	860	954	1002	21035	19866	20665
Despiste	9650	10130	10167	182	204	178	727	837	819	10734	10954	10972
Total	29635	29408	30226	397	410	392	1954	2076	2156	35935	34508	35484

Com base nesses dados, observa-se que as colisões representaram 53,5% do total de acidentes em 2024, enquanto os despistes corresponderam a 33,6%. Juntas, essas duas categorias foram responsáveis por mais de 85% das vítimas registadas no período analisado.

Essa evidência reforça a relevância de estudos voltados à construção de modelos preditivos da gravidade das lesões sofridas pelas vítimas de acidentes, com base em dados reais de sinistralidade rodoviária. Além disso, destaca-se a importância da simulação de medidas de

intervenção, com o objetivo de avaliar o impacto dessas ações na redução da gravidade dos acidentes.

3. Machine Learning

3.1 Enquadramento

No domínio da Inteligência Artificial (IA), o aprendizado de máquina (*Machine Learning* – ML) é um subcampo extremamente ativo e revolucionário. O ML pode ser compreendido como a capacidade de uma máquina aprender com experiências anteriores, ou seja, melhorar o seu desempenho com base nos resultados passados (Yao & Liu, 2005).

Segundo Yao e Liu (2005), o aprendizado de máquina envolve dois elementos principais: o elemento de aprendizagem e o elemento de desempenho. De forma simplificada, o processo ocorre em três etapas:

1. O ambiente fornece informações ao elemento de aprendizagem;
2. O elemento de aprendizagem utiliza essas informações para modificar o elemento de desempenho;
3. O elemento de desempenho, por sua vez, toma decisões mais eficazes, selecionando as ações adequadas para executar a tarefa.

A Figura 1 ilustra esse processo.

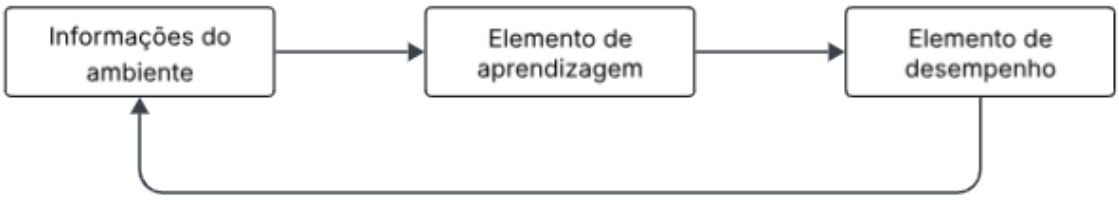


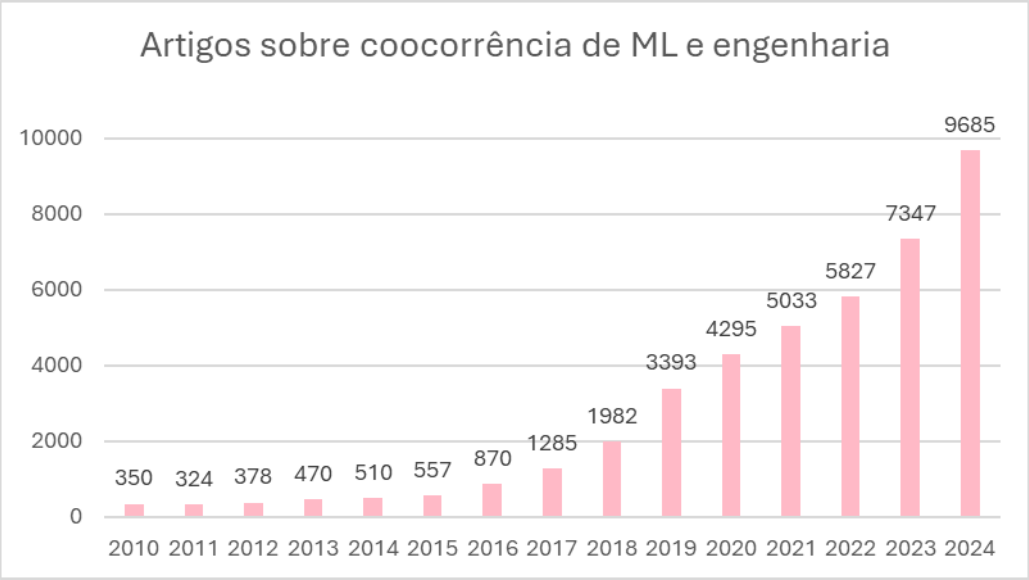
Figura 1 - Modelo de aprendizado de máquina (adaptado Yao & Liu, 2005).

O crescimento do uso do ML tem sido expressivo, principalmente devido à sua capacidade de realizar tarefas complexas, como: identificação de padrões, conclusões com base em dados, previsões e análises preditivas, descoberta de conexões ocultas, deteção de anomalias, além de apoio à tomada de decisões (Nithya et al., 2023).

Nos últimos anos, o uso do ML na engenharia, em suas mais diversas áreas, tem se expandido significativamente, com aplicações voltadas à previsão de resultados, otimização de processos, planejamento de layouts, auxílio à tomada de decisões, entre outras.

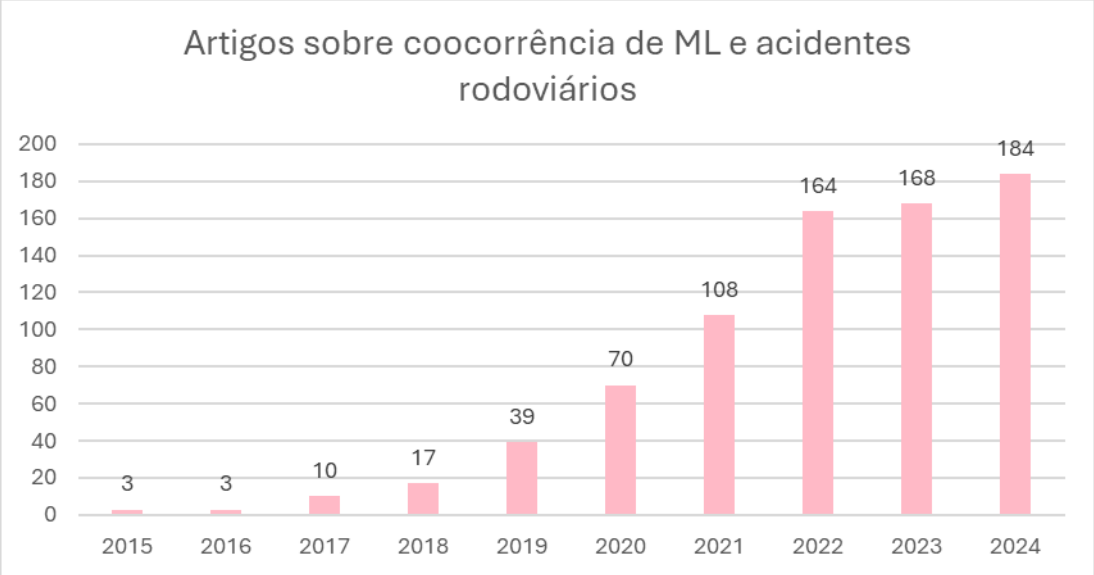
Esse avanço pode ser observado pela quantidade crescente de estudos publicados sobre esses temas. Utilizando o motor de busca Scopus e as palavras-chave "*machine learning*" e "*engineering*", verifica-se um aumento significativo nas publicações ao longo dos últimos sete anos, conforme ilustrado na Gráfico 2.

Gráfico 2 - Artigos sobre coocorrência de ML e engenharia.



Ao se realizar uma busca mais direcionada à temática em estudo, utilizando as palavras-chave "*machine learning*" e "*road accidents*", observa-se, nos últimos cinco anos, um crescimento relevante nas publicações relacionadas ao tema, conforme apresentado no Gráfico 3.

Gráfico 3 - Artigos sobre coocorrência de ML e acidentes rodoviários.



Os artigos citados anteriormente abordam de forma abrangente o uso de ML aplicado a acidentes rodoviários.

3.2 Algoritmos

O *Machine Learning* (ML) apresenta três vertentes principais de aprendizado: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Neste estudo, serão utilizados algoritmos baseados em aprendizagem supervisionada.

Segundo Nithya et al. (2023), a aprendizagem supervisionada (*supervised learning*) é definida como uma técnica de *machine learning* em que o modelo é treinado utilizando dados rotulados, ou seja, cada entrada no conjunto de dados está associada a uma saída conhecida.

O objetivo desse tipo de aprendizado é ensinar o modelo a aprender uma função de mapeamento entre as entradas e as saídas, de forma que ele seja capaz de prever corretamente os resultados para novos dados. Nas subsecções seguintes, detalham-se os algoritmos a aplicar no caso de estudo da presente dissertação.

3.2.1 Naive Bayes (NB)

O *Naive Bayes* (NB) é um algoritmo de classificação probabilística simples, porém preciso, baseado no Teorema de Bayes. Seu objetivo é estimar a classe de uma instância $X = (x_1, x_2, \ldots, x_n)$ como aquela que maximiza a probabilidade posterior $P(Y = y \mid \mathrm{X} = \mathrm{x})$ (ver Equação 1).

$$P(Y = y \mid \mathrm{X} = \mathrm{x}) = \frac{P(Y=y)\prod_{i=1}^{n} P(X=x_1|\mathrm{Y}=\mathrm{y})}{\sum_{i}^{C} P(y_1, X=x)} \qquad \text{(Equação 1)}$$

Em que: X é um vetor n-dimensional de números aleatórios, Y é uma variável aleatória e C representa o número de classes da variável Y.

A suposição principal do NB puro é que o valor de qualquer característica dada é independente do valor de qualquer outra característica, dado o valor da classe.

Os cálculos não dependem diretamente de $P(X = x)$, pois esse valor é apenas um fator de normalização que garante que a soma das probabilidades seja igual a 1.

$$P(Y = y \mid \mathrm{X} = \mathrm{x}) = 1 \qquad \text{(Equação 2)}$$

Além disso, supõe-se que os dados seguem uma distribuição Gaussiana, embora a literatura aponte algumas exceções.

A Figura 2 apresenta uma visão geral do funcionamento do algoritmo *Naive Bayes* (NB), onde se observa que cada uma das variáveis A_i é considerada estatisticamente independente entre si, tendo dependência apenas em relação à classe, C, à qual pertence.

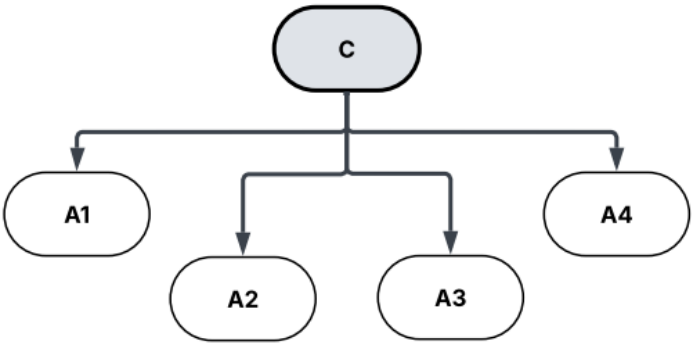


Figura 2 - Estrutura NB (adaptado de Zhang, 2004).

O Teorema de Bayes puro sofreu algumas modificações para se adequar à necessidade de considerar a independência condicional entre os atributos, dado o valor da classe. Com isso, a versão mais utilizada considera um fator w_i , que representa os pesos atribuídos a cada atributo. A probabilidade conjunta, então, é dada por:

$$P(Y = y \mid \mathrm{X} = \mathrm{x}) = \hat{P}\ (Y = y) \prod_{i=1}^{n} \hat{P}(X = x_1 \mid \mathrm{Y} = \mathrm{y})^{w_i} \qquad \text{(Equação 3)}$$

Em que: $\hat{P}$ é a probabilidade ponderada e w_i é o peso de cada atributo.

Essa simplificação torna o algoritmo computacionalmente eficiente, mesmo em conjuntos de dados grandes (Wickramasinghe e Kalutarage, 2020).

3.2.2 Artificial Neural Network (ANN)

Uma Rede Neural Artificial (*Artificial Neural Network* – ANN) é um modelo computacional inspirado no funcionamento do cérebro humano. Ela é composta por "neurônios" organizados em camadas (entrada, ocultas e saída), conectados por pesos, que são ajustados durante o treinamento da rede. A principal função da ANN é identificar padrões e prever resultados com base nos dados de entrada. As ANNs são compostas por redes em camadas de nós interligados (neurônios). Cada neurônio recebe informações, aplica uma função de ativação não linear e envia sua saída para a próxima camada.

A Figura 3 apresenta uma visão geral da estrutura de uma ANN.

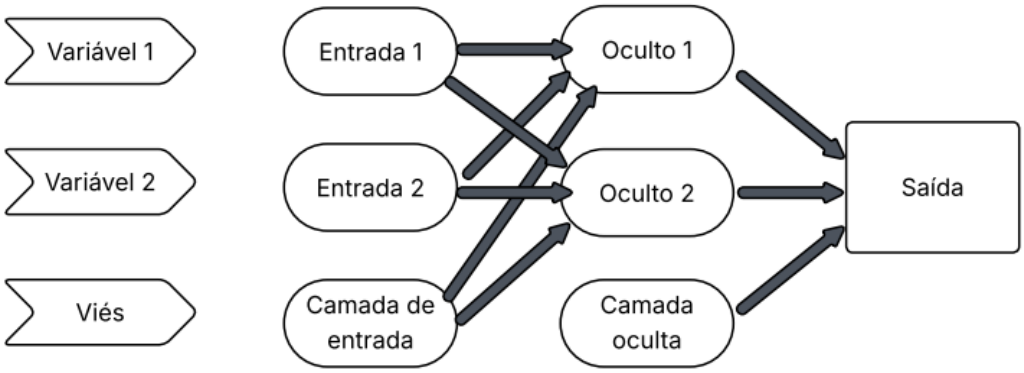


Figura 3 - Estrutura ANN (adaptado de TU, 1996).

Durante o treinamento da ANN, os dados são apresentados à rede, que realiza os cálculos e compara os resultados obtidos com os valores reais (saída desejada). A diferença entre eles (erro) é utilizada para ajustar os pesos, por meio de um processo chamado backpropagation, até que a rede consiga gerar saídas com erros mínimos (Afework & Sipos, 2020).

As ANNs são particularmente úteis em problemas que envolvem relações não lineares complexas entre variáveis, como a previsão de acidentes de trânsito (Tu, 1996).

3.2.3 Support Vector Machine (SVM)

A Máquina de Vetores de Suporte (*Support Vector Machine* – SVM) é um algoritmo de aprendizado supervisionado que pode ser utilizado tanto para classificação quanto para regressão. Seu objetivo é encontrar um hiperplano ótimo que separe os dados em diferentes classes com a maior margem possível. Em muitos casos, como ilustrado na Figura 4, pode haver diversas possibilidades de separação entre dois grupos de dados – por exemplo, entre círculos e quadrados.

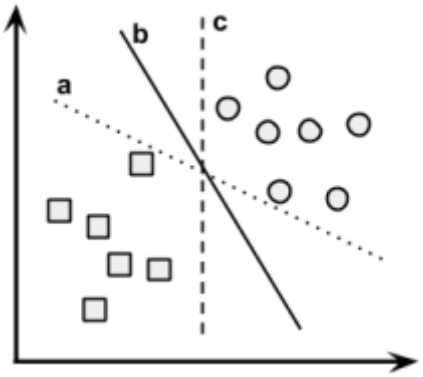


Figura 4 – Estrutura SVM (Lantz, 2013).

Na imagem, as linhas identificadas como a, b e c representam três possíveis divisórias que separam corretamente os dois grupos. Embora todas classifiquem corretamente os dados

apresentados, o SVM busca identificar a que oferece a maior margem de separação possível entre as classes. Essa linha ideal é conhecida como o Hiperplano de Margem Máxima (*Maximum Margin Hyperplane* – MMH) (Lantz, 2013).

No caso de dados que não são linearmente separáveis, o SVM utiliza funções kernel, como a função radial (RBF), para mapear os dados para um espaço de maior dimensão, onde se torne possível a separação das classes (Sun et al., 2017).

O SVM também pode ser adaptado para aprendizado incremental, permitindo atualizar o modelo com novos dados de forma rápida, o que é especialmente útil em contextos de previsão em tempo real.

3.2.4 Decision Tree (DT)

A Árvore de Decisão (*Decision Tree* – DT) é um método de aprendizado supervisionado, amplamente utilizado para tarefas de classificação e regressão. Ele utiliza uma estrutura em forma de árvore para tomar decisões com base nos atributos dos dados.

Cada nó interno representa um teste aplicado a um atributo, cada ramo corresponde a um resultado possível desse teste, e cada nó folha representa uma classe (no caso de classificação) ou um valor de saída (no caso de regressão) (Sarker, 2021).

A Figura 5 apresenta uma visão geral da estrutura DT.

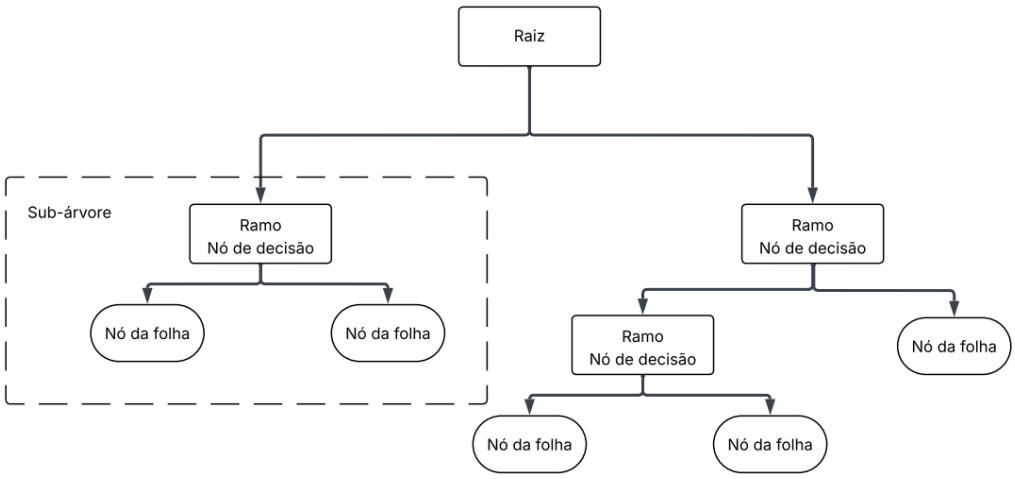


Figura 5 - Estrutura DT (adaptado Sarker, 2021).

A raiz é o nó principal onde a árvore se inicia, geralmente associada à variável mais relevante para o problema analisado. Os ramos são as divisões que ocorrem nos nós internos, guiadas pelos valores das variáveis do conjunto de dados, conduzindo o caminho para os próximos nós.

Cada nó intermediário representa uma variável e uma decisão baseada em seus valores, enquanto as folhas são os nós terminais que apresentam o resultado final da classificação (Abellán et al., 2013, Abdullah & Sipos, 2022, Nedjmedine & Tahar, 2022,)

3.2.5 Random Forest (RF)

A Floresta Aleatória (*Random Forest* – RF) é um algoritmo de aprendizado supervisionado baseado na aprendizagem em conjunto, que combina diversos modelos (no caso, árvores de decisão) para melhorar o desempenho do modelo final. Cada árvore é treinada com uma amostra aleatória dos dados (técnica de ensaio – *bagging*) e considera um subconjunto aleatório de atributos a cada divisão do nó, o que garante diversidade entre as árvores.

O RF funciona construindo várias árvores de decisão de forma paralela, cada uma treinada com uma amostra aleatória dos dados. Para fazer uma previsão, cada árvore "vota" e a classe mais votada é a escolhida como resultado final. Além disso, em cada nó da árvore, uma seleção aleatória de atributos é considerada para dividir os dados, o que introduz diversidade entre as árvores e evita correlação entre elas (Sarker, 2021).

A Figura 6 apresenta uma visão geral da estrutura RF.

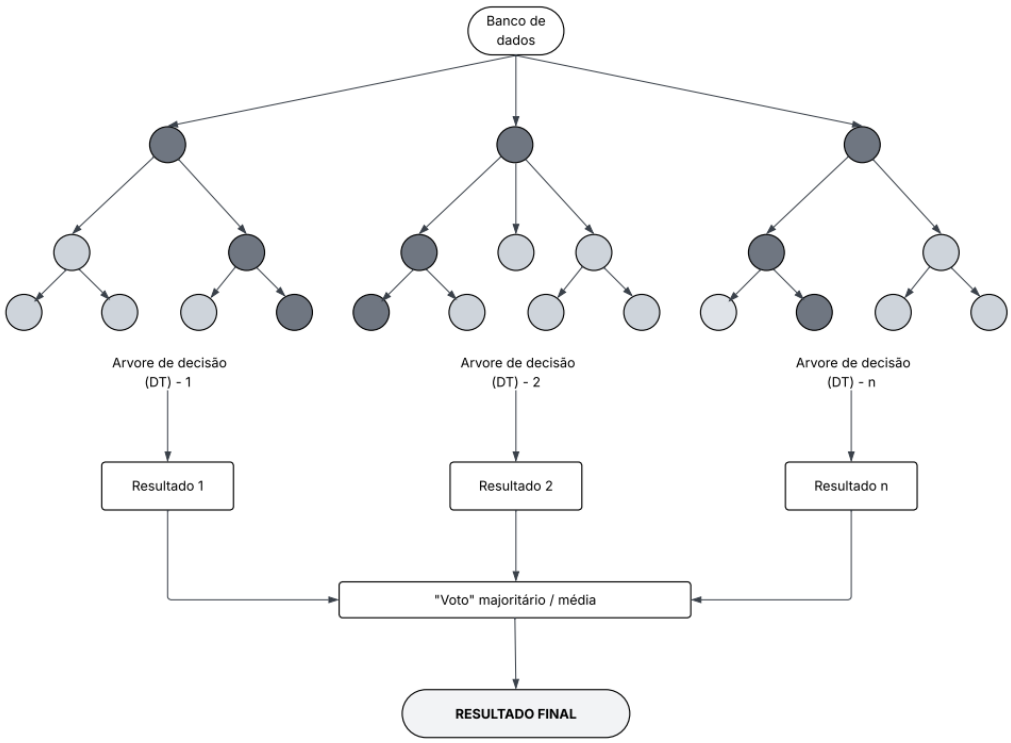


Figura 6 - Estrutura RF (adaptado Sarker, 2021).

3.2.6 K-Nearest Neighbors (KNN)

O *K-Nearest Neighbors* (KNN) é um algoritmo de aprendizado supervisionado utilizado para tarefas de classificação e regressão. Sua premissa é simples: dados pontos de entrada, o modelo classifica um novo exemplo com base nos K exemplos mais próximos do conjunto de treinamento, usando uma métrica de distância Euclidiana (M et al., 2022).

A Figura 7 apresenta uma visão geral da estrutura KNN.

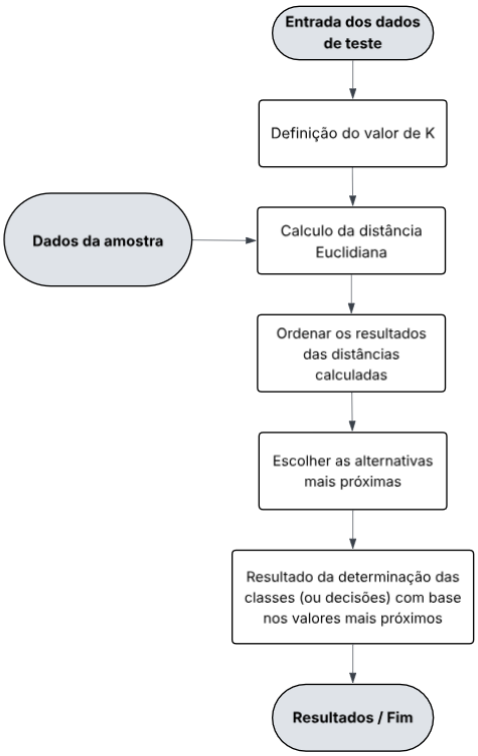


Figura 7 - Estrutura KNN (adaptado M et al., 2022).

3.2.7 Vantagens e desvantagens dos algoritmos considerados

Considerando os algoritmos de ML apresentados, seu uso e aplicação variam conforme a complexidade do problema e as características da base de dados de entrada, sendo que cada algoritmo possui vantagens e desvantagens específicas. A Tabela 4 apresenta uma visão geral comparativa dos algoritmos discutidos neste estudo, destacando suas principais vantagens e limitações, de forma a auxiliar na escolha do método mais adequado para diferentes contextos de análise.

A escolha do algoritmo a ser utilizado deve levar em conta as vantagens e desvantagens de cada método, e, principalmente, as características da base de dados disponível.

Tabela 4 - Vantagens e desvantagens os algoritmos ML.

ALGORITIMO	VANTAGENS	DESVANTAGENS	AUTORES
Naive Bayes (NB)	Estrutura simples e eficiente. Mantém bom desempenho mesmo com pequenos conjuntos de dados. Possui ampla aplicabilidade em diferentes tipos de problemas.	Apresenta baixa sensibilidade a variáveis pouco significativas.	Wickramasinghe & Kalutarage, 2020
Artificial Neural Networks (ANN)	Capaz de modelar relações complexas entre variáveis. Adapta-se bem às características específicas dos dados. Oferece bom desempenho em tarefas de previsão.	Requer grandes volumes de dados e tempo de treinamento em certos casos. Não torna evidente quais são as variáveis que influenciaram mais o resultado final.	Afework & Sipos, 2020 Tu, 1996
Support Vector Machine (SVM)	Lida adequadamente com dados complexos e não lineares. Pode ser ajustado para aprendizado incremental.	Não torna evidente quais são as variáveis que influenciaram mais o resultado final. Apresenta desempenho insatisfatório com dados desbalanceados.	Sun et al., 2017
Decision Tree (DT)	Fácil de entender e interpretar. Identifica automaticamente as variáveis mais relevantes. Proporciona boa precisão nos resultados.	Sensível a pequenas variações nos dados de entrada. Depende fortemente da estrutura da árvore para a extração de regras.	Abellán et al., 2013 Abdullah & Sipos, 2022 Nedjmedine & Tahar, 2022
Random Forests (RF)	Apresenta alto desempenho com grandes volumes de dados. Possui boa capacidade de generalização para dados fora da amostra.	Pode ser difícil interpretar individualmente as decisões tomadas em árvores profundas. Vulnerável ao desbalanceamento entre classes.	Manzoor et al., 2021 Xu & Huang, 2021
K-Nearest Neighbors (KNN)	Eficaz em contextos com grandes quantidades de dados. Estrutura simples.	Pode ter o desempenho comprometido quando os dados estão desbalanceados. Exige uma escolha criteriosa de parâmetros (como profundidade ou número de divisões).	Lv et al., 2009 M et al., 2022

3.3 Desequilíbrio do conjunto de dados

Quando se trabalha com dados em que uma classe está super-representada em relação às demais, estamos diante de um caso de desequilíbrio de dados. Esse cenário é comum em estudos envolvendo sinistralidade, nos quais, frequentemente, uma das classes de gravidade das lesões, normalmente os acidentes com feridos leves, possui uma quantidade de registos muito superior aos acidentes com feridos graves e vítimas mortais, caracterizando um desbalanceamento.

Diante desse problema, existem algumas abordagens específicas para lidar com o desequilíbrio, que geralmente seguem duas direções principais: os métodos sensíveis ao custo e os métodos de amostragem (sobreamostragem e subamostragem) (Haibo He & Garcia, 2009).

3.3.1 Métodos sensíveis ao custo

Os métodos sensíveis ao custo (*Cost-Sensitive Methods*) partem da ideia de associar diferentes custos à classificação incorreta das diferentes classes. Ou seja, consistem em ajustar os custos de erro atribuídos à classificação incorreta dos exemplos. Esses métodos utilizam uma matriz de custos que penaliza mais fortemente os erros cometidos sobre a classe minoritária (geralmente a mais crítica), incentivando o modelo a se concentrar mais nessa classe (Haibo He & Garcia, 2009). Dessa forma, o objetivo do aprendizado sensível ao custo não é apenas minimizar a taxa global de erro ou acurácia total, mas sim minimizar o custo total esperado dos erros de classificação.

Entre as principais vantagens desse método, destaca-se o fato de que não altera os dados originais, podendo ser aplicado em diversos algoritmos de aprendizado. Além disso, o "custo" associado a cada classe pode ser ajustado conforme a necessidade ou o contexto do problema, proporcionando maior flexibilidade ao modelo.

3.3.2 Métodos de amostragem

Os métodos de amostragem consistem na aplicação de técnicas como sobreamostragem e/ou subamostragem para balancear os conjuntos de dados, o que pode melhorar significativamente o desempenho dos modelos (Haibo He & Garcia, 2009).

A subamostragem aleatória (*Random Under-Sampling* – RUS) é uma técnica que busca reduzir o desequilíbrio na base de dados por meio da remoção aleatória de exemplos da classe majoritária, até que as classes fiquem equilibradas, ou seja, com o mesmo número de exemplos. Uma das principais desvantagens do RUS é que ele pode descartar informações importantes, prejudicando a representatividade da classe majoritária. Com isso, os modelos podem apresentar pior desempenho devido à perda de variabilidade (Leevy et al., 2018).

Já a sobreamostragem aleatória (*Random Over-Sampling* – ROS) consiste no aumento do número de casos da classe minoritária, por meio da replicação aleatória de exemplos existentes dessa classe. No entanto, a aplicação do ROS pode levar ao problema de *overfitting*, uma vez que a repetição de exemplos idênticos pode reforçar padrões específicos da amostra original, prejudicando a generalização do modelo (He & Garcia, 2009).

Uma alternativa para evitar o risco de *overfitting* e ainda equilibrar a base de dados é o uso da Técnica de Sobreamostragem de Minoria Sintética (*Synthetic Minority Over-sampling Technique* – SMOTE). Essa técnica consiste na criação de exemplos sintéticos a partir da distância entre os vizinhos mais próximos da classe minoritária, gerando novas amostras que mantêm a diversidade dos dados (Uttam & Sharma, 2021).

3.4 Parâmetros de validação

Existem diversos parâmetros estatísticos utilizados para validar a eficiência e a adequação de modelos de *Machine Learning* (ML). Neste estudo, foram considerados os seguintes indicadores: Coeficiente Kappa, Área sob a Curva ROC (ROC AUC) e a Matriz de Confusão.

O coeficiente Kappa (ou estatística Kappa) é uma medida utilizada para avaliar o grau de concordância entre avaliadores, especialmente quando as variáveis analisadas são categóricas. Introduzido por Cohen (1960), o Kappa quantifica o nível de concordância além do que seria esperado por acaso, oferecendo uma forma de avaliação mais robusta do que a simples percentagem de acerto.

Os valores do coeficiente variam entre -1 e 1, sendo que quanto mais próximo de 1, maior a concordância entre os resultados. A Tabela 5 apresenta os intervalos de valores e os respetivos níveis de interpretação atribuídos a cada faixa de Kappa.

Tabela 5 - Valor da estatística Kappa (adaptado de Landis e Koch 1977 apud Czodrowski, 2014).

Intervalo coeficiente Kappa	Correlação
< 0,0	Pobre
0,00 - 0,20	Leve
0,21 - 0,40	Razoavel
0,41 - 0,60	Moderada
0,61 - 0,80	Substancial
0,81 - 1,00	Quase perfeita
1,00	Perfeita

A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta gráfica utilizada para visualizar, organizar e selecionar classificadores com base em seu desempenho, ela é indicada principalmente para classificadores binários. Ela plota a taxa de verdadeiros positivos no eixo Y

contra a taxa de falsos positivos no eixo X, o gráfico ROC tem assim a função de representar as compensações relativas entre benefícios (verdadeiros positivos) e custos (falsos positivos) (Fawcett, 2006).

A área sob a curva ROC (AUC) é uma métrica derivada dessa curva que resume em um único número a capacidade do classificador discriminar entre as classes positiva e negativa. A AUC de um classificador é equivalente à probabilidade de que o classificador classifique uma instância positiva escolhida aleatoriamente em uma posição mais alta do que uma instância negativa escolhida aleatoriamente (Fawcett, 2006). De maneira geral, a área sob a curva ROC busca identificar bons modelos, com isso quanto mais próximo de 1 melhor o modelo.

A Tabela 6 apresentas os intervalos de valores e os respetivos níveis de interpretação atribuídos a cada faixa a área sob a curva ROC.

Tabela 6 - Valor da área sob a curva ROC (adaptado de Çorbacıoğlu & Aksel, 2023).

Intervalo AUC	Poder discriminativo do modelo
0,5 ≤ AUC ≤ 0,6	Sem poder discriminativo
0,6 ≤ AUC ≤ 0,7	Mau
0,7 ≤ AUC ≤ 0,8	Regular
0,8 ≤ AUC ≤ 0,9	Considerável
0,9 ≤ AUC	Excelente

A matriz de confusão fornece uma visualização simples dos erros cometidos pelo modelo, revelando possíveis interdependências entre as classes reais e previstas (Erbani et al., 2024). Essa matriz possui uma estrutura tabular, na qual se compara o rótulo de classe previsto com o rótulo de classe real para cada categoria de dados. Em uma configuração típica, as linhas da matriz representam os rótulos de classe reais, enquanto as colunas representam os rótulos previstos pelo modelo (Görtler et al., 2022). Um exemplo de matriz de confusão para uma variável com duas classes é apresentado na Figura 8.

Figura 8 - Matriz confusão para duas classes (Silva, 2023)

4. Estudo de caso

Este capítulo analisa a aplicação dos algoritmos de *Machine Learning* (ML) apresentados no capítulo anterior, para prever a gravidade de acidentes rodoviários, assim como os fatores que influenciam essa gravidade e o efeito de intervenções direcionadas à mitigação desses fatores. Os algoritmos utilizados serão: NB, ANN, SVM, DT, RF e KNN, todos aplicados por meio do software WEKA (Frank et al., 2017), utilizando uma base de dados oficial de acidentes do tipo colisão e despiste, com vítimas e ocorridos na rede nacional de estradas, registados em Portugal entre 2019 e 2023. A base de dados foi disponibilizada pela Autoridade Nacional de Segurança Rodoviária (ANSR).

O estudo dos aspetos mencionados permitirá delinear e validar os principais passos necessários à implementação de um Sistema de Apoio à Decisão baseado em técnicas de *Machine Learning*, voltado à análise e previsão da gravidade de acidentes rodoviários do tipo colisão e despiste.

4.1 Metodologia

O objetivo deste trabalho é avaliar a aplicabilidade do uso de ML na obtenção de modelos fiáveis de previsão da gravidade dos acidentes. Para isso, foram considerados diversos algoritmos com vista a identificar os que melhor traduzem o fenómeno estudado.

Esses algoritmos foram repetidamente testados com dados de diferentes categorias e tratamentos, conforme será detalhado no tópico 4.2 – Tratamento de Dados.

Apesar de a base de dados ser extensa, observou-se a necessidade de aplicar ferramentas do software WEKA (Frank et al., 2017) para lidar com o desequilíbrio das classes da variável dependente (gravidade do acidente), já que aproximadamente 90% dos registos correspondiam a acidentes com vítimas com ferimentos leves e cerca de 10% a acidentes com vítimas com ferimentos graves ou vítimas mortais. Dessa forma, foram realizados testes utilizando técnicas baseadas em métodos sensíveis ao custo, subamostragem (RUS) e SMOTE, com o intuito de alcançar os melhores resultados.

A Figura 9 apresenta a metodologia adotada.

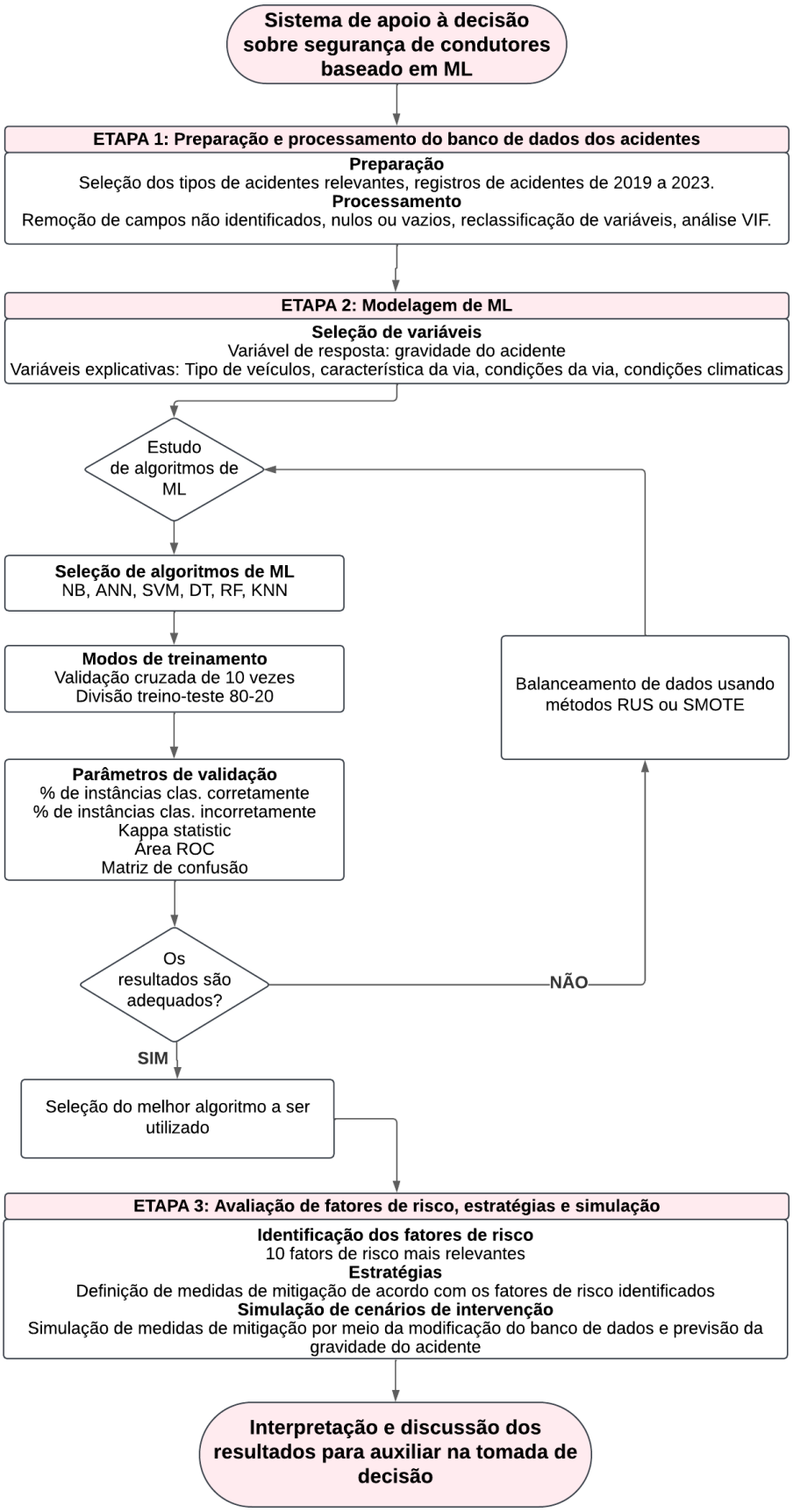


Figura 9 - Metodologia adotada no Sistema de apoio à decisão baseado em ML para análise de colisões e despistes.

O estudo foi desenvolvido em três etapas principais. A primeira etapa corresponde à preparação e processamento da base de dados, na qual foram realizadas as primeiras seleções a partir do banco original, restringindo-se apenas aos casos de acidentes do tipo colisão ou despiste, de acordo com os objetivos do estudo. Nessa fase, também foi feito o tratamento dos dados, com a remoção de campos irrelevantes e a análise do Fator de Inflação da Variância (VIF) para verificar a colinearidade entre as variáveis.

A segunda etapa refere-se à modelagem em *Machine Learning* (ML), em que foram desenvolvidos diversos modelos com o objetivo de identificar aqueles com melhor desempenho na previsão da variável dependente "gravidade dos acidentes". Nesta fase, foi realizado também o balanceamento da base de dados buscando otimizar os resultados obtidos pelos modelos.

A terceira e última etapa consiste na avaliação dos fatores de risco, definição de estratégias e simulações. Nessa fase, foram identificadas as variáveis com maior influência sobre acidentes de maior gravidade e definidas estratégias de intervenção, que foram posteriormente aplicadas em simulações. O objetivo foi verificar o potencial de redução nos acidentes graves por meio de ações direcionadas.

4.2 Preparação e processamento do banco de dados

4.2.1 Tratamento inicial dos dados

O tratamento dos dados pode ser considerado uma das etapas mais importantes deste estudo, tanto pela qualidade da base de dados utilizada nas modelagens quanto pela relevância em relação ao objetivo principal do trabalho.

Os dados disponibilizados pela Autoridade Nacional de Segurança Rodoviária (ANSR), correspondem aos acidentes notificados às autoridades competentes durante o período dos anos de 2019 a 2023. Esses registos seguem o modelo do Boletim Estatístico de Acidentes de Viação, modelo apresentado no anexo 1. Ao serem agrupados, totalizaram 166.538 acidentes de diferentes naturezas (colisão, despiste e atropelamento).

A fim de delimitar a área de interesse da pesquisa, foram selecionados apenas os acidentes ocorridos em vias classificadas como: Autoestrada (A), Itinerário Principal (IP), Itinerário Complementar (IC) e Estrada Nacional (EN), priorizando assim os sinistros ocorridos fora do ambiente urbano e, na maioria dos casos, com pouca influência do tráfego de peões. Outra decisão importante foi a seleção de apenas acidentes do tipo colisão ou despiste. Com esses dois critérios, a base foi reduzida para 25.635 registos de acidentes.

A base de dados passou ainda por uma etapa de limpeza, na qual foram excluídos registos com informações classificadas como “Não definido” em campos que poderiam influenciar nos resultados dos testes realizados, totalizando 25.416 casos.

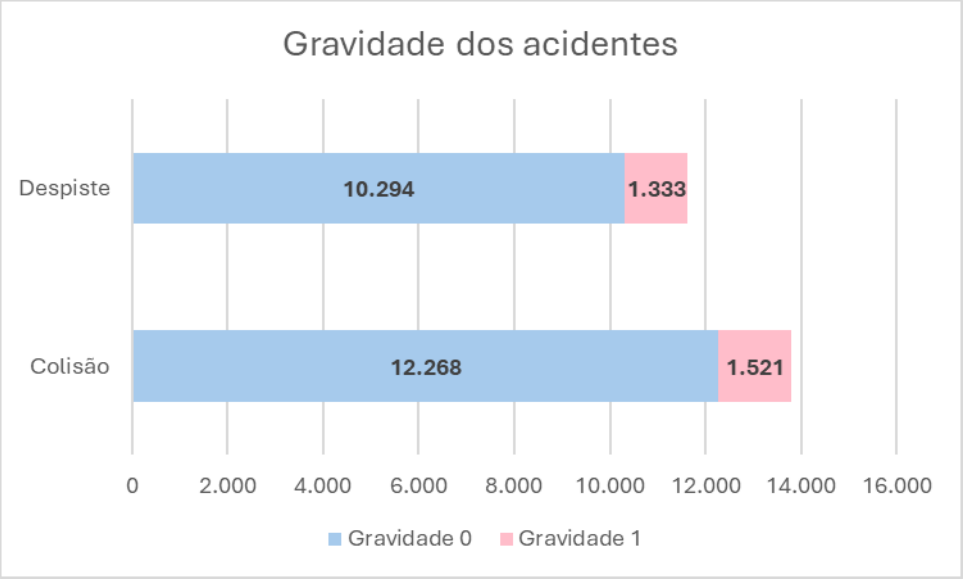
Foram removidas colunas com informações relacionadas com a localização (coordenadas, freguesia, concelho, indicação de ocorrência em Portugal continental, entre outras), data e hora, além de variáveis consideradas irrelevantes para o estudo ou que apresentavam uma quantidade significativa de campos preenchidos com a indicação "não definidos".

As informações sobre a quantidade e a gravidade das vítimas foram agrupadas com o objetivo de definir a severidade dos acidentes. Para isso, os casos de acidentes com vítimas mortais ou feridos graves a 30 dias foram classificados como Gravidade 1. Já os acidentes com apenas feridos leves a 30 dias foram classificados como Gravidade 0. Essa categorização da variável independente (variável de resposta) foi adotada considerando que o objetivo final do estudo é identificar formas de reduzir o número de casos com gravidade 1. A Tabela 7 e o Gráfico 4 apresentam o número de ocorrências e a respectiva classificação de severidade para os dois grupos de acidentes analisados.

Tabela 7 - Grupos e gravidade dos acidentes

	Gravidade 0	**Gravidade 1**	**Total**
Colisão	12.268	1.521	13.789
Despiste	10.294	1.333	11.627
Total	**22.562**	**2.854**	**25.416**

Gráfico 4 - Gravidade dos acidentes base de dados original.



4.2.2 Dados categóricos

A primeira abordagem de análise considerou o uso dos dados no formato original, categórico, utilizando as informações, após as manipulações descritas anteriormente, em uma forma próxima à original. Ou seja, após a exclusão de colunas e linhas que poderiam influenciar

negativamente os resultados, os dados restantes foram utilizados conforme disponibilizados pela ANSR.

Foram mantidas 27 variáveis independentes originais da base de dados, às quais foi acrescentada uma nova variável, a Gravidade, constituindo a variável independente. As variáveis consideradas na análise foram as seguintes:

- Dia da semana
- Localização
- Tipo natureza
- Número veículos ciclomotores
- Número veículos ligeiros
- Número de outros veículos
- Número veículos pesados
- Tipos de via – A, EN, IC, IP
- Traçado 1 – reta ou curva
- Traçado 2 – inclinação, patamar ou lombas
- Traçado 3 – situação da berma
- Traçado 4 – local: Via, berma
- Tipo de interseção
- Obras Arte
- Natureza
- Obstáculos e obras
- Sinais luminosos
- Regime de circulação
- Luminosidade
- Marca na via
- Condições de aderência
- Características Técnicas – autoestrada ou não
- Estado de conservação do pavimento
- Fatores atmosféricos
- Número de vias de trânsito por sentido
- Velocidade geral
- Velocidade local
- Gravidade

4.2.3 Dados binários

Com base nas 28 variáveis selecionadas, foi realizada uma recodificação das classes mais significativas das variáveis para o sistema binário, com o objetivo de simplificar a leitura pelo software, converter os dados num formato numérico para permitir a avaliação do Fator de Inflação da Variância – Variance Inflation Factor (VIF) e facilitar a interpretação dos resultados dos modelos de ML.

Após essa etapa, utilizou-se o software IBM SPSS Statistics para avaliar o VIF, ou seja, a colinearidade entre as variáveis independentes binárias. Os resultados finais da análise VIF estão apresentados nos Anexos II e III.

Uma vez que se pretende desenvolver modelos de previsão separados para cada tipo de natureza do acidente considerado, a análise foi realizada separadamente para os acidentes do tipo colisão e para os do tipo despiste, apresentando resultados distintos. As variáveis independentes identificadas como colineares foram removidas, com exceção da EN – Estrada Nacional consideradas pertinentes para a análise.

Com isso, as variáveis independentes binárias utilizadas nas análises para os acidentes do tipo colisão totalizaram 44, correspondendo às seguintes:

- Dia da semana
- Veículo Ciclomotor
- Veículos Ligeiro
- Outro veículo
- Veículo Pesado
- AE – Autoestrada
- EN – Estrada Nacional
- IC – Itinerário Complementar
- IP – Itinerário Principal
- Traçado em curva
- Traçado em patamar
- Berma não pavimentada
- Sem berma
- Traçado local
- Cruzamento
- Entroncamento
- Ramo de ligação - entrada
- Ramo de ligação – saída
- Rotunda
- Via de aceleração
- Via de desaceleração
- Obstáculo sinalizado
- Obstáculo não sinalizado
- Obstáculo insuficientemente sinalizado
- Sinais luminosos a funcionar
- Sinais luminosos desligados
- Sinais luminosos intermitentes
- Circulação nos dois sentidos
- Circulação reversível
- Aurora, crepúsculo
- Noite sem iluminação
- Noite com iluminação
- Sol encandeante
- Sinalização horizontal - sentido
- Sinalização horizontal - inexistente
- Pavimento com gelo, geada ou neve
- Pavimento molhado, húmido ou com água
- Pavimento em estado regular
- Pavimento em mau estado
- Chuva
- Nevoeiro
- 1 via
- Velocidade 90–99 km/h
- Velocidade 110–120 km/h
- Gravidade (variável dependente)

Já para os acidentes do tipo despiste foram consideradas 43 variáveis independentes, retirando apenas a categoria “Outros veículos”, das acima citadas para o caso das colisões.

4.2.4 Estatística descritiva e VIF

A seguir, são apresentadas, nas Tabelas 8 e 9, as análises descritivas das características dos acidentes, organizadas de acordo com as categorias originais fornecidas pela ANSR e as subcategorias utilizadas na base de dados reorganizada em formato binário. Para cada variável, são indicados as frequências relativas (percentuais de ocorrência) e os respetivos valores do VIF, que permitem avaliar a colinearidade entre os atributos.

Tabela 8 - Variáveis, análise descritiva e VIF (Parte 1/2).

		Colisão (13789)		Despiste (11627)	
		%	VIF	%	VIF
Dia da semana	Dia da semana	76%	1,04	67%	1,04
Gravidade do acidente	Mortos a 30 dias	3%	1,21	3%	1,55
	Feridos graves a 30 dias	9%	1,60	9%	2,78
	Feridos ligeiros a 30 dias	94%	1,83	91%	3,47
Tipos de veículo	Veículo Ciclomotor;	17%	1,20	17%	22,53
	Veículo Ligeiro	97%	1,15	77%	27,68
	Outros veículos	5%	1,14	3%	4,57
	Veículo Pesado	13%	1,13	4%	6,21
Tipo de via	AE – Autoestrada	36%	1599,95	34%	2611,30
	EN – Estrada Nacional	46%	4,26E+12	54%	4,86E+12
	IC – Itinerário Complementar	15%	1,71	9%	1,57
	IP – Itinerário Principal	3%	1,22	3%	1,25
Traçado 1 – reta ou curva	Traçado em curva	20%	1,13	47%	1,17
Traçado 2 – inclinação, patamar ou lombas	Traçado em patamar	72%	1,07	62%	1,12
Traçado 3 – situação da berma	Berma pavimentada	69%	-	57%	-
	Berma não pavimentada	10%	1,23	18%	1,66
	Sem berma	20%	1,21	25%	1,68
Traçado 4 – local: Via ou berma	Na via ou berma	100%	1,01	99%	1,01
Tipo de interseção	Cruzamento	3%	144,91	0%	1,02
	Entroncamento	5%	234,73	1%	1,02
	Ramo de ligação - entrada	0%	18,42	0%	1,04
	Ramo de ligação – saída	0%	5,71	0%	1,03
	Rotunda	1%	56,54	1%	1,10
	Via de aceleração	1%	59,78	1%	1,03
	Via de desaceleração	0%	23,09	1%	1,06
	Fora da intersecção	88%	490,04	95%	7,05E+11
Obstáculos e obras	Obstáculo sinalizado	2%	1,03	1%	1,03
	Obstáculo não sinalizado	0%	1,01	0%	1,03
	Obstáculo insuficientemente sinalizado	0%	1,01	0%	1,02
	Obstáculo inexistente	98%	-	99%	6,82E+12
Sinais luminosos	Sinais luminosos a funcionar	5%	1,08	2%	1,05
	Sinais luminosos desligados	2%	1,02	2%	1,01
	Sinais luminosos intermitentes	0%	1,01	0%	1,01
	Sinais luminosos inexistente	93%	2,36E+12	96%	6,82E+12
Regime de circulação	Circulação nos dois sentidos	47%	5,21	55%	2,93E+12
	Circulação reversível	0%	1,04	0%	1,01
	Circulação sentido único	53%	-2,21E+12	44%	6,05

Tabela 9 - Variáveis, análise descritiva e VIF (Parte 2/2).

		Colisão (13789)		Despiste (11627)	
		%	VIF	%	VIF
Luminosidade	Aurora, crepúsculo	4%	1,02	4%	1,02
	Noite sem iluminação	13%	1,10	18%	1,07
	Noite com iluminação	10%	1,08	8%	1,14
	Sol encandeante	1%	1,01	1%	1,01
	Pleno dia	71%	-5,80E+12	69%	-4,47E+12
Marcas na via	Sinalização horizontal - sentido e via	54%	-	46%	-
	Sinalização horizontal - sentido	36%	1,89	42%	2,26
	Sinalização horizontal - inexistente	9%	1,35	12%	1,64
Condições de aderência	Pavimento seco e limpo	83%	153,70	64%	19,06
	Pavimento com gelo, geada ou neve	0%	3,16	1%	1,65
	Pavimento molhado, húmido ou com água	17%	153,92	34%	21,00
Características Técnicas – autoestrada ou não	Autoestrada	36%	1602,85	34%	2610,89
	Estrada sem separador	46%	3,24E+12	56%	-
	Outras vias	17%	3,57	10%	2,85
Estado de conservação do pavimento	Pavimento bom	58%	-	59%	-
	Pavimento em estado regular	41%	1,03	40%	1,03
	Pavimento em mau estado	1%	1,03	1%	1,06
Fatores atmosféricos	Bom tempo	86%	58,85	70%	71,57
	Chuva	13%	56,16	29%	70,38
	Nevoeiro	1%	5,78	1%	5,30
Número de vias de trânsito por sentido	1 via	44%	119,51	55%	197,90
	2 via	25%	89,25	29%	162,85
	3 via	23%	84,54	14%	93,77
	4 via	6%	29,27	2%	15,93
	5 via	1%	5,70	0%	3,56
Velocidade local	Velocidade 50-89 km/h	40%	23,49	29%	9,15
	Velocidade 90–99 km/h	29%	20,71	41%	11,54
	Velocidade 110–120 km/h	30%	21,71	28%	9,60

4.3 Modelos

Foram desenvolvidos modelos utilizando a base de dados original, a base de dados original com aplicação de técnicas sensíveis ao custo, além das versões com subamostragem aleatória (RUS) e sobreamostragem (SMOTE). Cada modelo foi treinado com dois métodos distintos: validação cruzada com 10 subconjuntos (10-fold cross-validation) e divisão percentual de 80% dos dados utilizados para treinamento e 20% para teste (80/20 split). Esses procedimentos foram aplicados separadamente para cada tipo de acidente (colisão e despiste) e algoritmo considerado.

No total, foram realizados 64 testes com a base de dados de variáveis categóricas. Os testes com aplicação das técnicas RUS e SMOTE foram realizados apenas para os modelos que apresentaram os melhores desempenhos. conforme é apresentado na Tabela 10.

Tabela 10 - Modelos realizados para a base de dados de variáveis categóricas.

BASE DE DADOS CATEGÓRICA					
Algoritmo	Natureza do acidente	Base original	Técnicas sensíveis ao custo	RUS	SMOTE
Naive Bayes (NB)	Colisão	2	2	2	2
	Despiste	2	2	2	2
Artificial Neural Networks (ANN)	Colisão	2	2	2	2
	Despiste	2	2	2	2
Support Vector Machine (SVM)	Colisão	2	2	-	-
	Despiste	2	2	-	-
Decision Tree (DT)	Colisão	2	2	-	-
	Despiste	2	2	-	-
Random Forests (RF)	Colisão	2	2	-	-
	Despiste	2	2	-	-
K-Nearest Neighbors (KNN)	Colisão	2	2	-	-
	Despiste	2	2	-	-
Total de testes		24	24	8	8

O mesmo procedimento foi adotado para a base de dados das variáveis binárias, totalizando 80 testes. Assim como na base categórica, os testes com aplicação das técnicas de RUS e SMOTE foram realizados apenas para os modelos com melhor desempenho, conforme é apresentado na Tabela 11.

Tabela 11 - Modelos realizados para a base de dados de variáveis binárias.

BASE DE DADOS BINÁRIA					
Algoritmo	Natureza do acidente	Base original	Técnicas sensíveis ao custo	RUS	SMOTE
Naive Bayes (NB)	Colisão	2	2	2	2
	Despiste	2	2	2	2
Artificial Neural Networks (ANN)	Colisão	2	2	2	2
	Despiste	2	2	2	2
Support Vector Machine (SVM)	Colisão	2	2	-	-
	Despiste	2	2	-	-
Decision Tree (DT)	Colisão	2	2	2	2
	Despiste	2	2	2	2
Random Forests (RF)	Colisão	2	2	2	2
	Despiste	2	2	2	2
K-Nearest Neighbors (KNN)	Colisão	2	2	-	-
	Despiste	2	2	-	-
Total de testes		24	24	16	16

Com isso, nesta fase do estudo foram realizados 144 testes, considerando todas as variáveis como categóricas e como binárias.

4.4 Resultados e discussões

A seguir, serão apresentados os resultados dos testes realizados com a base de dados original para cada um dos algoritmos do ML considerados, tanto para os dados categóricos quanto para os dados binários.

Os resultados serão organizados de acordo com a natureza do sinistro — colisão ou despiste — e, em seguida, serão apresentados os resultados obtidos com os dados tratados utilizando a ferramenta SMOTE de sobreamostragem, por ter sido a técnica que apresentou os melhores desempenhos com vista a equilibrar as classes da variável de resposta. Os resultados das demais análises encontram-se disponíveis nos Anexos IV ao VII e nos Anexos XII ao XV.

4.4.1 Base de dados categórica

4.4.1.1 Acidentes do tipo colisão

A seguir, na Tabela 12, são apresentados os resultados obtidos para a base de dados categórica, considerando os acidentes do tipo colisão. Como mencionado anteriormente, essa base contém 13.789 registos.

Tabela 12 - Resultados dos algoritmos de *Machine Lerning* para base de dados categóricos em acidentes do tipo colisão.

Acidentes do tipo colisão									
Modelo	Teste	Inst. clas. corretamente	Inst. clas. incorretamente	*Kappa statistic*	Classe	Precisão	Área ROC	Matriz confusão	
		%	%					a	b
Naive Bayes (NB)	10 CV	80,4	19,6	0,183	0	0,916	0,696	10528	1740
					1	0,242	0,696	965	556
					Total	0,842	0,696		
	80/20	81,0	19,0	0,198	0	0,911	0,699	2120	317
					1	0,266	0,699	206	115
					Total	0,836	0,699		
Artificial Neural Networks (ANN)	10 CV	86,4	13,6	0,088	0	0,897	0,662	11741	527
					1	0,241	0,662	1354	167
					Total	0,824	0,662		
	80/20	83,8	16,2	0,158	0	0,901	0,664	2236	201
					1	0,269	0,664	247	74
					Total	0,827	0,664		
Support Vector Machine (SVM)	10 CV	89,0	11,0	0,003	0	0,890	0,501	12267	1
					1	0,750	0,501	1518	4
					Total	0,874	0,501		
	80/20	88,4	11,6	0,116	0	0,884	0,500	2437	0
					1	?	0,500	321	0
					Total	?	0,500		
Decision Tree (DT)	10 CV	89,0	11,0	0,000	0	0,890	0,500	12268	0
					1	?	0,500	1521	0
					Total	?	0,500		
	80/20	88,4	11,6	0,000	0	0,884	0,500	2437	0
					1	?	0,500	321	0
					Total	?	0,500		
Random Forests (RF)	10 CV	87,6	12,4	0,039	0	0,892	0,657	12016	252
					1	0,217	0,657	1451	70
					Total	0,818	0,657		
	80/20	87,5	12,5	0,036	0	0,886	0,662	2401	36
					1	0,250	0,662	309	12
					Total	0,812	0,662		
K-Nearest Neighbors (KNN)	10 CV	86,2	13,8	0,053	0	0,894	0,586	11761	507
					1	0,195	0,586	1398	123
					Total	0,817	0,586		
	80/20	86,4	13,6	0,057	0	0,888	0,601	2360	77
					1	0,230	0,601	298	23
					Total	0,811	0,601		

Por se tratar de uma base de dados extensa, mas com apenas 11% de acidentes na classe de gravidade 1 — conforme definido anteriormente acidentes com vítimas fatais ou feridos graves —, os algoritmos testados não obtiveram resultados satisfatórios. Todos os modelos apresentaram um coeficiente Kappa inferior a 0,200, indicando uma correlação fraca, e, em apenas um dos casos, a precisão para a classe 1 ultrapassou o valor de 0,300. Em relação à área ROC, todos os modelos apresentaram resultados inferiores a 0.699, indicando que não existe ou que existe um fraco poder discriminativos das classes de resposta da variável dependente (0,500<Área ROC≤0,700).

Os melhores resultados foram alcançados com o algoritmo *Naive Bayes*, no modelo que utilizou 80% dos dados para treinamento e 20% para teste. Nesse caso, o *Kappa Statistic* foi de 0,198, ainda assim inferior a 0,2 e conforme Landis e Koch 1977 apud Czodrowski, 2014, esse valor indica uma concordância leve entre os valores previstos e os observados. Por outro lado, esse mesmo modelo apresentou o melhor desempenho na área sob a curva ROC, com um valor de 0,699, o que permite classificá-lo como um modelo de desempenho mau.

Destaca-se, ainda, que os modelos *Decision Tree* e *Support Vector Machine* (com divisão 80/20 treino-teste) não apresentaram nenhuma precisão para os casos da classe 1, indicando que esses algoritmos falharam em identificar adequadamente os acidentes mais graves.

4.4.1.2 Acidentes do tipo despiste

A seguir, na tabela 13, são apresentados os resultados obtidos para os acidentes do tipo despiste, considerando a base de dados categórica. Esta base de dados contém 11.627 registos, como mencionado anteriormente.

Tabela 13 - Resultados dos algoritmos de *Machine Lerning* para base de dados categóricos em acidentes do tipo despiste.

Acidentes do tipo despiste									
Modelo	Teste	Inst. clas. corretamente	Inst. clas. incorretamente	*Kappa statistic*	Classe	Precisão	Área ROC	Matriz confusão	
		%	%					a	b
Naive Bayes (NB)	10 CV	86,0	14,0	0,084	0	0,892	0,643	9865	429
					1	0,246	0,643	1193	140
					Total	0,818	0,643		
	80/20	85,6	14,4	0,096	0	0,894	0,637	1958	101
					1	0,246	0,637	233	33
					Total	0,820	0,637		
Artificial Neural Networks (ANN)	10 CV	84,6	15,4	0,038	0	0,889	0,586	9718	576
					1	0,166	0,586	1218	115
					Total	0,806	0,586		
	80/20	86,2	13,8	0,027	0	0,888	0,588	1989	70
					1	0,167	0,588	252	14
					Total	0,805	0,588		
Support Vector Machine (SVM)	10 CV	88,5	11,5	0,000	0	0,885	0,500	10294	0
					1	?	0,500	1333	0
					Total	?	0,500		
	80/20	88,6	11,4	0,000	0	0,886	0,500	2059	0
					1	?	0,500	266	0
					Total	?	0,500		
Decision Tree (DT)	10 CV	88,5	11,5	0,000	0	0,885	0,500	10294	0
					1	?	0,500	1333	0
					Total	?	0,500		
	80/20	88,6	11,4	0,000	0	0,886	0,500	2059	0
					1	?	0,500	266	0
					Total	?	0,500		
Random Forests (RF)	10 CV	86,8	13,2	0,020	0	0,887	0,584	10037	257
					1	0,166	0,584	1282	51
					Total	0,804	0,584		
	80/20	86,5	13,5	0,023	0	0,887	0,573	1999	60
					1	0,167	0,573	254	12
					Total	0,805	0,573		
K-Nearest Neighbors (KNN)	10 CV	85,3	14,7	0,008	0	0,886	0,548	9857	437
					1	0,128	0,548	1269	64
					Total	0,799	0,548		
	80/20	85,2	14,8	0,006	0	0,886	0,541	1968	91
					1	0,125	0,541	253	13
					Total	0,799	0,541		

Neste conjunto de acidentes, observa-se que 11,5% dos casos são classificados com gravidade 1, valor bastante próximo ao encontrado na base de dados anterior (colisões), continuando assim, o desempenho dos modelos insatisfatório. Todos os algoritmos testados apresentaram coeficiente Kappa inferior a 0,200, o que indica correlação fraca e precisão para a classe 1 inferior a 0,300, sendo esta muito baixa. A área ROC não foi superior a 0,600 em nenhum dos casos, indicando, conforme Çorbacıoğlu & Aksel (2023), que não existe poder discriminativo para os modelos estudados.

O algoritmo *Naive Bayes*, com o modelo 80/20 treino-teste, apresentou o melhor desempenho para essa base de dados, com um *Kappa Statistic* de 0,0957 — indicando, portanto, leve concordância real entre os valores previstos e os valores observados. Os melhores resultados para a área sob a curva ROC foram de 0,643 (modelo 10 CV) e 0,637 (modelo 80/20), classificando o desempenho como mau.

Os testes com os algoritmos *Decision Tree* e *Support Vector Machine* falharam completamente na deteção dos casos da classe 1, apresentando *Kappa Statistic* igual a 0,000, nenhuma precisão para os casos de severidade 1, e desempenho geral do modelo igualmente insatisfatório.

4.4.2 Base de dados com variáveis binárias

4.4.2.1 Acidentes do tipo colisão

A seguir, na Tabela 14, são apresentados os resultados obtidos para a base de dados binários, considerando os acidentes do tipo colisão.

Tabela 14 - Resultados dos algoritmos de *Machine Lerning* para base de dados binária em acidentes do tipo colisão.

Acidentes do tipo colisão									
Modelo	Teste	Inst. clas. corretamente	Inst. clas. incorretamente	*Kappa statistic*	Classe	Precisão	Área ROC	Matriz confusão	
		%	%					a	b
Naive Bayes (NB)	10 CV	82,0	18,0	0,155	0	0,909	0,672	10873	1395
					1	0,236	0,672	1091	430
					Total	0,835	0,672		
	80/20	82,5	17,5	0,166	0	0,903	0,678	2188	249
					1	0,259	0,678	234	87
					Total	0,828	0,678		
Artificial Neural Networks (ANN)	10 CV	86,9	13,1	0,096	0	0,897	0,638	11820	448
					1	0,267	0,638	1358	163
					Total	0,827	0,638		
	80/20	86,4	13,6	0,088	0	0,890	0,605	2350	87
					1	0,269	0,605	289	32
					Total	0,818	0,605		
Support Vector Machine (SVM)	10 CV	89,0	11,0	0,000	0	0,890	0,500	12268	0
					1	?	0,500	1521	0
					Total	?	0,500		
	80/20	88,4	11,6	0,000	0	0,890	0,500	2437	0
					1	?	0,500	321	0
					Total	?	0,500		
Decision Tree (DT)	10 CV	89,0	11,0	0,000	0	0,890	0,500	12268	0
					1	?	0,500	1521	0
					Total	?	0,500		
	80/20	88,4	11,6	0,000	0	0,890	0,500	2437	0
					1	?	0,500	321	0
					Total	?	0,500		
Random Forests (RF)	10 CV	87,4	12,6	0,066	0	0,894	0,646	11950	318
					1	0,254	0,646	1413	108
					Total	0,824	0,646		
	80/20	86,8	13,2	0,073	0	0,889	0,639	2370	67
					1	0,272	0,639	296	25
					Total	0,817	0,639		
K-Nearest Neighbors (KNN)	10 CV	87,2	12,8	0,060	0	0,894	0,590	11914	354
					1	0,232	0,590	1414	107
					Total	0,821	0,590		
	80/20	86,657	13,343	0,077	0	0,889	0,587	2363	74
					1	0,267	0,587	294	27
					Total	0,817	0,587		

Conforme apresentado, o melhor desempenho foi do modelo *Naive Bayes* (80/20 treino-teste), embora os resultados indiquem apenas um desempenho mau a regular. A correlação entre os valores previstos e os valores reais foi nula, Mchugh (2012), com um índice Kappa de 0,165 e uma área sob a curva ROC de 0,678.

Já os modelos construídos com os algoritmos *Decision Tree* e *Support Vector Machine* classificaram todos os casos como pertencentes à classe 0 — acidentes com feridos leves — sem conseguir identificar os demais casos, apresentando assim um desempenho não satisfatório, com *Kappa Statistic* igual a 0,000 e área sob a curva ROC de 0,500.

4.4.2.2 Acidentes do tipo despiste

A seguir, na Tabela 15, são apresentados os resultados obtidos para a base de dados binários, considerando os acidentes do tipo despiste.

Tabela 15 - Resultados dos algoritmos de *Machine Lerning* para base de dados binária em acidentes do tipo despiste.

Acidentes do tipo despiste									
Modelo	Teste	Inst. clas. corretamente	Inst. clas. incorretamente	*Kappa statistic*	Classe	Precisão	Área ROC	Matriz confusão	
		%	%					a	b
Naive Bayes (NB)	10 CV	86,1	13,9	0,076	0	0,891	0,639	9889	405
					1	0,239	0,639	1206	127
					Total	0,816	0,639		
	80/20	85,6	14,4	0,076	0	0,892	0,639	1962	97
					1	0,224	0,639	238	28
					Total	0,815	0,639		
Artificial Neural Networks (ANN)	10 CV	86,1	13,9	0,028	0	0,887	0,581	9940	354
					1	0,169	0,581	1261	72
					Total	0,805	0,581		
	80/20	86,5	13,5	0,010	0	0,886	0,589	2003	56
					1	0,138	0,589	257	9
					Total	0,801	0,589		
Support Vector Machine (SVM)	10 CV	88,5	11,5	0,000	0	0,885	0,500	10294	0
					1	?	0,500	1333	0
					Total	?	0,500		
	80/20	88,6	11,4	0,000	0	0,886	0,500	2059	0
					1	?	0,500	266	0
					Total	?	0,500		
Decision Tree (DT)	10 CV	88,5	11,5	0,000	0	0,885	0,499	10294	0
					1	?	0,499	1333	0
					Total	?	0,499		
	80/20	88,6	11,4	0,000	0	0,886	0,500	2059	0
					1	?	0,500	266	0
					Total	?	0,500		
Random Forests (RF)	10 CV	86,7	13,3	0,007	0	0,886	0,572	10047	247
					1	0,133	0,572	1295	38
					Total	0,800	0,572		
	80/20	86,2	13,8	0,000	0	0,886	0,576	1997	62
					1	0,114	0,576	258	8
					Total	0,797	0,576		
K-Nearest Neighbors (KNN)	10 CV	86,3	13,7	0,002	0	0,885	0,549	9990	304
					1	0,119	0,549	1292	41
					Total	0,798	0,549		
	80/20	85,7	14,3	-0,009	0	0,885	0,557	1984	75
					1	0,096	0,557	258	8
					Total	0,795	0,557		

Assim como nos casos anteriores, o algoritmo *Naive Bayes* apresentou o melhor desempenho. A única diferença em relação aos resultados anteriores foi que, neste caso, o melhor resultado foi com o modelo obtido considerando um treino validação cruzada com 10 subconjuntos.

Os resultados mantêm o padrão observado anteriormente, com um coeficiente Kappa inferior a 0,200 (0,076) e área sob a curva ROC de 0,639, ou seja, correlação leve e poder discriminativo do modelo mau.

Os modelos baseados em *Decision Tree* e *Support Vector Machines* novamente não conseguiram identificar os casos da classe 1, classificando todos os registos como classe 0.

4.4.3 Análise geral

De forma geral, os resultados obtidos com as bases de dados categóricas variaram de maus a regulares. Nenhum dos modelos avaliados apresentou parâmetros de validação considerados satisfatórios, ou seja, coeficiente Kappa superior a 0,600 — que indica uma concordância modera a alta entre as classificações previstas e as reais — e área sob a curva ROC superior a 0,800, o que caracterizaria um modelo com poder discriminativo considerável.

Os modelos obtidos com o algoritmo *Naive Bayes* apresentaram os melhores resultados em todos os casos (colisão e despiste) (0,198< Kappa < 0,076 e 0,699< Área ROC <0,637). Ainda assim, o desempenho foi mau a regular e, de forma geral, os modelos não foram eficazes na identificação da classe menos representada, ou seja, os casos da classe de gravidade 1 (vítimas mortais e feridos graves).

Já os algoritmos *Decision Tree* e *Support Vector Machines* foram os que apresentaram os piores resultados, não conseguindo classificar corretamente os casos com vítimas mortais ou feridos graves. Os modelos identificam essencialmente a classe majoritária (gravidade 0), cuja quantidade é mais de oito vezes superior à da classe 1 (gravidade 1), o que evidencia a dificuldade desses algoritmos em lidar com bases de dados desbalanceadas.

Observou-se também que os modelos com divisão de 80% dos dados para treinamento e 20% para teste apresentaram resultados ligeiramente melhores aos obtidos com os modelos de validação cruzada com 10 subconjuntos.

Observa-se em todos os resultados apresentados que a matriz confusão demonstra que os modelos não conseguem identificar adequadamente os casos de acidentes apresentando maior gravidade. O acerto da classe 1 (acidentes com vítimas mortais ou feridos graves) é sempre baixo, mesmo para o modelo de melhor resultado, (inferior a 4%), sendo essa classe a que mais apresenta erros de classificação (superior a 76%).

Um dos principais fatores que podem ter contribuído para o desempenho insatisfatório dos modelos é o desequilíbrio da base de dados, uma vez que somente cerca de 11% dos registos pertencem à classe 1, em ambas as bases testadas. Como este percentual é baixo dificulta o aprendizado dos algoritmos em relação aos casos mais graves, limitando sua capacidade de previsão para essa classe. Tendo em conta este resultado, é estudado nas secções seguintes a aplicação das técnicas RUS e SMOTE com vista a diminuir o desequilíbrio das classes da variável de resposta (variável dependente).

4.4.4 Modelos com tratamento RUS e SMOTE

Conforme evidenciado anteriormente, devido à grande discrepância entre o número de casos com gravidade 0 e gravidade 1, foram aplicadas técnicas de sobreamostragem, SMOTE, e subamostragem aleatória, RUS, nas bases de dados disponíveis.

A aplicação da técnica RUS reduziu o número de casos dos acidentes do tipo colisão para 3.042, e dos acidentes do tipo despiste para 2.666 registos.

Já com o uso da técnica SMOTE foram adicionados novos casos sintéticos de forma a balancear a proporção entre as classes. Com isso, foram gerados seis vezes mais casos da classe de gravidade 1 em cada uma das bases de dados. Essa ampliação resultou em 21.394 casos no arquivo de acidentes por colisão e 18.292 casos no de acidentes por despiste.

As Tabelas 16 e 17 e os Gráficos 5 e 6 apresentam os resultados obtidos para as bases de dados dos acidentes do tipo colisão e despiste, respetivamente.

Tabela 16 - Comparação entre a base de dados original e as tratadas com as técnicas RUS e SMOTE para acidentes do tipo colisão.

Colisão			
	Gravidade 0	Gravidade 1	Total
RUS	1.521	1.521	3.042
Original	12.268	1.521	13.789
SMOTE	12.268	9.126	21.394

Gráfico 5 - Comparação entre a base de dados original e as tratadas com as técnicas RUS e SMOTE para acidentes do tipo colisão.

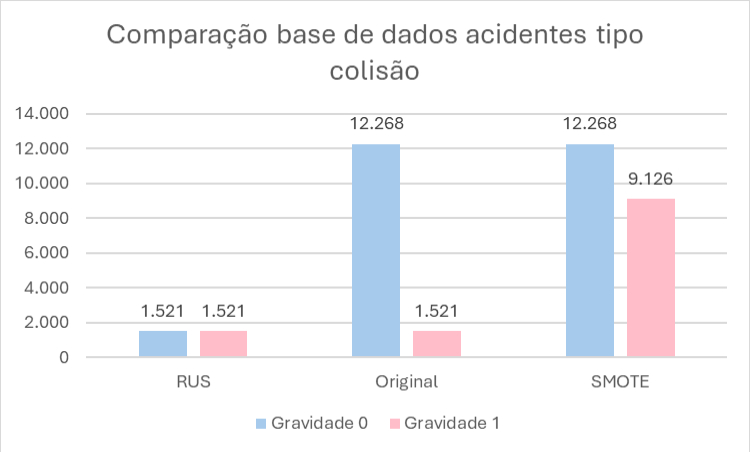
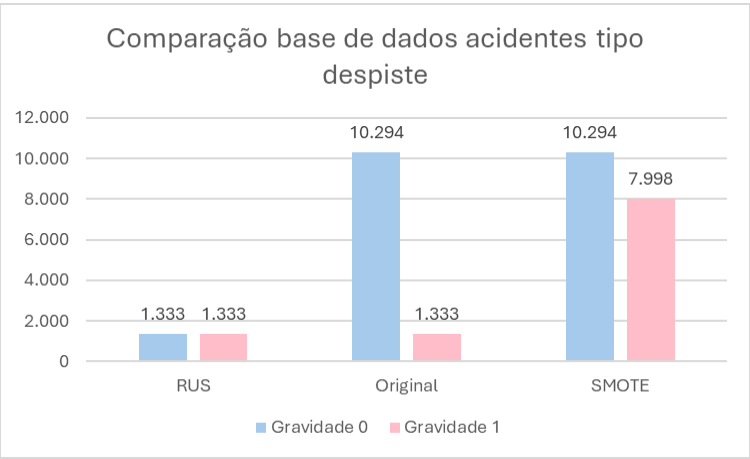


Tabela 17 - Comparação entre a base de dados original e as tratadas com as técnicas RUS e SMOTE para acidentes do tipo despiste.

	Despiste		
	Gravidade 0	Gravidade 1	Total
RUS	1.333	1.333	2.666
Original	10.294	1.333	11.627
SMOTE	10.294	7.998	18.292

Gráfico 6 - Comparação entre a base de dados original e as tratadas com as técnicas RUS e SMOTE para acidentes do tipo despiste.



A seguir, serão apresentados os resultados obtidos para as bases de dados com variáveis binárias, uma vez que estas demonstraram melhor desempenho em comparação às que

consideram variáveis do tipo categórico. Os resultados dos modelos desenvolvidos com a base de dados de variáveis categóricas encontram-se disponíveis nos Anexos VIII ao XI.

Foram selecionados para as etapas seguintes os algoritmos que apresentaram os melhores desempenhos nas análises iniciais, realizadas com a base de dados original. Os modelos escolhidos foram: *Naive Bayes, Artificial Neural Networks, Decision Tree* e *Random Forests*.

4.4.4.1 Acidentes do tipo colisão

A Tabela 18 apresenta um comparativo entre os resultados obtidos para os parâmetros instâncias corretamente classificadas, incorretamente classificadas, coeficiente Kappa e área sob a curva ROC, referentes aos modelos obtidos para a base de dados com variáveis binárias original dos acidentes do tipo colisão, bem como para os modelos obtidos a partir dos dados tratados com as ferramentas RUS e SMOTE.

Os demais resultados obtidos nas modelagens, assim como os valores dos restantes parâmetros analisados nesses testes, encontram-se disponíveis nos Anexo XVI e XVIII.

Os resultados apresentados referem-se aos testes realizados utilizando a técnica de validação cruzada com 10 subconjuntos (*10-fold cross-validation*), uma vez que esta abordagem demonstrou melhor desempenho.

Tabela 18 - Comparação entre os resultados obtidos com as diferentes bases de dados dos acidentes do tipo colisão.

	Base de dados	Inst. clas. corretamente	Inst. clas. incorretamente	Corret. clas. Gravidade 1	Kappa statistic	Área ROC
		%	%	%		
Naive Bayes (NB)	Original	82,0	18,0	28,3	0,155	0,672
	RUS	62,0	38,0	68,0	0,239	0,662
	SMOTE	67,1	32,9	64,9	0,334	0,724
Artificial Neural Networks (ANN)	Original	86,9	13,1	10,7	0,096	0,638
	RUS	59,1	40,9	58,7	0,183	0,627
	SMOTE	76,7	23,3	69,7	0,521	0,835
Decision Tree (DT)	Original	89,0	11,0	0,0	0,000	0,500
	RUS	62,6	37,4	61,5	0,253	0,642
	SMOTE	79,2	20,8	75,8	0,576	0,843
Random Forests (RF)	Original	87,4	12,6	7,1	0,066	0,646
	RUS	60,4	39,6	59,6	0,208	0,637
	SMOTE	80,8	19,2	76,9	0,607	0,886

Os melhores resultados foram obtidos nos testes com a base de dados submetida à sobreamostragem com recurso à técnica SMOTE. Já os testes com a base de dados tratada com a técnica de subamostragem (RUS) apresentaram desempenho ligeiramente superior ao da base

de dados original, mas ainda assim, nenhum dos modelos resultantes exibiu parâmetros considerados satisfatórios, isto é, *Kapppa* < 0,600 e Área ROC <0,800.

Como apresentado na Tabela 18, o melhor desempenho foi alcançado pelo modelo SMOTE executado com o algoritmo *Random Forest* e tipo de treino validação cruzada. A análise da estatística Kappa indicou uma concordância substancial entre os resultados previstos e os reais com Kappa = 0,607 e a área sob a curva ROC = 0,886, o que caracteriza um modelo com alto poder discriminativo entre as classes da variável de resposta. Além disso, o modelo conseguiu classificar corretamente mais de 80% dos mais de 21.000 casos e dos 9.126 casos classificados como gravidade 1 (feridos graves ou vítimas mortais), 7.021 foram corretamente identificados pelos modelos, o que corresponde a 76,9% de acerto para essa classe, demonstrando uma performance relevante na identificação dos acidentes de maior severidade.

4.4.4.2 Acidentes do tipo despiste

A Tabela 19 apresenta a comparação entre os resultados obtidos para os diferentes tratamentos da base de dados binária dos acidentes do tipo despiste, com testes realizados no formato 80/20 (treinamento/teste). Os demais resultados encontrados estão disponíveis nos Anexos XVII e XIV.

Tabela 19 - Comparação entre os resultados obtidos com as diferentes bases de dados dos acidentes do tipo despiste.

	Base de dados	Inst. clas. corretamente	Inst. clas. incorretamente	Corret. clas. Gravidade 1	Kappa statistic	Área ROC
		%	%	%		
Naive Bayes (NB)	Original	85,6	14,4	10,5	0,071	0,639
	RUS	61,5	38,5	67,3	0,229	0,664
	SMOTE	63,5	36,5	62,2	0,265	0,693
Artificial Neural Networks (ANN)	Original	86,5	13,5	3,4	0,010	0,589
	RUS	55,3	44,7	33,8	0,115	0,597
	SMOTE	73,9	26,1	66,7	0,465	0,822
Decision Tree (DT)	Original	88,5	11,5	0,0	0,000	0,500
	RUS	57,6	42,4	58,8	0,152	0,598
	SMOTE	77,2	22,8	79,5	0,534	0,828
Random Forests (RF)	Original	86,2	13,8	3,0	0,000	0,576
	RUS	51,6	48,4	51,1	0,032	0,543
	SMOTE	78,8	21,2	80,1	0,573	0,871

Assim como nos testes realizados com a base de dados dos acidentes do tipo colisão, o melhor resultado para os acidentes do tipo despiste também foram obtidos com a base submetida à sobreamostragem SMOTE. A base com subamostragem (RUS) apresentou uma leve melhoria

em relação à original, mas ainda assim, nenhum dos modelos resultantes apresentou desempenho totalmente satisfatório.

O melhor modelo, mais uma vez, foi o *Random Forest*, com Área ROC = 0,871, indicando alto poder discriminativo. No entanto, o coeficiente Kappa obtido demonstra que a concordância entre os resultados previstos e os reais é moderada (Kappa = 0,573). O modelo conseguiu classificar corretamente 1.231 dos 1.599 casos de gravidade 1, o que corresponde a mais de 80% de acerto na identificação dos acidentes com maior gravidade. De forma geral, obteve um desempenho expressivo, com 78,8% das instâncias corretamente classificadas, indicando boa capacidade preditiva e equilíbrio entre as classes analisadas.

4.4.4.3 Variáveis independentes explicativas mais significativas

Partindo da base de dados alterada com a técnica de sobreamostragem (SMOTE) e utilizando os modelos com melhor desempenho na avaliação da classe de gravidade dos acidentes (o *Random Forest* (RF) com treino *10-fold cross-validation* para os acidentes do tipo colisão e 80/20 para os despistes), foram obtidas as variáveis independentes mais significativas nos modelos obtidos através da aplicação da ferramenta *Attribute Selected Classifier* do programa WEKA. Estas variáveis são aquelas que apresentaram maior contribuição para a distinção entre os níveis de gravidade dos acidentes, sendo as listadas, por ordem de importância, na Tabela 20.

Tabela 20 - Variáveis independentes mais significativas.

N. de ordem	Colisão	Despiste
1	Circulação nos dois sentidos	Chuva
2	Uma via de trânsito	Pavimento molhado, húmido ou com água
3	Velocidade 90-99 km/h	Veículo ciclomotor
4	Auto estrada	Veículo ligeiro
5	Sinalização horizontal - delimitando sentido	Circulação em sentido único
6	Estrada Nacional	Noite sem iluminação
7	Veículo ciclomotor	Dia da semana
8	Pavimento molhado, húmido ou com água	Traçado em curva
9	Noite com iluminação	Uma via de trânsito
10	Veículo pesado	Sinalização horizontal inexistente

As variáveis mais significativas diferem entre os dois tipos de acidentes analisados, o que traduz as diferentes características associadas aos mesmos. Observa-se ainda que algumas variáveis não permitem uma abordagem do ponto de vista da engenharia aplicada às condições da infraestrutura, como aquelas relacionadas às condições climáticas, como por exemplo a variável Chuva. No entanto, outras variáveis identificadas são passíveis de intervenção ao nível das condições físicas e de circulação da infraestrutura, podendo ser melhoradas ou corrigidas com o

objetivo de reduzir a ocorrência de acidentes, como por exemplo a Velocidade e a Sinalização Horizontal da via (marcas rodoviárias).

Com base nas variáveis identificadas como mais influentes nos modelos, foram selecionadas duas relacionadas às condições físicas e de circulação da infraestrutura viária e duas associadas ao comportamento dos condutores, para cada tipo de acidente, com o objetivo de simular o impacto da sua alteração na redução da gravidade dos sinistros.

Para os acidentes do tipo colisão, as variáveis escolhidas foram: Velocidade 90–99 km/h, Noite com iluminação, Veículo Ciclomotor e Veículo Pesado. Já para os acidentes do tipo despiste, foram consideradas: Noite sem iluminação, Sinalização horizontal inexistente, Veículo Ciclomotor e Veículo Ligeiro.

Essas variáveis foram modificadas em simulações específicas, com o intuito de avaliar o potencial de mitigação da gravidade dos acidentes diante de diferentes cenários.

4.5 Cenários de simulação de medidas de intervenção

Com base nos dois modelos que apresentaram os melhores resultados — ou seja, *Decision Tree* e *Random Forest*, ambos aplicados à base de dados tratada com sobreamostragem (ROS) — e considerando as variáveis independentes mais significativas, foram realizadas simulações de medidas de intervenção na infraestrutura viária e no utente, para ambas as bases de dados (colisão e despiste).

Foram selecionadas duas variáveis passíveis de intervenção por meio da engenharia e duas relacionadas com o comportamento do utente da via. A partir disso, foram realizados testes considerando essas variáveis de forma isolada e também combinada.

Para os acidentes do tipo colisão, as medidas de intervenção de engenharia consideradas foram: redução da velocidade da via (variável velocidade 90-99 km/h) e melhoria das condições de iluminação da via durante o período noturno (variável noite com iluminação). As medidas com vista a alterar determinados comportamentos do utente foram: campanhas de formação e sensibilização dos condutores de veículos ciclomotores e veículos pesados (variáveis veículo ciclomotor e veículo pesado), com vista a reduzir a intervenção deste tipo de veículos nos acidentes.

Para os acidentes do tipo despiste, as medidas de intervenção selecionadas relacionadas com a infraestrutura viária (engenharia) foram: melhoria das condições de iluminação da via durante o período noturno (variável noite sem iluminação) e melhoria das marcas rodoviárias padrão (variável sinalização horizontal inexistente). As medidas com vista a alterar determinados comportamentos do utente foram: campanhas de formação e sensibilização dos condutores de veículos ciclomotores e veículos ligeiros (variáveis veículo ciclomotor e veículo ligeiro).

As simulações de implementação destas medidas foram realizadas considerando os acidentes de gravidade 1 das bases de dados dos acidentes por colisão e despiste originais (não sujeitas a sobreamostragem). As ferramentas utilizadas no programa WEKA para simular as medidas de intervenção foram *Supplied teste set* e o *Re-evaluate model on current test set*.

As seis simulações efetuadas consideraram as medidas de intervenção referidas anteriormente por alteração dos valores das variáveis de forma isolada ou combinada segundo o apresentado nas Tabelas 21 e 22

Tabela 21 - Cenários de simulações para os acidentes do tipo colisão.

Cenários de simulação - Colisão						
Cenário	1	2	3	4	5	6
Velocidade 90-99 km/h = 0	X		X			X
Iluminação noturna – Noite sem iluminação =0		X	X			X
Veículo ciclomotor = 0				X		X
Veículo pesado = 0					X	X

Tabela 22 - Cenários de simulações para os acidentes do tipo despiste.

Cenários de simulação - Despiste						
Cenário	1	2	3	4	5	6
Sinalização horizontal inexistente = 0	X		X			X
Iluminação noturna – Noite sem iluminação = 0		X	X			X
Veículo ciclomotor = 0				X		X
Veículo ligeiro = 0					X	X

Para cada uma das variáveis selecionadas, foram realizadas quatro simulações, utilizando os os modelos obtidos com os algoritmos *Decision Tree* (DT) e *Random Forest* (RF) para a base de dados tratada com a técnica ROS (SMOTE), aplicados tanto nos modelos treinados com *10-fold cross-validation* quanto opara os 80/20 *split*. Ao todo, foram realizadas 144 análises de simulação.

Durante as análises, observou-se que os resultados obtidos com os modelos obtidos com o mesmo algoritmo, independentemente do tipo de treinamento utilizado (validação cruzada com 10 subconjuntos ou 80/20 treino/teste), foram iguais.

Dessa forma, a seguir serão apresentados os resultados organizados por modelo/algoritmo.

4.5.1 Acidentes do tipo colisão

As Tabelas 23 e 24 apresentam os resultados obtidos nas análises de simulação realizadas com os modelos obtidos com os algoritmos *Decision Tree* (DT) e *Random Forest* (RF), respetivamente, para os acidentes do tipo colisão.

Tabela 23 - Resultado dos cenários de simulação para os modelos obtidos com o algoritmo *Decision Tree* - Colisão.

Decision Tree - Colisão						
		Previsão do N. e % de acidentes com Gravidade 0		Previsão do n. e % de acidentes com Gravidade 1		Probabilidade média da previsão
Cenário	Medida de intervenção	N.	%	N.	%	
0	Sem intervenção	1125	74	396	26	0,665
1	Velocidade 90-99 km/h	1125	74	396	26	0,680
2	Noite sem iluminação	1116	73	405	27	0,662
3	Velocidade 90-99 km/h + Noite sem iluminação	1116	73	405	27	0,685
4	Veículo ciclomotor	1308	86	213	14	0,704
5	Veículo pesados	1278	84	243	16	0,702
6	Todos	1478	97	43	3	0,737

Tabela 24 - Resultado dos cenários de simulação para os modelos obtidos com o algoritmo *Random Forest* – Colisão.

Random Forest - Colisão						
		Previsão do N. e % de acidentes com Gravidade 0		Previsão do n. e % de acidentes com Gravidade 1		Probabilidade média da previsão
Cenário	Medida de intervenção	N.	%	N.	%	
0	Sem intervenção	1228	81	293	19	0,656
1	Velocidade 90-99 km/h	1228	81	293	19	0,664
2	Noite sem iluminação	1243	82	278	18	0,671
3	Velocidade 90-99 km/h + Noite sem iluminação	1243	82	278	18	0,673
4	Veículo ciclomotor	1404	92	117	8	0,693
5	Veículo pesados	1341	88	180	12	0,690
6	Todos	1513	99	8	1	0,731

De maneira geral, observa-se que os cenários de simulação realizados com o modelo RF apresentaram resultados superiores aos apresentados pela simulação com o modelo DT, em relação à redução dos acidentes da classe de gravidade 1. O melhor resultado apresenta uma redução de 99% nos acidentes da classe de gravidade 1 (com vítimas mortais e feridos graves), correspondendo a uma intervenção conjunta ao nível da infraestrutura viária (velocidade e iluminação) e dos condutores de veículos ciclomotores e pesados. Contudo, a redução significativa observada no cenário sem intervenção sugere que a abordagem de simulação deve

ser revista e ajustada. Considerando o diferencial, verifica-se uma redução de 18% (99-81) nos acidentes da classe de gravidade 1. Verifica-se ainda que o impacto das medidas relacionadas com a formação e sensibilização dos condutores apresentam um impacto mais significativo na redução deste tipo de acidentes (7% e 11% considerando o diferencial com o cenário sem intervenção), quando comparado com o impacto das medidas direcionadas para a infraestrutura.

Também no modelo DT, as alterações na velocidade da via e na iluminação noturna indicaram um impacto menos significativo na diminuição de acidentes com vítimas mortais ou feridos graves, quando comparadas com as medidas direcionadas aos condutores de veículos ciclomotores e pesados. De igual modo, a conjugação de todas as medidas propostas apresenta melhores resultados, com uma redução de 25% (99-74).

A coluna de probabilidade média de previsão mostra que, as simulações envolvendo as variáveis de engenharia (velocidade e iluminação da via) e as relacionadas com os condutores de ciclomotores e pesados apresentaram entre 0.66 e 0.74 de acerto, com 18% a 35% no intervalo de probabilidade 0,75 – 1,00, o que, juntamente aos resultados obtidos evidencia que a conjugação de medidas pode contribuir efetivamente para a redução da gravidade dos acidentes do tipo colisão. O modelo RF, para intervenções direcionadas aos condutores de veículos ciclomotores e pesados, apresentou 25% a 28% dos casos no intervalo de probabilidade 0,75 – 1,00, evidenciando que um reforço na formação e sensibilização dos condutores pode contribuir significativamente para a diminuição da gravidade deste tipo de acidentes.

4.5.2 Acidentes do tipo despiste

As Tabelas 25 e 26 apresentam os resultados obtidos nas análises de simulação realizadas com os modelos obtidos com os algoritmos *Decision Tree* (DT) e *Random Forest* (RF), respetivamente, para os acidentes do tipo despiste.

Tabela 25 - Resultado dos cenários de simulação para os modelos obtidos com o algoritmo *Decision Tree* – Despiste.

Decision Tree - despiste						
		Previsão do N. e % de acidentes com Gravidade 0		Previsão do n. e % de acidentes com Gravidade 1		Probabilidade média da previsão
Cenário	Medida de intervenção	N.	%	N.	%	
0	Sem intervenção	934	70	399	30	0,696
1	Sinalização horizontal inexistente	928	70	405	30	0,696
2	Noite sem iluminação	1006	75	327	25	0,702
3	Sinalização horizontal – inexistente + noite sem iluminação	1000	75	333	25	0,702
4	Veículo ciclomotor	1078	81	255	19	0,737
5	Veículo ligeiro	990	74	343	26	0,739
6	Todos	1084	81	249	19	0,768

Tabela 26 - Resultado dos cenários de simulação para os modelos obtidos com o algoritmo *Random Forest* – Despiste.

Random Forest - despiste						
		Previsão do N. e % de acidentes com Gravidade 0		Previsão do n. e % de acidentes com Gravidade 1		Probabilidade média da previsão
Cenário	Medida de intervenção	N.	%	N.	%	
0	Sem intervenção	1109	83	224	17	0,648
1	Sinalização horizontal inexistente	1083	81	250	19	0,647
2	Noite sem iluminação	1109	83	224	17	0,648
3	Sinalização horizontal – inexistente + noite sem iluminação	1182	89	151	11	0,662
4	Veículo ciclomotor	1109	83	224	17	0,648
5	Veículo ligeiro	1084	81	249	19	0,675
6	Todos	1163	87	170	13	0,699

Os resultados obtidos nas simulações para este tipo de acidente foram, assim como apresentado para os acidentes do tipo colisão, melhores com o algoritmo *Random Forest* (RF). No entanto, também se verificara uma percentagem elevada de diminuição de acidentes da gravidade 1, para o cenário sem intervenção (83%).

Considerando o diferencial com o cenário sem intervenção, nos dois modelos (RF e DT), as medidas de intervenção direcionadas à infraestrutura viária (marcas rodoviárias e iluminação) apresentam um impacto maior na redução dos acidentes da classe 1 quando comparado com o verificado para os acidentes de colisão, apresentando reduções de 6% (89-83) para o modelo RF e de 5% (75-70) para o modelo DT. Para este tipo de acidentes as simulações sugerem que as medidas relacionadas com a infraestrutura viária, ou a conjugação de todas as medidas têm

maior impacto na redução dos acidentes de gravidade 1 (11% no modelo DT e 4% no modelo RF).

Para este tipo de acidentes, as medidas direcionadas aos condutores de veículos ciclomotores e ligeiros apresentam um impacto semelhante ou inferior às medidas relacionadas com a alterações da infraestrutura.

De forma geral, a probabilidade de previsão ficou entre 0.65 e 0.77, sendo que o modelo DT apresentara uma média ligeiramente superior de previsão em comparação ao modelos RF. O modelo com DT, para as intervenções de engenharia apresentou 18% a 38% dos casos no intervalo de probabilidade 0,75 – 1,00, evidenciando que modificações na infraestrutura viária podem contribuir significativamente para a diminuição da gravidade dos acidentes do tipo despiste. Tendo em conta o cenário envolvendo todas as medidas consideradas, esta percentagem é de 41% a 72%

Apesar de os modelos apresentarem uma acurácia de aproximadamente 75% na classificação dos casos da classe 1 e uma área sob a curva ROC superior a 0,800, observa-se que, ao serem utilizados para simulação com os casos reais da classe 1, no cenário sem intervenção, apresentam reduções da classe de gravidade 1 de 70% e 83%. Esse resultado evidencia a necessidade de se explorar abordagens de simulação mais adequadas em estudos futuros.

5. Conclusões e trabalhos futuros

É de conhecimento geral que os acidentes rodoviários podem ocorrer por diversos fatores, e que a sua gravidade não pode ser prevista com exatidão, uma vez que não segue um padrão único (fenómeno aleatório). No entanto, com o uso de abordagens baseadas em *Machine Learning* (ML), é possível identificar padrões complexos e os fatores com maior influência na gravidade dos acidentes e, a partir disso, dar suporte a estratégias de mitigação com vista a diminuir o número e a gravidade dos acidentes rodoviários. Ainda que essas medidas não eliminem completamente a ocorrência de acidentes, podem contribuir significativamente para que os sinistros apresentem menor gravidade.

O estudo sobre quais algoritmos de ML utilizar na análise de dados de sinistralidade e sobre a forma de estruturação da base de dados contendo as características principais dos acidentes, é de extrema importância para a obtenção de resultados mais precisos e confiáveis, especialmente diante da ampla variedade de algoritmos disponíveis. A escolha adequada desses elementos impacta diretamente no desempenho dos modelos e na qualidade das previsões geradas.

Entre os testes realizados com as bases de dados originais (colisões e despistes), tanto considerando as variáveis no formato categórico quanto binário, e para os seis algoritmos selecionados, observa-se que os modelos obtidos não foram satisfatórios, apresentando parâmetros de validação com valores baixos. Dessa forma, a capacidade de classificação e precisão dos modelos para detetar casos de gravidade 1, ou seja, acidentes com vítimas mortas ou feridos graves, não foi adequada.

Nos testes iniciais com as bases de dados originais, o melhor desempenho foi obtido com a base de dados com variáveis categóricas, utilizando o algoritmo *Naive Bayes*, no modelo com treinamento considerando a distribuição 80% treino e 20% teste. Constatou-se que, para ambas as bases de dados, os algoritmos *Support Vector Machine* (SVM) e *Decision Tree* (DT) não apresentaram validade, pois não foram capazes de identificar casos de gravidade 1.

Esses resultados estão muito provavelmente relacionados com o desequilíbrio da base de dados em relação às classes da variável de resposta, sendo que 89% dos casos pertencem à classe de gravidade 0 e apenas 11% à classe 1. Essa hipótese foi confirmada com a aplicação de técnicas de sobreamostragem da classe menos representada (classe 1), que resultaram em melhora significativa no desempenho dos modelos.

Os testes realizados com as bases de dados modificadas pelas técnicas de sobreamostragem (ROS) e subamostragem (RUS) apresentaram melhorias significativa nos resultados, devido ao maior equilíbrio entre as classes de gravidade da variável de resposta (gravidade do acidente), seja pelo acréscimo de casos da classe 1 (sobreamostragem), seja pela redução dos casos da classe 0 (subamostragem).

Os testes com sobreamostragem apresentaram os melhores desempenhos em comparação com os demais, sendo que os modelos obtidos com o algoritmo *Random Forest* (RF) foram os que apresentaram os melhores resultados em ambas as bases de dados.

O maior ganho de desempenho foi observado nos casos dos acidentes do tipo colisão, no modelo obtido com o algoritmo RF utilizando um treinamento de *10-fold cross-validation*, que obteve um coeficiente Kappa de 0,607, mais de nove vezes superior ao valor obtido com a base original (0,0658). Apresentou ainda uma área sob a curva ROC de 0,886, indicando alto poder discriminativo do modelo em relação às classes da variável de resposta.

Para os acidentes do tipo despiste, o melhor resultado foi obtido com o modelo *Random Forest* (RF) utilizando o formato de teste 80/20 (treinamento/teste), alcançando um coeficiente Kappa de 0,573, o que indica uma correlação moderada entre os resultados previstos e os reais, e uma área sob a curva ROC de 0,871, evidenciando o alto poder discriminativo do modelo.

Os fatores que influenciam a gravidade dos acidentes rodoviários podem ser agrupados em diferentes categorias, como fatores climáticos, de infraestrutura viária e relacionados ao condutor. Neste estudo, foram analisados os principais fatores ligados à infraestrutura viária, sendo observado que, para ambos os grupos de acidentes, a iluminação noturna da via destaca-se como um dos fatores mais relevantes. Além disso, a velocidade da via mostrou-se significativa nos acidentes do tipo colisão, enquanto a ausência de sinalização horizontal foi um dos principais fatores nos acidentes do tipo despiste. Quanto ao tipo de veículo interveniente no acidente, os ciclomotores foram identificados como fator de risco nos dois tipos de acidentes analisados, enquanto os veículos pesados foram mais comuns em colisões e os veículos ligeiros em acidentes por despiste.

Quanto aos cenários de simulação realizados, observou-se que modificações como a implementação/melhoria da sinalização horizontal ou o acréscimo de iluminação no período noturno, podem impactar significativamente a gravidade dos acidentes, sobretudo para os acidentes do tipo despiste. O comportamento dos condutores também pode contribuir significativamente para a redução da gravidade dos acidentes, em particular para os acidentes do tipo colisão envolvendo veículos ciclomotores e pesados.

No entanto, nas análises efetuadas, a consideração conjunta de medidas direcionadas à mudança de comportamento dos condutores, em conjunto com medidas de intervenção na infraestrutura são as mais promissoras no combate da sinistralidade com consequências graves. Contudo, de modo geral, torna-se evidente a necessidade de investigar abordagens de simulação mais apropriadas para aprimorar estudos futuros.

Destaca-se ainda que os resultados obtidos neste estudo, utilizando técnicas de *machine learning* para a construção de modelos preditivos da gravidade dos acidentes, representam uma contribuição valiosa para o estudo da segurança rodoviária. Eles fornecem orientações concretas

para investigações futuras mais aprofundadas sobre os diferentes tipos de acidentes, em especial despistes e colisões.

Por conseguinte, pode concluir-se que a abordagem baseada em ML (aprendizagem automática) é eficaz no apoio à tomada de decisões relativas à melhoria da segurança rodoviária para combater as consequências mais graves dos acidentes do tipo colisão e despiste, podendo ser replicada a outras realidades e alargada a outros tipos de acidentes.

A principal limitação observada ao longo do estudo foi o desequilíbrio entre as classes da variável dependente gravidade do acidente, com uma quantidade significativamente menor de casos na classe de gravidade 1 em relação à classe 0. Embora essa limitação possa ser vencida por meio das ferramentas de balanceamento existentes nos programas usados que permitem a obtenção de modelos de previsão com base em ML, os dados gerados aleatoriamente influenciam diretamente nos resultados obtidos.

Esse impacto é ainda mais relevante se considerarmos que, para que os modelos alcançassem parâmetros satisfatórios, foi necessário um acréscimo de 600% nos casos da classe de gravidade 1. Isso pode levantar questões quanto à representatividade e confiabilidade dos dados sintéticos gerados durante a sobreamostragem, pelo que devem ser exploradas abordagens alternativas para contornar o desequilíbrio das bases de dados. De entre as abordagens existentes sugere-se o estudo das baseadas na deteção de *outliers*, o ajuste do limiar de decisão para um valor diferente de 0,5 na previsão das probabilidades, ou a utilização de métodos adaptados ao desequilíbrio das bases de dados, como os algoritmos *XGBoost* e *LightGBM*.

Por outro lado, considerando que a base de dados disponibilizada pela ANSR inclui informações como a localização exata dos acidentes, é possível realizar estudos direcionados para identificar os pontos com maior concentração de sinistros com gravidade 1 (análise de *hotspots*), possibilitando o estudo e a execução de intervenções específicas e mais eficazes nesses locais.

6. Referências bibliográficas

Abdullah, P., & Sipos, T. (2022). Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method. Sustainability, 14(18), 11339. https://doi.org/10.3390/su141811339

Abellán, J., López, G., & de Oña, J. (2013). Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40(15), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027

Afework, A., & Sipos, T. (2020). Modelling of accidents for four lane non-urban highways using artificial neural networks technique. In 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE. https://doi.org/10.1109/saci49304.2020.9118819

ANRS (2021). Visão Zero 2030 - Home (s.d.). Visão Zero 2030. https://visaozero2030.pt/ Acesso em: 17 de maio de 2025

ANRS (2021). Visão Zero 2030 - Metodologia (s.d.). Visão Zero 2030. https://visaozero2030.pt/metodologia/ Acesso em: 17 de maio de 2025

ANRS (2021). Visão Zero 2030 - Sistema Seguro (s.d.). Visão Zero 2030. https://visaozero2030.pt/sistema-seguro/ Acesso em: 17 de maio de 2025

ANSR (Autoridade Nacional De Segurança Rodoviária) (2025). Relatório Outubro 2024. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Documents /2024/relatorio_de_sinistralidade_24h_e_fiscalizacao_rodoviaria_outubro_2024_vf.pdf

Atwah, A., & Al-Mousa, A. (2021). Car Accident Severity Classification Using Machine Learning. In 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). IEEE. https://doi.org/10.1109/3ict53449.2021.9581646

Çorbacıoğlu, Ş., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turkish Journal of Emergency Medicine, 23(4), 195. https://doi.org/10.4103/tjem.tjem_182_23

Erbani, J., Portier, P.-É., Egyed-Zsigmond, E., & Nurbakova, D. (2024). Confusion Matrices: A Unified Theory. IEEE Access, 1 https://doi.org/10.1109/access.2024.3507199

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Goniewicz, K., Goniewicz, M., Pawłowski, W., & Fiedor, P. (2015). Road accident rates: strategies and programmes for improving road traffic safety. European Journal of Trauma and Emergency Surgery, 42(4), 433–438. https://doi.org/10.1007/s00068-015-0544-6

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., & Patel, K. (2022). Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels. In CHI '22: CHI Conference on Human Factors in Computing Systems. ACM. https://doi.org/10.1145/3491102.3501823

Haibo He & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Hala, H., Anass, C., Rajaa, B., Youssef, B., & Garza-Reyes, J. (2021). Machine learning techniques for forecasting the traffic accident severity. In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA). IEEE. https://doi.org/10.1109/icdata52997.2021.00018

Lantz, B. Machine learning with R : learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications. Birmingham, Uk: Packt Publishing, 2013.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0151-6

Lv, Y., Tang, S., & Zhao, H. (2009). Real-Time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method. In 2009 International Conference on Measuring Technology and Mechatronics Automation. IEEE. https://doi.org/10.1109/icmtma.2009.657

M, A., K, A., K, A., M, A., & R, C. K. (2022). Accident Prediction Using KNN Algorithm. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). IEEE. https://doi.org/10.1109/icerect56837.2022.10059746

Mandrekar, J. N. (2011). Measures of Interrater Agreement. Journal of Thoracic Oncology, 6(1), 6–7. https://doi.org/10.1097/jto.0b013e318200f983

Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A., & Bisogni, C. (2021). RFCNN: Traffic Accident Severity Prediction based on Decision Level Fusion of Machine and Deep Learning Model. IEEE Access, 1. https://doi.org/10.1109/access.2021.3112546

Nedjmedine, O., & Tahar, M. (2022). Analysis of road accident factors using Decision Tree Algorithm: a case of study Algeria. In 2022 5th International Symposium on Informatics and its Applications (ISIA). IEEE. https://doi.org/10.1109/isia55826.2022.9993530

Organização Mundial da Saúde. Plano Global para a Década de Ação pela Segurança no Trânsito 2021–2030. OMS & Comissões Regionais da ONU; 2021. https://cdn.who.int/media/docs/default-source/documents/health-topics/road-traffic-injuries/global-plan-for-the-doa-of-road-safety-2021-2030-pt.pdf?sfvrsn=65cf34c8_35&download=true

P. E (2021). Parlamento Europeu: Quadro estratégico da UE em matéria de segurança rodoviária para o período 2021-2030 - Recomendações para as próximas etapas da campanha Visão Zero. https://www.europarl.europa.eu/doceo/document/TA-9-2021-0407_PT.html

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Silva, L. W. V. d. C. (2023). Análise de Sentimentos: Impacto da Tradução Neural na Avaliação de Desempenho. Monografia apresentada ao curso Ciência de dados e Inteligência. Universidade Federal da Paraíba. https://repositorio.ufpb.br/jspui/bitstream/123456789/31632/1/Lincoln%20Wallace%20Valentim%20da%20Costa%20Silva-TCC.pdf

Sun, P., Guo, G., & Yu, R. (2017). Traffic crash prediction based on incremental learning algorithm. In 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). IEEE. https://doi.org/10.1109/icbda.2017.8078803

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology, 49(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9

Uttam, A. K., & Sharma, G. (2021). A Comparison of Data Balancing Techniques for Credit Card Fraud Detection using Neural Network. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE. https://doi.org/10.1109/i-smac52330.2021.9640911

WHO (2023) Global status report on road safety 2023. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO. https://iris.who.int/bitstream/handle/10665/375016/9789240086517-eng.pdf?sequence=1

Wickramasinghe, I., & Kalutarage, H. (2020). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing. https://doi.org/10.1007/s00500-020-05297-6

Xu, H., & Huang, G. (2021). Analysis on Influencing Factors of accidents at national and provincial road intersections based on random forest. In 2021 6th International Conference on Transportation Information and Safety (ICTIS). IEEE. https://doi.org/10.1109/ictis54573.2021.9798574

Yao, X., Liu, Y. (2005). Machine Learning. In: Burke, E.K., Kendall, G. (eds) Search Methodologies. Springer, Boston, MA. https://doi.org/10.1007/0-387-28356-0_12

Zhang, H. (2004) The Optimality of Naive Bayes. FLAIRS2004 conference. https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf

Anexos

Anexo 1 - Modelo de Boletim Autoridade Nacional de Segurança Rodoviária (ANSR).

Anexo II –Fator de Inflação da Variância (VIF) obtido com o programa IBM SPSS Statistics - Base de dados: Acidentes do tipo colisão.

Anexo III –Fator de Inflação da Variância (VIF) obtido com o programa IBM SPSS Statistics - Base de dados: Acidentes do tipo despiste.

Anexo VI – Modelos e resultados: Base de dados original com variáveis categóricas – Acidentes do tipo colisão.

Anexo V – Modelos e resultados: Base de dados original com variáveis categóricas – Acidentes do tipo despiste.

Anexo VI – Modelos e resultados: Base de dados com variáveis categóricas penalizada – Acidentes do tipo colisão.

Anexo VII – Modelos e resultados: Base de dados com variáveis categóricas penalizada – Acidentes do tipo despiste.

Anexo VIII – Modelos e resultados: Base de dados com variáveis categóricas e RUS – Acidentes do tipo colisão.

Anexo IX – Modelos e resultados: Base de dados com variáveis categóricas e RUS – Acidentes do tipo despiste.

Anexo X – Modelos e resultados: Base de dados com variáveis categóricas e ROS (SMOTE) – Acidentes do tipo colisão.

Anexo XI – Modelos e resultados: Base de dados com variáveis categóricas e ROS (SMOTE) – Acidentes do tipo despiste.

Anexo XII – Modelos e resultados: Base de dados original com variáveis binárias – Acidentes do tipo colisão.

Anexo XIII – Modelos e resultados: Base de dados original com variáveis binárias – Acidentes do tipo despiste.

Anexo XIV – Modelos e resultados: Base de dados com variáveis binárias e penalizada – Acidentes do tipo colisão.

Anexo XV – Modelos e resultados: Base de dados com variáveis binárias penalizada – Acidentes do tipo despiste.

Anexo XVI – Modelos e resultados: Base de dados com variáveis binárias e RUS – Acidentes do tipo colisão.

Anexo XVII – Modelos e resultados: Base de dados com variáveis binárias e RUS – Acidentes do tipo despiste.

Anexo XVIII – Modelos e resultados: Base de dados com variáveis binárias e ROS (SMOTE) – Acidentes do tipo colisão.

Anexo XIX – Modelos e resultados: Base de dados com variáveis binárias e ROS (SMOTE) – Acidentes do tipo despiste.

Anexo I - Modelo de Boletim Autoridade Nacional de Segurança Rodoviária (ANSR).

Instrumento de notação registado no I.N.E., sob o n.º 10111, válido até 31/12/2014

Nº Boletim

Entidade Fiscalizadora

AUTORIDADE NACIONAL SEGURANÇA RODOVIÁRIA

Ministério da Administração Interna

BOLETIM ESTATÍSTICO DE ACIDENTES DE VIAÇÃO

(Reservado ao carimbo da unidade)

A - a preencher em todos os acidentes **B e seguintes - a preencher apenas em acidentes com vítimas**

A - IDENTIFICAÇÃO DO ACIDENTE

A1 DATA/HORA

Ano Mês Dia — Hora Min.

A2 LOCALIZAÇÃO

1. ☐ Fora das localidades
☐ Dentro das localidades

2. Distrito _____
Concelho _____
Freguesia _____
Povoação (ou a mais próxima) _____
_____
Coordenadas GPS
Latitude + ,
Longitude - ,

3. Designação de via
Km ,
Arruamento _____
_____ n.º _____

4. Se houver separador central indique em que sentido
1 ☐ Crescente
2 ☐ Decrescente

A3 TIPO DE ACIDENTE

1 ☐ Acidente só com danos materiais
2 ☐ Acidente com vítimas
Mortos
Feridos graves
Feridos leves

A4 NATUREZA DO ACIDENTE

1 ☐ Despiste
2 ☐ Colisão
3 ☐ Atropelamento

A5 NÚMERO DE VEÍCULOS INTERVENIENTES

Ciclomotor e motociclo
Veículo ligeiro
Veículo pesado
Outros

A6 CONDUTORES INTERVENIENTES

1. SEXO

	A	B	C	
1	☐	☐	☐	Masculino
2	☐	☐	☐	Feminino

2. DATA DE NASCIMENTO

A: Ano Mês Dia — B: Ano Mês Dia
C: Ano Mês Dia

B - CIRCUNSTÂNCIAS EXTERNAS

B1 CARACTERÍSTICAS TÉCNICAS DA VIA

1. ESTRADA COM SEPARADOR
1 Autoestrada – nº de vias de trânsito no sentido
2 Outra via – nº de vias de trânsito no sentido

2. ESTRADA SEM SEPARADOR – nº de vias no sentido

3. VIA DE TRÂNSITO
1 ☐ Esquerda
2 ☐ Direita
3 ☐ Central

B2 TRAÇADO DA VIA

1. EM PLANTA
1 ☐ Reta
2 ☐ Curva

2. EM PERFIL
1 ☐ Em patamar
2 ☐ Com inclinação
3 ☐ Em lomba

3. 1 ☐ Sem berma ou impraticável
2 ☐ Berma não pavimentada
3 ☐ Berma pavimentada

4. SITUAÇÃO DO ACIDENTE
1 ☐ Em plena via
2 ☐ Na berma
3 ☐ No passeio
4 ☐ Em via ou pista reservada
5 ☐ Em parque de estacionamento

5. INTERSEÇÃO DE VIAS
1 ☐ **Fora da interseção**
Em interseção de nível
2 ☐ Em cruzamento
3 ☐ Em entroncamento
4 ☐ Em rotunda
5 ☐ Em passagem de nível
Em interseção desnivelada
6 ☐ Em via de aceleração
7 ☐ Em via de desaceleração
8 ☐ Em ramo de ligação – entrada
9 ☐ Em ramo de ligação – saída

6. ACIDENTE EM OBRAS DE ARTE
1 ☐ Túnel
2 ☐ Viaduto/Ponte
3 ☐ Passagem estreita

B3 REGIME DE CIRCULAÇÃO

1. FAIXA DE RODAGEM COM
1 ☐ Sentido único
2 ☐ Dois sentidos
3 ☐ Reversível

2. VELOCIDADE PERMITIDA NO LANÇO
Limite geral _____ Km/h
Limite local _____ Km/h

B4 PAVIMENTO

1. TIPO DE PISO
1 ☐ Terra batida
2 ☐ Betuminoso
3 ☐ Betão de cimento
4 ☐ Calçada

2. ESTADO DE CONSERVAÇÃO
1 ☐ Em bom estado
2 ☐ Em estado regular
3 ☐ Em mau estado

3. OBSTÁCULOS OU OBRAS
1 ☐ Inexistentes
2 ☐ Não sinalizados
3 ☐ Insuficientemente sinalizados
4 ☐ Corretamente sinalizados

4. CONDIÇÕES DE ADERÊNCIA
1 ☐ Seco e limpo
2 ☐ Húmido
3 ☐ Molhado
4 ☐ Com água acumulada na faixa de rodagem
5 ☐ Com gelo, geada ou neve
6 ☐ Com lama
7 ☐ Com gravilha ou areia
8 ☐ Com óleo

B5 SINALIZAÇÃO

1. MARCAS NO PAVIMENTO
1 ☐ Sem marcas rodoviárias ou pouco visíveis
2 ☐ Com marcas – separadoras de sentido de trânsito
3 ☐ Com marcas – separadoras de sentido e de vias de trânsito

2. SINALIZAÇÃO LUMINOSA
1 ☐ Inexistente
2 ☐ A funcionar normalmente
3 ☐ Intermitente
4 ☐ Desligada

3. SINAIS
1 ☐ Stop
2 ☐ Cedência de passagem
3 ☐ Proibição de ultrapassagem
4 ☐ Passagem de peões
5 ☐ Outros

B6 LUMINOSIDADE

1 ☐ Em pleno dia
2 ☐ Sol encandeante
3 ☐ Aurora ou crepúsculo
4 ☐ Noite, sem iluminação
5 ☐ Noite, com iluminação

B7 FATORES ATMOSFÉRICOS

1 ☐ Bom tempo
2 ☐ Chuva
3 ☐ Vento forte
4 ☐ Nevoeiro
5 ☐ Neve
6 ☐ Nuvem de fumo
7 ☐ Granizo

C - NATUREZA DO ACIDENTE

DESPISTE
1 ☐ Despiste simples
Com transposição do separador central
2 ☐ Com dispositivo de retenção
3 ☐ Sem dispositivo de retenção
4 ☐ Com transposição do dispositivo de retenção lateral
5 ☐ Com capotamento
6 ☐ Com colisão com veículo imobilizado ou obstáculo
7 ☐ Com fuga

COLISÃO
8 ☐ Frontal
9 ☐ Traseira com outro veículo em movimento
10 ☐ Lateral com outro veículo em movimento
11 ☐ Com veículo ou obstáculo na faixa de rodagem
12 ☐ Choque em cadeia
13 ☐ Com fuga
14 ☐ Outras situações

ATROPELAMENTO
15 ☐ De peões
16 ☐ De animais
17 ☐ Com fuga

Incêndio posterior. A ☐ B ☐ C ☐ — A preencher no caso de se verificar

D - VEÍCULOS INTERVENIENTES

D1 CATEGORIA/CLASSE

1. VEÍCULOS A, B e C

	A	B	C	
1	☐	☐	☐	Velocípede
2	☐	☐	☐	Velocípede c/motor
3	☐	☐	☐	Ciclomotor
4	☐	☐	☐	Triciclo
5	☐	☐	☐	Motociclo cilindrada ≤ 125cc
6	☐	☐	☐	Motociclo cilindrada > 125cc
7	☐	☐	☐	Automóvel ligeiro
8	☐	☐	☐	Automóvel pesado
9	☐	☐	☐	Veículo agrícola
10	☐	☐	☐	Máquina industrial
11	☐	☐	☐	Veículo sobre carris
12	☐	☐	☐	Veículo de tração animal
13	☐	☐	☐	Quadriciclo
14	☐	☐	☐	Desconhecido

2. Se for automóvel ligeiro ou pesado, indicar o tipo:

	A	B	C	
1	☐	☐	☐	Passageiros
2	☐	☐	☐	Mercadorias
3	☐	☐	☐	Misto
4	☐	☐	☐	Trator
5	☐	☐	☐	Veículo especial. Qual?

3. A B C
1 ☐ ☐ ☐ Sem semireboque/reboque
2 ☐ ☐ ☐ Com semireboque/reboque

D2 TIPO DE SERVIÇO
A B C
1 ☐ ☐ ☐ Particular
2 ☐ ☐ ☐ Público

D3 ANO DE MATRÍCULA
A ____ B ____ C ____

D4 INSPEÇÃO PERIÓDICA
A B C
1 ☐ ☐ ☐ Não obrigatória
2 ☐ ☐ ☐ Válida
3 ☐ ☐ ☐ Sem validade

D5 CERTIFICADO ADR
1. Preencher apenas no caso de transporte de mercadorias perigosas
A B C
1 ☐ ☐ ☐ Válido
2 ☐ ☐ ☐ Sem validade
3 ☐ ☐ ☐ Inexistente
2. MATÉRIA/OBJETO PERIGOSO TRANSPORTADO ____

D6 CARGA/LOTAÇÃO/PNEUS
1. CARGA/LOTAÇÃO
A B C
1 ☐ ☐ ☐ Sem carga
2 ☐ ☐ ☐ Com excesso de carga
3 ☐ ☐ ☐ Carga bem acondicionada
4 ☐ ☐ ☐ Carga mal acondicionada
5 ☐ ☐ ☐ Com lotação excedida
2. PNEUS
A B C
1 ☐ ☐ ☐ Sem deficiência
2 ☐ ☐ ☐ Com deficiência
3. TACÓGRAFO
A B C
1 ☐ ☐ ☐ Sem tacógrafo ou desativado
2 ☐ ☐ ☐ Com tacógrafo

D7 SEGURO
A B C
1 ☐ ☐ ☐ Com seguro
2 ☐ ☐ ☐ Sem seguro
3 ☐ ☐ ☐ Isento

E - CONDUTORES INTERVENIENTES

E1 CARACTERÍSTICAS DA HABILITAÇÃO DE CONDUÇÃO
1. LICENÇA/CARTA DE CONDUÇÃO
A B C
1 ☐ ☐ ☐ Com licença/carta adequada ao veículo
2 ☐ ☐ ☐ Com licença/carta não adequada ao veículo
3 ☐ ☐ ☐ Em situação de instrução/exame
4 ☐ ☐ ☐ Caducada/suspensa
5 ☐ ☐ ☐ Sem licença/carta
6 ☐ ☐ ☐ Não necessária ao veículo que conduz
2. PAÍS DE EMISSÃO
A B C
1 ☐ ☐ ☐ Portugal
2 ☐ ☐ ☐ Outro(s) A ____ B ____ C ____
3. ANO DA HABILITAÇÃO
Relativamente ao veículo que conduzia
A ____ B ____ C ____
4. CERTIFICADO ADR
A B C
1 ☐ ☐ ☐ Válido
2 ☐ ☐ ☐ Sem validade
3 ☐ ☐ ☐ Inexistente

E2 CONDIÇÕES PSÍCO/FÍSICAS
1. CONTROLO DO NÍVEL DE ALCOOLEMIA
A B C
1 ☐ ☐ ☐ **Submetido ao teste de alcoolemia**
Não submetido por
2 ☐ ☐ ☐ Doença
3 ☐ ☐ ☐ Lesão ou morte decorrente do acidente
4 ☐ ☐ ☐ Condutor não contactado na altura do acidente
5 ☐ ☐ ☐ Fuga
6 ☐ ☐ ☐ Recusa
7 ☐ ☐ ☐ Outra

2. TAXA DE ALCOOLEMIA
A ____,__ B ____,__ C ____,__
3. OUTROS FATORES
A B C
1 ☐ ☐ ☐ Normal
2 ☐ ☐ ☐ Droga por despistagem
3 ☐ ☐ ☐ Sono/sonolência
4 ☐ ☐ ☐ Distração
5 ☐ ☐ ☐ Doença súbita
6 ☐ ☐ ☐ Fadiga
4. TEMPO DE CONDUÇÃO CONTINUADA
A B C
1 ☐ ☐ ☐ Menos de 1 hora
2 ☐ ☐ ☐ De 1 a 3 horas
3 ☐ ☐ ☐ De 3 a 5 horas
4 ☐ ☐ ☐ Mais de 5 horas
5 ☐ ☐ ☐ Ignorada

E3 AÇÕES E MANOBRAS ANTES DO ACIDENTE
1. A B C
1 ☐ ☐ ☐ Início de marcha
2 ☐ ☐ ☐ Saída de parqueamento ou rua particular
3 ☐ ☐ ☐ Em marcha normal
4 ☐ ☐ ☐ Ultrapassagem pela esquerda
5 ☐ ☐ ☐ Ultrapassagem pela direita
6 ☐ ☐ ☐ Mudança de direção para a esquerda
7 ☐ ☐ ☐ Mudança de direção para a direita
8 ☐ ☐ ☐ Marcha atrás
9 ☐ ☐ ☐ Circulação em sentido oposto ao estabelecido
10 ☐ ☐ ☐ Travagem brusca
11 ☐ ☐ ☐ Parado ou estacionado
12 ☐ ☐ ☐ Inversão do sentido de marcha
13 ☐ ☐ ☐ Trânsito em filas paralelas
14 ☐ ☐ ☐ Mudança de via de trânsito para a esquerda
15 ☐ ☐ ☐ Mudança de via de trânsito para a direita
16 ☐ ☐ ☐ Desvio brusco/saída de fila de trânsito
17 ☐ ☐ ☐ Atravessando a via
2. ESQUEMA ____ (Ver esquema em anexo)

E4 INFORMAÇÃO COMPLEMENTAR A AÇÕES E MANOBRAS
A B C
1 ☐ ☐ ☐ Desrespeito da sinalização vertical
2 ☐ ☐ ☐ Desrespeito das marcas rodoviárias
3 ☐ ☐ ☐ Desrespeito da sinalização semafórica
4 ☐ ☐ ☐ Manobra irregular
5 ☐ ☐ ☐ Velocidade excessiva para as condições existentes
6 ☐ ☐ ☐ Não sinalização da manobra
7 ☐ ☐ ☐ Desrespeito das distâncias de segurança
8 ☐ ☐ ☐ Circulação afastada da berma ou passeio
9 ☐ ☐ ☐ Rebentamento pneumático
10 ☐ ☐ ☐ Queda de carga ou objeto
11 ☐ ☐ ☐ Falha mecânica do veículo
12 ☐ ☐ ☐ Ausência de luzes quando obrigatórias
13 ☐ ☐ ☐ Obstáculo imprevisto na faixa de rodagem
14 ☐ ☐ ☐ Abertura de porta
15 ☐ ☐ ☐ Encandeamento
16 ☐ ☐ ☐ Não identificada

E5 ACESSÓRIOS DE SEGURANÇA
A B C
1 ☐ ☐ ☐ Capacete
2 ☐ ☐ ☐ Cinto de segurança
3 ☐ ☐ ☐ Sem uso de cinto/capacete
4 ☐ ☐ ☐ Isento

F - CONSEQUÊNCIAS DO ACIDENTE

F1 CONDUTORES VÍTIMAS
1. GRAU DE GRAVIDADE DAS LESÕES
A B C
1 ☐ ☐ ☐ Morto
2 ☐ ☐ ☐ Ferido grave
3 ☐ ☐ ☐ Ferido leve

F2 PASSAGEIROS VÍTIMAS
Veículo A **Veículo B** **Veículo C**
1. SEXO
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Masculino
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Feminino

2. IDADE
a b | i j | r s
c d | l m | t u
3. POSIÇÃO NO VEÍCULO
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ À frente
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ À retaguarda
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Desconhecido
4. USO DE ACESSÓRIOS DE SEGURANÇA
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ C/ capacete/cinto segurança
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ C/ sistema retenção de crianças
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ S/ uso capacete/cinto segurança
4 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ S/ sistema retenção de crianças
5. GRAU DE GRAVIDADE DAS LESÕES
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Morto
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ferido grave
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ferido leve
4 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ileso

F3 PEÕES VÍTIMAS
1. SEXO
a b c d
1 ☐☐☐☐ Masculino
2 ☐☐☐☐ Feminino
2. a b c d
1 ☐☐☐☐ Peão isolado
2 ☐☐☐☐ Peões em grupo
3 ☐☐☐☐ Conduzindo à mão velocípedes, carros de crianças ou de deficientes físicos
4 ☐☐☐☐ Deslocando-se sobre patins, trotinetes ou outros
3. IDADE
a ____ b ____ c ____ d ____
4. CONDIÇÕES PSÍCO-FÍSICAS
a b c d
1 ☐☐☐☐ Sem restrições
2 ☐☐☐☐ Com visão deficiente
3 ☐☐☐☐ Com audição deficiente
4 ☐☐☐☐ Com deficiência motora
Influenciada pelo álcool
a ____,__ b ____,__ c ____,__ d ____,__
5
5. AÇÕES
a b c d
1 ☐☐☐☐ A sair ou entrar num veículo
2 ☐☐☐☐ Surgindo inesperadamente na faixa de rodagem de trás de um obstáculo
3 ☐☐☐☐ Em plena faixa de rodagem
4 ☐☐☐☐ Em trabalhos na via
5 ☐☐☐☐ Atravessando fora da passagem de peões, a menos de 50 m de uma passagem
6 ☐☐☐☐ Atravessando fora da passagem de peões a mais de 50 m de uma passagem ou quando não exista passagem
7 ☐☐☐☐ Atravessando em passagem sinalizada
8 ☐☐☐☐ Atravessando em passagem sinalizada com desrespeito da sinalização semafórica
9 ☐☐☐☐ Em ilhéu ou refúgio na via
10 ☐☐☐☐ Transitando pela direita da faixa de rodagem
11 ☐☐☐☐ Transitando pela esquerda da faixa de rodagem
12 ☐☐☐☐ Transitando pela berma ou passeio
6. UTILIZAÇÃO DE MATERIAL REFLETOR
a b c d
1 ☐☐☐☐ Sim
2 ☐☐☐☐ Não
7. GRAVIDADE DAS LESÕES
a b c d
1 ☐☐☐☐ Morto
2 ☐☐☐☐ Ferido grave
3 ☐☐☐☐ Ferido leve

DATA ____/____/____
Número de boletins utilizados neste acidente ____
Nome _____
(Posto) _____

Anexo II –Fator de Inflação da Variância (VIF) obtido com o programa IBM SPSS Statistics - Base de dados: Acidentes do tipo colisão.

Modelo	Coeficiente não padronizado		Coeficiente padronizado	t	sig	Estatísticas de colinearidade	
	B	Erro Erro	Beta			Tolerância	VIF
(Constante)	0	6%		-12%	0,906		
Dia_ Semana	0	1%	0	-647%	0,000	0,9648	1,0365
Ciclomotor	0	1%	0	1520%	0,000	0,8583	1,1652
Ligeiros	0	2%	0	174%	0,082	0,8710	1,1481
Outros_Veic.	0	1%	0	571%	0,000	0,8817	1,1342
Pesados	0	1%	0	1349%	0,000	0,8993	1,1120
AE	0	1%	0	134%	0,182	0,2357	4,2427
IC	0	1%	0	163%	0,103	0,6202	1,6124
IP	0	2%	0	438%	0,000	0,8358	1,1965
Tracado_curva	0	1%	0	549%	0,000	0,8947	1,1177
Tracado_patamar	0	1%	0	-149%	0,136	0,9373	1,0669
B._nao_pavimentada	0	1%	0	-68%	0,499	0,8167	1,2244
Sem_b.	0	1%	0	-115%	0,249	0,8403	1,1901
Tracado_local	0	6%	0	-26%	0,791	0,9950	1,0050
Cruzamento	0	2%	0	-113%	0,258	0,9330	1,0719
Entroncamento	0	1%	0	-354%	0,000	0,9351	1,0695
R._ligacao_entrada	0	4%	0	-4%	0,966	0,9896	1,0105
R._ligacao_saida	0	8%	0	-161%	0,107	0,9963	1,0037
Rotunda	0	2%	0	-290%	0,004	0,9112	1,0974
Via_aceleracao	0	2%	0	-189%	0,058	0,9817	1,0186
Via_desaceleracao	0	4%	0	45%	0,655	0,9899	1,0102
Obst._sinalizados	0	2%	0	251%	0,012	0,9826	1,0177
Obst._nao_sinalizados	0	4%	0	-8%	0,938	0,9962	1,0038
Obst_insu._sinalizados	0	5%	0	46%	0,647	0,9961	1,0039
SL_funcionar	0	1%	0	-164%	0,101	0,9280	1,0776
SL_desligada	0	2%	0	-40%	0,690	0,9809	1,0194
SL_intermitente	0	6%	0	-65%	0,516	0,9940	1,0060
Circ._dois_sentidos	0	1%	0	792%	0,000	0,2550	3,9223
Circ._reversivel	0	5%	0	177%	0,077	0,9678	1,0333
Aurora_crepusculo	0	1%	0	46%	0,645	0,9814	1,0189
Noite_sem_iluminacao	0	1%	0	836%	0,000	0,9213	1,0854
Noite_com_iluminacao	0	1%	0	297%	0,003	0,9351	1,0694
Sol_escandeante	0	2%	0	213%	0,033	0,9920	1,0081
SH_sentido	0	1%	0	177%	0,077	0,5348	1,8700
SH_inexistente	0	1%	0	-300%	0,003	0,7493	1,3346
Pav._gelo_geada_neve	0	6%	0	122%	0,224	0,9939	1,0061
Pav._molhado_humido _com_agua	0	1%	0	-42%	0,677	0,2822	3,5438
Pav._regular	0	1%	0	105%	0,295	0,9716	1,0292
Pav._mau	0	3%	0	-120%	0,230	0,9794	1,0210
Chuva	0	1%	0	-94%	0,348	0,2863	3,4929
Nevoeiro	0	3%	0	108%	0,282	0,9405	1,0632
Vias_1	0	1%	0	348%	0,000	0,3067	3,2605
Vel._90-99	0	1%	0	771%	0,000	0,6837	1,4627
Vel.100-120	0	1%	0	550%	0,000	0,4592	2,1775

a. Variável Dependente: Severidade

Anexo III –Fator de Inflação da Variância (VIF) obtido com o programa IBM SPSS Statistics - Base de dados: Acidentes do tipo despiste.

Modelo	Coeficiente não padronizado		Coeficiente padronizado	t	sig	Estatísticas de colinearidade	
	B	Erro Erro	Beta			Tolerância	VIF
Dia_ Semana	-0,026	0,006	-0,038	-4,125	0,000	0,966	1,035
Ciclomotor	0,113	0,018	0,134	6,254	0,000	0,180	5,562
Ligeiros	0,034	0,017	0,044	1,940	0,052	0,157	6,349
Pesados	0,101	0,022	0,062	4,658	0,000	0,464	2,156
AE	0,014	0,015	0,021	0,954	0,340	0,173	5,792
IC	-0,003	0,012	-0,003	-0,267	0,790	0,665	1,503
IP	0,005	0,018	0,003	0,265	0,791	0,820	1,220
Tracado_curva	0,003	0,006	0,005	0,546	0,585	0,865	1,156
Tracado_patamar	1,822E-05	0,006	0,000	0,003	0,998	0,892	1,121
B._nao_pavimentada	0,000	0,010	0,000	-0,028	0,978	0,610	1,638
Sem_b.	-0,003	0,009	-0,004	-0,342	0,732	0,603	1,660
Tracado_local	-0,055	0,040	-0,012	-1,356	0,175	0,992	1,008
Cruzamento	-0,081	0,054	-0,014	-1,489	0,136	0,986	1,014
Entroncamento	0,002	0,029	0,001	0,067	0,946	0,986	1,014
R._ligacao_entrada	0,000	0,049	0,000	-0,010	0,992	0,981	1,020
R._ligacao_saida	0,050	0,066	0,007	0,755	0,450	0,979	1,021
Rotunda	-0,007	0,029	-0,002	-0,252	0,801	0,938	1,066
Via_aceleracao	-0,060	0,036	-0,015	-1,679	0,093	0,981	1,019
Via_desaceleracao	-0,007	0,027	-0,003	-0,268	0,789	0,947	1,056
Obst._sinalizados	-0,010	0,031	-0,003	-0,332	0,740	0,972	1,028
Obst._nao_sinalizados	0,009	0,048	0,002	0,199	0,842	0,982	1,018
Obst_insu._sinalizados	-0,064	0,077	-0,008	-0,838	0,402	0,982	1,018
SL_funcionar	-0,035	0,024	-0,014	-1,461	0,144	0,952	1,050
SL_desligada	-0,023	0,022	-0,009	-1,021	0,307	0,987	1,013
SL_intermitente	0,005	0,062	0,001	0,078	0,938	0,992	1,008
Circ._reversivel	-0,044	0,059	-0,007	-0,746	0,456	0,993	1,007
Circ._sentido_unico	-0,036	0,013	-0,056	-2,885	0,004	0,218	4,586
Aurora_crepusculo	0,041	0,015	0,025	2,750	0,006	0,978	1,022
Noite_sem_iluminacao	0,062	0,008	0,075	7,952	0,000	0,939	1,065
Noite_com_iluminacao	0,066	0,011	0,057	5,989	0,000	0,899	1,112
Sol_escandeante	-0,028	0,030	-0,008	-0,929	0,353	0,991	1,009
SH_sentido	-0,004	0,009	-0,006	-0,462	0,644	0,454	2,203
SH_inexistente	-0,025	0,011	-0,025	-2,193	0,028	0,623	1,604
Pav._gelo_geada_neve	-0,084	0,036	-0,021	-2,326	0,020	0,989	1,012
Pav._molhado_humido _com_agua	-0,042	0,012	-0,062	-3,470	0,001	0,261	3,827
Pav._regular	0,018	0,006	0,028	3,040	0,002	0,973	1,028
Pav._mau	-0,025	0,027	-0,009	-0,917	0,359	0,951	1,052
Chuva	-0,027	0,013	-0,039	-2,191	0,029	0,262	3,812
Nevoeiro	0,048	0,026	0,017	1,819	0,069	0,923	1,083
Vias_1	0,020	0,010	0,031	1,979	0,048	0,338	2,959
Vel._90-99	0,024	0,008	0,037	3,062	0,002	0,577	1,733
Vel.100-120	0,043	0,011	0,061	3,897	0,000	0,338	2,962
a. Variável Dependente: Severidade							

Anexo VI – Modelos e resultados: Base de dados original com variáveis categóricas – Acidentes do tipo colisão.

Acidentes do tipo colisão - base categórica "original													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	11084	80,382	2705	19,617	0,1829	0	0,858	0,634	0,916	0,696	10528	1740
							1	0,366	0,142	0,242	0,696	965	556
								0,804	0,58	0,842	0,696		
	Percentage split 80%	2235	81,037	523	18,963	0,1984	0	0,87	0,642	0,911	0,699	2120	317
							1	0,358	0,13	0,266	0,699	206	115
								0,81	0,582	0,836	0,699		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	11908	86,359	1881	13,641	0,0877	0	0,957	0,89	0,897	0,662	11741	527
							1	0,11	0,043	0,241	0,662	1354	167
								0,864	0,797	0,824	0,662		
	Percentage split 80%	2310	83,756	448	16,244	0,1579	0	0,918	0,769	0,901	0,664	2236	201
							1	0,231	0,082	0,269	0,664	247	74
								0,838	0,69	0,827	,827,664		
Support Vector Machines (SVM)	Cross-validation – folds 10	12270	88,984	1519	11,016	0,0034	0	1	0,998	0,89	0,501	12267	1
							1	0,002	0	0,75	0,501	1518	4
								0,89	0,888	0,874	0,501		
	Percentage split 80%	2437	88,361	321	11,639	0,1164	0	1	1	0,884	0,5	2437	0
							1	0	0	?	0,5	321	0
								0,884	0,884	?	0,5		
Decision Trees (DT)	Cross-validation – folds 10	12268	88,970	1521	11,031	0	0	1	1	0,89	0,5	12268	0
							1	0	0	?	0,5	1521	0
								0,89	0,89	?	0,5		
	Percentage split 80%	2437	88,361	321	11,639	0	0	1	1	0,884	0,5	2437	0
							1	0	0	?	0,5	321	0
								0,884	0,884	?	0,5		

Random Forests (RF)	Cross-validation – folds 10	12086	87,650	1703	12,350	0,0389	0	0,979	0,954	0,892	0,657	12016	252
							1	0,046	0,021	0,217	0,657	1451	70
								0,876	0,851	0,818	0,657		
	Percentage split 80%	2413	87,491	345	12,509	0,0358	0	0,985	0,963	0,886	0,662	2401	36
							1	0,037	0,015	0,25	0,662	309	12
								0,875	0,882	0,812	0,662		
K-Nearest Neighbors (KNN)	Cross-validation – folds 10	11884	86,185	1905	13,815	0,0532	0	0,959	0,919	0,894	0,586	11761	507
							1	0,081	0,041	0,195	0,586	1398	123
								0,862	0,822	0,817	0,586		
	Percentage split 80%	2383	86,403	375	13,597	0,0571	0	0,968	0,928	0,888	0,601	2360	77
							1	0,072	0,032	0,23	0,601	298	23
								0,864	0,824	0,811	0,601		

Anexo V – Modelos e resultados: Base de dados original com variáveis categóricas – Acidentes do tipo despiste.

Acidentes do tipo colisão - base categórica "original													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	10005	86,050	1622	13,950	0,0844	0	0,958	0,895	0,892	0,643	9865	429
							1	0,105	0,042	0,246	0,643	1193	140
								0,86	0,797	0,818	0,643		
	Percentage split 80%	1991	85,634	334	14,366	0,0957	0	0,951	0,876	0,894	0,637	1958	101
							1	0,124	0,049	0,246	0,637	233	33
								0,865	0,781	0,82	0,637		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	9833	84,570	1794	15,430	0,0384	0	0,944	0,914	0,889	0,586	9718	576
							1	0,086	0,056	0,166	0,586	1218	115
								0,846	0,815	0,806	0,586		
	Percentage split 80%	2003	86,151	322	13,850	0,0265	0	0,966	0,947	0,888	0,588	1989	70
							1	0,053	0,034	0,167	0,588	252	14
								0,862	0,843	0,805	0,588		
Support Vector Machines (SVM)	Cross-validation – folds 10	10294	88,535	1333	11,465	0	0	1	1	0,885	0,5	10294	0
							1	0	0	?	0,5	1333	0
								0,885	0,885	?	0,5		
	Percentage split 80%	2059	88,559	266	11,441	0	0	1	1	0,886	0,5	2059	0
							1	0	0	?	0,5	266	0
								0,886	0,886	?	0,5		
Decision Trees (DT)	Cross-validation – folds 10	10294	88,535	1333	11,465	0	0	1	1	0,885	0,5	10294	0
							1	0	0	?	0,5	1333	0
								0,885	0,885	?	0,5		
	Percentage split 80%	2059	88,559	266	11,441	0	0	1	1	0,886	0,5	2059	0
							1	0	0	?	0,5	266	0
								0,886	0,886	?	0,5		

Random Forests (RF)	Cross-validation – folds 10	10088	86,764	1539	13,236	0,02	0	0,978	0,962	0,887	0,584	10037	257
							1	0,038	0,025	0,166	0,584	1282	51
								0,868	0,854	0,804	0,584		
	Percentage split 80%	2011	86,495	314	13,505	0,0234	0	0,971	0,955	0,887	0,573	1999	60
							1	0,045	0,029	0,167	0,573	254	12
								0,865	0,849	0,805	0,573		
K-Nearest Neighbors (KNN)	Cross-validation – folds 10	9921	85,327	1706	14,673	0,0076	0	0,958	0,952	0,886	0,548	9857	437
							1	0,048	0,042	0,128	0,548	1269	64
								0,853	0,848	0,799	0,548		
	Percentage split 80%	1981	85,204	344	14,796	0,0064	0	0,956	0,951	0,886	0,541	1968	91
							1	0,049	0,044	0,125	0,541	253	13
								0,852	0,847	0,799	0,541		

Anexo VI – Modelos e resultados: Base de dados com variáveis categóricas penalizada – Acidentes do tipo colisão.

Acidentes do tipo colisão - base categórica penalizada														
Modelo	Penalização	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	4	Cross-validation – folds 10	8967	65,030	4822	34,970	0,1379	0	0,653	0,368	0,935	0,696	8005	4263
								1	0,632	0,347	0,184	0,696	559	962
									0,65	0,365	0,852	0,696		
		Percentage split 80%	1815	65,809	943	34,191	0,1467	0	0,663	0,377	0,93	0,698	1615	822
								1	0,623	0,337	0,196	0,698	121	200
									0,658	0,372	0,845	0,698		
Artificial Neural Networks (ANN)	7	Cross-validation – folds 10	6615	47,973	7174	52,027	0,0606	0	0,45	0,279	0,929	0,633	5518	6750
								1	0,721	0,55	0,14	0,633	424	1097
									0,48	0,309	0,842	0,633		
		Percentage split 80%	1609	58,339	1149	41,661	0,1305	0	0,563	0,259	0,943	0,71	1371	1066
								1	0,741	0,437	0,183	0,71	83	238
									0,583	0,279	0,854	0,71		
Support Vector Machines (SVM)	7	Cross-validation – folds 10	9878	71,637	39,11	28,363	0,1872	0	0,731	0,398	0,937	0,666	8962	3306
								1	0,602	0,268	0,217	0,666	605	916
									0,716	0,384	0,857	0,666		
		Percentage split 80%	2019	73,205	739	26,795	0,2171	0	0,748	0,386	0,936	0,681	1822	615
								1	0,614	0,252	0,243	0,681	124	197
									0,732	0,371	0,856	0,681		
Decision Trees (DT)	7	Cross-validation – folds 10	9867	71,557	3922	28,443	0,1191	0	0,749	0,553	0,916	0,576	9187	3081
								1	0,447	0,251	0,181	0,575	841	680
									0,716	0,52	0,835	0,576		
		Percentage split 80%	1986	72,009	772	27,991	0,1184	0	0,759	0,576	0,909	0,555	1850	587
								1	0,424	0,241	0,188	0,554	185	136
									0,72	0,537	0,825	0,555		

Random Forests (RF)	10	Cross-validation – folds 10	10909	79,114	2880	20,886	0,1182	0	0,853	0,711	0,906	0,658	10469	1799
								1	0,289	0,147	0,197	0,658	1081	440
									0,791	0,648	0,828	0,658		
		Percentage split 80%	2198	79,695	560	20,305	0,1507	0	0,86	0,685	0,905	0,681	2097	340
								1	0,315	0,14	0,229	0,681	220	101
									0,797	0,622	0,826	0,681		
K-Nearest Neighbors (KNN)	10	Cross-validation – folds 10	9700	70,346	4089	29,654	0,0753	0	0,744	0,621	0,906	0,587	9123	3145
								1	0,379	0,256	0,155	0,587	944	577
									0,703	0,58	0,823	0,587		
		Percentage split 80%	1970	71,429	788	28,571	0,0874	0	0,759	0,626	0,902	0,601	1850	587
								1	0,374	0,241	0,17	0,601	201	120
									0,714	0,581	0,817	0,601		

Anexo VII – Modelos e resultados: Base de dados com variáveis categóricas penalizada – Acidentes do tipo despiste.

Acidentes do tipo despiste - base categórica penalizada														
Modelo	Penalização	Teste	Clas. corretamente	%	Clas. incorretamente	%	Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
													a	b
Naive Bayes	4	Cross-validation – folds 10	7925	68,160	3702	31,840	0,1069	0	0,708	0,52	0,913	0,644	7285	3009
								1	0,48	0,292	0,175	0,644	693	640
									0,682	0,494	0,829	0,644		
		Percentage split 80%	1472	63,312	853	36,688	0,0977	0	0,644	0,447	0,918	0,638	1325	734
								1	0,553	0,356	0,167	0,638	119	147
									0,633	0,437	0,832	0,638		
Artificial Neural Networks (ANN)	7	Cross-validation – folds 10	3230	27,780	8397	72,220	0,0134	0	0,204	0,156	0,91	0,586	2105	8189
								1	0,844	0,769	0,121	0,586	208	1125
									0,278	0,229	0,82	0,586		
		Percentage split 80%	1898	81,634	427	18,366	0,1428	0	0,888	0,737	0,903	0,654	1828	231
								1	0,263	0,112	0,233	0,654	196	70
									0,816	0,665	0,826	0,654		
Support Vector Machines (SVM)	7	Cross-validation – folds 10	7413	63,757	4214	36,243	0,1067	0	0,674	0,434	0,92	0,607	6658	3636
								1	0,566	0,535	0,172	0,607	578	755
									0,638	0,424	0,834	0,607		
		Percentage split 80%	1476	63,484	849	36,516	0,0916	0	0,648	0,466	0,915	0,591	1334	725
								1	0,534	0,352	0,164	0,591	124	142
									0,635	0,453	0,829	0,591		
Decision Trees (DT)	8	Cross-validation – folds 10	7628	65,606	3999	34,394	0,0565	0	0,687	0,586	0,901	0,55	7076	3218
								1	0,414	0,313	0,146	0,55	781	552
									0,656	0,555	0,814	0,55		
		Percentage split 80%	1528	65,720	797	34,280	0,0621	0	0,687	0,575	0,902	0,55	1415	644
								1		0,425	0,313	0,149	0,55	153 113
										0,657	0,545	0,816	0,55	

Random Forests (RF)	10	Cross-validation – folds 10	8844	76,064	2783	23,936	0,0541	0	0,828	0,761	0,894	0,593	8525	1769
								1	0,239	0,172	0,153	0,593	1014	319
									0,761	0,693	0,809	0,593		
		Percentage split 80%	1785	76,774	540	23,226	0,06	0	0,836	0,763	0,895	0,591	1722	337
								1	0,237	0,164	0,158	0,591	203	63
									0,768	0,695	0,81	0,591		
K-Nearest Neighbors (KNN)	5	Cross-validation – folds 10	8375	72,031	3252	27,969	0,0403	0	0,777	0,719	0,893	0,547	8001	2293
								1	0,281	0,223	0,14	0,547	959	374
									0,72	0,662	0,807	0,547		
		Percentage split 80%	1641	70,581	684	29,419	0,0306	0	0,76	0,714	0,892	0,541	1565	494
								1	0,286	0,24	0,133	0,542	190	76
									0,706	0,66	0,805	0,542		

Anexo VIII – Modelos e resultados: Base de dados com variáveis categóricas e RUS – Acidentes do tipo colisão.

Acidentes do tipo colisão - base categórica RUS												
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix
			%		%							a b
Naive Bayes	Cross-validation – folds 10	1883	61,900	1159	38,100	0,238	0	0,553	0,315	0,637	0,675	841 680
							1	0,685	0,447	0,605	0,675	479 1042
								0,619	0,381	0,621	0,675	
	Percentage split 80%	362	59,540	246	40,461	0,1857	0	0,509	0,324	0,594	0,651	149 144
							1	0,676	0,491	0,597	0,651	102 213
								0,595	0,411	0,595	0,651	
Artificial Neural Networks (ANN)	Cross-validation – folds 10	1790	58,843	1252	41,157	0,1769	0	0,558	0,381	0,594	0,625	849 672
							1	0,619	0,442	0,583	0,625	580 941
								0,588	0,412	0,589	0,625	
	Percentage split 80%	356	58,553	252	41,447	0,1735	0	0,631	0,457	0,562	0,614	185 108
							1	0,543	0,369	0,613	0,614	144 171
								0,586	0,411	0,589	0,614	

Anexo IX – Modelos e resultados: Base de dados com variáveis categóricas e RUS – Acidentes do tipo despiste.

Acidentes do tipo despiste - base categórica RUS													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	1581	59,302	1085	40,698	0,186	0	0,539	0,353	0,604	0,631	719	614
							1	0,647	0,461	0,584	0,631	417	862
								0,593	0,407	0,594	0,631		
	Percentage split 80%	329	61,726	204	38,274	0,2325	0	0,566	0,324	0,622	0,663	145	116
							1	0,676	0,444	0,613	0,663	88	184
								0,617	0,638	0,618	0,663		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	1492	55,964	1174	44,036	0,1193	0	0,485	0,366	0,57	0,578	647	686
							1	634	0,515	0,552	0,578	488	845
								0,56	0,44	0,561	0,578		
	Percentage split 80%	288	54,034	245	45,966	0,0873	0	0,72	0,632	0,522	0,583	188	73
							1	0,368	0,28	0,578	0,583	172	100
								0,54	0,452	0,551	0,583		

Anexo X – Modelos e resultados: Base de dados com variáveis categóricas e ROS (SMOTE) – Acidentes do tipo colisão.

Acidentes do tipo colisão - base categórica SMOTE														
Modelo	SMOTE	Teste	Clas. corretamente	%	Clas. incorretamente	%	Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
													a	b
Naive Bayes	600% (100+100+50)	Cross-validation – folds 10	13964	65,271	7430	34,729	0,3047	0	0,636	0,325	0,725	0,729	7803	4465
								1	0,675	0,364	0,58	0,729	2965	6161
									0,653	0,342	0,663	0,729		
		Percentage split 80%	2754	64,361	1525	35,639	0,2883	0	0,62	0,325	0,721	0,721	1527	934
								1	0,675	0,38	0,568	0,721	531	1227
									0,644	0,348	0,656	0,721		
Artificial Neural Networks (ANN)	600% (100+100+50)	Cross-validation – folds 10	15513	72,511	5881	27,489	0,4428	0	0,736	0,29	0,774	0,805	9030	3238
								1	0,71	0,264	0,667	0,805	2643	6483
									0,725	0,279	0,728	0,805		
		Percentage split 80%	3092	72,260	1187	27,740	0,4283	0	0,779	0,354	0,749	0,802	1917	544
								1	0,646	0,221	0,684	0,802	643	1175
									0,723	0,297	0,721	0,802		

Anexo XI – Modelos e resultados: Base de dados com variáveis categóricas e ROS (SMOTE) – Acidentes do tipo despiste.

Acidentes do tipo despiste - base categórica SMOTE														
Modelo	SMOTE	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	2X (100 + 100)	Cross-validation – folds 10	10395	66,524	5231	33,47,62	0,2394	0	0,769	0,534	0,735	0,694	7911	2383
								1	0,466	0,231	0,51	0,694	2848	2484
									0,665	0,431	0,659	0,694		
		Percentage split 80%	2066	66,112	1059	33,888	0,2356	0	0,763	0,532	0,731	0,683	1561	484
								1	0,468	0,237	0,511	0,683	575	505
									0,661	0,43	0,655	0,683		
Artificial Neural Networks (ANN)	2X (100 + 100)	Cross-validation – folds 10	11319	75,437	4307	27,563	0,3822	0	0,799	0,42	0,786	0,779	8225	2069
								1	0,58	0,201	0,599	0,779	2238	3094
									0,724	0,345	0,722	0,779		
		Percentage split 80%	2242	71,744	883	28,256	0,3173	0	0,879	0,589	0,739	0,769	1798	247
								1	0,411	0,121	0,643	0,769	636	444
									0,717	0,427	0,705	0,769		

Anexo XII – Modelos e resultados: Base de dados original com variáveis binárias – Acidentes do tipo colisão.

Acidentes do tipo colisão - base numérica "original"													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	11303	81,971	2486	18,029	0,1554	0	0,886	0,717	0,909	0,672	10873	1395
							1	0,283	0,114	0,236	0,672	1091	430
								0,82	0,651	0,835	0,672		
	Percentage split 80%	2275	82,487	483	17,513	0,1655	0	0,898	0,729	0,903	0,678	2188	249
							1	0,271	0,102	0,259	0,678	234	87
								0,825	0,656	0,828	0,678		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	11983	86,903	1806	13,097	0,0957	0	0,963	0,893	0,897	0,638	11820	448
							1	0,107	0,037	0,267	0,638	1358	163
								0,869	0,798	0,827	0,638		
	Percentage split 80%	2382	86,367	376	13,633	0,088	0	0,964	0,9	0,89	0,605	2350	87
							1	0,1	0,036	0,269	0,605	289	32
								0,864	0,8	0,818	0,605		
Support Vector Machines (SVM)	Cross-validation – folds 10	12268	88,969	1521	11,031	0	0	1	1	0,89	0,5	12268	0
							1	0	0	?	0,5	1521	0
								0,89	0,89	?	0,5		
	Percentage split 80%	2437	88,361	321	11,639	0	0	1	1	0,89	0,5	2437	0
							1	0	0	?	0,5	321	0
								0,89	0,89	?	0,5		
Decision Trees (DT)	Cross-validation – folds 10	12268	88,969	1521	11,031	0	0	1	1	0,89	0,5	12268	0
							1	0	0	?	0,5	1521	0
								0,89	0,89	?	0,5		
	Percentage split 80%	2437	88,361	321	11,639	0	0	1	1	0,89	0,5	2437	0
							1	0	0	?	0,5	321	0
								0,89	0,89	?	0,5		

Random Forests (RF)	Cross-validation – folds 10	12058	87,447	1731	12,553	0,0658	0	0,974	0,929	0,894	0,646	11950	318
							1	0,071	0,026	0,254	0,646	1413	108
								0,874	0,829	0,824	0,646		
	Percentage split 80%	2395	86,838	363	13,162	0,073	0	0,973	0,922	0,889	0,639	2370	67
							1	0,078	0,027	0,272	0,639	296	25
								0,868	0,818	0,817	0,639		
K-Nearest Neighbors (KNN)	Cross-validation – folds 10	12021	87,178	1768	12,822	0,0597	0	0,971	0,93	0,894	0,59	11914	354
							1	0,07	0,029	0,232	0,59	1414	107
								0,872	0,83	0,821	0,59		
	Percentage split 80%	2390	86,657	368	13,343	0,0765	0	0,97	0,916	0,889	0,587	2363	74
							1	0,084	0,03	0,267	0,587	294	27
								0,867	0,813	0,817	0,587		

Anexo XIII – Modelos e resultados: Base de dados original com variáveis binárias – Acidentes do tipo despiste.

Acidentes do tipo despiste - base numérica "original"													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	10016	86,144	1611	13,856	0,0757	0	0,961	0,905	0,891	0,639	9889	405
							1	0,095	0,039	0,239	0,639	1206	127
								0,861	0,806	0,816	0,639		
	Percentage split 80%	1990	85,591	335	14,409	0,7056	0	0,953	0,895	0,892	0,639	1962	97
							1	0,105	0,047	0,224	0,639	238	28
								0,856	0,798	0,815	0,639		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	10012	86,110	1615	13,890	0,0279	0	0,966	0,946	0,887	0,581	9940	354
							1	0,054	0,034	0,169	0,581	1261	72
								0,861	0,841	0,805	0,581		
	Percentage split 80%	2012	86,538	313	13,462	0,0099	0	0,973	0,966	0,886	0,589	2003	56
							1	0,034	0,027	0,138	0,589	257	9
								0,865	0,859	0,801	0,589		
Support Vector Machines (SVM)	Cross-validation – folds 10	10294	88,535	1333	11,465	0	0	1	1	0,885	0,5	10294	0
							1	0	0	?	0,5	1333	0
								0,885	0,885	?	0,5		
	Percentage split 80%	2059	88,559	266	11,441	0	0	1	1	0,886	0,5	2059	0
							1	0	0	?	0,5	266	0
								0,886	0,886	?	0,5		
Decision Trees (DT)	Cross-validation – folds 10	10294	88,535	1333	11,465	0	0	1	1	0,885	0,499	10294	0
							1	0	0	?	0,499	1333	0
								0,885	0,885	?	0,499		
	Percentage split 80%	2059	88,559	266	11,441	0	0	1	1	0,886	0,5	2059	0
							1	0	0	?	0,5	266	0
								0,886	0,886	?	0,5		

Random Forests (RF)	Cross-validation – folds 10	10085	86,738	1542	13,262	0,0069	0	0,976	0,971	0,886	0,572	10047	247
							1	0,029	0,024	0,133	0,572	1295	38
								0,867	0,863	0,8	0,572		
	Percentage split 80%	2005	86,237	320	13,763	-0,0001	0	0,97	0,97	0,886	0,576	1997	62
							1	0,03	0,03	0,114	0,576	258	8
								0,862	0,862	0,797	0,576		
K-Nearest Neighbors (KNN)	Cross-validation – folds 10	10031	86,273	1596	13,727	0,0018	0	0,97	0,969	0,885	0,549	9990	304
							1	0,031	0,03	0,119	0,549	1292	41
								0,863	0,862	0,798	0,549		
	Percentage split 80%	1992	85,677	333	14,323	-0,0091	0	0,964	0,97	0,885	0,557	1984	75
							1	0,03	0,036	0,096	0,557	258	8
								0,857	0,863	0,795	0,557		

Anexo XIV – Modelos e resultados: Base de dados com variáveis binárias e penalizada – Acidentes do tipo colisão.

Acidentes do tipo colisão - base numérica penalizada														
Modelo	Penalização	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	10	Cross-validation – folds 10	7955	57,691	5834	42,309	0,107	0	0,562	0,304	0,937	0,672	6896	5372
								1	0,696	0,438	0,165	0,672	462	1059
									0,577	0,319	0,852	0,672		
		Percentage split 80%	1609	58,339	1149	41,661	0,114	0	0,569	0,308	0,933	0,678	1387	1050
								1	0,692	0,431	0,175	0,678	99	222
									0,583	0,323	0,845	0,678		
Artificial Neural Networks (ANN)	10	Cross-validation – folds 10	9343	67,757	4446	32,243	0,1294	0	0,693	0,449	0,926	0,653	8505	3763
								1	0,551	0,307	0,182	0,653	683	838
									0,678	0,433	0,844	0,653		
		Percentage split 80%	1898	68,818	860	31,182	0,117	0	0,715	0,514	0,913	0,635	1742	695
								1	0,486	0,285	0,183	0,635	165	156
									0,688	0,487	0,828	0,635		
Support Vector Machines (SVM)	10	Cross-validation – folds 10	8265	59,939	5524	40,061	0,1397	0	0,581	0,249	0,949	0,666	7123	5145
								1	0,751	0,419	0,182	0,666	379	1142
									0,599	0,268	0,865	0,666		
		Percentage split 80%	1654	59,971	1104	40,029	0,138	0	0,583	0,271	0,942	0,656	1420	1017
								1	0,729	0,417	0,187	0,656	87	234
									0,6	0,288	0,854	0,656		
Decision Trees (DT)	10	Cross-validation – folds 10	9331	67,670	4458	32,330	0,116	0	0,696	0,48	0,921	0,595	8540	3728
								1	0,52	0,304	0,175	0,595	730	791
									0,677	0,461	0,839	0,595		
		Percentage split 80%	1863	67,549	895	32,451	0,107	0	0,7	0,511	0,912	0,574	1706	731
								1	0,489	0,3	0,177	0,574	164	157
									0,675	0,486	0,827	0,574		

Random Forests (RF)	10	Cross-validation – folds 10	10203	73,994	3586	26,006	0,0969	0	0,788	0,646	0,908	0,624	9664	2604
								1	0,354	0,212	0,171	0,624	982	539
									0,74	0,598	0,827	0,624		
		Percentage split 80%	2044	74,112	714	25,888	0,0799	0	0,798	0,688	0,898	0,619	1944	493
								1	0,312	0,202	0,169	0,619	221	100
									0,741	0,632	0,813	0,619		
K-Nearest Neighbors (KNN)	10	Cross-validation – folds 10	9444	68,489	4345	31,511	0,0917	0	0,714	0,552	0,913	0,591	8763	3505
								1	0,448	0,286	0,163	0,591	840	681
									0,685	0,523	0,83	0,591		
		Percentage split 80%	1915	69,434	843	30,566	0,0922	0	0,73	0,579	0,905	0,586	1780	657
								1	0,421	0,27	0,17	0,587	186	135
									0,694	0,543	0,82	0,586		

Anexo XV – Modelos e resultados: Base de dados com variáveis binárias penalizada – Acidentes do tipo despiste.

Acidentes do tipo Despiste - base numérica penalizada														
Modelo	Penalização	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	10	Cross-validation – folds 10	6150	52,894	5477	47,106	0,0789	0	0,508	0,309	0,927	0,64	5229	5065
								1	0,691	0,492	0,154	0,64	412	921
									0,529	0,33	0,838	0,64		
		Percentage split 80%	1225	52,688	1100	47,312	0,0743	0	0,507	0,32	0,925	0,64	1044	1015
								1	0,68	0,493	0,151	0,64	85	181
									0,527	0,339	0,836	0,64		
Artificial Neural Networks (ANN)	10	Cross-validation – folds 10	5867	50,460	5760	49,540	0,06	0	0,483	0,327	0,919	0,601	4970	5324
								1	0,673	0,517	0,144	0,601	436	897
									0,505	0,349	0,83	0,601		
		Percentage split 80%	1823	78,409	502	21,591	0,0792	0	0,855	0,763	0,897	0,603	1760	299
								1	0,237	0,145	0,174	0,603	203	63
									0,784	0,692	0,814	0,603		
Support Vector Machines (SVM)	10	Cross-validation – folds 10	4433	38,127	7194	61,873	0,0051	0	0,32	0,142	0,946	0,589	3289	7005
								1	0,858	0,68	0,14	0,589	189	1144
									0,381	0,204	0,853	0,589		
		Percentage split 80%	891	38,323	1434	61,677	0,0508	0	0,324	0,162	0,94	0,581	668	1391
								1	0,838	0,676	0,138	0,581	43	223
									0,383	0,22	0,848	0,581		
Decision Trees (DT)	10	Cross-validation – folds 10	6534	56,197	5093	43,803	0,0469	0	0,565	0,458	0,905	0,557	5812	4482
								1	0,542	0,435	0,139	0,557	611	722
									0,562	0,456	0,817	0,557		
		Percentage split 80%	1326	57,032	999	42,968	0,0599	0	0,571	0,436	0,91	0,57	1176	883
								1	0,564	0,429	0,145	0,57	116	150
									0,57	0,435	0,823	0,57		

Random Forests (RF)	10	Cross-validation – folds 10	7697	66,199	3930	33,801	0,0369	0	0,701	0,637	0,895	0,563	7213	3081
								1	0,363	0,299	0,136	0,563	849	484
									0,662	0,598	0,808	0,563		
		Percentage split 80%	1575	67,742	750	32,258	0,0639	0	0,714	0,605	0,901	0,573	1470	589
								1	0,395	0,286	0,151	0,573	161	105
									0,677	0,569	0,815	0,573		
K-Nearest Neighbors (KNN)	10	Cross-validation – folds 10	7019	60,368	4608	39,632	0,0395	0	0,623	0,542	0,899	0,548	6409	3885
								1	0,548	0,377	0,136	0,548	723	610
									0,604	0,523	0,811	0,548		
		Percentage split 80%	1398	60,129	927	39,871	0,0429	0	0,618	0,53	0,9	0,557	1273	786
								1	0,47	0,382	0,137	0,557	141	125
									0,601	0,513	0,813	0,557		

Anexo XVI – Modelos e resultados: Base de dados com variáveis binárias e RUS – Acidentes do tipo colisão.

Acidentes do tipo colisão - base numérica RUS													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	1885	61,966	1157	38,034	0,2393	0	0,559	0,32	0,636	0,662	850	671
							1	0,68	0,441	0,607	0,662	486	1035
								0,62	0,38	0,621	0,662		
	Percentage split 80%	363	59,704	245	40,296	0,1889	0	0,509	0,321	0,596	0,633	149	144
							1	0,679	0,491	0,598	0,633	101	214
								0,597	0,409	0,597	0,633		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	1799	59,139	1243	40,861	0,1828	0	0,596	0,413	0,591	0,627	906	615
							1	0,587	0,404	0,592	0,627	628	893
								0,591	0,409	0,591	0,627		
	Percentage split 80%	351	57,730	257	42,270	0,1604	0	0,679	0,517	0,55	0,614	199	94
							1	0,483	0,321	0,618	0,614	163	152
								0,577	0,416	0,585	0,614		
Decision Trees (DT)	Cross-validation – folds 10	1905	62,623	1137	37,377	0,2525	0	0,638	0,385	0,623	0,642	970	551
							1	0,615	0,362	0,629	0,642	586	935
								0,626	0,374	0,626	0,642		
	Percentage split 80%	373	61,349	235	38,651	0,2306	0	0,686	0,545	0,584	0,612	201	92
							1	0,546	0,314	0,652	0,612	143	172
								0,613	0,381	0,619	0,612		
Random Forests (RF)	Cross-validation – folds 10	1837	60,388	1205	39,612	0,2078	0	0,611	0,404	0,602	0,637	930	591
							1	0,596	0,389	0,605	0,637	614	907
								0,604	0,396	0,604	0,637		
	Percentage split 80%	355	58,388	253	41,612	0,1701	0	0,628	0,457	0,561	0,602	184	109
							1	0,543	0,372	0,611	0,602	144	171
								0,584	0,413	0,587	0,602		

Anexo XVII – Modelos e resultados: Base de dados com variáveis binárias e RUS – Acidentes do tipo despiste.

Acidentes do tipo despiste - base numérica RUS													
Modelo	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
			%		%							a	b
Naive Bayes	Cross-validation – folds 10	1591	59,677	1075	40,323	0,1935	0	0,536	0,342	0,61	0,63	714	619
							1	0,658	0,464	0,586	0,63	456	877
								0,597	0,403	0,598	0,63		
	Percentage split 80%	328	61,538	205	38,462	0,2288	0	0,556	0,327	0,62	0,664	145	116
							1	0,673	0,444	0,612	0,664	89	183
								0,615	0,387	0,616	0,664		
Artificial Neural Networks (ANN)	Cross-validation – folds 10	1475	55,326	1191	44,674	0,1065	0	0,515	0,409	0,558	0,575	687	646
							1	0,591	0,485	0,55	0,575	545	788
								0,553	0,447	0,554	0,575		
	Percentage split 80%	295	55,347	238	44,653	0,11499	0	0,778	0,662	0,53	0,597	203	58
							1	0,338	0,222	0,613	0,597	180	92
								0,553	0,438	0,573	0,597		
Decision Trees (DT)	Cross-validation – folds 10	1493	56,002	1173	43,998	0,12	0	0,53	0,41	0,564	0,58	706	627
							1	0,59	0,47	0,557	0,58	546	787
								0,56	0,44	0,56	0,58		
	Percentage split 80%	307	57,598	226	42,402	0,1515	0	0,563	0,412	0,568	0,598	147	114
							1	0,588	0,437	0,584	0,598	112	160
								0,576	0,425	0,576	0,598		
Random Forests (RF)	Cross-validation – folds 10	1432	53,713	1234	46,287	0,0743	0	0,536	0,462	0,537	0,566	715	618
							1	0,538	0,464	0,537	0,566	616	717
								0,537	0,462	0,537	0,566		
	Percentage split 80%	275	51,595	258	48,405	0,0321	0	0,521	0,489	0,506	0,543	136	125
							1	0,511	0,479	0,527	0,543	133	139
								0,516	0,484	0,516	0,543		

Anexo XVIII – Modelos e resultados: Base de dados com variáveis binárias e ROS (SMOTE) – Acidentes do tipo colisão.

Acidentes do tipo colisão - base numérica SMOTE														
Modelo	SMOTE	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	600% (100+100+50)	Cross-validation – folds 10	14363	67,136	7031	32,864	0,334	0	0,688	0,351	0,725	0,724	8443	3825
								1	0,649	0,312	0,607	0,724	3206	5920
									0,671	0,334	0,675	0,724		
		Percentage split 80%	2831	66,160	1448	33,840	0,3159	0	0,671	0,351	0,721	0,716	1652	809
								1	0,649	0,329	0,593	0,716	639	1179
									0,662	0,342	0,667	0,716		
Artificial Neural Networks (ANN)	600% (100+100+50)	Cross-validation – folds 10	16416	76,732	4978	23,268	0,5206	0	0,82	0,303	0,784	0,835	10059	2209
								1	0,697	0,18	0,742	0,835	2769	6357
									0,767	0,251	0,766	0,835		
		Percentage split 80%	3263	76,256	1016	23,744	0,5142	0	0,794	0,279	0,794	0,838	1953	508
								1	0,721	0,206	0,721	0,838	508	1310
									0,763	0,248	0,763	0,838		
Decision Trees (DT)	600% (100+100+50)	Cross-validation – folds 10	16951	79,232	4443	20,768	0,5756	0	0,818	0,242	0,819	0,843	10038	2230
								1	0,758	0,182	0,756	0,843	2213	6913
									0,792	0,217	0,792	0,843		
		Percentage split 80%	3403	79,528	876	20,472	0,5815	0	0,82	0,238	0,824	0,85	2017	444
								1	0,762	0,18	0,757	0,85	432	1386
									0,765	0,213	0,795	0,85		
Random Forests (RF)	600% (100+100+50)	Cross-validation – folds 10	17284	80,789	4110	19,211	0,6068	0	0,837	0,231	0,83	0,886	10263	2005
								1	0,769	0,163	0,778	0,886	2105	7021
									0,808	0,202	0,808	0,886		
		Percentage split 80%	3474	81,187	805	18,813	0,6156	0	0,833	0,216	0,839	0,889	2049	412
								1	0,784	0,167	0,776	0,889	393	1425
									0,812	0,195	0,812	0,889		

Anexo XIX – Modelos e resultados: Base de dados com variáveis binárias e ROS (SMOTE) – Acidentes do tipo despiste.

Acidentes do tipo despiste - base numérica SMOTE														
Modelo	SMOTE	Teste	Clas. corretamente		Clas. incorretamente		Kappa statistic	Class	TP Rate	FP Rate	Precision	ROC Area	Confusion Matrix	
				%		%							a	b
Naive Bayes	600% (100+100+50)	Cross-validation – folds 10	11653	63,705	6639	36,295	0,2684	0	0,65	0,379	0,688	0,688	6687	3607
								1	0,621	0,35	0,579	0,688	3032	4966
									0,367	0,367	0,64	0,688		
		Percentage split 80%	2324	63,532	1334	36,468	0,2654	0	0,654	0,378	0,688	0,693	1329	730
								1	0,622	0,355	0,577	0,693	604	995
									0,635	0,368	0,639	0,693		
Artificial Neural Networks (ANN)	600% (100+100+50)	Cross-validation – folds 10	13418	73,354	4874	26,646	0,462	0	0,741	0,276	0,779	0,809	7625	2669
								1	0,724	0,259	0,685	0,809	2205	5793
									0,734	0,269	0,736	0,809		
		Percentage split 80%	2703	73,893	955	26,107	0,4654	0	0,795	0,333	0,755	0,822	1636	423
								1	0,667	0,205	0,716	0,822	532	1067
									0,739	0,277	0,738	0,822		
Decision Trees (DT)	600% (100+100+50)	Cross-validation – folds 10	14104	77,105	4188	22,895	0,5381	0	0,771	0,229	0,812	0,831	7938	2356
								1	0,771	0,229	0,724	0,831	1832	6166
									0,771	0,229	0,774	0,831		
		Percentage split 80%	2825	77,228	833	22,772	0,534	0	0,754	0,205	0,826	0,828	1553	506
								1	0,795	0,246	0,715	0,828	327	1272
									0,772	0,223	0,778	0,828		
Random Forests (RF)	600% (100+100+50)	Cross-validation – folds 10	14411	78,783	3881	21,217	0,5714	0	0,791	0,216	0,825	0,868	8142	2152
								1	0,784	0,209	0,744	0,868	1729	6269
									0,788	0,213	0,79	0,868		
		Percentage split 80%	2882	78,786	776	21,214	0,5731	0	0,778	0,199	0,834	0,871	1601	458
								1	0,801	0,222	0,737	0,871	318	1281
									0,788	0,209	0,792	0,871		