



**INSTITUTO
FEDERAL**

Ceará

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ

CAMPUS FORTALEZA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO

RODRIGO MESQUITA DE VASCONCELOS

**APRAVC: UMA NOVA ABORDAGEM DE APOIO A PREVENÇÃO DA
RECORRÊNCIA DO AVC**

FORTALEZA

2023

RODRIGO MESQUITA DE VASCONCELOS

APRAVC: UMA NOVA ABORDAGEM DE APOIO A PREVENÇÃO DA RECORRÊNCIA
DO AVC

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Sistemas de Computação

Linha de Pesquisa: Inteligência Artificial

Orientador: Dr. Carlos Maurício Jaborandy de Mattos Dourado Júnior

Coorientadora: Dra. Lidiane Andréa Oliveira Lima

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

- V331a Vasconcelos, Rodrigo.
APRAVC: UMA NOVA ABORDAGEM DE APOIO A PREVENÇÃO DA RECORRÊNCIA DO AVC /
Rodrigo Vasconcelos. - 2023.
113 f. : il. color.
- Dissertação (Mestrado) - Instituto Federal do Ceará, Mestrado em Ciência da Computação, Campus
Fortaleza, 2023.
Orientação: Prof. Dr. Carlos Maurício Jaborandy de Mattos Dourado Júnior.
Coorientação: Profa. Dra. Lidiane Andréa Oliveira Lima.
1. Acidente Vascular Cerebral. 2. Aprendizagem de máquina. 3. Inteligência Artificial. I. Título.
CDD 004
-

RODRIGO MESQUITA DE VASCONCELOS

APRAVC: UMA NOVA ABORDAGEM DE APOIO A PREVENÇÃO DA RECORRÊNCIA
DO AVC

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Sistemas de Computação

Linha de Pesquisa: Inteligência Artificial

Aprovada em:

BANCA EXAMINADORA

Dr. Carlos Maurício Jaborandy de Mattos Dourado Júnior (Orientador)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Dra. Lidiane Andréa Oliveira Lima (Coorientadora)
Universidade Federal do Ceará (UFC)

Dr. Pedro Pedrosa Rebouças Filho
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Dr. Antonio Carlos da Silva Barros
Universidade da Integração Internacional da Lusofonia Afro-Brasileira
(UNILAB)

Dr. Roger Moura Sarmiento
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

AGRADECIMENTOS

Nos difíceis caminhos da vida não estamos sozinhos, e nesta caminhada encontrei apoio que vale a pena ressaltar;

Antes de tudo o meu reconhecimento a Deus pelo seu infinito amor e cuidado despendido durante toda minha vida.

Ressalto a paciência e a fortaleza de minha esposa que com cuidado e dedicação administrou todo o entorno de nosso lar para proporcionar a tranquilidade necessária na tarefa de realizar os estudos imprescindíveis na conclusão deste trabalho.

Lembro aqui também meu núcleo familiar, meu Pai, minha Mãe e minha Avó Maria (inmemoriam), pela luta e pelo incentivo à educação, fazendo com que eu guardasse a premissa de que “a educação é a responsável pelo avanço da humanidade.”.

Não poderia deixar de citar aqueles que me incentivaram e deram valorosa contribuição para a conclusão deste trabalho. Na revisão destaco Amanda Ferreira e Tia Veronica. Aqui também quero lembrar de ser grato ao Laboratório Boitátá e ao Mardônio França que cooperaram diretamente, para que eu pudesse crescer e evoluir em conhecimento na área de aprendizagem de máquina. Meu agradecimento ao meu amigo Vladmir Bezerra que foi um grande incentivador pessoal para que eu pudesse concluir o programa do mestrado. Deixo também registrado o meu agradecimento a minha amiga de mestrado Neila Matos pelo companheirismo e amizade nos momentos mais difíceis do mestrado.

Quero agradecer aos profissionais da Liga Fisioneuro da UFC que foram verdadeiros companheiros na orientação e supervisão sobre o AVC. Agradecimento também a todos os amigos que incentivaram e me desejaram força e animo para que eu concluísse o mestrado.

Agradecimento aos professores Carlos Maurício e Lidiane Andréa meus orientadores, pela amizade, sabedoria vencendo comigo todos os obstáculos desta caminhada.

“Seja você quem for, seja qual for a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.”

(Ayrton Senna)

RESUMO

Nos últimos anos, o Brasil tem enfrentado uma crescente problemática em relação ao AVC e sua recorrência, associados a fatores de risco modificáveis e não modificáveis. O AVC é reconhecido como uma das principais causas de morte em todo o mundo, bem como um grande problema para a sociedade e governos, devido à grande quantidade de pessoas incapacitadas e internações graves. Diante dessa situação, esta dissertação apresenta uma nova abordagem para a prevenção da recorrência do AVC, com o objetivo de realizar um diagnóstico precoce utilizando técnicas de aprendizagem de máquina. Para isso, utilizou-se uma base de dados fornecida pela Liga Fisioneuro da UFC, contendo informações apenas de pacientes do estado do Ceará. Os dados foram preparados através da engenharia de atributos, resultando em uma base sólida para treinamento e teste. O estudo se concentrou em três grupos categóricos de técnicas de aprendizagem de máquina: métodos clássicos, métodos ensemble e redes neurais artificiais, em conjunto com quatro tipos de abordagens de balanceamento de cargas. Após treinamento e teste dos modelos, o AprAVC foi avaliado com métricas de performance de aprendizado de máquina selecionadas para este trabalho. O classificador *XGBoost*, utilizando a abordagem de balanceamento de cargas *oversampling*, obteve o melhor desempenho, com uma excelente capacidade de distinguir entre classes positivas e negativas, alcançando 97,31% de performance com a métrica da área sob a curva ROC. O AprAVC também obteve ótima performance nas outras três abordagens apresentadas no trabalho. Os resultados obtidos comprovam a viabilidade de utilizar algoritmos de aprendizagem de máquina para auxiliar na prevenção da recorrência do AVC mesmo com cargas desbalanceadas. O AprAVC pode contribuir significativamente para profissionais da área de saúde e pessoas em geral, ajudando na previsão da recorrência do AVC.

Palavras-chave: Acidente Vascular Cerebral, Inteligência Artificial, Aprendizagem de máquina

ABSTRACT

In recent years, Brazil has faced a growing problem in relation to stroke and its recurrence, associated with modifiable and non-modifiable risk factors. Stroke is recognized as one of the main causes of death worldwide, as well as a major problem for society and governments, due to the large number of disabled people and serious hospitalizations. Faced with this situation, this dissertation presents a new approach to preventing stroke recurrence, with the objective of performing an early diagnosis using machine learning techniques. For this, a database provided by Liga Fisioneuro of UFC was used, containing information only from patients in the state of Ceará. The data was prepared through feature engineering, resulting in a solid foundation for training and testing. The study focused on three categorical groups of machine learning techniques: classical methods, ensemble methods, and artificial neural networks, together with four types of load-balancing approaches. After training and testing the models, AprAVC was evaluated with machine learning performance metrics selected for this work. The classifier XGBoost, using the load balancing approach oversampling, achieved the best performance, with an excellent ability to distinguish between positive and negative classes, reaching 97.31% performance with the metric of the area under the ROC curve. AprAVC also achieved excellent performance in the other three approaches presented in work. The results obtained prove the feasibility of using machine learning algorithms to help prevent stroke recurrence even with unbalanced loads. AprAVC can significantly contribute to healthcare professionals and people in general, helping to predict stroke recurrence.

Keywords: Stroke, Artificial Intelligence, Machine Learning

LISTA DE FIGURAS

Figura 1 – Fluxo resumido de problemas de classificação	38
Figura 2 – Mapa de aprendizagem maquina	39
Figura 3 – Árvore de decisão (à esquerda) e o espaço de hipóteses (à direita).	40
Figura 4 – Árvore de Decisão.	41
Figura 5 – Curva de Regressão Logística.	42
Figura 6 – Como funciona o algoritmo KNN.	46
Figura 7 – Passos sequenciais do algoritmo KNN.	47
Figura 8 – Hiperplano - SVM	49
Figura 9 – Dados lineares e dados não lineares	50
Figura 10 – Função <i>Kernel</i>	51
Figura 11 – Crescimento de árvore em nível	55
Figura 12 – Esquema de funcionamento do <i>Random Forest</i>	58
Figura 13 – Rede Neural Artificial	60
Figura 14 – Conceito de Rede Neural Artificial	60
Figura 15 – Perceptron Multicamada - MLP	62
Figura 16 – Processo de Validação Cruzada	64
Figura 17 – Matriz de confusão.	66
Figura 18 – Curva de ROC.	69
Figura 19 – Etapa 1 - Conjunto de Dados	76
Figura 20 – Etapa 2 - Execução e aprimoramento do AprAVC	77
Figura 21 – Base de dados desbalanceada.	82
Figura 22 – Matriz de correlação dos atributos.	83

LISTA DE TABELAS

Tabela 1 – As métricas para avaliar métodos de classificação.	66
Tabela 2 – Apresentação dos trabalhos relacionados	74
Tabela 3 – Comparação dos trabalhos relacionados por métodos, métricas e resultados .	75
Tabela 4 – Descrição dos atributos.	81
Tabela 5 – Seleção de atributos por <i>score</i>	84
Tabela 6 – Aprendizagem de máquina	87
Tabela 7 – Parâmetros por algoritmo	88
Tabela 8 – Valores da Acurácia para a validação cruzada - BCO e BCU	89
Tabela 9 – Valores da Acurácia para a validação cruzada - BCOU e BCUC	89
Tabela 10 – Matriz de confusão para o problema de classificação binário proposto	91
Tabela 11 – Matriz de confusão para cada classificador - abordagem BCO	91
Tabela 12 – Matriz de confusão para cada classificador - abordagem BCU	92
Tabela 13 – Matriz de confusão para cada classificador - abordagem BCOU	92
Tabela 14 – Matriz de confusão para cada classificador - abordagem BCUC	93
Tabela 15 – Tempo de processamento de treino por classificador e tipo de abordagem . .	93
Tabela 16 – Tempo de processamento com a base de teste por classificador e tipo de abordagem	94
Tabela 17 – Tempo de processamento total por classificador e abordagem	94
Tabela 18 – Tabela de resultados (Acurácia) - Métodos Clássicos	95
Tabela 19 – Tabela de resultados (Acurácia) - Métodos Ensemble	95
Tabela 20 – Tabela de resultados (Acurácia) - Redes Neurais	95
Tabela 21 – Melhor método por abordagem (Acurácia)	96
Tabela 22 – Tabela de resultados (Precisão) - Métodos Clássicos	96
Tabela 23 – Tabela de resultados (Precisão) - Métodos Ensemble	96
Tabela 24 – Tabela de resultados (Precisão) - Redes Neurais	97
Tabela 25 – Melhor método por abordagem Precisão	97
Tabela 26 – Tabela de resultados (<i>Recall</i>) - Métodos Clássicos	97
Tabela 27 – Tabela de resultados (<i>Recall</i>) - Métodos Ensemble	98
Tabela 28 – Tabela de resultados (<i>Recall</i>) - Redes Neurais	98
Tabela 29 – Melhor método por abordagem <i>Recall</i>	98
Tabela 30 – Tabela de resultados (<i>F1 Score</i>) - Métodos Clássicos	99

Tabela 31 – Tabela de resultados (<i>F1 Score</i>) - Métodos Ensemble	99
Tabela 32 – Tabela de resultados (<i>F1 Score</i>) - Redes Neurais	100
Tabela 33 – Melhor método por abordagem (<i>F1 score</i>)	100
Tabela 34 – Tabela de resultados (AUC de ROC) - Métodos Clássicos	100
Tabela 35 – Tabela de resultados (AUC de ROC) - Métodos Ensemble	101
Tabela 36 – Tabela de resultados (AUC de ROC) - Redes Neurais	101
Tabela 37 – Melhor método por abordagem (AUC de ROC)	101
Tabela 38 – Melhores Performances	102

LISTA DE ABREVIATURAS E SIGLAS

Acc	Acurácia
AED	Análise Exploratória de Dados
AIT	Ataque Isquêmico Transitório
ASA	<i>American Stroke Association</i>
AUC	<i>Area Under the Curve</i>
AVC	Acidente Vascular Cerebral
BCO	Balanceamento de Cargas com <i>oversampling</i>
BCOU	Balanceamento de Cargas com <i>oversampling</i> mais <i>undersampling</i>
BCU	Balanceamento de Cargas com <i>undersampling</i>
BCUO	Balanceamento de Cargas com <i>undersampling</i> mais <i>oversampling</i>
BPNN	Back-propagation neural network
CART	Árvore de classificação e regressão
CatBoost	<i>Categorical Boosting</i>
DNN	<i>Deep Neural Network</i>
FN	Falsos negativos
FP	Falsos positivos
GBDT	<i>Gradient Boosting Decision Tree</i>
GOSS	<i>Gradient based one-sided sampling</i>
HAS	Hipertensão Arterial Sistêmica
IA	Inteligência Artificial
IMC	Índice de Massa Corporal
KNN	<i>K-Nearest Neighbor</i>
LightGBM	<i>Light Gradient Boosting Machine</i>
MLP	<i>Multi Layer Perceptron</i>
NB	<i>Neive Bayes</i>
OMS	Organização Mundial da Saúde
OPAS	Organização Pan-Americana da Saúde
PA	Pressão Arterial
Prec	Precisão
RFE	Recursive Feature Elimination
ROC	<i>Receiver Operating Characteristic</i>

Sens	Sensibilidade
SIM	Sistema de Informações sobre Mortalidade
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
SVM	<i>Support Vector Machine</i>
TFP	Taxa de Falsos Positivos
TPT	Tempo de Processamento Total
TVP	Taxa de Verdadeiros Positivos
UFC	Universidade Federal do Ceará
VN	Verdadeiros negativos
VP	Verdadeiros positivos
XGBoost	<i>Extreme Gradient Boosting</i>

LISTA DE SÍMBOLOS

s	Segundos
-----	----------

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Contextualização	16
1.2	Motivação	20
1.3	Objetivos	21
1.3.1	<i>Objetivos específicos</i>	21
1.4	Produção Científica	22
1.5	Organização deste trabalho	22
2	ACIDENTE VASCULAR CEREBRAL	23
2.1	Promoção da saúde e prevenção da doença	23
2.2	Tipos de AVC	24
2.3	Fatores de risco para AVC	25
2.4	Deteção e diagnóstico	26
2.5	Tratamento e recomendações	27
3	APRENDIZAGEM DE MÁQUINA	29
3.1	Engenharia de atributos	30
3.1.1	<i>Criação de Atributos</i>	31
3.1.2	<i>Transformação de Atributos</i>	31
3.1.3	<i>Seleção de Atributos</i>	34
3.2	Cargas desbalanceadas	36
3.3	Aprendizagem de máquina supervisionada para classificação	37
3.3.1	<i>Técnicas Clássicas</i>	39
3.3.2	<i>Métodos Ensemble</i>	53
3.3.3	<i>RNA</i>	59
3.3.4	<i>Validação cruzada</i>	63
3.3.5	<i>As métricas de aprendizagem de máquina</i>	64
4	REVISÃO BIBLIOGRÁFICA	70
4.1	Apresentação dos trabalhos	70
4.2	Comparativo dos Trabalhos Relacionados	73
5	METODOLOGIA	76
5.1	Conjunto de Dados	78

5.1.1	<i>Entendimento do domínio</i>	78
5.1.2	<i>Construção do dataset</i>	78
5.1.3	<i>Preparação do dataset</i>	81
5.2	Execução e aprimoramento do AprAVC	84
5.2.1	<i>Divisão da base</i>	84
5.2.2	<i>Geração do modelo</i>	85
5.3	Teste, avaliação e performance do modelo	90
6	RESULTADOS E DISCUSSÃO	91
6.1	Resultados - Matriz de confusão	91
6.2	Resultados - Tempo de processamento	93
6.3	Resultados - Acurácia	94
6.4	Resultados - Precisão	96
6.5	Resultados - Recall	97
6.6	Resultados - F1 Score	99
6.7	Resultados - AUC de ROC	100
6.8	Avaliação de desempenho	101
7	CONCLUSÕES	103
7.1	Trabalhos futuros	104
	REFERÊNCIAS	106

1 INTRODUÇÃO

1.1 Contextualização

A Organização Mundial da Saúde (OMS) classifica o Acidente Vascular Cerebral (AVC) como uma das principais causas de morte em todo o mundo. O relatório de 2019 da OMS mostrou que o AVC foi responsável por cerca de 13% das mortes em todo o mundo naquele ano, totalizando aproximadamente 6,5 milhões de óbitos. De acordo com Ren *et al.* (2021) em média por ano mundialmente cerca de 15 milhões de pessoas sofrem AVC. É importante destacar que o número real de mortes por AVC pode ser ainda maior, uma vez que nem todos os casos são diagnosticados ou reportados. Além disso, a pandemia de COVID-19 pode ter afetado o diagnóstico e tratamento adequado de outras doenças, incluindo as cerebrovasculares (OMS, 2020).

O estudo conduzido por Feigin *et al.* (2003) revelou que em 1999, houve 5,54 milhões de mortes por AVC em todo o mundo, sendo que 66,66% dessas mortes ocorreram em países menos desenvolvidos. Já em 2005, o número de mortes registradas, segundo Bonita e Beaglehole (2007), foi de 5,7 milhões, já Silva *et al.* (2020) informou que em 2015 o AVC foi responsável por 6,2 milhões de óbitos. Dados apontam que, sem intervenções, esse número pode aumentar chegando a 7,8 milhões de mortes em 2030, com a maioria das mortes ocorrendo em países pobres.

De acordo com o banco de dados da OMS, a cada ano, o número de mortes por AVC vem aumentando. No entanto, é importante notar que, sem ações efetivas de prevenção e tratamento, espera-se um aumento no número de mortes por AVC em todo o mundo nas próximas décadas, especialmente em países de baixa e média renda.

O AVC é reconhecido como uma das principais causas de morte em todo o mundo, bem como um grande problema para a sociedade e governos, devido à grande quantidade de pessoas incapacitadas e internações graves. De acordo com Botelho *et al.* (2016), o AVC é responsável por uma quantidade significativa de incapacidades. Na China, o país mais populoso do mundo, o AVC tornou-se a principal causa de morte e incapacidade. Para reduzir a incapacidade e a mortalidade causadas pelo AVC, muitos pesquisadores têm se dedicado a estudos em diversas áreas, incluindo etiologia, diagnóstico e tratamento. Como afirmado por Chang *et al.* (2019), o objetivo desses estudos é fazer avanços significativos na prevenção e tratamento do AVC.

De acordo com a *American Stroke Association* (ASA), cerca de 1 em cada 4 sobre-

viventes de AVC sofre uma recorrência do evento. Estatísticas médicas recentes indicam que a taxa de mortalidade e incapacitação após a ocorrência de um novo AVC é significativamente maior do que no primeiro evento, o que destaca a importância do monitoramento de risco mais próximo e ajustes na terapia em pacientes que já tiveram um AVC (CHANG *et al.*, 2019).

Após a alta hospitalar, muitos pacientes que sobreviveram a um AVC retornam para suas casas sem receber informações adequadas sobre os cuidados e recursos disponíveis para a prevenção secundária. Isso os torna suscetíveis a recorrências do AVC, o que enfatiza a importância de um monitoramento mais próximo e ajustes de terapia em pacientes que já tiveram um AVC (BOCCHI, 2004).

Durante o primeiro ano após a fase aguda, o risco de recorrência de AVC varia de 6% a 12%. Segundo Fernandez-Cadenas *et al.* (2017), os números sugerem que, dentro de um intervalo de 10 anos após o AVC, cerca de 60% dos pacientes morrerão e 54% terão um novo evento vascular. Portanto, é fundamental fornecer aos pacientes informações e orientações sobre os cuidados de prevenção secundária após a alta hospitalar para minimizar os riscos de recorrência do AVC.

No continente europeu a taxa de mortalidade ao longo de cinco anos após ao primeiro evento chega a 60%, mostrando assim a importância da prevenção secundária, já que existem múltiplos eventos que aumentam a probabilidade de gravidade de um risco de maior da recorrência do AVC, e consequentemente aumentam as chances de o indivíduo tornar-se dependente de pessoas, o que implica em custos financeiros e emocionais para a família e maior carga no sistema de saúde (KULESH *et al.*, 2010).

Nas últimas décadas, o Brasil tem apresentado uma importante transição no perfil epidemiológico com notoriedade para o AVC. O AVC é uma das principais causas de internações e mortalidade no país, dimensionando a sua magnitude como problema de saúde pública e retratando a alta prevalência na população. Segundo o MINISTÉRIO DA SAÚDE (2022a) o AVC é a causa mais frequente de óbito na população adulta no Brasil sendo responsável todo anos por 100 mil mortes e consistindo no diagnóstico de 10% das internações hospitalares públicas, sendo é a primeira causa em incapacidade e sequela (SCHMIDT *et al.*, 2019).

Segundo o Ministério da Saúde o Brasil possui uma das maiores taxas de mortalidade por AVC na América Latina. Segundo dados da Organização Pan-Americana da Saúde (OPAS), o Brasil tem a quarta maior taxa de mortalidade por AVC entre os países da América Latina e Caribe, ficando abaixo apenas de Cuba, Suriname e Belize. Um estudo prospectivo nacional

do AVC em 2013 indicou incidência anual de 108 casos por 100 mil habitantes com taxa de fatalidade aos 30 dias de 18,5% e aos 12 meses de 30,9%, e índice de recorrência após um 1 ano de 15,9%, sendo a recorrência mais severa do que a primeira vez (MINISTÉRIO DA SAÚDE, 2022b).

O estado do Ceará nos últimos anos, tem registrado um aumento na taxa de incidência do AVC. De acordo com dados do Ministério da Saúde, a taxa de incidência de AVC no estado do Ceará passou de 81,8 casos por 100.000 habitantes em 2010 para 100,8 casos por 100.000 habitantes em 2019. Isso representa um aumento de cerca de 23% na taxa de incidência da doença em apenas uma década. Segundo a Fundação Regional de Saúde do Ceará de 2009 a 2020, o Ceará registrou uma média anual de 7.981,5 internações por AVC, com aumento gradativo a partir de 2012 e ápice em 2019, com 9.485 internações por ano (FUNDAÇÃO REGIONAL DA SAÚDE DO CEARÁ, 2022).

De acordo com dados do Sistema de Informações sobre Mortalidade (SIM) do Ministério da Saúde, em 2019, o AVC é uma das principais causas de morte do Ceará. As doenças do sistema nervoso (que incluem o AVC) foram responsáveis por 2.369 óbitos em todo o estado, o que representa cerca de 3,3% do total de óbitos registrados no período. Segundo o jornal Diário do Nordeste no período de janeiro a agosto do ano de 2020, mais de 3.500 cearenses foram internados e 560 morreram após AVC (DIÁRIO DO NORDESTE, 2020).

A conscientização geral sobre a importância de exames preventivos regulares, adoção de hábitos saudáveis e diagnósticos precoces é crucial para prevenir a recorrência do AVC. Além disso, os governos e as autoridades de saúde devem implementar programas de saúde pública que incentivem a prevenção e o diagnóstico precoce, bem como garantir acesso a serviços de saúde de qualidade para todas as pessoas. Isso é importante para reduzir o impacto da doença na saúde da população e nos gastos do sistema de saúde. Portanto, é necessário um esforço conjunto para promover a prevenção e o tratamento adequado do AVC.

Nos últimos anos, a tecnologia tem avançado rapidamente, e a Inteligência Artificial (IA) tem se destacado como uma das áreas de maior crescimento. A IA é um campo da computação que busca desenvolver algoritmos e técnicas capazes de permitir que máquinas realizem tarefas que normalmente exigem inteligência humana, como reconhecimento de padrões, aprendizado, planejamento, raciocínio e autocorreção. A IA possui uma ampla gama de aplicações em diversos setores, tais como a automação de processos, o reconhecimento de voz e imagem e a análise de dados. Por exemplo, a automação de processos pode ser usada para otimizar o

desempenho de uma fábrica, tornando-a mais eficiente e reduzindo custos.

Com o avanço contínuo da IA, espera-se que essa tecnologia continue revolucionando a maneira como as empresas operam e como as pessoas interagem com o mundo digital. No entanto, é importante considerar os impactos éticos e sociais da implementação da IA para garantir que esses avanços sejam realizados de maneira responsável e benéfica para a sociedade como um todo.

A Inteligência Artificial desempenha um papel crucial em diversos campos, incluindo a medicina e a análise de dados. Na medicina, a IA possibilita o diagnóstico precoce de doenças e a prevenção de acidentes, contribuindo para um tratamento mais eficaz e salvando vidas. Na área da análise de dados, a IA está cada vez mais presente, permitindo que empresas obtenham *insights* valiosos e tomem decisões mais informadas e baseadas em dados. Isso pode resultar em um aumento de eficiência e competitividade para as empresas que adotam essa tecnologia.

A IA também é um componente chave na robótica, permitindo que os robôs sejam cada vez mais avançados e capazes de realizar tarefas complexas em setores como manufatura, transporte e logística. Isso tem ajudado a melhorar a eficiência e a segurança em vários setores.

No entanto, é importante destacar que a IA também traz consigo desafios éticos e de segurança, como a possibilidade de vieses e discriminação algorítmica, e o risco de falhas no sistema. É crucial que esses desafios sejam abordados e que a implementação da IA seja feita de forma responsável e transparente.

Com o desenvolvimento contínuo das pesquisas em IA, estatística e teoria da computação, surgiu uma subárea chamada de Aprendizado de Máquina. O Aprendizado de Máquina se concentra em desenvolver algoritmos e modelos capazes de aprender com dados e realizar tarefas sem serem explicitamente programados para isso. Na área da medicina, o Aprendizado de Máquina tem o potencial de gerar aplicações importantes, tais como:

- Sistemas de diagnóstico mais precisos: os algoritmos de Aprendizado de Máquina podem aprender com dados de diagnóstico médico para identificar padrões e sinais que possam indicar uma doença, auxiliando os profissionais médicos a realizar um diagnóstico mais preciso (VODENCAREVIC *et al.*, 2022).
- Previsão de riscos: o Aprendizado de Máquina pode ser aplicado para identificar quais pacientes correm maior risco de desenvolver uma doença ou complicação, permitindo que os profissionais de saúde tomem medidas preventivas e de tratamento antecipadamente (NÓBREGA *et al.*, 2018).

- Desenvolvimento de novos medicamentos: os algoritmos de Aprendizado de Máquina podem analisar grandes conjuntos de dados sobre as propriedades químicas e biológicas das moléculas para prever sua eficácia como medicamento, agilizando o processo de desenvolvimento de novos medicamentos e reduzindo os custos envolvidos.
- Personalização de tratamentos: Sistemas de aprendizado de máquina podem ser usados para personalizar tratamentos para pacientes individuais com base em suas características únicas, incluindo genética, estilo de vida e histórico médico.
- Análise de dados clínicos: A aprendizagem de máquina pode ajudar a identificar padrões e tendências nos dados clínicos, o que pode ser útil para a pesquisa médica e a melhoria da qualidade do atendimento (JUNIOR *et al.*, 2019; NÓBREGA *et al.*, 2018).

A aplicação da aprendizagem de máquina ao AVC pode gerar impactos significativos no desenvolvimento de soluções mais precisas e personalizadas para o diagnóstico, prevenção e tratamento da doença. Através da análise de grandes quantidades de dados médicos, incluindo imagens, resultados de exames e históricos clínicos, é possível identificar padrões e tendências que auxiliam no diagnóstico precoce e na prevenção do AVC. Além disso, essa abordagem pode ser aplicada na personalização de tratamentos, levando em consideração as características individuais e histórico médico dos pacientes.

1.2 Motivação

Considerando o exposto, verifica-se a existência de um problema a ser enfrentado. Este trabalho propõe uma nova abordagem de apoio para a prevenção da recorrência do AVC por meio da aplicação de técnicas de aprendizagem de máquina. As sugestões são embasadas em análises de dados que automatizam a construção de modelos analíticos. O intuito é permitir que a máquina aprenda com os dados, identifique padrões e apresente sugestões com o mínimo de intervenção humana necessária.

No entanto, o acesso limitado a bases de dados confiáveis sobre a recorrência do AVC é um obstáculo significativo para a pesquisa nesta área. De acordo com Sacco *et al.* (1982), informações sobre recorrência e letalidade do AVC em países de baixa ou média renda, como o Brasil, já eram escassas durante o período de sua pesquisa, e infelizmente isso ainda é verdade nos dias atuais. A maioria das bases de dados sobre AVC é coletada em hospitais e centros de pesquisa, e muitas vezes não está disponível para o público em geral ou para pesquisadores fora dessas instituições.

Além disso, é importante destacar que ainda se sabe pouco sobre a frequência e os determinantes da recorrência do AVC, e esses fatores podem variar consideravelmente entre diferentes populações, como aponta Chaves (2000). Por esse motivo, é fundamental investigar as causas desse fenômeno em todo o mundo.

O estudo utilizará uma base de dados composta por informações de três fontes diferentes, contendo informações de pacientes residentes no estado do Ceará. A primeira fonte é composta por dados coletados por uma pesquisa realizada pela Liga Fisioneuro da Universidade Federal do Ceará (UFC) para acompanhamento de pacientes por meio de anamnese. A segunda fonte é fornecida pela mesma Liga e refere-se ao acompanhamento de pacientes do Hospital Geral de Fortaleza. Já a terceira fonte é uma base gerada por meio de um formulário online, que foi acompanhado por profissionais da área de saúde, como médicos e fisioterapeutas. Essas bases de dados incluem informações de pacientes que sofreram um AVC como primeiro evento, pacientes que sofreram recorrência do AVC e pacientes que não passaram pelo AVC.

Nessa perspectiva, o presente trabalho busca apresentar uma nova abordagem que vai utilizar alguns métodos de aprendizado de máquina supervisionado de ordem classificatória. Será obtido de cada método utilizado a sua confiabilidade da estimativa consumindo os dados. De forma a determinar qual técnica foi mais precisa em identificar que as pessoas acometidas de AVC, tiveram uma recorrência.

1.3 Objetivos

O objetivo deste trabalho é apresentar uma nova abordagem para apoiar na prevenção da recorrência do AVC por meio de técnicas de aprendizagem de máquina para fins de um diagnóstico precoce.

1.3.1 Objetivos específicos

Os objetivos específicos deste trabalho são:

- Apresentar uma nova base de dados para estudo da recorrência do AVC;
- Apresentar uma nova abordagem utilizando técnicas de aprendizagem de máquina que tenha uma boa performance para bases desbalanceadas e com poucos dados;
- Determinar a técnica de aprendizagem de máquina que possui melhor performance para identificar uma possível recorrência do AVC com a base de dados apresentada;

- Ajudar as pessoas que já sofreram AVC, identificarem de forma antecipada uma possível recorrência do AVC.

1.4 Produção Científica

Durante o mestrado, o seguinte resumo foi publicado:

- **VASCONCELOS, R. M. d.; SILVA, L. K. C.; SOUZA, C. C. C. d.; COSTA, D. L. B.; GALVÃO, W. R.; JUCÁ, R. V. B. d. M.; VIANA, R. T.; LIMA, L. A. O.; JÚNIOR, C. M. J. d. M. D. Validation of an artificial intelligence in the prediction of stroke recurrence: preliminary results.** In: LETRA1, E. (Ed.). **XXX Congresso Brasileiro de Neurologia.** [S.l.]: Arquivos de Neuro-Psiquiatria, 2022. v. 80, n. 9, p. 228. ISSN 0004-282X. Fator de impacto 2.035.

1.5 Organização deste trabalho

O restante do trabalho está organizado da seguinte maneira: no capítulo 2 e no capítulo 3, apresenta-se a fundamentação teórica que sustenta esta dissertação de mestrado. No capítulo 2, destaca-se a subseção que discute os fatores de risco para o AVC. Já no capítulo 3, há ênfase nas seções de engenharia de atributos e cargas desbalanceadas. O capítulo 4 detalha a revisão bibliográfica e compara os trabalhos relacionados apresentados nos capítulos anteriores; o capítulo 5 descreve a construção do *dataset*, bem como a execução e o aprimoramento do AprAVC; o capítulo 6 apresenta os resultados obtidos nesta pesquisa, bem como suas respectivas análises. Por fim, no capítulo 7, são exibidas as considerações finais e as perspectivas para trabalhos futuros.

2 ACIDENTE VASCULAR CEREBRAL

O AVC, também conhecido como derrame cerebral é uma condição médica grave que afeta o fluxo sanguíneo para o cérebro e pode resultar em danos permanentes ou até mesmo morte. A epidemiologia do AVC é crucial para entender a natureza da doença e desenvolver estratégias eficazes de prevenção e tratamento (MINISTÉRIO DA SAÚDE, 2013; KASPER *et al.*, 2019; DEBAUN *et al.*, 2020).

Além disso, a epidemiologia do AVC também estuda a incidência e a prevalência da doença em diferentes populações e regiões geográficas. Esses dados são usados para identificar padrões de incidência e para desenvolver intervenções de saúde pública direcionadas a grupos de alto risco.

A compreensão da epidemiologia do AVC é crucial para aprimorar a prevenção, o diagnóstico e o tratamento dessa condição. Ao se analisarem os fatores de risco e as tendências da doença, os profissionais de saúde podem conceber intervenções mais efetivas para diminuir o impacto do AVC na saúde da população em geral.

2.1 Promoção da saúde e prevenção da doença

Para reduzir a incidência e o impacto do AVC na saúde pública, é essencial promover a saúde e prevenir a doença. A promoção da saúde envolve a adoção de estilos de vida saudáveis e a redução dos fatores de risco para o AVC. Além disso, a prevenção do AVC também inclui a identificação precoce e o tratamento de condições médicas que aumentam o risco da doença, permitindo o controle desses fatores e, assim, reduzindo o risco de AVC.

Para complementar, é importante que haja uma ampla disseminação de informações sobre os sinais de AVC e a importância de buscar atendimento médico imediato em caso de suspeita da doença. O tratamento precoce do AVC pode reduzir significativamente o risco de complicações graves e melhorar o prognóstico do paciente, sendo fundamental o conhecimento e a atuação rápida nesse sentido.

Existem diversas estratégias que podem ser adotadas para promover a saúde e prevenir o AVC, como campanhas de conscientização, programas de educação em saúde, intervenções em comunidades de alto risco, mudanças na política pública e desenvolvimento de novas tecnologias e tratamentos. Para serem efetivas, essas estratégias devem envolver profissionais de saúde, formuladores de políticas, líderes comunitários e pacientes, com o objetivo de reduzir a

incidência e o impacto do AVC na saúde pública.

Vale lembrar que o Dia Mundial do AVC foi instituído em 29 de outubro de 2006, durante o 1º Congresso Mundial de AVC, realizado em Genebra, na Suíça, pela OMS, com o objetivo de conscientizar sobre a importância da adoção de um estilo de vida saudável na prevenção do AVC e sobre a necessidade de atendimento médico imediato em caso de suspeita da doença, visando reduzir as chances de sequelas e complicações graves.

2.2 Tipos de AVC

Os AVCs são classificados em dois tipos, que ocorrem por motivos diferentes:

- AVC hemorrágico;
- AVC isquêmico.

O AVC isquêmico ocorre quando há obstrução de uma artéria que fornece sangue ao cérebro é bloqueada por um coágulo sanguíneo ou por um depósito de gordura, impedindo o fluxo de sangue e oxigênio para o tecido cerebral, que acabam morrendo. Essa obstrução pode acontecer devido a um trombo (trombose) ou a um êmbolo (embolia). O AVC isquêmico é o mais comum e representa 85% de todos os casos (MINISTÉRIO DA SAÚDE, 2015).

Os sintomas do AVC isquêmico podem incluir fraqueza ou paralisia em um lado do corpo, dificuldade de fala, perda de visão em um ou ambos os olhos, dor de cabeça intensa e repentina, tontura e falta de equilíbrio. Esses sintomas podem aparecer repentinamente e podem se agravar rapidamente (JUNIOR *et al.*, 2019; SOCIEDADE BRASILEIRA DE ACIDENTE VASCULAR CEREBRAL, 2023).

O tratamento do AVC isquêmico depende do tempo decorrido desde o início dos sintomas e da extensão do dano cerebral. Em alguns casos, o tratamento pode incluir medicações para dissolver o coágulo sanguíneo e restabelecer o fluxo de sangue para o cérebro. Em outros casos, pode ser necessário um procedimento cirúrgico para remover o coágulo ou abrir a artéria obstruída (MINISTÉRIO DA SAÚDE, 2015).

O AVC hemorrágico ocorre quando há rompimento de um vaso cerebral, provocando hemorragia. Esta hemorragia pode acontecer dentro do tecido cerebral ou na superfície entre o cérebro e a meninge. É responsável por 15% de todos os casos de AVC, mas pode causar a morte com mais frequência do que o AVC isquêmico (MINISTÉRIO DA SAÚDE, 2015).

Os sintomas do AVC hemorrágico são semelhantes aos do AVC isquêmico, mas podem ser mais intensos e aparecer repentinamente. Eles incluem dor de cabeça intensa e

repentina, náuseas, vômitos, perda de consciência, convulsões, confusão mental, fraqueza ou paralisia em um lado do corpo e dificuldade de fala (JUNIOR *et al.*, 2019; SOCIEDADE BRASILEIRA DE ACIDENTE VASCULAR CEREBRAL, 2023).

O tratamento do AVC hemorrágico depende da gravidade do sangramento e da área do cérebro afetada. Em casos mais leves, o tratamento pode incluir medicações para controlar a pressão arterial e reduzir o inchaço cerebral. Em casos mais graves, pode ser necessária uma cirurgia para drenar o sangue acumulado ou para reparar o vaso sanguíneo danificado (MINISTÉRIO DA SAÚDE, 2015).

2.3 Fatores de risco para AVC

Os fatores de risco são condições ou hábitos que aumentam a probabilidade de um indivíduo desenvolver uma doença ou condição médica, como o AVC. Alguns fatores de risco para o AVC incluem: idade avançada, histórico familiar de AVC, hipertensão arterial, tabagismo, diabetes, colesterol alto, doenças cardiovasculares, obesidade, sedentarismo, consumo excessivo de álcool, uso de anticoncepcionais hormonais, stress, e fibrilação atrial (arritmia cardíaca). Conhecer e gerenciar esses fatores de risco pode ajudar a prevenir ou minimizar o risco de AVC (SOCIEDADE BRASILEIRA DE ACIDENTE VASCULAR CEREBRAL, 2023).

Os fatores de risco para o AVC podem ser divididos em 2 grupos que são os modificáveis e não modificáveis. Os fatores de risco modificáveis para o AVC são aqueles que podem ser controlados ou mudados por meio de mudanças no estilo de vida ou tratamento médico. Alguns exemplos incluem: hipertensão arterial, tabagismo, diabetes, colesterol alto, obesidade, sedentarismo, consumo excessivo de álcool, stress. Os fatores de risco para o AVC que não podem ser modificados incluem: idade avançada, histórico familiar de AVC, raça, gênero, localização geográfica, histórico pessoal de AVC ou Ataque Isquêmico Transitório (AIT) (MINISTÉRIO DA SAÚDE, 2013; MARIANELLI *et al.*, 2020; REN *et al.*, 2021).

Os fatores de risco não modificáveis podem ser controlados ou tratados por meio de mudanças de estilo de vida, como uma dieta saudável, exercício físico regular, manutenção de um peso saudável, controle da pressão arterial, etc. É importante buscar aconselhamento médico para avaliar seus fatores de risco e desenvolver uma estratégia para prevenir o AVC.

A Hipertensão Arterial Sistêmica (HAS) é um dos principais fatores que aumenta o risco do AVC e está associada a piores desfechos. Portanto, a medição regular da Pressão Arterial (PA) é comumente realizada em tanto o tratamento clínico quanto o manejo de indivíduos com

AVC. As comorbidades também podem ter um grau de influência em acontecer o AVC. As comorbidades são consideradas o infarto do miocárdio, doença arterial coronariana, câncer, doenças pulmonares, osteoporose, artrite e sintomas de transtornos de humor, incluindo depressão e os sintomas de ansiedade (FOROUZANFAR *et al.*, 2017; REN *et al.*, 2021; STAMM *et al.*, 2019).

A ocorrência do AVC é resultante de uma complexa interação entre fatores de diversos tipos. A literatura tem apresentado estratégias analíticas que visam uma compreensão melhor estruturada dos fatores implicados na determinação desse evento. Todos os fatores de risco têm um fator de influência para que uma pessoa possa adquirir um AVC (KUMMER *et al.*, 2020).

2.4 Detecção e diagnóstico

É essencial que o diagnóstico e tratamento do AVC sejam realizados o mais rapidamente possível, pois isso aumenta as chances de uma recuperação completa. Atrasos no tratamento podem levar a consequências mais graves, tornando crucial estar atento aos sinais e sintomas do AVC e buscar atendimento médico imediato para minimizar o dano cerebral e aumentar as chances de recuperação (MINISTÉRIO DA SAÚDE, 2013).

O diagnóstico e detecção do AVC geralmente começam com uma avaliação clínica dos sintomas e sinais apresentados pelo paciente. Alguns sintomas comuns do AVC incluem: diminuição da sensibilidade e/ou fraqueza que começa de forma repentina na face, braço e/ou perna, especialmente se em apenas um lado do corpo, confusão mental, dificuldade para falar ou compreender a fala, alterações visuais súbitas em um ou ambos os olhos, dificuldade para andar, perda de equilíbrio e/ou coordenação repentina. Se houver dor de cabeça intensa sem causa conhecida, é essencial buscar atendimento emergencial, pois isso pode ser sugestivo de um AVC (MINISTÉRIO DA SAÚDE, 2013).

Existindo suspeita de AVC, o médico pode realizar exames complementares, como tomografia computadorizada ou ressonância magnética do cérebro, exame de sangue e avaliação de arritmias cardíacas, entre outros. O tratamento imediato é fundamental para minimizar o dano cerebral e maximizar a recuperação do paciente (MINISTÉRIO DA SAÚDE, 2013).

O AVC é uma emergência médica que pode causar danos permanentes ou até mesmo levar à morte. Por isso, é importante reconhecer os sintomas e procurar ajuda médica imediatamente. O tratamento precoce pode ajudar a reduzir os danos ao cérebro e melhorar as

chances de recuperação a detecção precoce do AVC é fundamental para o sucesso do tratamento, e é por isso que a sensibilização da população quanto aos sintomas e fatores de risco do AVC é tão importante para promover conscientização, rápida detecção e atendimento do AVC (SOCIEDADE BRASILEIRA DE ACIDENTE VASCULAR CEREBRAL, 2023; MINISTÉRIO DA SAÚDE, 2013).

2.5 Tratamento e recomendações

As sequelas provenientes de um AVC podem tornar o indivíduo parcial ou totalmente incapaz, impactando negativamente na vida produtiva e gerando altos custos financeiros relacionados ao tratamento de reabilitação. Um indivíduo com consequências decorrentes do AVC pode apresentar diversas limitações, e a recuperação é diferente em cada caso. Condições relacionadas ao AVC existem o déficit sensório-motor, limitação nas atividades de vida, comunicação, déficit cognitivo e distúrbio de humor. Dessa maneira, o paciente sofre como consequência uma importante queda na qualidade de vida na maioria dos casos (SMELTZER *et al.*, 2011).

O tratamento do AVC depende do tipo de AVC e da extensão do dano cerebral. No caso do AVC isquêmico, o tratamento geralmente inclui medicações para dissolver o coágulo sanguíneo e restabelecer o fluxo de sangue para o cérebro. Essas medicações devem ser administradas o mais rápido possível, de preferência nas primeiras horas após o início dos sintomas. Em alguns casos, pode ser necessário um procedimento cirúrgico para remover o coágulo ou abrir a artéria obstruída (MINISTÉRIO DA SAÚDE, 2013).

Já no caso do AVC hemorrágico, o tratamento pode incluir medicações para controlar a pressão arterial e reduzir o inchaço cerebral. Em casos mais graves, pode ser necessária uma cirurgia para drenar o sangue acumulado ou para reparar o vaso sanguíneo danificado (MINISTÉRIO DA SAÚDE, 2015).

Além do tratamento médico, é importante que o paciente receba cuidados de suporte, como a monitorização da pressão arterial, da respiração e da função cardíaca. A reabilitação é essencial para ajudar o paciente a recuperar a força muscular, a coordenação e a fala, bem como para prevenir complicações como a trombose e a pneumonia (MINISTÉRIO DA SAÚDE, 2023b).

O tratamento do AVC é geralmente realizado em Centros de Atendimento de Urgência, que são estabelecimentos hospitalares especializados em cuidados de emergência para pacientes com AVC. Esses centros seguem protocolos clínicos e diretrizes terapêuticas específicas

para fornecer um atendimento rápido e eficaz, incluindo o uso de medicamentos trombolíticos para dissolver coágulos sanguíneos (MINISTÉRIO DA SAÚDE, 2023b).

A reabilitação pós-AVC pode ser realizada em Centros Especializados em Reabilitação, onde uma equipe multidisciplinar de profissionais de saúde, incluindo médicos, fisioterapeutas, fonoaudiólogos e terapeutas ocupacionais, trabalha em conjunto para ajudar o paciente a recuperar a força muscular, a coordenação e a fala (MINISTÉRIO DA SAÚDE, 2023b).

É importante ressaltar que o tratamento, atendimento e reabilitação do AVC devem ser prescritos por médicos especialistas, levando em consideração as necessidades individuais de cada paciente. O uso de medicamentos e outras terapias pode ser necessário em casos específicos, e deve ser monitorado e ajustado regularmente para garantir a eficácia do tratamento.

Existem diversos fatores de risco que podem contribuir para o desenvolvimento de doenças crônicas, incluindo o AVC, alguns dos quais são inalteráveis, como idade, raça, constituição genética e sexo. No entanto, outros fatores podem ser controlados e são fundamentais para prevenir essas doenças (MINISTÉRIO DA SAÚDE, 2023a).

Adotar um estilo de vida saudável pode ajudar a prevenir o AVC e outras doenças crônicas, tais como não fumar, não consumir álcool ou drogas ilícitas, manter uma alimentação saudável, controle da pressão arterial, do colesterol e do diabetes, manter um peso saudável, beber bastante água, praticar atividades físicas regularmente, controlar a pressão arterial e a glicose (MINISTÉRIO DA SAÚDE, 2023a).

O Ministério da Saúde tem investido em diversas ações para promover hábitos saudáveis na população, incluindo o Programa Academia da Saúde, que incentiva a prática de atividade física através da criação de polos, e o Guia Alimentar para a População Brasileira, que fornece orientações sobre uma alimentação saudável e balanceada para evitar o desenvolvimento de doenças crônicas, como o AVC (MINISTÉRIO DA SAÚDE, 2023a).

3 APRENDIZAGEM DE MÁQUINA

A aprendizagem de máquina é uma subárea da inteligência artificial cujo objetivo é permitir que computadores aprendam e melhorem seu desempenho na realização de determinada tarefa a partir da experiência obtida nas execuções anteriores. Trata-se da construção de algoritmos que aprendem a partir dos dados, em que a estatística tem um papel fundamental, oferecendo as ferramentas necessárias para o processo de análise e a construção de modelos. Podemos dizer que a aprendizagem de máquina é o estudo científico de algoritmos e modelos estatísticos que os sistemas de computador usam para realizar uma tarefa específica (MITCHELL, 1997; LENZ *et al.*, 2020).

A aprendizagem de máquina depende de diferentes algoritmos para resolver problemas de dados. São muitos os processos e as tarefas que podem ser utilizados durante o aprendizado, geralmente, escolhidos de acordo com o tipo de problema, ou dados disponíveis, ou o número de variáveis, ou o tipo de modelo que melhor se adequaria a ele e assim por diante (LENZ *et al.*, 2020).

As aplicações de aprendizagem de máquina são amplas podendo ser utilizadas em diversas áreas como: educação, mecanismos de pesquisa, marketing digital, saúde, vendas, compras, etc. A aprendizagem de máquina possui amplos benefícios como a previsão do comportamento de um cliente, assumir tarefas rotineiras de funcionários de uma empresa, inteligência de negócios, treinamentos e capacitação de pessoas, diagnósticos de doenças e análise financeira de mercado.

Um fator muito importante na aplicação de aprendizagem de máquina são os dados, ou seja, as informações disponíveis para o sistema, tanto as recebidas quanto as geradas pela máquina, referenciados como dados de entrada e saída respectivamente. O grupo que engloba os dados de determinado sistema é denominado conjunto de dados. O conjunto de dados pode fazer menção a diferentes características de um sistema. Nem todos os dados de entrada referem-se à mesma característica, e isso também é válido para os dados de saída. Dessa forma, os dados também são organizados conforme a característica a que se referem. Essas características são conhecidas como atributos (LENZ *et al.*, 2020).

Existem três principais tipos de aprendizado de máquina: o aprendizado supervisionado, o aprendizado não supervisionado e o aprendizado por reforço. A seguir, apresentaremos uma breve descrição de cada um deles:

- **Aprendizado supervisionado:** Nesse tipo de aprendizado, os modelos são treinados em um

conjunto de dados rotulados, ou seja, dados que possuem uma variável alvo definida. O modelo é então capaz de prever a variável alvo em novos exemplos (ALPAYDIN, 2010);

- **Aprendizado não supervisionado:** Nesse tipo de aprendizado, os modelos são treinados em um conjunto de dados não rotulados, ou seja, dados que não possuem uma variável alvo definida. O modelo busca por padrões ou estruturas dentro dos dados e os usa para agrupar os exemplos em diferentes categorias ou clusters, com base em suas características compartilhadas (HASTIE *et al.*, 2009);
- **Aprendizado por reforço:** Nesse tipo de aprendizado, os modelos são treinados por meio da interação com um ambiente. O modelo aprende a tomar decisões em um ambiente dinâmico, recebendo feedback na forma de recompensas ou penalidades. O objetivo do modelo é maximizar a recompensa cumulativa ao longo do tempo (SUTTON; BARTO, 2018).

3.1 Engenharia de atributos

Modelos preditivos atuam identificando padrões em exemplos a partir de atributos disponíveis. Cada exemplo é composto de um ou mais atributos que, sozinhos ou em conjuntos, são preditivos para o atributo alvo, ou rótulo desejado. Em dados tabulares, atributos são tipicamente colunas na tabela, mas em outros domínios podem assumir outros formatos (como matrizes para imagens). A expressividade do modelo dita a complexidade dos padrões que podem ser extraídos, mas conjuntamente os atributos disponíveis podem habilitar ou facilitar a identificação dos padrões (ARAUJO, 2020).

Segundo Grus (2016) a engenharia de atributos é um processo de transformação de dados brutos em atributos úteis para o processo de aprendizado de máquina. É um processo fundamental para construção de uma aplicação em aprendizagem de máquina. Tipicamente é um processo amplamente manual que se beneficia de conhecimento e domínio do problema. Na engenharia de atributos existem basicamente três abordagens principais: criar atributos, transformar atributos e selecionar atributos (ARAUJO, 2020).

Essas abordagens são complementares e podem ser usadas em conjunto para criar modelos mais precisos e robustos. A escolha da abordagem depende do problema em questão e das características dos dados. A engenharia de atributos é uma das etapas mais importantes no desenvolvimento de modelos de aprendizado de máquina e é fundamental para obter bons resultados.

3.1.1 Criação de Atributos

A criação de atributos é uma etapa crucial no processo de desenvolvimento de modelos de aprendizado de máquina, pois os modelos são tão bons quanto as variáveis que os alimentam. Uma variável bem projetada pode aumentar significativamente a precisão do modelo e, portanto, é importante ter um bom entendimento do domínio do problema e da natureza dos dados para criar atributos relevantes (GRUS, 2016).

A criação de atributos envolve a geração de novas variáveis a partir de variáveis existentes. Discuta as diferentes técnicas de criação de atributos, como a combinação de variáveis, a geração de recursos de tempo, a criação de recursos de domínio e a geração de recursos baseados em texto, estabelecendo as informações que o modelo terá acesso não só durante o treinamento, mas também durante a inferência. Portanto, é importante que a criação de atributos seja reproduzível com o modelo em produção (GRUS, 2016).

Algumas técnicas comuns de criação de atributos incluem:

- Criação de novas variáveis: criação de novas variáveis a partir de combinações de variáveis existentes, como a soma, subtração, multiplicação, divisão ou concatenação de variáveis;
- Transformação de variáveis existentes: aplicação de funções matemáticas, como logaritmos, exponenciais, raiz quadrada, para transformar variáveis existentes e ajustar sua distribuição;
- *Binning* ou discretização: agrupamento de valores de variáveis contínuas em categorias ou intervalos discretos para torná-los mais fáceis de interpretar ou modelar;
- *One-Hot Encoding*: transformação de variáveis categóricas em variáveis numéricas binárias para que possam ser usadas em modelos de aprendizado de máquina;
- Extração de características de texto: a partir de dados de texto, a extração de recursos pode incluir técnicas como a criação de histogramas de palavras, a contagem de palavras-chave e a identificação de palavras-chave importantes.

3.1.2 Transformação de Atributos

A transformação de atributos é o processo de modificar ou mapear os valores dos atributos em um conjunto de dados. Isso pode ser feito com o objetivo de melhorar a qualidade dos dados, torná-los mais apropriados para análise ou modelagem, ou para adaptá-los a uma determinada técnica de análise, além de evitar que o algoritmo fique enviesado para as variáveis

com maior ordem de grandeza. Algumas técnicas comuns de transformação de atributos incluem normalização, codificação de variáveis categóricas, transformações matemáticas (como logaritmo ou raiz quadrada), escalonamento e agrupamento (CASTRO; FERRARI, 2016; GRUS, 2016).

Atributos derivados são novos atributos definidos a partir de atributos já criados. Este tem a intenção de introduzir conhecimento sobre a interação entre atributos e facilitar o processo de aprendizado pelo modelo. Em um problema de classificação, o modelo pode não ser capaz separar as classes a partir dos atributos originais, mas, com a adição de atributos derivados, isso pode passar a ser possível (ARAUJO, 2020).

Outras transformações visam corrigir discrepâncias ou problemas com os atributos. Por exemplo, alguns exemplos podem ter dados faltantes para algum atributo e pode ser desejável preencher estas lacunas, isto é, imputar valores que não apenas indicar a ausência (ARAUJO, 2020).

Alguns algoritmos de treinamento são sensíveis a diferenças de escala entre atributos. Por exemplo, se um atributo é idade de uma pessoa em anos, os valores estarão em uma escala de algumas dezenas, mas, se outro atributo for peso em gramas, os valores estarão em uma escala de muitos milhares. Isto pode levar o algoritmo de treinamento a dar mais atenção a peso do que a idade somente devido a escala. Para solucionar isso, podemos reescalar os atributos (GRUS, 2016).

Existem várias técnicas de normalização, porém as mais comuns são: *Simple Feature Scaling*, normalização Min-Max, normalização *Z-score*, normalização *Decimal Scaling* e normalização de Magnitude. A escolha da técnica adequada depende das características do conjunto de dados e do objetivo da análise. Em geral, é importante realizar a normalização de forma consciente e cuidadosa, para garantir que os dados estejam na escala apropriada para a análise e que os resultados sejam precisos e confiáveis (JO, 2019).

Isso é importante porque muitos métodos de aprendizado de máquina são sensíveis à escala dos dados, e valores muito diferentes entre atributos podem afetar negativamente a performance do modelo. Além disso, a normalização também pode ajudar a lidar com problemas de *features* dominantes ou problemas de colinearidade, em que duas ou mais *features* são altamente correlacionadas entre si (JO, 2019).

A técnica de normalização *Simple Feature Scaling* é uma das técnicas mais simples de normalização e consiste em dividir cada valor de uma variável pela sua amplitude (diferença entre o valor máximo e mínimo). A fórmula utilizada para normalizar um valor x é:

$$x_{norm} = \frac{x}{A} \quad (3.1)$$

Onde x é o valor original da variável e x_{norm} é o valor normalizado. A amplitude é calculada como a diferença entre o valor máximo e mínimo da variável:

$$A = \max_x - \min_x \quad (3.2)$$

Essa técnica de normalização é útil para dados que estão em uma escala muito grande e apresentam valores discrepantes (*outliers*), que podem afetar negativamente o desempenho do modelo. Ao normalizar os dados usando *Simple Feature Scaling*, os valores discrepantes são reduzidos, e os dados são trazidos para uma escala comparável.

No entanto, a normalização *Simple Feature Scaling* não garante que os dados estejam na mesma escala, uma vez que a amplitude é afetada pelos valores discrepantes. Além disso, se a distribuição dos dados não for uniforme, a normalização pode não ser tão eficaz.

No método Min-Max, os dados são dimensionados para um intervalo fixo, geralmente de 0 a 1, o custo de ter este intervalo limitado é que ele vai acabar com menor desvio padrão. Isso pode ser feito subtraindo cada valor do mínimo e dividindo o resultado pela diferença entre o máximo e o mínimo (JO, 2019). Segue a equação de escala Min-Max apresentada pela seguinte equação:

$$x_{norm} = \frac{x_{old} - x_{max}}{x_{max} - x_{min}} \quad (3.3)$$

Normalização *Z-score* é a técnica mais popular e é utilizada para padronizar os dados para ter média ($\mu = 0$) e desvio padrão ($\sigma = 1$) (JO, 2019). A fórmula utilizada é a seguinte:

$$x_{norm} = \frac{x - \mu(x)}{\sigma(x)} \quad (3.4)$$

Essa técnica de normalização é útil para dados que não possuem uma escala específica e distribuem-se de forma aproximadamente normal. A normalização *Z-score* ajuda a remover a influência de outliers e reduzir a variação nos dados. Quando a normalização *Z-score* é aplicada, os valores abaixo da média da distribuição assumem valores negativos e os valores acima da média assumem valores positivos.

No entanto, a normalização *Z-score* pode não ser adequada para conjuntos de dados com distribuições altamente assimétricas ou presença de outliers extremos, pois esses valores podem distorcer a média e o desvio padrão, influenciando negativamente o desempenho do modelo de aprendizado de máquina (ARAUJO, 2020).

O método de Normalização *Decimal Scaling* é a técnica que divide os dados pelo maior valor absoluto da variável e arredonda o resultado para um número decimal. A fórmula utilizada é a seguinte:

$$x_{norm} = \frac{x}{10^k} \quad (3.5)$$

Onde x é o valor original da variável e k é o número inteiro que representa o número de dígitos que devem ser usados para a normalização.

Normalização de Magnitude é a técnica que divide os dados pelo comprimento do vetor da variável. É usada principalmente em problemas de aprendizado de máquina envolvendo vetores. A fórmula utilizada é a seguinte:

$$x_{norm} = \frac{x}{||x||} \quad (3.6)$$

Onde x é o vetor original da variável e $||x||$ é o comprimento do vetor.

3.1.3 Seleção de Atributos

Em um cenário ideal, os atributos seriam apenas os suficientes e necessários para habilitar o aprendizado. Menos atributos impediriam o aprendizado com o desempenho desejado, enquanto mais atributos tornariam o aprendizado mais difícil devido ao maior espaço de hipóteses. Naturalmente, não sabemos de antemão que atributos são necessários e muito menos se são suficientes. A seleção de atributos visa reduzir o número de atributos disponíveis para o modelo (ARAUJO, 2020).

A seleção de atributos é necessária para evitar um problema comum em aprendizagem de máquina conhecido como maldição da dimensionalidade, que ocorre quando temos tantos atributos que acabamos com dados excessivamente esparsos. Dados esparsos são aqueles onde os pontos estão muito distantes uns dos outros no espaço de atributos, não permitindo estabelecer relações de similaridade e, portanto, discernir padrões nos exemplos. Segundo Varleysen e

François (2005) é um fenômeno em que a adição de mais dimensões aos dados e que pode levar a problemas significativos na análise e modelagem dos dados, na maioria das vezes, têm consequências no comportamento e desempenho dos algoritmos de aprendizado.

Existem diversas técnicas utilizadas na seleção de atributos em conjuntos de dados, dentre as quais se destacam:

- Seleção Univariada: avalia a relação entre cada atributo e a variável alvo, selecionando aqueles com maior correlação. Exemplos de técnicas: Teste Qui-Quadrado, ANOVA, F-test, *Mutual Information Test* e Correlação de Pearson;
- *Wrapper*: utiliza um modelo de aprendizado de máquina para avaliar o conjunto de atributos, selecionando aqueles que obtêm melhor desempenho na classificação. Exemplos de técnicas: *Recursive Feature Elimination*, *Forward Selection*, *Backward Elimination*;
- Incorporação de Modelo: seleciona atributos baseados na importância que têm para um modelo de aprendizado de máquina. Exemplos de técnicas: *Random Forest Feature Importance* e Coeficientes de Regressão Linear;
- Redução de Dimensionalidade: utiliza técnicas de projeção para reduzir a dimensionalidade do conjunto de dados, selecionando os atributos mais relevantes. Exemplos de técnicas: *Principal Component Analysis*, *Linear Discriminant Analysis*.

Uma metodologia complementar à seleção de atributos é a extração de atributos. Ao invés de simplesmente remover atributos, a extração de atributos busca encontrar alternativas para representação dos exemplos através da construção de novos atributos a partir dos atributos originais. Essa abordagem utiliza técnicas de redução de dimensionalidade para sintetizar a informação contida nos atributos originais (ARAUJO, 2020).

Reduzir o número de atributos contribui para ter modelos mais explicáveis, reduzir tempo de treinamento e inferência. Uma variação desta técnica é não remover o atributo, mas gerar uma permutação aleatória dos valores entre os exemplos. Outra alternativa é utilizar métricas como ganho de informação e correlações com o rótulo, mantendo apenas atributos com altos valores (LENZ *et al.*, 2020).

Outro passo importante é a atividade de correlação para cada *feature* no *dataset*, ou seja, para cada *feature X* do *dataset*, irá mostrar como ela se relaciona com a *feature Y*. A partir dessa visualização, você pode saber quais *features* possuem uma correlação positiva, quando se movem na mesma direção. E qual a correlação negativa quando se movem em direção opostas. Com isso consigo ter qual as *features* que exercem maior influência no meu modelo (GRUS,

2016).

3.2 Cargas desbalanceadas

As cargas desbalanceadas em aprendizagem de máquina ocorrem quando a distribuição dos dados não é uniforme entre as classes que estão sendo analisadas. Isso pode acontecer em diversas situações (AZANK, 2020).

As cargas desbalanceadas podem levar a problemas no treinamento do modelo, já que o algoritmo de aprendizagem pode acabar favorecendo a classe majoritária, ignorando a classe minoritária. Isso pode resultar em um modelo que é bom em prever a classe majoritária, mas ruim em prever a classe minoritária (CHAWLA, 2010).

Se desenvolvermos um modelo sem considerar essa desproporcionalidade nos dados, o modelo será induzido a ter uma acurácia alterada, o qual os parâmetros do algoritmo não diferenciarão a classe minoritária das demais categorias, acreditando que estão agregando resultado devido à aparente alta acurácia (AZANK, 2020; SHELKE *et al.*, 2017).

Existem várias técnicas que podem ser utilizadas para lidar com cargas desbalanceadas em aprendizagem de máquina. Entretanto cada uma das técnicas têm suas vantagens e desvantagens (CHAWLA, 2010). Algumas das técnicas mais comuns incluem:

- *Oversampling*: essa técnica consiste em duplicar ou triplicar os exemplos da classe minoritária, tornando a distribuição dos dados mais equilibrada entre as classes. O *oversampling* pode ser feito de várias maneiras, como por exemplo, replicando os exemplos existentes, gerando novos exemplos sintéticos ou ajustando os pesos dos exemplos durante o treinamento.
- *Undersampling*: essa técnica consiste em reduzir o número de exemplos da classe majoritária, tornando a distribuição dos dados mais equilibrada entre as classes. O *undersampling* pode ser feito de várias maneiras, como por exemplo, selecionando aleatoriamente exemplos da classe majoritária ou utilizando técnicas de clusterização para agrupar exemplos similares.
- Utilização de pesos: essa técnica consiste em atribuir pesos diferentes aos exemplos de cada classe durante o treinamento do modelo. Esses pesos podem ser ajustados de forma a dar mais importância aos exemplos da classe minoritária, tornando o modelo mais sensível a essa classe.
- Utilização de técnicas específicas de algoritmos: alguns algoritmos de aprendizagem de

máquina possuem técnicas específicas para lidar com cargas desbalanceadas, como por exemplo, o algoritmo *Support Vector Machine* (SVM) que permite ajustar o parâmetro de regularização para dar mais importância à classe minoritária.

3.3 Aprendizagem de máquina supervisionada para classificação

A aprendizagem de máquina supervisionada é uma abordagem para treinar algoritmos de aprendizado de máquina usando um conjunto de dados rotulados. Esses dados consistem em exemplos de entrada e saída esperada, que são usados para ensinar o modelo a fazer previsões precisas sobre novos dados. Cada coluna representa um atributo ou característica. Pelo menos uma dessas colunas deve referir-se a um atributo-alvo, que é um rótulo ou classe que identifica a instância. Cada linha representa uma instância com os valores correspondentes para cada um dos seus atributos (LENZ *et al.*, 2020).

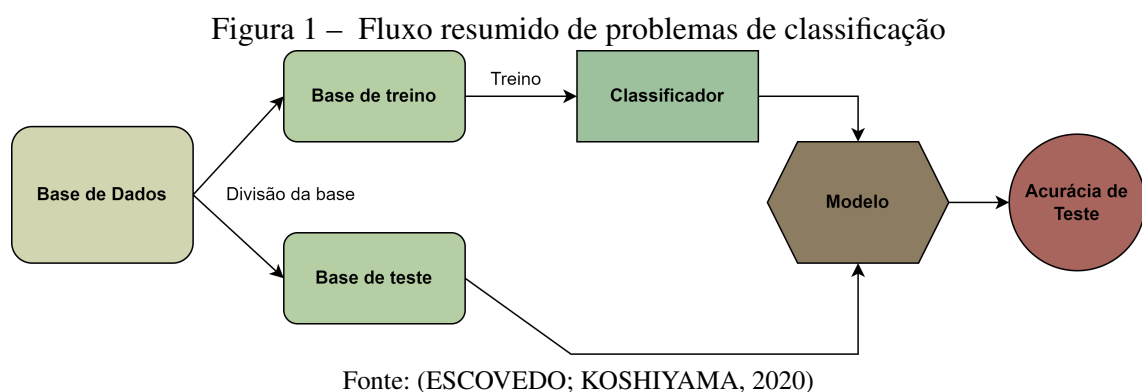
O aprendizado supervisionado possui algumas características próprias que são apresentadas como etapas. Uma etapa é o conjunto de dados que possuem os valores para o atributo preditor e alvo, outra etapa de aprimoramento onde acontece o treinamento do modelo com base nas instancias rotuladas e a última etapa que é a tarefa onde deve-se prever a resposta ou o rótulo completo (LENZ *et al.*, 2020).

Pode se afirmar que existem duas principais categorias de aprendizado supervisionado: classificação e regressão. A aprendizagem de máquina supervisionada por regressão é usada quando o objetivo é prever um valor numérico, contínuo ou discreto, para um exemplo de entrada. Por exemplo, um modelo de aprendizado de máquina pode ser treinado para prever o preço de uma casa com base em suas características, como número de quartos, tamanho do lote, localização etc. Nesse caso, a saída esperada seria um valor numérico, como 100.000, 200.000, 300.000, etc (HASTIE *et al.*, 2009).

Aprendizagem de máquina supervisionada por classificação é usada quando o objetivo é classificar os exemplos de entrada em diferentes categorias. O objetivo é encontrar um modelo que possa atribuir corretamente a cada exemplo de entrada a uma das categorias pré-definidas. Por exemplo, um modelo de aprendizado de máquina pode ser treinado para identificar se uma imagem contém um gato ou um cachorro. Nesse caso, a saída esperada seria uma das duas categorias: "gato" ou "cachorro" (KOTSIANTIS, 2007).

Para abordar o problema de classificação relacionado à prevenção da recorrência do AVC, apresentaremos um fluxo resumido de problemas de classificação na figura 1. O fluxo

começa com uma base de dados rotulada, que é dividida em dois subconjuntos disjuntos: a base de treino e a base de teste. Em seguida, o modelo é treinado: a base de treino é apresentada ao classificador para calibrar seus parâmetros de acordo com os dados apresentados. Depois disso, a etapa de predição de classes começa: os exemplos da base de teste são apresentados ao modelo treinado para que ele possa prever suas classes. A qualidade do modelo é medida comparando as classes previstas com as classes verdadeiras da base de teste, permitindo avaliar sua habilidade em classificar exemplos não vistos durante o treinamento (ESCOVEDO; KOSHIYAMA, 2020).



Nesta seção, vamos discutir as técnicas de aprendizado de máquina que serão aplicadas no desenvolvimento de uma nova abordagem para prevenção da recorrência de AVC. Cada algoritmo de aprendizado de máquina possui sistemas e funções matemáticas distintos, e a precisão pode variar dependendo dos tipos de dados utilizados em cada técnica. Nas subseções seguintes, apresentaremos técnicas supervisionadas de aprendizado de máquina para classificação, divididas em três tipos: técnicas Clássicas, métodos Ensemble e Redes Neurais Artificiais (RNA). A figura 2 apresenta um mapa resumido das técnicas de aprendizado de máquina classificadas por tipo que serão utilizadas neste trabalho.

Figura 2 – Mapa de aprendizagem maquina



Fonte: Elaborado pelo autor

3.3.1 Técnicas Clássicas

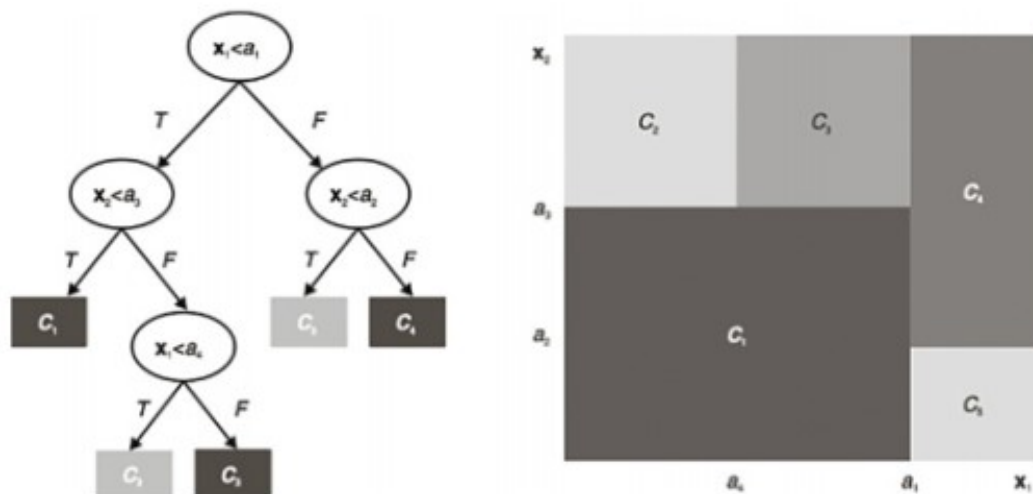
Neste tópico, serão apresentadas as principais técnicas clássicas de aprendizagem de máquina supervisionada para classificação: *Decision Tree*, *Logistic Regression*, *Naive Bayes*, *K-Nearest Neighbor* (KNN) e SVM. Esses algoritmos são amplamente utilizados em diversas aplicações e considerados clássicos por sua eficácia comprovada ao longo do tempo. De acordo com o artigo Jordan e Mitchell (2015) e o livro Alpaydin (2016), *Decision Tree*, *Logistic Regression*, *Naive Bayes*, SVM, KNN são considerados métodos clássicos de aprendizado de máquina e têm sido aplicados em áreas como finanças, medicina e reconhecimento de padrões. Cada técnica possui uma abordagem e objetivo específicos, mas todas compartilham o objetivo comum de prever ou classificar dados com base em modelos estatísticos ou algoritmos de aprendizado.

Decision Tree

A *Decision Tree* ou Árvore de Decisão é um dos métodos de modelagem preditiva mais conhecido em aprendizagem de máquina possuindo como sua principal característica a versatilidade e estrutura visual. O algoritmo possui uma semelhança com fluxograma, com etapas bem visíveis que facilita o entendimento. O método árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que pode ser utilizado para problemas de classificação ou regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação (BATISTA; FILHO, 2019; FACELI *et al.*, 2011).

Os algoritmos de decisão produzem grafos acíclicos, no sentido de um nó folha ou de outras divisões do espaço de hipóteses em que os nós de divisão expressam um teste condicional dos valores de atributos precursores. Como critério de aprendizado, são usadas funções de custo 0-1 para obter a moda ou minimizar a média do erro médio quadrático, para tarefas de classificação e regressão, respectivamente (FACELI *et al.*, 2011).

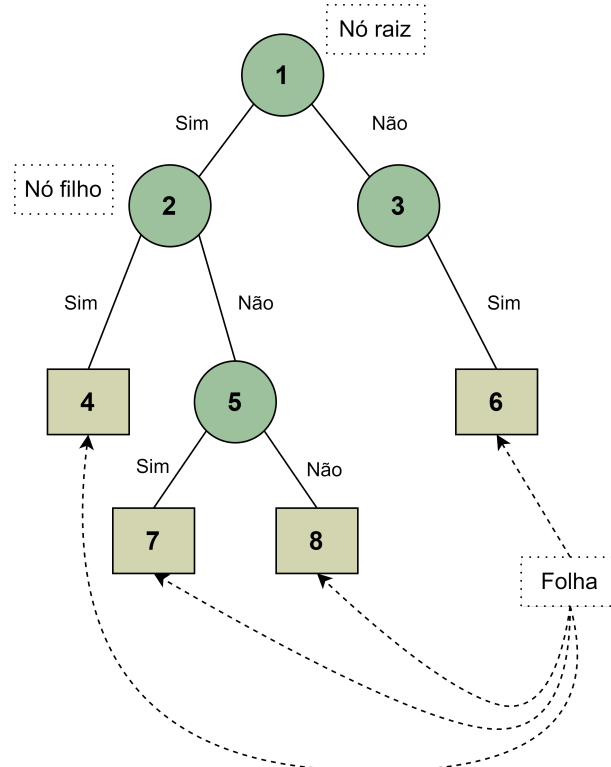
Figura 3 – Árvore de decisão (à esquerda) e o espaço de hipóteses (à direita).



Fonte: (FACELI *et al.*, 2011)

A Árvore de Decisão estabelece nós que se relacionam entre si seguindo uma hierarquia. Existe o nó-raiz, que é o mais importante, e os nós-folha, que são os resultados finais. No contexto de aprendizagem de máquina, o nó-raiz é um dos atributos da base de dados e o nó-folha é a classe. Nas ligações entre os nós existem regras de "sim-não" sempre seguindo com essa análise partindo para um nó ou para outro nó chegando no nó seguinte, e segue a mesma análise novamente. Na figura 4 apresenta um esquema de exemplo para um melhor entendimento do funcionamento.

Figura 4 – Árvore de Decisão.



Fonte: Elabora pelo autor.

O método é inspirado na forma como humanos tomam decisão e, por este motivo, se tornam um dos modelos mais simples de se entender. Uma das principais vantagens deste algoritmo é a apresentação visual da informação, facilitando o entendimento pelo ser humano. De forma resumida, uma árvore de decisão usa amostras das características dos dados para criar regras de decisão no formato de árvore, mapeando os dados em um conjunto de regras que podem ser usadas para uma decisão (ESCOVEDO, 2020).

Uma aplicação muito constante da árvore de decisão é quando existe um problema de diversas informações, ou seja, quando as categorias de classificação são múltiplas, e não apenas duas. O sistema apreende os dados, entende suas relações, realiza os cálculos a fim de entender quais são os nós mais importantes e vai ajustando as condições (ESCOVEDO, 2020).

De forma prática o método da árvore de decisão é construído a partir de um conjunto de dados de treinamento, onde cada nó da árvore representa uma pergunta sobre uma das entradas. O algoritmo decide qual pergunta fazer em cada nó, selecionando a pergunta que separa melhor os dados em relação à variável alvo. A árvore é construída recursivamente até que todos os dados estejam corretamente classificados ou não haja mais perguntas relevantes para

fazer. Depois de construir a árvore, ela pode ser usada para prever o valor da variável alvo para novos dados de entrada, percorrendo a árvore até chegar a uma folha, onde é fornecida a previsão.

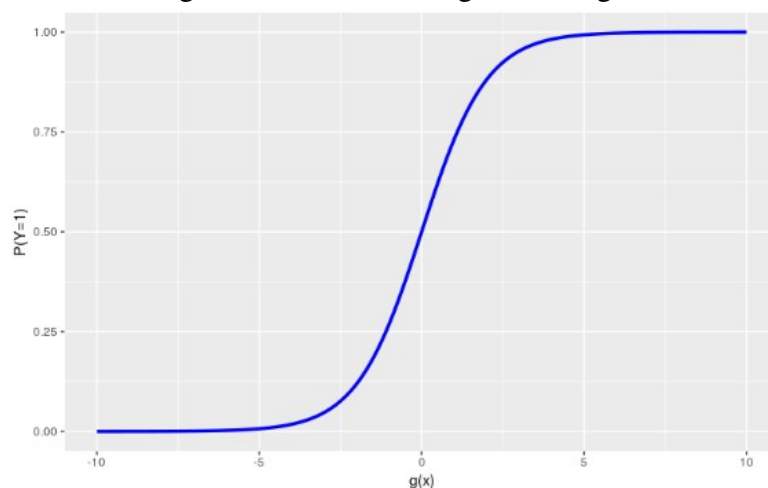
Logistic Regression

A *Logistic Regression* ou Regressão Logística é uma técnica estatística que visa produzir, a partir de um conjunto de observações. Apesar do nome, é um algoritmo utilizado exclusivamente para problemas de Classificação. A Regressão Logística é usada para estimar valores discretos de classes binárias (valores como 0/1, sim/não, verdadeiro/falso) com base em um conjunto de variáveis independentes. O método é empregado em diferentes áreas, sendo muito utilizada para problemas de duas classes. É um dos algoritmos de aprendizagem de máquina mais conhecidos e utilizados no mundo (RAY, 2019).

O objetivo de uma análise usando a Regressão Logística é encontrar o mais adequado e mais parcimonioso, ainda biologicamente razoável, modelo para descrever a relação entre um resultado (variável dependente ou resposta) e um conjunto de variáveis independentes (preditor ou explicativa). O que distingue a regressão logística do modelo de regressão linear é que a variável de resultado na regressão logística é categórica e geralmente binária ou dicotômico (LEMESHOW; JR, 2007).

O modelo de regressão logística possui sua representação gráfica em formato de "S", sendo esta conhecida como curva logística. Este modelo estabelece uma relação entre a probabilidade de ocorrência da variável categórica de interesse e as variáveis preditoras. A figura 5 apresenta um exemplo da curva logística (BATISTA; FILHO, 2019).

Figura 5 – Curva de Regressão Logística.



Fonte: (BATISTA; FILHO, 2019)

Logo abaixo uma das expressões mais simples da função que descreve a curva logística, onde e é conhecido como número de Euler ($e \cong 2,72$):

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.7)$$

A Regressão Logística tem as seguintes vantagens: simplicidade de implementação, eficiência computacional, eficiência de perspectiva de formação, facilidade de regularização. Nenhuma escala é necessária para recursos de entrada. Também logística regressão não é afetada por pequenos ruídos nos dados e multicolinearidade. As desvantagens são: incapacidade de resolver problemas não lineares como sua superfície de decisão é linear, propensa a sobreajuste, não funcionará bem, a menos que todas as variáveis independentes sejam identificadas (RAY, 2019).

Em aprendizagem de máquina, a regressão logística é utilizada para tarefas de classificação, ou seja, em que a variável de interesse é categórica. Nesse caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse. Uma nomenclatura bastante utilizada para os valores 0 e 1 é classe negativa e classe positiva, respectivamente (BATISTA; FILHO, 2019).

De forma simplificada, o modelo logístico pode ser definido da seguinte maneira:

$$\hat{y} = f\left(\frac{1}{Z}\right) = \frac{1}{1 + e^{-Z}} \quad (3.8)$$

onde $P(y = 1 \mid \frac{1}{Z})$ é a probabilidade de ocorrência da classe positiva dado Z . Sendo Z :

$$Z = \ln\left(\frac{p}{p-1}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} \quad (3.9)$$

onde p indica a probabilidade de ocorrência do evento de interesse, X é o conjunto das variáveis preditoras e α e β são parâmetros do modelo, os quais estão estimados pela máxima verossimilhança, encontrando uma combinação de coeficientes que maximiza a probabilidade de ocorrência do evento de interesse (FAVERO *et al.*, 2009).

Reescrevendo o modelo, obtém-se:

$$\hat{y} = f(Z) = \frac{1}{1 + e^{-Z}} = \text{prob}(y = \frac{1}{X_1, X_2, \dots, X_k}) = \frac{1}{1 + e^{-(\alpha - \sum B_i X_i)}} \quad (3.10)$$

Dessa forma, um modelo de regressão logística utilizado para a classificação binária de uma variável de interesse irá prever a probabilidade de pertencer à classe positiva. Tem-se, portanto, que $\hat{y}$ é dado por:

$$\hat{y} = \begin{cases} 0, & \text{se } Prob(y = \frac{1}{X}) < 0,5 \\ 1, & \text{se } Prob(y = \frac{1}{X}) \geq 0,5 \end{cases} \quad (3.11)$$

onde X é o vetor das variáveis preditoras, ou seja, uma vez que o modelo é treinado, ele pode ser usado para prever a probabilidade de um resultado positivo para novos valores das variáveis independentes. Um limiar de decisão é então escolhido para transformar essa probabilidade em uma previsão binária. Por exemplo, se o limiar for 0,5, uma probabilidade maior que 0,5 é considerada como um resultado positivo (1) e uma probabilidade menor que 0,5 é considerada como um resultado negativo (0) (BATISTA; FILHO, 2019).

Para finalizar um modelo de regressão logística é considerado de alta qualidade quando contém várias propriedades desejáveis: poder preditivo, interpretabilidade, significância, robustez a erros nos dados e esparsidade, entre outras. Para atingir esses objetivos concorrentes, os modeladores incorporam essas propriedades de forma iterativa à medida que aprimoram o modelo (BERTSIMAS; KING, 2017).

KNN

O KNN é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para classificação e regressão. Ele funciona encontrando os "k" exemplos de treinamento mais próximos (ou seja, os "k" vizinhos mais próximos) do exemplo de teste que se deseja classificar ou prever.

O KNN é um algoritmo não paramétrico, aonde a estrutura do modelo será determinada pelo *dataset* utilizado, ou seja, baseado nas suas instancias. O KNN usa um banco de dados que estão agrupados em várias classes e tenta classificar os dados de amostra. O KNN não assume nenhuma distribuição de dados subjacente e por isso é chamado de não paramétrico (BATISTA; FILHO, 2019; RAY, 2019).

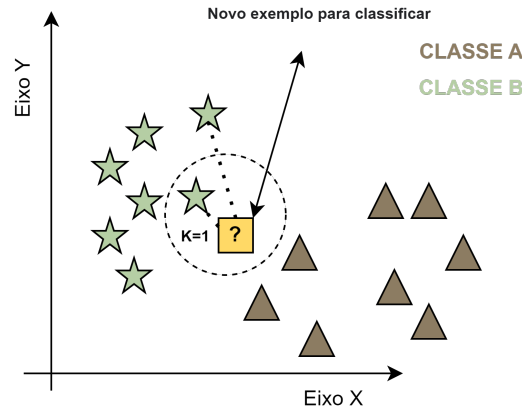
As vantagens do algoritmo KNN são: é um dos mais simples algoritmos a se implementar, boa precisão na maioria dos casos aplicados, a facilidade para se efetuar *tunning* dos poucos parâmetros existentes. Outra vantagem é o tempo para efetuar treinamento dos dados, sendo um dos mais rápidos para esta atividade. Já as desvantagens são: uma demora excessiva

na fase de teste e o alto consumo de memória para realizar esta atividade de teste, uma vez que o mesmo armazena tudo *dataset* em memória. O KNN não é indicado para conjuntos de dados grandes, uma vez que ele precisa calcular a distância entre o exemplo de teste e todos os exemplos de treinamento, quanto mais dados e mais dimensões desses registros, maior será o tempo para processar os cálculos. Uma vez que o mesmo trabalha com medidas de distância necessita sempre que nos atentemos a escala dos valores utilizados, para que não gere resultados equivocados. Além disso, o desempenho do KNN pode ser afetado por dados ruidosos ou desbalanceados, que podem resultar na degradação dos resultados (RAY, 2019).

O método KNN também é chamado de aprendizado lento ou melhor dizendo, é um algoritmo preguiçoso (*lazy*). Os algoritmos do tipo *lazy*, não necessitam de dados de treinamento para gerar o modelo, o que diminui em partes o processo inicial, mas em contrapartida gerará uma necessidade de análise posterior mais apurada. No caso de algoritmos que não necessitam de treinamento, todos os dados obtidos no *dataset* serão utilizados na fase de teste, resultando em um treinamento muito rápido e em um teste e validação lentos, momento o qual necessitamos estar bem atentos aos resultados gerados (LUZ, 2019).

Neste algoritmo possuímos uma variável K , este parâmetro significa a quantidade de vizinhos (*neighbors*). Em casos de modelos binários, aonde possuímos apenas duas classes, em geral aplicamos valores ímpares a K , mas lembre-se que cada caso é um caso. Imagine que temos um valor $P1$ o qual queremos prever, entre um grupo de duas classes aonde o valor atribuído a K foi 1, primeiro iremos identificar o ponto mais próximo a ele e depois qual a *label* que o identifica (classe A). Após identificar o ponto mais próximo e identificar a *label* deste ponto (classe A), iremos prever a que classe o ponto $P1$ faz parte. Para identificar de fato a que grupo o ponto $P1$ faz parte, iremos realizar uma votação aonde a maioria irá dizer a que classe este ponto $P1$ realmente faz parte. Na figura 6 é apresentado o funcionamento do KNN (LUZ, 2019).

Figura 6 – Como funciona o algoritmo KNN.



Fonte: (LUZ, 2019)

Vamos tentar identificar a distância existente entre o ponto P1 e os demais pontos do meu *dataset*, como $K=1$ o algoritmo irá verificar ponto a ponto, caso coloque o valor $K=3$ ele irá olhar a distância de P1 em relação a 3 pontos. Desta forma teremos a distância existente entre P1 e todos os pontos do meu *dataset*, assim conseguiremos saber a quais pontos P1 é mais próximo, ou seja, a classe mais similar. Assim a "votação" será concluída, e saberemos como classificar P1 (LUZ, 2019).

Para encontrar a distância entre os pontos pode ser utilizados as seguintes medidas de distância Euclidiana, Hamming, Manhattan e Markowski. A medida apresentada abaixo será a distância Euclidiana:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \dots + (x_n - x'_n)^2} \quad (3.12)$$

Dado um número inteiro positivo ímpar k , e uma instância a ser classificada x , sob uma métrica de distância d , o classificador KNN executa as seguintes ações:

- Para todo o conjunto de treinamento, calcula-se a distância d em relação a x ;
- Os k de treinamento mais próximos de x são agrupados em um conjunto A ;
- A probabilidade condicional para cada classe é definida por:

$$P(y = j | X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j) \quad (3.13)$$

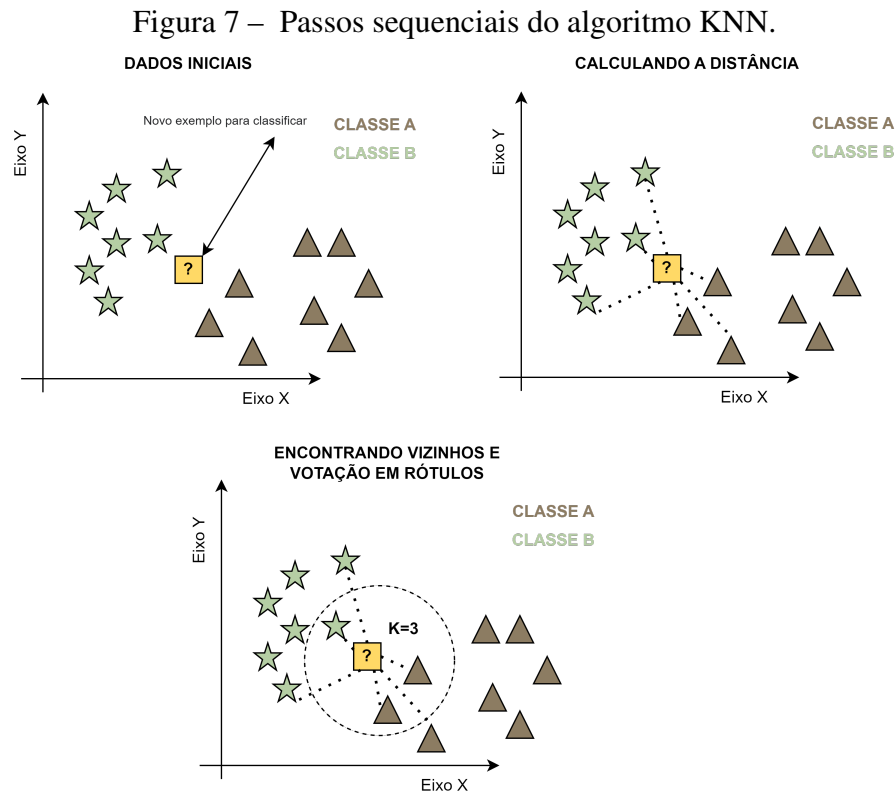
onde $I(x)$ é uma função que retorna 1 se seu argumento for verdadeiro e 0, caso contrário.

- Por último, a instância x é então classificada como a classe que apresentou maior probabilidade condicional (LUZ, 2019).

De forma resumida o KNN realiza os seguintes passos abaixo:

1. Calcula da distância;
2. encontra os pontos vizinhos mais próximos;
3. votar a *label* para o ponto a ser previsto.

A figura 7 apresenta os passos sequências do algoritmo KNN.



Fonte: (LUZ, 2019)

O número de vizinhos (*neighbor* = K) no algoritmo KNN é um hiperparâmetro o qual devemos escolher ao construir um modelo com base no KNN, então considere o K como a variável de controle para a predição/classificação. A máxima, não existe um número mágico, pois cada *dataset* é de um jeito, costumam dizer que dados são vivos e mutáveis e cada caso ele terá um comportamento, ou seja, um padrão. Valores atribuídos a K que sejam muito baixos serão influenciados por ruídos nos dados gerando assim influência no resultado final obtido, em contrapartida valores muito altos serão computacionalmente custosos.

Quando K apresenta um valor pequeno, o modelo não consegue observar a distribuição geral das classes nos conjuntos de dados, o que leva para uma situação de baixo viés, porém alta variância, tendo então o modelo apresentando sobreajuste. Já os resultados com valores altos atribuídos a K gerando vários vizinhos, resultarão em uma fronteira de decisão mais sensível,

resultando uma baixa variância dos resultados, mas um risco altíssimo de enviesamento.

Uma informação importante é que a maior parte do processamento no KNN é consumido durante a etapa de predição no algoritmo. Já o treinamento é um processo simples de memorização das instâncias e de seus rótulos. Um cuidado especial que deve se ter com o KNN, é que o mesmo tem seu desempenho prejudicado à medida que a quantidade de instâncias aumenta, uma vez que esse algoritmo é baseado em instâncias (FACELI *et al.*, 2011).

Como já mencionado nesta seção, os especialistas testam em sua grade maioria valores ímpares para K , pois os melhores resultados são valores não pares. A melhor maneira de se obter o melhor K é testando diversos valores. O trabalho de Jiang *et al.* (2007) sugere o uso da técnica de validação cruzada para estimar o valor adequado de k .

SVM

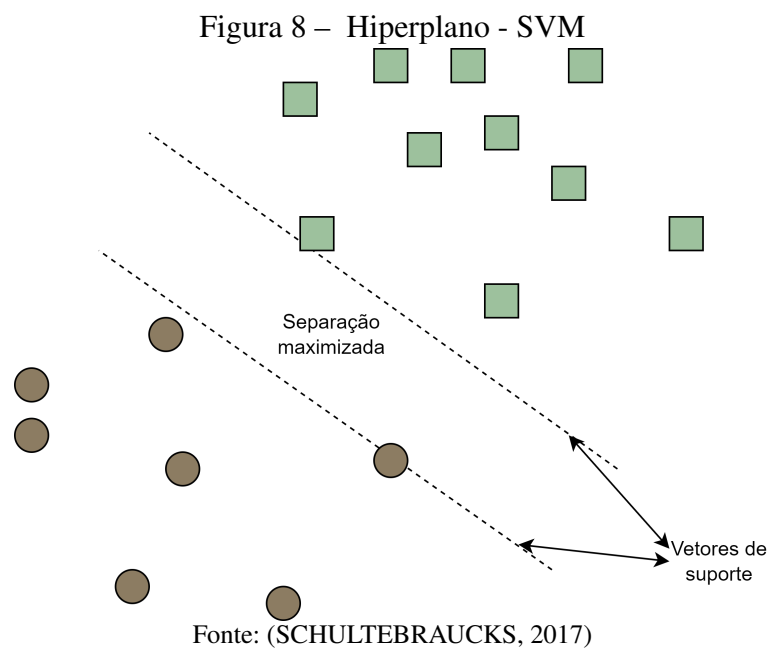
O SVM é um algoritmo de aprendizado de máquina supervisionado usado para regressão e classificação de vetores multidimensionais. O algoritmo funciona construindo um hiperplano que separa os exemplos de treinamento de diferentes classes no espaço de características. Originalmente, foi desenvolvido como uma classificação método linear, generalizado posteriormente para um classificador não linear e, por último, estendido para problemas de regressão (VAPNIK; YA., 1964; CORTES; VAPNIK, 1995).

O SVM de uma forma simplista pode ser definido como uma ferramenta de previsão em que procuramos um determinado linha ou limite de decisão denominado como hiperplano que separa facilmente os conjuntos de dados ou classes, portanto, evita o ajuste excessivo aos dados. Ele usa espaço de hipótese de um espaço linear em um alto espaço de recurso dimensional. Também é capaz de classificando os dados não lineares onde ele usa o kernel funções. O SVM pode usar diferentes funções kernel (como RBF, polinomial ou linear) para transformar os dados em um espaço de características de maior dimensionalidade, onde pode ser mais fácil encontrar um hiperplano separador. A escolha da função kernel adequada pode afetar o desempenho do modelo (SOMVANSHI *et al.*, 2016; FACELI *et al.*, 2011).

Embora o SVM tenham sido propostos por Vapnik e Ya. (1964), eles não receberam atenção significativa até os últimos anos, quando tornaram-se um estimador promissor em campos orientados a dados. O método SVM tem como ideia básica transformar os recursos de entrada em um espaço de dimensão superior, onde as duas classes podem ser separadas linearmente por uma superfície de alta dimensão, conhecida como hiperplano (RODRIGUES-GALIANO *et al.*,

2015).

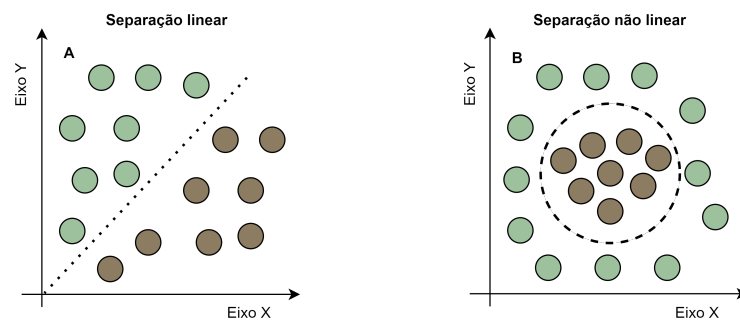
O SVM pode ser usado tanto para problemas de classificação e regressão. Para classificação temos muitos limites de decisão, que são capazes de classificar os conjuntos de dados, com isso surgem os hiperplanos. Na figura abaixo exigimos um hiperplano com alguns limites de decisão, o que significa que de todos os hiperplanos ou limites de decisão apenas um deles tem que ser selecionado para classificação (SCHULTEBRAUCKS, 2017). Na figura 8, podemos ver o hiperplano caracterizados com os dados divididos por classes.



Outra informação importante é que existem os dados lineares e os dados não lineares. Falamos de dados lineares, quando podemos classificar os dados com um classificador linear. O classificador linear toma sua decisão de classificação com base em uma combinação linear de características dos dados. As características também são conhecidas como recursos em aprendizado de máquina (SCHULTEBRAUCKS, 2017).

Na figura 9, podemos ver duas abordagens. A primeira no plano A podemos separar linearmente os rótulos alvo com uma linha de decisão. Poderíamos usar o SVM. Na figura B não podemos separar linearmente os rótulos alvo, pois os dados são mais complexos para ser divididos. Portanto, não podemos usar apenas um método de classificação linear. Felizmente, o SVM pode fazer ambas, classificação linear e não linear com o truque *kernel* (SCHULTEBRAUCKS, 2017).

Figura 9 – Dados lineares e dados não lineares

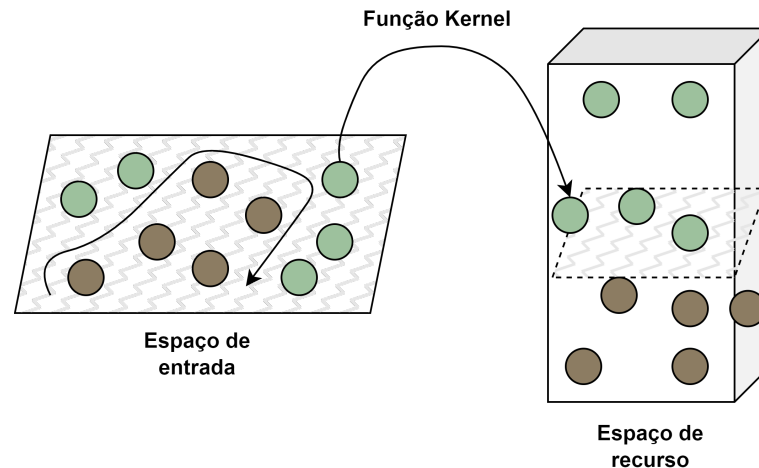


Fonte: (SCHULTEBRAUCKS, 2017)

No caso de dados não lineares, o truque do *kernel* permite SVM formar limites não lineares em lineares. Essa função mapeia o espaço de entrada separável não linear em um espaço de recurso separável linear de dimensão mais alta. O método do kernel tem como função principal ajudar a encontrar padrões e relações para alcançar uma alta precisão em seu modelo final de aprendizado de máquina (SCHULTEBRAUCKS, 2017). O método é executado em 3 etapas:

1. O algoritmo sempre usa apenas os produtos internos dos conjuntos de dados;
2. O algoritmo passa os dados originais através de mapas não lineares para formar novos dados em novas dimensões. Nestes espaços de características novas podemos então separar o conjunto de treinamento de forma linear. Quando mapeamos a solução de volta para o espaço de entrada, obtemos uma solução não linear, mas separada;
3. O algoritmo apresenta um produto escalar dos dados após as etapas 1 e 2 fazendo um mapeamento não linear no conjunto de dados, armazenando em tabelas e depois realizando uma pesquisa de tabela.

Resumindo o SVM tem como função projetar um hiperplano para construir um modelo de classificação. O hiperplano divide os rótulos de destino com uma margem maximizada. E usando o truque do *kernel*, o SVM pode classificar também os dados não lineares. Como apresentado na figura 10.

Figura 10 – Função *Kernel*

Fonte: (SCHULTEBRAUCKS, 2017)

Uma desvantagem do SVM é que ele pode ser sensível ao parâmetro de regularização C e à escolha da função kernel, o que pode levar a *overfitting* ou *underfitting* do modelo. Além disso, o treinamento de SVM pode ser computacionalmente caro para grandes conjuntos de dados.

Naive Bayes

O *Naive Bayes* (NB) é um algoritmo de aprendizado de máquina supervisionado usado para classificação. Ele se baseia no Teorema de Bayes para estimar a probabilidade de uma instância pertencer a uma determinada classe, dado um conjunto de características ou atributos (RUSSELL; P., 2009).

O algoritmo supõe que os atributos são independentes entre si, o que é uma suposição "ingênua" (daí o nome "Naive") mas muitas vezes eficaz em situações da vida real. Com base nessa suposição, o *Naive Bayes* pode calcular facilmente a probabilidade conjunta de todos os atributos, dada a classe, usando a probabilidade condicional de cada atributo dado a classe (ESCOVEDO; KOSHIYAMA, 2020).

O classificador NB é uma família de classificadores probabilísticos simples baseados em uma suposição comum de que todos as características são independentes umas das outras. Os diferentes classificadores NB, diferem principalmente pelas suposições que eles fazem em relação à distribuição de recursos. As suposições sobre a distribuição de características são chamadas de modelos de eventos do classificador NB. Para recursos discretos são populares, as distribuições multinomiais ou de Bernoulli. Esses pressupostos levam a dois modelos distintos, que muitas vezes são confundidos. Ao lidar com feições contínuas, uma suposição típica é a distribuição gaussiana (RUSSELL; P., 2009; JOHN; P., 1995; XU, 2016).

Os algoritmos de *Naive Bayes* derivam do Teorema de Bayes, a partir do qual é possível prever um evento relacionado a um conhecimento prévio. Entretanto, diferentes distribuições podem ser usadas para calcular a probabilidade de cada um dos atributos no processo de aprendizado. Os algoritmos de *Naive Bayes* são, essencialmente, previsores de atributos categóricos ou discretos (LENZ *et al.*, 2020).

Apesar das suposições aparentemente simplificadoras, o classificador NB funciona muito bem em muitas aplicações complexas do mundo real, como classificação de texto, como a extração de frase-chave e diagnóstico médico. Como um dos algoritmos de aprendizado indutivo mais eficientes, o classificador NB é frequentemente usado como linha de base na classificação de texto porque é rápido e fácil de implementar. Além disso, com pré-processamento adequado, é competitivo com mais métodos avançados incluindo o SVM (XU, 2016).

Classificador Bayesiano NB com modelo de evento multinomial é semelhante ao da contraparte clássica, mas NB classificador com modelo de evento gaussiano é obviamente melhor do que a contraparte clássica. O *Naive Bayes* Multinomial é comumente usado para classificação de texto, como em análise de sentimentos (XU, 2016).

Formalmente, seja $X(A_1, A_2, \dots, A_n, C)$ um conjunto de dados. Considere que $c_1, c_2, \dots, c_n$ são as classes do problema e valores possíveis do atributo C e que R é um novo exemplo que deve ser classificado. Sejam ainda $a_1, a_2, \dots, a_k$ os valores que R assume para os atributos previsores $A_1, A_2, \dots, A_k$, respectivamente. Resumidamente, o algoritmo consiste em dois passos:

1. Calcular as probabilidades condicionais $P(C = c_i | R), i = 1, 2, 3, \dots, k$;
2. Indicar como saída do algoritmo a classe c tal que $P(C = c | R)$ seja máxima, quando considerados todos os valores possíveis do atributo alvo C .

A intuição por trás do algoritmo é dar mais peso para as classes mais frequentes, considerando que os atributos são estatisticamente independentes entre si. Apesar de isto não ocorrer em muitos casos práticos, o método mostra-se bastante efetivo mesmo nos casos em que os atributos não sejam estatisticamente independentes (ESCOVEDO; KOSHIYAMA, 2020).

O *Naive Bayes* tem a vantagem de ser rápido e de exigir relativamente poucos exemplos de treinamento. No entanto, ele pode ser sensível a atributos correlacionados e a presença de *outliers*, o que pode afetar a precisão do modelo.

3.3.2 Métodos Ensemble

Uma das classes mais utilizada hoje em aprendizagem de máquina é o *ensemble learning*, um conjunto de técnicas que combinam as previsões de vários modelos fracos para obter um melhor desempenho preditivo. A seleção de recursos é um tópico importante no aprendizado de máquina. Nos últimos anos, através da integração da aprendizagem em conjunto, a abordagem de seleção de recursos baseada em aprendizado em conjunto foi proposta e estudada. Sua estratégia é simplesmente força na unidade, pois combinações eficientes de aprendizes fracos podem gerar modelos mais precisos e robustos, ou seja, sua ideia geral é gerar vários seletores de recursos diversos e combinar suas saídas. Essa abordagem é superior aos métodos convencionais de seleção de recursos em muitos aspectos. Entre eles, sua vantagem mais proeminente é a capacidade de lidar com problemas de estabilidade que geralmente são ruins nos métodos de seleção de recursos existentes (GUAN *et al.*, 2014).

Um classificador *ensemble*, consiste em um conjunto de classificadores treinados individualmente, classificadores de base, cujas decisões são de alguma forma combinadas para obter um melhor desempenho preditivo. Sua estratégia é simplesmente força na unidade, pois combinações eficientes de aprendizes fracos podem gerar modelos mais precisos e robustos. O algoritmo ensemble possui três tipos de algoritmos *Bagging*, *Boosting* e *Stacking* (MARQUES *et al.*, 2012).

O *Bagging* (*Bootstrap Aggregating*), gera um conjunto de dados por amostragem *bootstrap* (um tipo de técnica estatística) dos dados originais. O conjunto de dados gera um conjunto de modelos utilizando um algoritmo de aprendizagem simples por meio da combinação por votos para classificação. O seu uso é particularmente atraente quando a informação disponível é de tamanho limitado (BREIMAN, 1996).

No *Bagging* os classificadores são treinados de forma independente por diferentes conjuntos de treinamento através do método de inicialização. Para construí-los é necessário montar k conjuntos de treinamento idênticos e replicar esses dados de treinamento de forma aleatória para construir k redes independentes por reamostragem com reposição. Em seguida, deve-se agregar as k redes através de um método de combinação apropriada, tal como a maioria de votos (BREIMAN, 1996). Uma das técnicas mais conhecidas que é do tipo *Bagging* é o algoritmo *Random Forest*.

No *Boosting*, de forma semelhante ao *Bagging*, cada classificador é treinado usando um conjunto de treinamento diferente. A abordagem por *Boosting* original foi proposta por

Schapire em 1990. A principal diferença em relação ao *Bagging* é que os conjuntos de dados reamostrados são construídos especificamente para gerar aprendizados complementares e a importância do voto é ponderado com base no desempenho de cada modelo, em vez da atribuição de mesmo peso para todos os votos. Essencialmente, esse procedimento permite aumentar o desempenho de um limiar arbitrário simplesmente adicionando *learners* mais fracos. Dada a utilidade desse achado, *Boosting* é considerado uma das descobertas mais significativas em aprendizado de máquina (LANTZ, 2013).

Existem algumas técnicas mais conhecidas do tipo *Boosting* como o algoritmo *Categorical Boosting* (CatBoost), *Light Gradient Boosting Machine* (LightGBM) e *Extreme Gradient Boosting* (XGBoost). As seguintes técnicas de classificador ensemble serão apresentadas nesta seção: *CatBoost*, *LightGBM*, *Random Forest* e *XGBoost*.

CatBoost

O método *CatBoost* é um algoritmo de aprendizado de máquina supervisionado usado para classificação. Ele é baseado na técnica de *boosting* de árvore, assim como o *XGBoost* e o *LightGBM*, mas usa uma abordagem de codificação de categoria para lidar com recursos categóricos e uma técnica de processamento de ordenação para melhorar o desempenho. O *CatBoost*, é uma biblioteca nova de aumento de gradiente de código aberto que trabalha de forma bastante eficiente com recursos categóricos e supera o desempenho público existente de implementações disponíveis de aumento de gradiente em termos de qualidade (DOROGUSH *et al.*, 2018).

O CatBoost, assim como todas as implementações de aumento de gradiente padrão, ele constrói cada nova árvore para aproximar os gradientes do modelo atual. No entanto, todos os algoritmos de *boosting* clássicos sofrem de *overfitting* causado pelo problema de estimativas de gradientes pontuais tendenciosos. Gradientes usados em cada passo são estimados usando os mesmos pontos de dados em que o modelo atual foi construído. Isso leva a uma mudança de a distribuição de gradientes estimados em qualquer domínio do espaço de características em comparação com o verdadeiro distribuição de gradientes neste domínio, o que leva ao *overfitting* (DOROGUSH *et al.*, 2018).

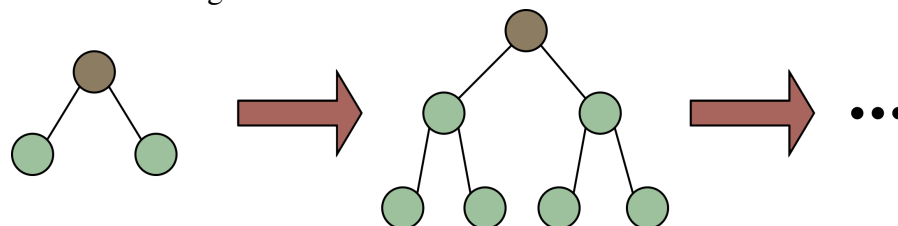
Em muitos algoritmos de *Gradient Boosting Decision Tree* (GBDT), a construção da próxima árvore compreende duas etapas: escolher a estrutura de árvore e definindo valores em folhas após a estrutura de árvore ser corrigida. Para escolher a melhor estrutura de árvore, o

algoritmo enumera através de diferentes divisões, constrói árvores com essas divisões, define valores nas folhas obtidas, pontua as árvores e seleciona a melhor divisão. Os valores da folha em ambas as fases são calculados com aproximações para gradientes (DOROGUSH *et al.*, 2018).

O *CatBoost* usa árvores inconscientes como preditores básicos. Em tais árvores, o mesmo critério de divisão é usado em todo um nível da árvore. Essas árvores são equilibradas e ficam menos propensas a *overfitting*. Árvores oblíquas impulsionadas por gradiente foram usadas com sucesso em várias tarefas de aprendizado. No esquecimento árvores cada índice de folha pode ser codificado como um vetor binário com comprimento igual à profundidade da árvore. Este fato é amplamente utilizado no avaliador de modelo *CatBoost*: primeiro binarizamos todos os recursos flutuantes usados, estatísticas e recursos codificados *one-hot* e, em seguida, use recursos binários para calcular as previsões do modelo (DOROGUSH *et al.*, 2018).

O *CatBoost* usa a estratégia *level-wise* conforme a figura 11, realizando um crescimento como uma árvore equilibrada.

Figura 11 – Crescimento de árvore em nível



Crescimento de árvore em nível

Fonte: (BANERJEE, 2020)

O *CatBoost* pode ser usado quando requer um tempo de treinamento curto em uma base robusta, trabalhando com um pequeno conjunto de dados e quando se possui um conjunto de dados categóricos. O único incomodo o pelo qual muitos não consideram o uso do *CatBoost* deve-se à pequena dificuldade em ajustar os parâmetros para otimizar o modelo para recursos categóricos (ADEBAYO, 2021).

Os principais recursos do *CatBoost* são:

1. *CatBoost* constrói árvores simétricas: Em cada passo, as folhas da árvore anterior são divididas usando a mesma condição. O par de divisão de recursos que representa a menor perda é selecionado e usado para todos os nós do nível. Essa arquitetura de árvore balanceada ajuda na implementação eficiente da CPU, diminui o tempo de previsão, torna os aplicadores de modelo rápidos e controla o *overfitting*, pois a estrutura serve como

regularização (JOHN, 2022);

2. Impulso ordenado: Os algoritmos de reforço clássicos são propensos a *overfitting* em conjuntos de dados pequenos e ruidosos devido a um problema conhecido como mudança de previsão. Ao calcular a estimativa de gradiente de uma instância de dados, esses algoritmos usam as mesmas instâncias de dados com as quais o modelo foi construído, portanto, não têm chances de experimentar dados não vistos. O *CatBoost*, por outro lado, usa o conceito de reforço ordenado, uma abordagem orientada por permutação para treinar o modelo em um subconjunto de dados enquanto calcula os resíduos em outro subconjunto, evitando assim o vazamento e o ajuste excessivo (JOHN, 2022);
3. Suporte a recursos nativos: *CatBoost* suporta todos os tipos de recursos, sejam numéricos, categóricos ou de texto, e economiza tempo e esforço de pré-processamento (JOHN, 2022).

Em resumo, o *CatBoost* é um algoritmo de *boosting* de árvore eficiente e preciso que usa estratégias inovadoras para lidar com recursos categóricos e melhorar o desempenho. Ele é amplamente utilizado em competições de ciência de dados e em aplicações do mundo real.

LightGBM

O método *LightGBM* é um algoritmo de aprendizado de máquina supervisionado usado para classificação. Ele é baseado na técnica de *boosting* de árvore, assim como o *XGBoost*, mas usa uma estratégia de crescimento de árvore em folhas (*leaf-wise*) e uma abordagem de amostragem por nó. O *LightGBM* é um algoritmo GBDT de código aberto da Microsoft. Ele usa o algoritmo baseado em histograma para acelerar o processo de treinamento, reduz o consumo de memória e combina comunicação de rede avançada para otimizar o aprendizado paralelo, chamado algoritmo de árvore de decisão de votação paralela (WANG *et al.*, 2017).

O *LightGBM* é um algoritmo relativamente novo e possui uma longa lista de parâmetros. Ele é prefixado com o "L" (*Light*) por causa de sua alta velocidade. Já o *GBM*, pois pode lidar com o grande tamanho dos dados e consumindo pouca memória para ser executado. De forma conceitual o Algoritmo *LightGBM* divide a folha da árvore com o melhor ajuste, embora outros algoritmos de árvore consigam realizar a divisão em profundidade. Não é aconselhável usar *LightGBM* em pequenos conjuntos de dados, pois é sensível ao *overfitting* (AHAMED; ARYA, 2021).

O algoritmo *LightGBM* usa dois conceitos. Eles são GBDT ou *Gradient based one-sided sampling* (GOSS). Esses são os métodos de aprendizado de máquina amplamente

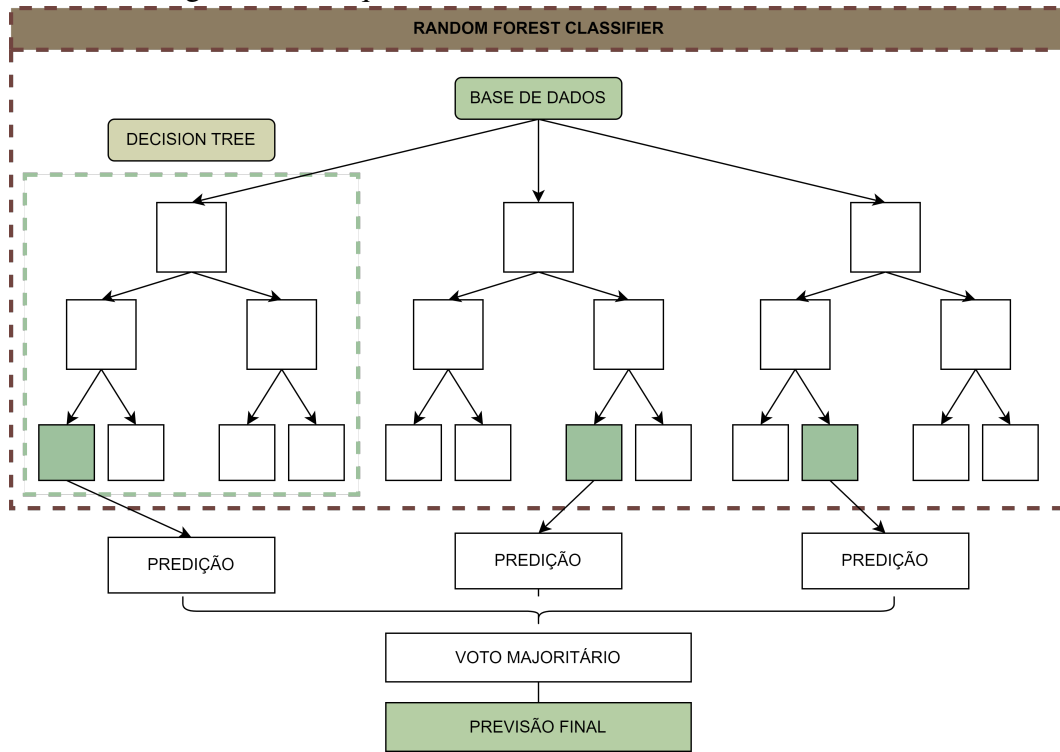
usados algoritmo envolvido para previsão de estudo. O GBDT é um algoritmo de *boosting*. O *boosting* é conseguido através da construção de um modelo a partir dos dados de trabalho, produzindo um segundo modelo que vai retificar os erros do modelo inicial estrutura. A eficácia, correção e interoperabilidade são seus fatores cruciais. É um avançar na agenda da árvore que é muito eficaz e acessível (AHAMED; ARYA, 2021).

Nos últimos anos, com a ascensão do *bigdata*, o GBDT está enfrentando novos desafios, especialmente no *trade-off* entre proficiência e precisão. As implementações convencionais do GBDT examinam todas as instâncias de dados necessárias para obter as informações necessárias. Isso vai levar em uma proporcionalidade entre o número de casos e o número de estruturas envolvidas (AHAMED; ARYA, 2021).

Random Forest

O *Random Forest* é um algoritmo de aprendizagem supervisionada em aprendizagem de máquina, baseado no método de árvores de decisão, podendo ser utilizado para problemas de classificação e regressão. O *Random Forest* foi apresentado pela primeira vez por Ho (1995). Trata-se de um *ensemble* de árvores de decisão que são construídas com base em *bagging* (*bootstrap aggregation*), utilizando a predição com a combinação de árvores de decisão, de modo que em que cada árvore é treinada em uma amostra aleatória dos dados de treinamento e em um subconjunto aleatório de recursos. Em seguida, a previsão final é feita por meio de uma votação das previsões de cada árvore. O objetivo é reduzir o sobreajuste e aumentar a precisão das previsões, tornando o modelo mais robusto e geralmente com melhor desempenho do que uma única árvore de decisão. (BREIMAN, 2001; BATISTA; FILHO, 2019).

Algumas características importantes do algoritmo *Random Forest* é que tecnicamente ele foi comprovado que é mais preciso do que o algoritmo clássico de *Decision Tree*, outra característica importante é que ele tem uma maneira eficaz de tratar os dados ausentes. Ele ainda pode produzir uma previsão razoável sem ser necessário um ajuste de hiperparâmetro e consegue resolver o problema de sobreajuste no momento que utiliza as árvores de decisão. Em cada árvore, um subconjunto de recursos é habilitado de forma aleatória no pronto divisório de cada nó. Abaixo a figura 12 apresenta o esquema de funcionamento do algoritmo (JUNIOR *et al.*, 2019).

Figura 12 – Esquema de funcionamento do *Random Forest*

Matematicamente segundo Lin e Jeon (2006) a equação da *Decision Tree* é dado por:

$$\hat{y} = \sum_{i=1}^n W(Xi, Xi') yi \quad (3.14)$$

Além disso, é possível estender a equação 3.14 para compor a previsão de *Random Forest*, que é dada por:

$$\hat{y} = \frac{1}{k} \sum_{j=1}^m \sum_{i=1}^n W j(Xi, Xi') yi \quad (3.15)$$

XGBoost

O XGBoost é um algoritmo de aprendizado de máquina supervisionado usado para classificação, regressão e outros tipos de tarefas de previsão. Ele funciona construindo uma sequência de árvores de decisão em que cada árvore é treinada para corrigir os erros das árvores anteriores. Durante o treinamento, o modelo otimiza uma função objetivo que combina a perda do modelo e um termo de regularização que ajuda a evitar o sobreajuste. O modelo também usa técnicas avançadas, como amostragem estocástica e divisão de dados em colunas, para aumentar a eficiência e o desempenho. Em resumo, o *XGBoost* é um algoritmo de *boosting* que combina várias árvores de decisão e usa técnicas avançadas para melhorar o desempenho e reduzir o

tempo de treinamento. Ele é um método recente e com resultados muito competitivos, e por isso frequentemente usado em competições de ciência de dados devido à sua eficácia e capacidade de lidar com grandes conjuntos de dados (CHEN; GUESTRIN, 2016; ??; BENTÉJAC *et al.*, 2019).

O *XGBoost* pode ser tratado como um algoritmo de localização de divisão ciente de esparsidade que pode lidar com precisão com valores ausentes. O algoritmo ajuda no processo de criação de um Árvore de classificação e regressão (CART) no *XGBoost* para calcular os valores ausentes diretamente. A formação da árvore CART é feita selecionando a melhor divisão para cada nó da árvore (RUSDAH; MURFI, 2020).

O modelo *XGBoost* é considerado mais eficaz em alguns estudos que *Deep Neural Network* (DNN). O estudo foi apresentado em uma conferência tinha como objetivo apresentar o desempenho de máquina superficial comparando *XGBoost* e DNN comparando os desempenhos de equipes. O artigo afirma que o *XGBoost* apresentou melhores resultados de aprendizado e previsão do que o DNN (GIANNAKAS *et al.*, 2021).

O *XGBoost* também é muito utilizado para auxiliar diagnóstico no campo da medicina inteligente por ser um algoritmo de aprendizado integrado. O *XGBoost* resulta em uma melhoria significativa no efeito de desempenho comparado a modelos básicos como de árvore e regressão logística. Para amostras com ausência de valores, o *XGBoost* pode aprender automaticamente para tomar sua decisão. Podemos considerar que ele é extremamente flexível uma vez que possui um grande número de hiperparâmetros passíveis de aperfeiçoamento, sendo possível conseguir ajustar adequadamente para o cenário do problema enfrentado (SUN *et al.*, 2020).

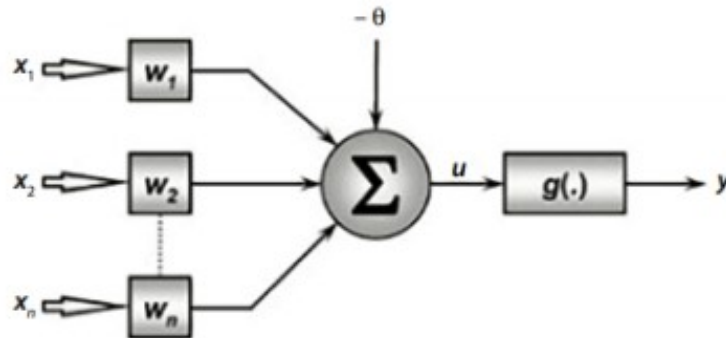
3.3.3 RNA

Uma rede neural é uma série de algoritmos que se esforçam para reconhecer relacionamentos subjacentes em um conjunto de dados por meio de um processo que imita o funcionamento do cérebro humano. As redes neurais referem-se a sistemas de neurônios, de natureza orgânica ou artificial. As redes neurais podem adaptar-se à mudança de entrada, para que a rede gere o melhor resultado possível sem a necessidade de redesenhar a saída critério (RAY, 2019).

Uma rede neural artificial funciona através de três camadas interconectadas: a camada de entrada, que recebe os dados de entrada; a camada oculta, que processa esses dados; e a camada de saída, que envia o resultado calculado. Essas camadas trabalham juntas para

permitir que a rede neural artificial execute tarefas complexas, como reconhecimento de voz, imagem e aprendizado de máquina (SILVA *et al.*, 2019). A figura 13 ilustra o processo descrito.

Figura 13 – Rede Neural Artificial

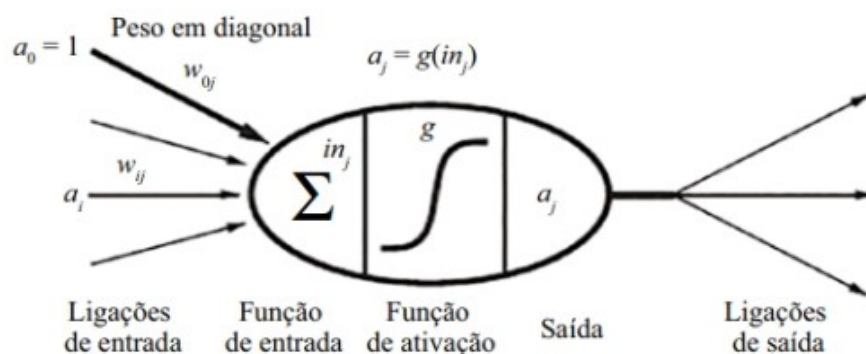


Fonte: (SILVA *et al.*, 2019)

As redes neurais artificiais realizam a construção de modelos matemáticos que simulam as atividades do cérebro, descrevendo as atividades mentais por meio da eletroquímica das células cerebrais (os neurônios) (SILVA *et al.*, 2019).

De acordo com McCulloch e Pitts (1943), ele foi responsável por apresentar o primeiro modelo matemático de um neurônio, conhecido como modelo M-P. A partir desse modelo, as redes neurais artificiais começaram a demonstrar a interconexão de neurônios através de ligações de entrada e saída, onde a ativação de entrada pode ser qualquer estímulo que desencadeie funções de entrada e de ativação. A figura 14 apresenta de forma conceitual uma rede neural artificial.

Figura 14 – Conceito de Rede Neural Artificial



Fonte: (SILVA *et al.*, 2019)

As redes neurais são compostas de nós ou unidades, que se conectam por ligações direcionadas. Uma ligação da unidade i para a unidade j serve para propagar uma ativação a_i de i

j , sendo que cada ligação tem um peso numérico w_{ij} associado a ele, que, por sua vez, determina o sinal de conexão entre as unidades ou os nós de uma rede neural (SILVA *et al.*, 2019). Assim, essas redes implementam o seguinte somatório:

$$in_j = \sum_{i=0}^n w_{ij} a_i \quad (3.16)$$

O somatório anterior representa a função fundamental das redes neurais, chamada de função de ativação, que é uma representação do somatório das ligações de entrada em uma rede neural artificial. A seguir, o somatório para se obter a saída:

$$a_j = g(in_j) = g\left(\sum_{i=0}^n w_{ij} a_i\right) \quad (3.17)$$

As redes neurais artificiais possuem 3 categorias de aprendizagem: rede neural supervisionada, rede neural não supervisionada e rede neural reforçada. Na rede neural supervisionada, a saída da entrada é já conhecida. Já em relação a estrutura existem dois tipos: as redes neurais de camada única e as redes neurais de multicamada.

A estrutura de rede de camada única foi proposta em 1958, por Rosenblatt, e é a forma mais simples de estruturação de rede neural artificial, considerando que se constitui de uma camada neural e de um neurônio. Essas redes podem ter n entradas, porém, apenas uma saída com valores de 0 e 1 ou de -1 e 1. Já as estruturas de redes neurais de multicamada ou *perceptrons* de multicamadas são semelhantes aos *perceptrons* que servem como modelos de neurônios artificiais, presentes em somente uma camada ou em multicamadas. Uma das principais vantagens na construção da multicamada é que, nela, há a entrada de estímulos e uma camada intermediária de construção da resposta para encaminhar a saída, formando os grafos que simulam neurônios em multicamadas (SILVA *et al.*, 2019).

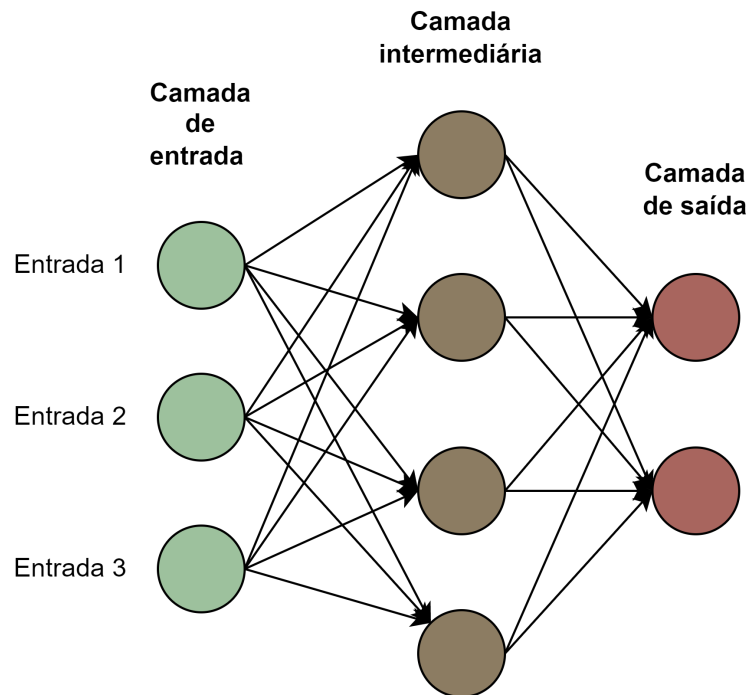
Multi Layer Perceptron (MLP)

O MLP evoluíram ao longo dos anos como uma técnica muito poderosa para resolver uma ampla variedade de problemas. Muito progresso foi feito na melhoria do desempenho e no entendimento como essas redes neurais operam (DELASHMIT; MANRY, 2005).

O MLP uma rede neural artificial composta por múltiplas camadas de neurônios, que são organizados em uma arquitetura *feedforward*. Cada neurônio em uma camada recebe entradas

de neurônios na camada anterior e produz uma saída que é enviada para os neurônios na camada seguinte. Cada camada é composta por nós e nas redes totalmente conectadas consideradas neste artigo, cada nó se conecta a cada nó nas camadas subsequentes. Cada MLP é composto por um mínimo de três camadas consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (DELASHMIT; MANRY, 2005). A figura 15 abaixo apresenta o MLP.

Figura 15 – Perceptron Multicamada - MLP



Fonte: (MOHANTY, 2019)

O Minsky e Papert (1969) analisaram matematicamente o *perceptron* e demonstraram que redes de uma camada não são capazes de solucionar problemas que não sejam linearmente separáveis. Como não acreditavam na possibilidade de se construir um método de treinamento para redes com mais de uma camada, eles concluíram que as redes neurais seriam sempre suscetíveis a essa limitação.

Contudo, o desenvolvimento do algoritmo de treinamento *backpropagation*, por Rumelhart *et al.* (1986), precedido por propostas semelhantes ocorridas nos anos 70 e 80, mostrou que é possível treinar eficientemente redes com camadas intermediárias, resultando no modelo de Redes Neurais Artificiais mais utilizado atualmente, as redes MLP, e treinadas com o algoritmo *backpropagation*.

O MLP treina iterativamente, pois a cada passo de tempo as derivadas parciais da função de perda em relação aos parâmetros do modelo são calculadas para atualização de parâmetros. Ele também pode ter um termo de regularização adicionado à função de perda que

reduz os parâmetros do modelo para evitar o *overfitting* (PEDREGOSA *et al.*, 2011).

Nessas redes, cada camada tem uma função específica. A camada de saída recebe os estímulos da camada intermediária e constrói o padrão que será a resposta. As camadas intermediárias funcionam como extratoras de características, seus pesos são uma codificação de características apresentadas nos padrões de entrada e permitem que a rede crie sua própria representação, mais rica e complexa, do problema (PEDREGOSA *et al.*, 2011).

O MLP possui suas vantagens e as suas desvantagens:

- Vantagens: Capacidade de aprender modelos não lineares e a capacidade de aprender modelos em tempo real (PEDREGOSA *et al.*, 2011);
- Desvantagens: O MLP com camadas ocultas tem uma função de perda não convexa onde existe mais de um mínimo local (PEDREGOSA *et al.*, 2011).

3.3.4 Validação cruzada

A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de um modelo, ou seja, a habilidade de um modelo de se adaptar a novos dados que não foram utilizados no treinamento. A validação cruzada é realizada dividindo-se os dados disponíveis em um conjunto de treinamento e um conjunto de teste, onde o conjunto de treinamento é utilizado para treinar o modelo e o conjunto de teste é utilizado para avaliar a performance do modelo em dados nunca antes vistos, com o objetivo geral de melhorar a confiabilidade estatística dos resultados. A utilização da validação cruzada tem altas chances de detectar se o seu modelo está sobreajustado aos seus dados de treinamento, ou seja, sofrendo *overfitting* (RABELLO, 2022; RAMEZAN *et al.*, 2013).

Overfitting é um problema comum em aprendizagem de máquina, que ocorre quando um modelo é treinado para se ajustar excessivamente aos dados de treinamento, ao ponto de perder a capacidade de generalização para novos dados. Em outras palavras, o modelo aprende os padrões presentes nos dados de treinamento tão bem que começa a memorizá-los ao invés de generalizá-los, tornando-se incapaz de fazer previsões precisas em dados que não foram vistos anteriormente.

Existem diversas formas de validação cruzada, sendo a mais comum a validação cruzada *k-fold*. Nessa técnica, os dados são divididos em k partições, onde $k - 1$ partições são utilizadas para treinar o modelo e a partição restante é utilizada para testar o modelo. Esse processo é repetido k vezes, de forma que cada partição seja utilizada como conjunto de teste em

uma das iterações. O resultado final é a média da performance do modelo em todas as iterações.

No caso da figura 16, temos um exemplo de validação cruzada com $k=5$, onde a base de dados foi dividida em 5 grupos. Em cada uma das 5 iterações, um dos grupos é selecionado como conjunto de teste e os outros 4 grupos são utilizados como conjunto de treinamento. Essa técnica é especialmente útil quando a quantidade de dados é limitada, pois permite que se obtenha uma estimativa mais precisa do desempenho do modelo em novos dados, ao mesmo tempo que se utiliza todo o conjunto de dados disponível para o treinamento do modelo.

Figura 16 – Processo de Validação Cruzada

Interação 1	Teste	Treino	Treino	Treino	Treino
Interação 2	Treino	Teste	Treino	Treino	Treino
Interação 3	Treino	Treino	Teste	Treino	Treino
Interação 4	Treino	Treino	Treino	Teste	Treino
Interação 5	Treino	Treino	Treino	Treino	Teste

Fonte: (RABELLO, 2022)

Validar a estabilidade do modelo de aprendizado de máquina é essencial para garantir sua eficácia em dados não vistos anteriormente. Não é suficiente ajustar o modelo aos dados de treinamento e esperar que ele funcione bem em dados reais, pois isso pode levar a um problema de sobreajuste, no qual o modelo se adapta demais aos dados de treinamento e não generaliza bem para novos dados. É importante, portanto, garantir que o modelo capture a maioria dos padrões dos dados corretos, sem capturar muito o ruído, ou seja, que apresente baixo viés e baixa variância (GUPTA, 2017).

3.3.5 As métricas de aprendizagem de máquina

As métricas são utilizadas para a avaliação do modelo, ou seja, o desempenho do aprendizado supervisionado pode ser medido usando métricas de avaliação. A métrica utilizada para determinação do melhor modelo depende do problema analisado. Aqui nessa seção serão apresentadas as principais métricas para avaliação de modelos de classificação de dados. apresentaremos nesta seção também a matriz de confusão (RUSDAH; MURFI, 2020).

A matriz de confusão é uma medição de desempenho do modelo de classificação. Dentro uma matriz de confusão, a classe prevista será comparada com a classe real. Cada coluna da matriz mostra os resultados previstos para a classe correspondente à coluna, enquanto cada linha mostra a classe real. A matriz de confusão é usada para calcular desempenho de algumas métricas como acurácia, precisão e sensibilidade (RUSDAH; MURFI, 2020).

Um modelo de classificação de dados visa realizar uma previsão com base em ocorrências passadas. Para isso, o modelo utiliza um conjunto de dados com entradas (indivíduos) e atributos (propriedades). Além disso, é necessário conhecer o resultado esperado para esse conjunto de dados (rótulos). Todas essas informações serão usadas para treinar um modelo que será utilizado para prever resultados esperados para novos dados que surgirem no futuro. Ao treinar esse modelo deve-se utilizar um conjunto de dados (não usados no treinamento) para testar o quanto o modelo acerta. Entretanto, não basta apenas contar a quantidade de acertos que seu modelo teve para dizer se ele é bom ou não. Dependendo do problema estudado, métricas diferentes devem ser utilizadas para essa avaliação. Entretanto, antes de apresentarmos essas métricas, precisamos entender alguns conceitos para classificações binárias: as classes que os dados preditos poderão receber (MARIANO, 2021).

Em um problema de classificação, há duas soluções possíveis: acerto ou erro. Entretanto, para um problema de classificação binária temos ainda duas outras classes possíveis, vamos chamá-las de classes positiva e negativa. De acordo com Ferrari e Silva (2017), em problemas de classificação binária, predições podem ter quatro possíveis classes que serão apresentadas logo abaixo:

- Verdadeiros positivos (VP) - quando o método diz que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva;
- Verdadeiros negativos (VN) - quando o método diz que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa;
- Falsos positivos (FP) - quando o método diz que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa;
- Falsos negativos (FN) - quando o método diz que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva.

Vamos definir siglas para algumas métricas como Sensibilidade (Sens), Acurácia (Acc) e Precisão (Prec). Diante disso a tabela 1 apresentará a fórmula matemática de cada métrica agrupada, mas cada fórmula será apresentada nas próximas seções posteriores.

Tabela 1 – As métricas para avaliar métodos de classificação.

MÉTRICA	FÓRMULA
Sensibilidade	$S = \frac{VP}{VP+FN}$
Especificidade	$E = \frac{VN}{FP+VN}$
Acurácia	$A = \frac{VP+VN}{VP+FP+VN}$
Precisão	$P = \frac{VP}{VP+FP}$
F1 score	$F1\ score = 2x \frac{P \times S}{P+S}$

Fonte: Elaborado pelo autor

Matriz de confusão

A matriz de confusão indica a quantidade de ocorrências que o programa teve para cada uma das quatro categorias. A tabela abaixo segue um modelo representativo de uma matriz de confusão. A matriz de confusão indica a quantidade de ocorrências que o programa teve para cada uma das quatro categorias como apresentado na figura 17.

Figura 17 – Matriz de confusão.

Matriz de confusão		Classe predita	
		Positiva	Negativa
Classe original	Positiva	VP	FN
	Negativa	FP	VN

Fonte: (FERRARI; SILVA, 2017)

A soma dos quatro campos deve ser igual a N .

$$N = VP + VN + FP + FN \quad (3.18)$$

Para obtermos o total de predições realizadas em cada classe, podemos somar os valores presentes em cada coluna. Observe como obtemos o total de valores preditos como positivos:

$$pred_p = VP + FP \quad (3.19)$$

Para obter os valores preditos como negativos, usamos:

$$pred_n = VN + FN \quad (3.20)$$

Para obtermos o total de valores reais somamos os valores em cada linha. Assim, para obter os valores reais positivos, calculamos:

$$real_p = VP + FN \quad (3.21)$$

Para obter os valores reais negativos, calculamos:

$$real_n = VN + FP \quad (3.22)$$

Para calcular acertos do método, usamos:

$$acertos = VP + VN \quad (3.23)$$

Para calcular os erros, usamos:

$$erros = FP + FN \quad (3.24)$$

Acurácia

A acurácia é considerada uma das métricas mais simples e importantes. Ela avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas, ou seja, **acurácia = total de acertos / número total de elementos**. Utilizando como base a matriz de confusão, podemos obter a acurácia pela fórmula:

$$A = \frac{VP + VN}{VP + FN + VN + FP} \quad (3.25)$$

Sensibilidade

Outra métrica que pode ser utilizada é a sensibilidade (também conhecida como *recall*). Essa métrica avalia a capacidade do método de detectar com sucesso resultados classificados como verdadeiros positivos em relação ao total de exemplos positivos. É muito útil

quando o foco é minimizar os falsos negativos (HALIMU *et al.*, 2019). Ela pode ser obtida pela equação:

$$S = \frac{VP}{VP + FN} \quad (3.26)$$

Especificidade

Por outro lado, a especificidade avalia a capacidade do método de detectar resultados negativos. Podemos calculá-la usando a equação:

$$E = \frac{VN}{VN + FP} \quad (3.27)$$

Precisão

A precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos:

$$P = \frac{VP}{VP + FP} \quad (3.28)$$

É útil quando o foco é minimizar os falsos positivos.

F1 score

A *F1 score* é uma média harmônica calculada com base na precisão e sensibilidade. Ela pode ser obtida com base na equação:

$$F1 = 2x\left(\frac{precisao * sensibilidade}{precisao + sensibilidade}\right) \quad (3.29)$$

O *F1 Score* leva em consideração tanto os falsos positivos quanto os falsos negativos. É uma métrica útil quando as classes estão desequilibradas.

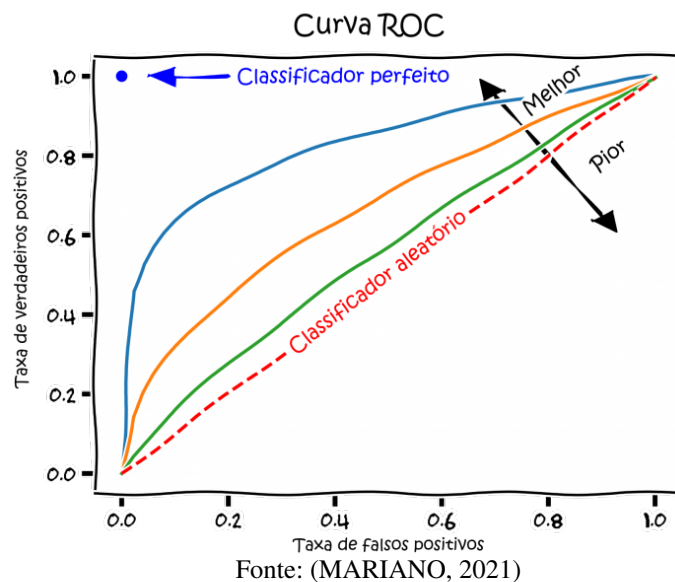
Curva de Receiver Operating Characteristic (ROC) e Area Under the Curve (AUC)

A curva ROC, ou "Curva Característica de Operação do Receptor" é um gráfico que permite avaliar um classificador binário. Essa visualização leva em consideração a Taxa de Verdadeiros Positivos (TVP) (sensibilidade - 0) e a Taxa de Falsos Positivos (TFP)

(especificidade - 1). Esse gráfico permite comparar diferentes classificadores e definir qual o melhor com base em diferentes pontos de corte. Na prática, quanto mais próximo do topo do eixo Y melhor o classificador (HALIMU *et al.*, 2019).

Na ilustração 18 é apresentado a curva ROC, onde o eixo Y armazena a TVP (sensibilidade - 0). O eixo X armazena a TFP (especificidade - 1). O ponto azul representa um classificador perfeito, isto é, um classificador que atinge 100% de verdadeiros positivos e 0% de falsos positivos. A linha azul claro indica um resultado melhor do que os apresentados pelas linhas laranja e verde. A linha tracejada vermelha indica o limiar aleatório. Resultados abaixo da linha diagonal vermelha são considerados classificadores ruins (MARIANO, 2021).

Figura 18 – Curva de ROC.



Uma curva ROC pode ser avaliada pela métrica AUC. A visibilidade com o AUC é bem melhor. O AUC calcula a área da forma bidimensional formada abaixo da curva. Essa métrica indica a probabilidade de duas previsões serem corretamente ranqueadas. A AUC será um valor entre 0 e 1, mas a interpretação do valor da AUC é uma probabilidade. Quanto maior esse valor, ou seja mais próximo de 1, melhor a capacidade do modelo em separar classes. Um bom modelo é um modelo que pode separar bem as duas classes, portanto, quanto menor a interseção entre as duas classes que nosso modelo tiver, melhor será e consequentemente maior será a AUC. Quando a AUC é igual a 0,5, o modelo não tem capacidade de separação de classe e é tão bom quanto uma escolha aleatória (SILVA, 2019).

4 REVISÃO BIBLIOGRÁFICA

Entender o quadro clínico das pessoas que já tiveram AVC é essencial para desenvolver um estudo que ajude a prevenir a recorrência do AVC. Para os especialistas na área da saúde, as maiores preocupações estão em identificar as características que mais contribuem para possível recorrência do AVC. Nesse contexto, este capítulo apresenta uma revisão bibliográfica sobre pesquisas com foco na utilização de técnicas de aprendizagem de máquina com o objetivo de desenvolver um método que tenha um bom desempenho e que busque colaborar na prevenção da recorrência do AVC. Em particular, é dada uma atenção especial aos trabalhos publicados entre os anos de 2015 e 2022. A pesquisa bibliográfica utilizou as bases dos Periódicos Capes, IEEE e Google Scholar. As palavras utilizadas na pesquisa foram: "*machine learnig*", "*stroke*", "*artificial neural network*", "*recurrence*", "aprendizagem de maquina", "AVC", "recorrência", e "redes neurais artificiais". Dentro da pesquisa realizada vários trabalhos foram apresentados, apenas quatro pesquisas se destacaram e ambas foram realizadas em países diferentes, onde dois trabalhos são de origem chinesa.

A metodologia apresentada por todos os trabalhos possuem diferenças na escolha dos algoritmos, na forma de configuração como eles estão sendo utilizados, e também na forma de seleção dos atributos, onde alguns dos trabalhos utilizam todos os atributos da base estudada, outros fazem seleção por correlação e outros fazem de forma mesclada dependendo da técnica de aprendizagem de máquina utilizada. Outra diferença encontrada nos trabalhos são que as bases de dados utilizadas em cada trabalho possuem as suas particularidades devido as suas regionalidades e também nos atributos das bases, tanto na quantidade como no tipo. Uma observação relevante é que alguns dos trabalhos estudam o AVC sem diferenciar o tipo e outros trabalhos enfatizam o tipo de AVC estudado.

4.1 Apresentação dos trabalhos

O trabalho Abedi *et al.* (2021) foi baseado em dados extraídos de um sistema de banco de dados chamado Geisinger Health da região da Pensilvânia nos Estados Unidos. A base de dados incluía dados demográficos, clínicos e laboratoriais de pacientes com acidente vascular cerebral isquêmico de 2003 a 2019, ao total a base possuía 2091 registros e 53 atributos.

Abedi (2021) usou uma metodologia que dividiu os dados em grupos por período de tempo entre o primeiro evento e a recorrência do AVC. O estudo organizou os pacientes em um

total de seis grupos. Um dos grupos possuía pacientes que não tiveram a recorrência do AVC nos últimos cinco anos e os outros grupos nomeados em 1, 2, 3, 4 e 5 eram compostos por pacientes que tiveram a recorrência do AVC entre 1, 2, 3, 4 e 5 anos.

O durante desenvolvimento do trabalho o Abedi (2021) utilizou a seleção de recursos em diferentes estratégias em cada conjunto. A seleção de recursos foi realizada com base em três abordagens orientadas aos dados para cada conjunto de caso de controle. A seleção utilizou alguns filtros incluindo correlação de Pearson e filtragem univariada, ou seja, examinando uma variável por vez. Os conjuntos apresentaram classes desbalanceadas onde foi necessário utilizar a combinação de técnicas de *oversampling* e *undersampling*.

Todos os grupos foram treinados e testados por seis técnicas de aprendizagem de máquina escolhidas pelo autor, *Extreme Gradient Boosting*, *Gradient Boosting Machine*, *Random Forest*, *Support Vector Machine* e *Decision Tree*. O resultado de cada grupo aplicado a cada técnica foi comparado aos resultados de todos os grupos treinados e testados utilizando o método Regressão Logística. Com o resultado o autor afirmou que todas as técnicas alcançaram um índice satisfatório capaz de prever a recorrência do AVC, mas o melhor resultado foi obtido pela técnica *Random Forest* aplicada no grupo com pacientes que tiveram a recorrência após um ano. O resultado foi medido pela métrica AUC de ROC e alcançou um valor de 0,79.

O trabalho de Vodencarevic *et al.* (2022) apresentou modelos para a previsão da recorrência do AVC no prazo de um ano. O modelo foi desenvolvido utilizando uma base de registro populacional contínuo de pacientes com foco em acidente vascular cerebral. Os dados analisados foram no total de 384 pacientes do Erlangen Stroke Registry da Alemanha. Os pacientes apresentados foram acompanhados por profissionais especializados em três meses e doze meses após o primeiro evento do AVC e depois anualmente se estendendo até dois anos. Os dados continham 250 variáveis, incluindo dados demográficos, comorbidades, interações com o sistema de saúde, e parâmetros clínicos limitados coletados durante entrevistas com pacientes ou seus representantes. Especialistas em dois cursos de medicina selecionaram dentre as 250 apenas 93 potenciais preditores com base nos indicadores de qualidade para *Stroke Care* do *German Stroke Registers Study*.

Segundo Vodencarevic *et al.* (2022) a construção do modelo apresentou três grandes desafios, o desequilíbrio de classe, dados ausentes e maldição de dimensionalidade. Para lidar com os desafios foram implantadas a técnica de SMOTE (*undersampling*), para lidar com os dados ausentes foi aplicado modelos de regressão linear para estimar os valores ausentes. Um

conjunto de 25 técnicas de aprendizagem de máquina foram utilizados onde todos possuíam hiperparâmetros específicos configurados.

O desenvolvimento e avaliação do modelo como o resultado era binário, a tarefa de previsão foi tratada como um problema de classificação binária com as seguintes classes: recorrência e nenhuma recorrência dentro de um ano a partir da data inicial do acompanhamento. A métrica de desempenho do algoritmo selecionado pelo o autor para a melhor classificação foi o AUC de ROC.

Em resumo, o autor concluiu que uma predição individual da recorrência do AVC dentro de um ano após a entrevista inicial do paciente é viável e que o modelo desenvolvido é dinâmico e aplicável a qualquer momento em que os dados do paciente se tornam disponíveis, e não apenas no momento inicial do acidente vascular cerebral. O melhor desempenho de previsão foi alcançado pelo algoritmo SVM linear alcançando o valor de 0,70 de AUC de ROC.

Chang *et al.* (2019) apresenta um modelo de predição da recorrência do AVC baseado na técnica de SVM. Ele justifica a utilização do método SVM por ser bastante usado em previsão de doenças por reconhecimento de padrões e previsão por regressão. Os dados para treino e teste foram coletados de 198 prontuários de um hospital regional na cidade de Pequim na China.

No desenvolvimento metodológico do trabalho o autor utilizou um método de processamento de dados para tratamento e ausência dos dados. Foi realizada outra abordagem que se baseava na idade para diminuir a quantidade de *features* já que a quantidade era relativamente grande, logo em seguida uma outra ação foi aplicada baseada em correlação, onde reduziu ainda mais as *features* resultando em apenas quatro.

Segundo o autor as variáveis de entrada fortemente correlacionadas trariam efeitos adversos, como redundância do modelo e a diminuição da precisão da previsão, motivada pela existência de uma forte correlação linear entre a saúde médica individual e as variáveis, e por isso era necessário reduzir a dimensão das variáveis de saúde médica. E por ultimo foi utilizado um método de Recursive Feature Elimination (RFE) para identificar qual a entrada ideal de variáveis para o modelo.

O modelo de predição de Chang *et al.* (2019) foi aplicado à técnica de SVM por regressão e por classificação. A técnica de SVM para classificação foi utilizado três atributos de entrada onde foram obtidos os seguintes resultados obtendo o resultado de recall de 71,83%, precisão de 81,22% e acurácia de 76% e a técnica de SVM por regressão foram obtidos os seguintes resultados recall de 68,28%, precisão de 81,78% e acurácia de 75,19%.

Já o trabalho de Lu e Wang (2022) apresenta um modelo para previsão da recorrência do AVC tipo Isquêmico para efeito da educação de saúde individual utilizando as técnicas de aprendizagem de máquina Back-propagation neural network (BPNN) e Regressão Logística. O conjunto de dados utilizado no trabalho não é informado de onde foi retirado, mas o autor informa que os dados podem ser disponibilizados através de solicitação e uma análise para liberação. A base no seu total constava com 700 registros e 16 atributos.

A metodologia utilizada por Lu e Wang (2022) teve duas abordagens, onde a primeira utilizou como método para teste e treino a técnica de Regressão Logística. Nesta abordagem foi realizada a seleção de atributos com maior peso de influencia, o qual no total foi selecionado apenas sete atributos. A segunda abordagem foi utilizando o método o BPNN que uma rede neural artificial de multicamada utilizando um algoritmo de aprendizado supervisionado, nesta abordagem foi selecionada todos os atributos sem seleção por peso.

O autor apresentou na conclusão do trabalho a comparação nos resultados das duas abordagens apresentando as vantagens de cada. As vantagens do modelo de regressão logística foram à simplicidade e facilidade em usar, a interpretação quantitativa dos efeitos individuais dos fatores, pode ser obtida diretamente a estimativa aproximada do risco relativo e pode ser estabelecida a metodologia da dependência quantitativa de variáveis. O modelo BPNN utilizou a informação, juntamente com um modo de pensamento humano, para desenvolver a rede aprendendo exemplos existentes. Possuindo uma forte capacidade de resolver o efeito colinear e interação entre variáveis, não possuindo restrições na distribuição de dados e pode fazer pleno uso das informações de dados, com forte tolerância a falhas.

Os resultados experimentais foram apresentados com as seguintes métricas e seus respectivos valores, a área sob a curva ROC do modelo de predição em regressão logística foi de 0,735, a taxa de precisão da previsão é 82,5% e a sensibilidade e especificidade são de 63,2% e 75,6%, respectivamente. Os resultados obtidos em BPNN foram à área sob a curva ROC de 0,796. A predição foi de 86,8%, a sensibilidade de 81,2%, a especificidade de 82,1%. Lu (2022) considerou os resultados obtidos em BPNN melhor que o obtido por Regressão Logística em todas as métricas utilizadas.

4.2 Comparativo dos Trabalhos Relacionados

A tabela 2 apresenta os trabalhos de forma categórica e mostrando os tipos de técnicas aprendizagem de máquina utilizadas em cada trabalho, a quantidade de registros nas

bases, a quantidade de atributos, se foi utilizado técnicas de *undersampling* e *oversampling* e as métricas escolhidas para os resultados.

Tabela 2 – Apresentação dos trabalhos relacionados

Trabalho	País	Registros	Atributos	Balanceamento	Algoritmos	Métricas
Abedi (2021)	EUA	2091	53	U ¹ e O ²	Gradient boost classifier, XGBoost, Random Forest, Árvores de decisão, SVM e Logistic Regression	AUC de ROC
Vondencarevic (2022)	Alemanha	384	250	U e O	25 técnicas ³	AUC de ROC
Chang (2019)	China	117	19	-	SVM classifier e SVM regression	Recall, precisão, acurácia
Lu (2022)	China	700	16	-	BPNN e Logistic regression	AUC de ROC, precisão, sensibilidade e especificidade

Fonte: Elaborado pelo autor

A tabela 3 apresenta os trabalhos para mostrar o melhor método de aprendizagem de máquina utilizado em cada trabalho, por métrica utilizada para performance e os seus respectivos valores.

¹ *Undersampling*

² *Oversampling*

³ Most-frequent dummy, Logistic regression, Naive Bayes, Linear SVM, Ridge classifier, Linear discriminant analysis, Decision tree, KNN, Nonlinear SVM, MLP, Gaussian process classifier, Random forest, Extra tree, AdaBoost, XGBoost, Stacking metaclassifier, Voting classifier, SGB classifier, Elliptic envelope, One-class SVM, Isolation forest, Balanced bagging, Balanced random forest, Easy ensemble e RUSBoost

Tabela 3 – Comparação dos trabalhos relacionados por métodos, métricas e resultados

Trabalho	País	Método com melhor performance	Métricas	Resultados
Abedi (2021)	EUA	Random Forest	AUC de ROC	0,79
Vondencarevic (2022)	Alemanha	SVM	AUC de ROC	0,70
Chang (2019)	China	SVM (classificação)	<i>Recall</i> , precisão, acurácia	71,83%, 81,22%, 76%
Lu (2022)	China	BPNN	AUC de ROC, precisão, sensibilidade, especificidade	0,796, 86,8%, 81,2%, 82,1%

Fonte: Elaborado pelo autor

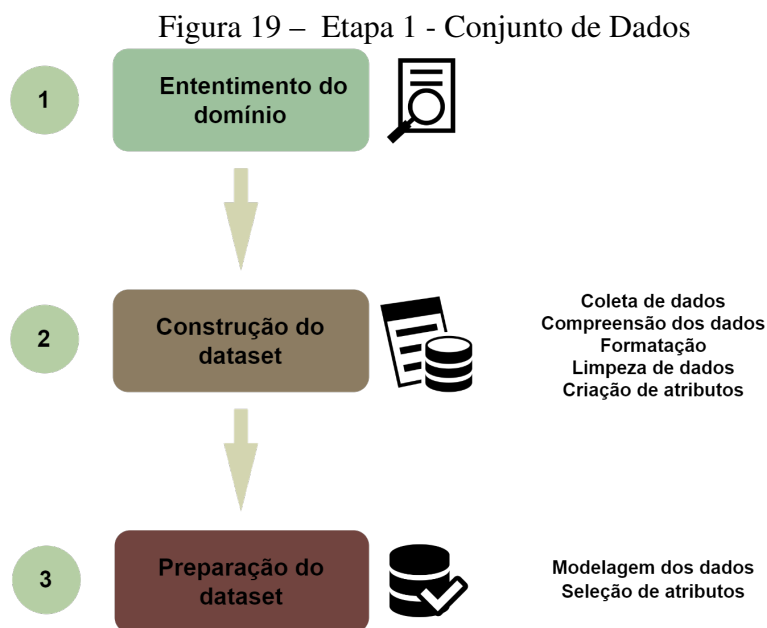
O AprAVC apresentará uma metodologia inovadora em relação aos estudos já realizados sobre o tema. A pesquisa utilizará uma nova base regional de dados do estado do Ceará e empregará técnicas de aprendizado de máquina supervisionado, que permitirão a categorização por tipo e a classificação das informações coletadas. Após a apuração dos resultados, será realizada uma comparação entre as técnicas utilizadas de acordo com a metodologia determinando qual delas obteve melhor desempenho.

Com isso, espera-se que o AprAVC traga novas perspectivas para a compreensão e prevenção do AVC, fornecendo informações mais precisas e confiáveis sobre a doença. A utilização de técnicas de aprendizado de máquina pode contribuir significativamente para aprimorar a eficácia dos tratamentos e intervenções, reduzindo o impacto do AVC na saúde pública.

5 METODOLOGIA

No presente capítulo, será exposta a metodologia que será utilizada para a execução do projeto. Para isso, ela será dividida em duas etapas distintas: a primeira etapa, que será nomeada de "Conjunto de Dados", e a segunda etapa, denominada de "Execução e aprimoramento do AprAVC".

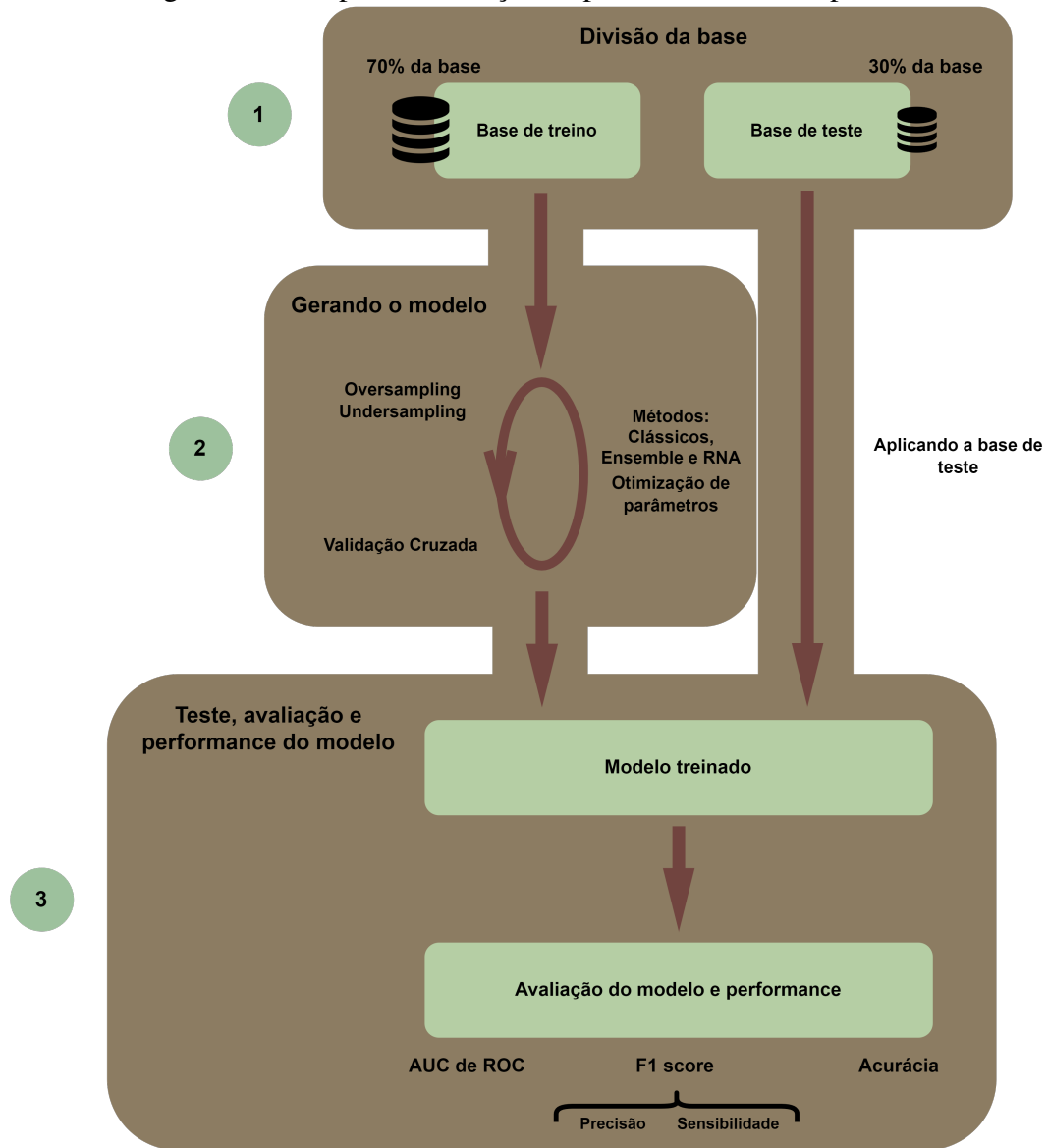
A primeira etapa será dividida em três fases, cada uma delas com um objetivo específico. A primeira fase consistirá no entendimento do domínio em questão, a fim de que se possa compreender de maneira mais profunda as particularidades e nuances da área em que se está trabalhando. Em seguida, será realizada a construção do *dataset*, que será a base para a execução do projeto, como pode ser observado na figura 19. Por fim, a preparação do *dataset* será realizada, de modo que o conjunto de dados esteja adequado para ser utilizado na próxima etapa (Execução e aprimoramento do AprAVC).



Fonte: Elaborado pelo autor

A segunda etapa do projeto é fundamental para que se possa alcançar os objetivos propostos. Ela consistirá em três fases distintas, cada uma delas desempenhando um papel importante na execução e no aprimoramento do sistema conforme apresentado na figura 20.

Figura 20 – Etapa 2 - Execução e aprimoramento do AprAVC



Fonte: Elaborado pelo autor

A primeira fase será a divisão do *dataset*, que será realizada a fim de separar as informações disponíveis em conjuntos de treinamento e validação. Essa divisão é fundamental para que se possa avaliar a eficácia do modelo que será gerado na próxima fase. A segunda fase será a geração do modelo em si, que será desenvolvido utilizando técnicas de aprendizado de máquina. Essa fase é crucial para que se possa criar um modelo capaz de analisar os dados e identificar possíveis casos da recorrência do AVC.

Após a geração e treinamento do modelo, a terceira fase do processo será iniciada. Nessa etapa, os dados de teste, separados na primeira fase, serão utilizados para avaliar a capacidade do modelo na detecção de casos de recorrência do AVC. Em seguida, será realizada a avaliação da performance do modelo, a fim de verificar sua eficácia e precisão.

5.1 Conjunto de Dados

5.1.1 Entendimento do domínio

Nesta etapa, é importante o entendimento do domínio e a definição dos objetivos a serem alcançados. Além disso, também entender o que deve ser feito nas fases seguintes ao processo de descoberta de conhecimento: construção do *dataset* e preparação do *dataset*.

Na primeira etapa do desenvolvimento deste trabalho, é fundamental considerar os estudos realizados para a construção da Motivação (Seção 1.2). Além disso, é preciso levar em conta as pesquisas realizadas para a construção da Fundamentação Teórica (Capítulos 2 e 3) e a revisão bibliográfica sobre a recorrência do AVC (Capítulo 4). Os dados utilizados como fonte para este trabalho foram obtidos a partir da Liga Fisioneuro da UFC¹.

5.1.2 Construção do dataset

A segunda etapa do processo consiste na coleta dos dados, compreensão dos dados, formatação, limpeza dos dados e criação de atributos.

Coleta de dados

A primeira base de dados foi criada por profissionais do curso de fisioterapia da UFC, associados à Liga Fisioneuro. Esses profissionais acompanharam pacientes que sofreram um AVC, com o objetivo de monitorar o seu estado evolutivo ao longo dos meses, em relação às sequelas pós-AVC e à prevenção da recorrência. Todas as informações coletadas foram registradas em formulários de papel preenchidos à mão.

A segunda base de dados foi obtida por meio de uma pesquisa de campo conduzida por profissionais da Liga Fisioneuro, com o objetivo de monitorar a evolução de pacientes que sofreram AVC e foram atendidos pela primeira vez no Hospital Geral de Fortaleza. Além disso, a pesquisa buscou prevenir a recorrência do AVC nos pacientes estudados. Os dados foram coletados em planilhas, porém ainda não haviam sido tabulados e padronizados.

A terceira base foi criada por meio de um formulário digital acompanhado e analisado por especialistas da Liga Fisioneuro. O formulário tinha perguntas objetivas e subjetivas, e só podia ser preenchido por pessoas nascidas e residentes no estado do Ceará.

¹ <https://defisio.ufc.br/pt/extensao/neurofuncional/liga-fisioneuro/>

Compreensão dos dados

As três bases constavam com dados relacionados a fatores de riscos modificáveis (hipertensão, problema cardíaco, diabetes, depressão, etilista, fumante, peso) e não modificáveis (idade, sexo, estado civil, tipo de residência, altura, tipo de trabalho, primeiro evento do AVC, recorrência do AVC).

As duas bases fornecidas pela Liga Fisioneuro não possuíam dados registrados de pessoas que não tiveram AVC, a ausência deste tipo classe poderia acarretar no desempenho e consequentemente impedir que os objetivos do AprAVC fossem alcançados. Por isso, foi necessário a formação da base de dados digital que tivesse uma outra classe que representaria as pessoas que não tinham passado pelo evento do AVC.

Realizamos uma revisão dos dados categóricos e numéricos presentes nas três bases, comparando seus atributos. O objetivo dessa revisão foi explorar as bases e identificar diferenças entre os atributos, além de verificar se algum atributo presente em uma das bases estava faltando nas outras.

Formatação

Nesta etapa, foi necessário converter a primeira base de dados da Liga Fisioneuro do formato de papel para o formato digital de linhas por colunas. Esse processo foi bastante desafiador, uma vez que alguns dados não estavam organizados de forma clara e precisavam ser lidos com atenção. Além disso, algumas comorbidades não estavam identificadas diretamente, mas foram recuperadas a partir da prescrição de medicamentos nos formulários dos pacientes. A conversão de dados de papel para o formato digital foi uma tarefa bem complexa, que exigiu muita atenção e cuidado para garantir a veracidade e confiabilidade dos dados.

A segunda base fornecida pela Liga Fisioneuro já estava em formato digital de linhas por colunas, mas ainda precisava ser organizada para seguir a mesma estrutura da primeira base, sendo exigido muito tempo e esforço para a padronização. A terceira base foi preparada com base nos atributos das bases anteriores, para facilitar a conversão dos dados para o formato padrão.

Após a padronização e organização das três bases de dados, elas foram concatenadas para formar uma única base, contendo um total de 713 registros de pessoas nascidas e residentes no estado do Ceará.

Limpeza de dados

A limpeza de dados é um processo crucial que tem como objetivo remover ou corrigir informações inválidas, inconsistentes ou ausentes. Isso pode envolver diversas tarefas, como identificação e tratamento de valores ausentes, correção de dados duplicados e remoção de *outliers*. A qualidade dos dados é fundamental para garantir resultados precisos e confiáveis em análises futuras (CASTRO; FERRARI, 2016).

Durante a execução do processo de limpeza de dados, foi identificado que nem todos os atributos do conjunto original eram relevantes para a previsão de recorrência do AVC. Por exemplo, os atributos "id" e "nome" foram considerados irrelevantes e removidos da base de dados. Além disso, foram removidos registros duplicados e registros sem informações preenchidas em nenhuma das classes.

Para tratar os campos vazios nos atributos de comorbidades, como diabetes e depressão/ansiedade, optou-se por adicionar o valor 0 para todos os registros em que o campo estava vazio. Essa decisão foi tomada com base no conhecimento dos profissionais envolvidos, que indicaram que a ausência de preenchimento significava a ausência da respectiva comorbidade. Vale ressaltar que todos os outros atributos relevantes, como idade, sexo, hipertensão, problemas cardíacos, estado civil, tipo de trabalho, tipo de residência, status de tabaco, status de álcool, avc1 e avc2 estavam preenchidos corretamente.

Criação de atributos

Para padronizar os atributos, utilizou-se de algumas técnicas na etapa de criação de atributos. Primeiramente, os fatores de risco modificáveis foram marcados com valor 1 em caso positivo e com valor 0 em caso negativo, enquanto os fatores de risco não modificáveis e comportamentais foram padronizados seguindo as regras de escrita de letras minúsculas.

Além disso, foi realizada a codificação de variáveis categóricas em dados numéricos, a fim de que algoritmos de aprendizado de máquina e técnicas de análise pudessem trabalhar com esses dados. Para isso, foi utilizada a técnica *label encoder*, que converte os valores categóricos em ordem crescente em cada atributo em dados numéricos de 0 a 1. Por exemplo, na coluna sexo, os valores "masculino" e "feminino" foram codificados como 0 e 1, respectivamente. Outro exemplo foi a coluna "estado civil", os valores "casado", "divorciado", "outro", "solteiro" e "viúvo" foram codificados como 0, 1, 2, 3, 4, respectivamente.

Ainda nessa etapa, foi criado um novo atributo a partir dos atributos peso e altura,

através de uma operação matemática utilizando a fórmula $IMC = \frac{peso}{altura^2}$, gerando o atributo Índice de Massa Corporal (IMC).

Com essa transformação, a base de dados ficou ajustada e pronta para a criação de um modelo robusto. O *dataset* final continha 713 registros, com 12 atributos (*features*) e um atributo *target* indicando se houve reincidência do AVC. Na tabela 4 estão apresentados os atributos e suas respectivas descrições, após a conclusão desta etapa.

Tabela 4 – Descrição dos atributos.

Atributo	Descrição do atributo
Sexo	Identifica o sexo
Idade	Identifica a idade do paciente
Hipertensão	Identificada se a pessoa possui hipertensão
Problema cardíaco	Identifica se a pessoa possui problema cardíaco
Estado civil	Identifica o estado civil da pessoa
Tipo de trabalho	Identifica o tipo de trabalho da pessoa
Tipo de residência	Identifica o tipo local onde a pessoa mora
IMC	Identifica o índice de massa corpórea
Status para tabaco	Identifica se a pessoa é fumante, não fumante e nunca fumou
Status para álcool	Identifica se a pessoa é etilista, não etilista e nunca foi etilista
Depressão/Ansiedade	Identifica se a pessoa tem depressão ou ansiedade
Stroke1	Identifica se a pessoa já passou por um primeiro evento de AVC
Stroke2	Identifica se a pessoa já passou por uma recorrência do AVC

Fonte: Elaborado pelo autor

5.1.3 Preparação do dataset

A terceira etapa do processo consiste na modelagem dos dados e a seleção de atributos.

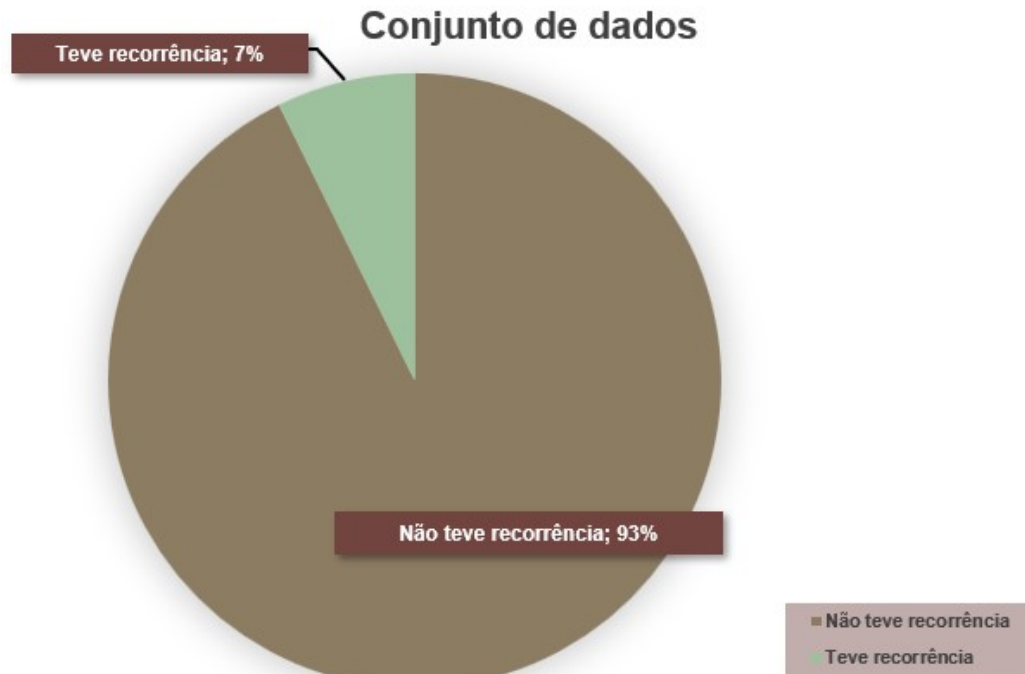
Modelagem dos dados

Durante esta etapa do trabalho, utilizou-se a Análise Exploratória de Dados (AED) para compreender as informações contidas no conjunto de dados e explorar suas propriedades. A técnica foi aplicada com o objetivo de criar novas hipóteses, encontrar padrões nos dados e identificar os atributos mais influentes na recorrência do AVC.

Para isso, realizou-se uma comparação entre os atributos e sua relação com a coluna alvo, ou seja, a recorrência do AVC, por meio da análise de gráficos sintéticos. Assim, ficou evidente quais atributos possuem maior peso de influência no acontecimento da recorrência do AVC. Durante essa etapa, também se observou um fator muito importante: o desbalanceamento

de cargas no atributo alvo. A figura 21 apresenta o desbalanceamento em formato de gráfico para uma melhor visualização e compreensão.

Figura 21 – Base de dados desbalanceada.



Fonte: Elaborado pelo autor

Outra observação compreendida foi a necessidade da normalização de alguns atributos, cujo os valores excediam de 1. Para realizar a normalização, utilizou-se a técnica de "min-max", que ajustou os valores dos atributos para uma escala entre 0 e 1, já que não existem valores negativos. Os atributos que receberam a aplicação da técnica foram a idade, estado civil, tipo de trabalho, IMC, status para tabaco e status para álcool.

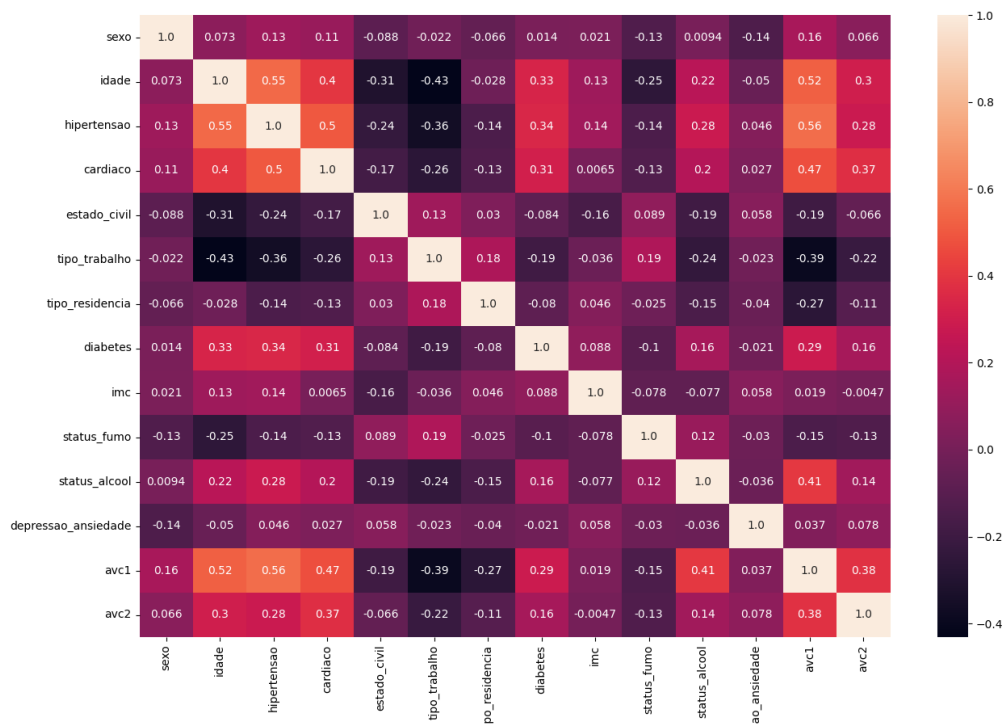
Seleção dos atributos

Nesta etapa foram utilizados dois métodos para verificar a importância dos atributos: a matriz de correlação utilizando a correlação de Pearson e a técnica de seleção univariada de atributos.

A matriz de correlação mostrou como as variáveis estavam correlacionadas entre si, através da visualização gráfica por meio de um mapa de calor. Se o coeficiente é próximo de 1, significa que as variáveis estão fortemente correlacionadas e aumentam ou diminuem juntas. Se o coeficiente é próximo de 0, significa que as variáveis não estão correlacionadas. Se o coeficiente é próximo de -1, significa que as variáveis estão fortemente correlacionadas, mas

aumentam e diminuem de forma oposta. Com o auxílio da matriz de correlação, foi possível identificar as *features* com maior relação com o atributo alvo. Na figura 22 é apresentado a matriz de correlação de todos os atributos do conjunto de dados.

Figura 22 – Matriz de correlação dos atributos.



Fonte: Elaborado pelo autor

Em seguida, utilizou-se o método de seleção univariada de atributos, por meio da técnica "*SelectKBest*", para selecionar os atributos mais relevantes na construção do modelo. Essa técnica utiliza diversos testes estatísticos para avaliar a importância dos atributos. Os atributos receberam um valor em *score*, onde quanto maior o *score*, maior é a relação significativa com a variável alvo.

Conforme destacado por Kelleher e Tierney (2018), valores abaixo de 10 *scores* indicam uma correlação fraca entre o atributo e a variável alvo. Isso significa que a presença ou ausência desse atributo tem pouca influência na variação da variável alvo e, portanto, esses atributos podem ser considerados irrelevantes para o modelo preditivo. Nesse sentido, é possível removê-los sem afetar significativamente o desempenho do modelo.

Foram realizados vários testes modificando os itens correlacionados e observou-se que valores abaixo de 10 *scores* não tiveram um grande impacto, ou seja, o relevante são os classificadores que possuem valor de correlação acima de 10 *scores* conforme destacado por

Kelleher e Tierney (2018). Abaixo na tabela 5 é apresentado os atributos com os valores de correlação por *scores* com a recorrência do AVC.

Tabela 5 – Seleção de atributos por *score*

Atributo	Score
avc1	123,021295
cardiaco	113,139920
idade	71,552552
hipertensao	61,360395
tipo_trabalho	34,720504
diabetes	18,645097
status_alcool	14,394291
status_fumo	12,297098
tipo_residencia	8,444451
depressao_ansiedade	4,330743
estado_civil	3,126482
sexo	3,101842
imc	0,015788

Fonte: Elaborado pelo autor

5.2 Execução e aprimoramento do AprAVC

5.2.1 Divisão da base

Para este trabalho, adotou-se a proporção de 70% para treinamento e 30% para teste da base de dados, uma prática comum em aprendizado de máquina e análise de dados. Essa divisão tem o propósito de avaliar a capacidade do modelo em generalizar os resultados para dados não observados anteriormente.

O conjunto de treinamento, correspondente a 70% dos dados, será utilizado para ajustar os parâmetros do modelo e aprender as relações entre as variáveis. Essa etapa será fundamental para garantir que o modelo seja capaz de generalizar os resultados para novos dados.

Por outro lado, o conjunto de teste, que contém 30% dos dados, será reservado para avaliar o desempenho do modelo. Esses dados não são utilizados durante o treinamento, e a finalidade é verificar se o modelo pode generalizar os resultados para dados não vistos anteriormente.

5.2.2 Geração do modelo

O processo de geração do modelo foi dividido em três fases. Na primeira fase, foram definidas as técnicas de *oversampling* e *undersampling* que seriam utilizadas. Em seguida foi selecionado os algoritmos de aprendizagem e definimos os parâmetros do modelo. Depois disso, treinamos o modelo utilizando os dados de treinamento.

Durante o processo de treinamento, o modelo foi exposto a diversos exemplos de dados e os parâmetros foram ajustados para minimizar a diferença entre as previsões do modelo e os resultados reais. Utilizamos a técnica de validação cruzada para avaliar o desempenho do modelo em diferentes conjuntos de dados. Com isso, conseguimos otimizar o modelo para obter previsões mais precisas.

Na terceira e última fase, apresentamos o modelo que foi treinado e pronto para ser utilizado na base de teste, a fim de avaliar sua capacidade de generalizar para dados desconhecidos. Essa avaliação foi feita por meio do cálculo de métricas de desempenho, que fornecem informações sobre a qualidade das previsões do modelo.

Balanceamento de cargas

Como apresentado na modelagem dos dados por meio da técnica de AED, o conjunto de dados possuía uma classe maior que a outra, ou seja, classe alvo da recorrência do AVC tinha um número muito limitado de dados totalizando 7% do total de toda a base, enquanto os outros 93% são de pessoas que não tiveram a recorrência do AVC.

Para evitar sérios problemas, motivado pela a base de dados desbalanceada, foi necessária uma ação utilizando as técnicas de *oversampling* e *undersampling* para ajudar na distribuição de classes no conjunto de dados.

O método *oversampling* utilizado foi o *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE é uma técnica estatística para aumentar o número de casos em seu conjunto de dados de maneira equilibrada. O componente funciona gerando novas instâncias de casos minoritários existentes que você fornece como entrada.

Inicialmente, utilizou-se apenas a técnica de *Random Under Sampler* no método de *undersampling*. No entanto, essa abordagem resultou em uma perda significativa de informações. Para remediar esse problema, optou-se por adotar uma fusão de métodos, que consistiu na união de duas ou mais observações das classes majoritárias em apenas uma. Essa técnica se mostrou eficaz na redução da perda de informações.

A primeira técnica utilizada foi o *Random Under Sampler*, que consiste em selecionar aleatoriamente um conjunto de exemplos da classe majoritária e removê-los do conjunto de dados até que a distribuição das classes esteja equilibrada. O tamanho do conjunto de dados resultante é determinado pelo número de exemplos da classe minoritária e pela proporção desejada entre as classes. A principal vantagem do *Random Under Sampler* é a sua simplicidade e eficácia em lidar com conjuntos de dados desbalanceados.

A segunda técnica utilizada foi *Near Miss* versão 1, essa técnica consiste em selecionar exemplos da classe majoritária que estão "próximos" dos exemplos da classe minoritária, com base na distância Euclidiana entre os exemplos, e remover esses exemplos selecionados do conjunto de dados, a fim de equilibrar a distribuição das classes.

A terceira técnica foi a *One Sided Selection*, essa técnica consiste em selecionar exemplos da classe majoritária que estão mais próximos dos exemplos da classe minoritária e remover os exemplos da classe majoritária que estão distantes dos exemplos da classe minoritária, a fim de equilibrar a distribuição das classes. A técnica *One Sided Selection* possui duas etapas:

- Seleção inicial: nesta etapa, a técnica seleciona aleatoriamente um subconjunto dos exemplos da classe minoritária. Em seguida, a técnica utiliza esses exemplos selecionados para selecionar exemplos da classe majoritária que estão próximos aos exemplos da classe minoritária.
- Seleção final: nesta etapa, a técnica seleciona exemplos da classe majoritária que estão próximos dos exemplos selecionados na etapa anterior. Os exemplos da classe majoritária que estão distantes dos exemplos selecionados na etapa anterior são removidos do conjunto de dados.

A principal vantagem da técnica *One Sided Selection* é que ela pode preservar as informações importantes contidas nos exemplos da classe majoritária, enquanto ainda equilibra a distribuição das classes.

Para a construção do AprAVC, foram utilizadas quatro abordagens diferentes com o objetivo de identificar qual delas resultaria em uma melhor predição da recorrência do AVC, quando aplicadas em conjunto com métodos de aprendizagem de máquina. As abordagens utilizadas foram as seguintes:

- Aplicação do método de Balanceamento de Cargas com *oversampling* (BCO);
- Aplicação do método de Balanceamento de Cargas com *undersampling* (BCU);
- Aplicação dos dois métodos, Balanceamento de Cargas com *oversampling* mais *undersampling*

(BCOU) na respectiva ordem;

- Aplicação dos dois métodos, Balanceamento de Cargas com *undersampling* mais *oversampling* (BCUO) na respectiva ordem.

Para lidar com o desbalanceamento dos dados e determinar a abordagem mais eficaz para a predição da recorrência do AVC, foram selecionadas diversas estratégias. O objetivo foi explorar diferentes formas de abordagem e determinar o método que mais se generalize para cada uma delas.

Nesse sentido, foram escolhidas quatro abordagens distintas, cada uma com seus respectivos métodos, a fim de avaliar o desempenho de cada um na tarefa de predição da recorrência do AVC.

O propósito final é identificar a abordagem mais adequada para o contexto do estudo, considerando a performance de cada método e as particularidades do problema. Além disso, será avaliado se todas as abordagens são adequadas para a tarefa de predição da recorrência do AVC.

Métodos, estimação de parâmetros e validação do modelo

Foram utilizados dez métodos de aprendizado de máquina para desenvolver modelos preditivos de aprendizado supervisionado. Na Tabela 6, é possível verificar os algoritmos utilizados e classificados por tipo.

Tabela 6 – Aprendizagem de máquina

Tipo	Algoritmo
Aprendizado Clássico	Decision Tree KNN Logistic Regression Naive Bayes SVM
Metodo Ensemble	CatBoost LightGBM Random Forest XGBoost
Redes Neurais	MLP

Fonte: Elaborado pelo autor

Um processo preditivo de dados ocorre por meio do ajuste de parâmetros do modelo escolhido. Desta forma, quando os algoritmos de aprendizagem de máquina são usados para esse fim, o processo é denominado de aprendizado indutivo ou treinamento (aprendizagem supervisionada). Após a etapa de treinamento, o modelo é aplicado em novos exemplos (fase de teste), a partir da qual o modelo é avaliado quanto à sua capacidade preditiva.

Já o treinamento foi o processo onde foi ajustado os parâmetros do modelo aos dados de treinamento, o modelo aprendeu a realizar previsões de forma mais precisa. O treinamento também proporcionou encontrar os melhores parâmetros para o AprAVC, tornando o AprAVC pronto para ser executado consumindo base de teste e também estando apto a reconhecer os padrões para a recorrência do AVC.

Foram realizadas as escolhas de quais hiperparâmetros deveriam ser utilizados de acordo com cada método de aprendizagem de máquina utilizado no trabalho. Foi levado em consideração também diferentes aspectos na escolha desses hiperparâmetros, como desempenho preditivo, tempo de processamento, sensibilidade, entre outros de forma empírica durante a realização do trabalho.

A tabela 7 apresenta os algoritmos com os principais parâmetros utilizados e também os parâmetros testados com variação. Os demais parâmetros de cada algoritmo são os originalmente disponibilizados no *sklearn*².

Tabela 7 – Parâmetros por algoritmo

Algoritmo	Parâmetro
Decision Tree	criterion = entropy max_depth = 15
KNN	n_neighbors = [33 e 100]
Logistic Regression	random_state = 1 C = 500
Naive Bayes	Não se aplica
SVM	kernel = linear C = 1 probability = True
CatBoost	n_estimators = 1000 learning_rate = 0.1
LightGBM	n_estimators = 500 max_depth = 50
Random Forest	criterion = entropy n_estimators = 100 random_state = 1
XGBoost	n_estimators = 1000 learning_rate = 0.001 max_depth = 20
MLP	learning_rate_init = 0.001 max_iter = 1000 hidden_layer_sizes = 100

Fonte: Elaborado pelo autor

Em relação às variações de parâmetros dos algoritmos da tabela 7, o valor de

² <https://scikit-learn.org/stable/>

$n_neighbors$ do KNN é de 30 quando utilizar apenas a abordagem de balanceamento de cargas com *oversampling* e 100 para as outras três abordagens.

Neste trabalho, utilizamos algoritmos de análise preditiva e aplicamos a técnica de validação cruzada para avaliar a performance do modelo em dados não vistos. Os algoritmos foram treinados e testados 5 vezes para cada uma das 4 abordagens de balanceamento de cargas, usando validação cruzada com 5 *folds* para gerar uma estimativa média de acertos e erros de cada algoritmo. No total, foram realizadas 20 execuções para cada método. Os resultados obtidos demonstraram que o modelo conseguiu identificar corretamente a maioria dos padrões dos dados e apresentou baixa variância, ou seja, não estava captando muito ruído. A tabela 8 apresenta a média de acurácia e o desvio padrão de cada modelo nas abordagens BCO e BCU.

Tabela 8 – Valores da Acurácia para a validação cruzada - BCO e BCU

Categoria do Classificador	Classificador	Acurácia (Média $\pm$ Desvio Padrão)	
		BCO	BCU
Classico	Decision Tree	0,9421 $\pm$ 0,0137	0,8192 $\pm$ 0,0350
	KNN	0,8799 $\pm$ 0,0173	0,8705 $\pm$ 0,0601
	Logistic Regression	0,8864 $\pm$ 0,0085	0,8038 $\pm$ 0,1234
	Naive Bayes	0,8573 $\pm$ 0,0054	0,8859 $\pm$ 0,0379
	SVM	0,8605 $\pm$ 0,0145	0,8859 $\pm$ 0,0379
Ensemble	CatBoost	0,9636 $\pm$ 0,0190	0,8692 $\pm$ 0,0841
	LightGBM	0,9518 $\pm$ 0,0289	0,7872 $\pm$ 0,0652
	Random Forest	0,9582 $\pm$ 0,0204	0,8526 $\pm$ 0,0619
	XGBoost	0,9475 $\pm$ 0,0217	0,8662 $\pm$ 0,0390
Redes Neurais	MLP	0,9313 $\pm$ 0,0077	0,8192 $\pm$ 0,1229

Fonte: Elaborado pelo autor

A tabela 9 apresenta o valor médio da acurácia e o desvio padrão para cada modelo nas abordagens BCOU e BCUC.

Tabela 9 – Valores da Acurácia para a validação cruzada - BCOU e BCUC

Categoria do Classificador	Classificador	Acurácia (Média $\pm$ Desvio Padrão)	
		BCOU	BCUC
Classico	Decision Tree	0,8395 $\pm$ 0,0815	0,8637 $\pm$ 0,0750
	KNN	0,8294 $\pm$ 0,0490	0,8648 $\pm$ 0,0524
	Logistic Regression	0,8350 $\pm$ 0,0858	0,8363 $\pm$ 0,1136
	Naive Bayes	0,8238 $\pm$ 0,0537	0,8945 $\pm$ 0,0352
	SVM	0,8216 $\pm$ 0,0739	0,8945 $\pm$ 0,0352
Ensemble	CatBoost	0,8893 $\pm$ 0,0984	0,8330 $\pm$ 0,1234
	LightGBM	0,8749 $\pm$ 0,0757	0,7725 $\pm$ 0,0491
	Random Forest	0,8782 $\pm$ 0,0948	0,8802 $\pm$ 0,0559
	XGBoost	0,8871 $\pm$ 0,0890	0,8505 $\pm$ 0,0905
Redes Neurais	MLP	0,8760 $\pm$ 0,0692	0,8813 $\pm$ 0,0967

Fonte: Elaborado pelo autor

5.3 Teste, avaliação e performance do modelo

Agora que o modelo está treinado, precisamos medir sua performance ao aplicá-lo em dados não utilizados durante o treinamento. Isso será feito usando o conjunto de dados reservado para testes, que corresponde a 30% do total da base. O objetivo é avaliar a capacidade do modelo de generalizar para dados desconhecidos.

Cada método de aprendizagem de máquina será submetido à base de testes juntamente com as quatro abordagens utilizadas para o balanceamento de cargas apresentadas. Em seguida, serão avaliados utilizando as métricas de avaliação escolhidas. As métricas de avaliação foram selecionadas com base na literatura da área de aprendizado de máquina, bem como nos trabalhos apresentados no Capítulo 4. As métricas escolhidas para avaliar o modelo foram: Matriz de confusão, AUC de ROC, Precisão, *Recall*, *F1-score*, Acurácia e Tempo de Processamento Total (TPT).

A escolha da métrica Acurácia se deve à sua popularidade e facilidade de interpretação. Ela mede a proporção de previsões corretas em relação ao total de previsões, sendo bastante útil em problemas de classificação binária. Além disso, a Precisão foi adotada por sua capacidade de apresentar a proporção de verdadeiros positivos em relação a todos os positivos previstos pelo modelo. Essa métrica é especialmente importante em problemas em que a detecção de falsos positivos pode ter um impacto significativo, como é o caso da recorrência do AVC.

Já a *Recall* foi adicionada por apresentar a proporção de verdadeiros positivos em relação ao total de positivos reais na base de dados. Essa métrica é útil em problemas em que a detecção de falsos negativos pode ter um impacto significativo. O *F1-score* foi adotado como uma média harmônica ponderada da Precisão e da *Recall*, dando igual peso a ambas as métricas. Ela leva em consideração tanto os falsos positivos quanto os falsos negativos, o que a torna útil em problemas com cargas desbalanceadas.

A Matriz de Confusão foi escolhida por ser uma tabela que mostra a frequência de classificações corretas e incorretas do modelo em cada classe. Ela é útil para entender como o modelo está se saindo em termos de FP, FN, VP e VN. Já a AUC de ROC foi selecionada como a métrica principal para este trabalho, devido à sua capacidade de medir a habilidade do modelo em distinguir entre classes positivas e negativas. Isso a torna uma métrica importante em problemas de classificação binária com classes desbalanceadas. No entanto, todas as métricas serão avaliadas em conjunto para fornecer uma visão mais completa da performance do modelo. Os resultados por abordagem e tipo de método serão apresentados no capítulo 6.

6 RESULTADOS E DISCUSSÃO

A seguir, apresentaremos os resultados da análise de desempenho dos classificadores, destacando qual obteve o melhor desempenho no AprAVC por abordagem.

6.1 Resultados - Matriz de confusão

Após a execução dos algoritmos de classificação, foi gerada uma matriz de confusão para cada modelo com o resultado da classificação binária. As colunas representam as classes previstas pelo modelo, enquanto as linhas representam as classes reais de cada amostra. A diagonal principal da matriz mostra o número de classificações corretas para cada classe, ou seja, os VP e VN. Já a diagonal secundária apresenta as classificações incorretas de cada classe, que são os FP e FN. Na tabela 10, é possível observar a representação correta da matriz de confusão."

Tabela 10 – Matriz de confusão para o problema de classificação binário proposto

Classificador	Modelo (Previsto)	
Referência (Real)	Sem recorrência do AVC (0)	Com recorrência do AVC (1)
Sem recorrência do AVC (0)	VN	FP
Com recorrência do AVC (1)	FN	VP

Fonte: Elaborado pelo autor

Na tabela 11, os valores da matriz de confusão para a abordagem BCO.

Tabela 11 – Matriz de confusão para cada classificador - abordagem BCO

Decision Tree	Previsto	
Real	(0)	(1)
(0)	179	16
(1)	44	151

KNN	Previsto	
Real	(0)	(1)
(0)	152	43
(1)	21	174

Logistic Regression	Previsto	
Real	(0)	(1)
(0)	153	42
(1)	12	183

Naive Bayes	Previsto	
Real	(0)	(1)
(0)	134	61
(1)	0	195

SVM	Previsto	
Real	(0)	(1)
(0)	134	61
(1)	0	195

CatBoost	Previsto	
Real	(0)	(1)
(0)	185	10
(1)	51	144

LightGBM	Previsto	
Real	(0)	(1)
(0)	186	9
(1)	55	140

Random Forest	Previsto	
Real	(0)	(1)
(0)	184	11
(1)	46	149

XGBoost	Previsto	
Real	(0)	(1)
(0)	184	11
(1)	30	165

MLP	Previsto	
Real	(0)	(1)
(0)	181	14
(1)	125	70

Fonte: Elaborado pelo autor

Uma análise dessas matrizes pode ser feita tomando como exemplo o classificador *XGBoost* dos 195 casos classificados como "sem recorrência do AVC", o modelo acertou 184 (94%). Para os classificados como "com recorrência do AVC", o modelo acertou 165 (85%).

Na tabela 12, os valores da matriz de confusão para a abordagem BCU.

Tabela 12 – Matriz de confusão para cada classificador - abordagem BCU

Decision Tree	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	3	12

KNN	Previsto	
Real	(0)	(1)
(0)	16	3
(1)	1	14

Logistic Regression	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	3	12

Naive Bayes	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	0	15

SVM	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	0	15

CatBoost	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	1	14

LightGBM	Previsto	
Real	(0)	(1)
(0)	14	5
(1)	1	14

Random Forest	Previsto	
Real	(0)	(1)
(0)	15	4
(1)	0	15

XGBoost	Previsto	
Real	(0)	(1)
(0)	13	6
(1)	1	14

MLP	Previsto	
Real	(0)	(1)
(0)	14	5
(1)	2	13

Fonte: Elaborado pelo autor

Na tabela 13, os valores da matriz de confusão para a abordagem BCOU.

Tabela 13 – Matriz de confusão para cada classificador - abordagem BCOU

Decision Tree	Previsto	
Real	(0)	(1)
(0)	176	19
(1)	53	133

KNN	Previsto	
Real	(0)	(1)
(0)	151	44
(1)	28	158

Logistic Regression	Previsto	
Real	(0)	(1)
(0)	155	40
(1)	17	169

Naive Bayes	Previsto	
Real	(0)	(1)
(0)	134	61
(1)	0	186

SVM	Previsto	
Real	(0)	(1)
(0)	155	40
(1)	17	169

CatBoost	Previsto	
Real	(0)	(1)
(0)	183	12
(1)	44	142

LightGBM	Previsto	
Real	(0)	(1)
(0)	189	6
(1)	47	139

Random Forest	Previsto	
Real	(0)	(1)
(0)	185	10
(1)	50	136

XGBoost	Previsto	
Real	(0)	(1)
(0)	187	8
(1)	43	143

MLP	Previsto	
Real	(0)	(1)
(0)	180	15
(1)	138	48

Fonte: Elaborado pelo autor

Na tabela 14, os valores da matriz de confusão para a abordagem BCUO.

Tabela 14 – Matriz de confusão para cada classificador - abordagem BCUO

Decision Tree		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		5	14

KNN		Previsto	
Real		(0)	(1)
(0)		16	3
(1)		0	13

Logistic Regression		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		3	16

Naive Bayes		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		0	19

SVM		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		0	19

CatBoost		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		2	17

LightGBM		Previsto	
Real		(0)	(1)
(0)		14	5
(1)		2	17

Random Forest		Previsto	
Real		(0)	(1)
(0)		16	3
(1)		1	18

XGBoost		Previsto	
Real		(0)	(1)
(0)		13	6
(1)		2	17

MLP		Previsto	
Real		(0)	(1)
(0)		14	5
(1)		3	16

Fonte: Elaborado pelo autor

6.2 Resultados - Tempo de processamento

Nesta seção, serão apresentados os resultados do tempo de processamento para o treinamento e teste dos classificadores, bem como o TPT, que é a soma do tempo de treinamento e do tempo de teste. Os tempos serão medidos em segundos (*s*) e separados por classificador e tipo de abordagem. O tempo de processamento de treinamento de cada modelo será exibido na tabela 15.

Tabela 15 – Tempo de processamento de treino por classificador e tipo de abordagem

Tempo de processamento - Treinamento				
Classificador / Tipo	BCO	BCU	BCOU	BCUO
Decision Tree	0,3606s	0,1844s	0,2988s	0,3717s
KNN	1,2901s	0,3625s	0,7521s	0,9225s
Logistic Regression	0,6092s	0,3081s	0,6680s	0,8943s
Naive Bayes	0,4364s	0,6845s	0,4633s	0,5488s
SVM	2,2071s	0,3401s	3,3669s	0,3389s
CatBoost	99,5888s	43,9823s	95,4266s	41,8981s
LightGBM	7,6145s	1,0406s	6,5246s	12,1318s
Random Forest	6,0566s	4,8985s	5,7252s	4,8344s
XGBoost	320,0438s	75,8067s	286,3654s	58,3742s
MLP	163,5791s	11,9552s	142,4487s	12,2254s

Fonte: Elaborado pelo autor

O tempo de processamento de teste é referente ao tempo de processamento para consumir a base de teste e apresentar a performance por métrica, a tabela 16 apresenta os devidos valores.

Tabela 16 – Tempo de processamento com a base de teste por classificador e tipo de abordagem

Tempo de processamento - Teste				
Classificador / Tipo	BCO	BCU	BCOU	BCUO
Decision Tree	1,5796s	0,4909s	1,4144s	0,9232s
KNN	1,2974s	0,5434s	1,1116s	1,1687s
Logistic Regression	1,1120s	1,3086s	1,1287s	1,6818s
Naive Bayes	1,2019s	2,0216s	1,0850s	1,3401s
SVM	0,8876s	0,8642s	1,9008s	0,9854s
CatBoost	3,6963s	1,4274s	4,1273s	1,9801s
LightGBM	0,7680s	0,6995s	0,7695s	0,5476s
Random Forest	1,0998s	0,6794s	1,0764s	1,9801s
XGBoost	15,2757s	2,5627s	16,3621s	5,2918s
MLP	5,8533s	0,9448s	6,8992s	0,9544s

Fonte: Elaborado pelo autor

Para o TPT temos a soma do tempo de processamento de teste mais o tempo de processamento de treino. A tabela 17 vai apresentar os valores.

Tabela 17 – Tempo de processamento total por classificador e abordagem

Tempo de processamento total				
Classificador / Tipo	BCO	BCU	BCOU	BCUO
Decision Tree	1,9402s	0,6753s	1,7132s	1,2749s
KNN	2,5875s	0,9059s	1,8637s	2,0912s
Logistic Regression	1,7212s	1,6167s	1,7967s	2,5761s
Naive Bayes	1,6383s	2,7061s	1,5483s	1,8889s
SVM	3,0947s	1,2043s	5,2677s	1,3243s
CatBoost	103,2851s	45,4097s	99,5539s	43,8782s
LightGBM	8,3825s	1,7401s	7,2941s	12,6794s
Random Forest	7,1564s	5,5779s	6,8016s	5,7891s
XGBoost	335,3195s	78,3694s	302,7275s	63,6660s
MLP	169,4324s	12,9s	149,3469s	13,1798s

Fonte: Elaborado pelo autor

6.3 Resultados - Acurácia

Na tabela 18 observamos que utilizando a abordagem CBO o método *Logistic Regression* obteve a melhor acurácia, quando utilizando a abordagem CBU a melhor performance foi o KNN. A abordagem CBOU o melhor resultado foi SVM e a abordagem BCUO o KNN conseguiu a melhor performance.

Tabela 18 – Tabela de resultados (Acurácia) - Métodos Clássicos

Acc					
Métodos Clássicos					
Abordagem	<i>Decision tree</i>	KNN	<i>Logistic Regression</i>	<i>Naive Bayes</i>	SVM
BCO	84,62%	83,59%	86,15%	84,36%	84,36%
BCU	73,83%	88,24%	73,53%	82,35%	82,35%
BCOU	81,10%	81,10%	85,04%	83,99%	83,99%
BCUO	71,05%	92,11%	76,32%	84,21%	84,21%

Fonte: Elaborado pelo autor

Na tabela 19, observa-se que na abordagem CBO, o *XGBoost* obteve o melhor resultado em termos de desempenho, enquanto que na abordagem CBU, o *Random Forest* apresentou a melhor performance. Ao se aplicar a abordagem CBOU, o *XGBoost* também teve o melhor desempenho, enquanto que na abordagem CBUO, o *Random Forest* se destacou como o melhor modelo.

Tabela 19 – Tabela de resultados (Acurácia) - Métodos Ensemble

Acc				
Métodos Ensemble				
Abordagem	<i>CatBoost</i>	<i>LightGBM</i>	<i>Random Forest</i>	<i>XGBoost</i>
BCO	84,36%	83,59%	85,38%	89,49%
BCU	79,41,24%	82,35%	88,24%	79,41%
BCOU	85,30%	86,09%	84,25%	86,61%
BCUO	78,95%	81,58%	89,47%	78,95%

Fonte: Elaborado pelo autor

A tabela 20 podemos visualizar o desempenho do MLP para as quatro abordagens.

Tabela 20 – Tabela de resultados (Acurácia) - Redes Neurais

Acc	
Redes Neurais	
Abordagem	MLP
BCO	64,36%
BCU	79,41%
BCOU	59,84%
BCUO	78,95%

Fonte: Elaborado pelo autor

A tabela 21 é apresentado o resultado final para a métrica acurácia com o método de aprendizagem de máquina que melhor performou com cada tipo de abordagem.

Tabela 21 – Melhor método por abordagem (Acurácia)

Acc		
Resultado final		
Abordagem	Método de ML	Performance
BCO	<i>XGBoost</i>	89,49%
BCU	<i>Random Forest e KNN</i>	88,24%
BCOU	<i>LightGBM</i>	86,09%
BCUO	<i>KNN</i>	92,11%

Fonte: Elaborado pelo autor

6.4 Resultados - Precisão

Na tabela 22 observamos que utilizando a abordagem CBO o método *Decision Tree* atingiu a melhor performance em precisão, quando utilizando a abordagem CBU a melhor performance foi o KNN. A abordagem CBOU o melhor resultado foi alcançado pelo método *Decision Tree* e a abordagem CBUO a melhor performance foi o KNN.

Tabela 22 – Tabela de resultados (Precisão) - Métodos Clássicos

Precisão					
Métodos Clássicos					
Abordagem	<i>Decision Tree</i>	KNN	<i>Logistic Regression</i>	<i>Naive Bayes</i>	SVM
BCO	90,42%	81,23%	81,33%	76,17%	76,17%
BCU	66,67%	82,35%	66,67%	71,43%	71,43%
BCOU	87,50%	78,22%	80,86%	75,30%	80,49%
BCUO	70%	86,36%	72,73%	76%	76%

Fonte: Elaborado pelo autor

A tabela 23 a abordagem BCO o *LightGBM* desempenhou o melhor resultado, quando a abordagem BCU o *Random Forest* apresentou a melhor performance. Aplicando a abordagem BCOU a melhor performance foi o *LightGBM* e a abordagem BCUO o *Random Forest* teve a melhor performance.

Tabela 23 – Tabela de resultados (Precisão) - Métodos Ensemble

Precisão				
Métodos Ensemble				
Abordagem	<i>CatBoost</i>	<i>LightGBM</i>	<i>Random Forest</i>	<i>XGBoost</i>
BCO	93,51%	93,96%	93,12%	93,75%
BCU	70%	73,68%	78,95%	70%
BCOU	92,21%	95,86%	93,15%	94,70%
BCUO	73,91%	77,27%	85,71%	73,91%

Fonte: Elaborado pelo autor

A tabela 24 podemos visualizar o desempenho do MLP para as quatro abordagens.

Tabela 24 – Tabela de resultados (Precisão) - Redes Neurais

Precision	
Redes Neurais	
Tipo de análises	MLP
BCO	83,33%
BCU	72,22%
BCOU	76,19%
BCUO	76,19%

Fonte: Elaborado pelo autor

Fazendo uma consolidação dos resultados utilizando a métrica Precisão, a tabela 25 vai apresentar o método que teve melhor performance diante das quatro abordagens.

Tabela 25 – Melhor método por abordagem Precisão

Precision		
Resultado final		
Abordagem	Método de ML	Performance
BCO	<i>LightGBM</i>	93,96%
CBU	<i>KNN</i>	82,35%
BCOU	<i>LightGBM</i>	95,86%
BCUO	<i>KNN</i>	86,36%

Fonte: Elaborado pelo autor

6.5 Resultados - Recall

Na tabela 26 observamos que utilizando a abordagem CBO os métodos *Naive Bayes* e SVM conseguiram a melhor *recall*, quando utilizando a abordagem CBU as melhores performances ficaram com *Naive Bayes* e SVM. A abordagem CBOU o melhor resultado foi alcançado por *Naive Bayes* e a abordagem BCUC três métodos alcançaram a melhor performance em resultados iguais KNN, *Naive Bayes* e SVM.

Tabela 26 – Tabela de resultados (Recall) - Métodos Clássicos

Recall					
Métodos Clássicos					
Abordagem	Decision tree	KNN	Logistic Regression	Naive Bayes	SVM
BCO	77,44%	80,18%	93,85%	100%	100%
BCU	80%	93,33%	80%	100%	100%
BCOU	71,51%	84,95%	90,86%	100%	88,71%
BCUC	73,68%	100%	84,21%	100%	100%

Fonte: Elaborado pelo autor

Na tabela 27, é possível verificar que na abordagem BCO, o *XGBoost* obteve o

melhor desempenho em termos de *Recall*, enquanto que na abordagem BCU, o *Random Forest* se destacou como o modelo com melhor performance. Ao aplicar a abordagem BCOU, novamente o *XGBoost* apresentou a melhor performance, enquanto que na abordagem BCUO, o *Random Forest* obteve o melhor resultado.

Tabela 27 – Tabela de resultados (*Recall*) - Métodos Ensemble

<i>Recall</i>				
Métodos Ensemble				
Abordagem	<i>CatBoost</i>	<i>LightGBM</i>	<i>Random Forest</i>	<i>XGBoost</i>
BCO	73,85%	71,79%	76,41%	84,62%
BCU	93,33%	93,33%	100%	93,33%
BCUO	76,34%	74,73%	73,12%	76,88%
BCOU	89,47%	89,47%	94,74%	89,47%

Fonte: Elaborado pelo autor

Na tabela 28 podemos observar que ao executar o AprAVC utilizando as abordagens com *oversampling* e com *oversampling* mais *undersampling* o método MLP não performou tão bem na proporção de verdadeiros positivos.

Tabela 28 – Tabela de resultados (*Recall*) - Redes Neurais

<i>Recall</i>	
Redes Neurais	
Abordagem	<i>MLP</i>
BCO	35,90%
BCU	86,67%
BCUO	25,81%
BCOU	84,21%

Fonte: Elaborado pelo autor

A tabela 29 apresenta o resultado final para a métrica *Recall* com a melhor performance para cada tipo de abordagem.

Tabela 29 – Melhor método por abordagem *Recall*

<i>Recall</i>		
Resultado final		
Abordagem	Método de ML	Performance
BCO	<i>Neive Bayes</i> e SVM	100%
BCU	<i>Neive Bayes</i> , SVM e <i>Random Forest</i>	100%
BCUO	<i>Naive Bayes</i>	100%
BCOU	<i>Neive Bayes</i> , SVM e KNN	100%

Fonte: Elaborado pelo autor

6.6 Resultados - *F1 Score*

Na tabela 30 os resultados da nossa análise mostram que a abordagem CBO foi a mais eficaz para o método *Logistic Regression*, enquanto que para a abordagem CBU, o KNN teve a melhor performance. Já na abordagem BCOU, o *Naive Bayes*, e na abordagem BCUC, foi o KNN que obteve a melhor resultado.

Tabela 30 – Tabela de resultados (*F1 Score*) - Métodos Clássicos

<i>F1 Score</i>					
Métodos Clássicos					
Abordagem	<i>Decision tree</i>	KNN	<i>Logistic Regression</i>	<i>Naive Bayes</i>	SVM
BCO	81,43%	84,47%	87,14%	86,47%	86,47%
BCU	72,73%	87,50%	72,73%	83,33%	83,33%
BCOU	78,70%	81,44%	85,57%	85,91%	84,40%
BCUC	71,79%	92,68%	78,05%	86,36%	86,36%

Fonte: Elaborado pelo autor

A tabela 31 apresenta os resultados para os métodos *Ensemble*. A abordagem CBO alcançou melhor performance com o *XGBoost*, já a abordagem CBU a melhor performance foi o método *Random Forest*. A abordagem BCOU a melhor performance foi *XGBoost* e a BCUC o *Random Forest* teve a melhor performance.

Tabela 31 – Tabela de resultados (*F1 Score*) - Métodos Ensemble

<i>F1 Score</i>				
Métodos Ensemble				
Abordagem	<i>CatBoost</i>	<i>LightGBM</i>	<i>Random Forest</i>	<i>XGBoost</i>
BCO	82,52%	81,40%	83,94%	88,95%
BCU	80%	82,35%	88,24%	80%
BCOU	83,53%	83,99%	81,93%	84,87%
BCUC	80,95%	82,93%	90%	80,95%

Fonte: Elaborado pelo autor

A tabela 32 apresenta os resultados do método MLP para as quatro abordagens.

Tabela 32 – Tabela de resultados (*F1 Score*) - Redes Neurais

<i>F1 Score</i>	
Redes Neurais	
Abordagem	MLP
BCO	50,18%
BCU	78,79%
BCOU	38,55%
BCUO	80,00%

Fonte: Elaborado pelo autor

Consolidando os resultados da métrica *F1 Score*, a tabela 33 apresenta o método que teve melhor performance diante dos quatro tipos de abordagens.

Tabela 33 – Melhor método por abordagem (*F1 score*)

<i>F1 Score</i>		
Resultado final		
Abordagem	Método de ML	Performance
BCO	<i>XGBoost</i>	88,95%
BCU	<i>Random Forest</i>	88,24%
BCOU	<i>Naive Bayes</i>	85,91%
BCUO	<i>KNN</i>	92,68%

Fonte: Elaborado pelo autor

6.7 Resultados - AUC de ROC

Ao analisarmos a tabela 34, é possível observar que na abordagem CBOU, o método *Logistic Regression* alcançou o melhor desempenho em termos de AUC. No entanto, nas outras três abordagens (CBU, CBO e CBUO), o método *Naive Bayes* apresentou o melhor resultado em todas elas.

Tabela 34 – Tabela de resultados (AUC de ROC) - Métodos Clássicos

<i>AUC</i>					
Métodos Clássicos					
Abordagem	<i>Decision tree</i>	KNN	<i>Logistic Regression</i>	<i>Naive Bayes</i>	SVM
BCO	84,62%	88,13%	91,95%	92,33%	91,03%
BCU	74,21%	91,58%	87,02%	94,74%	93,68%
BCOU	80,88%	88,24%	91,35%	91,34%	90,20%
BCUO	71,05%	92,38%	87,26%	94,46%	92,80%

Fonte: Elaborado pelo autor

Na tabela 35 podemos observar que a abordagem CBO o melhor resultado foi alcançado pelo método *XGBoost*. Já o CBU a melhor performance foi *Random Forest*. A BCOU

o melhor resultado foi LightGBM e BCUO o *Random Forest* obteve a melhor performance.

Tabela 35 – Tabela de resultados (AUC de ROC) - Métodos Ensemble

AUC				
Métodos Ensemble				
Abordagem	<i>CatBoost</i>	<i>LightGBM</i>	<i>Random Forest</i>	<i>XGBoost</i>
BCO	95,62%	96,51%	96,17%	97,31%
BCU	79,65%	89,82%	91,93%	85,96%
BCOU	95,86%	96,87%	95,86%	96,85%
BCUO	77,56%	87,40%	90,30%	84,21%

Fonte: Elaborado pelo autor

A tabela 36 apresenta os resultados do método MLP para as quatro abordagens.

Tabela 36 – Tabela de resultados (AUC de ROC) - Redes Neurais

AUC	
Redes Neurais	
Abordagem	<i>MLP</i>
BCO	86,87%
BCU	85,26%
BCOU	86,22%
BCUO	84,76%

Fonte: Elaborado pelo autor

Fazendo uma consolidação dos resultados utilizando a métrica AUC a tabela 37 vai apresentar o método que teve melhor performance diante dos quatro tipos de abordagens.

Tabela 37 – Melhor método por abordagem (AUC de ROC)

AUC de ROC		
Resultado final		
Abordagem	Método	Performance
BCO	<i>XGBoost</i>	97,31%
BCU	<i>Naive Bayes</i>	94,74%
BCOU	<i>LightGBM</i>	96,87%
BCUO	<i>Naive Bayes</i>	94,46%

Fonte: Elaborado pelo autor

6.8 Avaliação de desempenho

Na tabela 38, são apresentados os algoritmos que obtiveram o melhor desempenho para cada tipo de abordagem de balanceamento de cargas. É importante ressaltar que a métrica AUC de ROC é a mais apropriada para este tipo de problema, uma vez que avalia a precisão das

classificações em vez de seus valores absolutos, tornando-se invariante em escala e permitindo medir a qualidade das previsões do modelo.

Tabela 38 – Melhores Performances

Melhor performance							
Abordagem	Método	AUC	Precisão	Recall	F1 score	Acc	TPT
BCO	<i>XGBoost</i>	97,31%	93,75%	84,62%	88,95%	89,49%	335,32s
BCU	<i>Naive Bayes</i>	94,74%	71,43%	100%	83,33%	82,35%	2,71s
BCOU	<i>LightGBM</i>	96,87%	95,85%	74,73%	83,99%	86,09%	7,29s
BCUO	<i>Naive Bayes</i>	94,46%	76%	100%	86,36%	84,21%	1,89s

Fonte: Elaborado pelo autor

Os resultados positivos e de alto desempenho obtidos pelo AprAVC comprovam que é possível construir uma nova abordagem utilizando algoritmos de aprendizagem de máquina para auxiliar na prevenção da recorrência do AVC, mesmo em situações de bases com cargas desbalanceadas. Os melhores resultados obtidos para cada abordagem demonstraram ter alcançado um índice satisfatório de desempenho, inclusive quando comparados aos trabalhos apresentados no capítulo 4.

De maneira geral, a abordagem que utilizou o algoritmo *XGBoost* em conjunto com a técnica BCO apresentou a melhor performance, alcançando um valor de 97,31% na métrica AUC de ROC. Isso demonstra uma excelente capacidade de distinguir entre classes positivas e negativas. Portanto, essa abordagem é altamente recomendada para a tarefa em questão.

7 CONCLUSÕES

A saúde é um aspecto universal que afeta toda a sociedade em diferentes graus. A complexidade da saúde envolve desde a prevenção e promoção da saúde até o tratamento de doenças e condições, abrangendo diversos setores e profissionais da área.

No Brasil e no mundo, o AVC é uma das principais causas de morte e incapacidade, representando um grande desafio para a sociedade e para o sistema de saúde. A prevenção da recorrência do AVC é fundamental para evitar suas consequências graves e impactantes na vida dos pacientes e suas famílias.

Nesse contexto, a aplicação de técnicas de aprendizagem de máquina tem se mostrado uma alternativa promissora para contribuir com a prevenção e tratamento do AVC. Por meio da análise de dados e identificação de padrões, é possível desenvolver modelos preditivos que auxiliem os profissionais de saúde na tomada de decisão e na prevenção da recorrência do AVC.

Este trabalho apresentou uma nova abordagem para apoiar a prevenção da recorrência do AVC. Utilizando métodos de aprendizagem de máquina supervisionados por classificação para treinamento e teste, a base de dados final foi obtida pela união de três base de dados fornecidos pela Liga Fisioneuro da UFC. Durante o desenvolvimento do AprAVC, foram utilizadas técnicas de balanceamento de cargas, permitindo a execução de quatro tipos de abordagens distintas:

- *oversampling*;
- *undersampling*;
- *undersampling* e *oversampling*;
- *oversampling* e *undersampling*.

Após a execução de cada método de aprendizagem de máquina, o AprAVC gerou diversos resultados que foram apresentados no capítulo 6. A análise dos resultados permitiu constatar que o desempenho do AprAVC foi satisfatório quando comparado com os trabalhos apresentados na revisão bibliográfica, demonstrando a viabilidade de alcançar uma performance ótima para auxiliar na prevenção da recorrência do AVC, mesmo diante de uma base de dados desbalanceada. Além disso, os resultados evidenciaram que o AprAVC pode contribuir diretamente para a prevenção da recorrência do AVC em pacientes que já tiveram o primeiro episódio, possibilitando um diagnóstico precoce que pode auxiliar no tratamento e evitar complicações futuras.

Com base na base de dados e na metodologia aplicada, as quatro abordagens utilizadas alcançaram um excelente desempenho na detecção da recorrência do AVC. Na abordagem

BCO, o método *XGBoost* apresentou uma excelente performance com AUC de ROC de 97,13%. Enquanto na abordagem BCU, o método *Naive Bayes* alcançou um resultado satisfatório com AUC de ROC de 94,74%. Na abordagem BCOU, o método *LightGBM* apresentou uma AUC de ROC de 96,87%, enquanto na abordagem BCU o método *Naive Bayes* obteve uma AUC de ROC de 94,46%. Conclui-se que o AprAVC apresentou uma performance superior em relação aos trabalhos apresentados na literatura e é capaz de contribuir significativamente na detecção precoce de recorrência do AVC.

A construção de uma nova base de dados regional representa uma importante contribuição para pesquisas futuras relacionadas à recorrência do AVC no estado do Ceará. A base de dados foi construída de forma a garantir a privacidade dos pacientes e está disponível no repositório do *Kaggle*¹. Para assegurar a confidencialidade dos dados dos pacientes, nomes e quaisquer números de documentos foram removidos da base, contendo apenas informações sobre fatores de risco modificáveis e não modificáveis, bem como a indicação se o paciente teve AVC, se teve recorrência e se nunca teve AVC.

7.1 Trabalhos futuros

Como perspectivas futuras, pretende-se manter um acompanhamento constante dos pacientes por meio de pesquisas de campo realizadas pelos fisioterapeutas da Liga Fisioneuro, com o objetivo de aumentar o número de registros na base de dados. Além disso, planeja-se uma ação conjunta com a Liga Fisioneuro para aumentar a quantidade de atributos na base de dados, como por exemplo, o tempo decorrido desde o episódio do AVC e se o paciente apresentou perda de peso em relação ao período pré-AVC. Tais informações são relevantes para aprimorar a análise do perfil dos pacientes e a efetividade das intervenções realizadas para prevenção da recorrência do AVC.

Será desenvolvido um aplicativo com o objetivo de sensibilizar usuários leigos sobre a problemática do AVC e também permitir o cadastro de pacientes por meio de uma tela com informações detalhadas. Essa iniciativa se tornará mais um canal de contribuição para o aumento da quantidade de registros na base de dados.

O AprAVC terá um novo módulo desenvolvido para mensageria em colaboração com a Liga Fisioneuro. Essa nova funcionalidade permitirá a integração do AprAVC em postos de saúde na cidade de Fortaleza. O AprAVC enviará alertas via mensagens para os assistentes

¹ <https://www.kaggle.com/datasets/digaomesquita/avcrecurrence>

sociais dos postos de saúde de cada região, informando uma possível recorrência do AVC em pacientes analisados. Essa iniciativa tem como objetivo possibilitar uma intervenção mais rápida e efetiva, reduzindo o risco de danos permanentes e aumentando as chances de recuperação do paciente.

Para aprimorar o AprAVC, pretende-se utilizar outros métodos além dos já empregados neste estudo como trabalho futuro. Adicionalmente, serão explorados mais métodos de redes neurais para encontrar as melhores combinações de algoritmos e abordagens, com o objetivo de maximizar a acurácia do AprAVC.

REFERÊNCIAS

- ABEDI, V.; AVULA, V.; CHAUDHARY, D.; SHAHJOUEI, S.; KHAN, A.; GRIESSENAUER, C. J.; LI, J.; ZAND, R. Prediction of long-term stroke recurrence using machine learning models. **Journal of Clinical Medicine**, v. 10, n. 6, p. 1286–2002, 2021.
- ADEBAYO, S. **HOW CATBOOST ALGORITHM WORKS IN MACHINE LEARNING**. 2021. Disponível em: <<https://dataaspirant.com/catboost-algorithm/>>. Acesso em: 16 NOV. 2021.
- AHAMED, S.; ARYA, M. S. Lgbm classifier based technique for predicting type-2diabetes. **European Journal of Molecular & Clinical Medicine**, v. 8, p. 454–467, 2021.
- ALPAYDIN, E. **Introduction to machine learning**. 2. ed. Cambridge, Massachusetts, EUA: MIT Press, 2010. v. 1.
- ALPAYDIN, E. **Machine learning: The new AI**. Cambridge, Massachusetts, EUA: MIT Press, 2016. v. 1.
- ARAUJO, R. **Engenharia de Atributos**. 2020. Disponível em: <<https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/>>. Acesso em: 17 NOV. 2021.
- AZANK, F. **Dados Desbalanceados — O que são e como lidar com eles?** 2020. Disponível em: <<https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b>>. Acesso em: 27 JUN. 2022.
- BANERJEE, P. **LightGBM Classifier in Python**. 2020. Disponível em: <<https://www.kaggle.com/code/prashant111/lightgbm-classifier-in-python/notebook>>. Acesso em: 16 NOV. 2021.
- BATISTA, A. F. d. M.; FILHO, A. D. P. C. Machine learning aplicado a saúde. In: **Livro de Minicursos do Simposio Brasileiro de Computação Aplicado a saúde 2019**. Porto Alegre, Rio Grande do Sul, Brasil: SBC, 2019. p. 1–42.
- BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of xgboost. **Machine Learning**, 2019.
- BERTSIMAS, D.; KING, A. Logistic regression: From art to science. **Statistical Science**, v. 32, n. 3, p. 367–384, 2017.
- BOCCHI, S. C. M. O papel do enfermeiro como educador junto a cuidadores familiares de pessoas com AVC. **Revista Brasileira de Enfermagem**, v. 5, p. 569–573, 2004.
- BONITA, R.; BEAGLEHOLE, R. Stroke prevention in poor countries time for action. **Stroke**, v. 38, n. 11, p. 2871–2872, 2007.
- BOTELHO, T. d. S. B.; NETO, C. D. M.; ARAUJO, F. L. C. d.; ASSIS, S. C. Epidemiologia do acidente vascular cerebral no Brasil. **Temas em Saúde**, v. 16, n. 2, p. 2447–2131, 2016.
- BREIMAN, L. Bagging predictors. **Machine Learning**, Academic Publishers, v. 24, p. 123–140, 1996.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001.

- CASTRO, L. N. d.; FERRARI, D. G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. 1. ed. São Paulo, Brasil: Editora Saraiva, 2016. v. 1.
- CHANG, W.; LIU, Y.; XU, X.; ZHOU, S. Prediction model of stroke recurrence based on support vector machine. In: **The Second International Conference on Physics, Mathematics and Statistics**. Bristol, England: IOP, 2019. v. 1324.
- CHAVES, M. L. F. Acidente vascular encefálico: conceituação e fatores de risco. **Stroke**, v. 13, p. 290–295, 2000.
- CHAWLA, N. V. Data mining for imbalanced datasets: An overview. In: **Data Mining and Knowledge Discovery Handbook**. 2. ed. Berlim, Alemanha: Springer, 2010.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. Nova Iorque, Estados Unidos: ACM, 2016. p. 785–794.
- CORTES, C.; VAPNIK, V. N. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995.
- DEBAUN, M. R.; JORDAN, L. C.; KING, A. A.; SCHATZ, J.; VICHINSKY, E.; FOX, C. K.; MCKINSTRY, R. C.; TELFER, P.; KRAUT, M. A.; DARAZ, L.; KIRKHAM, F. J.; MURAD, M. H. American society of hematology 2020 guidelines for sickle cell disease: prevention, diagnosis, and treatment of cerebrovascular disease in children and adults. **Blood Advances**, v. 4, n. 8, p. 1554–1588, 2020.
- DELASHMIT, W. H.; MANRY, M. T. Recent developments in multilayer perceptron neural networks. In: **7th Annual Memphis Area Engineering and Science Conference (MAESC)**. Berlim, Alemanha: Springer, 2005. p. 1–16.
- DIÁRIO DO NORDESTE. **Um a cada quatro casos de AVC no Ceará atinge pessoas com menos de 60 anos**. 2020. Disponível em: <<https://diariodonordeste.verdesmares.com.br/metro/um-a-cada-quatro-casos-de-avc-no-ceara-atinge-pessoas-com-menos-de-60-anos-1.3006492>>. Acesso em: 28 DEZ. 2021.
- DOROGUSH, A. V.; ERSHOV, V.; YANDEX, A. G. Catboost: gradient boosting with categorical features support. **Computer Science, Machine Learning**, Cornell University, 2018.
- ESCOVEDO, T. **Machine Learning: Conceitos e Modelos — Parte I: Aprendizado Supervisionado**. 2020. Disponível em: <<https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445>>. Acesso em: 10 JUN. 2022.
- ESCOVEDO, T.; KOSHIYAMA, A. S. **Introdução a Data Science — Algoritmos de Machine Learning e métodos de análise**. 1. ed. São Paulo, Brasil: Editora Casa do Código, 2020.
- FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 1. ed. Rio de Janeiro, Brasil: LTC, 2011.
- FAVERO, L. P. L.; BELFIORE, P. P.; SILVA, F. L. d.; CHAN, B. L. **Análise de dados: modelagem multivariada para tomada de decisões**. 1. ed. Amsterdam, Holanda: Elsevier, 2009.

FEIGIN, V. L.; LAWES, C. M.; A, B. D.; ANDERSON, C. S. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. **The lancet Neurology**, v. 2, n. 1, p. 43–53, 2003.

FERNANDEZ-CADENAS, I.; MENDIOROZ, M.; GIRALT, D.; NAFRIA, C.; GARCIA, E.; GALLEG0-FABREGA, C.; DOMINGUES-MONTANARI, S.; DELGADO, P. Grecos project (genotyping recurrence risk of stroke). **Stroke**, v. 48, n. 5, p. 1147–1153, 2017.

FERRARI, D. G.; SILVA, L. N. d. C. **Introdução a mineração de dados**. São Paulo, Brasil: Editora Saraiva, 2017.

FOROUZANFAR, M. H.; LIU, P.; ROTH, G. A.; BIRYKOV, S.; MARCZAK, L.; ALEXANDER, L.; MURRAY, C. J. Global burden of hypertension and systolic blood pressure of at least 110 to 115 mm hg, 1990–2015. **JAMA**, Chicago, Illinois, Estados Unidos, v. 317, n. 2, p. 165–182, 2017.

FUNDAÇÃO REGIONAL DA SAÚDE DO CEARÁ. **Em 11 anos, hospitalizações por AVC no Ceará aumentam e óbitos de pacientes internados pela doença reduzem**. 2022. Disponível em: <<https://www.funsaude.ce.gov.br/2022/07/01/>>. Acesso em: 2022-07-22.

GIANNAKAS, F.; TROUSSAS, C.; KROUSKA, A.; SGOUROPOULOU, C.; VOYIATZIS, I. Xgboost and deep neural network comparison: The case of teams performance. In: **17th International Conference on Intelligent Tutoring Systems**. Berlim, Alemanha: Springer, 2021. p. 343–349.

GRUS, J. **Data science do zero: primeiras regras com o Python**. 1. ed. Rio de Janeiro, Brasil: Editora Alta Books, 2016. v. 1.

GUAN, D.; YUAN, W.; LEE, Y.-K.; NAJEEBULLAH, K.; RASEL, M. K. A review of ensemble learning based feature selection. **IETE Technical Review**, Taylor & Francis, v. 31, p. 190–198, 2014.

GUPTA, P. **Cross-Validation in Machine Learning**. 2017. Disponível em: <<https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>>. Acesso em: 09 DEZ. 2021.

HALIMU, C.; KASEM, A.; NEWAZ, S. H. S. Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In: **3rd International Conference on Machine Learning and Soft Computing (ICMLSC)**. Nova Iorque, Estados Unidos: ACM, 2019. p. 1–6.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. Berlim, Alemanha: Springer, 2009. v. 1.

HO, T. K. Random decision forests. In: **3rd International Conference on Document Analysis and Recognition**. Manhattan, Nova Iorque, Estados Unidos: IEEE, 1995. v. 1, p. 278–282.

JIANG, L.; CAI Z. AND WANG, D.; JIANG, S. Passos sequenciais do algoritmo knn. In: **Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)**. Manhattan, Nova Iorque, Estados Unidos: IEEE, 2007. v. 1, p. 679–683.

JO, J.-M. Effectiveness of normalization pre-processing of big data to the machine learning performance. **Journal of the KIECS**, v. 14, p. 547–552, 2019.

JOHN, B. **When to Choose CatBoost Over XGBoost or LightGBM [Practical Guide]**. 2022. Disponível em: <<https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm>>. Acesso em: 16 NOV. 2021.

JOHN, G.; P., L. Estimating continuous distributions in bayesian classifiers. In: **11th international conference on uncertainty in artificial intelligence**. São Mateo, California, Estados Unidos: Morgan Kaufmann, 1995. v. 11, p. 338–345.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. In: **Artificial Intelligence**. Pensilvânia, EUA: AAAS, 2015. v. 349, p. 255–260.

JUNIOR, C. M. J. M. D.; SILVA, S. P. P. D.; NÓBREGA, R. V. M. D.; BARROS, A. C. D. S.; FILHO, P. P. R.; ALBUQUERQUE, V. H. C. D. Deep learning iot system for online stroke detection in skull computed tomography images. **Computer Networks**, Elsevier, v. 152, p. 25–39, 2019.

KASPER, D. L.; JAMESON, J. L.; FAUCI, A. S.; LONGO, D. L.; LOSCALZO, J. **Medicina interna de Harrison**. 19. ed. Porto Alegre, Rio Grande do Sul, Brasil: AMGH, 2019. v. 2.

KELLEHER, J.; TIERNEY, B. **Data Science**. Cambridge, Massachusetts, EUA: MIT Press, 2018. v. 1.

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. In: **Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies**. Amsterdam, Holanda: IOS Press, 2007. v. 31, n. 3, p. 249–268.

KULESH, S. D.; FILINA, N. A.; FRANTAVA, N. M.; ZHYTKO, N. L.; KASTSINEVICH, T. M.; KLIATSKOVA, L. A.; SHUMSKAS, M. S.; HILZ, M. J.; SCHWAB, S.; KOLOMINSKY-RABAS, P. L. Risk factors for early recurrence after ischemic stroke: the role of stroke syndrome and subtype. **Stroke**, v. 29, p. 2726–2760, 2010.

KUMMER, B. R.; HAZAN, R.; MERKLER, A. E.; KAMEL, H.; WILLEY, J. Z.; MIDDLESWORTH, W.; BOEHME, A. A multilevel analysis of surgical category and individual patient-level risk factors for postoperative stroke. **The Neurohospitalist**, v. 10, n. 1, p. 22–28, 2020.

LANTZ, B. **Machine Learning with R**. 2. ed. Michigan, Estados Unidos: Packt Publishing, 2013. v. 1.

LEMESHOW, S.; JR, D. W. H. Logistic regression. In: **Wiley Encyclopedia of Clinical Trials**. [S.l.]: WILEY, 2007.

LENZ, M. L.; NEUMANN, F. B.; SANTARELLI, R.; SALVADOR, D. **Fundamentos de Aprendizagem de Máquinas**. 1. ed. Porto Alegre, Rio grande do Sul, Brasil: Editora Sagah Educação, 2020.

LIN, Y.; JEON, Y. Random forests and adaptive nearest neighbors. **Taylor & Francis: Journal of the American Statistical Association**, v. 101, n. 474, p. 578–590, 2006.

LU, T.; WANG, Y. Prediction model construction for ischemic stroke recurrence with bp network and multivariate logistic regression and effect of individualized health education. **Computational and mathematical methods in medicine**, Hindawi, v. 2022, p. 1–10, 2022.

LUZ, F. **Algoritmo KNN para classificação**. 2019. Disponível em: <<https://inferir.com.br/artigos/algoritmo-knn-para-classificacao/>>. Acesso em: 10 JUN. 2022.

MARIANELLI, M.; MARIANELLI, C.; NETO, T. P. L. Principais fatores de risco do AVC isquêmico: Uma abordagem descritiva. **Brazilian Journal of health Review**, v. 3, n. 6, p. 19679–19690, 2020.

MARIANO, D. **Métricas de avaliação em machine learning**. 2021. Disponível em: <<https://ricardomatsumura.medium.com/engenharia-de-atributos-87370c31d0eb>>. Acesso em: 16 NOV. 2021.

MARQUES, A. I.; GARCIA, V.; SANCHEZ, J. S. A review of ensemble learning based feature selection. **Expert Systems with Applications**, Elsevier, v. 39, p. 10244–10250, 2012.

MBAABU, O. **Introduction to Random Forest in Machine Learning**. 2020. Disponível em: <<https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>>. Acesso em: 29 NOV. 2021.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, p. 115–133, 1943.

MINISTÉRIO DA SAÚDE. **Diretrizes de atenção à reabilitação da pessoa com acidente vascular cerebral**. Brasília, Distrito Federal, Brasil, 2013. Disponível em: <https://bvsmis.saude.gov.br/bvsmis/publicacoes/diretrizes_atencao_reabilitacao_acidente_vascular_cerebral.pdf>.

MINISTÉRIO DA SAÚDE. **Acidente Vascular Cerebral**. 2015. Disponível em: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc>>. Acesso em: 02 ABR. 2023.

MINISTÉRIO DA SAÚDE. **Dia Mundial do AVC: Ministério da Saúde alerta para os tipos, sintomas e prevenção**. 2022. Disponível em: <<https://www.gov.br/saude/pt-br/assuntos/noticias/2022/outubro/dia-mundial-do-avc-ministerio-da-saude-alerta-para-os-tipos-sintomas-e-prevencao>>.

MINISTÉRIO DA SAÚDE. **Prevenção de doenças cardiovasculares na Atenção Primária é tema de destaque em congresso Global Stroke Alliance**. 2022. Disponível em: <<https://aps.saude.gov.br/noticia/18448>>.

MINISTÉRIO DA SAÚDE. **Como prevenir o AVC?** 2023. Disponível em: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc/como-prevenir-o-avc>>. Acesso em: 02 ABR. 2023.

MINISTÉRIO DA SAÚDE. **Tratamento e reabilitação do AVC**. 2023. Disponível em: <<https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc/tratamento-e-reabilitacao-do-avc>>. Acesso em: 02 ABR. 2023.

MINSKY, M. L.; PAPERT, S. A. **Perceptrons. An Introduction to Computational Geometry**. Cambridge, Massachusetts, EUA: MIT Press, 1969.

MITCHELL, T. M. **Machine Learning**. 1. ed. Nova Iorque, Estados Unidos: McGraw-Hill Science, 1997.

MOHANTY, A. **Multi layer Perceptron (MLP) Models on Real World Banking Data**. 2019. Disponível em: <<https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f>>. Acesso em: 02 DEZ. 2021.

NÓBREGA, R. V. M. D.; FILHO, P. P. R.; RODRIGUES, M. B.; SILVA, S. P. P. d.; JUNIOR, C. M. J. M. D.; ALBUQUERQUE, V. H. C. D. Lung nodule malignancy classification in chest computed tomography images using transfer learning and convolutional neural networks. **Neural Computing and Applications**, Springer, v. 1, p. 1–18, 2018.

OMS. **The top 10 causes of death**. 2020. Disponível em: <<https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death>>. Acesso em: 22 JAN. 2022.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COUNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RABELLO, E. B. **Cross Validation: Avaliando seu modelo de Machine Learning**. 2022. Disponível em: <<https://medium.com/@edubrazrabello/cross-validation-avaliando-seu-modelo-de-machine-learning-1fb70df15b78>>. Acesso em: 09 JUN. 2022.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. **Remote Sensing**, v. 11, p. 185–206, 2013.

RAY, S. A quick review of machine learning algorithms. In: **2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon 2019)**. Manhattan, Nova Iorque, Estados Unidos: IEEE, 2019. v. 1, p. 35–39.

REN, L.; GAO, Y.; JIANG, Y.; WANG, G.; LI, Q.; GU, Y.; YU, H.; LI, J.; ZHANG, L. Association between blood pressure indicators and stroke in aged population: A community-based nested case-control study. **Original Research**, v. 2021, n. 16, p. 997–1005, 2021.

RODRIGUES-GALIANO, V.; SANCHEZ-CASTILLO, M.; CHICA-OLMO, M.; CHICA-RIVAS, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804–818, 2015.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, p. 533–536, 1986.

RUSDAH, A. D.; MURFI, H. Xgboost in handling missing values for life insurance risk prediction. **SN Applied Sciences**, v. 2, n. 8, 2020.

RUSSELL, S.; P., N. **Artificial intelligence: a modern approach**. 3. ed. Nova Jersey, Estados Unidos: Pearson Education, 2009.

SACCO, R. L.; WOLF, P. A.; KANNEL, W. B.; MCNAMARA, P. M. Survival and recurrence following stroke: the framingham study. **Rev Bras Hipertens**, v. 7, n. 4, p. 372–382, 1982.

SCHMIDT, M. H.; SELAU, C. M.; SOARES, P. d. S.; FRANCHI, E. F.; PIBER, V. D.; QUATRIN, L. B. Acidente vascular cerebral e diferentes limitações: Uma análise interdisciplinar. **Arquivos de Ciências da Saúde da UNIPAR**, v. 23, n. 2, p. 139–144, 2019.

SCHULTEBRAUCKS, L. **Introduction to Support Vector Machines**. 2017. Disponível em: <<https://medium.com/@LSchultebras/ introduction-to-support-vector-machines-9f8161ae2fcb>>. Acesso em: 11 JUN. 2022.

SHELKE, M. S.; DESHMUKH, P. R.; SHANDILYA, V. K. A review on imbalanced data handling using undersampling and oversampling technique. **International Journal of Recent Trends in Engineering & Research (IJRTER)**, v. 3, p. 444–449, 2017.

SILVA, E. S.; BORGES, J. W. P.; MOREIRA, T. M. M.; RODRIGUES, M. T. P.; SOUZA, A. C. C. Prevalência e fatores de risco associados ao acidente vascular cerebral em pessoas com hipertensão arterial: uma análise hierarquizada. **Revista de Enfermagem Referência**, v. 5, n. 3, p. 1–8, 2020.

SILVA, F. M.; LENZ, M. L.; FREITAS, P. H. C.; SANTOS, S. C. B. **Inteligência artificial**. 1. ed. Porto Alegre, Rio grande do Sul, Brasil: Editora Sagah Educação, 2019.

SILVA, M. **Matriz de Confusão e AUC ROC**. 2019. Disponível em: <<https://medium.com/data-hackers/matriz-de-confus%C3%A3o-e-auc-roc-f7e446dca107>>. Acesso em: 19 NOV. 2021.

SMELTZER, S. C.; BARE, B. G.; HINKLE, J. L.; CHEEVER, K. H. **Tratado de enfermagem médico-cirúrgica**. 12. ed. Rio de Janeiro, Rio de Janeiro, Brasil: Guanabara Koogan, 2011. v. 1.

SOCIEDADE BRASILEIRA DE ACIDENTE VASCULAR CEREBRAL. **Acidente Vascular Cerebral**. 2023. Disponível em: <<https://avc.org.br/pacientes/acidente-vascular-cerebral/>>. Acesso em: 02 ABR. 2023.

SOMVANSHI, M.; SHITAL, T.; CHAVAN, P.; SHINDE, S. V. A review of machine learning techniques using decision tree and support vector machine. In: **International Conference on Computing Communication Control and automation (ICCUBEA)**. Manhattan, Nova Iorque, Estados Unidos: IEEE, 2016.

STAMM, B. J.; BURKE, J. F.; LIN, C. C.; PRICE, J. R.; LESLI, S. E. Disability in community-dwelling older adults: Exploring the role of stroke and dementia. **Journal of Primary Care & Community Health**, v. 10, n. 1, p. 1–11, 2019.

SUN, X.; XU, Z.; FENG, Y.; YANG, Q.; XIE, Y.; WANG, D.; YU, Y. Rbc inventory-management system based on xgboost model. **Indian Journal of Hematology and Blood Transfusion**, v. 37, n. 1, p. 126–133, 2020.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. 2. ed. Cambridge, Massachusetts, EUA: MIT Press, 2018. v. 1.

VAPNIK, V. N.; YA., C. A. On a class of perceptrons. **Automation and Remote Control**, v. 25, p. 103–109, 1964.

VARLEYSEN, M.; FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. In: **8th International Work-Conference on Artificial Neural Networks**. Berlim, Alemanha: Springer, 2005. p. 758–770.

VODENCAREVIC, A.; WEINGARTNER, M.; CARO, J. J.; UKALOVIC, D.; ZIMMERMANN-RITTEREISER, M.; SCHWAB, S.; KOLOMINSKY-RABAS, P. Prediction of recurrent ischemic stroke using registry data and machine learning methods: The erlangen stroke registry. **Stroke**, AHA Journals, v. 53, n. 7, p. 2299–2309, 2022.

WANG, D.; ZHANG, Y.; ZHAO, Y. Lightgbm: An effective mirna classification method in breast cancer patients. **Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics (ICCB)**, p. 7–11, 2017.

XU, S. Bayesian naive bayes classifiers to text classification. **Journal of Information Science**, v. 44, p. 48–59, 2016.