PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

ESCOLA POLITÉCNICA E DE ARTES

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DIAGNÓSTICO MÉDICO USANDO RACIOCÍNIO BASEADO EM CASOS E APRENDIZAGEM DE MÁQUINA

ALISSON VIANA DE ANDRADE

GOIÂNIA
2023

ALISSON VIANA DE ANDRADE

DIAGNÓSTICO MÉDICO USANDO RACIOCÍNIO BASEADO EM CASOS E APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso apresentado à Escola Politécnica, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Banca Examinadora:

Orientador: Prof. Dr. Clarimar José Coelho

Coorientador: Eng. Douglas Vieira do Nascimento

Prof. Dr. Rafael Viana de Carvalho

GOIÂNIA
2023

RESUMO

O objetivo do trabalho é o desenvolvimento de uma metodologia para o diagnóstico de doenças usando o Raciocínio Baseado em Casos (RBC). O pré-processamento da base de casos inclui a normalização de dados, tratamento de valores ausentes e seleção de características significativas, contribuindo para uma base de casos precisa. O algoritmo k-NN (k-Nearest Neighbors) é usado para determinar a proximidade de casos passados e identificar padrões similares. A escolha criteriosa do parâmetro k desempenha um papel crucial nesse processo, influenciando diretamente na sensibilidade do diagnóstico. A adaptação de casos utiliza redes neurais artificiais e explora a flexibilidade e a capacidade de aprendizado das redes neurais para ajustar os casos recuperados, levando em consideração nuances específicas do paciente em questão. A adaptação dinâmica melhora a capacidade do sistema em lidar com variações individuais e complexidades presentes em diferentes casos clínicos. Os resultados obtidos demonstram a eficácia da abordagem proposta no diagnóstico de doenças. A integração das etapas de pré-processamento, recuperação de casos com k-NN e adaptação de casos com redes neurais contribui para um sistema robusto e flexível, capaz de lidar com uma variedade de cenários clínicos

Palavras chaves: Raciocínio Baseado em Casos (RBC), Heurística de Diferença de Casos (CDH), Aprendizado de Máquina, Aprendizagem Profunda

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ML - Machine Learning: aprendizado de máquina

RBC - Raciocínio Baseado em Casos

K-NN - Algoritmo de k-nearest neighbors: algoritmo de k-vizinhos mais próximos

CDH - Adaptação de casos usando heurística de diferença de casos

SUMÁRIO

1 INTRODUÇÃO

Diagnóstico médico no contexto da aprendizagem de máquina (Machine Learning, ML) refere-se à aplicação de algoritmos e modelos computacionais para identificar e classificar condições médicas com base em dados clínicos (Ahsan, et al., 2022). Ferramentas ML analisam informações como sintomas, exames laboratoriais, imagens médicas e histórico do paciente para fornecer uma avaliação objetiva sobre a presença ou ausência de uma doença específica (Shehab, et al., 2022). O diagnóstico automatizado compreende a identificação, pelo computador, de padrões correspondentes nos dados, proporcionando suporte ao médico no diagnóstico de uma condição patológica. Quando se identifica uma anormalidade no paciente, este é direcionado a realizar um exame específico (Santos, et al., 2019). A automação do diagnóstico possibilita a detecção precoce da doença, por meio de abordagens e procedimentos computacionais, nos quais os dados são examinados e categorizados para determinar se o paciente apresenta alguma patologia (Nia, Kaplanoglu, Nasab, 2023).

O raciocínio baseado em casos (RBC) é um paradigma de resolução de problemas que se inspira na maneira como os seres humanos aprendem e aplicam conhecimento prático para resolver novos problemas (Kolodner, 1993). Na abordagem RBC, os problemas são resolvidos encontrando soluções análogas em situações passadas que são armazenadas como casos (Watson, Marir, 1994). Cada caso consiste em uma descrição de um problema passado e a solução associada (Aamodt, Plaza, 1994). A recuperação de casos ocorre quando um novo problema é apresentado, o sistema busca em sua base de casos por situações semelhantes que foram previamente resolvidas (Jian, Zhe, Zhenxing, 2015). A similaridade entre os casos é avaliada com base em características específicas. A reutilização do conhecimento é feita quando casos similares são identificados, o conhecimento contido nesses casos é aplicado para resolver o novo problema. Isso envolve a adaptação da solução do caso anterior às características específicas do problema atual. A revisão e retenção de casos é feita após a resolução do novo problema, a solução é revisada e, se apropriado, o novo caso é adicionado à base de casos (Golding, 1995). Isso contribui para a expansão contínua da base de conhecimento do sistema. O RBC é frequentemente utilizado em situações em que há uma variedade de casos passados que podem ser utilizados para orientar a resolução de problemas semelhantes no futuro (Popa, A., Wood, 2011). Essa abordagem é particularmente eficaz em domínios onde existem padrões recorrentes e experiências passadas podem ser aplicadas de maneira útil. No contexto da saúde, por exemplo, o

RBC pode ser empregado para o diagnóstico médico, como discutido anteriormente. A base de casos pode conter informações sobre sintomas, testes e tratamentos associados a condições médicas específicas, permitindo que o sistema aprenda com casos anteriores e aplique esse conhecimento ao analisar novas situações clínicas (Sharma, Sharma, 2020).

Um Sistema de Raciocínio Baseado em Casos (RBC) possui dois principais processos dentro das etapas de Recuperação e Reutilização. Na etapa de Recuperação de caso é buscado da Base de Casos o caso mais similar ao novo caso que estamos querendo resolver. A similaridade de um problema com outro é baseado no quão se parecem por um cálculo da distância euclidiana entre os dois. E na etapa de Reutilização é a etapa que será necessário a Adaptação de casos, será realizado uma adaptação do novo problema em cima do problema mais similar recuperado da base de casos. O processo de Adaptação é uma das partes mais complexas em um sistema de Raciocínio Baseado em Casos (RBC), pois depende da complexidade dos dados e do problema que queremos solucionar (Policastro, 2004).

Um dos maiores desafios da área de RBC é o desenvolvimento de métodos eficientes para a adaptação de casos. Uma alternativa para superar as dificuldades associadas à aquisição de conhecimento para adaptação de casos tem sido a utilização de abordagens híbridas e de algoritmos de aprendizado automático para a aquisição do conhecimento utilizado para a adaptação (Policastro, 2004).

Para a abordagem de Recuperação de Casos, é usado no trabalho algumas propostas de algoritmos que de fato obtêm a similaridade entre o caso atual e um outro caso, que segundo Policastro (2004) Um caso pode ser considerado útil ao problema atual se ele pode ser facilmente aplicado para a resolução desse novo problema. As propostas para serem implementadas no processo de recuperação de casos são: *kd-tree* que possibilita a recuperação por similaridade baseada em Árvores de Decisão; *Fish and Shrink* possibilita a redução sucessiva dos intervalos de similaridades possíveis entre uma consulta e os casos da base de casos; Case Retrieval Nets utilizam conceitos de redes neurais para construção de uma estrutura de memória com nós e arestas, que as arestas possuem pesos indicando a similaridade. As estratégias de adaptação por substituição trocam os valores de uma solução prévia por valores apropriados para uma nova situação. As estratégias de adaptação por transformações alteram a estrutura da solução pela inclusão ou remoção de componentes da solução recuperada para atender às

necessidades do novo problema. As estratégias de adaptação por geração repetem os passos que foram executados para se obter uma solução recuperada, no contexto do novo problema. Quando um novo problema é apresentado, a solução desse novo problema é obtida diretamente repetindo-se o processo executado para a obtenção da solução do problema recuperado (Policastro, 2004)

E para a etapa de adaptação, o artigo aborda uma metodologia de modificação de atributos no problema atual. Ou seja, fazendo a recuperação pela similaridade, a adaptação do novo caso/problema pode ser adaptado apenas com mudanças de atributos.

A RBC provou ser especialmente útil para resolução de problemas e aplicações de decisão nos domínios das ciências da saúde. As motivações para a aplicação dos domínios do CB estão presentes desde o início, com o objetivo de melhorar a saúde (Bichindaritz e Marling, 2010, tradução nossa).

Por conta disso, o objetivo do trabalho é aplicar a técnica de RBC na área da saúde voltado para a diabetes. Foi escolhido uma base de dados mais simples de diabetes para ser testada a aplicação do RBC.

2 MATERIAIS E MÉTODOS

A aplicação de RBC no diagnóstico de diabetes apresenta alguns desafios específicos, principalmente devido à natureza complexa e variável dessa condição. Alguns dos problemas associados a essa aplicação são a heterogeneidade da diabetes. A diabetes é uma condição médica que abrange diferentes tipos, como diabetes tipo 1 e tipo 2, cada um com características distintas. Além disso, a condição pode se manifestar de maneira diferente em cada indivíduo. A heterogeneidade da diabetes torna desafiador encontrar casos passados verdadeiramente representativos para a aplicação do RBC. Evolução temporal da diabetes sugere que o quadro clínico dos pacientes com diabetes pode evoluir ao longo do tempo, exigindo adaptações constantes no tratamento e no diagnóstico. Casos passados podem não refletir adequadamente as mudanças na condição do paciente, prejudicando a eficácia do RBC, que geralmente assume que casos semelhantes levam a soluções semelhantes. A variedade de fatores contribuintes contribui para que a diabetes seja influenciada por uma variedade de fatores, incluindo genética, estilo de vida, dieta e outros. O RBC pode ter dificuldade em lidar com a complexidade dessas variáveis e em identificar casos passados que considerem adequadamente esses fatores múltiplos. Necessidade de dados atualizados para um diagnóstico preciso, é essencial contar com dados clínicos atualizados sobre os dados pessoais das pessoas; peso, altura, índice de glicemia, etc. O RBC depende de uma base de casos que reflita a evolução contínua do conhecimento médico sobre a diabetes. Se a base de casos estiver desatualizada, as conclusões do RBC podem não refletir as práticas e descobertas médicas mais recentes. Limitações na quantidade de dados disponíveis, qualidade e quantidade de dados disponíveis para treinar e alimentar o sistema RBC são fundamentais. Em alguns casos, pode haver uma escassez de dados específicos para determinadas variações da diabetes, prejudicando a capacidade do RBC de fornecer diagnósticos precisos.

2.1 Coleta de Dados

A Base de Dados de diabetes que será utilizada neste trabalho, é possível ser encontrada no site da Kaggle, que é um site de repositórios de várias bases de dados. O conjunto de dados é composto por 768 pessoas divididas em 9 atributos, o conjunto é composto apenas por mulheres com pelo menos 21 anos de idade. Esses dados são originalmente do *National Institute of Diabetes and Digestive and Kidney Diseases*.

Na imagem abaixo podemos visualizar as cinco primeiras e últimas linhas representando cada linha um paciente e cada paciente um casos

Figura 1 - Imagem da base de dados antes do tratamento dos dados, oito atributos e uma Alvo binária.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Fonte: autoria própria

O conjunto de dados possui 9 atributos para cada paciente, sendo assim, segue abaixo o dicionário dos atributos.

Pregnancies = Grávida.
Glucose = Nível de glicose no sangue.
BloodPressure = Pressão arterial.
SkinThickness = Espessura da pele.
Insulin = Insulina.
BMI = índice de massa corporal.
DiabetesPedigreeFunction = Avalia a probabilidade de diabetes com base na família.
Age = Idade.

Por fim, temos nossa variável Alvo, que basicamente é o resultado de se o paciente possui diabetes ou não. É uma Alvo binário sendo **1** para diabetico e *0* para não diabetico.

Outcome = Resultado da nossa variável Alvo.

A partir da análise de dados desse dataset, busco por meio de um sistema de Raciocínio Baseado em Casos (RBC) encontrar casos similares ao novo caso e adaptá-los para fazer o diagnóstico desse novo caso para identificar se o novo caso é de um paciente com diabetes ou não.

2.1.1 Pré-processamento dos dados

A etapa de pré-processamento dos dados é de extrema importância, pois dados vazios ou nulos podem prejudicar o desempenho do nosso modelo de aprendizado de máquina e ter uma métrica de acurácia bastante baixa.

2.1.1.1 Tratamento de valores desconhecidos

Um problema em pré-processamento de dados é o tratamento de valores desconhecidos. Valores ausentes consistem na não medição de valores para um ou mais atributos em determinadas situações como recusa por parte de entrevistados em responder determinadas perguntas, defeitos em equipamentos, entre outras (Policastro, 2004)

Nessa etapa é verificado se existem valores nulos ou vazios para os seguintes atributos: Glucose, BloodPressure, SkinThickness, Insulin, BMI. Com isso, o conjunto de dados foi reduzido de 768 para 392 pacientes.

Figura 2 - Base de dados depois da verificação de valores nulos e vazios, o conjunto de dados foi reduzido de 768 para 392 pacientes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
6	3	78	50	32	88	31.0	0.248	26	1
8	2	197	70	45	543	30.5	0.158	53	1
13	1	189	60	23	846	30.1	0.398	59	1
...	...	...	...	...	...	...	...	...	...
753	0	181	88	44	510	43.3	0.222	26	1
755	1	128	88	39	110	36.5	1.057	37	1
760	2	88	58	26	16	28.4	0.766	22	0
763	10	101	76	48	180	32.9	0.171	63	0
765	5	121	72	23	112	26.2	0.245	30	0

392 rows × 9 columns

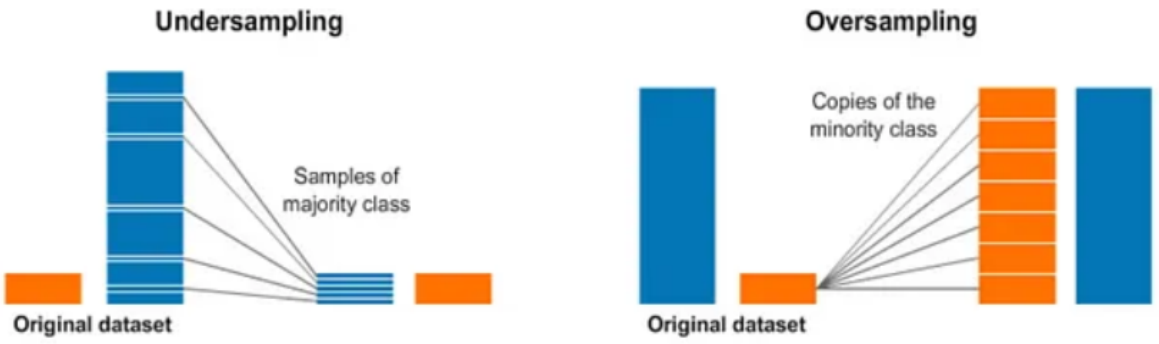
Fonte: autoria própria

2.1.1.2 Balanceamento de Classes

O balanceamento de classes é uma etapa importante no pré-processamento de dados, pois uma base de dados com classes desbalanceadas, afetam diretamente no treinamento de qualquer modelo de aprendizado de máquinas.

Existem dois processos de balanceamento de classes, sendo eles o over-sampling e under-sampling.

Figura 3 - a imagem do lado esquerdo mostra o processo de Undersampling, que consiste em pegar elementos da classe majoritária e igualar com a classe minoritária. E do lado direito é o processo ao contrário.

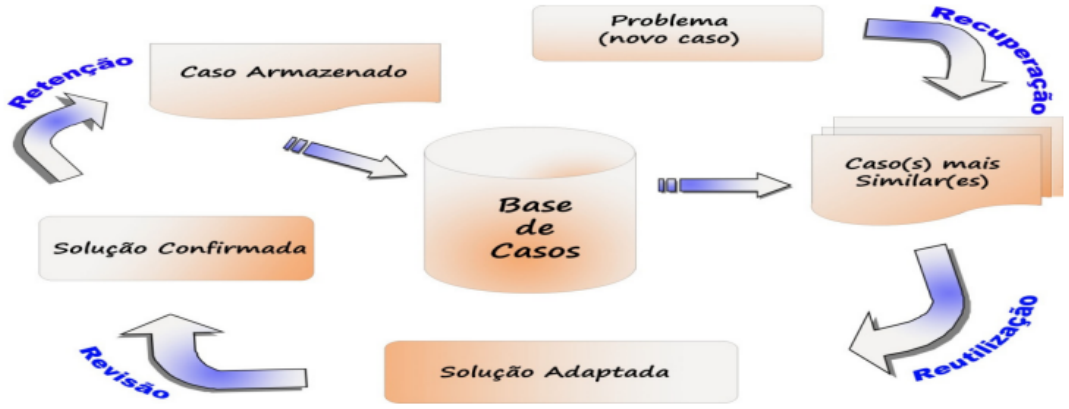


Fonte: Medium, 2019

2.2 Raciocínio Baseado em Casos (RBC)

O Raciocínio Baseado em Casos (RBC) é criado em cima do comportamento humano, como a capacidade de lembrar de um evento passado para solucionar um evento atual. As etapas de um ciclo de RBC são baseados em quatro etapas, sendo elas a recuperação, que consiste em buscar casos similares ao novo caso da base de caso; a reutilização, que consiste em entender que não existe casos idênticos e com isso é necessário uma adaptação da solução do caso recuperado; a revisão que consiste em ser o aprendizado com falhas ou acerto, no qual será julgado de acordo com a solução; e por fim a retenção, que após a revisão, o novo caso solucionado será armazenado na Base de Casos.

Figura 4 – Ciclo do Raciocínio Baseado em Casos



Fonte: Adaptado de Wangenheim e Wangenheim (2003).

2.2.1 Representação de casos

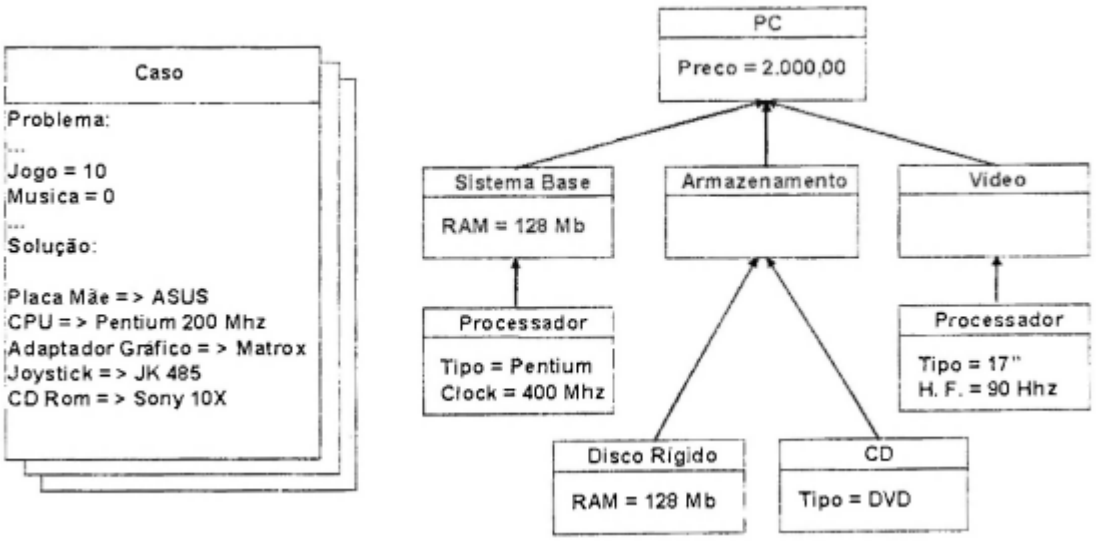
Um sistema de RBC é dependente da estrutura e conteúdo de sua coleção de casos. O problema de representação é primariamente o problema de decidir o que armazenar em um caso, encontrar uma estrutura apropriada para descrever o conteúdo dos casos e decidir como a memória de casos deve ser organizada e indexada para possibilitar a recuperação e a reutilização de soluções prévias (Policastro, 2004)

A representação de um caso é uma etapa importante, pois a partir dessa definição, todo o processo do RBC poderá encaminhar com sucesso.

O formato de representação dos casos e da BC deve facilitar a recuperação de

casos ocorridos em momentos apropriados. Memória Plana e Estrutura Hierárquica são duas das abordagens mais comuns. A escolha do modelo de memória para representação dos casos deve considerar alguns fatores: A representação utilizada para a base de casos; o propósito para o qual o sistema está sendo desenvolvido; O número e complexidade dos casos que serão armazenados; o número de atributos que serão utilizados para recuperação dos casos prévios; se algumas características são similares o suficiente para serem agrupadas; o conhecimento que se tem sobre o domínio da aplicação para determinar as similaridades entre os casos (Policastro, 2004)

Figura 5 - Representações de Casos



Fonte: Policastro, 2004.

2.2.2 Etapa de Recuperação

A etapa de recuperação é a etapa inicial de um ciclo Raciocínio Baseado em Casos (RBC), consiste em buscar o caso armazenado na base de casos mais similar ao novo problema/caso apresentado.

A similaridade entre casos consiste em um cálculo de geometria que mede a distância entre dois casos, sendo o caso atual apresentado com o caso mais similar recuperado da base de casos.

Segundo Urnau et al, 2014, para a etapa de Recuperação, o processo de similaridade é pré-definido com valores de 0 a 10, sendo do menos similar ao mais similar. O cálculo utiliza o valor da similaridade entre os atributos definidos para a consulta e multiplica pelo respectivo peso, calculando o somatório de todos os atributos.

Figura 6: Função da Similaridade do artigo de Urnau el al, 2014

$$Função\ de\ Similaridade\ (N,F) = \frac{\sum_{i=1}^{n} (N_i, F_i) * W_i}{\sum_{i=1}^{n} W_i}$$

Fonte: Urnau,Kipper e Frozza, 2014

Porém, uma métrica de similaridade conhecida e simples é a medida de distância euclidiana, que consiste em calcular a distância do novo caso com o caso que será recuperado. Uma abordagem simples com aprendizado de máquina é a utilização do algoritmo de K-Vizinhos Mais Próximos (K-Nearest Neighbour). O algoritmo de K-NN (K-Nearest Neighbour) basicamente faz a medida da distância do novo caso com algum outros K casos e obtendo os K vizinhos mais similares (próximos) ao caso atual.

2.2.3 Etapa de Reutilização

A etapa de reutilização é considerada a etapa mais desafiadora e complexa, nessa etapa ocorre o processo de adaptação. Quando um caso é recuperado da base de casos por ser o caso mais similar ao problema atual, existirá uma diferença mesmo que a similaridade seja alta, ou seja, os dois casos (atual e recuperado) podem ser de um grau de similaridade bastante alta, mas não significa que vão ser idênticos. Não existem problemas iguais ou casos iguais, por conta disso, na etapa de reutilização é necessário o processo de adaptação da solução do caso recuperado. A adaptação consiste em partir do ponto de uma solução previamente recuperada para a construção de uma nova solução.

2.2.4 Etapa de Revisão

Essa etapa tem como finalidade a revisão e avaliação da solução criada na fase de reutilização. Sendo essa revisão capaz de ser feita pelo usuário do sistema; por simulações; pela aplicação da solução no ambiente real.

Na fase de revisão é feita a avaliação se a solução criada está correta de acordo com o problema, ou se está errada. É de extrema importância a etapa de revisão, pois será um parâmetro para a avaliação do correto funcionamento do sistema.

2.2.5 Etapa de Retenção

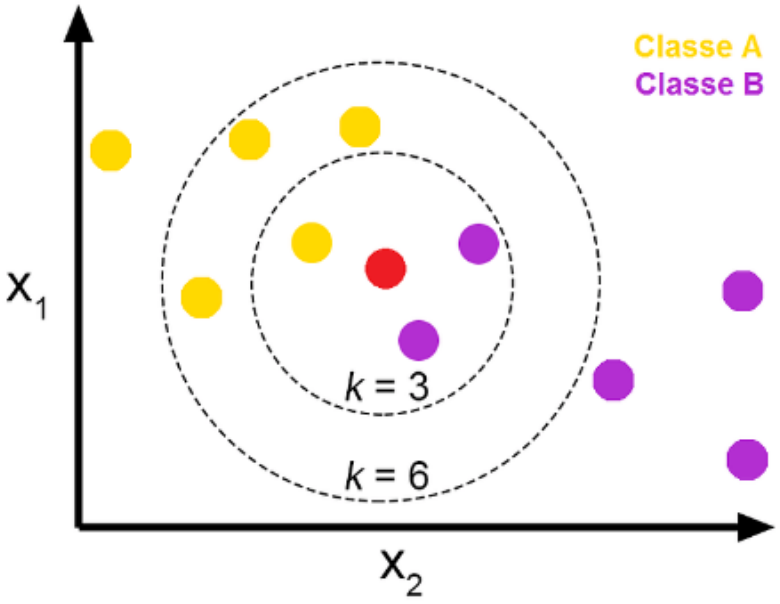
A etapa de retenção tem como finalidade a atualização da Base de Casos (BC) pois quando mais casos existirem na BC mais o sistema ficará aguçado em calcular a similaridade. Quanto mais casos adicionados na BC, melhora as soluções que o sistema pode oferecer. Essa etapa consiste em obter as soluções revisadas e adicionar na BC para que nos próximos casos obtenham soluções mais aprimoradas e com a melhor similaridade.

2.3 Algoritmo de k-nearest neighbors (k-nn)

KNN(K — Nearest Neighbors) é um dos muitos algoritmos (de aprendizagem supervisionada) usado no campo de data mining e machine learning, ele é um classificador onde o aprendizado é baseado “no quão similar” é um dado (um vetor) do outro. O treinamento é formado por vetores de n-dimensões. (MEDIUM, 2018)

O objetivo do Algoritmo é classificar um novo dado por meio do cálculo de sua distância e K-Vizinhos Mais Próximos com os outros dados das classes. A imagem abaixo mostra que um elemento da cor vermelha (sem classificação) quer saber de qual classe pertence.

Gráfico 1 - Gráfico exemplificando o Algoritmo de KNN, duas classes e um novo objeto a ser classificado, classificação entre classe A e classe B com K = 3 e K = 6



Assim que um dado chega para ser classificado, é calculado a sua distância com todos os outros dados já classificados, com isso teremos todas as distâncias (podendo ser calculadas pela métricas Euclidiana, Manhattan, Minkowski ou Ponderada) entre dois pontos, do dado atual com todos os outros dados. Com todas as distâncias já calculadas, é obtido os K - Vizinhos mais similares (de menor distância) ao dado. Na imagem utilizando o K = 3 o dado de cor vermelha será classificado como a classe B (roxa), porém se o K aumentar para 6 (K = 6) o dado de cor vermelha será classificado como a Classe A (amarela).

O algoritmo de K-NN retorna a classe com K-vizinhos que mais aparecem após o cálculo da distância. Com isso, ele será utilizado na etapa de recuperação de casos no sistema RBC.

2.3.1 Cálculo da distância

O cálculo de distância no algoritmo de K-NN pode ser feito de várias maneiras, sendo ele a distância Euclidiana, Manhattan, Minkowski ou Ponderada.

Porém é bastante comum a utilização da distância Euclidiana para o cálculo de distância entre dois pontos no algoritmo de K-Vizinhos Mais Próximos (K-NN).

A definição da distância Euclidiana entre dois pontos P = (p1,...pn) e Q = (q1,..qn) é dado pela fórmula.

Figura 7 - Fórmula matemática da distância euclidiana.

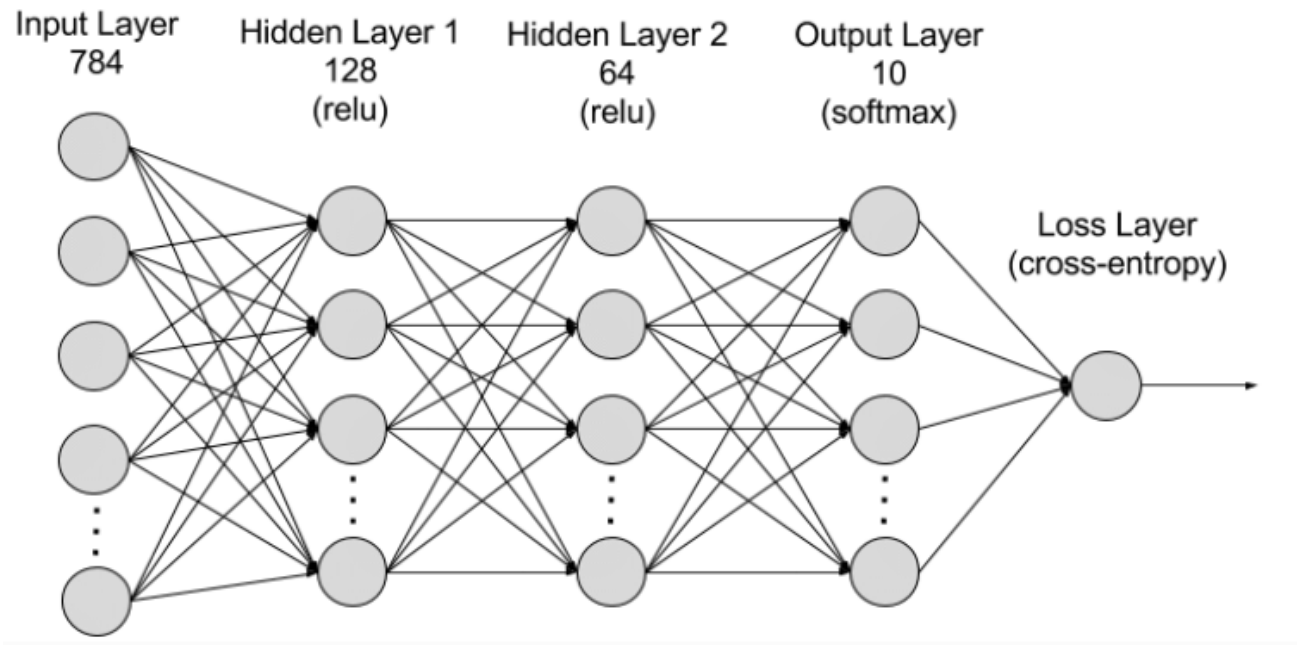
$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

2.4 Rede Neural (MLP)

Uma rede neural é um método de inteligência artificial que ensina computadores a processar dados de uma forma inspirada pelo cérebro humano. É um tipo de processo de machine learning, chamado aprendizado profundo, que usa nós ou neurônios interconectados em uma estrutura em camadas, semelhante ao cérebro humano. A rede neural cria um sistema adaptativo que os computadores usam para aprender com os erros e se aprimorar continuamente. As redes neurais artificiais tentam solucionar problemas complicados, como resumir documentos ou reconhecer rostos com grande precisão (AWS AMAZON, 2022)

As redes neurais artificiais (ANNs) são compostas por camadas de um nó, contendo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada nó, ou neurônio artificial, conecta-se a outro e tem um peso e um limite associados. Se a saída de qualquer nó individual estiver acima do valor do limite especificado, esse nó será ativado, enviando dados para a próxima camada da rede. Caso contrário, nenhum dado será transmitido junto à próxima camada da rede (IBM, 2022)

Figura 8 - Exemplificação de uma rede MLP



Uma Rede Neural (NN) é composta por uma camada de entrada, uma ou mais camadas ocultas que são passadas por funções de ativação (relu,softmax) e por camadas de saídas.

As redes neurais feedforward processam dados em uma direção, do nó de entrada para o nó de saída. Cada nó de uma camada está conectado a todos os nós da próxima camada. Com o passar do tempo, uma rede feedforward usa o processo de feedback para aprimorar as previsões. (AWS AMAZON,2022).

O trabalho de Ye, et al (2021) é um dos mais recentes e tem como foco resolver o processo de Reutilização de Casos que é considerado a etapa mais complexa do RBC. Com isso é abordado o método de Heurística de Diferença de Casos, com a sigla em inglês (CDH).

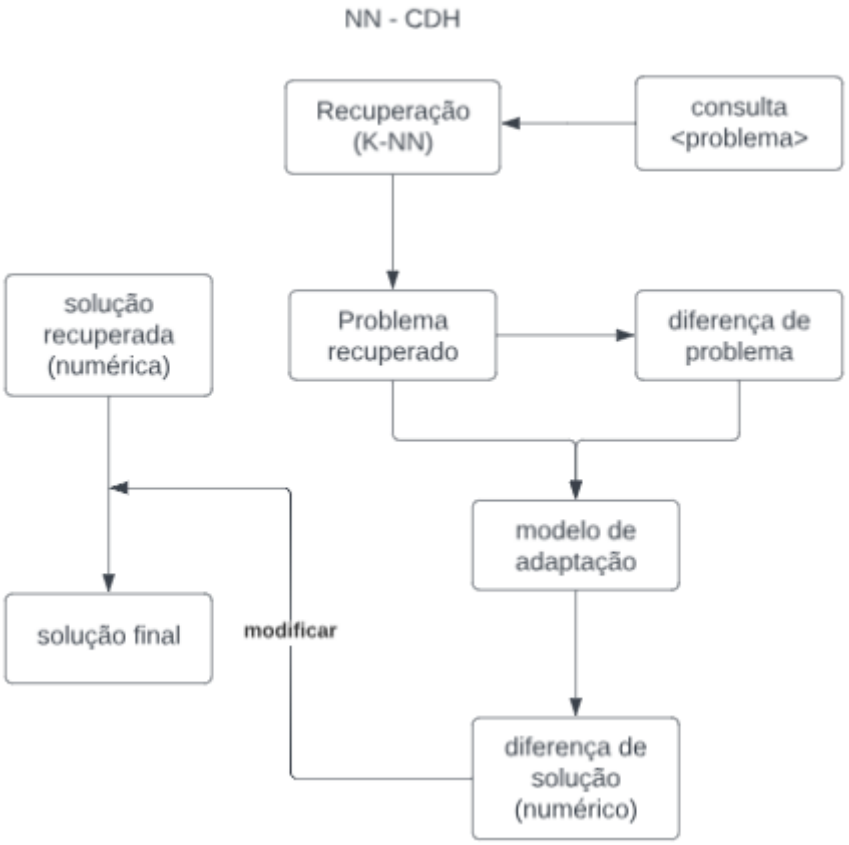
Um método popular para abordá-lo é a abordagem heurística de diferença de casos (CDH), que aprende adaptações de pares de casos com base nas diferenças de problemas e nas diferenças de soluções. A abordagem CDH tem sido frequentemente usada para gerar regras de adaptação, mas pesquisas recentes de RBC sobre regressão baseada em casos investigaram a substituição de regras de aprendizagem por modelos de rede de aprendizagem baseados em CDH para adaptação (Ye at el, 2021)

O artigo busca tratar o processo de reutilização de casos, mas também mostra como foi feita a etapa de recuperação de casos, que parte do princípio da similaridade. O autor mostra que dois principais modos de utilizar o algoritmo de K-NN para buscar a similaridade entre os casos.

Como é feito o aprendizado de diferença de pares de casos, o autor aborda que os casos a serem utilizados são os casos mais próximos, ou seja, os casos mais similares são usados para poder descobrir as diferenças entre o par de casos e assim setar uma nova regra de adaptação, aprender uma nova regra de adaptação de acordo com as diferenças. Para buscar o caso mais similar, segundo o autor, pode ser feito de duas maneiras, buscando um 1-NN (Vizinhos Mais Próximos com k = 1) e com com 3-NN (Vizinhos Mais Próximos com k = 3) calculando a média das classificações dos três casos mais semelhantes.

O fluxo da metodologia sobre o Classificador com Redes Neurais com abordagem do método de Heurística de Diferença de Casos (C-NN-CDH) é colocado pelo autor:

Figura 9: Fluxo de como uma abordagem de Heurística de Diferença de Casos para um sistema RBC funciona de acordo com o autor. Busca o caso mais similar com o KNN, faz a adaptação usando Heurística de Diferença de Casos, para obter regras de adaptação para pares de casos.



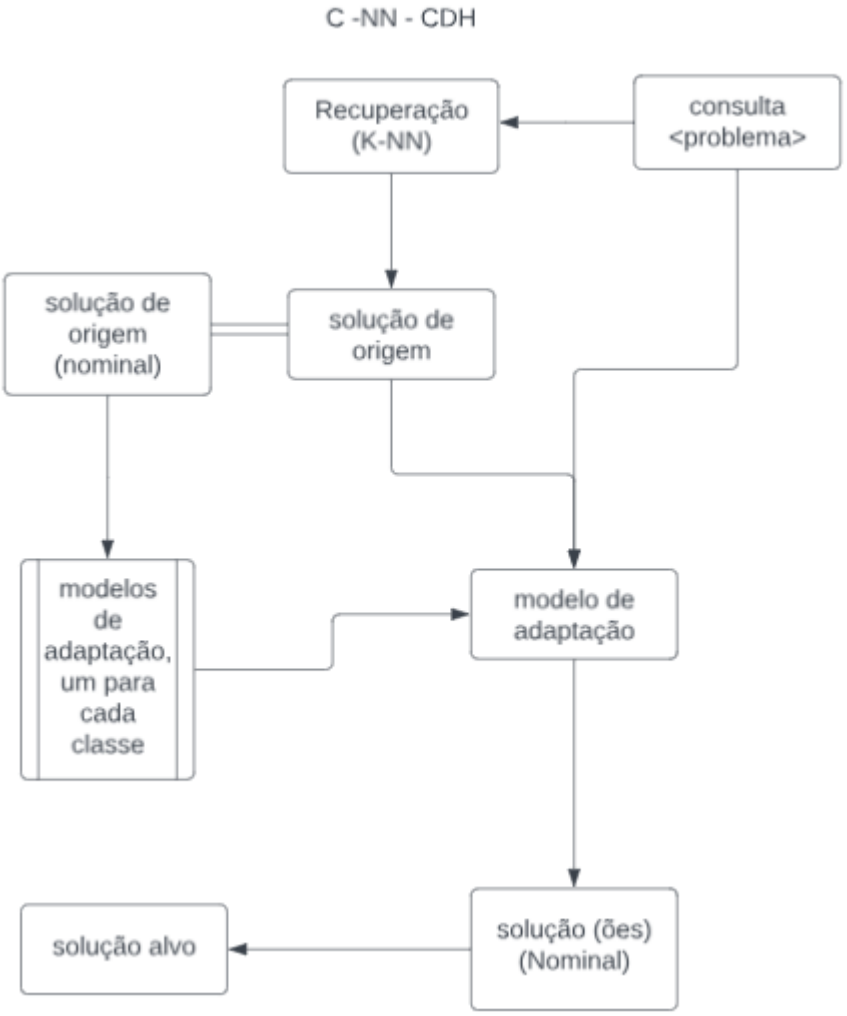
Fonte: Ye at el, 2021

NN-CDH e outros métodos CDH aprendem com pares de casos. Um dos pares é tratado como o caso fonte (com seu problema fonte e solução) e o outro como o caso alvo (com seu problema alvo e solução), onde a fonte deve ser adaptada ao alvo. Para simplificar, nos referimos a eles como um par de casos. Um método CDH aprende uma regra de adaptação para adaptar a solução do caso fonte para fornecer uma solução para o caso alvo. Para uma abordagem NN-CDH, o sistema RBC primeiro recupera um caso de origem semelhante à consulta (caso de destino) e calcula a diferença do problema entre o problema de origem e o problema de destino (Ye at el, 2021).

Com esse primeiro fluxo, a abordagem não é de um classificador ainda, mas uma aplicação da Heurística de Diferença de Casos (CDH) para o aprendizado de regras de pares de casos similares e logo depois colocados em uma rede neural.

A diferença do problema é então passada para uma rede neural, que é previamente treinada nas diferenças de problemas e soluções dos pares de casos de treinamento. A rede neural prevê a diferença de solução entre a solução de origem e a solução de destino.

Figura 10: Esse fluxograma detalha como será feito a classificação utilizando as abordagens passadas como o KNN e CDH, para depois que aprender regras de adaptação, poder classificar.



Fonte: Ye at el, 2021

O método substitui o cálculo tradicional da diferença CDH pelo cálculo implícito de uma técnica de aprendizado de máquina (por exemplo, rede neural), potencialmente levando em conta não apenas a diferença, mas o contexto do próprio caso fonte. Chamamos essa abordagem geral de tratamento de pares de casos como abordagem heurística de diferença de casos para classificação (“C-CDH”) (Ye at el, 2021).

Este trabalho tem como proposta usar os recursos da Inteligência Artificial, com o uso de técnicas e abordagem de um Sistema de Raciocínio Baseado em Casos (RBC) a partir de informações sobre pacientes que possuem ou não diabetes,com isso obter um

diagnóstico de um novo paciente com base na adaptação dos casos similares de pessoas que possui ou não diabetes. A abordagem de um Sistema de Raciocínio Baseado em Casos é utilizar de informações ou experiências passadas para poder resolver novos problemas. Segundo Ganascia (1997) o Sistema de Raciocínio Baseado em Casos (RBC) tem como objetivo solucionar novos problemas com base na adaptação e solução de problemas anteriores que são similares.

A abordagem Heurística de Diferença de Casos (CDH) é um dos métodos mais utilizados para aprender conhecimento de adaptação. A abordagem CDH pega pares de casos da base de casos e de cada par aprende uma regra para adaptar um caso a outro.

Como exemplo simplificado, considere a aplicação do RBC à previsão de preços de apartamentos. Suponha que dois apartamentos A e B sejam muito semelhantes, exceto que A tem piso acarpetado, enquanto B tem piso de madeira, e que o aluguel de B é $ 200 a mais. Ao comparar os apartamentos A e B, uma abordagem CDH pode aprender a regra de que mudar do carpete para o piso de madeira aumenta o aluguel de um apartamento em US$ 200. Uma questão para as abordagens CDH é qual generalização aprender com os pares de casos: em vez de aprender uma diferença absoluta, uma abordagem CDH pode aprender a variação percentual ou outra caracterização da diferença observada (Ye et al, 2021)

Um método CDH aprende uma regra de adaptação para adaptar a solução do caso fonte para fornecer uma solução para o caso alvo. Para uma abordagem NN-CDH, o sistema RBC primeiro recupera um caso de origem semelhante à consulta (caso de destino) e calcula a diferença do problema entre o problema de origem e o problema de destino. A diferença do problema é então passada para uma rede neural, que é previamente treinada nas diferenças de problemas e soluções dos pares de casos de treinamento. A rede neural prevê a diferença de solução entre a solução de origem e a solução de destino. Finalmente, o sistema RBC aplica a diferença de solução prevista à solução de origem e utiliza o resultado adaptado como previsão final (Ye et al, 2021)

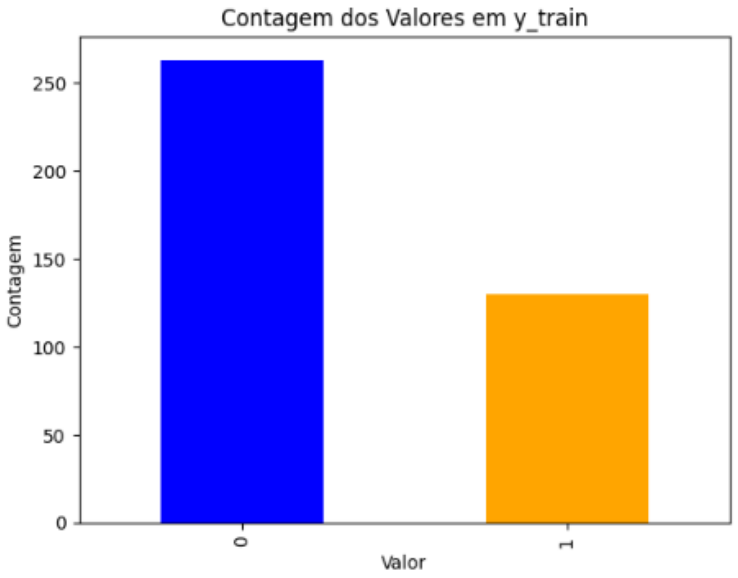
3 RESULTADOS

O resultado para o sistema RBC para a etapa de Recuperação usando o item 3.3 Algoritmo de k-nearest neighbors (k-nn) para buscar o caso mais similar na BC e o uso do item 3.4 Rede Neural para a adaptação de casos, estão descritos abaixo:

3.1 Balanceamento de classes

Antes de utilizar a técnica de balanceamento, foi feito um comparativo de classes para ilustrar o desbalanceamento.

Gráfico 2 - O Gráfico mostra o desbalanceamento de da classe 0 e a 1.



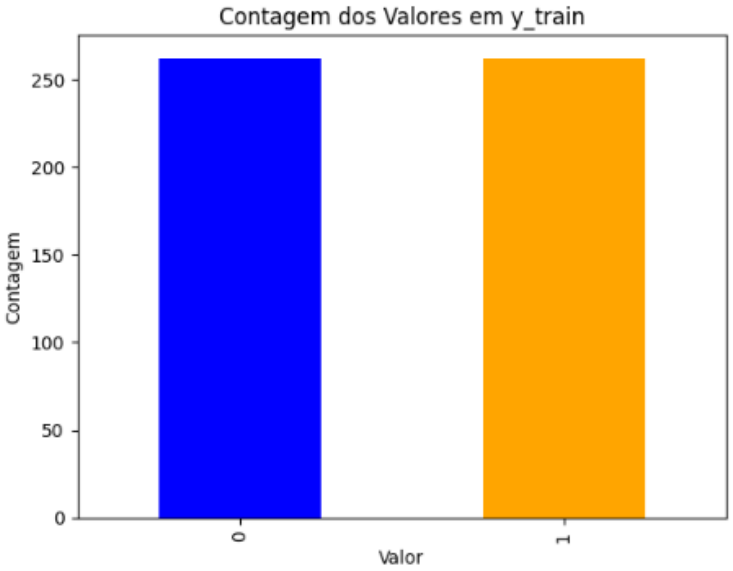
Fonte: autoria própria

Na imagem temos 262 casos para a Classe 0 e 130 casos para a Classe 1, com isso, temos uma proporção de 2,02:1 de classe 0 para classe 1. No entanto, não é considerado um desbalanceamento grave, já que a proporção é de 2:1, mas feito o balanceamento tivemos ótimos resultados para o algoritmo K-NN para etapa de Recuperação de Casos.

Na base de dados deste trabalho, foi utilizado o método de Oversampling, que consiste em utilizar o algoritmo de KNN para recuperar exemplos similares da classe minoritária, com isso cada exemplo recuperado é comparado com a classe minoritária e é criado um exemplo sintético.

Após a aplicação da técnica Oversampling de balanceamento de classes, tivemos a seguintes classes balanceadas:

Gráfico 3 - Classes 0 e 1 balanceadas após a aplicação da técnica de Oversampling



Fonte: autoria própria

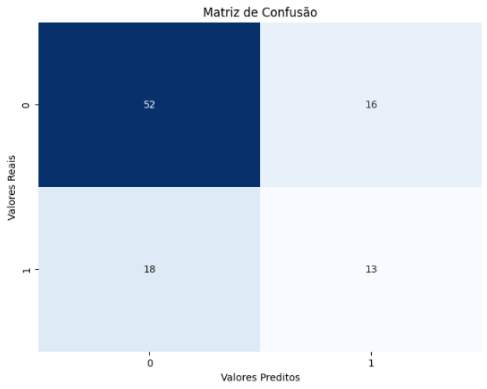
Com essa aplicação de balanceamento, obtivemos 262 casos para a Classe 0 e 262 casos para a classe 1, tendo uma proporção de 1:1.

3.2 Recuperação usando K-NN para similaridade

Com o uso da técnica de balanceamento de classes, o algoritmo de K-NN teve uma performance melhor de assertividade na recuperação de casos mais similares.

Para a classe desbalanceada, tivemos uma matriz de confusão um pouco mais confusa, com poucos acertos nos casos da Classe 1.

Figura 11: Matriz de confusão do algoritmo de KNN a partir do desbalanceamento de classes.



Fonte: autoria própria

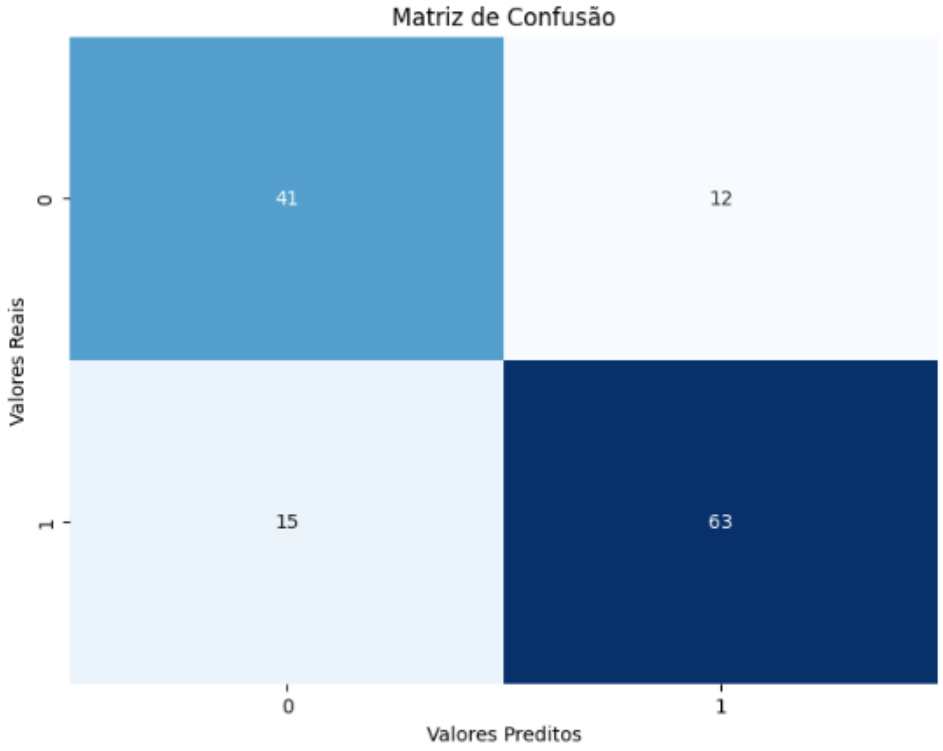
Com as classes desbalanceadas, o algoritmo de K-NN teve uma acurácia

aproximadamente de 0,6836 para o K = 3, com intuito de classificação.

Para a etapa de Recuperação de Casos usando o K-NN no RBC, o algoritmo de K-NN teve uma acurácia de 0.6734 para o K = 1, com intuito de buscar o caso mais similar.

Com base nisso, os resultados a partir do balanceamento de classes, teve um desempenho bastante superior ao desbalanceamento.

Figura 12: Matriz de confusão do algoritmo de KNN a partir do balanceamento de classes.



Fonte: autoria própria

Com a aplicação da técnica de balanceamento, o algoritmo de K-NN pode ter resultados melhores de acurácia na parte de classificação e de recuperação de casos no RBC.

Para a classificação, o algoritmo de K-NN com o K=3 tivemos uma acurácia de 0.7938. E para a Recuperação de Casos no RBC, a acurácia foi de 0.8854 com o K=1 buscando a similaridade.

Tabela 1 - Tabela de comparação de acurácia do algoritmo de K-NN com as classes balanceadas e não balanceadas.

Acurácia	K = 3 (Classificação)	K = 1 (Recuperação de Casos - Caso mais Similaridade)
Desbalanceada	0,6836 (68,3%)	0,6734 (67,3%)
Balanceada	0,7938 (79,3%)	0,8854 (88,5%)

Fonte: autoria própria

Com nossa base de dados treinada com o algoritmo de K-NN e possuindo as duas classes sendo elas diabetes ou não, podemos classificar novos dados/casos assim que for consultado. Com isso podemos realizar a etapa de recuperação no sistema de RBC, pois o algoritmo retorna a classificação a partir dos K-Vizinhos Mais Próximos, porém a métrica que buscamos é saber qual é o Vizinho Mais Próximo ao novo caso que temos no sistema RBC. O valor da distância é o ponto principal da similaridade, o menor valor ou a distância mais curta é o objetivo de caso ser o mais similar com o caso atual.

Obtendo o caso mais similar, teremos dois casos parecidos mas não idênticos, por isso entra a etapa de adaptação de casos

3.3 Adaptação de casos usando heurística de diferença de casos (CDH) e redes neurais.

A etapa de reutilização tem um processo chamado de Adaptação de casos, que é considerado o processo mais complexo para se resolver, porém no Estado da Arte temos abordagem de heurística de diferença de casos (CDH), que é a mais atual usada e com melhor desempenho.

Com o processo de calcular a diferença entre dois casos para fazer a adaptação da solução, assim que buscado da base de dados o caso mais similar com a ajuda do algoritmo de K-NN, teremos dois casos para analisar, o caso atual em busca de uma solução e o outro é o caso recuperado com a maior similaridade.

A abordagem de heurística de diferença de casos (CDH) consiste em calcular a diferença de um par de casos similares e usar a rede neural para aprender com essas diferenças, o aprendizado de diferenças será feito para associar as regras de adaptação de um sistema de RBC.

Após calcular a diferença de cada par de casos, é colocado em uma rede neural

feedforward com 2 camadas ocultas (128 e 64 nós com funções de ativação ReLU) e uma camada de saída com função de ativação softmax. A função de perda é entropia cruzada categórica e o modelo é otimizado usando Adam (Ye et al, 2021).

Para efeito de comparação, um classificador de rede neural é implementado com a mesma configuração. Observe que a rede neural de adaptação é um componente do sistema RBC (C-NN-CDH) e produz a classificação final com base em um caso recuperado e na consulta, enquanto o classificador de rede neural produz diretamente a classificação final com base apenas na consulta (Ye et al, 2021).

A rede neural de adaptação é uma rede feedforward com 2 camadas ocultas (128 e 64 nós com funções de ativação ReLU) e uma camada de saída com função de ativação softmax. A função de perda é entropia cruzada categórica e o modelo é otimizado usando Adam.

Tabela 2 - Tabela de acurácia para as etapas de validação do algoritmo de KNN, Recuperação de Casos mais similares e Processo de Adaptação de Casos.

KNN (K=3)	RBC Recuperação (KNN) K=1	Adaptação
0.7922 (79,2%)	0.7402 (74%)	0.8051 (80%)

4 Conclusão

O método de calcular a similaridade com o algoritmo de K-NN é bastante usado e teve uma acurácia de 0.80

Um sistema RBC para resolver novos problemas depende da sua capacidade de adaptar soluções anteriores às novas circunstâncias.

A abordagem heurística de diferença de caso é um método de conhecimento leve para aprender conhecimento de adaptação

Para trabalhos futuros, concluir a etapa de revisão e retenção.

REFERÊNCIAS

Ahsan, M. M., Luna, S. A. , Siddique, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3):541. doi: 10.3390/healthcare10030541. PMID: 35327018; PMCID: PMC8950225.

BEGUM, S. et al. Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5586661>. Acessado em: 20 junho. 2023.

BICHINDARITZ,I; MARLING, C. Case-Based Reasoning in the Health Sciences: Foundations and Research Directions. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-14464-6_7> Acessado em: 21 junho. 2023

BROWNLEE, J. Regression Tutorial with the Keras Deep Learning Library in Python. Disponivel em: <https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/>. Acessado em: 23 outubro. 2023.

BROWNLEE, J. How to Develop a Cost-Sensitive Neural Network for Imbalanced Classification. Disponível em: <https://machinelearningmastery.com/cost-sensitive-neural-network-for-imbalanced-classification/> Acessado em: 23 outubro. 2023.

IDF Diabetes Atlas. Disponível em: <https://diabetesatlas.org/atlas/tenth-edition/>. Acesso em: 20 outubro. 2023.

JOSÉ, I. KNN (K-Nearest Neighbors). Disponibel em: <https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e> Acessado em 21 setembro. 2023.

KAHNEMAN, D. (2021) Ruído: Uma falha no julgamento humano 1ª Edição.

Nia, G., N., Kaplanoglu, E., Nasab, A. Evaluation of artificial intelligence techniques in

disease diagnosis and prediction. Discov Artif Intell 3, 5 (2023). https://doi.org/10.1007/s44163-023-00049-5

POLICASTRO, C. Estratégias de adaptação de casos em sistemas de raciocínio baseado em casos. Disponível em: <https://bdtd.ibict.br/vufind/Record/USP_72b44ece6a62ca3543eb04a288be9e11> Acessado em: 10 maio. 2023.

Popa, A., Wood, W., Application of case-based reasoning for well fracturing planning and execution, Journal of Natural Gas Science and Engineering, v. 3, n. 6, p. 687-696, ISSN 1875-5100, 2011, https://doi.org/10.1016/j.jngse.2011.07.013.

RAMOS, T. Uso de Machine Learning para predição de pacientes com Diabetes Mellitus. Disponível em: <https://thiagoramos20042.medium.com/uso-de-machine-learning-para-predi%C3%A7%C3%A3o-de-pacientes-com-diabetes-mellitus-426a063cb121>. Acesso em: 22 outubro. 2023.

Santos, M. K., Ferreira Júnior, J. R., Wada, D. T, Tenório, A. P. M., Barbosa, M. H. N., Marques, P. M. de A., Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine, Radiol Bras. 2019 Nov/Dez;52(6):387–396

Sharma, M., Sharma, C., A Review on Diverse Applications of Case-Based Reasoning. In: Sharma, H., Govindan, K., Poonia, R., Kumar, S., El-Medany, W. (eds) Advances in Computing and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore, 2020. https://doi.org/10.1007/978-981-15-0222-4_48.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., Gandomi, A. H., Machine learning in medical applications: A review of state-of-the-art methods, Computers in Biology and Medicine, v. 145, 2022, 105458, ISSN 0010-4825,
https://doi.org/10.1016/j.compbiomed.2022.105458.

URNAU, E. et al. Desenvolvimento de um sistema de apoio à decisão com a técnica de raciocínio baseado em casos. Disponível em: <https://www.scielo.br/j/pci/a/QPLpXjmkCySpqXZVxbswGYs/>. Acessado em: 8 maio. 2023.

VAZ, A. Como lidar com dados desbalanceados em problemas de classificação. Disponivel em: <https://medium.com/data-hackers/como-lidar-com-dados-desbalanceados-em-problemas-de-classifica%C3%A7%C3%A3o-17c4d4357ef9> Acessado em: 27 outubro. 2023.

WANGENHEIM, C. (2002) Raciocínio Baseado em Casos 1ª Edição.

YE, X. et al. Learning Adaptations for Case-Based Classification: A Neural Network Approach. Disponivel em: <(PDF) Learning Adaptations for Case-Based Classification: A Neural Network Approach (researchgate.net)>. Acessado em: 9 maio. 2023.