UNIVERSIDADE FEEVALE

GABRIEL EDUARDO MARTINI

O APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO STATUS GLICÊMICO DE PACIENTES

Novo Hamburgo

2020

GABRIEL EDUARDO MARTINI

O APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DO STATUS GLICÊMICO DE PACIENTES

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale

Orientador: Prof. Dr. Rodrigo Rafael Villarreal Goulart

Novo Hamburgo

2020

AGRADECIMENTOS

Agradeço a todos que, de alguma maneira contribuíram para a realização deste trabalho e chegada até este estágio da graduação, em especial:

A meus pais Ivan e Tania e ao meu irmão Cássio, que sempre incentivaram a priorização e dedicação aos estudos.

Ao professor Dr. Rodrigo, por todas as sugestões e colaborações antes e durante o desenvolvimento deste trabalho.

Muito Obrigado!

RESUMO

A Organização Mundial da Saúde afirma que 1 em cada 11 pessoas do mundo tem diabetes, mas somente 50% dos pacientes conhecem seu diagnóstico. O exame laboratorial mais solicitado pelos profissionais é o hemograma, que não apresenta o nível glicêmico do paciente, indicador da diabetes. Portanto é levantada a hipótese de identificar pacientes com potencial de diabetes a partir de dados de exames correlatos. Esta situação apresenta um grande desafio: detectar ou classificar o índice glicêmico do paciente através de dados hematológicos. Nesse contexto, sistemas de apoio à decisão clínica têm demonstrado alto grau de assertividade no auxílio ao diagnóstico. A tecnologia com maior relevância da atualidade é a Inteligência Artificial, com foco principal no Aprendizado de Máquina, implementando conceitos computacionais que possibilitam aprendizado automatizado por meio de dados pré-existentes, o que se apresenta viável no estudo das publicações correlatas a esta pesquisa. Com foco no diagnóstico da diabetes, a pesquisa apresenta a implementação de métodos de aprendizado de máquina para classificação do status glicêmico de pacientes a partir de dados hematológicos. Através do uso de máquina de vetores de suporte, foi desenvolvido com base em trabalhos correlatos, um classificador capaz de indicar o status glicêmico dos pacientes a partir das variáveis hematológicas, com sua acurácia apurada através de validação cruzada de 10 vezes. Os resultados foram comparados com pesquisas correlatas e considerações acerca da aplicação da inteligência artificial na área do diagnóstico são apontadas.

Palavras-chave: Inteligência artificial. Aprendizado de Máquina. Diagnóstico. Classificação Glicêmica.

ABSTRACT

The World Health Organization states that 1 in 11 people in the world have diabetes, but only 50% of patients know their diagnosis. The laboratory test most requested by professionals is the blood count, which does not show the patient's glycemic level, an indicator of diabetes. Therefore, the hypothesis of identifying patients with diabetes potential is raised based on data from related tests. This situation presents a great challenge: to detect or classify the patient's glycemic index through hematological data. In this context, clinical decision support systems have demonstrated a high degree of assertiveness in assisting diagnosis. A technology with greater relevance today is Artificial Intelligence, with a main focus on Machine Learning, implementing computational concepts and enabling automated learning through pre-existing data, or it is viable in the study of publications related to this research. With a focus on the diagnosis of diabetes, research presents the implementation of machine learning methods for classifying the glycemic status of patients based on hematological data. Through the use of the support vector machine, a classifier capable of indicating the glycemic status of patients based on hematological variables was developed based on related work, with its precision after 10 folds cross-validation. The results were compared with related researches and considerations on the application of artificial intelligence in the diagnosis area are pointed out.

Keywords: Artificial intelligence. Machine Learning. Diagnosis. Glycemic classification.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ACC	Acurácia
AA	Análise de Associação
AM	Aprendizado de máquina
AUC	*Area Under the Curve*
CSV	*Comma Separated Values*
CHCM	Concentração de Hemoglobina Corpuscular Média
DM	Diabetes Mellitus
DM1	Diabetes Mellitus Tipo 1
DM2	Diabetes Mellitus Tipo 2
RI	Resistência à ação da insulina no corpo
RBC	Hemácias
HCT	Hematócrito
Hb	Hemoglobina
HCM	Hemoglobina Corpuscular Média
A1c	Hemoglobina Glicada
ID	Identificação
IEEE	*Institute of Electrical and Electronic Engineers*
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*
WBC	Leucócitos
SVM	Máquina de vetores de suporte
AND	Operador lógico E
PLT	Plaquetas
RF	*Random Forest*
RDW	*Red Cell Distribution Width*
RNA	Rede Neural Artificial
SMO	*Sequential minimal optimization*
VPM	Volume Plaquetário Médio
VCM	Volume Corpuscular Médio
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

Os avanços tecnológicos das últimas décadas vêm trazendo contribuições importantes para a precisão do diagnóstico médico, refletindo em benefícios para os pacientes. Este acelerado processo de inovação resulta em maior segurança na tomada de decisão clínica, tanto em casos agudos ou urgentes, quanto em doenças crônicas. Dentro desta nova realidade, torna-se facilitado o acesso a diversos procedimentos de análise e apoio ao diagnóstico, em especial, os exames laboratoriais. A Medicina Laboratorial possui papel fundamental na tomada de decisão por parte dos médicos pois traz indicadores confiáveis sobre o estado de saúde do paciente. Exames laboratoriais são ferramentas eficazes para reduzir as incertezas da prática clínica, contribuir para a preservação e restauração da saúde e melhorar a qualidade dos cuidados de saúde (ANDRIOLO, 2008).

Diversas patologias comuns na população deixam de ser diagnosticadas ou são diagnosticadas de forma tardia devido à falta de requisições adequadas de exames clínicos pelos médicos. É sabido que diagnósticos tardios aumentam o risco de disseminação das doenças e levam a complicações, tornando o tratamento mais difícil (SHCOLNIK, 2018).

A Organização Mundial da Saúde acredita que 1 em cada 11 pessoas no mundo tem diabetes. Segundo o *International* Diabetes *Federation* (IDF), o Brasil é o quarto país com mais diabéticos no mundo, chegando a 7% da população, e o desafio passa pela falta de controle glicêmico dos pacientes: 50% dos diabéticos desconhecem o diagnóstico (SBAC, 2018). A diabetes é uma doença metabólica que pode ser classificada em duas categorias: Tipo 1 e Tipo 2. A primeira é uma forma de diabetes relacionada ao sistema autoimune, em geral identificada na infância ou adolescência. Na diabetes tipo 1, as células responsáveis pela defesa do organismo acabam atacando outras, capazes de sintetizar insulina, por causa de um defeito no sistema imunológico. Na diabetes do tipo 2, o organismo não produz insulina suficiente para controlar a taxa de açúcar no sangue, sendo este o tipo que concentra a maior quantia de incidência da doença em pacientes (SBAC, 2018).

Na área do diagnóstico através de exames laboratoriais, o parâmetro mais solicitado pelos profissionais da medicina é o Hemograma, seja para pacientes ambulatoriais ou hospitalizados (CONTI, 2018). Segundo Kawamoto et al. (2013), parâmetros hematológicos, incluindo contagem de glóbulos vermelhos, hematócrito e

hemoglobina estão associados à resistência à insulina, o que caracteriza a diabetes tipo I. Em estudo publicado por Kutlu et al. (2008), afirma-se que existe uma relação significativa entre os níveis de hemoglobina e funções celulares em pacientes com diabetes tipo 2.

Oportunizar melhorias no diagnóstico de uma doença que faz parte da rotina de um grande número de pessoas torna-se uma contribuição relevante no meio científico. Com um exame laboratorial de grande número de requisições médicas e suas correlações com variáveis hematológicas fundamentadas por achados científicos, oportuniza-se a classificação do status glicêmico destes pacientes através do uso de tecnologias da computação, possibilitando a entrega de diagnóstico a pacientes que, caso dependessem da realização de um exame específico, desconheceriam seu diagnóstico.

Em busca de trabalhos correlatos, encontra-se uma comunidade científica aquecida em trabalhos que visam o apoio no diagnóstico da diabetes. Apesar disso, não foram encontradas publicações que apresentem o desenvolvimento de modelos classificadores de status glicêmicos somente com o uso de variáveis hematológicas, trazendo exclusividade para a pesquisa que está sendo proposta neste trabalho.

As práticas do setor de saúde são baseadas em informações adquiridas de forma sistemática e constante, necessitando métodos de apoio na organização e formação do conhecimento com base nessas informações. Em análise de pesquisas correlatas na área da saúde, é perceptível certa dificuldade dos autores para a obtenção de grandes volumes de dados para uso em pesquisa, principalmente tratando de bases de dados com resultados de exames clínicos. Para o adequado desenvolvimento desta pesquisa, o autor firmou parceria com a empresa Laboratório Bom Pastor LTDA, situada na cidade de Igrejinha – RS, que disponibilizou uma base de dados de resultados de exames laboratoriais, necessária para a realização do estudo. Dessa forma, a etapa de coleta dos dados possui como objetivos a extração e compreensão da base de resultados de exames, bem como a realização de tratamentos de pré-processamento dos dados.

A tecnologia de finalidade geral com maior relevância nessa era é a Inteligência Artificial (IA), com foco principal no Aprendizado de Máquina (AM) (BRYNJOLFSSON, 2017). O AM constitui-se de programação computacional de análise de dados para a construção de modelos analíticos. Esta programação baseia-se no conceito de que o sistema computacional pode aprender com dados, identificar padrões e tomar

decisões com o mínimo de intervenção humana. A “máquina” pode ser protagonista no seu aprendizado e capaz de se adaptar sozinha quando exposta a novos dados (OSAKI, 2018).

Sistemas computadorizados de apoio à decisão clínica têm indicado um alto grau de acurácia em suas propostas diagnósticas. Usando algoritmos, estratégias de tomada de decisão e um volume de dados considerável, sistemas de Inteligência Artificial são capazes de propor ações, entender e classificar informações. O Aprendizado de Máquina pode oferecer indicadores de risco e de implicações da correlação entre diagnóstico e terapia, dados que poderão ser posteriormente confirmadas por estudos randomizados e controlados em uma amostra de pacientes (LOBO, 2018).

É notável o aumento e a popularização do uso de práticas de Inteligência Artificial em diversas áreas. De acordo com o relatório Tendências da Tecnologia, divulgado pela Organização Mundial de Propriedade Intelectual (OMPI) em 31 de janeiro de 2019, 50% de todas as patentes para Inteligência Artificial foram publicadas desde 2013, somando mais de 170 mil ideias (ONU – Organização das Nações Unidas, 2019).

Diante deste contexto, esta pesquisa traz como objetivo geral a coleta e análise de dados de resultados de exames laboratoriais para o pré-processamento e desenvolvimento de um modelo de classificador de status glicêmico de pacientes através da aplicação de algoritmos de aprendizado de máquina. Como dado de treinamento para a correta classificação das informações, realizou-se a correlação dos dados hematológicos com o real status glicêmico desses pacientes. Ao final, a validação dos resultados obtidos foi realizada através do consolidado método de *Cross-validation* de 10 *folds*.

Para conhecimento do leitor, na sequência deste documento o capítulo 2 apresenta a contextualização do ambiente do problema e dos termos técnicos da área da saúde utilizados no desenvolver do projeto e os trabalhos relacionados ao tema da pesquisa. O capítulo 3 apresenta o contexto dos dados a serem utilizados e a exploração das informações que compõem o conjunto de dados. No capítulo 4, são apresentados os métodos utilizados e o modelo classificador proposto, com os resultados obtidos. Por fim, são expostas considerações sobre a pesquisa e os tópicos futuros a este trabalho.

2 CORRELAÇÕES COM A ÁREA DA SAÚDE

A interligação entre tecnologia e saúde está em constante processo de fortalecimento. O uso de novas informações sobre saúde e tecnologias tem o potencial de reduzir o custo e melhorar a pesquisa e os resultados em saúde. Essas inovações podem oferecer suporte ao monitoramento contínuo da saúde, possibilitando a prevenção ou o diagnóstico de patologias (KUMAR, NILSEN, 2014). Nessa interligação de áreas de conhecimento, torna-se necessário que profissionais da área da computação tenham embasamento sobre termos técnicos e conceitos da área da saúde, utilizados no decorrer deste trabalho. Este capítulo traz um referencial teórico sobre conceitos relevantes da área da saúde, que serviram como base para desenvolvimento da pesquisa.

2.1 DIABETES

Segundo Varella (2019), Diabetes Mellitus (DM) não se trata de uma doença única, mas de uma composição de doenças com características em comum: concentrações de glicose aumentadas no sangue, que podem ser provocadas por diferentes situações. Esse conjunto de doenças caracteriza-se por hiperglicemia, resultante de defeitos no metabolismo da insulina no organismo. A hiperglicemia quando na forma crônica do diabetes associa-se a danos a longo prazo, como disfunção e falha de órgãos, como rins, olhos, nervos, coração e vasos sanguíneos (ADA, 2020). Ao ser diagnosticada, a Diabetes é classificada em dois grupos: Tipo 1 e Tipo 2.

O Tipo 1, também chamado de Diabetes Imunoimediado, caracteriza-se pelos indivíduos onde a causa da doença é uma deficiência absoluta na secreção de insulina. Essa classificação de diabetes atinge de 5% a 10% das pessoas que contém a doença e é caracterizada pela destruição autoimune das células β do pâncreas. As células β (beta) são responsáveis por secretar a insulina no pâncreas, regulando os níveis de glicose no sangue (GALLEGO, 2013). Segundo Delves (2018), em indivíduos normais, o sistema imunológico reage somente aos antígenos de substâncias estranhas ou perigosas e não aos antígenos das substâncias produzidas no próprio corpo humano. Porém, em alguns indivíduos com disfunções, o sistema imunológico produz células que atacam outros tipos de células do próprio organismo. Essa resposta defeituosa é denominada reação autoimune. Pessoas com maior risco

para o desenvolvimento deste tipo de diabetes podem ser identificadas facilmente por processo patológico ocorrendo no pâncreas e em marcadores genéticos.

O grupo Tipo 2 é mais frequente na população, abrangendo 90% a 95% das pessoas que contém a doença, composto pelos indivíduos onde a causa é uma composição de resistência à ação da insulina no corpo (RI) e a secreção de insulina inadequada. A RI é a patologia para quando o indivíduo possui resposta biológica reduzida à uma concentração de insulina, fazendo com que essa substância perca a eficiência no corpo e as células comecem a resistir à insulina, ou seja, ignorá-la, fazendo com que este paciente apresente níveis de glicose elevados (FONSECA et al., 2018). Essa forma de diabetes pode permanecer por anos no indivíduo sem apresentar sintomas porque o aumento nos níveis de glicose ocorre gradualmente e, em fases iniciais, as quantidades de glicose no sangue não são graves o suficiente para que sejam percebidos sintomas da doença. Pessoas com idade avançada, baixa atividade física ou obesos possuem maior risco na contração da diabetes. Segundo a ADA (2020), defeitos genéticos na função das células β do pâncreas também podem ser causadores de hiperglicemia, caracterizando a diabetes. O aumento de glicose, chamado Hiperglicemia, é a condição que ocorre quando há pouca insulina no organismo, ou quando o corpo não consegue usá-la apropriadamente. É o aumento dos níveis de glicose no sangue (SBD, 2017).

Glicose é um tipo de açúcar que atua como a principal fonte energética do corpo humano. Os carboidratos que ingerimos são digeridos até a obtenção de glicose e absorvidos pelo intestino delgado para a corrente sanguínea. A grande maioria das células precisa da glicose em níveis adequados para o seu funcionamento normal. A insulina age como um controlador dos níveis da glicose, transportando-a para dentro das células na quantidade adequada e encaminhando o excesso para armazenamento no corpo. Portanto, para uma vida saudável, a glicose e a insulina devem estar em equilíbrio (SBPC, 2020).

A Hemoglobina é uma proteína transportadora de oxigênio, presente nos glóbulos vermelhos do nosso sangue. Diversos tipos de hemoglobina estão presentes na corrente sanguínea, mas a forma predominante é a hemoglobina A e suas frações, que representa 95% a 98% do total de hemoglobina circulante. Uma parte da glicose que circula em nosso sangue forma ligações espontâneas com a hemoglobina, gerando moléculas nomeadas de Glicadas, sendo que quanto maior a concentração de glicose no sangue, mais hemoglobina glicada se forma. Uma vez glicada, a glicose

continua ligada até o resto da vida da hemácia, que dura cerca de 120 dias (SBPC, 2020).

Diversos procedimentos de diagnóstico associam-se à diabetes e seus estágios. Segundo a Associação Americana de Diabetes (2020), a dosagem de Hemoglobina Glicada (A1c) é o procedimento de padrão-ouro para diagnosticar diabetes em indivíduos com fatores de risco e para estimativa de riscos do desenvolvimento da doença no futuro, devido à essa dosagem estimar a taxa média de ligação da glicose na hemoglobina nos últimos três meses, o que se torna um indicador confiável para a confirmação dos níveis glicêmicos do paciente. Diante de um parâmetro considerado confiável para o diagnóstico da diabetes, verificamos que o laboratório parceiro deste estudo dispõe da realização deste exame, que será adicionado ao conjunto de dados da pesquisa para uso como a classe-alvo do modelo classificador.

A dosagem de Hemoglobina Glicada mensura o percentual de ligação da glicose à hemoglobina no sangue do paciente. A American Diabetes Association (ADA, 2020) determina que o valor de A1c possa ser interpretado em três faixas distintas, representadas na Figura 1 abaixo.

Figura 1 - Classificação de Status Glicêmico baseado no valor de A1c

Fonte: ADA (2020)

Em interpretação à Figura 1, podemos determinar que o resultado de A1c de um paciente pode ser classificado em três níveis. Se incluem no nível 1 os indivíduos que possuem valor de A1c abaixo de 5,7%, rotulados como normais. O segundo nível é composto pelos indivíduos na zona pré-diabetes, aqueles com A1c superior a 5,7% e inferior a 6,5%. O nível dois caracteriza aqueles pacientes que possuem tendência ou risco a desenvolverem diabetes. O terceiro nível, identificando indivíduos com diagnóstico confirmado da diabetes, é composto pelos resultados de A1c maiores que 6,5%.

Os resultados do exame de Hemoglobina Glicada, adicionados à base de dados utilizada neste estudo trarão um registro do real status glicêmico de cada paciente registrado nos dados, variável que será utilizada como o valor alvo do classificador a ser construído.

2.2 HEMOGRAMA

O hemograma é o exame de medicina diagnóstica que avalia de forma quantitativa e qualitativa todos os elementos celulares do sangue. É o exame mais solicitado nas consultas médicas, fazendo parte de todas as revisões e check-ups de saúde (FAILACE, 2018). Essa preferência dos médicos pela realização do hemograma denota que, além de ser um exame fundamental na triagem de problemas de saúde, este é um indicador indispensável no controle evolutivo de doenças infecciosas, doenças crônicas em geral, das emergências médicas, cirúrgicas e traumatológicas e, ainda, no acompanhamento de radioterapia, tendo relação com toda a patologia (FAILACE, 2018).

A grande quantidade de informações que o hemograma pode fornecer torna este exame um dos mais solicitados nas práticas clínica e cirúrgica (SPSP, 2014). É dinâmico e variável, de acordo com as condições clínicas do paciente no momento da realização do exame. Hemograma é a definição do conjunto de avaliação das células do sangue que, junto com dados clínicos do paciente, permite a dedução de prognósticos e diagnósticos de um grande número de patologias. Failace (2018) define que o hemograma consiste na análise dos três principais grupos de células do sangue: Eritrócitos, Leucócitos e Plaquetas.

O Eritrograma é a divisão do hemograma que faz a avaliação das hemácias, caracterizando o elemento presente em maior quantidade no sangue humano. A função das hemácias é realizar o transporte de oxigênio do pulmão para os demais tecidos do corpo, tarefa realizada pelo conteúdo hemoglobínico presente na grande massa de eritrócitos (FAILACE, 2018). Os eritrócitos, também chamados de hemácias ou glóbulos vermelhos, são contabilizados no hemograma por:

- Eritrócitos (RBC): Representa a contagem total das hemácias por microlitro da amostra de sangue do paciente;
- Hemoglobina (Hb): Componente das hemácias responsável pelo carregamento e transporte do oxigênio no sangue.

Além das contagens quantitativas de células do tipo eritrócito, são estabelecidas algumas medidas de análise qualitativa e quantitativa dos glóbulos vermelhos no sangue, avaliando a morfologia das células:

- Hematócrito (Hct): Representa a porcentagem do volume ocupado pelas hemácias no volume total da amostra de sangue;

- Volume Corpuscular Médio (VCM): Representa o volume médio dos eritrócitos, discriminando se há eritrócitos maiores ou menores que o tamanho normal. Correlaciona-se inversamente proporcional à contagem de eritrócitos;
- *Red blood cell Distribution Width* (RDW): Amplitude de distribuição dos eritrócitos na amostra sanguínea. Representa, em níveis aumentados, excessiva heterogeneidade volumétrica na população de eritrócitos;
- Hemoglobina Corpuscular Média (HCM): Representa a quantidade média de hemoglobina por eritrócito. Parâmetro calculado, representado pela divisão da quantidade de hemoglobina pela quantidade de eritrócitos presentes em um mesmo volume de sangue;
- Concentração hemoglobínica corpuscular média (CHCM): Concentração média de volume da hemoglobina nos eritrócitos. Calculada pelo quociente do HCM pelo VCM.

O segundo grupo de células que compõe o hemograma são os leucócitos. Contabilizados no Leucograma (WBC), a contagem dos leucócitos e suas frações analisa a composição das células brancas do sangue. Produzidas na medula óssea, a quantificação e classificação dessas células verifica a imunidade e como o organismo reage a inflamações e infecções (FAILACE, 2018). Os leucócitos podem ser classificados em:

- Neutrófilos: Tipo de leucócito mais comum, representando 60% a 70% dos leucócitos em circulação nos pacientes normais. São responsáveis pelo combate a infecções bacterianas;
- Linfócitos: Representam 20% a 30% dos leucócitos totais. Combatem vírus, tumores e produzem anticorpos. Ocasionalmente, podem agredir o próprio organismo, criando doenças autoimunes, como no caso da diabetes tipo 1 (ABBAS, LICHTMAN, 2012);
- Monócitos: Compondo de 2% a 10% do total de leucócitos em circulação, são células de defesa responsáveis por agir contra microrganismos invasores;
- Eosinófilos: Representam 3% a 5% dos leucócitos em circulação nos pacientes normais. São responsáveis por agir em casos de alergias ou infecções por parasitas;

- Basófilos: Compõem em média 2% dos leucócitos em circulação. São células de defesa, ativadas em inflamações crônicas ou alergias prolongadas.

O plaquetograma é composto pelos fragmentos de células, de alta importância no início do processo de coagulação. Contabiliza as plaquetas (PLT), segundo componente mais numeroso no sangue e tem como principal função o estancamento de sangramentos após lesões na parede dos vasos sanguíneos. O Volume Plaquetário Médio (VPM) é um indicador do diâmetro médio das plaquetas e é obtido através de cálculo matemático. (MOSKALENSKY et. al., 2018).

O nível elevado de glicose no sangue contribui para a ocorrência de distúrbios nas células sanguíneas e seus índices (MILOSEVIC, PANIN. 2019). Em estudo desenvolvido por Milosevic e Panin (2019), avaliando diversas publicações que realizam a correlação dos sinais hematológicos apresentados pela diabetes, foram apontadas mudanças hematológicas correlacionadas à diabetes, capazes de serem diagnosticadas através do hemograma.

Segundo estes autores, parâmetros hematológicos como WBC, RBC e PLT podem sofrer alterações em quadros de diabetes, fornecendo informações que trazem indícios para o diagnóstico e acompanhamento da doença. Os glóbulos brancos são biomarcadores bem estabelecidos de inflamação e doenças cardiovasculares, comuns em Diabetes. Estudos apresentam redução na vida útil e quantidade de glóbulos vermelhos devido ao aumento da glicemia. Níveis reduzidos de hemoglobina e hematócrito associam-se ao aumento do risco de progressão de microangiopatia, que são lesões isquêmicas no cérebro e podem ser causadas pela diabetes tipo 2. O Volume Plaquetário Médio encontra-se aumentado em pacientes com infarto agudo do miocárdio, apontando o potencial do VPM na fisiopatologia à doença cardiovascular. Abbas e Litctman (2012) e Alisson (2016) citam que em casos de diabetes tipo 1, linfócitos atuam no ataque de células beta do pâncreas, impedindo a produção de insulina e desencadeando o desenvolvimento de diabetes.

O estudo de Milosevic e Panin (2019) conclui que parâmetros do hemograma são úteis para o acompanhamento da diabetes e das complicações ocasionadas em pacientes portadores da doença. Embora a hemoglobina glicada seja o exame padrão-ouro para o diagnóstico e acompanhamento da diabetes, outros testes diagnósticos podem ser utilizados para esta funcionalidade, através dos impactos

gerados pela doença no metabolismo. Explorar as mudanças sutis em um exame de comum realização na área da medicina laboratorial, buscando encontrar um padrão de mudanças capaz de caracterizar o diagnóstico da diabetes vai de encontro com o objetivo geral deste trabalho de pesquisa.

2.3 SISTEMAS COMPUTACIONAIS DE APOIO AO DIAGNÓSTICO DE DIABETES

Embora o tempo de diagnóstico da diabetes esteja melhorando, certos casos podem levar até 10 anos para que a existência da doença seja confirmada, o que exige do médico a análise de diversos indicadores e pode se tornar complexo em certos casos. Sistemas computacionais com foco no apoio ao diagnóstico ganham grande valor nesses casos, pela possibilidade do processamento de um grande volume de informações, com maior detalhamento em um curto espaço de tempo, quando comparados aos seres humanos. Isso possibilita uma maior qualidade nos serviços médicos, além da difusão dos conhecimentos especializados. (BASSO et al., 2014).

A área da saúde é uma das mais visadas para aplicação e testes de novas tecnologias devido a seu grande leque de possibilidades de implantação de melhoramentos tecnológicos, seja no campo de diagnósticos ou tratamentos. Naturalmente, as áreas de maior abrangência na população mundial são as mais fomentadas com propostas de soluções tecnológicas para apoio ao diagnóstico, como é o caso da diabetes.

Pesquisas relacionadas ao tema deste trabalho foram realizadas nos motores de pesquisa IEEE e PubMed. Buscando identificar áreas de interesse relacionadas à classificação dos níveis de glicose através de IA, aplicou-se a pesquisa: (“*Glucose*” AND “*Artificial Intelligence*”), resultando em 145 documentos na base IEEE e 581 documentos na base PubMed. Separou-se a procura por trabalhos relacionados em duas etapas, sendo uma para cada motor de pesquisa utilizado.

Na base PubMed, os resultados foram ordenados pela melhor combinação com a *string* de busca. Realizou-se a leitura do resumo de cada documento para identificar o nível de correlação com o assunto do presente trabalho e foram selecionados os 20 documentos com maior correlação para leitura dos artigos.

Na base IEEE, os resultados foram ordenados por relevância. Seguindo o padrão adotado para a pesquisa no PubMed, foi realizada a leitura do resumo dos documentos e selecionados os 20 com maior correlação para a leitura dos artigos completos.

Foram encontradas diversas propostas de uso de inteligência artificial associada aos níveis de Glicose, mas a característica de aplicação de dados hematológicos no estudo foi encontrada somente em uma publicação. Após a leitura dos 40 artigos, identificou-se três com forte correlação a este trabalho, sendo citados em sequência.

2.3.1 Aprendizado de máquina na intercorrelação de parâmetros hematológicos e níveis de glicose

Com objetivo de classificar o status glicêmico de indivíduos participantes do estudo, Worachartcheewan et al. (2013) propôs o uso de abordagens de aprendizado de máquina, como as máquinas de vetor de suporte, analisando uma amostra populacional de 190 indivíduos residentes na Tailândia. Os pacientes foram descritos conforme um conjunto de exames químicos e hematológicos realizados no sangue, incluindo Glicose, Colesterol total, Triglicerídeos, Colesterol LDL, Colesterol HDL, WBC, RBC, Hb e Hct.

Modelos de classificação foram construídos utilizando o software WEKA, implementando Máquina de Vetores de Suporte (SVM), com o algoritmo de otimização mínima sequencial de John Platt (Witten et al., 2011) e transformação dos dados em um espaço hiperplano dimensional pelo *Radial Basis Function kernel* (WORACHARTCHEEWAN et al., 2013). Os parâmetros hematológicos e o nível de glicose foram utilizados como variáveis de entrada e a classificação do status glicêmico (Normal, Pré-Diabetes ou Diabetes) foi a variável de saída esperada. O modelo classificador foi validado através de validação cruzada de dez vezes (*Cross validation* de 10 *folds*).

Como citado no capítulo anterior, a busca pela correlação de resultados que possibilite o diagnóstico da diabetes através de exames correlatos é um tema vastamente abordado em estudos científicos. Em busca da correlação entre os parâmetros da base de dados, a análise de associação foi aplicada através do algoritmo Apriori para descobrir parâmetros que ocorrem com frequência em

indivíduos com DM. As regras foram obtidas com valores mínimos de suporte = 5% e confiança = 70%. Gerou-se então uma tabela com as características dos indivíduos e seus valores de P, sendo que a amostra dos pacientes foi estratificada conforme os níveis de glicose:

- Normal: Glicose inferior a 100 mg/dL;
- Pré-Diabetes: Glicose entre 100 e 125 mg/dL;
- Diabetes: Glicose superir a 125 mg/dL.

Tabela 1 - Características clínicas da população da amostra

Etapa	Normal	Pré-diabetes	Diabetes	P
Número de casos	107 (56,32)	59 (31,05)	24 (12,63)	-
Masculino	33 (17,37)	27 (16,84)	11 (5,79)	-
Feminino	74 (38,95)	32 (14,21)	13 (6,84)	-
Glicose (mg/dL)	91,82 ± 4,93	108,25 ± 6,02	185,29 ± 67,87	< 0,05
Colesterol (mg/dL)	202,89 ± 43,42	188,9 ± 36,17	178,38 ± 40,75	< 0,05
Triglicerideos (mg/dL)	112,89 ± 57,43	117,68 ± 46,29	144,54 ± 79,23	0,053
HDL-C (mg/dL)	59,93 ± 16,92	45,44 ± 15,87	51,17 ± 14,82	< 0,05
LDL-C (mg/dL)	125,96 ± 39,18	114,76 ± 28,66	105,00 ± 134,92	< 0,05
WBC ($_x10^9$/L)	6,16 ± 1,74	6,69 ± 1,96	7,28 ± 2,14	< 0,05
RBC ($_x10^9$/L)	4,56 ± 0,67	4,57 ± 0,63	4,49 ± 0,69	0,728
Hb (g/dL)	12,87 ± 1,39	12,87 ± 1,57	12,48 ± 1,99	0,179
Hct (%)	39,80 ± 4,44	39,80 ± 4,95	38,55 ± 6,09	0,151

Fonte: Worachartcheewan et al. (2013), traduzido pelo autor.

A partir da Tabela 1 é possível a visualização de algumas informações sobre a amostra:

- No grupo de indivíduos com DM, os valores médios de glicose, triglicerídeos e leucócitos aumentaram;
- Nos grupos Pré-DM e DM, os níveis de triglicerídeos aumentaram enquanto o colesterol HDL teve seus níveis diminuídos;
- Os níveis de Hb, Hct e RBC não foram significantes nos três grupos devido aos valores de P elevados.

Diante dessas considerações e correlações encontradas, foi construída a matriz de intercorrelação para discernir as relações entre os parâmetros analisados, possibilitando a visualização de que o status glicêmico está diretamente correlacionado com RBC, Hb e HCT, como pode ser observado na crescente correlação do grupo normal para o grupo Pré-DM e, por fim, o grupo DM como mostra a Tabela 2.

Tabela 2 - Matriz de intercorrelação de parâmetros bioquímicos e hematológicos

Normais	Glicose	Colesterol Total	Triglicerídeos	Colesterol HDL	Colesterol LDL	WBC	Hb	Hct	RBC
RBC	0,1	0,18	0,24	0,12	0,2	0,14	0,51	0,67	1
Hct	0,17	0,28	0,25	0,12	0,3	0,14	0,95	1	0,67
Hb	0,14	0,26	0,22	0,08	0,28	0,11	1	0,95	0,51
WBC	0,2	0,03	0,32	0,2	0	1	0,11	0,14	0,14
Colesterol LDL	0,26	0,94	0,42	0,1	1	0	0,28	0,3	0,2
Colesterol HDL	0,14	0,2	0,4	1	0,1	0,2	0,08	0,12	0,12
Triglicerideos	0,25	0,39	1	0,4	0,42	0,32	0,22	0,25	0,24
Colesterol Total	0,23	1	0,39	0,2	0,94	0,03	0,26	0,28	0,18
Glicose	1	0,23	0,25	0,14	0,25	0,2	0,14	0,17	0,1

Pré-Diabetes	Glicose	Colesterol Total	Triglicerídeos	Colesterol HDL	Colesterol LDL	WBC	Hb	Hct	RBC
RBC	0,1	0,08	0,23	0,13	0,06	0,25	0,62	0,74	1
Hct	0,2	0,08	0,1	0,04	0,09	0,04	0,95	1	0,74
Hb	0,22	0,13	0,06	0,09	0,13	0,13	1	0,95	0,62
WBC	0,24	0,27	0,12	0,38	0,15	1	0,13	0,04	0,25
Colesterol LDL	0,05	0,93	0,16	0,2	1	0,15	0,13	0,09	0,06
Colesterol HDL	0,06	0,53	0,33	1	0,2	0,38	0,09	0,04	0,13
Triglicerideos	0,06	0,11	1	0,33	0,16	0,12	0,05	0,1	0,23
Colesterol Total	0,07	1	0,11	0,53	0,93	0,27	0,13	0,08	0,08
Glicose	1	0,07	0,06	0,06	0,058	0,24	0,22	0,2	0,1

Diabéticos	Glicose	Colesterol Total	Triglicerídeos	Colesterol HDL	Colesterol LDL	WBC	Hb	Hct	RBC
RBC	0,36	0,53	0,03	0,21	0,5	0,08	0,8	0,88	1
Hct	0,5	0,57	0,06	0,22	0,56	0,12	0,99	1	0,88
Hb	0,53	0,55	0,1	0,19	0,54	0,17	1	0,99	0,8
WBC	0,32	0,08	0,65	0,34	0,14	1	0,17	0,12	0,08
Colesterol LDL	0,03	0,95	0,01	0,23	1	0,14	0,54	0,56	0,5
Colesterol HDL	0,02	0,42	0,5	1	0,23	0,34	0,19	0,22	0,21
Triglicerideos	0,27	0,07	1	0,5	0,01	0,65	0,1	0,06	0,03
Colesterol Total	0,04	1	0,07	0,42	0,95	0,08	0,55	0,57	0,53
Glicose	1	0,04	0,27	0,02	0,03	0,32	0,53	0,5	0,36

Fonte: Worachartcheewan et al. (2013), traduzido pelo autor.

No modelo classificador proposto, parâmetros hematológicos compostos por WBC, RBC, Hb, Hct e glicose foram utilizados como variáveis de entrada, enquanto os grupos Diabetes, Pré-Diabetes e Normal foram utilizados como variáveis de saída. No desenvolvimento dos modelos de classificação através de Máquinas de Vetor de Suporte, os parâmetros C e gama (ɣ) foram otimizados em um processo de *Grid Search*. Este processo realiza a execução do modelo classificador variando os parâmetros C e ɣ, para obter os valores ótimos, capazes de gerar a melhor acurácia possível. Os resultados indicaram que os valores ótimos para os parâmetros C e ɣ foram 2^{23} e $2^{-8,5}$ respectivamente, gerando uma precisão de 100% e 98,42%, respectivamente, como mostra a Tabela 3.

Tabela 3 - Resumo do desempenho para identificação de DM utilizando SVM

Etapa	Normal	Pré-diabetes	Diabetes	Acurácia
Treinamento				100
Normal	107	0	0	
Pré-diabetes	0	59	0	
Diabetes	0	0	24	
Validação cruzada 10x				98,42
Normal	107	0	0	
Pré-diabetes	0	58	1	
Diabetes	0	2	22	

Fonte: Worachartcheewan et al. (2013), traduzido pelo autor.

Além do modelo de SVM, análise de associação (AA) foi aplicada para descobrir ocorrências frequentes em indivíduos com ou sem DM. Geraram-se 533 regras que podem ser estratificadas em grupos: 411 regras para o grupo de pacientes normais, 111 regras para o grupo de pacientes com pré-diabetes e 11 regras para o grupo de diabéticos. Foi observado que as regras para o grupo dos diabéticos englobaram somente níveis anormais de glicose. Regras para o grupo pré-diabetes foram associadas ao aumento de WBC, RBC, Hb e Hct, além da glicose elevada, o que é característico da doença.

Por fim, aponta-se que as variáveis Hb e Hct aumentaram em pacientes com alta viscosidade do sangue, conforme encontrado nos grupos DM e Pré-DM. RBC também foi correlacionada à condição glicêmica (JUNG et al., 2013). Conforme visualizado na Tabela 2, os parâmetros hematológicos foram associados ao status glicêmico, coincidindo com achados anteriores (JUNG et al., 2013).

A inclusão de parâmetros hematológicos na classificação do estado de DM levou a uma precisão acima de 98%, estando implícito que os parâmetros hematológicos são variáveis importantes, juntamente com o nível de glicose, para a identificação no estado de DM. As abordagens de aprendizado de máquina por Máquinas de Vetor de Suporte empregadas no estudo demonstraram ser capazes de classificar corretamente o status do DM.

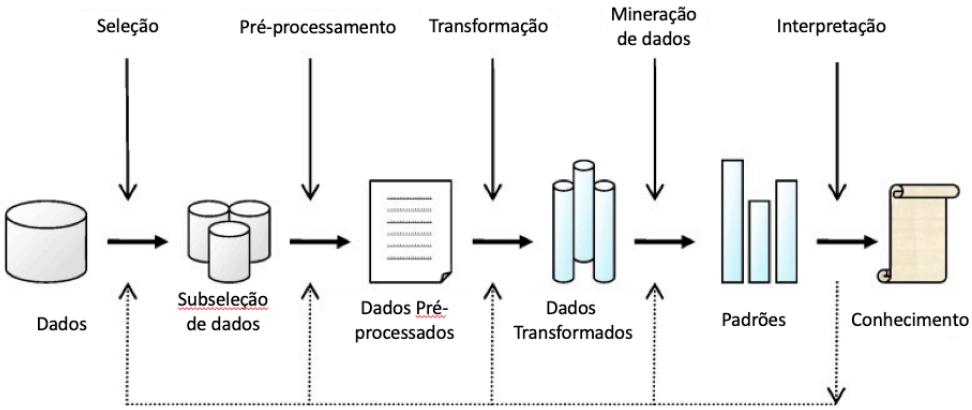
2.3.2 Métodos de aprendizado de máquina na pesquisa de diabetes

A aplicação de métodos de aprendizado de máquina e de mineração de dados na pesquisa de DM é uma abordagem fundamental para a utilização de grandes volumes de dados em pesquisas relacionados à diabetes. O severo impacto social da doença faz do DM uma das principais prioridades de pesquisa em ciências médicas,

que inevitavelmente gera enormes quantidades de dados. Nesse contexto, abordagens de aprendizado de máquina e de mineração de dados de Diabetes Mellitus são grande preocupação quando se trata de diagnóstico e outros aspectos ligados á administração clínica. Kavakiotis et al. (2017) apresenta uma revisão literária sobre o status atual do aprendizado de máquina e abordagens de mineração de dados na pesquisa sobre diabetes.

A descoberta de dados (KDD) em bases de dados é um campo que abrange teorias, métodos e técnicas, tentando entender os dados e extrair conhecimento. Caracteriza-se em um processo de várias etapas (seleção, pré-processamento, transformação, mineração de dados e avaliação de resultados), como demonstra a Figura 2:

Figura 2 - As etapas básicas do processo KDD



Fonte: Kavakiotis et al. (2017).

Uma definição completa de KDD é dada por Fayyad et al. (1996): KDD é o processo não trivial que identifica padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis nos dados.

As tarefas de aprendizado de máquina são normalmente classificadas em três categorias: a) aprendizado supervisionado, no qual o sistema deduz uma função dos dados de treinamento; b) aprendizado não supervisionado, onde o sistema tenta inferir a estrutura dos dados não rotulados; c) aprendizado de reforço, onde o sistema interage com um ambiente dinâmico. No aprendizado supervisionado, o sistema “aprende” intuitivamente uma função, chamada de função alvo, descrevendo os dados. A função objetivo é utilizada para prever o valor de uma variável, denominada a variável de saída, a partir de um conjunto de variáveis de entrada. Um subconjunto de todos os casos com valor de saída conhecido é utilizado como conjunto de dados

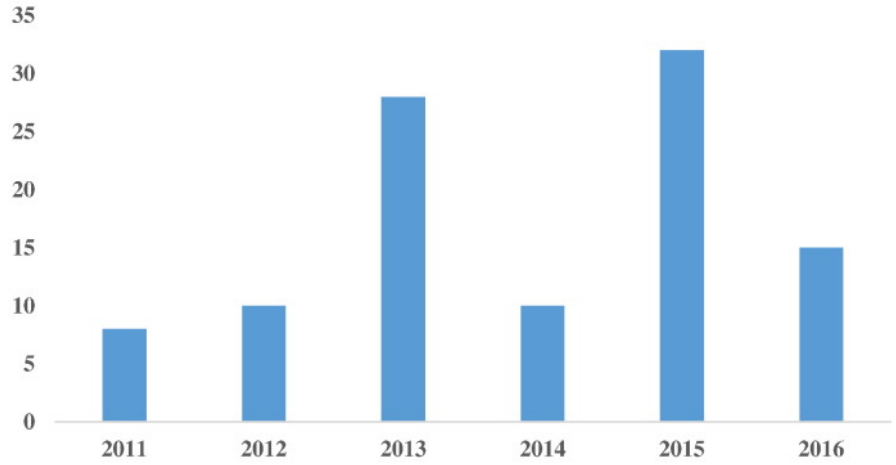
de treinamento. Existem dois tipos de tarefas de aprendizado: regressão e classificação. Os modelos de regressão preveem modelos numéricos, enquanto os modelos de classificação geram classes distintas. Algumas das técnicas mais utilizadas são Árvores de Decisão, Algoritmos genéticos, Redes Neurais Artificiais e Máquinas de Vetores de Suporte.

O diagnóstico de Diabetes Mellitus é realizado através de vários testes, como a dosagem de Hemoglobina Glicada, o teste aleatório de glicose no sangue ou o teste oral de tolerância à glicose. Existem evidências de que, tanto no tipo 1 da doença, quanto no tipo 2, o diagnóstico precoce e a previsão do início da doença são vitais para o retardo dos sintomas, a seleção do medicamento adequado e prolongamento da expectativa de vida do indivíduo.

Inúmeros algoritmos e diferentes abordagens foram aplicados na previsão do DM, como algoritmos tradicionais de aprendizado de máquina, técnicas de aprendizado por conjunto e aprendizado por regras de associação, com a finalidade de obter a melhor precisão de classificação. As abordagens de conjuntos que utilizam vários algoritmos de aprendizado provaram ser uma maneira eficaz de melhorar a precisão da classificação. Os autores Ozcift e Gulten (2011) propuseram um modelo utilizando o *Rotation Forest*, algoritmo de conjunto desenvolvido recentemente, para combinar 30 algoritmos de aprendizado de máquina. Han et al. (2015) apresentam uma abordagem de aprendizado em conjunto explicitando as decisões de máquinas de vetores de suporte de forma compreensível e transparente. Regras de associação são aplicadas para identificar fatores de risco ou detectar combinações de variáveis que ocorrem frequentemente juntas em pacientes diabéticos.

No seu fechamento, o estudo traz um gráfico de comparativo anual de quantidade de estudos publicados na área de uso de inteligência artificial para a diabetes, utilizados na composição da revisão bibliográfica, apresentados na Figura 3 abaixo.

Figura 3 - Artigos por ano empregados no estudo



Fonte: Kavakiotis et al. (2017).

Grande parte dos estudos realizaram análises comparativas em diferentes algoritmos de aprendizado de máquina para avaliar seu desempenho preditivo e, então, escolher o mais eficiente. Tratando-se de estudos que não tenham uma avaliação ou premissa já existente sobre a escolha do algoritmo a ser utilizado, a testagem de vários algoritmos deve ser a base nos estudos a serem realizados, levando em consideração que várias características do conjunto de dados podem afetar de forma significativa o desempenho do algoritmo. O Quadro 1 representa estudos comparando algoritmos de aprendizado de máquina em vários conjuntos de dados diferentes:

Quadro 1- Comparação entre diferentes algoritmos de Machine Learning

Autor	Tipo de dados	Nº de sujeitos	Algoritmos	Validação	Melhor Precisão
Cai et al.	Gut microbiota	499	Regressão logística, Análise discriminante linear, *Naive Bayes*, Máquina de vetores de suporte	Validação cruzada 10 vezes	Máquina de vetores de suporte
Malik et al.	Medições eletroquímicas da saliva	175	Regressão logística, Máquina de vetores de suporte, Redes neurais artificiais	Validação cruzada 3 vezes	Máquina de vetores de suporte ACC = 84,09
Ferran et al.	Medições demográficas, antropométricas, sinais vitais, diagnóstico de laboratório clínico	10632	Regressão logística, K vizinhos mais próximos, Redução de dimensionalidade multifatorial, Máquina de vetores de suporte	Validação cruzada 5 vezes	Máquina de vetores de suporte ACC = 31,3
Mani et al.	Valores demográficos e diagnóstico de laboratório clínico	2280	*Naive Bayes*, Regressão logística, K vizinhos mais próximos, Florestas aleatórias (RF), Máquina de vetores de suporte	Validação cruzada 5 vezes	Florestas Aleatórias (RF) AUC = 0,803 / 0,807 / 0,877

Tapak et al.	Medições demográficas, antropométricas e diagnóstico de laboratório clínico	6500	Redes neurais artificiais, Máquina de vetores de suporte, Florestas aleatórias	Validação cruzada 10 vezes	Máquina de vetores de suporte ACC = 0,986 AUC = 0,979

Fonte: Kavakiotis et al. (2017).

A categoria mais popular entre os artigos que englobam o tema foi a previsão e diagnóstico de biomarcadores do DM. Como a pesquisa tem um processo orientado a dados, as limitações presentes na pesquisa de aprendizado de máquina do DM estão relacionadas à disponibilidade de dados. Dados biológicos são mais difíceis e caros de gerar, portanto, são menos disponíveis para a comunidade científica.

Quando se trata de aprendizado de máquina e mineração de dados, ressalta-se que a grande maioria de artigos associados nesta pesquisa relataram precisão acima de 80% na classificação ou previsão do DM. Na tarefa de previsão, os algoritmos mais utilizados foram Máquinas de Vetor de Suporte, Redes Neurais Artificiais e Árvores de Decisão. Salienta-se que as máquinas de vetores de suporte surgem como o algoritmo de maior sucesso em conjuntos de dados referentes à Diabetes Mellitus. Mais de 85% dos artigos utilizaram as abordagens de aprendizado supervisionado e, nos 15% restantes, regras de associação foram empregadas principalmente no descobrimento de associação entre biomarcadores.

Com o advento da biotecnologia e a grande massa de dados que vem sendo produzida, juntamente com o aumento e disseminação das técnicas de Aprendizado de Máquina, ocorreu uma exploração mais aprofundada em direção ao diagnóstico e tratamento do DM através do emprego de técnicas de aprendizado de máquina e mineração de dados, com o objetivo de sugerir novos biomarcadores e detectar aspectos-chave da doença. Por fim, o estudo relata que todas as pesquisas relacionadas refletem um processo orientado à base de dados disponível, salientando que lacunas e limitações resultantes nas pesquisas de aprendizado de máquina no diabetes estão diretamente relacionadas à disponibilidade de dados. Aponta-se que um conjunto de dados suficientemente grande é imprescindível para que o algoritmo de aprendizado seja treinado adequadamente.

2.3.3 Índices hematológicos e sua correlação com nível de glicose no sangue

O número de pessoas que sofrem de DM tipo 2 vem apresentando um aumento significativo devido ao envelhecimento da população mundial e ao sedentarismo, que ocorre em grande parte da população, principalmente nas grandes cidades. A carga temporária da hiperglicemia é responsável por complicações do DM e resultados adversos, como o aumento do risco de desenvolvimento de doença cardiovascular e infarto do miocárdio. O DM2 compõe uma parte da síndrome metabólica que compreende dislipidemia, hipertensão e índices hematológicos comprometidos.

As variações hematológicas apresentadas pela doença afetam glóbulos vermelhos, glóbulos brancos e os fatores de coagulação. Outras anormalidades registradas em pacientes com DM são apontadas nas hemácias, leucócitos e disfunção plaquetária. Em revisão sistemática e metanálise de estudos correlatos, Biadgo et al. (2016) aponta que o número de leucócitos periféricos, caracterizado por basófilos, eosinófilos e neutrófilos se caracterizou aumentado, enquanto o número de monócitos não teve alteração em pacientes com DM2.

Os dados hematológicos são indicadores de alta importância na avaliação de variações de tamanho, número e maturidade de diferentes células sanguíneas. São importantes para a avaliação e tratamento dos pacientes com DM. O trabalho possuiu como objetivo a determinação de índices hematológicos e sua correlação com o nível de glicemia em jejum em pacientes com DM2, para comparar com pacientes aparentemente saudáveis.

O estudo foi aplicado de maneira comparativa no período de fevereiro a abril de 2015, na clínica de doenças crônicas do Hospital Universitário Gondar, na cidade de Gondar, Etiópia. O hospital possui 500 leitos e presta serviços de referência na saúde para mais de 5 milhões de habitantes da região. Como hospital de ensino, desempenha papel importante na pesquisa e serviço comunitário.

Foram analisados 148 pacientes com DM2, sendo 59 homens e 89 mulheres, com idades entre 25 e 70 anos. Como sujeitos de controle, foram inclusos no estudo mais 59 homens e 89 mulheres aparentemente saudáveis, que não possuíam histórico prévio de doenças crônicas. Foram excluídos do estudo pacientes gestantes, fumantes, alcoólatras, hipertensos ou em tratamento com aplicação de insulina.

Através de questionário, foram coletados dados como altura, peso e circunferência da cintura. A pressão arterial foi aferida por pessoal qualificado, além

da medição dos níveis de batimentos cardíacos. Foi coletada uma amostra de 5 mililitros de sangue para determinação da Glicose em jejum e uma amostra de sangue para dosagem dos parâmetros hematológicos.

Os dados foram digitados e analisados através do software *Statistical Package for the Social Sciences* (SPSS), versão 20 (IBM *Corporation, Armonk*, NY, EUA). Testes de normalidade dos dados foram realizados através de histogramas, por comparação de médias, medianas e pela realização de testes de simetria através do teste de Kolmogorov – Smirnov. Foram criados valores médios e de desvio padrão para variáveis contínuas, porcentagens para variáveis categóricas e intervalo interquartil para as variáveis com distribuição normal. A força da associação entre os pares de variáveis foi avaliada pelo coeficiente de correlação de Pearson e Spearman, onde valores inferiores a 0,05 foram considerados estatisticamente significativos.

Na análise dos parâmetros hematológicos, observou-se incremento estatisticamente significante nos índices de leucócitos, neutrófilos e linfócitos nos pacientes com DM2, em comparação com o grupo de controle. Nos índices de hemácias, apenas o RDW apresentou-se significativo. Incrementos calculados como o VPM e PDW também se demonstraram significantes nos pacientes com DM2. Os valores de correlação *P* podem ser observados na Tabela 4 abaixo, sendo que $P < 0,05$ foi considerado significativo para o estudo.

Tabela 4 - Comparação dos índices hematológicos entre pacientes DM2 e pacientes de controle.

Variáveis	Média ± Desvio padrão (DM2)	Média ± Desvio Padrão (Controle)	*P*
Série de células brancas			
WBC (Leucócitos)	6,59 ± 1,42	5,56 ± 1,38	0,000
Linfócitos	2,60 ± 0,70	2,04 ± 0,63	0,000
Neutrófilos	3,57 ± 1,46	3,11 ± 1,04	0,012
Série de células vermelhas			
Hemácias	5,12 ± 0,57	5,1 ± 0,54	0,7555
Hemoglobina	15,2 ± 1,7	15,1 ± 1,5	0,739
Hematócrito	46,7 ± 5,1	46,4 ± 4,2	0,609
VCM	91,7 ± 5,0	90,7 ± 4,3	0,056
MCH	29,7 ± 2,6	30,0 ± 5,3	0,086
CHCM	32,6 ± 2,0	32,5 ± 1,0	0,772
RDW	47,3 ± 2,6	45,2 ± 3,0	0,000
Série plaquetária			
Plaquetas	255,7 ± 82,0	246,3 ± 67,4	0,280
VPM	10,4 ± 1,1	9,9 ± 1,1	0,001
PDW	14,5 ± 2,1	13,4 ± 2,1	0,000

Fonte: Biadgo et al. (2016).

A glicemia de jejum apontou correlação positiva com o total de leucócitos, linfócitos e neutrófilos para o grupo DM2. No entanto, não foi observada a mesma correlação entre as variáveis no grupo de controle. Os valores de *P* podem ser visualizados na Tabela 5, que apresenta a correlação das variáveis hematológicas com o dado de Glicose para os pacientes do estudo.

Tabela 5 - Correlação de Pearson (r) entre os índices hematológicos e glicose de jejum

Variáveis	Grupo DM2 Glicose Jejum r(*P*)	Grupo Controle Glicose Jejum r(*P*)
Série de células brancas		
WBC (Leucócitos)	0,221 (0,007)	0,008 (0,923)
Linfócitos	0,174 (0,034)	0,119 (0,149)
Neutrófilos	0,175 (00,033)	-0,033 (0,692)
Série de células vermelhas		
Hemácias	0,129 (0,118)	0,047 (0,57)
Hemoglobina	0,106 (0,201)	0,01 (0,905)
Hematócrito	0,149 (0,077)	0,026 (0,758)
VCM	0,036 (0,663)	-0,081 (0,328)
MCH	0,083 (0,315)	-0,045 (0,587)
CHCM	0,159 (0,053)	-0,021 (0,798)
RDW	-0,12 (0,559)	-0,094 (0,254)
Série plaquetária		
Plaquetas	-0,048 (0,559)	0,029 (0,729)
VPM	0,038 (0,643)	0,008 (0,927)
PDW	0,614 (0,042)	0,007 (0,936)

Fonte: Biadgo et al. (2016).

Ainda no estudo foram realizadas comparações entre as variáveis de IMC e relação cintura/quadril, onde não se obteve correlações estatisticamente significantes. Evidências da pesquisa apontam que os índices hematológicos estão alterados em pacientes com DM2. O estudo também aponta que os índices de hemácias apresentaram incremento em pacientes diabéticos, quando relacionados ao grupo de controle, mas a diferença não foi estatisticamente significante. Entre os índices da série vermelha, encontrou-se diferença entre os grupos de DM2 e controle no parâmetro RDW, correlacionado a achados anteriores dos autores da pesquisa. Os índices de leucócitos aumentaram significativamente no grupo DM2, em comparação ao grupo de controle. Artigos técnicos apontam a correlação devido ao aumento do estresse oxidativo do organismo, desencadeado pelos altos níveis de glicose. O estudo aponta que não foi encontrada diferença significativa nos níveis de plaquetas entre os dois grupos estudados, mas foi detectado um incremento significativo no VPM em pacientes com DM2.

Por fim, a pesquisa encontrou correlação significativa entre as variáveis de Glicose de Jejum, VPM, Linfócitos, Neutrófilos, WBC e RDW aumentadas entre os pacientes diabéticos, quando comparados aos níveis de controle, explicitando um reflexo dos distúrbios no controle glicêmico dos pacientes.

Em busca bibliográfica, registrada com seus principais resultados nessa sessão de trabalhos correlatos, foi possível identificar publicações confirmando a existência de correlações entre o status glicêmico e seu impacto em alterações nos parâmetros hematológicos. A partir dos trabalhos citados, percebe-se que a aplicação de técnicas de aprendizado de máquina com foco no diagnóstico ou previsão da diabetes é uma área que vem ganhando atenção na comunidade científica. Com base no conteúdo abordado nas publicações correlatas, o presente trabalho busca satisfazer uma área ainda não explorada, que é a classificação glicêmica com base somente em índices hematológicos. As informações encontradas nos trabalhos correlatos apresentam alta relevância devido à constante correlação com achados no decorrer da pesquisa, que serão reportados nos capítulos seguintes.

3 EXTRAÇÃO E COMPREENSÃO DA BASE DE DADOS

Na medida em que nossa saúde se molda através do contexto em que vivemos, não surpreende que a inovação digital que transformou nossas rotinas de vida também tenha um papel a desempenhar na formação da saúde da população (GALEA, VAUGHAN, 2017). As novas tecnologias têm real potencial para melhorar o que fazemos e como fazemos, a partir da abertura de ideias emergentes, propiciando a maior eficácia das ações ou apresentando novos métodos, a fim de melhorar cada vez mais a saúde das populações (NGUYEN et al., 2017). Nesse contexto, apesar da grande ênfase no potencial das novas tecnologias e seus benefícios para a saúde, é notável que o uso apropriado e o desenvolvimento de melhorias em tecnologias já existentes são uma área de grande poder para avanços na medicina diagnóstica.

Pesquisadores médicos apontam que a familiaridade com ferramentas de aprendizado de máquina voltadas a análises de grandes volumes de dados será um requisito indispensável para a próxima geração de clínicos, na qual algoritmos poderão concorrer ou substituir os médicos em áreas de exames minuciosos, como anatomia e exames de imagem (CHAR et al., 2018). Segundo Beam e Kohane (2018), exemplos demonstraram que o *big data* e o aprendizado de máquina são capazes de criar algoritmos com desempenho aproximado aos médicos humanos.

Para a aplicação de aprendizado de máquina em processos de descoberta de conhecimento, Witten et al. (2017) aponta duas etapas aplicadas diretamente aos dados, predecessoras ao modelo classificador, que são de grande importância para a consistência dos resultados obtidos: Subseleção de dados (também chamada de composição da base de dados) e Pré-processamento dos dados. Este capítulo apresenta a primeira parte do desenvolvimento desta pesquisa, que compreende pela composição e extração da base de dados a ser utilizada. Em seguida, análises foram realizadas no conjunto de dados, expondo uma visão geral da população de dados obtida.

3.2 COMPOSIÇÃO DA BASE DE DADOS

Um dos recursos essenciais para a execução de algoritmos de Aprendizado de máquina são os dados que alimentam o modelo. Compor uma base de dados adequada à finalidade para qual está se desenvolvendo o modelo classificador é um

dos pontos-chave para que seja obtida uma boa acurácia. Os trabalhos correlatos à esta pesquisa apresentaram contribuições essenciais na composição da base de dados. Biadgo et al. (2016), afirma que dados hematológicos são indicadores de alta importância na avaliação de variações de tamanho, número e maturidade de diversas células sanguíneas, importantes na avaliação e tratamento de pacientes com Diabetes. Partindo dessas informações e as associações da DM com parâmetros hematológicos referenciadas na sessão de correlações com a área da saúde, a etapa de seleção das variáveis hematológicas para a composição da base de dados foi iniciada.

A seleção das variáveis hematológicas para a composição da base de dados foi iniciada pelo índice de correlação do dado hematológico com a diabetes, segundo a análise de referências da área da saúde. Milosevic e Panin (2019) apontam que os dados hematológicos de Leucócitos, Eritrócitos e Plaquetas podem sofrer alterações em quadros de diabetes, fornecendo indícios para o diagnóstico e acompanhamento da doença. Referenciado na sessão de correlações com a área da saúde, o hemograma é composto por três grandes agrupamentos de informações, correspondentes às classes de dados hematológicos citados pelos autores acima. Desse modo, compreende-se que reflexos da diabetes podem ser encontrados em todas as sessões de um exame diagnóstico Hemograma.

Na série branca dos dados hematológicos, alteração nos quantitativos de Leucócitos, marcadores de infecção e doenças cardiovasculares, são comuns em quadros de diabetes (MILOSEVIC; PANIN, 2019). Os indicadores quantitativos de leucócitos são subdivididos em cinco tipos, sendo que cada um possui suas funções específicas e de grupo. Considerando que as citações em estudos são de alteração no total de leucócitos em quadros de diabetes, consequentemente os quantitativos de suas classes também serão alterados. Deste modo, todos os dados da série branca foram incluídos no subconjunto de dados.

Milosevic e Panin (2019) também apontam que pode haver reduções na vida útil e quantidade de glóbulos vermelhos devido ao aumento da glicemia. Microangiopatia, causada em casos severos de Diabetes tipo 2, são associados aos níveis reduzidos de hemoglobina e hematócrito, variáveis que também são componentes da série de glóbulos vermelhos do hemograma. Nesta série, medidas de análise quantitativa e qualitativa dos glóbulos vermelhos são estabelecidas com base nos valores obtidos das contagens de glóbulos vermelhos, que

consequentemente também se representarão alterados em casos de DM. Seguindo o mesmo raciocínio utilizado na série anterior, todas as variáveis da série vermelha foram inclusas no subconjunto de dados.

A série do plaquetograma é afetada através do volume plaquetário médio (VPM), que se encontra aumentado em pacientes com infarto agudo do miocárdio. Sabendo que o volume plaquetário médio é um valor calculado, que engloba o quantitativo de plaquetas do paciente, as variáveis de Contagem de Plaquetas e VPM foram adicionadas ao subconjunto de dados.

Segundo as colocações citadas, variáveis hematológicas de Leucócitos, Hemácias e Plaquetas seriam indispensáveis na base de dados. Embora os estudos correlatos não utilizem o detalhamento nas frações dos grupos de células, entende-se que ao utilizar as frações que compõem um grande grupo de informações, a precisão das classificações seria aumentada ou se encontraria nas frações o parâmetro que sofre alterações, refletindo no total do grande grupo. Deste modo, foram adicionadas as variáveis hematológicas à composição da base de dados, contemplando o maior detalhamento das células. A composição do detalhamento dos grupos pode ser visualizada no Quadro 2:

Quadro 2 - Composição da base de dados segundo grupos de variáveis hematológicas

Geral	Série Branca	Série Vermelha	Plaquetograma
Sexo	Leucócitos	Hemácias	Plaquetas
Idade	Neutrófilos	Hematócrito	VPM
	Linfócitos	Hemoglobina	
	Monócitos	VCM	
	Eosinófilos	HCM	
	Basófilos	CHCM	
		RDW	

Fonte: Elaborado pelo autor.

No grupo denominado como geral, alocaram-se as informações de contexto dos pacientes. Em relação ao texto apresentado na entrega do Trabalho de Conclusão I, houve a exclusão da variável ID, presente no grupo Geral. Este dado foi removido pois era um ID sequencial do registo no sistema de informações da empresa que cedeu a base de dados, não interferindo no modelo classificador. Houve também a troca de Data de Nascimento pela variável de Idade, por ser facilitado o uso dos valores numéricos no classificador. A idade do paciente para cada registro do conjunto de dados foi calculada subtraindo a data de nascimento da data de realização do exame, tendo o resultado convertido para Anos. A data de nascimento e o sexo dos indivíduos são informações relevantes pois os valores podem ser influenciados por

fatores como idade, sexo ou presença de fatores de risco (ROSENFELD, 2019). Para o grupo de Leucócitos, Hemácias e Plaquetas, suas frações foram adicionadas à base de dados.

Como cita o capítulo 2.1.3, a principal forma de diagnóstico de DM na atualidade é através da dosagem da hemoglobina A1c, procedimento de diagnóstico laboratorial também disponível na base de dados. Para maior acurácia no treinamento dos algoritmos de aprendizado de máquina, na composição da base de dados foi incluso o parâmetro de resultado de A1c para os pacientes estudados. Esta informação será utilizada na avaliação da acurácia do modelo classificador gerado neste trabalho, visto que é o padrão-ouro na representação do status glicêmico do paciente (SBPC, 2020).

O grande volume de informações presentes na base de dados disponível para uso facilitou a composição do conjunto de dados, pois todas as informações referenciadas pelos trabalhos correlatos estavam disponíveis no banco de dados. O uso das variáveis que representam as frações de cada um dos três grandes grupos de células é um detalhe específico que não foi utilizado nos trabalhos correlatos. A disponibilidade da obtenção do resultado da dosagem de Hemoglobina Glicada para cada paciente é um diferencial, pois este dado representa o real status glicêmico de cada paciente no momento da realização do hemograma, informação importante para uso no treinamento do modelo classificador estudado.

3.3 CONTEXTO DOS DADOS

Para execução do trabalho de pesquisa, firmou-se parceria com a empresa Laboratório Bom Pastor, na qual o autor da pesquisa está empregado, que forneceu a base de dados para realização do estudo. Atuando no ramo de análises clínicas e medicina laboratorial há mais de 45 anos e presente em 9 cidades da encosta da serra gaúcha, esta empresa presta serviço a 8 hospitais e atende a pacientes ambulatoriais em 14 unidades de coleta de exames, operando de forma ininterrupta em um regime de 24 horas, 7 dias por semana.

Atendendo mais de 20 mil pacientes por mês, a empresa possui datacenter próprio, onde todas as informações ficam armazenadas em um banco de dados de arquitetura pós-relacional Caché, específico para armazenamento de informações da área da saúde e finanças, desenvolvido pela *InterSystems*. Sob esta base de dados

opera o sistema nomeado Shift, responsável por informatizar todo o fluxo de operação laboratorial, desde os cadastros de pacientes e exames até os sistemas de automação e comunicação com os equipamentos analisadores. No fluxo do laboratório, todos os tubos de amostras biológicas são identificados com códigos de barras, que são lidos pelos equipamentos de automação durante o processamento dos exames, para garantir a segurança do processo. Os resultados produzidos pelos equipamentos de análise são enviados através de automação para o sistema Shift, que realiza a validação por expressões regulares e compõe o laudo dos exames de forma automática.

Os dados hematológicos coletados para uso neste trabalho foram gerados por um equipamento de análises do fabricante Beckman Coulter, modelo DXH-800, sendo que todos os exames do setor de hematologia do laboratório passam por revisão microscópica de um profissional habilitado antes da liberação ao laudo no sistema. Controles de qualidade têm seus valores aferidos três vezes ao dia em três níveis (Baixo, Médio e Alto) para cada parâmetro hematológico analisado. Os valores de A1c que compõem a base de dados são resultados do equipamento BioRad *Variant* II Turbo, que possui dois níveis de controle de qualidade e valores aferidos três vezes ao dia.

Na arquitetura do sistema de informação laboratorial implementado, cada variável de resultados de exames fica armazenada em um campo denominado parâmetro. O valor do parâmetro alterna-se a cada paciente, conforme seus resultados. O resultado de um exame poderá ser composto por um ou vários parâmetros, dependendo de sua complexidade.

Para a exportação dos resultados de exames do sistema de informação laboratorial, utilizou-se uma funcionalidade já desenvolvida no sistema Shift, denominada Exportação de Resultados. Nesta funcionalidade, é possível realizar uma exportação de dados, selecionando os dados que se deseja exportar. Foram selecionados os parâmetros correspondentes à cada variável do conjunto de resultados de exames a ser gerado, juntamente com os dados de idade e sexo, que correspondem ao paciente de cada ordem de serviço. A base de informações foi gerada em um arquivo com formatação separada por vírgulas (CSV), que pode ser importado em diversas plataformas de software para visualização e tratamento das informações.

Diante da definição de variáveis e dados necessários para a pesquisa, foram exportados os resultados dos exames de Hemograma e A1c realizados entre 01/01/2018 e 30/09/2019. Todas as informações foram exportadas sem características de identidade dos pacientes, mantendo-os anônimos, preconizando o código de ética da empresa fornecedora das informações.

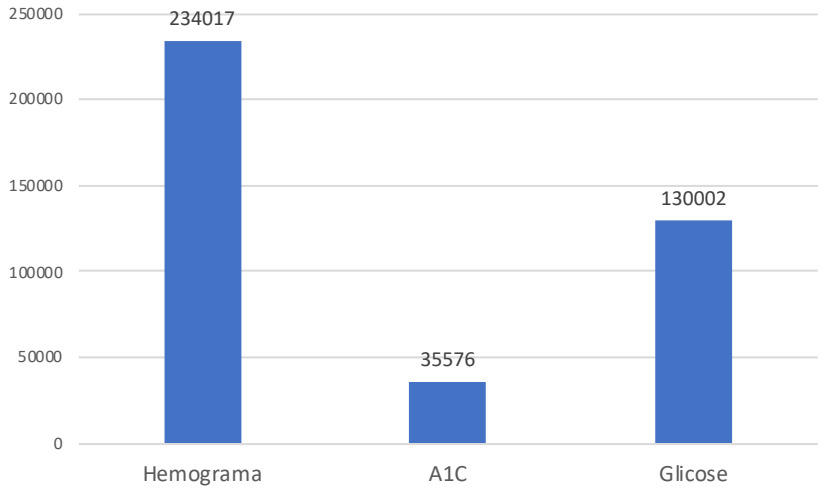
3.4 EXPLORAÇÃO E PRÉ-PROCESSAMENTO

A etapa de exploração da base de dados tem como objetivo a tomada de conhecimento das informações que compõem o grande conjunto de dados, bem como a identificação de possíveis pontos de necessidade de tratamento de informações na etapa de pré-processamento dos dados. A amostra completa de dados foi extraída e aberta no Microsoft Excel, para que fosse possível uma primeira visualização do grande bloco de informações.

A base é composta por uma tabela com 252727 linhas e 18 colunas, sendo que os parâmetros de resultados são caracterizados pelas colunas e os dados de resultados de pacientes estão expressos nas linhas. O quantitativo de linhas corresponde à quantidade de pedidos de exames contemplados na base de dados.

Em análises iniciais dos dados, notou-se que algumas linhas estavam sem os valores de resultados para o Hemograma, para A1c ou para Glicose. Havia linhas com resultado para um só exame. Suspeitando de possíveis problemas de exportação, uma amostra destes casos foi separada para análise das informações e comparação com os dados registrados no sistema Shift. Constatou-se através dos registros do sistema que nos casos em que alguma das informações estava vazia, o paciente não havia realizado o exame. Como os três exames que compõem a base de dados não são correlatos diretamente, médicos solicitantes podem realizar solicitação somente de um ou outro procedimento, causando o registro de resultados somente para o exame solicitado e ficando os demais registros inexistentes na base de dados. Nessa condição, somando separadamente os quantitativos de resultados do bloco de informações exportado, obtém-se um total de 399595 exames. A distribuição está representada pela Figura 4.

Figura 4 - Distribuição da base de dados por tipo de exame



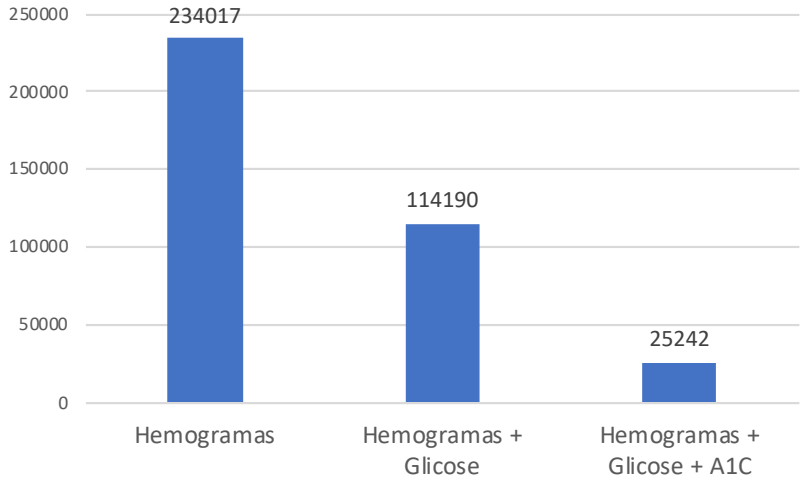
Fonte: Elaborado pelo autor.

Os dados expressos na Figura 4 compactuam com o referenciado na introdução desta pesquisa, onde informa que o Hemograma é o exame mais solicitado nas consultas médicas, fazendo parte de todas as revisões e check-up de saúde (FAILACE, 2018). Há um grande volume a maior na solicitação do Hemograma, quando comparado aos exames de Glicose e, principalmente A1c, parâmetro ainda pouco utilizado pelos médicos em razão de ser relativamente novo, com posicionamentos oficiais na comunidade médica a partir de 2009 (Neto et al., 2009), ou devido ao valor agregado mais alto, quando comparado às dosagens de Glicose. Este gráfico deixa explícita a lacuna onde esta pesquisa visa atuar: A classificação do status glicêmico através dos dados do Hemograma, este que é largamente mais realizado quando comparado a A1c, possivelmente entregando diagnóstico a muitos pacientes que não realizariam o procedimento específico para correlação com a diabetes.

A proposta desta pesquisa possui como objetivo o uso dos resultados de A1c e Hemograma em algoritmos de aprendizado de máquina. Para que esta aplicação seja implementada de modo coerente, é necessário que a base de dados contenha todos os parâmetros preenchidos para cada paciente a ser analisado, o que trará uma redução grande no número de instâncias visto que o quantitativo de informações de A1c é muito inferior ao quantitativo de Hemogramas, representado na Figura 4. Partindo deste princípio, foi realizada a redução da base de dados, de modo a manter

somente as linhas que possuem todas as informações, cujos quantitativos de registros são apresentados na Figura 5.

Figura 5 - Quantidade de registros por etapa do processo de filtragem dos dados.



Fonte: Elaborado pelo autor

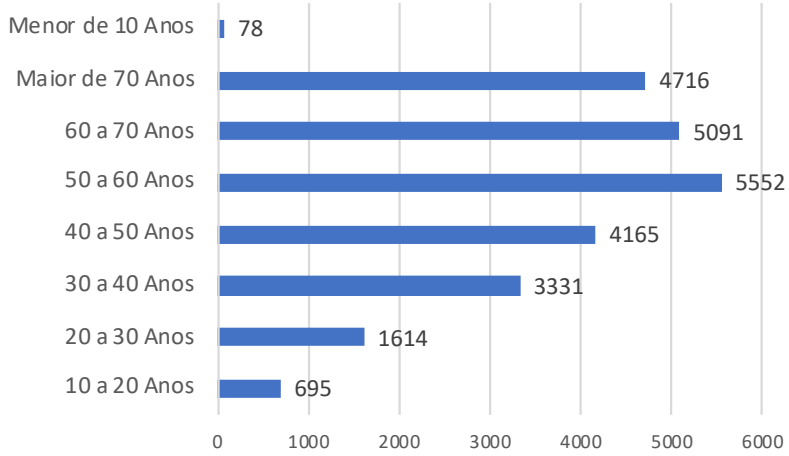
O processo de redução partiu da base de dados original, analisando o Hemograma, que é a informação predominante no conjunto de dados. Na primeira etapa, foi realizada uma busca correlacionando todos os pacientes que possuem as informações de Hemograma e Glicose. A informação de Glicose foi estabelecida para esta etapa por ser o segundo elemento mais presente no conjunto de dados. Ao final da busca, foram encontrados 114190 registros adequados ao padrão de busca realizado. Na segunda etapa, o resultado de A1c foi adicionado como obrigatório na busca, trazendo a base de dados correta para a realização do estudo. Ao final do processo de redução, a base de dados ficou com um total de 25242 registros. Este processo foi realizado utilizando o Microsoft Excel, que dispõe de ferramentas de filtros avançados para aplicação das buscas necessárias na base de dados.

Em análises clínicas, por motivos técnicos, alguns procedimentos acabam não podendo ser executados e, consequentemente, não são obtidos resultados para estes exames. Por padrão, a empresa que cedeu a base de dados insere um caractere asterisco no parâmetro de resultado quando não é possível finalizar a análise ou obter-se de um resultado confiável para determinados procedimentos. No conjunto de informações analisadas, foram encontradas algumas ocorrências dessa situação no hemograma, sendo que todas foram removidas da base de dados.

Durante o processo de exportação, considerou-se exportar a data de nascimento dos indivíduos, devido à idade ser um fator importante para a análise de

riscos e evolução da Diabetes Mellitus. Prezando manter a identidade dos pacientes anônima, os valores da coluna Data de Nascimento foram substituídos pelo valor da idade em anos. Para obter a idade, foi utilizada a função DATADIF do Excel, que permite realizar o cálculo de diferença de tempo entre duas datas e personalizar o formato de saída, neste caso utilizado em anos. Assim, o cálculo das idades foi definido como a diferença entre a data de realização dos exames e a data de nascimento de cada indivíduo, para obter a idade do paciente no momento em que os exames foram realizados. Para visualização da composição de idades no conjunto de dados, foi gerada uma escala de faixas etárias com intervalo de 10 anos, onde atribuiu-se cada indivíduo à sua respectiva classe.

Figura 6 - Quantidade de indivíduos por faixa etária

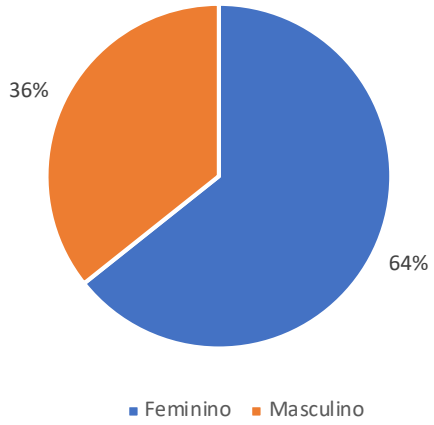


Fonte: Elaborado pelo autor.

Analisando a Figura 6 acima, é visível que a grande fração dos pacientes se concentra entre maiores de 40 anos, devido à maior necessidade da realização de exames de rotina, seja de forma corretiva ou preventiva. Idades inferiores a 20 anos são o menor grupo, possivelmente porque o tipo 2 de diabetes, que ocorre em 95% dos pacientes com DM, apresenta sintomas após os 40 anos (FONSECA et al., 2018).

Gerou-se também um gráfico da distribuição percentual dos indivíduos por sexo, reportado na Figura 7 abaixo.

Figura 7 - Distribuição da amostra de indivíduos por sexo



Fonte: Elaborado pelo autor.

Modena, 2015, aponta que o número de homens que procura um médico para consultas é 30% menor que o número de mulheres, o que é representado na base de dados estudada, que também é predominada por pessoas do sexo feminino.

Com base nos trabalhos correlatos e na revisão de literatura de termos técnicos da área da saúde, foram selecionados os parâmetros hematológicos para composição da base de dados. A inserção das frações para cada área do hemograma na base de dados poderá ser um diferencial na próxima etapa desta pesquisa, pois podem se encontrar correlações apresentadas em parâmetros específicos do hemograma nos pacientes diabéticos ou pré-diabéticos.

A extração da base de dados foi realizada com sucesso e o formato de exportação foi adequado para a visualização das informações. A consolidação dos dados e geração de gráficos oportunizou encontrar a diferença no quantitativo de Hemogramas, quando comparado a Glicoses e A1c, o que culminou no processo de redução da base, para que os registros contenham todas as informações necessárias, garantindo que a confiança nos resultados gerados nas próximas etapas do trabalho de pesquisa não seja afetada.

4. MODELO CLASSIFICADOR

Abordagens de aprendizado de máquina e mineração de dados sobre Diabetes Mellitus são áreas de grande interesse por profissionais médicos, quando se trata de aspectos ligados ao diagnóstico ou administração clínica. (KAVAKIOTIS et al., 2017).

Scott, 2018, afirma que técnicas de aprendizado de máquina são práticas computacionais capazes de extrair conhecimento de grandes bases de dados, apoiando o diagnóstico clínico, a previsão de riscos e assessorando o acompanhamento de doenças graves. Algoritmos são capazes de relacionar experiências clínicas, classificando pacientes com diabetes ou outras patologias previstas no modelo classificador.

No desenvolvimento de um modelo de classificador do status glicêmico, a etapa inicial foi a escolha das técnicas de aprendizado de máquina mais adequadas ao tipo de dados utilizado. Em pesquisa bibliográfica, encontrou-se um trabalho desenvolvido por Worachartcheewan et al. (2013), que apresentou uma proposta semelhante à desta pesquisa, o qual foi selecionado para uso como base para composição do modelo de classificador. Sendo assim, realizou-se inicialmente a tentativa de reprodução do método classificador implementado neste trabalho correlato e, após obter os resultados, se apresenta um modelo que mais se adequa ao conjunto de dados utilizado nesta pesquisa.

Este capítulo apresenta as técnicas e ferramentas utilizadas para o desenvolvimento de um modelo de classificador de status glicêmico. Atuando de modo interdisciplinar com a área da saúde neste trabalho de pesquisa, o capítulo aborda também conceitos de Inteligência Artificial que serão utilizados no decorrer do trabalho, para embasamento de leitores que não sejam da área da computação.

4.1 APRESENTAÇÃO DOS MÉTODOS

A Máquina de Vetores de Suporte (*Support Vector Machines* ou SVMs), desenvolvida por Vapnik (1995) é considerada a primeira aplicação prática do aprendizado estatístico (SOUZA, 2014). É uma das implementações de Aprendizado de Máquina mais utilizadas atualmente, apresentando resultados satisfatórios na área de bioinformática, principalmente na análise de grandes volumes de dados (DIAS, PASCUTTI, SILVA, 2016). Em seu funcionamento, as SVMs buscam realizar a

classificação de padrões em bases de dados, empregando um princípio de indução, obtendo o aprendizado e gerando conclusões genéricas a partir de um modelo com resultados conhecidos. Deste modo, podemos classificar este aprendizado em dois modos: Supervisionado e não supervisionado.

O modelo de aprendizado supervisionado é caracterizado pela apresentação do conhecimento do ambiente de dados por um conjunto de exemplos, que determinam as entradas e saídas desejadas. O algoritmo de aprendizado constitui a representação do modelo classificador a partir dos exemplos apresentados. No aprendizado não supervisionado, não há apresentação de um conjunto de exemplos ao modelo, fazendo com que o algoritmo de aprendizado constitua a representação das saídas com base nos dados de entrada (BUNKER; THABTAH, 2017). Na base de dados utilizada para a realização deste trabalho, adicionamos uma coluna com o resultado da dosagem de Hemoglobina Glicada de cada registro, compondo o conjunto de exemplos a serem apresentados para o algoritmo de aprendizado de máquina, caracterizando desta forma, o modelo de aprendizado supervisionado.

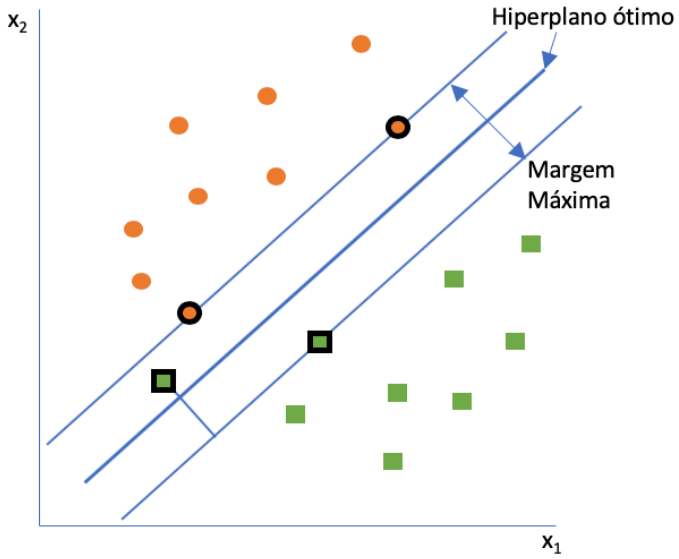
O conjunto de exemplos apresentado para a máquina de vetores de suporte pode ser expressado na forma ($\mathbf{x}_i$, y_i), onde $\mathbf{x}_i$ representa os dados de entrada e y_i seu exemplo de classificação, nomeado de rótulo. A partir deste conjunto de exemplos, se gera um classificador preditor, capaz de prever o valor de y_i a partir dos dados de entrada $\mathbf{x}_i$. Esta fase de geração do modelo classificador a partir dos dados ($\mathbf{x}_i$, y_i) é nomeada como treinamento do algoritmo de aprendizado de máquina (SHUANG et al., 2019).

Os problemas a serem resolvidos pelos algoritmos de SVM podem ser classificados em dois grupos: Binários ou Multiclasses, este último, tornando o ciclo de aprendizado e previsão consideravelmente mais complexo. É comum autores realizarem a combinação de saídas geradas por classificadores binários para compor a resposta de um problema de multiclasses, com o objetivo de reduzir sua complexidade (NETO, 2017).

O funcionamento básico de uma SVM pode ser descrito da seguinte maneira: dado um problema linearmente separável, onde o conjunto y_i rótulo é composto por duas classes, uma Máquina de Vetores de Suporte realiza o mapeamento do espaço de entrada, encontrando um hiperplano otimizado, que maximiza a margem de divisão das classes. Nesta divisão entre as classes, os pontos situados sobre os extremos da margem do espaço hiperplano são os chamados vetores de suporte (NETO, 2017). A

Figura 8 abaixo demonstra um hiperplano ótimo, partindo de uma entrada de dados $\mathbf{x}_i$, o conjunto binário de rótulos y_i está representado pelos quadrados e círculos, divididos pelas linhas de separação traçadas pelo algoritmo. A SVM visa encontrar a maior margem de separação entre as duas classes, para garantir uma boa acurácia do algoritmo (KINTO, 2011).

Figura 8 - Hiperplano de Vetores de Suporte

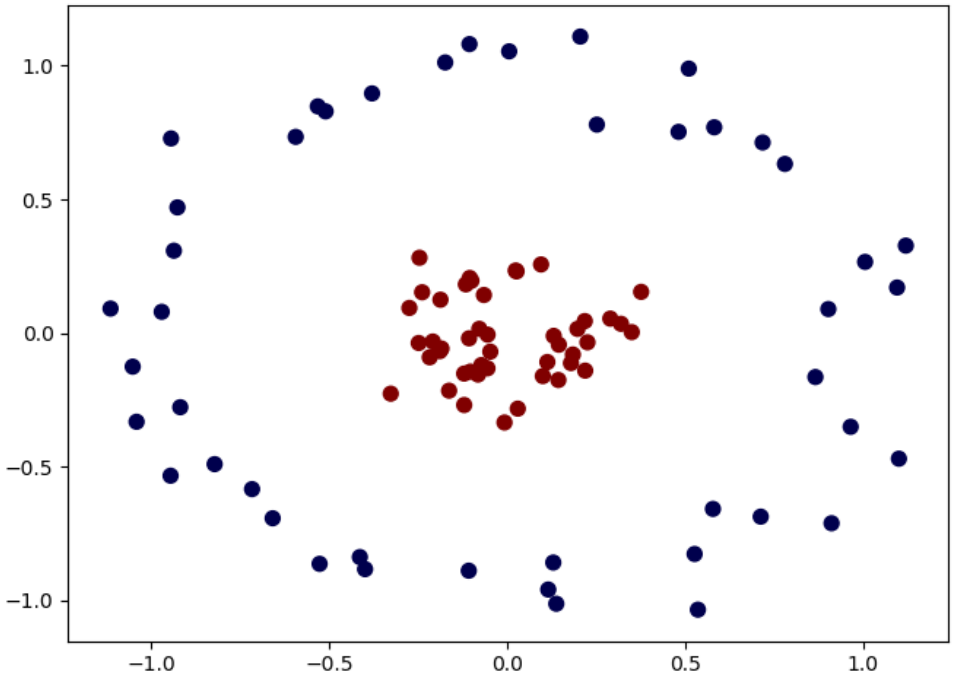


Fonte: Boechat, 2012. Adaptado pelo autor.

Na Figura 8, é possível visualizar os vetores de suporte, representados pelos quadrados e círculos com bordas pretas. É perceptível que a Figura 8 representa um hiperplano ótimo realizando a separação entre duas classes de dados lineares, onde é possível traçar uma única linha realizando a divisão entre as duas classes. No caso de implementação deste trabalho, o conjunto de dados de entrada $\mathbf{x}_i$ não é composto por dados linearmente separáveis, o que será apresentado nos resultados do modelo classificador e faz a necessidade do uso de SVMs não lineares, que possibilitam a criação de uma divisão curva entre os dados do hiperplano (PUPALE, 2018).

Para a realização do mapeamento em casos de dados não lineares, utiliza-se uma função *kernel*, capaz de mapear o domínio do espaço de entrada ($\mathbf{x}_i$) para um novo espaço.

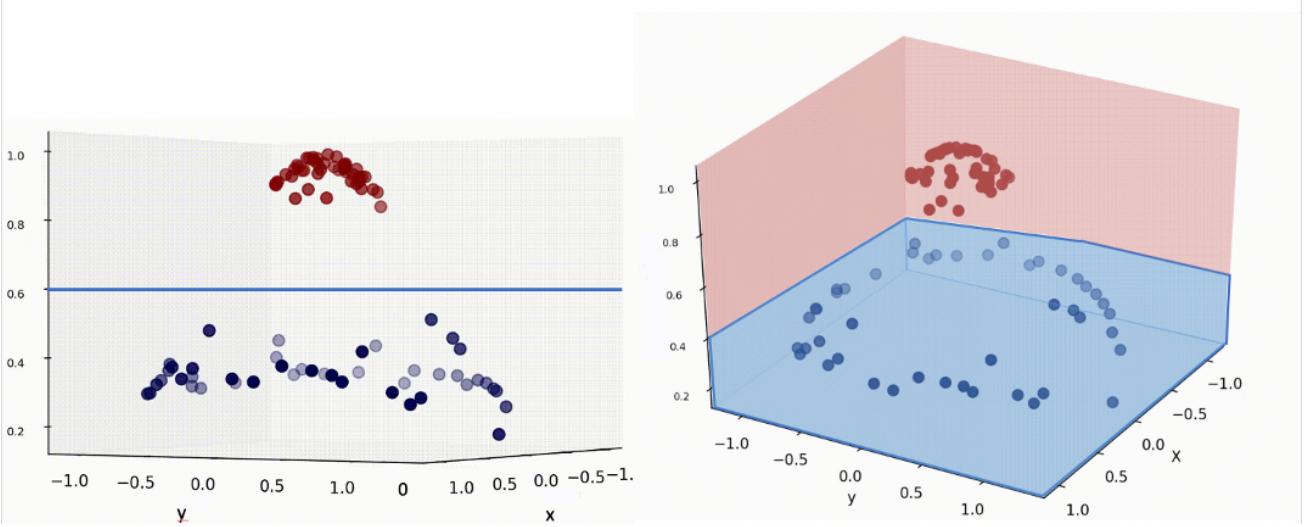
Figura 9 - Dados não linearmente separáveis



Fonte: Bhattacharyya, 2018.

A Figura 9 representa dados não linearmente separáveis, onde se faz necessário o uso de *kernel* para apoio na classificação. O princípio de uso de um *kernel* é a realização do mapeamento do conjunto de dados em um espaço dimensionável mais alto, onde fica possível encontrar um hiperplano capaz de separar as amostras de dados (BHATTACHARYYA, 2018). A Figura 10 abaixo relaciona os dados apresentados na Figura 9, separados utilizando um *kernel*.

Figura 10 - Separação de dados por *kernel*



Fonte: Bhattacharyya, 2018. Adaptado pelo autor.

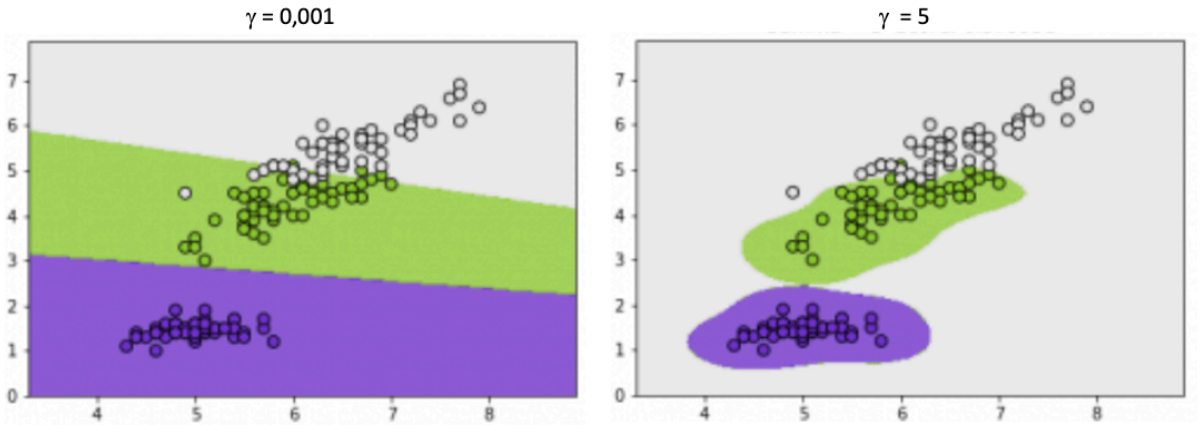
Analisando a Figura 10, é possível visualizar que o *kernel* realizou a criação de mais uma dimensão no hiperplano, que possibilitou a separação dos dados por uma linha, o que não era possível no hiperplano de duas dimensões, representado na Figura 9.

Woracharcheewan et al. (2013), em seu estudo de aplicação de aprendizado de máquina à previsão dos níveis de diabetes a partir de dados hematológicos, faz o uso de *Radial Basis Function* (RBF) *kernel* para otimizar a classificação dos dados, devido à mesma necessidade de uso de uma base de dados não linearmente separáveis.

O desempenho e acurácia dos modelos classificadores que utilizam Máquinas de Vetores de Suporte está diretamente ligado aos parâmetros definidos previamente à execução dos algoritmos. Drawson (2019) aponta que o parâmetro mais importante na execução de uma SVM é o C, que define o equilíbrio entre a correta classificação dos dados e a adaptabilidade do modelo em receber dados diferentes dos de treinamento. Portanto, em uma execução com valores elevados no parâmetro C, está predito que a amostra de dados possui os exemplos mais extremos possíveis. Assim, os futuros dados inseridos neste modelo classificador estarão mais distantes dos pontos em que o modelo foi treinado.

De mesmo modo que temos parâmetros influenciando no desempenho da SVM, o RBF *kernel* também tem um parâmetro que define sua precisão, chamado gama (γ). Os limites das classes separadas pelo *kernel* são estabelecidos pelo valor de gama, sendo que, quanto maior o valor, maior o isolamento dos dados feito pelo *kernel* no hiperplano (Drawson, 2019). Um exemplo de variação da eficácia do *kernel* em razão da variação de gama é exposto pela Figura 11 abaixo.

Figura 11 - Efeito do parâmetro gama no *kernel*



Fonte: Drawson, 2019. Adaptado pelo autor.

Analisando a Figura 11 acima, é possível visualizar o estreitamento da seleção dos dados pelo *kernel*, em razão da variação do parâmetro gama. Quanto mais justa a seleção do *kernel* nos dados, maior a precisão do classificador. Porém, ao inserir um novo conjunto de dados, o valor de gama deverá ser revisto pois, como mudam os dados, a seleção e estreitamento realizado pelo *kernel* também serão alterados. A otimização dos parâmetros tanto da SVM quanto do *kernel* são fatores essenciais para aprimorar a acurácia do classificador.

Para todo modelo de aprendizado de máquina proposto, um plano de validação é necessário, para garantir a estabilidade do modelo e aferir a precisão do classificador. Seguindo o modelo de validação implementado por Worachartcheewan et al. (2013), utilizou-se a validação cruzada (*cross-validation*) de 10 vezes para aferir a precisão dos modelos classificadores executados no decorrer deste trabalho. A prática de validação cruzada divide os dados em *k* subconjuntos. Assim, o método de validação é repetido *k* vezes, sendo que, a cada vez um dos *k* subconjuntos é utilizado como conjunto de teste/validação e os outros *k-1* subconjuntos são consolidados para formar um só conjunto de treinamento. A estimativa de erro é calculada sobre todas as *k* tentativas de obter a eficácia do modelo. O uso do modelo de validação cruzada é eficaz para aferir a precisão de um classificador pois utiliza a maioria dos dados para ajuste e, de mesma forma, utiliza também a maioria dos dados no conjunto de validação. Em geral, são utilizados *k=5* ou *k=10* para a validação cruzada (GUPTA, 2017).

4.2 MODELO PROPOSTO E RESULTADOS

Dado o embasamento teórico acerca dos algoritmos utilizados, esta sessão apresentará a aplicação da Máquina de Vetores de Suporte aos dados, tarefa realizada com o emprego do *Waikato Environment for Knowledge Analysis*, popularmente chamado de Weka, um pacote de ferramentas para aplicação de técnicas de aprendizado de máquina, pré-processamento de dados e análise de resultados, desenvolvido pela Universidade de Waikato, Nova Zelândia (WITTEN; FRANCK, 2017).

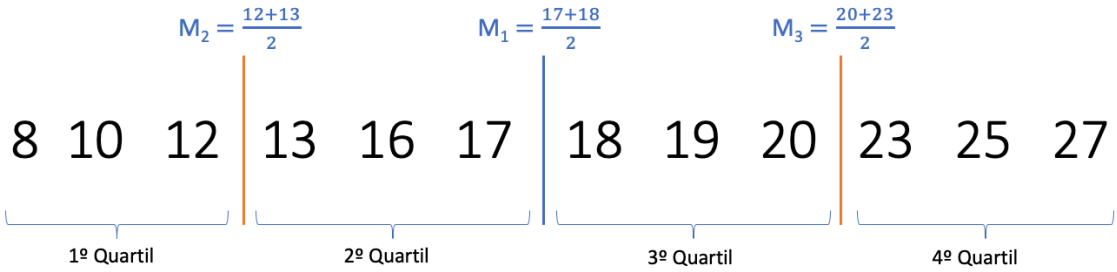
A área do diagnóstico da Glicose através de métodos de Inteligência Artificial tem apresentado diversos trabalhos publicados, cada um empregando a metodologia que mais se adequa ao seu conjunto de dados e objetivos. Worachartcheewan et al.

(2013), apresenta um modelo de classificador de status glicêmico a partir dos dados hematológicos de pacientes, correlacionando com a proposta deste trabalho. No primeiro experimento desta pesquisa, o trabalho de Worachartcheewan et al. (2013) foi reproduzido em seus mesmos padrões, mas com a base de dados utilizada para este estudo.

Citado detalhadamente na sessão de trabalhos correlatos à esta pesquisa, Worachartcheewan et al. (2013) realizou sua pesquisa em dados obtidos através de uma restrita quantia de pacientes que realizaram procedimentos do tipo check-up de saúde em uma clínica na Tailândia. Nesta etapa de reprodução do trabalho correlato, observou-se que foram utilizadas somente as variáveis hematológicas de Leucócitos, Eritrócitos, Hemoglobina e Hematócrito. Sabendo que o subconjunto de dados obtido para utilização neste trabalho possui algumas variáveis além das utilizadas no trabalho correlato, essas foram removidas do conjunto de dados na realização deste experimento, para que a representação seja aproximada do trabalho original.

Worachartcheewan et al. (2013), apresenta a divisão dos parâmetros hematológicos em quatro grupos, baseados nos quartis, como pré-processamento dos dados hematológicos. Bakker et al. (2004) define os quartis como um método de divisão de valores em conjuntos iguais. Martins (2014), define o processo de divisão dos quartis, que inicia ordenando o conjunto de dados em ordem crescente e calculando o valor da mediana dos dados(M_1). Então, o conjunto de dados fica dividido em dois grupos, sendo um à esquerda da mediana M_1 e outro à direita da mediana M_1. Em seguida, calcula-se a mediana dos números do conjunto à esquerda da mediana M_1, formando a mediana M_2. O primeiro quartil (Q_1) tem seu limite definido do início do conjunto de dados até o valor da mediana M_2. O segundo quartil (Q_2) define-se a partir do valor de M_2 até M_1. Para a definição dos quartis terceiro (Q_3) e quarto (Q_4), utiliza-se a mesma sistemática dos primeiros. Calcula-se a mediana dos números à direita da mediana M_1, formando a mediana M_3. O conjunto de valores pertencentes à Q_3 parte do valor da mediana M_2 até o valor de M_3. O quarto quartil possui seus valores determinados à direita da mediana M3 até o final do conjunto de dados. A figura 12 abaixo demonstra a divisão entre os quartis.

Figura 12 - Divisão entre quartis



Fonte: Elaborado pelo autor

Após a adequação dos dados em quartis, o processamento dos dados implementado por Worachartcheewan et al. (2013), Máquina de Vetores de Suporte foi empregada por meio da implementação do algoritmo de Mínima Otimização Sequencial (SMO), desenvolvido por John Platt. Treinar uma Máquina de Vetores de suporte requer grande capacidade de programação quadrática e recursos de processamento disponíveis. Esta solução de otimização divide o problema em séries de pequenos problemas, que são resolvidos com menor complexidade e custo computacional (PLATT, 1999). Na aplicação Weka, o algoritmo de SVM com a Mínima Otimização Sequencial está nomeado como “SMO”.

Representando fidedignamente o modelo proposto por Worachartcheewan et al. (2013), os parâmetros da SVM e do *kernel* foram ajustados para os mesmos valores utilizados nos processamentos originais. Utilizou-se um valor de C = 2^{23} na máquina de vetores de suporte, e $\gamma = 2^{-8,5}$ no *kernel*.

Com a base de dados readequada e o classificador definido, realizou-se o processamento dos dados, no qual se obteve 68,60% das instâncias classificadas corretamente, valor bastante abaixo do registrado por Worachartcheewan et al. (2013), que foi de 98,42%. Entretanto, no estudo de maior acurácia, é utilizada uma base de dados com apenas 119 registros, quantidade de registros bastante diferente da utilizada neste estudo, onde a amostra de dados possui 25076 registros. A Tabela 6 abaixo mostra as quantidades de registros por classe, comparando as duas fontes de dados.

Tabela 6 - Comparação da quantidade de registros nas bases de dados

Classificação	Quantidade de registros	% Em relação ao total
Base de dados utilizada no estudo de Worachartcheewan et al.	190	
Normal	107	56%

Pré-Diabetes	59	31%
Diabetes	24	13%
Base de dados utilizada neste estudo	25076	
Normal	10754	43%
Pré-Diabetes	6995	28%
Diabetes	7327	29%

Fonte: Elaborado pelo autor.

A Tabela 6 representa o comparativo dos volumes de registros nas bases de dados utilizadas nos dois estudos correlatos. É possível notar que Worachartcheewan et al. (2013) utilizou uma base de dados com uma quantidade de registros consideravelmente menor do que a base utilizada neste estudo. Outra questão importante é que o percentual de registros da classe “Diabetes” no estudo de Worachartcheewan et al. (2013) é 45% menor ao percentual de registros na mesma classe, da base utilizada neste estudo.

Diante da situação de divergência nos volumes de dados, realizou-se a redução da base de dados em bases menores, para comparação de resultados. A redução foi feita utilizando a função *Resample*, disponível no pacote Weka. Esta função tem como resultado uma sub amostra randômica do *dataset*, que foi reduzido de forma percentual linear, até chegar ao número de registros correspondente ao utilizado no estudo comparado. Com as reduções realizadas nas bases de dados, os algoritmos foram executados novamente e os resultados estão expostos na Tabela 7, abaixo.

Tabela 7 - Comparação de resultados de processamentos

% de Redimensionamento	Quantidade de registros	% de instâncias classificadas corretamente
100%	25076	68,60%
10%	2507	61,90%
1%	250	58,40%
0,75%	190	57,36%

Fonte: Elaborado pelo autor.

Analisando a Tabela 7 acima, é possível visualizar que reduzindo o tamanho da amostra de dados utilizada no algoritmo, se reduziu também a precisão na classificação. Diante desses dados, levanta-se um questionamento acerca da grande diferença de precisão entre as duas representações do modelo classificador. A baixa quantidade de registros da classe “Diabetes” presente na base de dados utilizada por

Worachartcheewan et al. (2013) poderia gerar uma tendência no classificador, visto que há poucos registros dessa classe para aprendizado. A Tabela 8 abaixo apresenta as matrizes de confusão dos classificadores executados.

Tabela 8 - Comparação entre matrizes de confusão

Classificação	Normal	Pré-Diabetes	Diabetes
Worachartcheewan et al. (2013)			
Normal	107	0	0
Pré-Diabetes	0	58	1
Diabetes	0	2	22
Este estudo			
Normal	72	6	0
Pré-Diabetes	46	2	3
Diabetes	20	4	37

Fonte: Elaborado pelo autor.

Analisando as matrizes de confusão apresentadas na Tabela 8, aplicando o modelo classificador em uma base de dados com maior quantidade de instâncias rotuladas como “Diabetes”, é perceptível que esta foi a classe com maior número de erros do classificador. O trabalho de Worachartcheewan et al. (2013) possui 8,3% de erro na classificação das instâncias rotuladas como “Diabetes”, enquanto o resultado obtido por este trabalho foi de 39,3% de erro na mesma classe. Comparando os volumes de dados para treinamento e classificação, a quantidade de instâncias pertencentes à classe “Diabetes” submetidas ao algoritmo foi 2,54 vezes maior do que a base de dados utilizada por Worachartcheewan et al. (2013), o que poderia representar uma relação da redução da acurácia do classificador com a complexidade da base de dados utilizada. A classe “Pré-Diabetes” representou um percentual de erro elevado para a base utilizada neste estudo, com somente 2 instâncias corretamente classificadas, representando somente 4% de acerto nesta classe.

Inicialmente, propomos a reprodução do modelo classificador desenvolvido por Worachartcheewan et al. (2013), na base de dados obtida para realização deste estudo, buscando avaliar as correlações e reprodutibilidade do estudo realizado na Tailândia, nos dados da população da Encosta da Serra do Rio Grande do Sul. Obtivemos uma diferença de 29,82% na acurácia entre os resultados obtidos nas duas bases de dados, diferença bastante grande considerando que os estudos seguiram os mesmos padrões de processamento dos dados. Diante desta grande diferença,

iniciou-se o processo da otimização do classificador desta pesquisa, buscando a melhoria da acurácia de classificação obtida para a base de dados da população estudada nesta pesquisa.

Huang et al. (2019), aponta que, devido ao classificador SVM ser baseado nas distâncias entre os vetores de suporte do hiperplano, este método não pode ser aplicado diretamente em dados categóricos. Complementa que ao utilizar dados numéricos, a precisão do classificador pode ser aumentada devido ao aumento da sensibilidade da posição dos dados no hiperplano. Nas classificações realizadas através do método de Worachartcheewan et al. (2013), as classes de dados foram divididas em quartis, tendo seus valores numéricos transformados em valores nominais de 1 a 4, o que segundo o estudo de Huang et al. (2019), poderia influenciar negativamente a precisão do classificador. Diante desta premissa, a base de dados foi reimportada ao software Weka, fazendo com que os valores dos parâmetros do hemograma fossem interpretados em formato numérico.

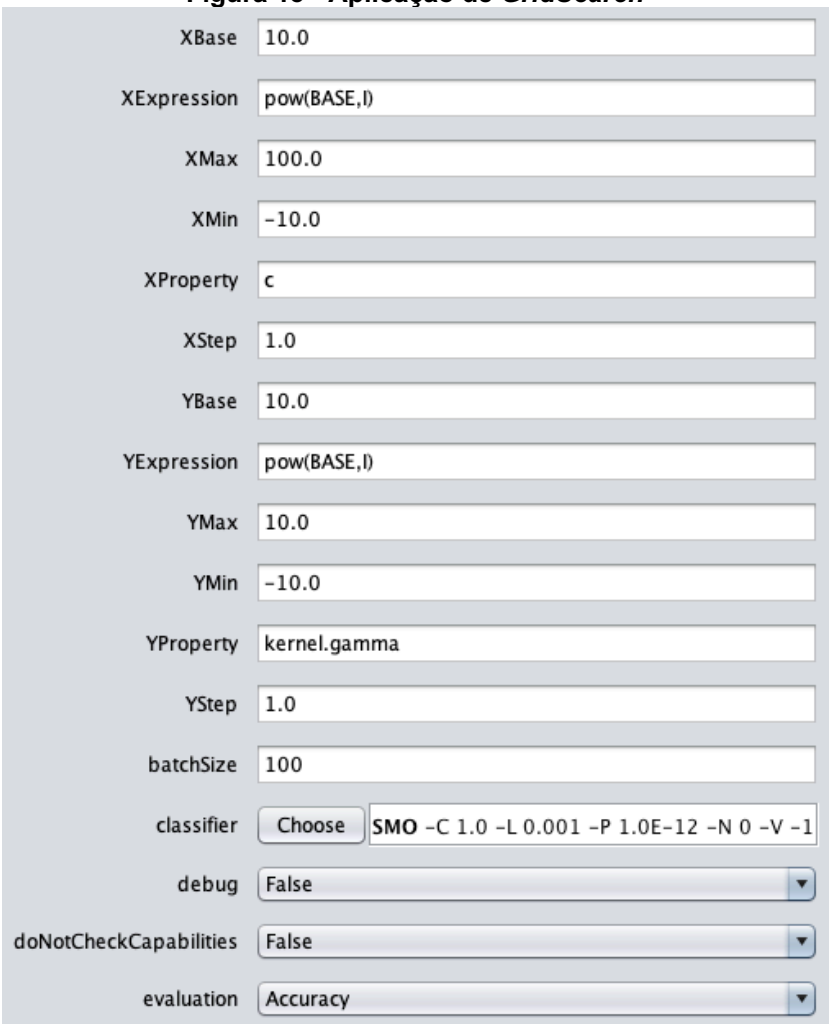
Na etapa em que o processamento foi realizado utilizando o modelo de Worachartcheewan et al. (2013), os parâmetros adequados citados pelo autor foram utilizados na máquina de vetores de suporte e no *kernel*, utilizando a otimização da performance já elaborada pelo autor em seu artigo original, visto que as bases de dados eram semelhantes. Sabemos que a essência dos algoritmos de aprendizagem é construir um modelo de aprendizado e classificar dados conforme a função objetivo a eles desenvolvida. Segundo Sun et al. (2019), a população de dados e os parâmetros de processamento influenciam drasticamente a eficácia de um modelo classificador de dados. Para prover o desenvolvimento de uma máquina de aprendizagem consistente, métodos eficazes de otimização devem ser utilizados para garantir o melhor desempenho e eficiência do modelo.

Partindo da premissa de Sun et al. (2019), sabendo que o modelo proposto por Worachartcheewan et al. (2013) não se demonstrou satisfatoriamente adequado aos dados utilizados por esta pesquisa, verificou-se a necessidade de implementação de um método de otimização dos parâmetros do classificador, buscando aumentar a acurácia. Algoritmos de otimização são largamente utilizados em casos de aprendizado de máquina. Uma das implementações mais comuns e poderosas com essa finalidade é o *GridSearch* (Khalid, 2020).

GridSearch consiste em um algoritmo de otimização aplicado à algoritmos de *Machine Learning* que possibilita selecionar os melhores parâmetros para o problema

a ser resolvido. O processo de implementação desta técnica é realizado por meio do fornecimento de um grupo de parâmetros a serem testados em diversas iterações no modelo classificador em desenvolvimento, para que os parâmetros capazes de fornecer a maior acurácia sejam expostos. Embora não seja um método disponível por padrão no software Weka (WITTEN; FRANCK, 2017), uma biblioteca de código aberto pode ser adicionada ao através de seu gerenciador de pacotes, para possibilitar a aplicação de *GridSearch* aos modelos de aprendizado de máquina disponíveis no software. A Figura 13 abaixo demonstra uma parte dos parâmetros a serem fornecidos à interface para execução do método de *GridSearch* no software Weka.

Figura 13 - Aplicação de *GridSearch*



Fonte: Elaborado pelo autor.

Como apresenta a Figura 13, na implementação do método de *GridSearch* são selecionados dois parâmetros do modelo classificador para serem otimizados, sendo inseridos nos campos *XProperty* e *YProperty*. Para cada parâmetro a ser otimizado,

são definidos os seus limites máximo e mínimo e o valor que será alterado a cada iteração. No final, o classificador a ser utilizado é selecionado e o parâmetro a ser otimizado é informado no campo “*evaluation*”. O algoritmo de SVM foi utilizado como modelo classificador, seguindo o citado no início do trabalho. Como buscamos um aumento da acurácia de classificação do modelo, a validação do *GridSearch* foi realizada pela acurácia resultante de cada iteração. Os parâmetros de maior influência no modelo classificador proposto são o C da máquina de vetores de suporte, e gama da função *kernel*. Portanto, estes foram os parâmetros inseridos para validação através do *GridSearch*. Como resultados da etapa de otimização dos parâmetros, foram obtidos valores de C = 1.0 e gama = 0.01.

Em nova execução, considerando a alteração no formato de entrada dos dados e a otimização dos parâmetros da SVM, se obteve um aumento considerável na precisão do classificador, chegando aos 73,34% de instâncias corretamente classificadas. A matriz de confusão da nova execução do algoritmo está apresentada na tabela 9, abaixo.

Tabela 9 – Matriz de confusão dos resultados obtidos

Classe	Normal	Pré Diabetes	Diabetes	% de acerto na classe
Normal	8995	1664	105	76,05%
Pré-Diabetes	2533	3831	637	55,08%
Diabetes	299	1460	5598	88,30%

Fonte: Elaborado pelo autor.

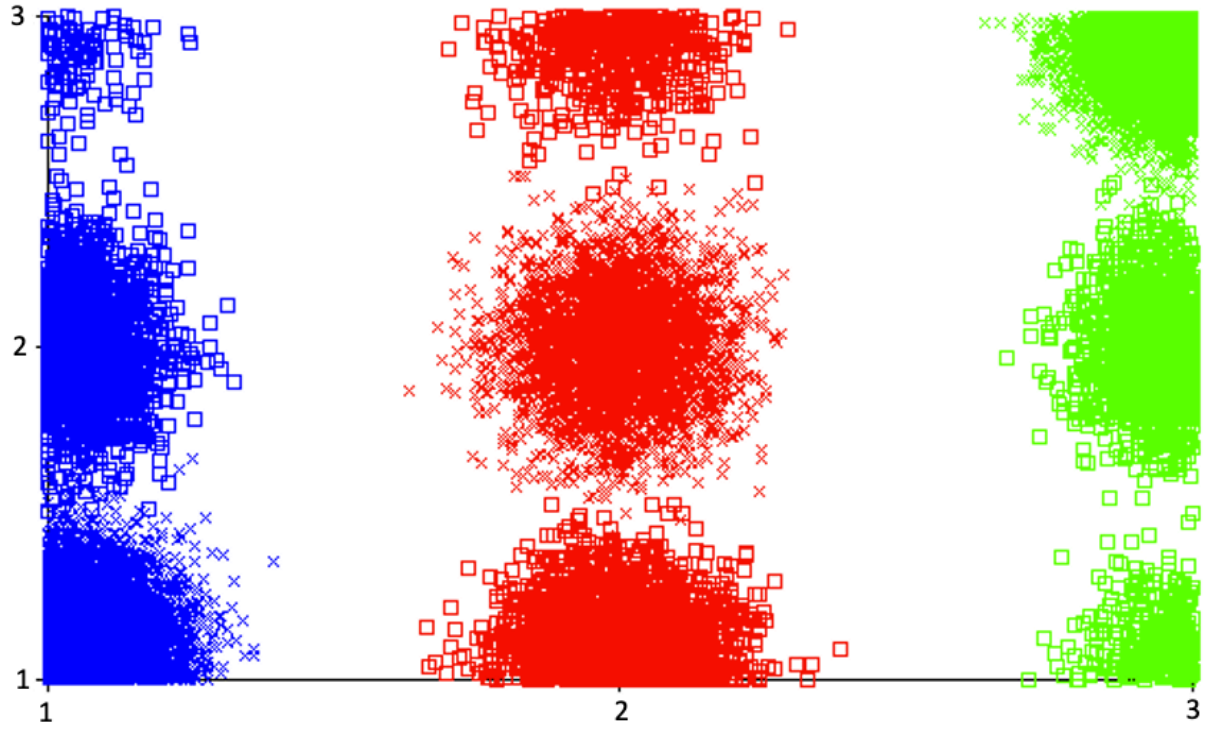
A Tabela 9 acima apresenta a matriz de confusão da execução do classificador com os dados numéricos. Foi adicionada a coluna “% de acerto na classe”, calculando o percentual de instâncias corretamente classificadas em relação ao total de instâncias em cada classe, facilitando a análise dos erros e acertos expressos pela matriz de confusão.

É possível visualizar que a classe Pré-Diabetes apresentou 55,08% de acurácia, o menor percentual de acerto na classificação, trazendo um desvio de classificação à classe Normal. Pode ser feita uma relação com SBD (2016), que relata muitos casos de pacientes que estão passando pela fase Pré-Diabetes, mas ainda não apresentaram sintomas clínicos e, consequentemente, possuem os resultados de

seus exames clínicos em uma zona limítrofe, o que poderia gerar uma área de incerteza do classificador. O mesmo vale para casos normais que foram classificados como Pré-Diabetes.

A classe Diabetes apresentou maior precisão em sua classificação, com 88,30% de instâncias corretamente classificadas, o que demonstra eficácia do modelo classificador na identificação daqueles casos em que o paciente é diagnosticado com a doença já evoluída. As classes Normal e Diabetes apresentaram o maior percentual de acerto na classificação, o que faz relação com a eficácia do modelo proposto na separação de pessoas saudáveis e pessoas com a doença em estágio evoluído. Correspondendo à matriz de confusão apresentada, a Figura 14 expõe a dispersão das classificações dos resultados no modelo classificador implementado.

Figura 14 - Gráfico de Dispersão do classificador



Fonte: Elaborado pelo autor.

A Figura 14 acima é a representação gráfica da matriz de confusão obtida na execução do classificador. Os números no eixo vertical representam as classes reais de cada instância. Os números do eixo horizontal representam as classes preditas para cada instância. Os pontos marcados com o símbolo X representam as classificações corretas e os pontos marcados com o símbolo □ representam as

classificações incorretas em cada classe. Para a interpretação dos eixos, o número 1 corresponde à classe Normal, 2 à classe Pré-diabetes e 3 para a classe Diabetes.

Como já reportado pela Matriz de confusão, é possível visualizar de forma gráfica a zona de indecisão entre as classes Normal e Pré-diabetes, registrada entre os pontos 1 horizontal e 2 vertical do gráfico, onde se sobrepõem registros corretamente e incorretamente classificados. A classe diabetes (3), apesar de apresentar registros próximos à classe Pré-diabetes no gráfico, apresentou a acurácia mais alta na sua classificação.

Neste capítulo, apresentamos a implementação do modelo classificador de status glicêmico, atingindo o objetivo geral deste trabalho. Em primeiro momento, seguimos a implementação do modelo apresentado por Worachartcheewan et al. (2013), que utilizou uma pequena amostra de dados da população da Tailândia para implementação de um modelo de classificador, que ao ser aplicado à base de dados utilizada neste estudo, apresentou baixa acurácia. Kavakiotis et al. (2017) aponta que um algoritmo com ótima precisão na classificação da diabetes para um conjunto de dados pode facilmente apresentar menor precisão em outra base de dados, o que gerou um viés de adequações no modelo classificador.

A alteração no formato dos dados de entrada do modelo classificador e aplicação do método de otimização de parâmetros na máquina de vetores de suporte possibilitou o aumento da acurácia em 5%. Analisando a matriz de confusão e o gráfico de plotagem das classificações, foi possível visualizar que se obteve uma melhor precisão na classificação dos indivíduos de fato portadores de Diabetes, quando comparada à precisão obtida na classificação dos indivíduos pré-diabéticos. Achados como o de SBD (2016) apontam que um possível motivo para esta diferença na acurácia da classificação pode ser devido a pacientes que estão passando por uma fase branda da diabetes não apresentarem sintomas clínicos ou alterações claras nos dados de seus exames laboratoriais.

A aplicação de métodos de aprendizado de máquina na pesquisa de diabetes é uma abordagem fundamental para o uso de grandes volumes de dados no apoio ao diagnóstico. O próximo capítulo descreve os apontamentos e conclusões acerca do processo de pesquisa e implementação do modelo classificador.

4 CONCLUSÃO

Diversas patologias comuns na população deixam de ser diagnosticadas ou são diagnosticadas de forma tardia devido à falta da realização de exames clínicos adequadamente. A diabetes é uma doença cujo número de portadores vêm crescendo a cada ano, atingindo 1 em cada 11 pessoas no mundo (OMS, 2019).

Profissionais médicos usam IA para maior assertividade e agilidade nos diagnósticos. Na medicina, IA utiliza de ciência de dados sob informações do paciente, possivelmente gerando resultados melhores que especialistas podem realizar (MURALI; SIVAKUMARAN, 2018).

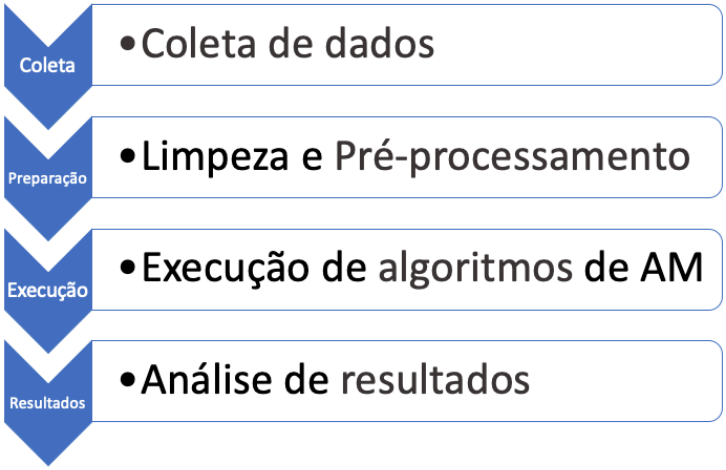
O meio científico está aquecido em pesquisas utilizando a aplicação de aprendizado de máquina na área da saúde. Foram encontradas diversas publicações de trabalhos científicos abordando o uso de inteligência artificial no tratamento de glicose, cujas principais e mais relevantes foram citadas no Capítulo 2. A pesquisa e leitura dos trabalhos correlatos demonstrou grande importância no embasamento teórico-prático para desenvolvimento deste trabalho, visualizando-se métodos e conceitos com seus resultados em aplicações diretamente relacionadas ao foco de estudo desta pesquisa. A análise das publicações através da revisão de estado da arte realizada oportunizou a visão de que, embora a diabetes seja um tema extremamente comum a nível mundial, poucos autores trabalharam propostas semelhantes à deste trabalho. De mesmo modo, não foram encontrados trabalhos relacionando índices hematológicos e glicêmicos com o uso de ferramentas de *Machine Learning* publicados ou com participação de autores brasileiros.

A tecnologia de finalidade geral com maior relevância nessa era é a Inteligência Artificial, com foco principal no Aprendizado de Máquina (BRYNJOLFSSON, 2017). A área da medicina acompanha diretamente os avanços tecnológicos em nível mundial, sendo que uma das áreas com maior objetivo de desenvolvimento e aplicação de novas tecnologias é a área da saúde. Embora o surgimento de implementações e práticas de inteligência artificial não seja relativamente novo, esta tecnologia ainda não se representa presente na rotina de grande parte dos profissionais da saúde, principalmente na região caracterizada pela base de dados utilizada neste estudo.

O processo de obtenção de conhecimento através do uso de *Machine Learning* neste estudo foi composto, basicamente, pelas etapas: 1) Coleta, análise e compreensão do conjunto de dados; 2) Aplicação das regras de pré-processamento

na base de dados; 3) Aplicação das técnicas de Aprendizado de Máquina (processamento); 4) Avaliação dos resultados. Na Figura 15 apresenta-se o fluxo básico de aplicação de ML.

Figura 15 - Etapas da aplicação de Aprendizado de Máquina



Fonte: Bell, 2014, traduzido pelo autor.

A etapa de coleta dos dados a serem utilizados neste trabalho de pesquisa foi viabilizada pela parceria com a empresa Laboratório Bom Pastor, que cedeu a grande base de dados. Em relação aos estudos correlatos, é notável uma grande diferença nos volumes de dados utilizados, o que representa uma certa dificuldade da comunidade científica para a obtenção da base de dados a ser utilizada neste tipo de estudo. Em comparação, a maior quantidade de registros encontrada em uma base de dados utilizada nos estudos correlatos foi de 6500 registros. Neste estudo, após os refinamentos necessários, a quantidade de registros da base de dados foi de 25122. Essa premissa compactua com a pesquisa de Kavakiotis et al. (2017), onde se apontam limitações nas pesquisas relacionando aprendizado de máquina e Diabetes, relacionadas à dificuldade de acesso e disponibilidade de dados biológicos à comunidade científica.

Inicialmente, o modelo classificador utilizado seguiu os padrões já implementados por Worachartcheewan et al. (2013), que analisando uma amostra de 190 registros, realizou a divisão dos dados hematológicos por quartis e utilizou máquina de vetores de suporte para a geração das classificações. Na implementação deste modelo na base de dados utilizada neste estudo, não foram obtidos resultados satisfatórios, tendo um percentual de registros corretamente classificados de 68,6%, muito abaixo do obtido pelo estudo original, que foi de 98,4%. Diante desta grande

diferença de acurácia, geraram-se alguns questionamentos acerca das diferenças nas amostras de dados, visto que Worachartcheewan et al. (2013) analisa uma amostra de dados com tamanho 93% menor do que a deste estudo, sendo que estes dados foram colhidos em um ambiente no qual os pacientes estavam realizando procedimentos de Check-up de rotina, trazendo um total de apenas 24 pacientes diabéticos presentes na base de dados, podendo facilitar a tarefa do classificador. Um estudo publicado por Sun et al. (2019) aponta que um dos pontos-chave para a obtenção de um bom modelo classificador é a consistência e volume da amostra de dados, consideração que vem de encontro direto ao achado nesta pesquisa.

A partir do resultado insatisfatório utilizando o modelo proposto pelo trabalho correlato, a recriação da base de informações a ser utilizada no classificador utilizando dados hematológicos em formato numérico e a aplicação de métodos de otimização dos parâmetros do modelo classificador possibilitou a obtenção de um aumento percentual na acurácia do modelo, apresentando 73,34% de instâncias corretamente classificadas após o ajuste.

A sessão de correlações com a área da saúde apresenta achados científicos que comprovam a existência de uma correlação representando indícios dos níveis glicêmicos nos resultados do hemograma dos pacientes. Embora não existam valores preconizados ou faixas de referência para investigação da diabetes deste modo, foi possível visualizar através dos resultados apresentados pelos trabalhos correlatos e por esta pesquisa, que é possível realizar a identificação de pacientes diabéticos a partir de seus dados hematológicos. Atualmente, órgãos de saúde preconizam que o método padrão-ouro para o diagnóstico da diabetes é a dosagem de Hemoglobina Glicada. Diante de um modelo classificador que nos permita a definição do status glicêmico nas mesmas classes que o método preconizado, mas sem a realização deste exame, gera-se um possível novo método de diagnóstico da diabetes a partir da aplicação da tecnologia de inteligência artificial.

Na sessão de exploração da base de dados, é possível visualizar que o número de pacientes que realiza a dosagem de Hemoglobina Glicada corresponde à 15% do total de pacientes que realizou o exame Hemograma. Sabemos que a avaliação das variáveis hematológicas é realizada de modo mais frequente pois traz indícios de vários tipos de doenças ao passo de que a A1c, apresenta somente um resumo dos níveis glicêmicos. Com um modelo que permita a classificação dos níveis glicêmicos a partir das variáveis hematológicas, um pré-diagnóstico ou alerta poderia ser emitido

à fração restante de 85% de pacientes que não realizaram a dosagem de A1c para a verificação do status de seus níveis glicêmicos, contribuição relevante para a prática médica e comunidade em geral, visto que aproximadamente 50% dos diabéticos desconhecem seu diagnóstico (SBAC, 2018).

Tratando-se do envolvimento com a área do diagnóstico, onde preconiza-se sempre a maior probabilidade de acerto possível por proporcionar decisões clínicas através dos resultados dos exames laboratoriais, entende-se que para o uso de um modelo classificador que visa entregar o diagnóstico ou um pré-diagnóstico a pacientes e profissionais da saúde, seria necessário que o percentual de acerto do classificador fosse maior. Partindo dessa premissa, uma oportunidade para melhoria futura deste trabalho é a busca pelo aumento da acurácia de classificação. Embora as SVMs sejam o método mais adequado para utilização nas características deste trabalho, comparativos em estudos realizados por Kavakiotis et al. (2017) apontam que outros algoritmos de aprendizado de máquina podem apresentar bons resultados na classificação de valores numéricos, tarefa realizada neste estudo. Esta é uma possível metodologia a ser utilizada para aumentar a precisão de classificação, chegando mais próximo de um cenário ideal para aplicação prática de um classificador do status glicêmico na área do diagnóstico.

Estudos acerca da obtenção de conhecimento através do aprendizado de máquina se apresentam em alta na área da saúde, com foco maior no processamento e diagnóstico de exames de imagem para detecção do câncer ou outras patologias. Enxerga-se que a área da aplicação do *Machine Learning* em resultados de exames laboratoriais é ainda pouco explorada, embora demonstre grande potencial no apoio ao diagnóstico, principalmente na intercorrelação de exames que possuam correlações clínicas e técnicas com diversos tipos de patologias. Esta deixa abre espaço para futuros trabalhos de pesquisa associando dados hematológicos na busca do diagnóstico de outras doenças além da diabetes, ou até o uso de resultados de outros exames laboratoriais em busca do auxílio ao diagnóstico com o apoio de técnicas computacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. **Imunologia celular e molecular.** 7 ed. Rio de Janeiro: Elsevier, 2012.

ALISSON, Elton. **Estudo avança no conhecimento da genética molecular do diabetes mellitus.** Disponível em: < http://agencia.fapesp.br/estudo-avanca-no-conhecimento-da-genetica-molecular-do-diabetes-mellitus/22701/ > Acesso em: 25 mai. 2020.

AMERICAN DIABETES ASSOCIATION. **Classification and Diagnosis Of Diabetes.** Diabetes Care Journal, v 43, p.14-31, 2020.

AMERICAN DIABETES ASSOCIATION. **Understanding A1c.** Disponível em: < https://www.diabetes.org/a1c >. 2020. Acesso em: 30/04/2020.

ANDRIOLO, Adgamar. **Princípios básicos de medicina laboratorial.** In: Schor N, editor. Guias de medicina ambulatorial e hospitalar da UNIFESP. 2 ed. São Paulo: Manole, 2008. cap. 1, p. 1-10.

BAKKER, Arthur. BIEHLER, Rolf. KONOLD, Cliff. **Should Young Students Learn About Box Plots?.** Roundtable, v. 29, 2004**.**

BASSO, Maik. VIEIRA, João Paulo. PARREIRA, Fábio José et al. **Sistema inteligente para Apoio ao Diagnóstico de Diabetes empregando Redes Neurais.** Anais do Encontro Anual de Tecnologia da Informação. A. 4, n. 1, p. 56-63. 2014.

BEAM, Andrew. KOHANE, Isaac. **Big Data and Machine Learning in Health Care.** JAMA. Ed. 319, p. 1317-1318. 2018.

BELL, Jason. **Machine Learning: Hands-on for developers and technical professionals**. Hoboken, NJ: Wiley, 2014.

BIADGO, Belete. MELKU, Mulugeta. ABEBE, Molla. **Hematological índices and their correlation with fasting blood glucose level na anthropometric measurements in type 2 diabetes mellitus patients in Gondar, Northwest Ethiopia.** Diabetes, metabolic syndrome and obesity: targets and therapy. V. 9, p. 91-99, 2016.

BHATTACHARYYA, Saptashwa. **Support Vector Machine: Kernel Trick.** 2018. Disponível em: < https://towardsdatascience.com/understanding-support-vector-machine-part-2-kernel-trick-mercers-theorem-e1e6848c6c4d > Acesso em: 25 mai 2020.

BRYNJOLFSSON, Erik; ROCK, Daniel; Syverson, Chad. **Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics.** NBER Working Paper Series, p. 24001, 2017.

BOECHAT, Gláucya. **MÁQUINAS DE VETORES DE SUPORTE.** Unicamp, Workshop sobre teoria de conjuntos Fuzzy. 2012. Disponível em: < http://www.dca.fee.unicamp .br/~glaucya/Apresentacao-workshop-fuzzy2012.pdf >. Acesso em: 30/04/2020.

BUNKER, Rory. THABTAH, Fadi. **A machine learning framework for sport result prediction.** Applied Computing and Informatics, v. 15, p. 27-33, 2019.

CHAR, Danton. SHAH, Nigam. MAGNUS, David. **Implementing Machine Learning in Health Care – Addressing Ethical Challenges.** The New England journal of medicine. E. 378, p. 981-983. 2018.

CONTI, Adah. **O hemograma completo: Novas ferramentas para um exame tradicional.** 2018. Disponível em: < https://alvaroapoio.com.br/inovacao/o-hemograma-completo-novas-ferramentas-para-um-exame-tradicional >. Acesso em: 23/04/2020.

DAWSON, Carl. **SVM Parameter Tuning.** Disponível em: < https://towardsdata science.com/a-guide-to-svm-parameter-tuning-8bfe6b8a452c >. 2019. Acesso em: 25 mai. 2020.

DELVES, Peter. **Autoimune Diseases.** University College London, 2018. Disponível em: <https://www.msdmanuals.com/pt-br/casa/doenças-imunológicas /reações-alérgicas-e-outras-doenças-relacionadas-à-hipersensibilidade/doenças-autoimunes>. Acesso em: 29 out. 2019.

DIAS, Maria Fernanda Ribeiro. PASCUTTI, Pedro Geraldo. SILVA, Manuela Leal. **Aprendizado de máquina e suas aplicações em bioinformática.** Revista Semioses, v.10, n.1, p. 23-31. 2016.

FAILACE, Renato. **Hemograma: Manual de Interpretação.** Centro Universitário Cesmac. Editora Artmed, ed. 6. 2018.

FAYYAD, U. M., Piatetsky Shapiro, G., Smyth, P. & Uthurusamy, R. **Advances in Knowledge Discovery and Data Mining.** 1996, AAAIPress, The Mit Press.

FONSECA, Érika. ROCHA, Tânia Pavão Oliveira. NOGUEIRA, Iara Antônia et al. **Síndrome metabólica e resistência insulínica pelo Homa-IR.** International Journal of Cardiovascular Sciences. Ed. 31, p. 201-208. 2018.

GALEA, Sandro. VAUGHAN, Roger. **Complexity in Public Health Research: A Public Health of Consequence.** American Journal of Public Health. Ed. 107, p. 1367-1368. 2017.

GALLEGO, Franciane Quintanilha. **Análise morfológica das células beta-pancreáticas de ratas diabéticas em diferentes idades de vida. 2014.** 76 f. Dissertação (mestrado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Medicina de Botucatu, 2014. Disponível em: <http://hdl.handle.net/11449/108868>. Acesso em: 29 out. 2019.

GUPTA, Prashant. **Cross-Validation in Machine Learning.** 2017. Disponível em: < https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f >. 2017. Acesso em: 26 mai 2020.

HAN, Longfei, LUO, Senlin. YU, Jianmin et al. **Rule extraction for support vector machines using ensemble learning approach: An application for diagnosis of diabetes.** IEEE Journal of Biomedical and Health Informatics, v. 19, i. 2, p. 728-734, 2015.

HUANG, Cheng-Lung. WANG, Chieh-Jen. **A GA-based feature selection and parameters optimization for support vector machines.** Expert Systems with Applications, v. 31. p. 231-240. 2019.

INTERNATIONAL DIABETES FEDERATION, **IDF Diabetes Atlas**, ed. 9, 2019.

JUNG, Chan. LEE, Won. KANG, Sung. **The risk of Metabolic Syndrome According to the White Blood Cell Count in Apparently Healthy Korean Adults.** Yonsei Medical Journal, ed. 54, p. 6115-620, 2013.

KAVAKIOTIS, Ioannis, TSAVE, Olga, SALIFOGLOU, Athanasios, MAGLAVERAS, Nicos et al. **Machine Learning and Data Mining Methods in Diabetes Research.** Computational and Structural Biotechnology Journal, v.15, p. 104-116, 2017.

KAWAMOTO, Ryuichi et al. **Hematological parameters are associated with metabolic syndrome in Japanese community.** International Journal of Basic and Clinical Endocrinology, v. 43, i. 2, p. 334-341, 2013.

KHALID, Muhammad Junaid. **Grid Search Optimization Algorithm in Python.** 2020. Disponível em: < https://stackabuse.com/grid-search-optimization-algorithm-in-python/ > Acesso em: 25 mai. 2020.

KINTO, Eduardo Akira. **Otimização e análise das máquinas de vetores de suporte aplicadas à classificação de documentos.** Disponível em: < https://www.teses.usp.br/teses/disponiveis/3/3142/tde-04112011-151337/publico/Eduardo_Kinto_Final_ PosDefesa.pdf >. 2011. Acesso em: 30/04/2020.

KUMAR, Santosh. NILSEN, Wendy. **Mobile Health Technology Evaluation.** American Journal of Preventive Medicine, v.45, i. 2, p. 228-236, 2013.

KUTLU, Mustafa. GOK, Deniz Engin. MUSABAK, Ilgen. et al. **The relationship between anxiety, coping strategies and charasteristics of patients with diabetes.** Health Qual Life Outcomes, 2008. E. 6, p.79.

LOBO, Luiz C. **Inteligência Artificial, o Futuro da Medicina e a Educação Médica.** In: Revista brasileira de Educação Médica, p. 3-8, 2018.

MARTINS, Maria Eugênia Graça. **Quartis.** Revista de ciência elementar, v. 2, p. 268. 2013.

MILOSEVIC, Dragna. PANIN, Violeta Lukic. **Relationship Between Hematological Parameters and Glycemic Control in Type 2 Diabetes Mellitus Patents.** Journal of Medical Biochemistry, ed. 38, p. 164-171, 2019.

MOSKALENSKY, Alexander. **Method of the simulation of blood platelet shape and its evolution during activation.** PLos Computational Bioloby. v14, i. 3, 2018.

MURALI, Nivetha; SIVAKIMARAN, Nivethika. **Artificial intelligence in Healthcare - A Review**. International Journal of Modern Computation, Information and Communication Technology, a. 1, e. 6, p. 103-110, 2018.

NETO, Ajalmar da Rocha. **Máquinas de vetores-suporte: Uma revisão. Disponível em: < h**ttp://abricom.org.br/lnlm/wpcontent/uploads/sites/4/2019/03/vol15 -no1-art2.pdf >. 2017. Acesso em: 26/04/2020.

NETO, Augusto Pimazoni. ANDRIOLO, Adagmar. FRALGE, Fadlo et al. **Atualização sobre hemoglobina glicada (HbA1c) para avaliação do controle glicêmico e para o diagnóstico do diabetes: aspectos clínicos e laboratoriais.** Jornal Brasileiro de Patologia.e Medicina Laboratorial. V. 45, n. 1, p. 31-48. 2009.

NYGUYEN, Thin. LARSEN, Mark. O'DEA, Bridianne et al. **Kernel-based features for predicting population health indices from geocoded social media data.** Decision Support Systems, v. 102, p. 22-31, 2017.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS, **WIPO Technology Trends 2019 – Artificial Intelligence.** 2019. World Intellectual Property Organization. v. 34, p. 146-150. 2019.

OSAKI, Milton M. **Inteligência artificial, prática médica e a relação médico-paciente.** In: Revista de Administração em Saúde, v. 18, n. 72, 2018.

OZCIFT, Akin. GULTEN, Arif. **Classifier ensemble Construction with rotation forest to improve medical diagnosis performance of machine learning algorithms.** Computer Methods and Programs in Biomedicine. v. 104, i. 3, p. 443-451, 2011.

PLATT, John. **Using Analytic QP and Sparseness to Speed Training of Support Vector Machines.** MIT Press, p. 557-563. 1999.

PUPALE, Rushikesh. **Support Vector Machines (SVM) – An Overview.** 2018. Disponível em: < https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 >. Acesso em: 26/04/2020.

ROSENFELD, Luiz Gastão. MALTA, |Deborah Carvalho. DUNCAN, Bruce Bartholow et al. **Prevalência de diabetes mellitus determinada pela hemoglobina glicada na população adulta brasileira, Pesquisa Nacional de Saúde.** Revista Brasileira de Epidemiologia. V. 22, s. 2, 2019.

SCHOLNIK, Wilson. **Exames laboratoriais: necessidade ou desperdício?.** 2018. Disponível em: < http://sbpc.org.br/wp-content/uploads/2018/06/ExamesLaboratoriaisNecessidadeOuDesperdicio20mai2018.pdf>. 2018. Acesso em: 21 ago. 2019.

SHUANG, Yu. XIONGFEI, Li. ZHANG, Xiaoli. WANG, Hancheng. **The OCS-SVM: An Objective-Cost-Sensitive SVM with Sample-Based Misclassification Cost Invariance.** IEEE, v. 7, p. 118931-118942, 2019.

SOCIEDADE BRASILEIRA DE ANÁLISES CLÍNICAS. **Qual a situação da diabetes no Brasil?.** 2018. **Disponível em: <**http://www.sbac.org.br/blog/2018/ 11/26/qual-a-situacao-da-diabetes-no-brasil/>. Acesso em: 23 ago. 2019.

SOCIEDADE BRASILEIRA DE DIABETES. **Hiperglicemia.** 2017. Disponível em: < https://www.diabetes.org.br/publico/diabetes/hiperglicemia> Acesso em: 28 out. 2019.

SOCIEDADE BRASILEIRA DE PATOLOGIA CLÍNICA E MEDICINA LABORATORIAL; **Hemoglobina glicada e glicemia média estimada.** 2020. Disponível em: < https://labtestsonline.org.br/tests/hemoglobina-glicada-e-glicemia-media-estimada >. Acesso em: 23/04/2020

SOCIEDADE DE PEDIATRIA DE SÃO PAULO. **Importância da interpretação do hemograma.** Recomendações e atualização de Condutas de Pediatria, v68, p. 3-6, 2014.

SOUZA, Elaine Barbosa. **Características das hemácias, componentes e glóbulos vermelhos do sangue.** Universidade Metodista de São Paulo, Curso de Ciências Biológicas, 2019. Disponível em: <https://www.todabiologia.com/anatomia/hemacias .htm>. Acesso em: 30 out. 2019.

SOUZA, Luis Antonio. **Interface para classificação de dados por máquina de vetores de suporte.** UNESP-Bauru, Departamento de Computação. 2014.

SUN, Shiliang. CAO, Zehui. ZHU, Han et al. **A Survey of Optimization Methods from a Machine Learning Perspective.** IEEE Transactions on Cybernetics, doi: 10.1109/TCYB.2019.2950779. p. 1-14, 2019.

VARELLA, Dráuzio. **Diabetes.** Disponível em: < https://drauziovarella.uol.com. br/doencas-e-sintomas/diabetes/>. 2019. Acesso em: 11 set. 2019.

WITTEN, Ian. FRANK, Eibe. HALL, Mark. et al. **Data mining – Pratical Machine Learning Tools and Techniques.** Morgan Kauffman, ed. 4, 2017.

WORACHARTCHEEWAN, Apilak. NANTASENAMAT Chanin. PRASERTSRITHONG, Pisit et al. **Machine Learning approaches for discerning intercorrelation of hematological parameters and glucose level for identification of diabetes mellitus.** EXCLI Journal, v. 12 p. 885-893, 2013.