



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Henrique Rezer Mosquér da Silva

**Sistema para a Medição de Glicose *In Vitro*, Baseado em Espectroscopia
Óptica e Inteligência Artificial**

Florianópolis

2023

Henrique Rezer Mosquér da Silva

**Sistema para a Medição de Glicose *In Vitro*, Baseado em Espectroscopia
Óptica e Inteligência Artificial**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica.

Orientador(a): Prof. Jefferson Luiz Brum Marques, PhD

Florianópolis

2023

Rezer Mosquér, Henrique
Sistema para a Medição de Glicose In Vitro, Baseado em
Espectroscopia Óptica e Inteligência Artificial / Henrique Rezer
Mosquér ; orientador, Jefferson Luiz Brum Marques, 2023.
133 p.

Dissertação (mestrado profissional) - Universidade Federal de
Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em
Engenharia Elétrica, Florianópolis, 2023.

Inclui referências.

1. Engenharia Elétrica. 2. Espectroscopia infravermelho. 3.
Medição in vitro. 4. Aprendizado de Máquina. 5. Glicose. I. Luiz
Brum Marques, Jefferson. II. Universidade Federal de Santa
Catarina. Programa de Pós-Graduação em Engenharia Elétrica. III.
Título.

Henrique Rezer Mosquér da Silva

**Sistema para a Medição de Glicose *In Vitro*, Baseado em
Espectroscopia Óptica e Inteligência Artificial**

O presente trabalho em nível de Mestrado foi avaliado e aprovado, pela banca examinadora composta pelos seguintes membros:

Prof. Glauco Cardozo, Dr.
Instituto Federal de Santa Catarina

Prof. Cristiano Rodrigo Bohn Rhoden, Dr.
Universidade Franciscana

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Engenharia Elétrica.

Insira neste espaço a
assinatura digital

Prof. Telles Brunelli Lazzarin, Dr
Coordenação do Programa de Pós-Graduação

Insira neste espaço a
assinatura digital

Prof. Jefferson Luiz Brum Marques, PhD
Orientador

Florianópolis, 2023.

Aos meus amados familiares, com dedicação especial à memória eterna da minha querida avó, cujo amor e inspiração permanecerão comigo para sempre.

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todas as pessoas e instituições que desempenharam um papel fundamental na realização deste mestrado. Primeiramente, minha gratidão vai para minha família, que sempre esteve ao meu lado, apoiando e incentivando meus estudos. A minha companheira, Ingrid Rosales, merece um agradecimento especial por seu constante apoio e compreensão ao longo dessa jornada acadêmica.

Ao meu orientador, Jefferson Luiz Brum Marques, expresso minha profunda gratidão pela sua paciência, orientação e apoio inestimável durante todo o período do mestrado. Sua orientação foi crucial para o desenvolvimento desta pesquisa.

Não posso deixar de reconhecer o apoio financeiro da CAPES, que tornou este projeto possível do mesmo modo que torna possível a pesquisa científica brasileira.

Aos professores do programa de pós-graduação do IEB-UFSC, agradeço pelos valiosos ensinamentos e orientações ao longo deste caminho acadêmico. Suas contribuições foram fundamentais para o meu crescimento como pesquisador.

Em resumo, agradeço a todos que de alguma forma contribuíram para a conclusão deste mestrado.

Muito obrigado!

“Há uma forma de fazer isso melhor – encontre-a.”

(Thomas Edison)

RESUMO

A Diabetes Mellitus é uma preocupação de saúde global crescente, afetando mais de meio bilhão de pessoas em todo o mundo. Os métodos convencionais de medição de glicose são dispendiosos e dolorosos, envolvendo fitas reagentes e punções nas extremidades dos dedos. Portanto, a busca por métodos não invasivos de medição de glicose se torna imperativa. O objetivo deste estudo é desenvolver uma metodologia não invasiva para medir a concentração de glicose em testes *in vitro*, com potencial aplicação futura em testes *in vivo*. Para isso, utilizamos amostras de glicose produzidas de acordo com a metodologia proposta neste estudo, a partir da glicose desidratada e água ultrapura do tipo I (Milli-Q). Inicialmente, foi definida uma solução padrão e, em seguida, produzidas amostras diluindo essa solução em água Milli-Q seguindo metodologia proposta neste estudo, que foram posteriormente medidas usando o sistema proposto. O sistema de medição envolveu o uso de quatro LEDs em comprimentos de onda de 940, 840, 750 e 660 nm como emissores, e um nanoespectrômetro com um range de detecção de 600 a 1000 nm em um intervalo de 5 nm. Neste estudo, foram utilizadas duas abordagens de aprendizado de máquina supervisionadas: modelos de regressão e modelos de classificação, com o intuito de identificar a técnica que oferecesse as métricas mais promissoras. Os modelos de regressão foram usados para prever valores contínuos da concentração de glicose na solução *in vitro*. O modelo Light Gradient Boosting Machine (LGBM) obteve as métricas mais favoráveis, com um RMSE de 34.03 (mg/dL) e um R² de 0.88 para o conjunto de teste, com aproximadamente 92% dos dados previstos no conjunto de teste dentro da zona A e zona B da grade de erro de Clarke. Já o modelo de classificação com as melhores métricas foi o Extremely Randomized Trees (EXTRA TREE), que definiu classes discretas para a concentração de glicose na solução *in vitro*. Foram exploradas três abordagens distintas: a definição de 3 classes (hipoglicemia, normoglicemia e hiperglicemia), 23 classes com intervalos de 15 mg/dL e 23 classes com rótulos representando os intervalos de cada classe (por exemplo, 85-100 mg/dL). Os resultados indicaram que a abordagem com 3 classes obteve uma acurácia e um F1-score de 95%, enquanto ambas as abordagens com 23 classes alcançaram uma acurácia e um F1-score de 86%. Comparando as classes de valores contínuos com os valores originais de concentração de glicose, observamos um RMSE de 8,88 (mg/dL) para os dados de teste. É importante destacar que ambas as abordagens de classificação obtiveram 100% dos dados na zona A da grade de erro de Clarke. Este estudo demonstrou que a abordagem com modelos de classificação possui um grande potencial para pesquisas futuras. Recomenda-se explorar uma maior variabilidade de valores de teste nos dados e a definição de intervalos menores para cada classe, como, por exemplo, 10 mg/dL, em trabalhos subsequentes.

Palavras-chave: Espectroscopia infravermelho. Medição *in vitro*. Aprendizado de Máquina. Glicose.

ABSTRACT

Diabetes Mellitus is a growing global health concern, affecting over half a billion people worldwide. Conventional methods for glucose measurement are costly and painful, involving test strips and finger pricks. Therefore, the quest for non-invasive glucose measurement methods becomes imperative. The aim of this study is to develop a non-invasive methodology for measuring glucose concentration in in vitro tests with potential future application in in vivo tests. To achieve this, glucose samples were produced following the methodology proposed in this study, using d-anhydrous glucose and type I ultrapure water (Milli-Q). Initially, a standard solution was defined, and samples were produced by diluting this solution in Milli-Q water according to the methodology proposed in this study, which were subsequently measured using the proposed system. The measurement system involved the use of four LEDs at wavelengths of 940, 840, 750, and 660 nm as emitters and a nano-spectrometer with a detection range of 600 to 1000 nm at 5 nm intervals. In this study, two supervised machine learning approaches were employed: regression models and classification models, with the aim of identifying the technique that offered the most promising metrics. Regression models were used to predict continuous values of glucose concentration in the in vitro solution. The Light Gradient Boosting Machine (LGBM) model yielded the most favorable metrics, with an RMSE of 34.03 (mg/dL) and an R² of 0.88 of the predicted data in the test set, with approximately 92% of the test dataset predictions falling within the A zone and B zone of the Clarke error grid. On the other hand, the classification model with the best metrics was Extremely Randomized Trees (EXTRA TREE), which defined discrete classes for glucose concentration in the in vitro solution. Three distinct approaches were explored: defining 3 classes (hypoglycemia, normoglycemia, and hyperglycemia), 23 classes with 15 mg/dL intervals, and 23 classes with labels representing the intervals of each class (e.g., 85-100 mg/dL). The results indicated that the 3-class approach achieved an accuracy and F1-score of 95%, while both 23-class approaches achieved an accuracy and F1-score of 86%. Comparing the classes of continuous values with the original glucose concentration values, observers noted an RMSE of 8.88 (mg/dL) for the test data. It is important to note that both classification approaches achieved 100% of the data within the A zone of the Clarke error grid. This study demonstrated that the classification model approach has great potential for future research. It is recommended to explore a wider range of test values and define smaller intervals for each class, such as 10 mg/dL, in subsequent work.

Keywords: Infrared Spectroscopy. In vitro Measurement. Machine Learning. Glucose.

LISTA DE FIGURAS

Figura 1: Evolução de indivíduos com diabetes mellitus	23
Figura 2: Método enzimático-amperométrico para medição da concentração de glicose in vitro.....	25
Figura 3: Diagrama de blocos do dispositivo para medição de glicose com método de picada no dedo.....	26
Figura 4: Esquema simplificado ilustrando a espectroscopia Raman.	30
Figura 5: (a) Um esquema simplificado ilustrando a espectroscopia de impedância tecidual, (b) Um modelo elétrico para uma única célula vermelha do sangue.	31
Figura 6: Princípio da polarimetria óptica no olho para monitoramento de glicose. ...	33
Figura 7: O espectro eletromagnético vai desde os raios gama (0,0001 nm) até as ondas de rádio (100 m).	34
Figura 8: Tipos de interações entre luz (fótons) e tecido.....	35
Figura 9: Modos de medição NIRS. (a) Transmissão transparente, (b) transmissão transparente dupla, (c) reflectância, (d) reflectância difusa e (e) transmissão difusa.	37
Figura 10: Projeção Fischer L-Glicose E D-Glicose	38
Figura 11: Esquema da mutarrotação da D-glicose. A interconversão entre as formas em anel α -D-glicose e β -D-glicose ocorre por meio da forma linear γ -D-glicose.....	39
Figura 12: Coeficiente de absorção molar (hemoglobinas).	40
Figura 13: Estruturas Químicas da Glicose	42
Figura 14: Diferentes estratégias subjacentes para sistemas de espectrômetros miniaturizados surgiram nos últimos 30 anos. (A) Óptica dispersiva miniaturizada. (B) Filtros sintonizáveis ou em matriz de banda estreita. (C) Sistemas baseados em transformada de Four	45
Figura 15: (Figura à esquerda) Pequeno Conjunto de treinamento bidimensional composto por duas classes distintas, vermelha e azul. (Figura direita) Um modelo linear é treinado para separar as duas classes.	50
Figura 16: (Esquerda) Conjunto de dados de treinamento de brinquedo de dez corporações com seus preços de ações e valores de receita associados. (Direita) Modelo linear é ajustado aos dados.	52
Figura 17: Ilustração conceitual de um estudo associação ampla do genoma empregando regressão.	53

Figura 18: Representação de regressão por SVM.	54
Figura 19: Representação de uma regressão por DT.	55
Figura 20: Exemplificação Rede neural feedforward com uma camada oculta. (FFNN)	56
Figura 21: Divisão Hold-out.	60
Figura 22: Divisão dos dados utilizando Validação Cruzada, 3 camadas.	61
Figura 23: Diagrama de blocos do sistema proposto para predição de concentração de glicose em meio <i>in-vitro</i>	62
Figura 24 : Pinout NSP32m DBK Development Board Kit.	64
Figura 25: Estrutura do chip-scale espectrômetro.	65
Figura 26: Circuito Módulo emissor.	68
Figura 27: Circuito módulo receptor	69
Figura 28: Circuito módulo digital.	70
Figura 29: Gráfico equação da reta, realizado no Statistica.	72
Figura 30: Bloco de plástico para medição <i>in vitro</i>	73
Figura 31: Fluxograma de criação de modelo e processo de teste de modelo.	76
Figura 32: Predição modelos com melhores hiperparâmetros.	81
Figura 33: Atributos com maiores relevância frente aos modelos.	83
Figura 34: Predição modelos com melhores hiperparâmetros e 22 atributos.	84
Figura 35: Predição do conjunto de teste.	86
Figura 36: Predição do conjunto de teste com 22 atributos.	87
Figura 37: Predição do conjunto de Novas Leituras.	89
Figura 38: Predição do conjunto de Novas Leituras abordagem de 22 atributos.	90
Figura 39: Resposta espectral para LED 640, 750, 860 e 940 nm, para 3 regiões de glicemia.	92
Figura 40: Grade de Erro de Clarke.	93
Figura 41: Predições conjunto de teste, aplicados na Grade de erro de Clarke.	94
Figura 42: Conjunto de Teste com 22 Atributos - Representação nas Grades de Erro de Clarke.	95
Figura 43: Conjunto de Novas Leituras com todos os atributos: Representação nas Grades de Erro de Clarke.	97
Figura 44: Conjunto de Novas Leituras com 22 atributos: Representação nas Grades de Erro de Clarke.	98

Figura 45: Grade de Erro de Clarke Modelo de Regressão LGBM abordagem com 22 atributos.....	100
Figura 46: Gráfico de B&A para o Modelo de Regressão LGBM no Conjunto de Teste.	101
Figura 47: Histograma dos Resíduos das Predições do Modelo XGB no Conjunto de Teste.	102
Figura 48: Predição dados de validação.	105
Figura 49: Matriz de confusão predição dos dados de validação.....	106
Figura 50: Predições conjunto de teste. (A) Predições com métricas de RMSE e R2. (B) Matriz de confusão para o conjunto de teste.	107
Figura 51: Grade de Erro de Clarke (valores contínuos).....	108
Figura 52: Gráfico de Bland-Altman para o Modelo Extra Tree no Conjunto de Teste.	109
Figura 53: Histograma dos Resíduos das Predições do Modelo Extra Tree no Conjunto de Teste.	109
Figura 54: Matriz de confusão para o conjunto de teste.....	111
Figura 55: Matriz de confusão para 3 classes para o conjunto de teste.....	112
Figura 56: Melhores modelos ajustados aos dados de teste.....	114
Figura 57: Modelo LGBM predições conjunto de teste, grade de erro de Clarke....	115
Figura 58: Grade de Erro de Clarke divindade em 3 regiões. Esquerda superior (Hipoglicemia), direita superior (Normoglicemia) e inferior (Hiperglicemia).....	117
Figura 59: Glicemia entre de 140 mg/dL e 300 mg/dL.	118
Figura 60: Comparativo grade de erro de Clarke Modelos de Regressão(esquerda) e Classificação (Direita).	119
Figura 61: Grade de Erro de Clarke predições conjunto Novas Leituras. (Esquerda) Modelo KNN. (Direita) Modelo XGB abordagem de 22 atributos.	121

LISTA DE TABELAS

Tabela 1: Principais características dos emissores.....	63
Tabela 2: Principais características do sensor digital (HDC1080).....	66
Tabela 3: Principais características da plataforma (Portenta H7 Lite).....	67
Tabela 4: Diluições a partir da solução padrão.	71
Tabela 5: Preparação das amostras.	72
Tabela 6: Resumo do banco de dados.....	74
Tabela 7: Métricas de Validação.	82
Tabela 8: Métricas de Validação para 22 atributos.	83
Tabela 9: Hiperparâmetros definidos após busca em grade e validação cruzada com 10-Fold.	85
Tabela 10: Distribuição das predições na grade de erro de Clarke.....	95
Tabela 11: Distribuição das predições para 22 atributos na grade de erro de Clarke.	96
Tabela 12: Distribuição das predições do conjunto de Novas Leituras para todos os atributos na grade de erro de Clarke.....	98
Tabela 13: Distribuição das predições do conjunto de Novas Leituras para 22 atributos na grade de erro de Clarke.....	99
Tabela 14: Estudos que utilizaram modelos de classificação para predição do nível de glicose em meio in vitro.	103
Tabela 15: Definição das 23 Classes contínuas para Modelo de Classificação.....	104
Tabela 16: Hiperparâmetros definidos após busca em grade e validação cruzada com 10-Fold.	105
Tabela 17: Distribuição das predições para modelo Extra Tree modelo Extra Tree	108
Tabela 18: Definição das 23 Classes discretas para Modelo de Classificação.	110
Tabela 19: Métricas modelo de classificação Extra tree valores discretos.....	111
Tabela 20: Definição de 3 Classes para modelo de Classificação.	111
Tabela 21: Métricas modelo de classificação Extra tree.	112
Tabela 22: Métricas para as 3 regiões definidas.	116
Tabela 23: RMSE abordagem de classificação vs abordagem de regressão para o conjunto de teste.	118
Tabela 24: Métricas para o conjunto de Novas Leituras Modelos de regressão.	120

Tabela 25: Métricas para o conjunto de teste e Novas Leituras Modelos de regressão.
..... 122

LISTA DE ABREVIATURAS E SIGLAS

ANN	Redes Neurais Artificiais
BT	Árvores De Reforço
CGM	Monitoramento Contínuo De Glicemia
CMOS	Óxido <i>Complementary Metal-Oxide-Semiconductor</i>
CV	Validação Cruzada
DM	<i>Diabetes Mellitus</i>
DM1	Diabetes Tipo 1
DM2	Diabetes Tipo 2
DMD	Matriz De Microespelhos Digitais
DMG	Diabetes Gestacional
DT	Árvore De Decisão
EGA	Error Grid Analysis
Extra tree	<i>Extremely Randomized Trees Classifier</i>
IFG	Glicemia De Jejum Prejudicada
IGT	Tolerância À Glicose Prejudicada
RF	Floresta Aleatória
FT	Transformada De Fourier
GB	Gradient Boosting
GDH	Glicose Desidrogenase
GO _x	Glicose Oxidase
ISO	Organização Internacional de Normalização
KNN	K-Nearest Neighbors
LGBM	LightGBM Regressor
LEDs	Diodos Emissores De Luz
LVF	Filtro Linear Variável
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Médio Absoluto
MIR	Espectroscopia De Infravermelho Médio
MS	MinMax Scaler
NCGM	Autoavaliação: Monitoramento Não Contínuo
NIRS	Espectroscopia De Infravermelho Próximo
OP	Polarimetria Óptica
R ²	Coefficiente De Determinação
RMSE	Erro Quadrático Médio
SMBG	Dispositivos De Automonitoramento De Glicose No Sangue
RS	Robust Scaler
SN	Sem Normalização
SS	Standart Scaler
SVM	Máquina De Vetor De Suporte
Voting	<i>Ensemble Voting</i>
XGB	Extreme Gradient Boosting Regressor

SUMÁRIO

1	INTRODUÇÃO	19
2	OBJETIVOS	20
2.1	OBJETIVO GERAL	20
2.2	OBJETIVOS ESPECIFICOS	20
3	FUNDAMENTAÇÃO TEÓRICA.....	21
3.1	DIABETES	21
3.1.1	Descrição	21
3.1.2	Uma Breve História da Automonitorização	24
3.2	MEDIÇÃO NÃO INVASIVA	27
3.2.1	Espectroscopia de Infravermelho.....	28
3.2.2	Espectroscopia Raman	29
3.2.3	Espectroscopia de Impedância.....	30
3.2.4	Polarimetria Óptica	31
3.3	ESPECTROSCOPIA DE INFRAVERMELHO.....	33
3.3.1	Aspectos Teóricos	33
3.3.2	Radiação Eletromagnética	34
3.3.3	Lei de Beer-Lambert	36
3.3.4	Modos de Medição	37
3.3.5	Glicose d-Anidra (Dextrose).....	38
3.3.6	Absorção Luz pela Glicose e Água	40
3.3.7	Bandas de Absorção da Glicose	41
3.3.8	Absorção de Luz por Componentes Sanguíneos e Tecidos	42
3.4	ESPECTRÔMETRO	44
3.5	CALIBRAÇÃO	48
3.5.1	Aprendizado Supervisionado.....	49
<i>3.5.1.1</i>	<i>Classificação</i>	<i>49</i>
<i>l.</i>	<i>Extremely randomized trees classifier (Extra tree).....</i>	<i>50</i>
<i>3.5.1.1.1</i>	<i>Métricas Modelos de Classificação</i>	<i>51</i>
<i>3.5.1.1.2</i>	<i>Acurácia</i>	<i>51</i>
<i>3.5.1.1.3</i>	<i>Recall.....</i>	<i>51</i>
<i>3.5.1.1.4</i>	<i>Precisão.....</i>	<i>51</i>
<i>3.5.1.1.5</i>	<i>F1-Score</i>	<i>51</i>

3.5.1.1.6	Matriz de Confusão.....	52
3.5.1.2	<i>Regressão</i>	52
I.	<i>Máquinas de Vetores de Suporte</i>	53
I.	<i>Decision Tree Regressor</i>	54
II.	<i>Random Forest Regressor</i>	55
III.	<i>Artificial Neural Network</i>	56
3.5.1.2.1	Métricas Modelos de Regressão.....	57
3.5.1.2.2	Coeficiente de Determinação (R ²).....	57
3.5.1.2.3	Erro Quadrático Médio (RMSE).....	57
3.5.1.2.4	Erro Absoluto Médio (MAE).....	58
3.5.1.2.5	Erro Percentual Médio Absoluto (MAPE).....	58
3.5.2	Requisitos Para Desempenho <i>In Vitro</i> e ISO 15197	58
3.5.3	Conjuntos de Treinamento, Validação e Teste	59
3.5.3.1	<i>Hold-out</i>	59
3.5.3.2	<i>Validação Cruzada</i>	60
4	MATERIAIS E MÉTODOS.....	62
4.1	EMISSOR DE INFRAVERMELHO	62
4.2	MÓDULO RECEPTOR	63
4.3	MÓDULO DIGITAL	66
5	DESENVOLVIMENTO.....	67
5.1	MÓDULO EMISSOR	68
5.2	MÓDULO RECEPTOR	68
5.3	MÓDULO DIGITAL	70
6	METODOLOGIA DA AMOSTRA.....	70
6.1	MATERIAIS	71
6.2	PREPARO DA SOLUÇÃO PADRÃO	71
6.3	LINEARIDADE	71
6.4	PREPARAÇÃO DAS DILUIÇÕES PARA PREDIÇÃO	72
6.5	MEDIÇÃO <i>IN VITRO</i>	73
6.6	PREPARAÇÃO DOS DADOS	75
7	PREDIÇÃO.....	75
7.1	CRIAÇÃO E TESTE DOS MODELOS	75
7.2	DEFINIÇÃO DE ALGORITMOS PARA A CRIAÇÃO DOS MODELOS	76
7.3	CRIAÇÃO DO MODELO	78

7.4	TESTE DO MODELO	79
7.5	VALIDAÇÃO DOS MODELOS DE REGRESSÃO	80
7.6	TESTE DO MODELO	86
7.7	NOVAS LEITURAS	89
7.8	GRADE DE ERRO DE CLARKE E ANÁLISE ESTATÍSTICA DO DESEMPENHO.....	91
7.8.1	Assinatura Óptica	91
7.8.2	Grade de Erro de Clarke	93
7.8.3	Análise Estatística do Desempenho e Distribuição dos Erros	100
7.9	MODELO DE CLASSIFICAÇÃO	103
7.9.1	Conjunto de Treinamento e Validação Classes Contínuas	105
7.9.2	Conjunto de Teste.....	106
7.9.2.1	<i>Análise Estatística do Desempenho e Distribuição dos Erros.....</i>	<i>108</i>
7.9.3	Utilizando 23 Classes (Valores Discretos).....	110
7.9.4	Utilizando Apenas 3 Classes.....	111
8	DISCUSSÃO	113
9	CONCLUSÃO	122
	REFERÊNCIAS	125

1 INTRODUÇÃO

O diabetes mellitus é um distúrbio metabólico que afeta a capacidade do corpo de processar a glicose no sangue. Pacientes diabéticos monitoram seus níveis de glicose no sangue em um esforço para mantê-los na faixa normal (aproximadamente 70 a 180 mg/dL) por meio de medicamentos, exercícios, dieta adequada etc, (SHOKREKHODAEI *et al.*, 2021). A diabetes é um grande problema de saúde que atingiu níveis alarmantes. Hoje, mais de meio bilhão de pessoas vivem com diabetes em todo o mundo (DIABETES, 2021). A prevalência global de diabetes em pessoas de 20 a 79 anos em 2021 foi estimada em 10,5% (536,6 milhões de pessoas), com projeção de crescimento de 12,2% (783,2 milhões) em 2045. A prevalência de diabetes se apresenta semelhante em homens e mulheres e prevalece mais alta naqueles com 75 a 79 anos.

Segundo (DIABETES, 2021), estima-se que a prevalência (em 2021) seja maior nas áreas urbanas (12,1%) do que nas rurais (8,3%), e nas áreas de alta renda (11,1%) em comparação com os países de baixa renda (5,5%). Espera-se um maior aumento relativo à prevalência de diabetes entre 2021 e 2045 ocorra em países de renda média (21,1%) em comparação com países de alta renda (12,2%) e baixa renda (11,9%). Os gastos globais com saúde relacionados ao diabetes foram estimados em 966 bilhões de dólares em 2021 e estão projetados para atingir 1.054 bilhões de dólares em 2045 (SUN *et al.*, 2022).

A medição da glicose no sangue geralmente exige que os pacientes usem uma lanceta e uma tira de teste de açúcar no sangue para coletar sangue de seus dedos, após um dispositivo eletrônico é usado para medir e converter os dados em um valor de glicose no sangue (YEN *et al.*, 2022).

Os medidores de glicose possuem duas partes essenciais: uma reação enzimática e um detector. A porção enzimática do medidor de glicose é geralmente embalada em um estado desidratado em uma tira descartável ou cuvette de reação. A glicose na amostra de sangue do paciente reidrata e reage com as enzimas para produzir um produto que pode ser detectado. Alguns medidores geram peróxido de hidrogênio ou um intermediário que pode reagir com um corante, resultando em uma mudança de cor proporcional à concentração de glicose na solução. Outros medidores incorporam as enzimas em um biossensor que gera um elétron que é detectado pelo medidor. Existem três reações enzimáticas principais utilizadas pelos medidores de

glicose atuais: glicose oxidase (GOx), glicose desidrogenase (GDH) e hexoquinase (HK) (TONYUSHKINA; NICHOLS, 2009).

Um sistema baseado em espectroscopia de infravermelho próximo é composto, simplificada, por emissores e receptores de radiação infravermelha. Um feixe de radiação é emitido, atravessa parte do corpo de um indivíduo e incide no receptor. A radiação infravermelha sofre modificações em suas propriedades ao interagir com substâncias (e.g., glicose, hemoglobina, gordura, osso, pele) contidas em seu caminho óptico. As modificações das propriedades são representações quantitativas características de cada uma das substâncias. Mediante a análise dos efeitos produzidos na radiação infravermelha, pressupõe-se que seja possível estimar a concentração da glicose em amostras *in vitro* e corrente sanguínea.

2 OBJETIVOS

2.1 OBJETIVO GERAL

O objetivo geral do trabalho consiste em contribuir para o desenvolvimento de uma metodologia de medição de concentração de glicose por meio não invasivo, em teste *in vitro*, baseado em princípios de espectroscopia de infravermelho próximo.

2.2 OBJETIVOS ESPECIFICOS

Para atingir o objetivo geral proposto, é necessária a conclusão dos seguintes objetivos específicos:

1. Descrever a técnica de espectroscopia de infravermelho próximo aplicada à medição não invasiva de glicose *in vitro*;
2. Desenvolver uma metodologia de instrumentação eletrônica com a capacidade de medição de glicose (*in vitro*).
3. Estabelecer métricas de desempenho e avaliar o instrumento em medição de glicose em água desmineralizada (*in vitro*);

3 FUNDAMENTAÇÃO TEÓRICA

Expõe-se na presente seção o referencial teórico necessário para o apoio à compreensão e ao desenvolvimento do presente trabalho. Inicialmente, trata-se de questões relativas à diabetes e apresentam-se as particularidades da medição não invasiva de glicose sanguínea. Em seguida, descrevem-se o método de espectroscopia de infravermelho próximo e aspectos relacionados aos sistemas de instrumentação eletrônica baseados nessa técnica.

3.1 DIABETES

Tratam-se, a seguir, de questões da doença de diabetes e da automonitorização da glicose sanguínea.

3.1.1 Descrição

O *diabetes mellitus* (DM) é uma doença crônica que ocorre quando o pâncreas de um paciente não é mais capaz de produzir insulina ou o corpo do paciente não consegue utilizar totalmente a insulina que produz (YEN *et al.*, 2022).

A insulina é um hormônio essencial produzido no pâncreas. Ela permite que a glicose da corrente sanguínea entre nas células do corpo, onde é convertida em energia ou armazenada. A insulina também é essencial para o metabolismo de proteínas e gorduras. A falta de insulina, ou a incapacidade das células de responder a ela, leva a altos níveis de glicose no sangue (hiperglicemia), que é o indicador clínico de diabetes (DIABETES, 2021).

As complicações do diabetes variam de condições agudas com risco de vida, como hipoglicemia grave ou cetoacidose, a complicações crônicas e debilitantes que afetam vários sistemas orgânicos, como retinopatia, nefropatia, neuropatia e doença cardiovascular. As estimativas da prevalência de complicações diabéticas são desafiadoras, em parte porque não existem padrões internacionalmente acordados para o diagnóstico (NICKERSON; DUTTA, 2012).

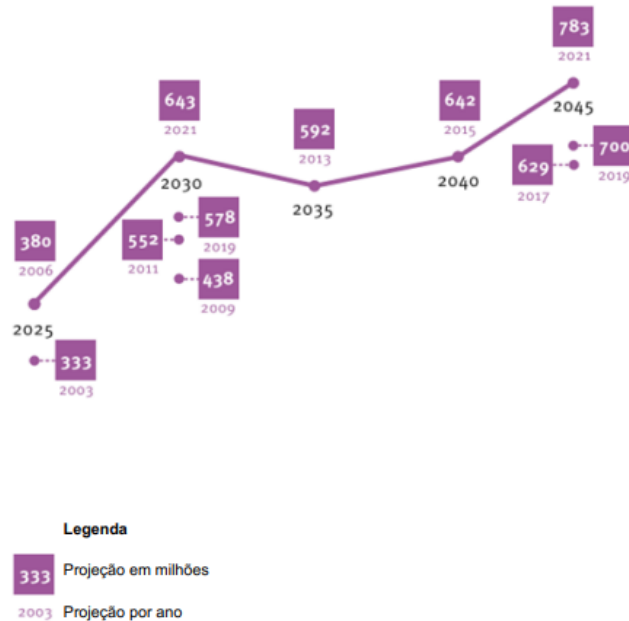
Existem 3 tipos principais de diabetes: diabetes tipo 1 (DM1), diabetes tipo 2 (DM2) e diabetes gestacional (DMG). O primeiro tipo, também conhecido como

“diabetes juvenil/infantil” ou “diabetes insulínodépendente” é definido pela deficiência na produção de insulina no corpo humano e seu tratamento requer a administração regular de análogos de insulina (JOHNSON *et al.*, 2021) (TAO; SHI; ZHAO, 2015). A causa exata do DM1 ainda não foi descoberta. No entanto, é geralmente aceito pela sociedade científica que um complexo número de fatores ambientais e genéticos interagem para o desenvolvimento do DM1 em seus estágios iniciais.

O DM2, anteriormente chamado de “diabetes de início adulto” ou “diabetes não insulínodépendente”, é caracterizado pela resistência à insulina (JOHNSON *et al.*, 2021). Em outras palavras, é o resultado de uma resposta ineficaz do organismo à insulina produzida. À medida que a doença progride, a deficiência de insulina também pode ocorrer. Etnia, histórico familiar de diabetes combinado com obesidade, dieta não saudável e exercícios físicos limitados são as principais causas de DM2 (TAO; SHI; ZHAO, 2015).

A zona cinzenta da transição de normoglicêmica para diabetes é frequentemente caracterizada como “Tolerância à glicose prejudicada” (IGT) ou “glicemia de jejum prejudicada” (IFG). As duas últimas condições são geralmente reconhecidas como precursoras do DM2, uma vez que se estima que cerca de 1 em cada 4 indivíduos com IGT/IFG progredirá para diabetes dentro de um período de 3-5 anos (NATHAN *et al.*, 2007). Finalmente, o DMG é a condição de elevação da glicemia durante a gravidez em mulheres sem histórico prévio de diabetes (JOHNSON *et al.*, 2021). Neste caso, a criança está em alto risco de desenvolver diabetes mais tarde na vida adulta.

Figura 1: Evolução de indivíduos com diabetes mellitus



Fonte: Modificada de (DIABETES, 2021).

A Figura 1 expõe a projeção de indivíduos com diabetes no mundo, com projeção de mais de meio bilhão de indivíduos com diabetes em 2035. O diabetes impõe um fardo econômico substancial aos países, sistemas de saúde, pessoas com diabetes e suas famílias (YANG, W. *et al.*, 2018).

O custo total estimado do diabetes diagnosticado em 2017 é de US\$ 327 bilhões, incluindo US\$ 237 bilhões em custos médicos diretos e US\$ 90 bilhões em produtividade reduzida. Para as categorias de custos analisadas, o atendimento a pessoas com diabetes diagnosticada representa 1 em cada 4 dólares de assistência médica nos EUA, e mais da metade desse gasto é diretamente atribuível ao diabetes. Pessoas com diabetes diagnosticada incorrem em despesas médicas médias de \$ 16.750 por ano, dos quais \$ 9.600 são atribuídos ao diabetes. Pessoas com diabetes diagnosticada, em média, têm despesas médicas 2,3 vezes maiores do que as despesas que seriam na ausência de diabetes (YANG, W. *et al.*, 2018).

3.1.2 Uma Breve História da Automonitorização

Indivíduos com DM1 precisam de injeções diárias de insulina para manter o nível de glicose no sangue dentro de um intervalo adequado. Sem insulina, elas não sobreviveriam. No entanto, com tratamento diário de insulina, monitoramento regular da glicose no sangue, educação e apoio, elas podem viver uma vida saudável e retardar ou prevenir muitas das complicações associadas ao diabetes (DIABETES, 2021).

Os primeiros métodos de medição da glicose foram o teste de Trommer (1841) e o teste de Fehling (1848), que utilizavam a presença de glicose na urina e suas propriedades redutoras com reagentes de sulfato de cobre alcalino. De acordo com Claude Bernard, em 1870 havia três métodos comumente usados para medir a glicose: polarimetria, redução de Cu (II) a Cu (I) por açúcares redutores e formação de dióxido de carbono como resultado da fermentação de uma solução que contém glicose (DZIERGOWSKA *et al.*, 2019).

Em 1850, Maumené descreveu as primeiras tiras de teste feitas de lã merino contendo cloreto de estanho. Em contato com urina contendo glicose e calor, a tira muda de cor para preto. O reagente desenvolvido com base nessas propriedades redutoras foi criado em 1908 por Benedict. Em 1919, esse método foi aprimorado por Folin e Wu e em 1940 por Nelson. A desvantagem desses métodos era a falta de especificidade, porque a reação poderia ocorrer não apenas devido ao contato com glicose, mas também com frutose, galactose, ácido úrico, ácido ascórbico, corpos cetônicos ou salicilatos (MOODLEY *et al.*, 2015).

O primeiro biossensor enzimático de glicose foi criado em 1962 por Clark e Lyons. Foi feito para monitorar a glicose durante a cirurgia cardíaca. A partir de então, muitos outros biossensores foram criados. Utilizam métodos eletroquímicos, espectrofotométricos, fluorescentes dentre outros. Em 2013, os biossensores de glicose representavam 85% de todos os biossensores disponíveis no mercado, e o diabetes era uma doença modelo para a geração de novos conceitos de biossensores (DZIERGOWSKA *et al.*, 2019).

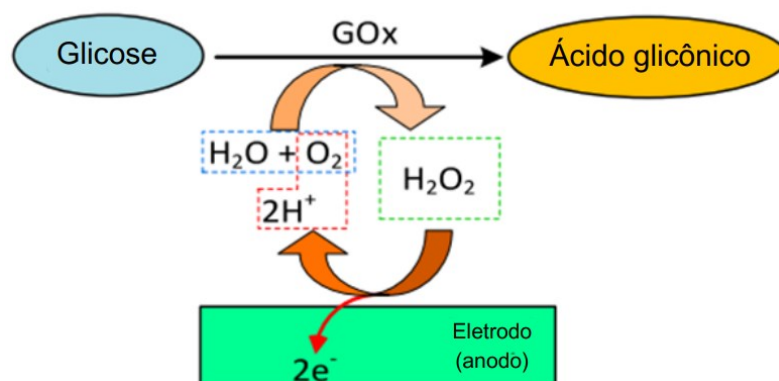
Existem dois tipos de dispositivos destinados ao uso pessoal e autoavaliação: monitoramento não contínuo (NCGM) e monitoramento contínuo de glicose (CGM). Como o nome sugere, dispositivos NCGM (comumente conhecidos como dispositivos de automonitoramento de glicose no sangue SMBG) são usados para monitorar os

níveis de glicose apenas em pontos específicos durante o dia, com uma frequência dependente do tipo de diabetes, dieta, dosagem de medicamentos e condição clínica da pessoa. Por outro lado, os dispositivos CGM podem monitorar os níveis de glicose automaticamente em pequenos intervalos de tempo, tornando possível monitorar mudanças rápidas e tendências perdidas nos testes SMBG. No entanto, a precisão e confiabilidade de ambos os sistemas são adequadas em situações de atendimento e autoavaliação (GONZALES; MOBASHSHER; ABBOSH, 2019).

Os dispositivos SMBG são os típicos glicosímetros que requerem a punção do dedo com uma lanceta para acessar o sangue capilar. O método de medição de glicose é fundamentalmente o mesmo que a técnica eletroquímica (Método Enzimático-Amperométrico). A reação completa e a detecção ocorrem em uma tira de teste de glicose conectada a um medidor. Após colocar uma gota da amostra de sangue na tira de teste, a glicose oxida na presença de uma enzima para produzir uma determinada quantidade de corrente proporcional ao nível de glicose. Os elétrons então viajam para o medidor contendo um conversor de corrente para tensão para fornecer uma tensão proporcional ao nível de glicose (GONZALES; MOBASHSHER; ABBOSH, 2019).

Técnica eletroquímica, considerando que a enzima glicose oxidase (GO_x) é específica para a glicose, neste método, a oxidação da glicose ocorre na presença de GO_x , oxigênio (O_2) e água (H_2O) para formar ácido glicônico e peróxido de hidrogênio (H_2O_2). O peróxido de hidrogênio é então oxidado eletroquimicamente no anodo de uma sonda eletroquímica, produzindo um sinal amperométrico (corrente) proporcional à concentração de glicose na amostra (GONZALES; MOBASHSHER; ABBOSH, 2019).

Figura 2: Método enzimático-amperométrico para medição da concentração de glicose in vitro.



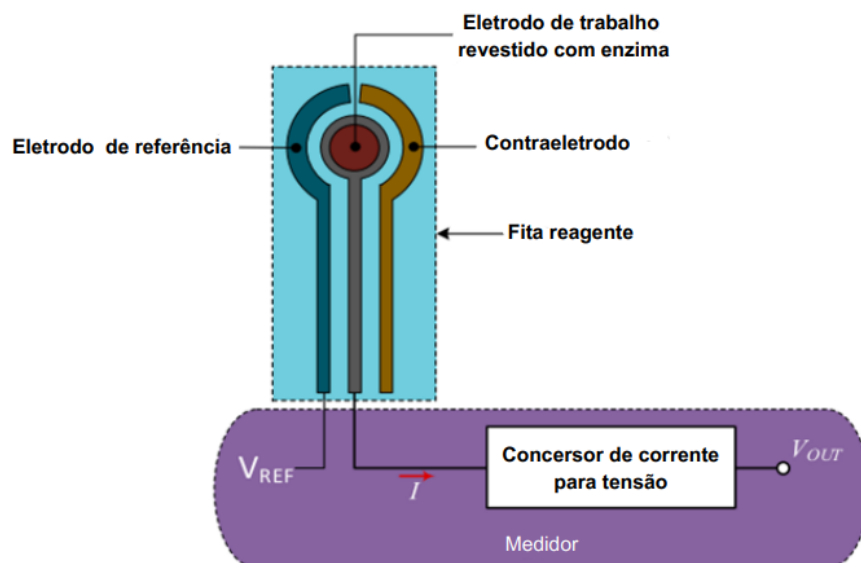
Fonte: Modificada de (GONZALES; MOBASHSHER; ABBOSH, 2019).

De acordo com a American Diabetes Association (2022), as tiras de teste (também conhecidas como fitas de reagente) são um componente essencial dos SMBG. Possuem um reagente que reage com a glicose presente na amostra de sangue, produzindo uma mudança de cor que pode ser lida por um medidor de glicose. A qualidade e a precisão dessas tiras podem variar consideravelmente entre os diferentes fabricantes e modelos, e é importante que os pacientes e os profissionais de saúde sejam orientados sobre como armazenar e utilizar corretamente essas tiras (Bode et al., 2015; Wu et al., 2018).

Além disso, as tiras de teste podem representar um custo significativo para os pacientes, especialmente para aqueles que fazem um monitoramento frequente da glicose no sangue (IDF, 2019). Portanto, é importante que os pacientes tenham acesso a informações precisas e atualizadas sobre as melhores práticas para o uso das tiras de teste, a fim de otimizar a qualidade do monitoramento e reduzir os custos associados ao tratamento do diabetes (Heinemann et al., 2018; Rodbard, 2017).

As características das tiras de teste, como a sensibilidade e a especificidade, são cruciais para garantir a precisão do monitoramento da glicemia e, conseqüentemente, para um tratamento eficaz do diabetes (Bode et al., 2015). Além disso, é importante considerar a estabilidade das tiras de teste durante o armazenamento e o transporte, a fim de garantir que elas mantenham sua integridade e funcionem corretamente (Wu et al., 2018).

Figura 3: Diagrama de blocos do dispositivo para medição de glicose com método de picada no dedo.



Fonte: Modificada de (GONZALES; MOBASHSHER; ABBOSH, 2019).

3.2 MEDIÇÃO NÃO INVASIVA

Em seguida, será abordado o tema da medição de glicemia sem a necessidade de invasão, juntamente com as dificuldades enfrentadas na implementação de um sistema não invasivo e as principais técnicas utilizadas atualmente.

De acordo com o fato de o teste de glicose no sangue causar lesões na pele humana, ele pode ser simplesmente dividido em monitoramento de glicose no sangue invasivo e não invasivo, em relação ao tempo ou frequência de operação, os sistemas podem ser divididos em duas categorias: discretos ou contínuos. No entanto, na literatura, a classificação de tempo contínuo não é usada de forma rigorosa. A denominação de tempo contínuo se refere a um dispositivo capaz de medir a glicemia em intervalos onde a variação é pequena o suficiente. Dispositivos que medem a glicemia em intervalos iguais ou inferiores a 10 minutos são classificados como contínuos (TANG, L. *et al.*, 2020)(SILVA; LUIZ, 2017).

Atualmente, a tecnologia invasiva de detecção de glicose no sangue é dominante. Nos hospitais, o sangue retirado dos indivíduos em jejum pela manhã é medido com precisão pelo analisador bioquímico automático. Embora os resultados desse método sejam precisos e possam ser usados como uma importante base para o diagnóstico de diabetes, não é adequado para o monitoramento contínuo de diabéticos devido ao seu processo tedioso, tempo longo de detecção e grande quantidade de extração de sangue venoso. O SMBG é um método amplamente utilizado para monitorar a concentração de glicose no sangue, mas tem desvantagens como feridas dolorosas nos dedos, alto custo e adesão inadequada às diretrizes. Métodos aprimorados são necessários para prevenir complicações graves do diabetes (TANG, L. *et al.*, 2020)

Atualmente diversos pesquisadores estão focando seus estudos, em métodos não-invasivos de medição de glicemia, utilizando métodos ópticos, como espectroscopia de infravermelho próximo (NIRS), espectroscopia de infravermelho médio (MIR), espectroscopia RAMAN, dentre outras, e métodos transdérmicos, como espectroscopia de impedância, iontoforese reversa, dentre outros, como podemos ver em, (NAWAZ *et al.*, 2016)(GONZALES; MOBASHSHER; ABBOSH, 2019)(WANG, Z. *et al.*, 2022)(PRIYOTI *et al.*, 2019)(YEN *et al.*, 2022)(SONG, K. *et al.*, 2015a).

Métodos ópticos, como a espectroscopia, são comumente utilizados, mas métodos contínuos que testam tecidos e fluidos secretados pelo corpo humano estão sendo estudados também. Embora os métodos contínuos possam fornecer um diagnóstico rápido, ainda há um atraso na mudança dos níveis de glicose em comparação com os níveis sanguíneos. Métodos não invasivos podem ser indiretos, envolvendo a análise de propriedades indiretas da glicose, ou diretos, baseados nas propriedades intrínsecas das moléculas de glicose. Um dos principais problemas com muitos métodos não invasivos é a necessidade de calibração contínua, tornando as medidas tradicionais de sangue ainda necessárias. No entanto, sensores miniaturizados usados permanentemente podem ser um avanço nos testes não invasivos (DZIERGOWSKA *et al.*, 2019).

3.2.1 Espectroscopia de Infravermelho

Baseia-se na mensuração da intensidade de radiação infravermelha absorvida pelos tecidos. Quando radiação é focada sobre uma amostra, ocorrem fenômenos como absorção, dispersão e emissão, os quais estão relacionados com a composição estrutural e química da amostra. A molécula de glicose possui bandas de absorção específicas (*e.g.*, assinatura óptica), de modo que se pode direcionar radiação através de tecidos biológicos ou amostras *in vitro* e analisá-la para determinar a concentração de glicose. As radiações de infravermelho próximo (NIR) e de infravermelho médio (MIR) são as mais utilizadas nessa técnica (SILVA; LUIZ, 2017).

A tecnologia de espectroscopia no infravermelho próximo (NIRS) depende da absorção e dispersão de comprimentos de onda na faixa de 780 nm a 2500 nm devido às vibrações moleculares e rotação de ligações dentro da molécula. Utiliza três modos básicos de medição: transmitância, reflectância (incluindo reflectância difusa) e interação. No entanto, todos eles dependem da mesma tecnologia central, um espectrômetro dispersivo (GONZALES; MOBASHSHER; ABBOSH, 2019).

No NIRS, a região espectral que se situa entre 750-2500 nm corresponde a transições vibracionais fundamentais e combinações de sobretons de grupos (CH-OH-NH). Essa região espectral (700-1100 nm), conhecida como janela terapêutica, apresenta intensidades de absorção de melanina, banda de absorção de água e hemoglobina suficientemente baixas para que a luz possa penetrar em tecidos

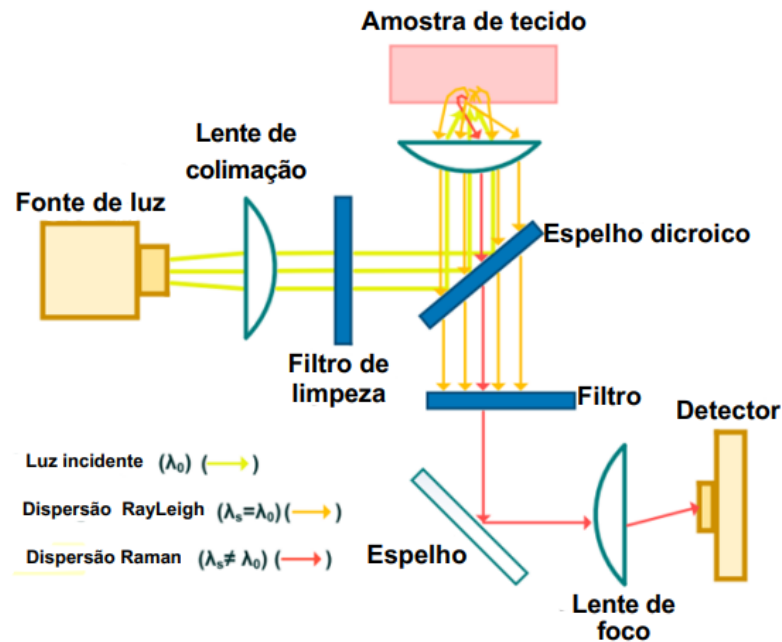
profundos com eficiência de até 90-95%. A glicose apresenta picos de absorção em 939, 970 , 1197 nm na região de sobretom mais alta, 1408, 1536 e 1688, 1925 nm na região de sobretoms mais baixa e 2100, 2261, 2326 nm na região de combinação (NAWAZ *et al.*, 2016).

3.2.2 Espectroscopia Raman

A espectroscopia Raman permite a avaliação da dispersão da luz de uma única frequência. A vantagem desse método, em comparação com MIR e NIR, é a ausência de ruído causado pela presença de água, enquanto a desvantagem é um sinal fraco e susceptibilidade a fatores como turbidez, hematócrito, espessura da pele e presença de melanina. Como no caso de outras técnicas ópticas, a análise multidimensional é necessária. Este método pode ser usado tanto para estudos *in vitro* quanto *in vivo*. A espectroscopia Raman pode ser usada para medições usando outros meios além do sangue, incluindo o uso de fluido lacrimal. Uma das empresas que utiliza a espectroscopia Raman infravermelha para estudar medidas não invasivas de glicose é a Light Touch Medical. Em 2018, Lundsgaard-Nielsen, descreveram o dispositivo para medir a concentração de glicose por espectroscopia Raman que poderia ser usado para medições independentes em casa (DZIERGOWSKA *et al.*, 2019).

Na dispersão Raman, a mudança no comprimento de onda e na frequência entre a luz incidente e emitida ocorre devido à interação da luz com as moléculas espalhadoras. Conforme Figura 4, os elétrons dentro das moléculas absorvem energia suficiente da luz incidente para fazer transições para estados de energia mais altos, e quando os elétrons fazem a transição de volta para diferentes níveis de energia, as transições resultam na emissão de fótons com frequências diferentes das dos fótons incidentes, caracterizadas como deslocamento Raman (SHOKREKHODAEI; QUINONES, 2020).

Figura 4: Esquema simplificado ilustrando a espectroscopia Raman.



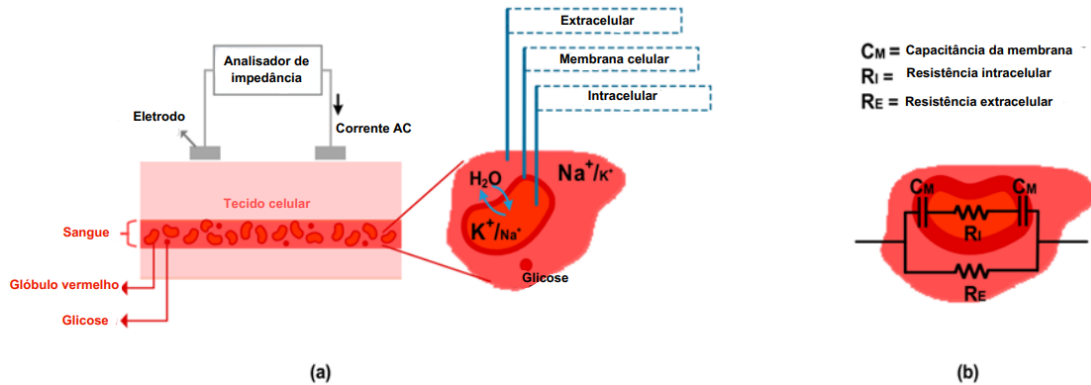
3.2.3 Espectroscopia de Impedância

A espectroscopia de impedância mede o sinal de infravermelho gerado pelo corpo humano como resultado das mudanças na concentração de glicose. É usada para medir as propriedades dielétricas do tecido. Uma pequena corrente alternada passa pelo tecido, e a impedância é lida como uma função da frequência. A glicose é medida indiretamente por meio de interações com eritrócitos que dependem da sua concentração. Este método é vulnerável a flutuações de temperatura, movimento e presença de suor na pele (DZIERGOWSKA *et al.*, 2019).

A Figura 5a ilustra os componentes fisiológicos que desempenham um papel no comportamento da impedância tecidual. A impedância tecidual depende das características elétricas das membranas celulares. Como os glóbulos vermelhos suspensos no sangue ocupam cerca de 45% do volume sanguíneo, a membrana dos glóbulos vermelhos tem um papel importante na impedância total do tecido e afeta o valor da capacitância no circuito equivalente do tecido. Em uma visão microscópica, o comportamento elétrico de um único glóbulo vermelho é modelado como ilustrado na Figura 5b. A membrana do glóbulo vermelho contribui para a capacitância, e o fluido intracelular e extracelular contribuem para a resistência do modelo elétrico. Em uma

escala macroscópica, o modelo elétrico para o tecido humano é mais complexo, como descrito em (SHOKREKHODAEI; QUINONES, 2020).

Figura 5: (a) Um esquema simplificado ilustrando a espectroscopia de impedância tecidual, (b) Um modelo elétrico para uma única célula vermelha do sangue.



Fonte: Modificada de (SHOKREKHODAEI; QUINONES, 2020).

A desidratação celular promove a redistribuição de íons de potássio $[K^+]$ do espaço intracelular para o espaço extracelular, aumentando o nível de íons $[K^+]$ no soro no espaço extracelular. Conforme os íons $[K^+]$ e $[Na^+]$ estão em equilíbrio, ocorrem mudanças associadas na permitividade e na condutividade do meio circundante e das membranas celulares, incluindo a membrana dos **RBCs**. Essas atividades resultam em mudanças na impedância do tecido. A permitividade e a condutividade de uma solução aquosa diminuem quando a concentração de glicose aumenta para uma faixa de frequência entre 1 kHz e 1 MHz. A magnitude da impedância aumenta com a diminuição da permitividade e da condutividade. A espectroscopia de impedância enfrenta desafios de medição e precisão, e estes são descritos abaixo, juntamente com possíveis soluções (SHOKREKHODAEI; QUINONES, 2020).

3.2.4 Polarimetria Óptica

Polarimetria óptica (OP) é um método promissor não invasivo para detecção de glicose no sangue, mas sua precisão pode ser limitada por fatores como propriedades de espalhamento da pele, artefatos de movimento e presença de outras biomoléculas com características de rotação óptica no sangue.

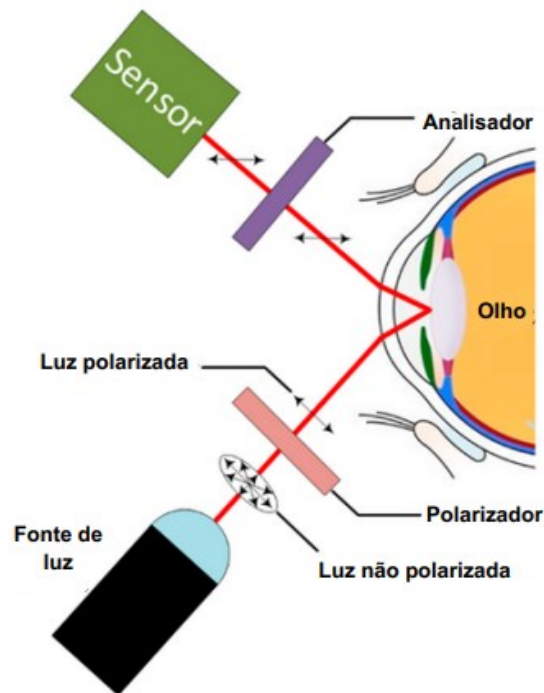
A OP é um dos primeiros métodos de detecção não invasivos de glicose no sangue. Uma vez que a glicose é uma substância opticamente ativa e tem uma rotação óptica estável, quando um feixe de luz polarizada ilumina uma solução contendo solutos de glicose, a presença de glicose causa uma certa rotação do plano polarizado da luz incidente.

Neste ponto, a direção de polarização forma um ângulo de deflexão com a direção incidente original, que é proporcional à quantidade de glicose. Depois que a luz polarizada passa pela amostra, o plano da luz polarizada também pode ser medida com um polarímetro. Quando o eixo de polarização do analisador é combinado com o ângulo de rotação do campo elétrico, o fotodetector detectará a intensidade máxima de luz. Quando o eixo de polarização é perpendicular ao ângulo de rotação do campo elétrico, a luz não será detectada pelo fotodetector.

Embora o método seja diretamente detectado pela luz visível, não seja complicado de operar e seja fácil obter resultados, a precisão de erro do método pode não ser satisfatória, principalmente devido às seguintes três razões: (a) Afetado pelas características de dispersão dos objetos a serem medidos, a luz polarizada é propensa a perder sua polarização, enquanto a pele humana tem uma forte propriedade de dispersão, e a atividade óptica da glicose é quase completamente inibida pela birrefringência linear da derme, por isso os olhos (especialmente o humor aquoso na frente) são geralmente selecionados para a medição. (b) Quando a glicose no sangue é detectada através do humor aquoso do olho, artefatos de movimento podem causar alterações na birrefringência córnea.

Essa birrefringência variável no tempo é uma das fontes de interferência mais importantes, que pode causar mudanças significativas no estado de polarização da luz detectada e, portanto, confundir as características ópticas da glicose. (c) Quando o feixe de luz atravessa a pele e atinge o sangue, a presença de biomoléculas com características de rotação óptica no sangue (como a vitamina C e a albumina) pode causar um desvio que não pode ser ignorado no ângulo de deflexão (TANG, L. *et al.*, 2020).

Figura 6: Princípio da polarimetria óptica no olho para monitoramento de glicose.



Fonte: Modificada de (GONZALES; MOBASHSHER; ABBOSH, 2019).

3.3 ESPECTROSCOPIA DE INFRAVERMELHO

A seguir, serão apresentados detalhes sobre a técnica de espectroscopia de infravermelho para medição de glicemia, abrangendo conceitos teóricos e práticos. Os aspectos teóricos da técnica serão abordados para esclarecer seus princípios fundamentais, enquanto serão descritos também alguns aspectos práticos importantes para a implementação de um sistema não invasivo de medição de glicose sanguínea como também em amostras, com base nessa técnica.

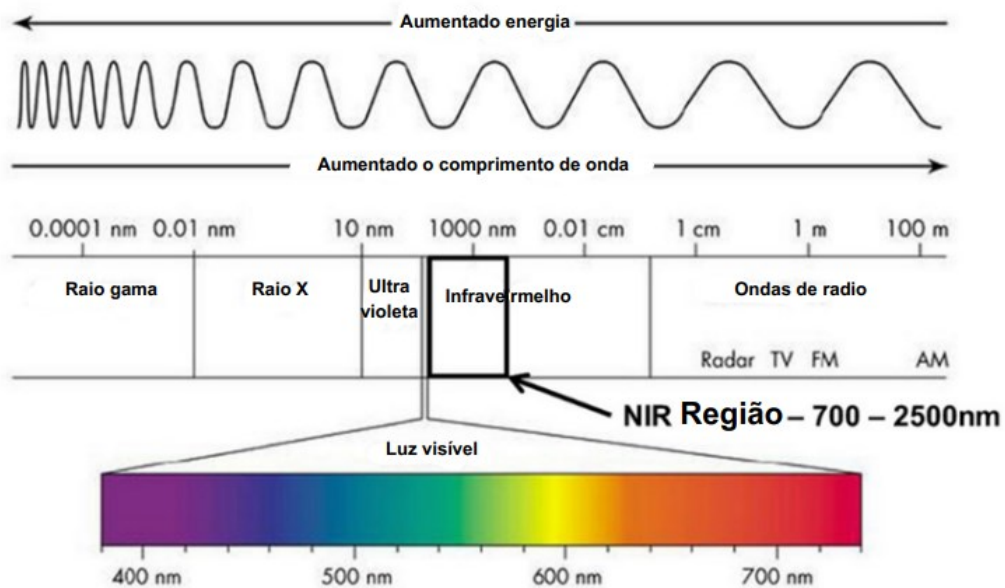
3.3.1 Aspectos Teóricos

Discute-se, na sequência, a respeito dos aspectos teóricos de espectroscopia de infravermelho. Apresentam-se conceitos essenciais para a compreensão da absorção de radiação eletromagnética pela matéria.

3.3.2 Radiação Eletromagnética

A espectroscopia de absorção no infravermelho médio (MIR) e no infravermelho próximo (NIRS) são técnicas de medição usadas para adquirir informações quantitativas sobre uma amostra e investigar seus componentes. Como ilustra a Figura 7, a faixa de comprimento de onda para NIR está entre 700 e 2500 nm e para MIR, a faixa está entre 2500 e 25.000 nm. A detecção e medição de NIR são possíveis tanto nos modos de reflexão quanto nos modos de transmissão devido às profundidades de penetração associadas de 0,5 mm ou mais (SHOKREKHODAEI; QUINONES, 2020).

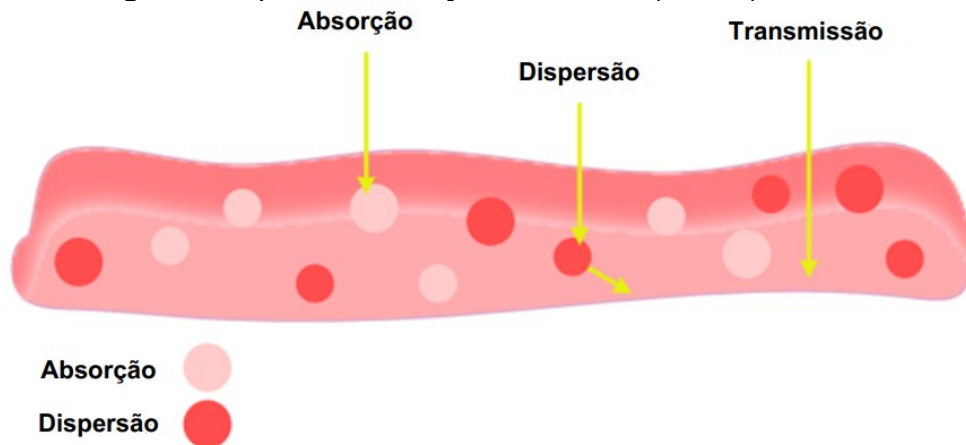
Figura 7: O espectro eletromagnético vai desde os raios gama (0,0001 nm) até as ondas de rádio (100 m).



Fonte: Modificada de (WHITAKER, 2002).

As localizações estratégicas do sensor de glicose no corpo humano incluem os dedos, as orelhas, os lábios, o antebraço, a câmara anterior do olho e a língua. À medida que a luz entra no corpo, ela interage com átomos dentro do tecido e é absorvida, transmitida ou dispersada, conforme ilustrado na Figura 8.

Figura 8: Tipos de interações entre luz (fótons) e tecido.



Fonte: Modificada de (SHOKREKHODAEI; QUINONES, 2020).

Quando a luz é absorvida por um material, a energia associada à luz ou ao fóton é consumida como resultado da interação entre a luz e o material. Dependendo da energia do fóton, pode haver diferentes tipos de transições de energia em moléculas, incluindo transição de energia entre estados vibracionais, estados rotacionais, estados eletrônicos etc. (SHOKREKHODAEI; QUINONES, 2020).

Dependendo do comprimento de onda da luz incidente (λ_1), a dispersão pode ser dominante. A dispersão elástica e inelástica são dois tipos de dispersão da luz. Na dispersão elástica, a energia da luz dispersa é igual à energia da luz incidente (ou seja, *i. e.*, $\lambda_1 = \lambda_{luz_dispersa}$), enquanto na dispersão inelástica, a energia da luz dispersa é menor ou maior do que a luz incidente (ou seja, *i. e.*, $\lambda_1 \neq \lambda_{luz_dispersa}$), (BOUSTANY; BOPPART; BACKMAN, 2010). A luz dispersa pode estar em qualquer direção/ângulo (desde a retro espalhada até a frente), e é afetada pela estrutura do material e pela energia da luz incidente. A luz que interage com os componentes do tecido tende a se dispersar na direção frontal para uma única dispersão. No entanto, a dispersão múltipla pode resultar em dispersão para trás no tecido (SHOKREKHODAEI; QUINONES, 2020).

3.3.3 Lei de Beer-Lambert

A Lei de Beer-Lambert Equação (1) fornece uma formulação matemática do método que permite o cálculo da absorvância de uma amostra a partir da concentração e da espessura (TANG, L. *et al.*, 2020).

$$I = I_0 10^{(-l.\varepsilon.c)} = I_0 e^{(-l.\mu a)} \quad (1)$$

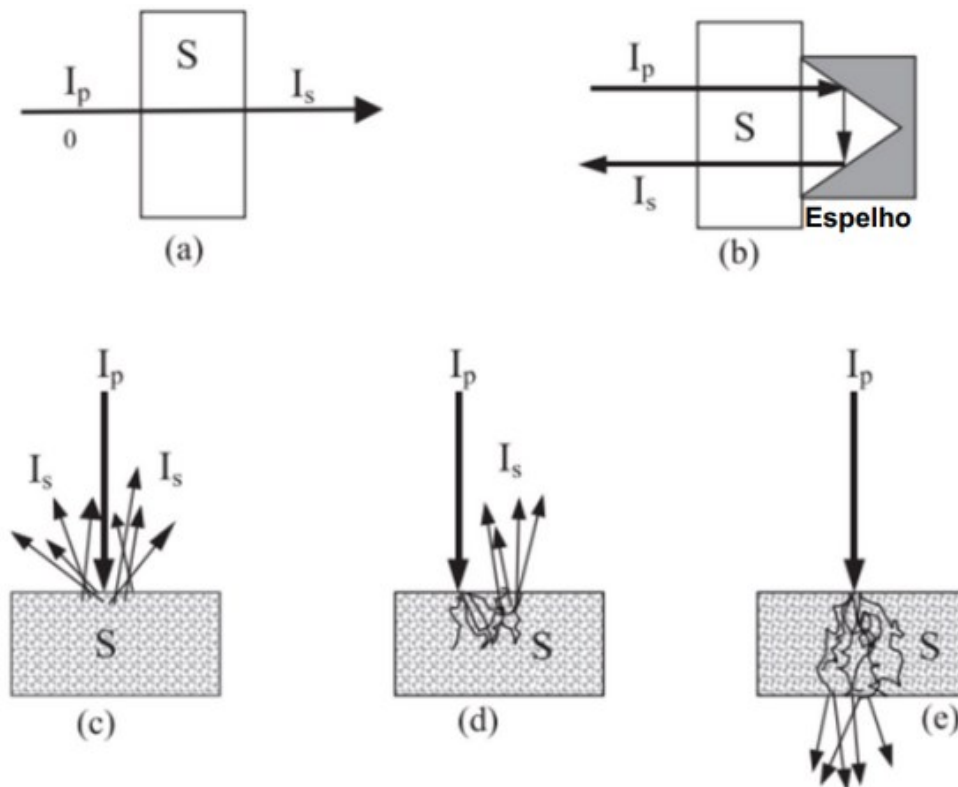
Essa equação foi baseada em diversos fatores: a complexidade dos componentes do sangue, as interações entre eles, sua distribuição irregular causada pela presença de macromoléculas (e.g, proteínas ou gorduras) e o efeito do ruído colorido e do desvio da linha de base. A estratégia não linear utilizada na análise desses estudos é chamada mUVE-KPLS (eliminação modificada de variáveis não informativas - mínimos quadrados parciais com kernel) (DZIERGOWSKA *et al.*, 2019).

O modelo mostra a intensidade da luz refletida/transmitida como uma função da espessura da amostra, concentração e coeficiente de absorção, onde o efeito da luz dispersa é negligenciado. A absorvância é definida como $\log(I_0/I)$ (TANG, L. *et al.*, 2020).

3.3.4 Modos de Medição

A transmitância difusa é o método de medição mais utilizado, mas outros métodos podem ser usados, como mostrado na Figura 9.

Figura 9: Modos de medição NIRS. (a) Transmissão transparente, (b) transmissão transparente dupla, (c) reflectância, (d) reflectância difusa e (e) transmissão difusa.



Fonte: Modificada de (SANTOS, 2018).

A Figura 9 (a) é usada para materiais transparentes ou para amostras em cubetas de vidro. A Figura 9 (b) é semelhante à Figura 9 (a), mas neste caso o feixe de luz passa duas vezes pelo material. Por causa disso, a atenuação é produzida duas vezes. A reflectância difusa é usada para sólidos, mostrada na Figura 9 (c). Na reflectância difusa, a mudança da intensidade está relacionada com a dispersão e absorção produzida pelo sólido. Nesse caso, a lei de Beer não é aplicável, porque é válida apenas para meios transparentes.

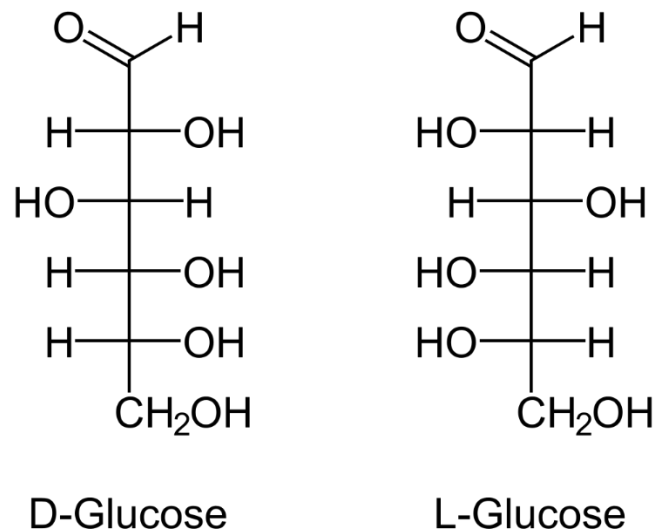
A Figura 9 (d) mostra uma reflectância difusa onde o ponto de medição é feito em um ponto separado da injeção do feixe de luz. Devido à interação com os tecidos ser maior, esse método fornece mais informações. Em meios homogêneos, pode ser

positivo, mas para meios heterogêneos pode ser impossível diferenciar a fonte de absorção ou dispersão. A Figura 9 (e) mostra o caminho mais longo para o feixe de luz com a maior atenuação. Este método tem a característica de fornecer informações da amostra completa e é normalmente usado quando o meio é isotrópico ou pode ser tratado como isotrópico. Um exemplo conhecido de transmissão difusa é a oximetria de pulso, onde o dedo, na maioria dos casos, é tratado como um único material para detectar a relação entre a presença dos cromóforos Hb e HbO₂, medindo a absorção em seus respectivos picos de absorção (SANTOS, 2018; YAMASHITA, 2013).

3.3.5 Glicose d-Anidra (Dextrose)

Duas formas de glicose existem: L-glicose e D-glicose (ou dextrose), Figura 10. Embora compartilhem a fórmula C₆H₁₂O₆, suas estruturas moleculares variam. Em organismos complexos como o sangue humano, a quantidade de L-glicose é insignificante (na ordem de nm/dL), contrastando com a presença mais substancial de D-glicose (na ordem de mg/dL) (SHOKREKHODAEI *et al.*, 2021). Dispositivos de glicose disponíveis comercialmente focalizam a medição dos níveis de D-glicose no sangue (STEFAN-VAN STADEN; MITROFAN, 2018).

Figura 10: Projeção Fischer L-Glicose E D-Glicose

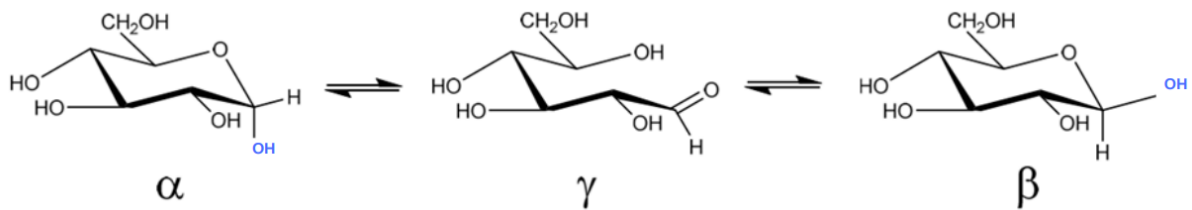


Fonte: Construção do próprio autor (2023).

Segundo (VAN ENTER; VON HAUFF, 2018), a cadeia γ -D-glicose atua como intermediária na conversão entre as formas cíclicas, embora não seja

termodinamicamente estável. Ela pode fechar em dois ângulos distintos, formando os anômeros α -D-glicose ou β -D-glicose com centros quirais R ou S. A mudança entre essas formas ocorre por meio da forma linear γ -D-glicose. Os anômeros α e β têm diferentes orientações de grupos OH e propriedades químicas, sendo a α -D-glicose mais estável devido ao efeito anômero. A preferência pela posição equatorial na β -D-glicose é influenciada pela polaridade do solvente e grupos substituintes, sendo mais prevalente em condições fisiológicas. A conformação em cadeira estabiliza o grupo OH no centro anomérico do β -anômero, minimizando impedimentos estéricos e repulsões elétricas, conforme Figura 11.

Figura 11: Esquema da mutarrotação da D-glicose. A interconversão entre as formas em anel α -D-glicose e β -D-glicose ocorre por meio da forma linear γ -D-glicose.



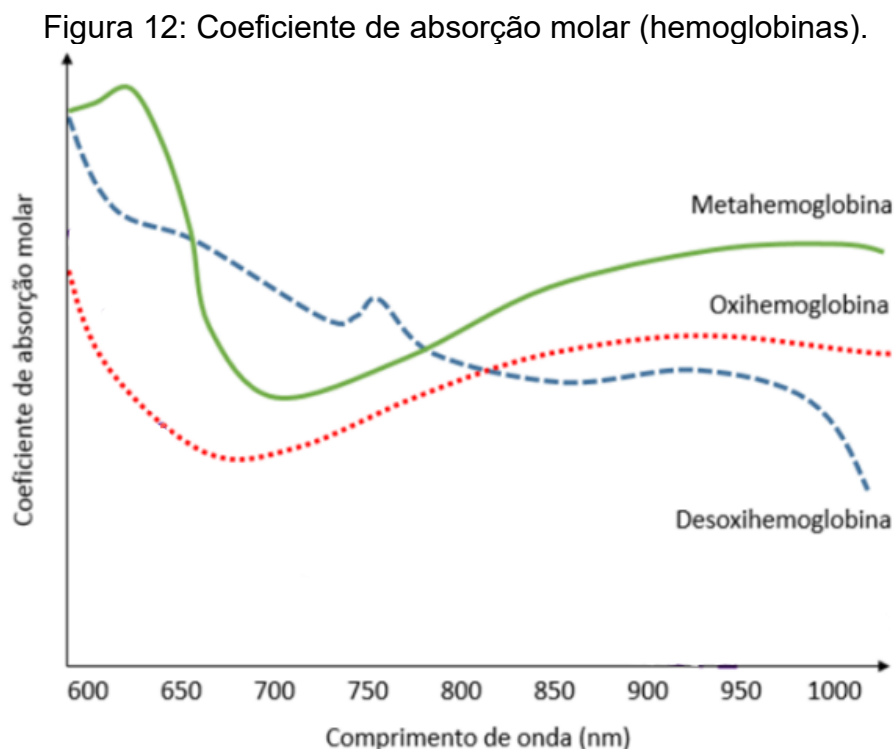
Fonte: Modificada de (VAN ENTER; VON HAUFF, 2018)

Conforme (VAN ENTER; VON HAUFF, 2018), a maioria dos sensores de glicose usa enzimas como a glicose oxidase (GO_x) e a glicose desidrogenase (GDH) para medir a glicose indiretamente. A GO_x mede apenas o anômero β -D-glicose, enquanto a GDH mede ambos os anômeros. A proporção entre eles deve ser conhecida para calcular a concentração total de glicose. Sensores comerciais detectam mudanças eletrônicas ou ópticas causadas pelas reações enzimáticas. O centro catalítico da GO_x é o FAD, que é oxidado e reduzido para medir a glicose. As diferentes gerações de biossensores usam métodos variados, como peróxido de hidrogênio e mediadores, para converter $FADH_2$ de volta em FAD. Em ambientes corporais estáveis, as proporções de anoméricos são consistentes, facilitando o cálculo da concentração de glicose. No entanto, em fluidos com variações de pH e temperatura, esse cálculo é mais complexo.

3.3.6 Absorção Luz pela Glicose e Água

As moléculas mais abundantes nos fluidos biológicos no corpo humano são a água, enquanto a glicose compreende apenas 0,07-0,1% do plasma sanguíneo, moléculas de água absorvem uma porcentagem significativa da luz incidente, especialmente dentro do intervalo de comprimentos de onda do MIR. Quando isso ocorre, a absorção de luz pela água é independente e adicional à absorção de luz pelas moléculas de glicose e, portanto, diminui a sensibilidade às moléculas de glicose. Diante disto, é necessário identificar uma janela de comprimento de onda que minimize a absorção de luz pela água e maximize a absorção de luz pelas moléculas de glicose (SHOKREKHODAEI; QUINONES, 2020).

A Figura 12, exibe o coeficiente de absorção molar de três tipos de hemoglobina em relação ao comprimento de onda.



Fonte: Modificada de (SILVA; LUIZ, 2017).

Segundo, (KOZMA *et al.*, 2014)(MARUO; YAMADA, 2015)(TURGUL; KALE, 2018)(VURAL; YOLERI; TORUN, 2019), existem dois picos de absorção para água na faixa NIR, localizado entre 1350 e 1520 nm e outro entre 1790 e 2000 nm, A absorção de luz pela glicose é maior na faixa de comprimento de onda entre 2000 e

2400 nm em comparação com as faixas mais curtas (700-1100 nm). Assim, o uso de comprimentos de onda mais curtos pode resultar em maior seletividade para as moléculas de glicose, minimizando o efeito interferente da água.

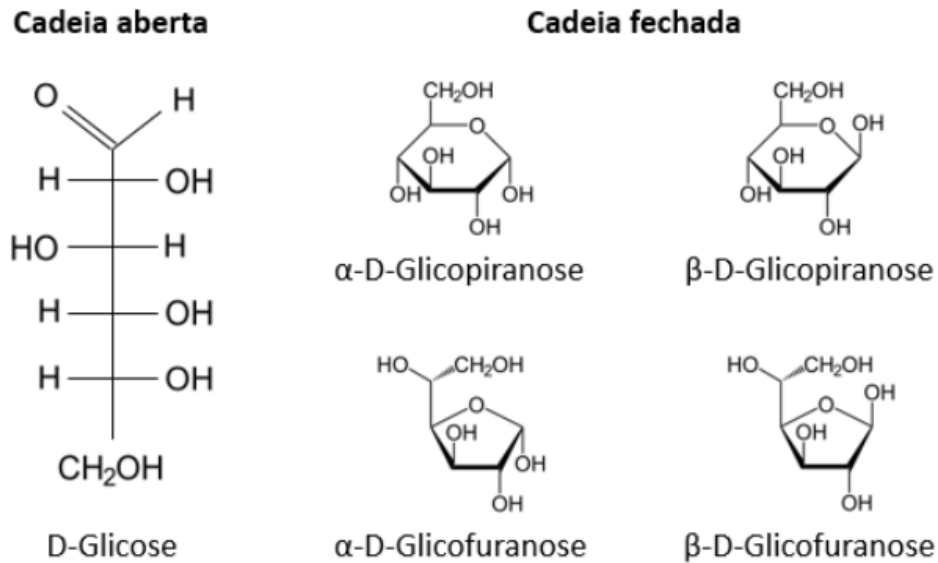
A abordagem de múltiplos comprimentos de onda foi aplicada em um dispositivo de medição de glicose não invasivo chamado TensorTip Combo Glucometer. Projetado pela Cnoga Medical Ltd (Caesarea, Israel) e o processo de aprovação do conceito começou em 2006. O dispositivo é capaz de medir glicose na faixa entre 70 e 440 mg/dL (SHOKREKHODAEI; QUINONES, 2020).

Conforme (PFÜTZNER *et al.*, 2018; SEGMAN, 2018), o TensorTip Combo Glucometer utiliza quatro LEDs com comprimentos de onda entre 600 e 1000 nm, juntamente com uma câmera de sensor de imagem em cores, para capturar a luz transmitida através da ponta do dedo e obter a leitura de glicose. Essa leitura é baseada em um sinal de seis dimensões que inclui posição (x, y), tempo (t) e cor (vermelho, verde e azul), que estão correlacionados com a concentração de glicose no sangue. Para avaliar o desempenho do dispositivo, um estudo foi realizado com 14 indivíduos saudáveis, seis pacientes com diabetes tipo 1 e 16 pacientes com diabetes tipo 2, e a análise dos resultados utilizando a grade de erro de consenso mostrou que todos os dados estavam na zona A (96,6%) e B (3,4%).

3.3.7 Bandas de Absorção da Glicose

A glicose é um monossacarídeo, um tipo de carboidrato simples, que é uma das principais fontes de energia para o corpo humano. Ela pertence ao grupo dos açúcares, mais especificamente ao grupo das hexoses, pois possui seis átomos de carbono em sua molécula. Sua fórmula química é $C_6H_{12}O_6$. Conforme podemos ver na Figura 13.

Figura 13: Estruturas Químicas da Glicose



Fonte: (SILVA; LUIZ, 2017)

As bandas de absorção da glicose na região do infravermelho próximo consideradas são a banda de sobretom de segunda ou maior ordem (750 a 1400 nm), a banda de primeiro sobretom (1400 a 2000 nm) (YADAV *et al.*, 2014).

3.3.8 Absorção de Luz por Componentes Sanguíneos e Tecidos

A dispersão da luz devido a múltiplos componentes dos tecidos resulta em desvio da lei de Beer-Lambert de absorção, o que resulta em erro de medição. Na verdade, tanto os componentes dos tecidos quanto os componentes do sangue causam dispersão da luz, o que atenua a intensidade da luz medida (SHOKREKHODAEI; QUINONES, 2020). A atenuação total da luz depende do coeficiente de atenuação total, μ_{total} (1/cm), que é a soma do coeficiente de absorção das espécies absorventes, μ_a , e do coeficiente de espalhamento reduzido das espécies que causam espalhamento, μ'_s (Equação (2)). O coeficiente de atenuação total representa o quão fortemente a luz é atenuada por espécies moleculares para um comprimento de onda específico. O coeficiente de espalhamento reduzido é uma contribuição do coeficiente de espalhamento, μ_s , e da anisotropia da propagação da luz no tecido biológico, g , que é a média do cosseno do ângulo de espalhamento (ROGERS *et al.*, 2014).

$$\mu_{total} = \mu_a + \mu'_s = \mu_a + \mu_s(1 - g) \quad (2)$$

Se g estiver próximo de 1, uma maior quantidade de luz é espalhada na direção para frente em comparação com a dispersão para trás. O valor de g para tecido biológico encontra-se entre 0,65 e 0,95 (ROGERS *et al.*, 2014).

Isso indica que quando a luz interage com os componentes dos tecidos, tende a se dispersar na direção frontal durante um único evento de dispersão. No entanto, após múltiplos eventos de dispersão, a dispersão geral da luz pode resultar em uma dispersão posterior. A dispersão da luz devido à interferência dos tecidos causa erros na medição da glicose, já que a dispersão da luz varia entre indivíduos devido à variedade de gordura nos tecidos. Além disso, as variações na dispersão da luz podem ocorrer devido a diferenças no colágeno, proteína, fluxo sanguíneo e estado de hidratação de uma pessoa (ZIRK; POETZSCHKE, 2004).

Segundo (CHUNG *et al.*, 2012; JONASSON *et al.*, 2018), é importante reduzir o espalhamento de luz, selecionando o comprimento de onda da fonte de luz que tenha menos espalhamento. A intensidade da luz dispersa devido aos componentes do tecido é negativamente correlacionada com o comprimento de onda da luz incidente, e é reduzida aumentando o comprimento de onda da luz incidente. Como resultado, a luz NIR resulta em mais espalhamento pelo tecido em comparação com a luz MIR (CHUNG *et al.*, 2012). No entanto, a luz NIR resulta em maior profundidade de penetração através do tecido, em comparação com a luz MIR.

De acordo com, (TANG, W. *et al.*, 2019; WAN *et al.*, 2020), a presença de componentes de espalhamento em uma amostra causa uma relação multivariada não linear entre os espectros de absorção e a concentração dos componentes absorventes. Vários métodos de calibração, como regressão local, redes neurais artificiais (ANN) e máquina de vetor de suporte (SVM), foram desenvolvidos para modelar essa não linearidade e melhorar a precisão da leitura de glicose no sangue.

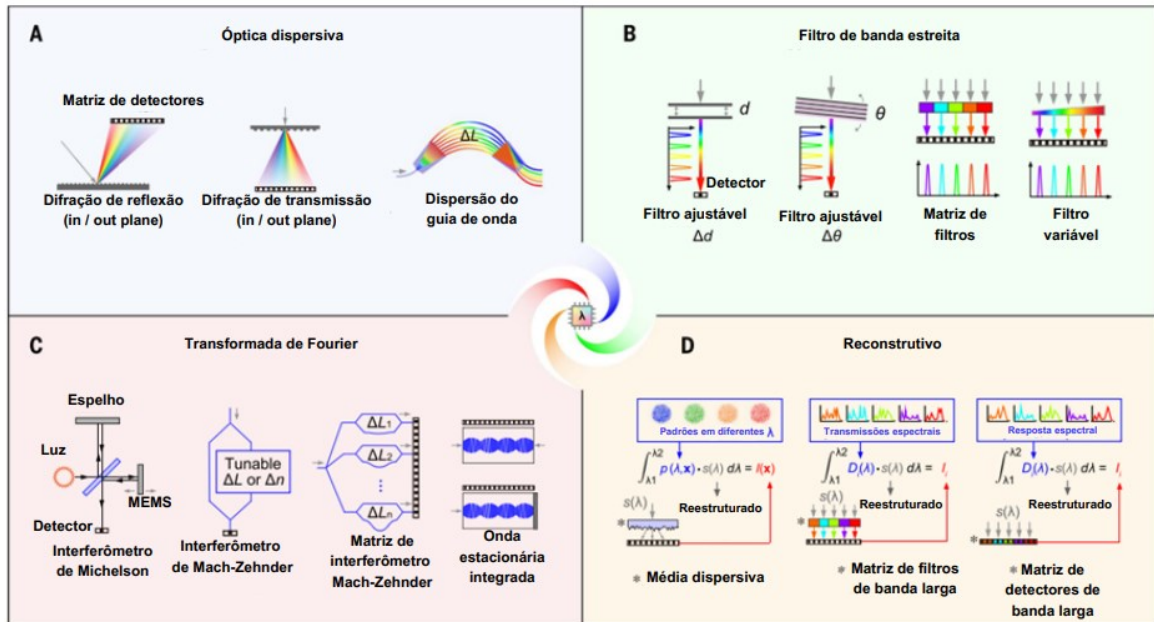
A não linearidade pode ser minimizada por 1) identificar o comprimento de caminho ótimo seguido pelo feixe de luz; 2) combinação de múltiplas aquisições e 3) aplicação de algoritmos de pré-processamento nas informações espectrais, como correção de espalhamento multiplicativo (MSC), correção de sinal ortogonal (OSC) e pré-processamento de analito líquido (TANG, W. *et al.*, 2019; WAN *et al.*, 2020).

3.4 ESPECTRÔMETRO

Segundo (SONG, H. *et al.*, 2020), A primeira detecção espectral foi feita por Isaac Newton em 1665., utilizando um prisma triangular para dividir a luz solar em um padrão colorido de arco-íris. No entanto, a medição quantitativa do espectro só foi possível em 1859, quando o primeiro espectrômetro prático foi desenvolvido por Kirchhoff e Bunsen (1861). Na década de 1960, com o desenvolvimento de dispositivos semicondutores e fotoeletrônicos, o espectrômetro de leitura direta surgiu, tornando mais conveniente o armazenamento e processamento de dados espectrais.

Os espectrômetros convencionais geralmente consistem em três partes: uma fonte de luz, componentes de divisão de luz e detecção. Conforme (WANG, J.; ZHENG; WANG, 2021), Espectrômetros compactos, leves e de tamanho reduzido são extremamente necessários para uma ampla variedade de aplicações. Por exemplo, dispositivos portáteis e aplicativos de análise química baseados em smartphones no setor ambiental (e.g., análise de água e alimentos), aplicações biomédicas (e.g., análise de sangue, plasma, urina etc.), operações industriais, segurança nacional, entre outros.

Figura 14: Diferentes estratégias subjacentes para sistemas de espectrômetros miniaturizados surgiram nos últimos 30 anos. (A) Óptica dispersiva miniaturizada. (B) Filtros sintonizáveis ou em matriz de banda estreita. (C) Sistemas baseados em transformada de Four



Fonte: Modificada de (YANG, Z. *et al.*, 2021)

A Figura 14 é dividida em quatro partes, rotuladas A, B, C e D, cada uma representando uma estratégia subjacente diferente para caracterização espectral.

A parte A demonstra óptica dispersiva miniaturizada, que usa versões em pequena escala dos mesmos componentes encontrados em espectrômetros de bancada convencionais. A luz interage com essas ópticas dispersivas miniaturizadas e diferentes componentes espectrais são separadas espacialmente ao chegar a um conjunto de detectores.

A parte B demonstra filtros de banda estreita ajustáveis ou dispostos, que podem ser usados para transmitir luz seletivamente com comprimentos de onda específicos. A análise de espectros completos pode ser realizada com um único filtro (cujas propriedades transmissivas podem variar ao longo do tempo) ou passando a luz por uma série de vários filtros exclusivos de banda estreita, cada um montado em seu próprio detector.

A parte C demonstra sistemas baseados na transformada de Fourier, onde interferômetros integrados (como aqueles baseados em componentes de sistemas micro eletromecânicos (MEMS) podem ser usados para produzir interferogramas temporais ou espaciais, que são então convertidos computacionalmente em um espectro legível.

A parte D demonstra sistemas baseados em reconstrução espectral computacional, que aproveitam o poder de processamento de computador mais facilmente disponível e as reduções no tamanho e no custo do microprocessador. Esses sistemas normalmente apresentam um conjunto de detectores codificados com características distintas de resposta espectral que, quando medidos em paralelo, podem ser combinados usando algoritmos complexos para aproximar ou “reconstruir” um espectro de luz incidente. Esses sistemas podem aproveitar não apenas os avanços tecnológicos em hardware, mas também o desenvolvimento de novas abordagens computacionais, em particular, aquelas baseadas em sensoriamento compressivo e aprendizado de máquina.

Conforme, (BEĆ; GRABSKA; HUCK, 2021), a instrumentação de bancada baseada nesse esquema foi amplamente marginalizada por dispositivos de transformada de Fourier (FT) com o interferômetro de Michelson, o mais popular, ou interferômetro de polarização, menos comum. Espectrômetros FT permitem que toda a região de comprimento de onda alcance o detector na forma de um interferograma, uma quantidade dependente de frequência. Em princípio, isso proporciona um aumento direto na capacidade óptica do espectrômetro, resultando em uma relação sinal-ruído (SNR) melhor. No entanto, deve-se observar que a maioria dos espectrômetros NIR portáteis utiliza tecnologias e elementos espectroscópicos distintamente diferentes para adquirir espectros.

Existem dois tipos de fontes de luz comumente utilizadas em espectrômetros miniaturizados. O primeiro é uma lâmpada de tungstênio halógena, conhecida por sua radiação térmica e alta intensidade luminosa. Ela requer temperaturas mais elevadas para a emissão no infravermelho próximo (NIR), mas apresenta desafios em termos de estabilidade térmica. A segunda opção são os diodos emissores de luz (LEDs), que oferecem tamanho compacto, baixo consumo de energia e preço acessível. No entanto, sua largura de banda de emissão é limitada, especialmente na região do NIR. A escolha da fonte de luz depende dos requisitos específicos do espectrômetro. (BEĆ; GRABSKA; HUCK, 2021).

A seleção de comprimento de onda é essencial em um espectrômetro. Diferentes técnicas estão disponíveis, como interferômetro de Fabry-Pérot, máscara de Hadamard, filtro linear variável (LVF) e matriz de microespelhos digitais (DMD). Esses mecanismos podem ser miniaturizados por meio de tecnologias como MEMS e microssistemas optoeletrônicos (MOEMS). Cada técnica possui vantagens e

limitações, como custo, tamanho, resolução e capacidade de processamento de luz. O LVF é altamente compacto e econômico, mas requer um detector de matriz complexo, enquanto o DMD permite a obtenção de espectrômetros de varredura sem partes móveis em macro escala. A escolha da técnica depende das necessidades específicas do espectrômetro miniaturizado (BEĆ; GRABSKA; HUCK, 2021).

Dois tipos de detectores são comumente usados em espectrômetros miniaturizados. Diodos fotovoltaicos de silício são adequados para regiões visíveis e SW-NIR, porém possuem baixa relação sinal/ruído. Para as regiões de onda média e longa do NIR, os fotodetectores de InGaAs são indispensáveis devido à alta relação sinal/ruído, tempo de resposta rápido e boa eficiência quântica. Existe uma variante estendida de InGaAs para comprimentos de onda acima de 1700 nm, porém com menor relação sinal/ruído e necessidade de resfriamento (BEĆ; GRABSKA; HUCK, 2021).

Segundo (WANG, J.; ZHENG; WANG, 2021), o uso de arranjos de orifícios *subwavelength* como filtros plasmônicos oferece uma rota mais simples para a produção de filtros de cor de banda estreita. Esses arranjos metálicos plasmônicos podem ser integrados a espectrômetros miniaturizados devido ao seu tamanho reduzido.

A NanoLambda desenvolveu um sensor de espectro em escala de chip com base nessa tecnologia de arranjo de filtros plasmônicos, juntamente com um método de reconstrução de espectro usando regularização adaptativa. O sensor de espectro baseado em filtro óptico nano-plasmônico possui um arranjo de filtros nano-ópticos integrado a um *imager* CMOS regular. Ao introduzir estruturas nanométricas em filmes metálicos, esses dispositivos plasmônicos oferecem controle sobre a polarização e comprimento de onda da luz transmitida. O processo de reconstrução do espectro emprega técnicas de regularização e métodos de otimização para uma estimativa precisa do espectro. A tecnologia de arranjo de filtros plasmônicos da NanoLambda tem potencial para sensores de espectro compactos e eficientes.

3.5 CALIBRAÇÃO

O aprendizado de máquina é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que podem aprender a partir de dados, sem serem explicitamente programados para fazê-lo (ZHANG *et al.*, 2021).

Isso é feito através da construção de modelos matemáticos e estatísticos que podem ser usados para fazer previsões ou tomar decisões com base nos dados disponíveis.

Podendo ser usado para ensinar computadores a realizar tarefas úteis, incluindo detecção de objetos, reconhecimento de fala, descoberta de conhecimento médico e análise preditiva. (WATT; BORHANI; KATSAGGELOS, 2016).

A modelagem, por sua vez, é o processo de criar um modelo matemático ou estatístico que pode ser usado para descrever o comportamento de um sistema ou processo. Isso envolve identificar variáveis importantes, selecionar o tipo de modelo a ser usado e ajustar seus parâmetros para otimizar o desempenho do modelo em relação aos dados observados. O aprendizado de máquinas consiste em 4 principais etapas são elas, segundo (WATT; BORHANI; KATSAGGELOS, 2016).

- I. **Coleta dos Dados:** Assim como os humanos, o modelo precisa ser treinado para reconhecer a diferença entre os dados, (e.g., a diferença entre cachorros e gatos). Esses dados são tipicamente chamados de dados de treinamento, onde devemos ter uma quantidade de dados homogênea. Intuitivamente, quanto maior e mais diversos os dados de treinamento, melhor será a performance na tarefa de aprendizado.
- II. **Importância de Atributos:** Isso geralmente não é uma tarefa simples, pois projetar recursos de qualidade pode depender muito da aplicação, e.g., um recurso como "número de pernas" seria inútil para discriminar entre gatos e cães (já que ambos têm quatro pernas!), mas seria muito útil para diferenciar gatos e cobras. Além disso, extrair os recursos de um conjunto de dados de treinamento também pode ser desafiador. E.g, se algumas de nossas imagens de treinamento estivessem borradas ou tiradas de um ângulo em que não pudessemos ver a cabeça do animal, os recursos que projetamos podem não ser extraídos adequadamente.

- III. Treinamento do Modelo:** Após definir os atributos mais importantes para a tarefa, dos dados de treinamento, o passo final para ensinar um computador a diferenciar entre gatos e cachorros é um problema geométrico simples: fazer com que o computador encontre uma linha ou modelo linear que separe claramente gatos e cachorros em nosso espaço de recursos cuidadosamente criado. Já que uma linha (em um espaço de duas dimensões) tem dois parâmetros, uma inclinação e uma interceptação, isso significa encontrar os valores apropriados para ambos.
- IV. Testando o Modelo:** Para testar a eficácia do aprendizado do modelo, será mostrado para o modelo dados inéditos de cães e gatos, este conjunto de dados inéditos ao modelo é comumente chamado de dados de teste, e a avaliação de métricas é feita em cima destes dados. Caso as métricas não sejam satisfatórias, é necessário, coletar mais dados, com maior diversidade, avaliar novos atributos e novos modelos, e testar novamente com os dados de teste.

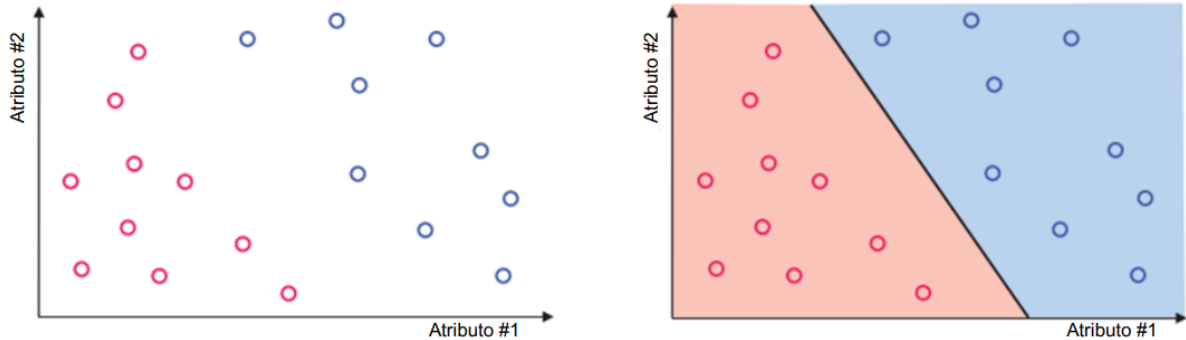
3.5.1 Aprendizado Supervisionado

Nesta seção será discutido as duas classes de aprendizado supervisionado, o qual é o tipo de aprendizado de máquina mais utilizado na prática.

3.5.1.1 Classificação

Segundo (BOROVCHNIK; BENTZ; KAPADIA, 2012), o objetivo é aprender um mapeamento das entradas X para as saídas, onde $y \in \{1, \dots, C\}$, sendo C o número de classes. Se $C = 2$, é definido como classificação binária ($y \in \{0,1\}$), se $C > 2$, então é definido como classificação multiclasse. Na Figura 15 temos a representação de uma classificação binária.

Figura 15: (Figura à esquerda) Pequeno Conjunto de treinamento bidimensional composto por duas classes distintas, vermelha e azul. (Figura direita) Um modelo linear é treinado para separar as duas classes.



Fonte: Modificada de (WATT; BORHANI; KATSAGGELOS, 2016).

I. *Extremely randomized trees classifier (Extra tree)*

Conforme (SHARMA; KUMAR; JAIN, 2022), o aprendizado neste método é realizado por meio da agregação das saídas recebidas de várias árvores de decisão não correlacionadas montadas em uma floresta para obter os resultados de segregação. A forma como ele opera é sinônimo do acaso da floresta. No entanto, se diferencia especificamente disso na maneira como as árvores de decisão são construídas na floresta. A árvore de decisão é construída na Floresta de Extra Árvores, onde é gerada a partir do treinamento em suas fases iniciais.

O algoritmo utiliza o processo de construção de árvores de decisão, onde cada nó é dividido em subconjuntos com base em critérios como Gini impurity equação 3 ou entropia equação 4. A diferença chave do Extra Tree está na seleção dos pontos de corte para essas divisões. Em vez de escolher o ponto de corte ótimo, ele escolhe os pontos de corte aleatoriamente dentro de um intervalo definido. A aleatoriedade introduzida pelo método de construção e amostragem de recursos ajuda a reduzir o *overfitting*, tornando o Extra Tree menos sensível a ruídos nos dados de treinamento e melhorando sua capacidade de generalização para novos dados. F_j é a frequência do rótulo j em um nó e 0 é o número de rótulos únicos.

$$\text{Gini Impurity} = \sum_{j=1}^0 f_j(1 - f_j) \quad (3)$$

$$Entropy = \sum_{j=1}^0 - f_j \log (f_j) \quad (4)$$

3.5.1.1.1 Métricas Modelos de Classificação

Onde TP = verdadeiro positivo, FP = falso positivo, TN = verdadeiro negativo e FN = falso negativo, e esses valores são derivados da matriz de confusão (SHARMA; KUMAR; JAIN, 2022).

3.5.1.1.2 Acurácia

A acurácia refere-se a quanto o algoritmo treinado prevê as classes corretas. A acurácia pode ser definida pela equação 5.

$$ACURÁCIA = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

3.5.1.1.3 Recall

A Recall também é conhecida como a completude de um classificador. Recall também pode ser definido como o número de verdadeiros positivos dividido pelo número de verdadeiros positivos e o número de falsos negativos, definido pela equação 6.

$$RECALL = \frac{TP}{TP+FN} \quad (6)$$

3.5.1.1.4 Precisão

A precisão também é conhecida como a exatidão dos classificadores. A precisão também pode ser definida como o número de verdadeiros positivos dividido pelo número de verdadeiros positivos e falsos positivos, definido pela equação 7.

$$PRECISÃO = \frac{TP}{TP+FP} \quad (7)$$

3.5.1.1.5 F1-Score

A F1-score refere-se ao equilíbrio entre a precisão e a Recall. A F1-score também pode ser definida como a média harmônica de precisão e Recall, definida pela equação 8.

$$F1 - SCORE = 2 * \frac{PRECISÃO*RECALL}{PRECISÃO+RECALL} \quad (8)$$

3.5.1.1.6 Matriz de Confusão

Uma matriz de confusão é uma matriz bidimensional onde suas linhas mostram os rótulos verdadeiros e suas colunas mostram os rótulos previstos por um classificador. Criando uma matriz de confusão para um classificador multiclasse é simples e direto, como mostra na equação 9.

$$M(r, c) = \sum_{i=1}^m (I(y_i = r) I(h(x_i) = c)),$$

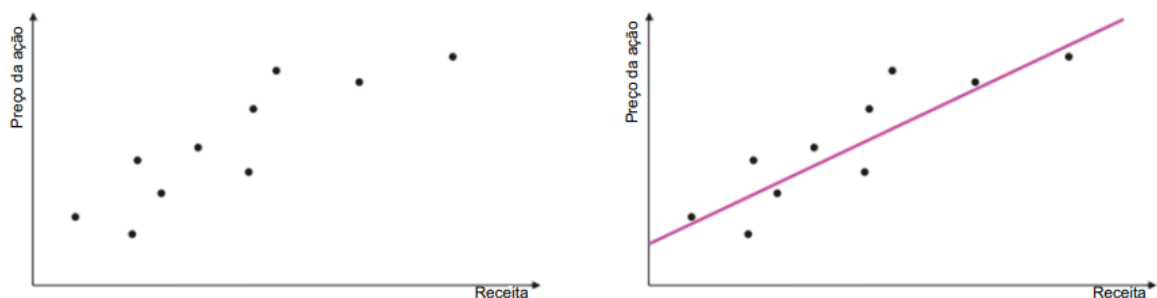
$$\forall r, c \in \{0, \dots, q - 1\} \quad (9)$$

Onde M é a matriz de confusão, r e c são a linha e a coluna da matriz de confusão, m é o número de instâncias no conjunto de dados de teste, $I(\cdot)$ é a função indicadora, x_i é a i -ésima entrada para o classificador $h(\cdot)$, y_i é o rótulo verdadeiro atribuído à entrada x_i , e q é o número de classes.

3.5.1.2 Regressão

Segundo (BOROVCHNIK; BENTZ; KAPADIA, 2012), a regressão é como a classificação, com a exceção que a variável de resposta se trará de um valor contínuo. A Figura 16 demonstra um simples exemplo. Há uma única entrada de valor real $X_i \in R$ e uma única resposta de valor real. Várias extensões desse problema básico podem surgir, $y_i \in R$.

Figura 16: (Esquerda) Conjunto de dados de treinamento de brinquedo de dez corporações com seus preços de ações e valores de receita associados. (Direita) Modelo linear é ajustado aos dados.



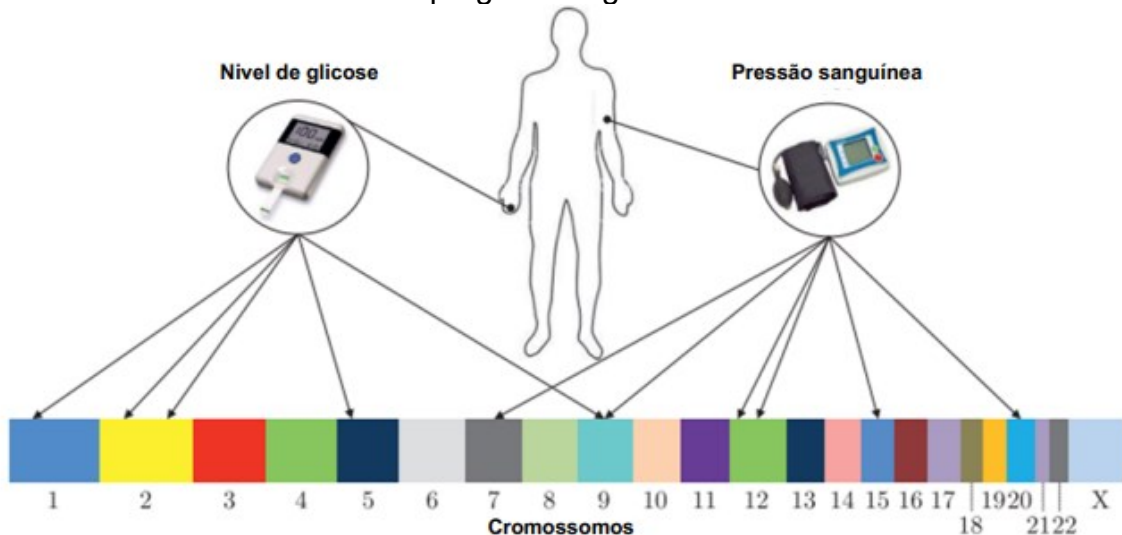
Fonte: Modificada de (WATT; BORHANI; KATSAGGELOS, 2016).

Conforme BOROVCHNIK; BENTZ; KAPADIA (2012), modelos de regressão podem ser utilizados nas seguintes tarefas, previsão da idade do indivíduo que vai

assistir um vídeo no Youtube, previsão da localização no espaço 3D de um vetor final do braço do robô, dados os sinais de controle (torques) enviados para seus vários motores, dentre outras ocasiões.

WATT; BORHANI; KATSAGGELOS (2016), apresenta a aplicação de regressão, no estudo de associação ampla do genoma, visando compreender as conexões entre dezenas de milhares de marcadores genético, obtidos de todo o genoma humano de vários indivíduos, com doenças como pressão alta, problemas cardíacos, diabetes, diferentes formas de câncer, com o objetivo de no futuro, produzir terapias direcionadas a genes, conforme podemos visualizar na Figura 17.

Figura 17: Ilustração conceitual de um estudo associação ampla do genoma empregando regressão.



Fonte: Modificada de (WATT; BORHANI; KATSAGGELOS, 2016).

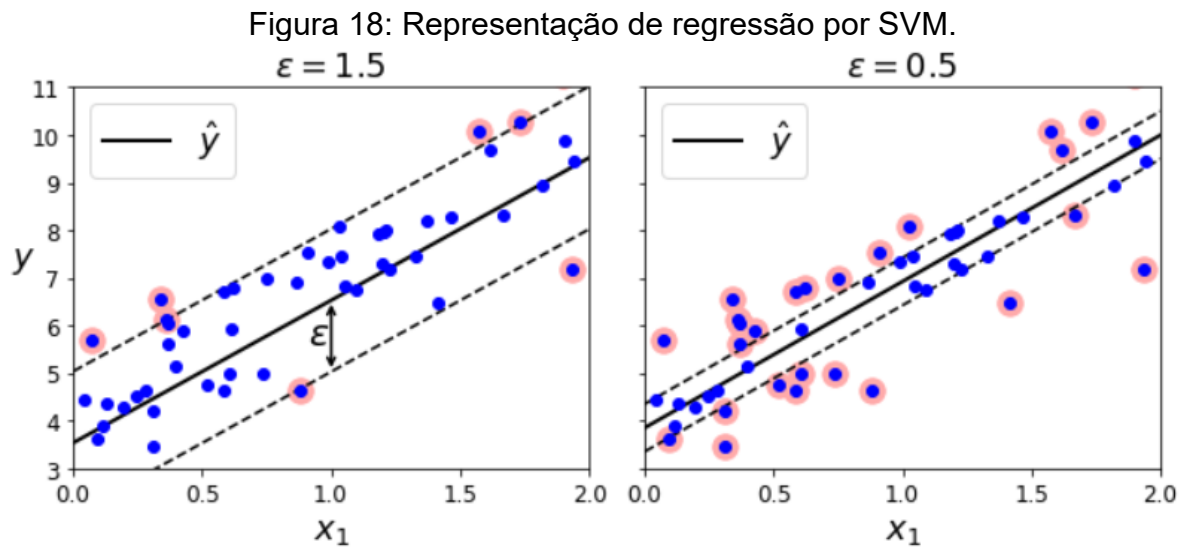
I. Máquinas de Vetores de Suporte

Segundo (GÉRON, 2021)(DEISS *et al.*, 2020), Máquinas de Vetores de Suporte (SVM), é um modelo robusto e versátil, capaz de fazer classificações lineares ou não lineares, e de regressão, geralmente tem equilíbrio adequado entre precisão preditiva e a capacidade de generalizar modelos treinados para dados não vistos.

Modelos de regressão SVM possuem duas *features* principais, que podem ser otimizadas, a primeira é a seleção do Kernel e a segunda é a tolerância ao ruído na função de perda insensível ao epsilon (ϵ) para cada kernel.

As funções de kernel retornam o produto interno entre dois pontos em um espaço de recursos adequado, definindo assim uma noção de similaridade em espaços de alta dimensão (Karatzoglou et al., 2006).

Conforme, (DEISS *et al.*, 2020) , na regressão SVM, existem quatro famílias principais de kernels: linear, polinomial, radial e sigmoide, e cada kernel tem seus próprios parâmetros de otimização e possíveis cenários de adequação. Na Figura 18, temos uma representação da utilização de regressão SVM.



Fonte: Produção do próprio autor (2023).

I. Decision Tree Regressor

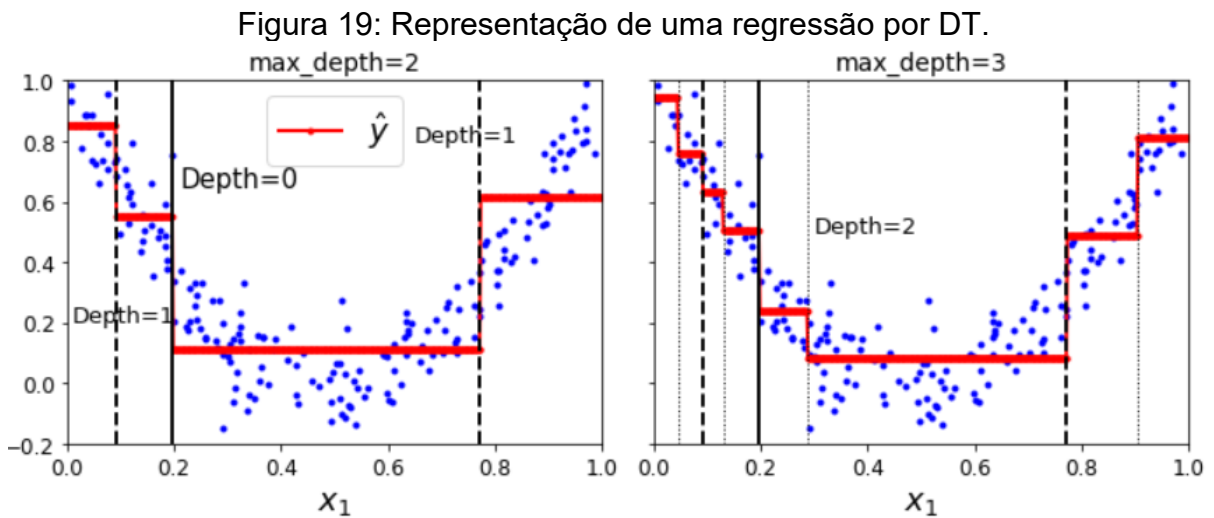
Segundo, (AMASYALI; EL-GOHARY, 2018), os algoritmos de Árvore de decisão (DT) usam uma árvore para mapear instâncias em previsões, em um modelo de árvore de decisão, cada nó representa um recurso, cada ramificação da árvore representa um valor diferente para um recurso e cada nó folha representa uma classe de predição.

As árvores de decisão são um algoritmo flexível que pode crescer com uma quantidade maior de dados de treinamento. As árvores de classificação e regressão floresta aleatória (RF) e árvores de reforço (BT) são os métodos de árvore de decisão mais amplamente utilizados na área de previsão.

Segundo, (SHOKREKHODAEI *et al.*, 2021) , O processo de divisão de um nó continua até atingir um critério de parada. Exemplos desses critérios são quando a

entropia de um nó é menor que um limite pré-especificado ou quando a profundidade de uma árvore atinge um nível pré-especificado. Quando isso ocorre, os nós tornam-se folhas, ou nós terminais de uma árvore.

O desempenho preditivo de um modelo DT também depende dos valores dos hiperparâmetros DT. Estes incluem a profundidade máxima de uma árvore, denotada como *max_depth*, o número mínimo de amostras necessárias para dividir um nó de decisão, denotado como *min_samples_split* e o número máximo de folhas em uma árvore, denotado como *max_leaf_nodes*, conforme demonstra (YANG, L.; SHAMI, 2020). Na figura 19 temos uma representação da utilização de regressão por DT.



Fonte: Produção do próprio autor (2023).

II. Random Forest Regressor

A Regressão por Floresta Aleatória (RF) é um método de aprendizado em conjunto que utiliza o método de bagging para combinar várias árvores de decisão (YANG, L.; SHAMI, 2020). Cada árvore é amostrada de forma independente e aleatória, e seus resultados são combinados usando a regra da maioria (INJADAT et al., 2018). O algoritmo RF opera da seguinte forma: escolhe-se T número de árvores para crescer em número de variáveis para dividir cada nó, onde m é menor ou igual a M , o número de variáveis de entrada (MALIK; SHAHZAD; KHAN, 2011). Durante o crescimento de cada árvore, uma amostra de tamanho N é construída a partir dos N casos de treinamento com reposição, e uma árvore é desenvolvida a partir dessa nova amostra. Em cada nó, m variáveis são selecionadas aleatoriamente dentre as M disponíveis para encontrar a melhor divisão. As árvores crescem até o tamanho

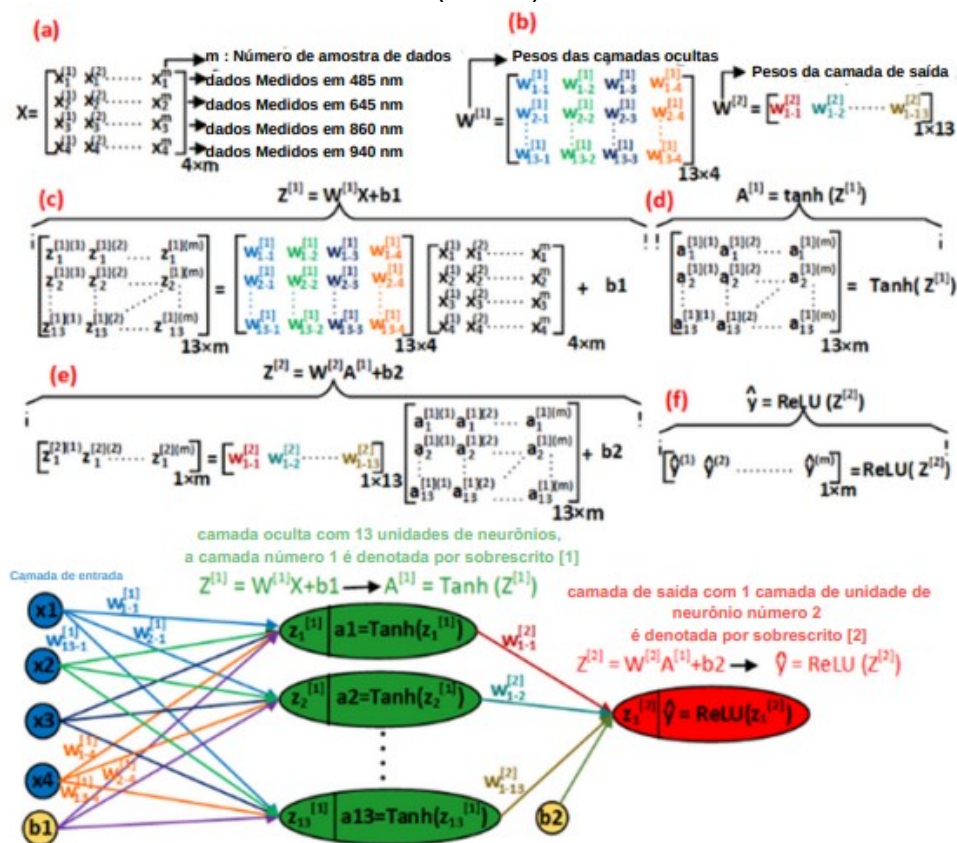
máximo sem poda. Para fazer previsões para um ponto X, os votos de cada árvore na floresta são coletados e a decisão final é tomada por meio de votação majoritária

III. Artificial Neural Network

Redes neurais artificiais são um subconjunto de aprendizado de máquina e utilizam um grupo de algoritmos para resolver problemas tanto lineares quanto não lineares (SHOKREKHODAEI *et al.*, 2021).

Existem diferentes tipos de redes neurais, e cada uma possui sua própria estrutura e aplicação, como a rede neural de alimentação direta (FFNN), a rede neural recorrente (RNN), a rede neural convolucional (CNN), a rede neural híbrida etc.

Figura 20: Exemplificação Rede neural feedforward com uma camada oculta. (FFNN)



Fonte: Modificada de (SHOKREKHODAEI *et al.*, 2021).

Segundo (SHOKREKHODAEI *et al.*, 2021), FFNN consiste em várias unidades de processamento chamadas neurônios, que são organizados em diferentes camadas. A rede geralmente é composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada camada contém neurônios que são

conectados para modelar os dados e fazer uma previsão na camada de saída, conforme ilustrado na Figura 20.

3.5.1.2.1 Métricas Modelos de Regressão

3.5.1.2.2 Coeficiente de Determinação (R²)

O R² é uma medida de quão bem os resultados observados são reproduzidos pelo modelo, com base na proporção da variação total explicada pelo modelo (DEISS *et al.*, 2020), o R² mede a porcentagem de variação nos dados de saída que é explicada pela variável independente. O R² varia de 0 a 1, onde um valor mais próximo de 1 indica um ajuste melhor do modelo aos dados observados.

Conforme a equação 10 desta medida abaixo, na qual \hat{y} representa o valor predito, y representa o valor médio das amostras e y_i representa o valor real.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - y)^2} \quad (10)$$

3.5.1.2.3 Erro Quadrático Médio (RMSE)

O RMSE é a diferença entre os valores observados e os valores previstos (DEISS *et al.*, 2020).

Conforme equação 11, o RMSE é calculado como a média dos erros quadráticos e, em seguida, é tirada a raiz quadrada para obter a mesma unidade de medida dos valores originais. O RMSE fornece uma medida da dispersão dos erros entre os valores observados e os valores previstos. Quanto menor o valor do RMSE, melhor é o ajuste do modelo aos dados, o que a torna uma métrica de avaliação excelente para problemas nos quais grandes erros não são tolerados, como é o caso de exames médicos e projeções de preços.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - y_i)^2} \quad (11)$$

3.5.1.2.4 Erro Absoluto Médio (MAE)

De acordo com a equação 5, o MAE consiste na média das distâncias entre valores preditos e reais. Diferentemente do RMSE, essa métrica não “pune” tão severamente os outliers do modelo.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (12)$$

3.5.1.2.5 Erro Percentual Médio Absoluto (MAPE)

Em contraste com as métricas anteriores, essa medida exprime uma porcentagem, obtida por meio da equação 13, pela divisão da diferença entre predito (\hat{y}) e real pelo valor real (y_i).

Assim como o RMSE e o MAE, quanto menor o valor, mais preciso seria o modelo de regressão.

Devido a sua formulação, essa métrica não lida tão bem se tratando de problemas com um grande alcance de números, como uma regressão que prevê uma variável que vai de 10 a 20.000.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (13)$$

3.5.2 Requisitos Para Desempenho *In Vitro* e ISO 15197

Os testes usando o limite de detecção são baseados na experiência empírica da espectroscopia. Tecidos vivos são mais complexos que soluções de glicose em água pura e meios líquidos turvos. Considerando a precisão requerida para dispositivos de glicemia (20 mg/dL), limites de detecção são obtidos: água pura (1 mg/dL) e meio líquido turvo (5 mg/dL).

A Organização Internacional de Normalização (ISO) é uma organização independente e não governamental que define e desenvolve especificações para procedimentos, serviços e produção de produtos de alta qualidade, confiáveis e seguros em diversas indústrias, incluindo dispositivos médicos (GONZALES; MOBASHSHER; ABBOSH, 2019).

De acordo com a ISO 15197:2013, os critérios mínimos de acurácia requerem que:

- 95% dos resultados inferiores a 100 mg/dL estejam dentro de ± 15 mg/dL do valor de referência;
- 95% dos resultados iguais ou acima de 100 mg/dL estejam dentro de $\pm 15\%$ do valor de referência;
- 99% dos resultados estejam dentro das zonas A ou B da grade.

3.5.3 Conjuntos de Treinamento, Validação e Teste

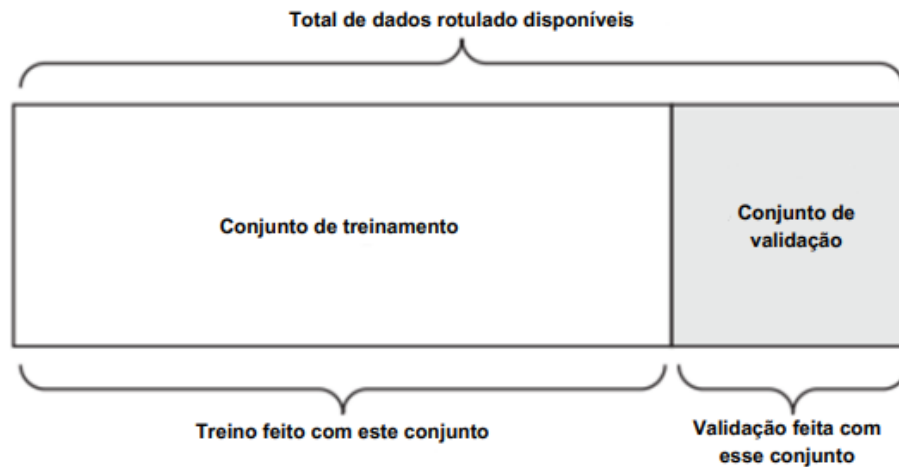
Avaliar um modelo sempre se resume a dividir os dados disponíveis em três conjuntos:

Treinamento, validação e teste. A calibração é feita com os dados de treinamento e avaliação do algoritmo é realizada com os dados de validação. Após a verificação das métricas do algoritmo, é realizado o teste do modelo com os dados de teste, este conjunto não deve ser conhecido pelo algoritmo.

3.5.3.1 *Hold-out*

Hold-out é quando é dividido o conjunto de dados em um conjunto treino e 'teste'. O conjunto de treinamento é o qual o modelo é treinado e o conjunto de teste é utilizado para avaliar o desempenho desse modelo em dados não vistos. Uma divisão comum ao usar o método hold-out é usar 80% dos dados para treinamento e os 20% restantes dos dados para teste, conforme podemos ver na Figura 21.

Figura 21: Divisão Hold-out.



Fonte: Modificada de (SVYATKOVSKIY; KATES-HARBECK; TANG, 2017).

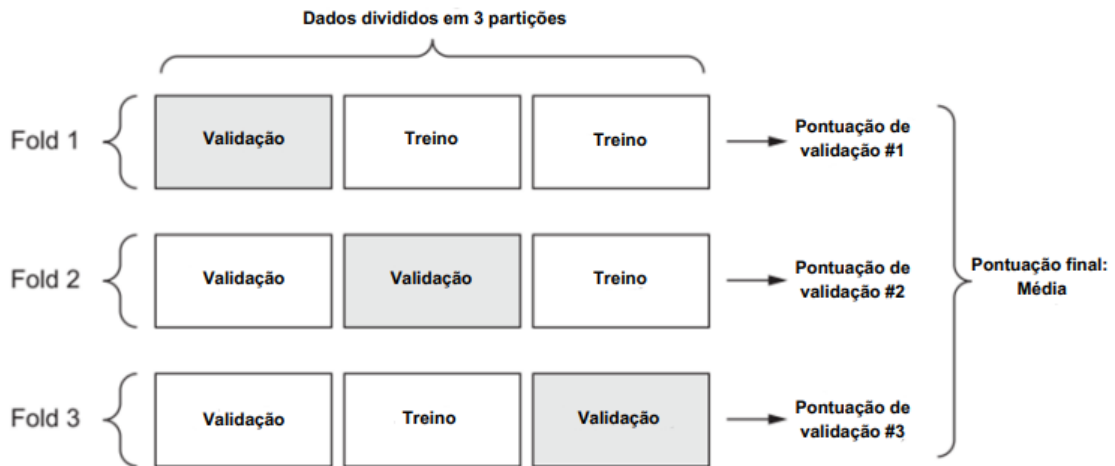
3.5.3.2 Validação Cruzada

Validação cruzada (CV) é uma técnica comumente utilizada para avaliação de desempenho de modelos de aprendizado de máquina. O CV consiste em particionar os dados em conjuntos (partes), onde um conjunto é utilizado para treino e outro conjunto é utilizado para teste e avaliação do desempenho do modelo.

Segundo, (Drokow et al., 2021; Parraga et al., 2021), na técnica de validação cruzada CV, a avaliação pode diferir grandemente dependendo de como é feita a divisão entre o conjunto de treinamento e o conjunto de teste. Assim, o CV com *K-fold* é um dos métodos preferencialmente usados e eficientes que se destacam na CV.

A maior vantagem de usar a técnica de *K-Fold* CV é que ela não se preocupa com como os dados são divididos (PRUSTY; PATNAIK; DASH, 2022). Conforme podemos na Figura 22.

Figura 22: Divisão dos dados utilizando Validação Cruzada, 3 camadas.



Fonte: Modificada de (SVYATKOVSKIY; KATES-HARBECK; TANG, 2017).

Conforme, (BERRAR, 2019), são criados k subconjuntos disjuntos de tamanho aproximadamente igual. Aonde, "*fold*" se refere ao número de subconjuntos resultantes. Essa divisão é realizada amostrando aleatoriamente os casos do conjunto de aprendizado sem reposição. O modelo é treinado usando $k - 1$ subconjuntos, que juntos representam o conjunto de treinamento. Em seguida, o algoritmo é aplicado ao subconjunto restante, que é denominado conjunto de validação, e o desempenho é medido. Esse procedimento é repetido até que cada um dos k subconjuntos tenha servido como conjunto de validação. A média das k medições de desempenho nos k conjuntos de validação é o desempenho validado cruzada.

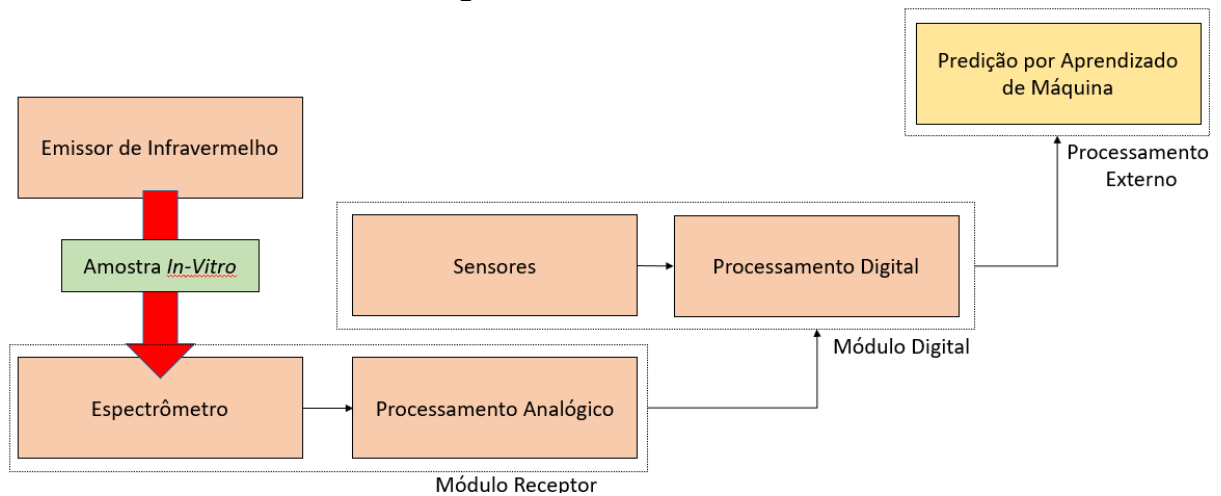
A validação cruzada frequentemente envolve amostragem aleatória estratificada, o que significa que a amostragem é realizada de tal forma que as proporções de classes nos subconjuntos individuais reflitam as proporções no conjunto de aprendizado (BERRAR, 2019).

Por exemplo, suponha que o conjunto de aprendizado contenha $n = 100$ casos de duas classes, a classe positiva e a classe negativa, com $n^+ = 80$ e $n^- = 20$. Se a amostragem aleatória for feita sem estratificação, é bastante possível que alguns conjuntos de validação contêm apenas casos positivos (ou apenas casos negativos). No entanto, com a estratificação, cada conjunto de validação na validação cruzada com 5 *folds* é garantido de conter aproximadamente dezesseis casos positivos e quatro casos negativos, refletindo assim a proporção de classes no conjunto de aprendizado.

4 MATERIAIS E MÉTODOS

Inicialmente, propôs-se o diagrama de blocos para o sistema de predição da concentração de glicose, de uma amostra *in-vitro*, com base em dispositivos semelhantes apresentados na revisão literatura, conforme podemos ver na Figura 23.

Figura 23: Diagrama de blocos do sistema proposto para predição de concentração de glicose em meio *in-vitro*.



Fonte: Produção do próprio autor (2023).

4.1 EMISSOR DE INFRAVERMELHO

O módulo emissor consiste em um conjunto de diodos emissores de luz (LED), que abrangem o comprimento de onda da luz visível (400 à 700nm) e infravermelho (750nm à 1000nm).

O emprego da tecnologia LED é crescente nos mercados médicos e de eletrônica comerciais de oximetria de pulso, além de aplicações em nível de pesquisa como medidor de glicemia.

Tabela 1: Principais características dos emissores

Modelo (Fabricante)	Comprimento (nm)	Potência (mW)	Corrente (mA)	HPBW (°)
KEL5002A (Kodenshi)	940 ± 22,5	5,7	50 – 100	20
L-7113SF6C (Kingbright)	860 ± 25	3,8	20 – 50	20
LED750-03AU (Roithner)	750	25	20 – 50	20
OS5RBA5111A (OptoSupply)	660	3,3	20 – 70	15

Fonte: Produção do próprio autor (2023).

O principal aspecto considerado na escolha dos LEDs foi o comprimento de onda. A Tabela 1, indica os valores de comprimento de onda dominante. Neste aspecto é possível escolher entre duas abordagens, para a medição de glicose, os LEDs de infravermelho próximo na banda de combinação (2000 a 2500 nm), apresentam ótima resposta a glicose, porém além de serem afetados com o ruído da água, possuem cotações elevadas, deste modo não foram empregados neste projeto.

Sendo assim, optou-se pelo uso de três LEDs de infravermelho para a banda de sobretom de segunda ou maior ordem (750 a 1000 nm). E o LED complementar na cor vermelha (660 nm), os comprimentos de onda foram definidos baseados na literatura, como podemos ver em Tabela 1.

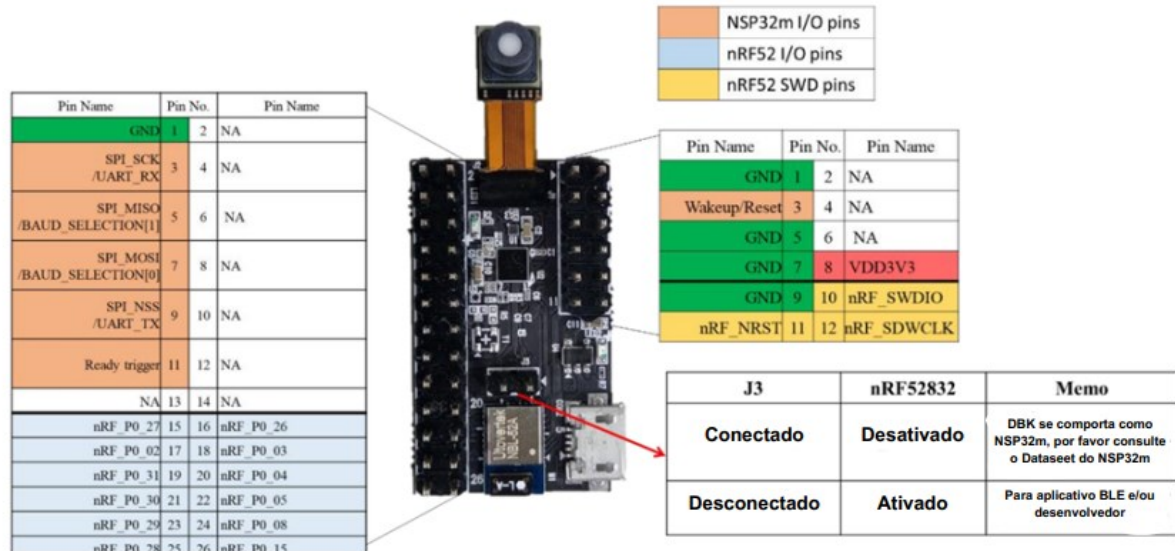
4.2 MÓDULO RECEPTOR

O módulo receptor de infravermelho, se trata de um espectrômetro da empresa *nanolambda*, o qual possui uma faixa do espectro eletromagnético sensível de 600 a 1000 nm.

Segundo (CHEN; CHUANG; CHANG, 2020), o espectrômetro NSP32 (*nanolambda*, Daejeon, Coreia), tem como sua tecnologia central do sensor filtros plasmônicos, que podem ser integrados a um fotodetector regular, como um óxido metálico complementar de metal (CMOS). Em resumo, pode-se considerar que o sensor é composto por muitos fotodetectores de espectro estreito. A vantagem desse sensor é seu tamanho pequeno em escala de chip e sua capacidade de detectar o

espectro em uma ampla faixa de 600 a 1000 nm. Com um intervalo de medida de 5 nm.

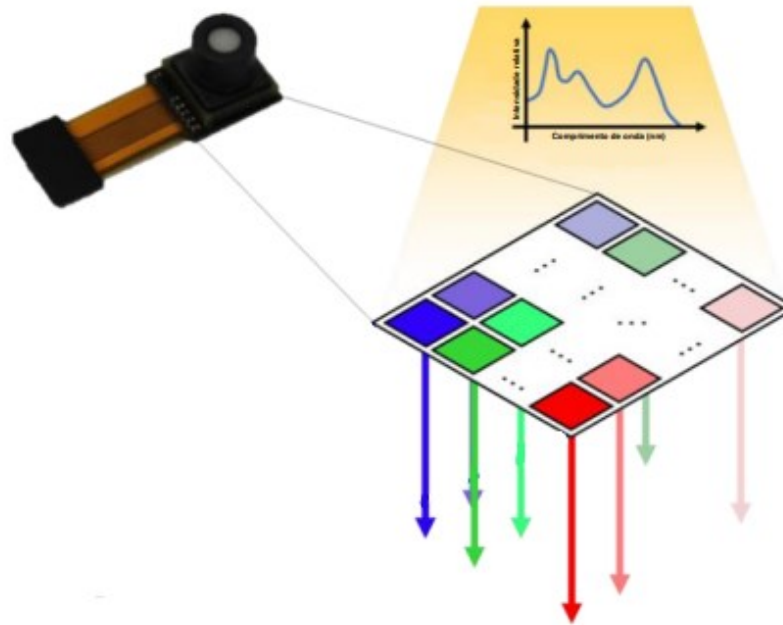
Figura 24 : Pinout NSP32m DBK Development Board Kit.



Fonte: Modificada de Datasheet DBK (2018).

De acordo com (SARKAR; ASSAAD, 2022), CMOS significa *Complementary Metal-Oxide-Semiconductor*, que é um tipo de sensor de imagem usado em câmeras digitais e outros dispositivos de imagem. É caracterizado por baixo consumo de energia, alta velocidade e alta densidade de integração, tornando-o adequado para dispositivos portáteis e compactos.

Figura 25: Estrutura do chip-scale espectrômetro.



Fonte: Modificada de (CHEN; CHUANG; CHANG, 2020).

De acordo com, (CHEN; CHUANG; CHANG, 2020)(CHANG *et al.*, 2019)(HE *et al.*, 2015)(SARKAR; ASSAAD, 2022)(LEE *et al.*, 2016)(LEE *et al.*, 2017), O sensor MW-PPG foi projetado usando filtros plasmônicos que podem ser integrados a um fotodetector comum, como um gerador de imagens semicondutor CMOS. Os filtros plasmônicos introduzem estruturas em nanoescala em filmes metálicos, que fornecem uma maneira única de controlar a polarização e o comprimento de onda da luz que passa pelas estruturas. O comprimento de onda de transmissão só pode ser controlado pelas estruturas laterais em uma única camada, o que possibilita a produção de um dispositivo contendo diferentes canais de filtro de maneira econômica.

O sensor MW-PPG foi projetado para detectar sinais MW-PPG de forma síncrona na faixa de 600 a 1000 nm em um intervalo de 5nm. A integração dos filtros plasmônicos com o sensor CMOS permite o controle da polarização e comprimento de onda da luz, o que é essencial para a medição precisa da saturação de oxigênio no sangue periférico (SpO2) e da pressão arterial.

4.3 MÓDULO DIGITAL

O módulo digital consiste em duas partes: sensores digitais e unidade de processamento digital de sinal.

Sensores digitais de temperatura e umidade foram considerados, pela influência desses parâmetros em sistemas de instrumentação eletrônica de medição de glicemia (SMITH, 2020), com objetivo do aumento de acurácia na predição. O modelo optado, foi o circuito integrado modelo HDC1080, do fabricante Texas Instruments, cujas características são apresentadas na Tabela 2:

Tabela 2: Principais características do sensor digital (HDC1080).

Características	Descrição	Observação
Faixa de temperatura	-40 a 120 °C	Valores recomendáveis
Acurácia de temperatura	± 0,2 °C	Típica
Faixa de umidade	0 a 100%	Não condensante
Acurácia de umidade	± 2%	Típica
Tensão de alimentação	2,7 a 5 V	Valores recomendáveis
Frequência de comunicação	10 a 400 kHz	Via I ² C
Encapsulamento	PWSON DMB	SMD

Fonte: Produção do próprio autor (2023).

O HDC1080 é um sensor digital de umidade com sensor de temperatura integrado que fornece excelente precisão de medição com consumo de energia muito baixo. ("HDC1080 LOW POWER, HIGH ACCURACY DIGITAL HUMIDITY SENSOR WITH TEMPERATURE SENSOR", 2016). Com aplicações em sensores vestíveis e dispositivos médicos.

Optou-se pelo uso da plataforma Portenta H7 Lite, do fabricante Arduino, cujas principais características são apresentadas na Tabela 3:

Tabela 3: Principais características da plataforma (Portenta H7 Lite)

Características	Descrição
Microcontrolador	Arm® Cortex®-M7 480 MHz + Arm® 32-bit Cortex®-M4 240 MHz
Memória Interna	8 Mb SDRAM / 16 MB QSPI Flash
Interface de Comunicação	4x I ² C, 6x SPIs, 4x UARTs, USB-C
Compilador	Online ou Instalável (C/ C++/ Micropython)
Tensão de alimentação	3,3 V

Fonte: Produção do próprio autor (2023)

O Portenta H7 Lite é capaz de realizar de forma simultânea a execução de código de alto nível e tarefas em tempo real, graças à presença de seus dois processadores. Um exemplo disso é a capacidade de executar tanto o código compilado pelo Arduino quanto o código MicroPython simultaneamente, permitindo a comunicação entre os dois núcleos. Com o H7 Lite, é possível facilmente executar processos criados com o TensorFlow™ Lite, onde um dos núcleos pode ser dedicado ao processamento de algoritmos de visão computacional em tempo real, enquanto o outro núcleo se encarrega de operações de baixo nível, como o controle de um motor ou o desempenho de funções de interface com o usuário.

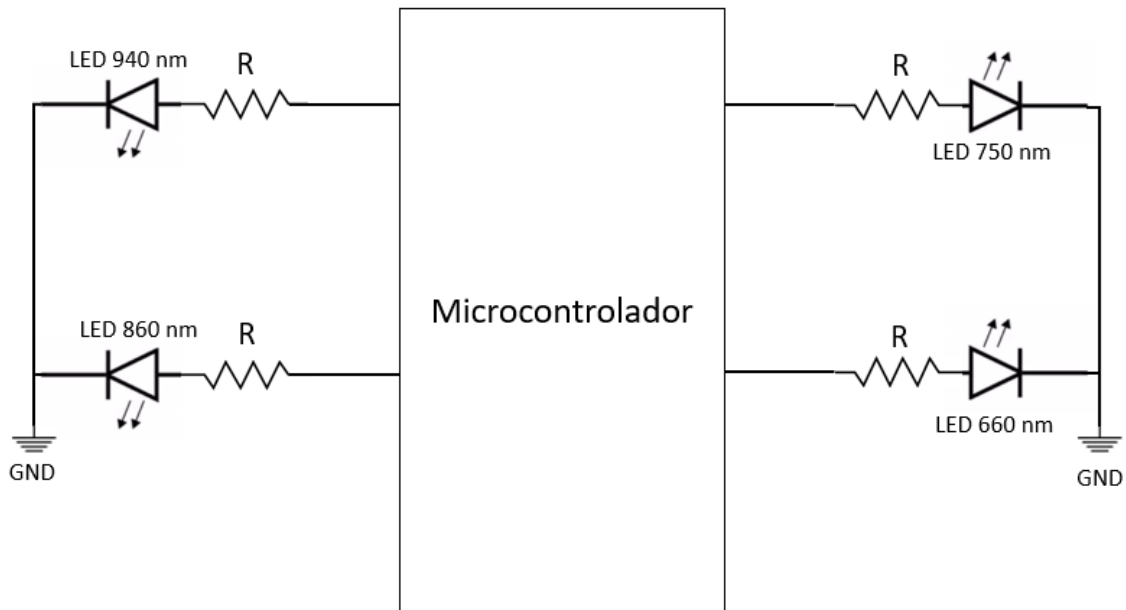
5 DESENVOLVIMENTO

A etapa de desenvolvimento teve início tomando-se as partes individuais do sistema em implementação concomitante de hardware e software. Após isso, integraram-se as partes individuais no sistema como todo. Nos tópicos subsequentes, apresentam-se os detalhes das partes individuais, bem como a montagem integrada final.

5.1 MÓDULO EMISSOR

O circuito do módulo emissor consiste em conexões entre os LEDs e a plataforma Portenta H7 Lite. As conexões são divididas em: Sinais de comunicação (PORTAS INPUT), e sinais de alimentação (GND). Conforme Figura 26.

Figura 26: Circuito Módulo emissor.



Fonte: Produção do próprio autor (2023).

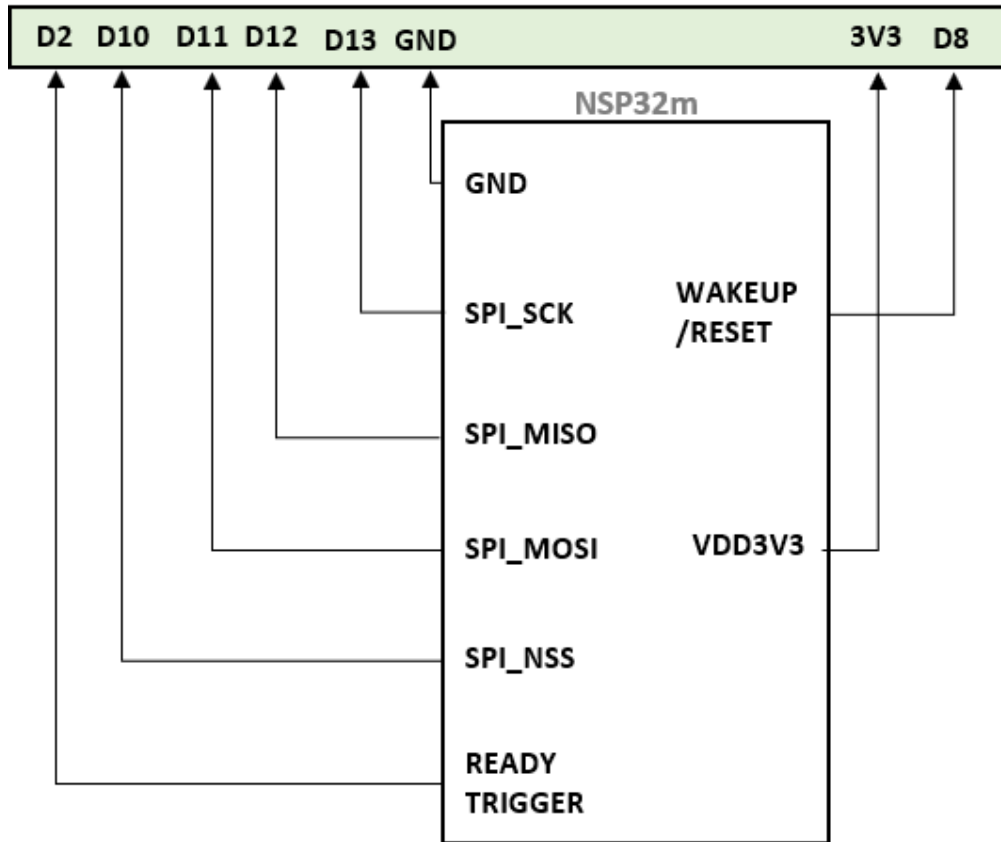
Para a alimentação positiva dos LEDs, utilizaram-se os pinos de comunicação do Portenta H7 Lite de tensão de 3,3 V (VOUT). Utilizou-se um resistor externo de 150 Ω para o ajuste da tensão.

5.2 MÓDULO RECEPTOR

O circuito do módulo receptor consiste em conexões entre o Nano espectômetro NSP32m e a plataforma Portenta H7 Lite.

Utilizou-se o pino de tensão 3 V (VOUT) para alimentação do espectômetro, para a alimentação positiva. Foram utilizadas as portas digitais da plataforma Portenta H7 Lite, para a comunicação por SPI, conforme Figura 27.

Figura 27: Circuito módulo receptor
Portenta H7 Lite

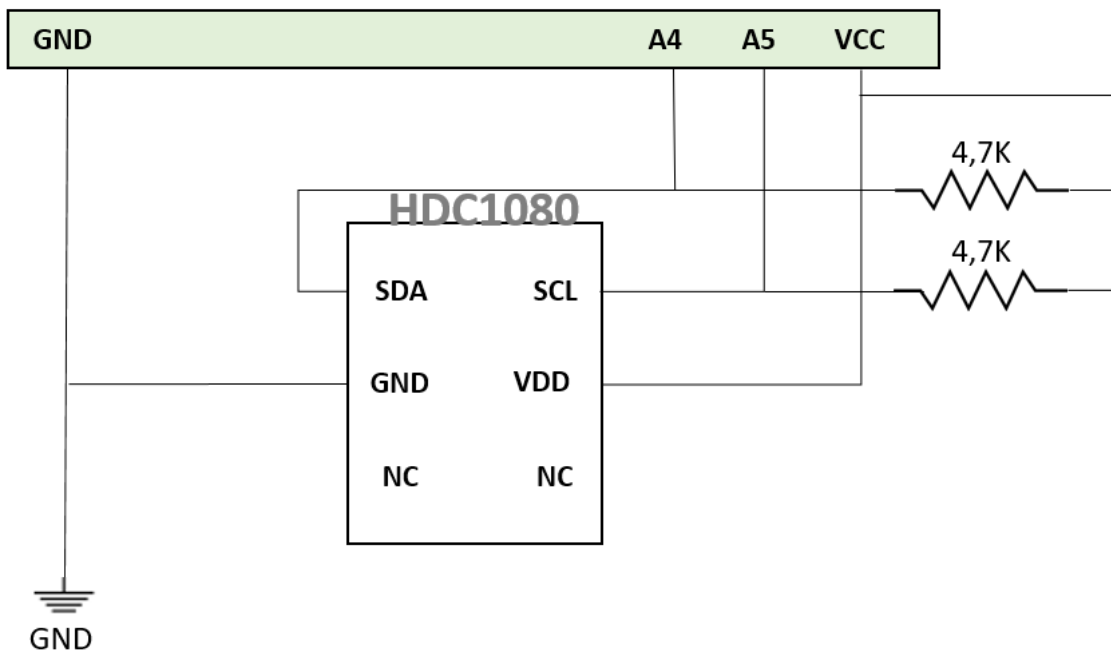


Fonte: Produção do próprio autor (2023).

5.3 MÓDULO DIGITAL

O circuito do módulo digital consiste em conexões entre o sensor HDC1080 e a plataforma Portenta H7 Lite. Essas conexões são divididas em: alimentação do HDC1080 (GND e VDD) e sinais de comunicação (SDA e SCL), conforme a Figura 28.

Figura 28: Circuito módulo digital.
Portenta H7 Lite



Fonte: Produção do próprio autor (2023).

Para a alimentação positiva do sensor, utilizou-se o pino da plataforma Portenta H7 Lite de tensão de 5 V. Para comunicação foram utilizadas as portas analógicas (A4 e A5) I²C.

6 METODOLOGIA DA AMOSTRA

Nesta seção serão debatidos os materiais para o desenvolvimento da solução da amostra e as etapas de obtenção das amostras utilizadas no projeto.

6.1 MATERIAIS

Os materiais utilizados para a produção da amostra foram, D-(+) - Glicose Anidra teor 99.7% (Neon), glicosímetro Guide Me (Accu Chek®), água de alta pureza (Milliq®) e vórtex mixer (KASVIbasic®) utilizado para a solubilização da amostra.

6.2 PREPARO DA SOLUÇÃO PADRÃO

A metodologia utilizada para a produção da amostra foi adaptada e otimizada de (SHOKREKHODAEI *et al.*, 2021), onde foram pesados 250mg para 100mL de água Milliq e solubilizados em vórtex mixer durante 10 minutos, a qual foi mantida em temperatura ambiente (24°C) durante 1 hora.

6.3 LINEARIDADE

A linearidade foi determinada por meio da equação da reta. A partir da solução padrão descrita no item 6.2, a reta foi construída nas concentrações de 9 a 0,5 mL, conforme descrito na Tabela 4. O diluente utilizado para o preparo das diluições foi água milliq e solubilizadas em vórtex mixer durante 1 minuto e mantida em descanso por 1 minuto.

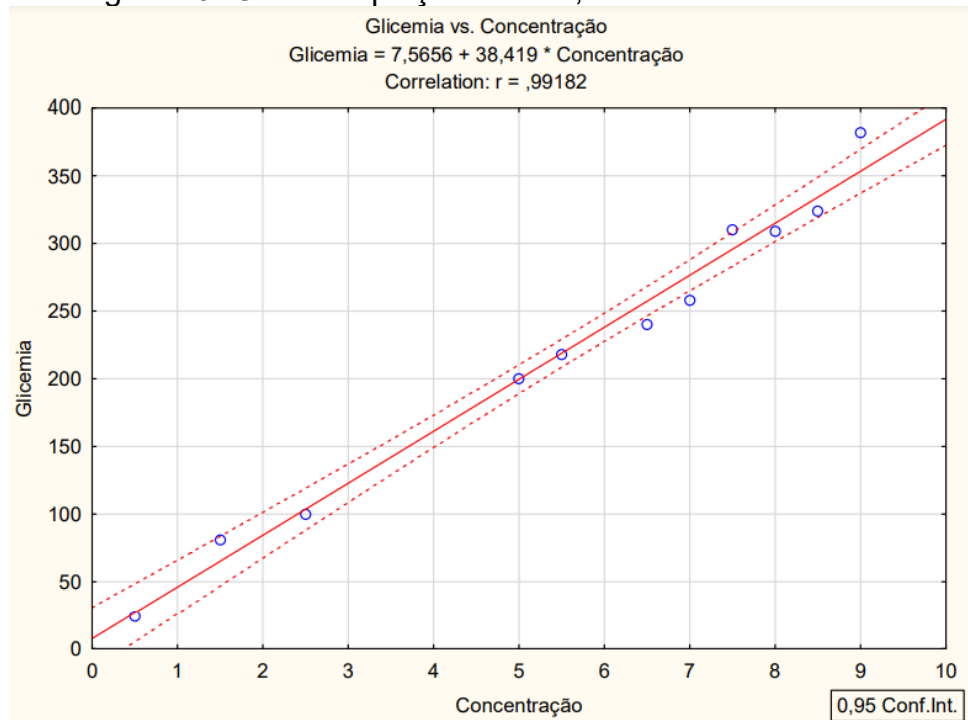
Tabela 4: Diluições a partir da solução padrão.

Amostra	Concentração de padrão (mL)	Concentração de água milliq (mL)	Glicosímetro (mg/dL)
1	9	1	382
2	8,5	1,5	324
3	8	2	309
4	7,5	2,5	310
5	7	3	258
6	6,5	3,5	240
7	5,5	4,5	218
8	5	5	200
9	2,5	7,5	100
10	1,5	8,5	81
11	0,5	9,5	24

Fonte: Produção do próprio autor (2023).

A equação da reta foi determinada por meio do estudo de regressão linear, pelos métodos de mínimos quadrados. Os valores experimentais foram tratados estatisticamente, utilizando o software Statística®.

Figura 29: Gráfico equação da reta, realizado no Statistica.



Fonte: Produção do próprio Autor (2023).

6.4 PREPARAÇÃO DAS DILUIÇÕES PARA PREDIÇÃO

Para as diluições procedeu-se da seguinte forma:

9 mL da solução padrão foi pipetado em um balão volumétrico de 10 mL, completando-se o volume com água milliq, e levados ao vórtex mixer durante 1 minuto e mantido em descanso durante 1 minuto. O mesmo processo seguiu-se para as demais diluições conforme Tabela 5.

Tabela 5: Preparação das amostras.

Amostra	Concentração de padrão (mL)	Concentração de água milliq (mL)	Equação da reta (mg/dL)
1	9	1	353
2	8,7	1,3	342
3	8,5	1,5	334
4	8,2	1,8	323
5	8	2	315
6	7,7	2,3	303
7	7,5	2,5	296
8	7,2	2,8	284
9	7	3	276

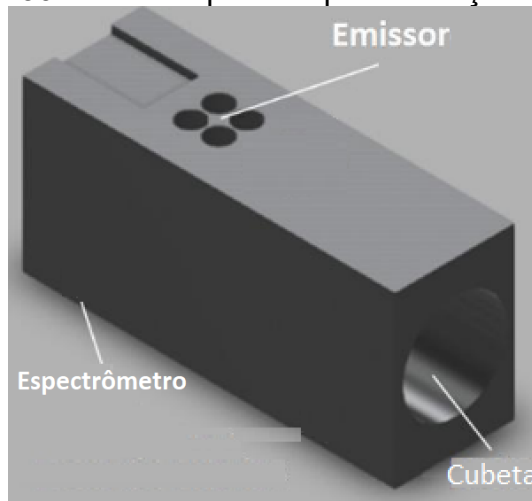
⋮	⋮	⋮	⋮
33	1	9	46
34	0,7	9,3	34
35	0,5	9,5	27
36	0,2	9,8	15

Fonte: Produção do próprio autor (2023).

6.5 MEDIÇÃO *IN VITRO*

Para a realização do experimento, fixou-se uma cubeta no interior de um bloco de plástico, o qual possui orifícios para o posicionamento dos LEDs e do espectrômetro, conforme Figura 30. A cubeta utilizada é de quartzo, que possui formato retangular, com 2 faces polidas e 2 faces foscas, caminho óptico de 10mm e 3,5mL de volume interno. A cubeta foi posicionada de forma que a face polida fosse mantida perpendicular a incidência dos LEDs e leitura do espectrômetro.

Figura 30: Bloco de plástico para medição *in vitro*.



Fonte: Produção do próprio autor (2023).

Após posicionar a cubeta dentro do bloco de plástico, foram feitas leituras com o módulo emissor, e armazenadas em um banco de dados. Após o término das leituras, a cubeta era retirada e lavada com água milliq, para eliminar qualquer resíduo da amostra anterior, e então inserida a amostra seguinte, contendo a próxima diluição. Esse processo se seguiu até ser completado as diluições.

O módulo receptor, obtém resposta do espectro na faixa de 600 a 1000 nm, em um intervalo de 5 nm, sendo assim temos 81 intervalos de respostas, para cada emissão do conjunto de LED do módulo emissor. Foram usados 4 LEDs, com isso, temos 81 respostas para cada LED. Também foram obtidos 3 repostas do espaço de cores CIE 1931, o qual retorna os valores de X, Y e Z, as quais representam a cor em um determinado ponto no espaço, esses valores são utilizados para descrever a cor de maneira objetiva e são obtidos a partir de medições espectrofotométricas. Por fim teremos duas leituras do módulo digital, as quais são a temperatura ambiente e humidade. Logo, ao final de uma leitura teremos 336 respostas do módulo receptor, mais 2 repostas do módulo digital, totalizando 338 as quais serão caracterizadas como *features*, no momento da predição dos valores.

Foram feitas 243 leituras, destas 197 foram utilizadas para o conjunto de teste, 46 foram utilizadas para o conjunto de treino. Em outro momento foram realizadas 20 leituras apenas com objetivo de teste do modelo, servindo como “Novas Leituras”, a Tabela 6 representa o banco de dados produzido neste estudo.

Tabela 6: Resumo do banco de dados.

LED_660_600n m	LED_660_605n m	LED_660_610n m	·	TEMP	HUMIDAD E	mg/d L
0.00000000e+0 0	0.00000000e+0 0	2.96968380e+0 7	·	23.70	84.70	353
0.00000000e+0 0	0.00000000e+0 0	2.76083160e+0 7	·	23.63	82.62	353
0.00000000e+0 0	6.92990500e+0 6	3.66356310e+0 7	·	23.78	83.21	353
⋮	⋮	⋮	·	⋮	⋮	⋮
0.00000000e+0 0	3.70103800e+0 6	4.88348310e+0 7	·	24.84	71.18	15
0.00000000e+0 0	1.66541360e+0 7	5.81976280e+0 7	·	25.03	77.91	15

0.00000000e+0	3.45914700e+0	4.51200860e+0	.			
0	6	7	.	25.21	74.03	15
			.			

Fonte: Produção do próprio autor (2023).

6.6 PREPARAÇÃO DOS DADOS

O banco de dados é dividido em dois conjuntos, conjunto de treino/validação e conjunto de teste. O conjunto de treinamento/validação é usado para criar, ajustar e medir o desempenho do modelo. Quando o modelo está pronto, ele é testado usando o conjunto de teste, o qual não foi submetido na fase de criação do modelo, não havendo vazamento de informações deste conjunto. Normalmente o conjunto de treinamento/validação é dividido em dois conjuntos independentes: conjunto de treinamento e conjunto de validação. Neste caso esta divisão não é produzida porque o processo de validação é realizado através da validação cruzada fundindo os processos de treinamento e validação. Por fim o modelo também foi submetido ao conjunto “Novas Medidas”, o qual produzido em separado.

7 PREDIÇÃO

Nesta seção, está exposta a etapa de criação de modelos e a avaliação da capacidade do instrumento na predição da glicose in vitro. Os dados foram divididos em um conjunto de treinamento/validação e um conjunto de teste. Para cada modelo estabelecido, na abordagem de regressão, foram calculados o R2 e o RMSE. Já na abordagem de classificação, foi calculada a acurácia e o F1-Score para a avaliação dos modelos.

7.1 CRIAÇÃO E TESTE DOS MODELOS

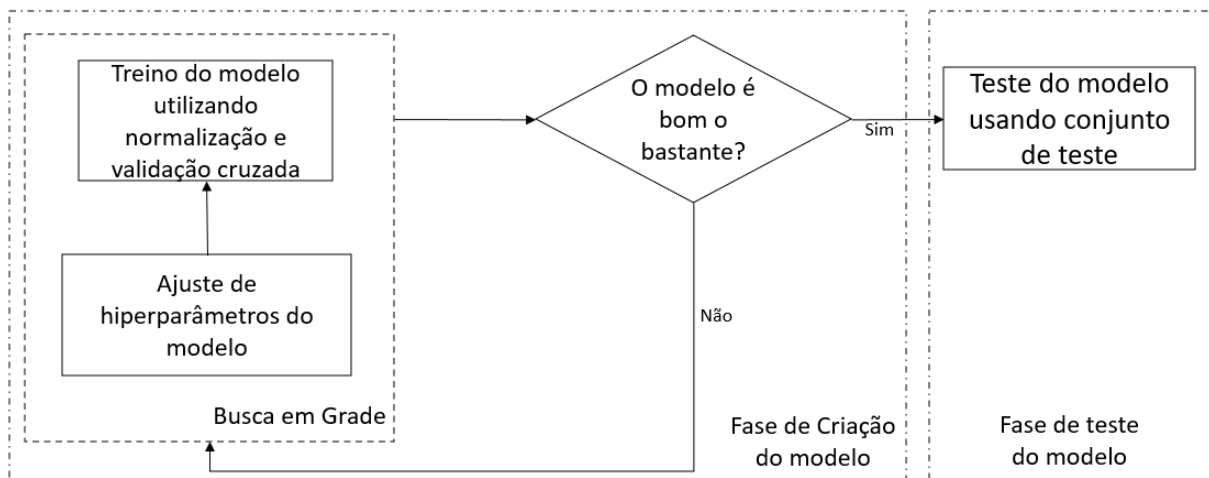
Para a criação e avaliação dos modelos foi utilizado, *jupyter notebook*, que se trata de uma aplicação que permite a criação e compartilhamento de documentos interativos contendo código executável, compatível com diversas linguagens de programação, incluindo *Python*. Sendo possível importar as bibliotecas de cada

algoritmo, e métricas para avaliação. Todos os códigos em *python* e *c++* usados neste trabalho, podem ser encontrados neste repositório online¹.

Também foram testadas diferentes abordagens de normalização de dados, sendo estas, *Standart scaler*, *minmax scaler* e *robust scaler*. *Standard Scaler* ajusta os dados para ter uma média zero e um desvio padrão de um, *MinMax Scaler* dimensiona os dados para um intervalo específico, geralmente entre 0 e 1 e *Robust Scaler* é uma transformação resistente a *outliers*, escalando os dados com base nos quartis.

A Figura 31 mostra o processo de criação e teste dos modelos da esquerda para a direita. É possível dividir esse processo em duas partes. A primeira parte envolve a criação do melhor modelo possível, com o melhor ajuste aos dados possível. Foram selecionados 6 modelos de regressão e 1 modelo de classificação para a predição do nível de glicose por meio não invasivo durante o teste *in vitro*:

Figura 31: Fluxograma de criação de modelo e processo de teste de modelo.



Fonte: Produção do próprio autor (2023).

7.2 DEFINIÇÃO DE ALGORITMOS PARA A CRIAÇÃO DOS MODELOS

O modelo de *Ensemble Voting (Voting)*, utilizando os algoritmos *Random Forest* e *Gradient Boosting*, é uma técnica eficaz para problemas de regressão. O *Random Forest* é um método que combina múltiplas árvores de decisão independentes, e suas previsões são agregadas por meio de votação ou média,

¹ Repositório Online: <https://github.com/Henriquerezer/Dissertacao>

proporcionando uma estimativa robusta. Por outro lado, o *Gradient Boosting* é uma abordagem sequencial que constrói um conjunto de modelos fracos, geralmente árvores de decisão simples, onde cada modelo é ajustado aos resíduos do modelo anterior, resultando em um refinamento iterativo das previsões. Ao combinar esses dois algoritmos em um modelo de *Ensemble Voting*, é possível aproveitar a diversidade e as capacidades de aprendizado de ambos, resultando em uma previsão final mais precisa e confiável para problemas de regressão.

O KNN (K-Nearest Neighbors) é um modelo de aprendizado de máquina amplamente utilizado tanto para problemas de regressão quanto de classificação. Em problemas de regressão, esse modelo prevê os valores de saída com base nas variáveis de entrada e na proximidade com os K vizinhos mais próximos no conjunto de dados de treinamento. O KNN para regressão é apreciado por sua simplicidade e capacidade de lidar com dados não lineares. No entanto, é importante mencionar que o KNN pode ser sensível a outliers e requer a escolha adequada do valor de K, que representa o número de vizinhos considerados para a predição. Além disso, o KNN também é aplicável a problemas de classificação, onde ele atribui rótulos de classe com base na maioria dos K vizinhos mais próximos.

O Random Forest (RF) pertence à categoria de algoritmos baseados em árvores de decisão, mas se destaca por criar um conjunto de árvores de decisão independentes e combiná-las para obter uma predição final. Nesse modelo, várias árvores de decisão são treinadas com amostras aleatórias dos dados de treinamento e com uma seleção aleatória de recursos em cada árvore. Essa abordagem assegura a diversidade entre as árvores, o que ajuda a evitar o overfitting e, ao mesmo tempo, aumenta a capacidade de generalização do modelo. Além disso, é importante destacar que o Random Forest também é aplicável a problemas de classificação, onde ele combina as previsões das árvores para atribuir rótulos de classe.

O Gradient Boosting (GB) é um modelo de aprendizado de máquina amplamente utilizado tanto para problemas de regressão quanto de classificação. Em problemas de regressão, ele constrói um conjunto de árvores de decisão de forma sequencial, onde cada árvore é ajustada para corrigir os erros residuais das árvores anteriores. Isso permite que o modelo melhore gradativamente suas previsões, combinando as contribuições de várias árvores. O GB é altamente eficaz em lidar com relações complexas entre variáveis, capturando interações não lineares e fornecendo previsões precisas. Além disso, quando aplicado a problemas de classificação, o GB

segue a mesma estratégia de construção sequencial de árvores, mas para atribuir rótulos de classe em vez de valores contínuos. Portanto, o GB é uma ferramenta versátil e poderosa que pode ser usada tanto para problemas de regressão quanto de classificação, com adaptações apropriadas para cada aplicação.

O XGB (Extreme Gradient Boosting Regressor) é um modelo de aprendizado de máquina baseado em gradient boosting que utiliza a biblioteca XGBoost. Compartilhando muitas semelhanças com o Gradient Boosting Regressor, mas possui algumas características adicionais. O XGB aprimora o gradient boosting tradicional com técnicas avançadas, como amostragem estocástica, regularização e otimização eficiente. Essas técnicas melhoram o desempenho, a estabilidade e a velocidade do treinamento do modelo. Suportando também a utilização de funções de perda personalizadas e possui opções para lidar com dados ausentes de forma mais eficiente.

O LightGBM Regressor é um modelo de aprendizado de máquina baseado em gradient boosting, que faz parte da família de algoritmos de boosting de árvores, utiliza uma abordagem "leaf-wise" para construir as árvores de decisão. Nesse caso, a divisão ocorre na direção que mais reduz a função de perda, em vez de expandir em todas as direções. Isso permite que o LightGBM atinja uma maior profundidade com menos divisões, tornando-o mais eficiente e capaz de lidar com conjuntos de dados grandes.

O algoritmo Extra Trees é uma extensão das Árvores de Decisão que se destaca em tarefas de classificação multiclases. Suas principais vantagens incluem a redução de *overfitting*, a robustez a dados ruidosos, eficiência computacional e a capacidade de lidar com conjuntos de dados desbalanceados. Isso é alcançado por meio da introdução de aleatoriedade na seleção de dados e características, tornando-o uma escolha sólida em muitos cenários. No entanto, ele pode apresentar maior variância nas previsões e ser menos interpretável do que alguns outros métodos devido à sua natureza de ensemble. A escolha do algoritmo depende das necessidades específicas do problema em questão.

7.3 CRIAÇÃO DO MODELO

Normalmente modelos são treinados primeiro e depois validados, porém no caso da CV, ambas as etapas são realizadas ao mesmo tempo, *grid search*, ou busca

em grade, é uma técnica amplamente utilizada na otimização de hiperparâmetros em modelos de aprendizado de máquina. Consiste em definir um conjunto de valores possíveis para cada hiperparâmetro do modelo e, em seguida, realizar uma busca exaustiva por combinações desses valores.

Durante essa busca, o modelo é treinado e avaliado com cada conjunto de hiperparâmetros, geralmente usando uma métrica de desempenho, como a acurácia ou o RMSE. O objetivo é encontrar a combinação de hiperparâmetros que resulta no melhor desempenho do modelo. É uma abordagem sistemática e confiável para a seleção de hiperparâmetros, porém, pode ser computacionalmente custoso devido ao número de combinações a serem testadas. No entanto, sua utilidade na obtenção de um desempenho otimizado do modelo o torna uma ferramenta valiosa na construção de modelos de aprendizado de máquina.

Além da definição de diversos hiperparâmetros do modelo, é possível incluir um conjunto de combinações, de técnicas de normalização, filtrando diferentes abordagens de importância de atributos e de diferentes divisões da validação cruzada, buscando assim o melhor desempenho do modelo.

Uma parte importante do ajuste do modelo é chamada de seleção de recursos. Contrariando o senso comum, usar mais preditores do que o estritamente necessário para criar um modelo pode não levar a um modelo melhor. Alguns preditores podem ser considerados como "ruído". Reduzir o número de preditores diminui a complexidade dos modelos, reduz o tempo de treinamento e diminui possíveis problemas de sobreajuste do modelo.

7.4 TESTE DO MODELO

Essa etapa é de suma importância para avaliar a capacidade de generalização do modelo. Uma capacidade adequada de generalização implica na habilidade do modelo em realizar previsões precisas com base em novas medidas. Os modelos são ajustados visando obter o melhor desempenho durante o processo de validação cruzada. Contudo, um baixo valor do RMSE obtido durante a validação cruzada não garante um desempenho similar com novos dados. Com o intuito de verificar essa capacidade, o modelo é testado com medidas que não foram utilizadas no processo de treinamento.

Por isso, o conjunto de teste é separado no início e posteriormente utilizado para avaliar a capacidade de generalização dos modelos. Caso o RMSE do conjunto de teste seja significativamente maior do que o obtido na validação cruzada, significa que o modelo está super-ajustado, sendo necessário reiniciar a fase de criação do modelo desde o início. É importante que os valores do RMSE obtidos durante a previsão no conjunto de teste e durante a validação cruzada sejam semelhantes.

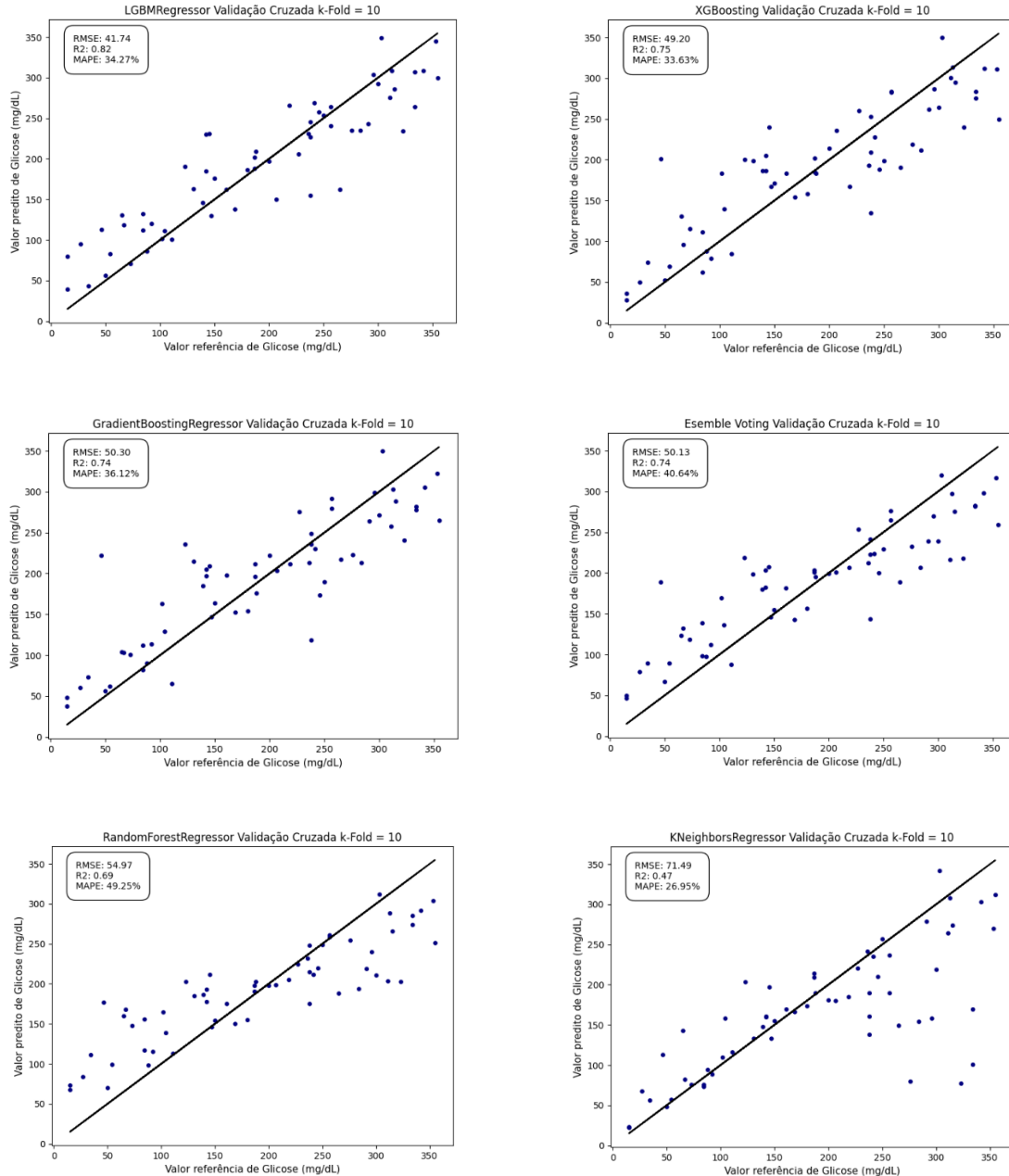
Caso contrário, uma das possíveis razões é que o conjunto de teste é pequeno, sendo necessário um banco de dados mais abrangente. A fase de teste é obrigatória, uma vez que a sua negligência impossibilita a verificação do funcionamento adequado do modelo na realidade, quando utilizado para prever novos dados, mesmo que as métricas da validação cruzada sejam promissoras.

7.5 VALIDAÇÃO DOS MODELOS DE REGRESSÃO

A Figura 32 mostra as previsões geradas em função da concentração de glicose nas amostras *in vitro*, utilizando validação cruzada com 10 divisões (número de dados do conjunto de treinamento), onde os dados de treinamento foram divididos em dois conjuntos de treino e validação, na proporção de 70% e 30%, esta abordagem visa trazer maior generalização nas previsões do modelo.

Após a fase de criação do modelo, foi obtido como as 3 melhores métricas de RMSE, os modelos LGBM, XGB e GB, com RMSE correspondentes de 41.74, 49.20 e 50.30. A reta representa os pontos referência de glicose, e os pontos em azul representam os valores preditos pelo modelo.

Figura 32: Predição modelos com melhores hiperparâmetros.



Fonte: Produção do próprio autor (2023).

A Tabela 7, expõe as métricas de avaliação dos modelos durante a validação cruzada, onde os modelos estão organizados frente ao valor do RMSE, em ordem decrescente. Os dados são divididos entre predição gerada sem normalização (SN) dos dados, utilizando como normalização a abordagem standard scaler (SS), minmax scaler (MS) e robust scaler (RS).

Tabela 7: Métricas de Validação.

Modelo	RMSE (SN)	R ² (SN)	RMSE (SS)	R ² (SS)	RMSE (MS)	R ² (MS)	RMSE (RS)	R ² (RS)
LGBM	41.74	0.82	45.80	0.78	45.01	0.79	48.10	0.76
XGB	49.20	0.75	50.49	0.74	50.14	0.74	50.09	0.74
GB	50.25	0.74	51.09	0.73	52.82	0.72	51.50	0.72
Voting	50.68	0.74	50.13	0.74	50.31	0.74	52.22	0.72
RF	54.97	0.69	55.02	0.69	55.30	0.70	55.05	0.69
KNN	71.49	0.47	76.86	0.74	75.44	0.41	82.09	0.30

Fonte: Produção do próprio autor (2023).

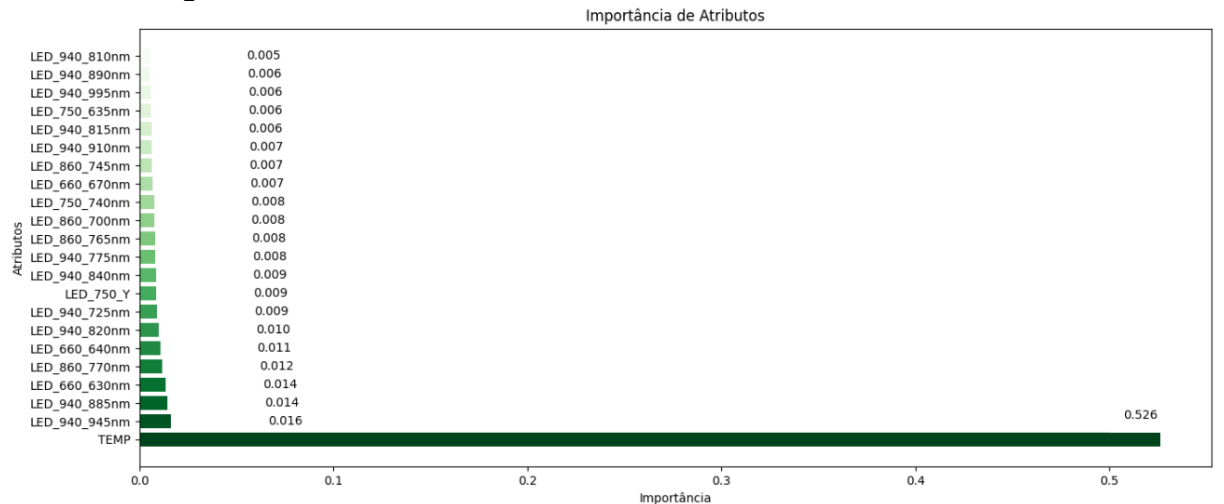
Os modelos computacionais têm a capacidade de explorar as relações entre diversos preditores, conferindo maior valor aos preditores de baixa correlação do que quando considerados individualmente no modelo final. Para avaliar a importância de cada preditor, é empregado um método que consiste em treinar o modelo excluindo um preditor por vez. Quanto mais significativo for o aumento do RMSE maior será a relevância desse preditor para o modelo. Com o objetivo de obter uma perspectiva geral da importância dos preditores, são aplicadas normalizações através da divisão dos incrementos do RMSE pelo maior incremento observado. Essa abordagem permite tirar conclusões relevantes a partir do gráfico gerado.

Na Figura 33, temos os 22 atributos que possuem maior impacto frente a adequação do modelo aos dados, os atributos foram escolhidos por meio do método importância de atributos de uma RF.

O princípio é transformar o valor de um parâmetro de característica em um número aleatório, calcular seu impacto na precisão do modelo e mensurar a importância desse parâmetro com base no valor médio da diminuição da precisão obtida a partir de múltiplos cálculos. Quanto maior o valor, maior é a importância da variável (ZHAO *et al.*, 2022).

Nota-se que o atributo com maior impacto é a temperatura, em seguida a de duas leituras do LED de 940 nm, no comprimento de onda de 945nm e 895nm.

Figura 33: Atributos com maiores relevância frente aos modelos.



Fonte: Produção do próprio autor (2023).

A Tabela 8, expõe as métricas de avaliação dos modelos durante a validação cruzada, onde os modelos estão organizados frente ao valor do RMSE, em ordem decrescente. Os dados são divididos entre predição gerada sem normalização (SN) dos dados, utilizando como normalização a abordagem (SS), (MS) e (RS), utilizando como os 22 atributos na Figura 33, como atributos preditores.

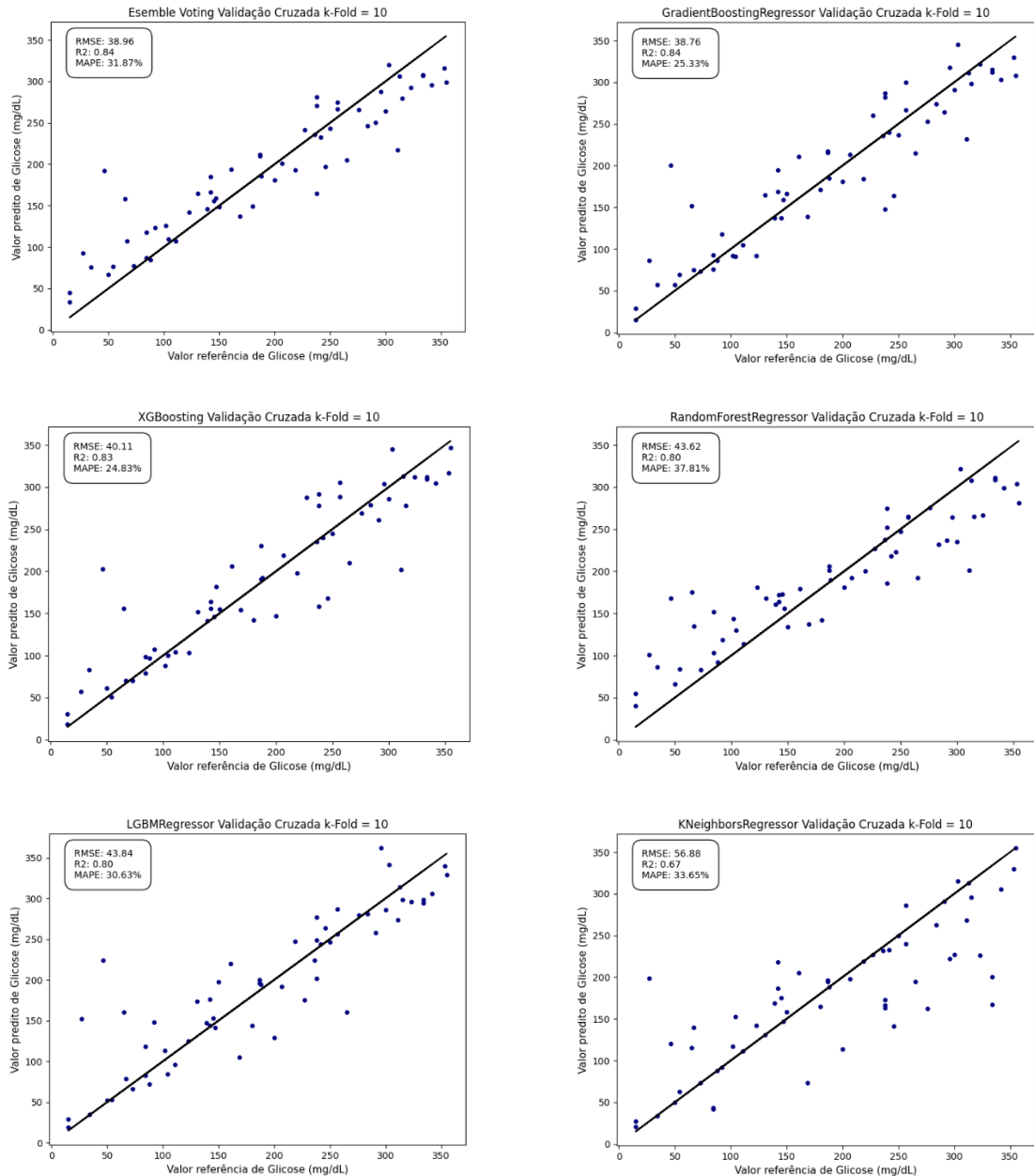
Tabela 8: Métricas de Validação para 22 atributos.

Modelo	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²
	(SN)	(SN)	(SS)	(SS)	(MS)	(MS)	(RS)	(RS)
Voting	39.30	0.84	40.25	0.84	38.96	0.84	39.43	0.83
GB	39.37	0.84	39.48	0.84	38.76	0.84	40	0.93
XGB	40.11	0.83	40.11	0.83	40.11	0.83	40.11	0.83
RF	45.09	0.79	44.80	0.79	43.62	0.80	45.37	0.78
LGBM	45.45	0.78	46.55	0.77	43.84	0.80	47	0.75
KNN	75.17	0.41	62.09	0.60	56.88	0.67	57.01	0.66

Fonte: Produção do próprio autor (2023).

A Figura 34 mostra as predições geradas para os 22 melhores atributos em função da concentração de glicose nas amostras *in vitro*, utilizando CV com 10 divisões, onde os dados de treinamento foram divididos em dois conjuntos de treino e validação, na proporção de 70% e 30%, esta abordagem visa trazer maior generalização nas predições do modelo.

Figura 34: Predição modelos com melhores hiperparâmetros e 22 atributos.



Fonte: Produção do próprio autor (2023).

Após a frase de criação do modelo, foi obtido como as 3 melhores métricas de RMSE, os modelos Ensemble Voting, GB e XGB, com RMSE correspondentes de 38.96, 38,76 e 40,11. A reta representa os pontos referência de glicose, e os pontos em azul representam os valores preditos pelo modelo.

Na tabela 9, encontram-se os hiperparâmetros estabelecidos após a conclusão da busca em grade para cada modelo. A busca em grade foi conduzida para ambas as abordagens: a primeira englobando todos os atributos e a segunda com apenas 22 atributos.

Tabela 9: Hiperparâmetros definidos após busca em grade e validação cruzada com 10-Fold.

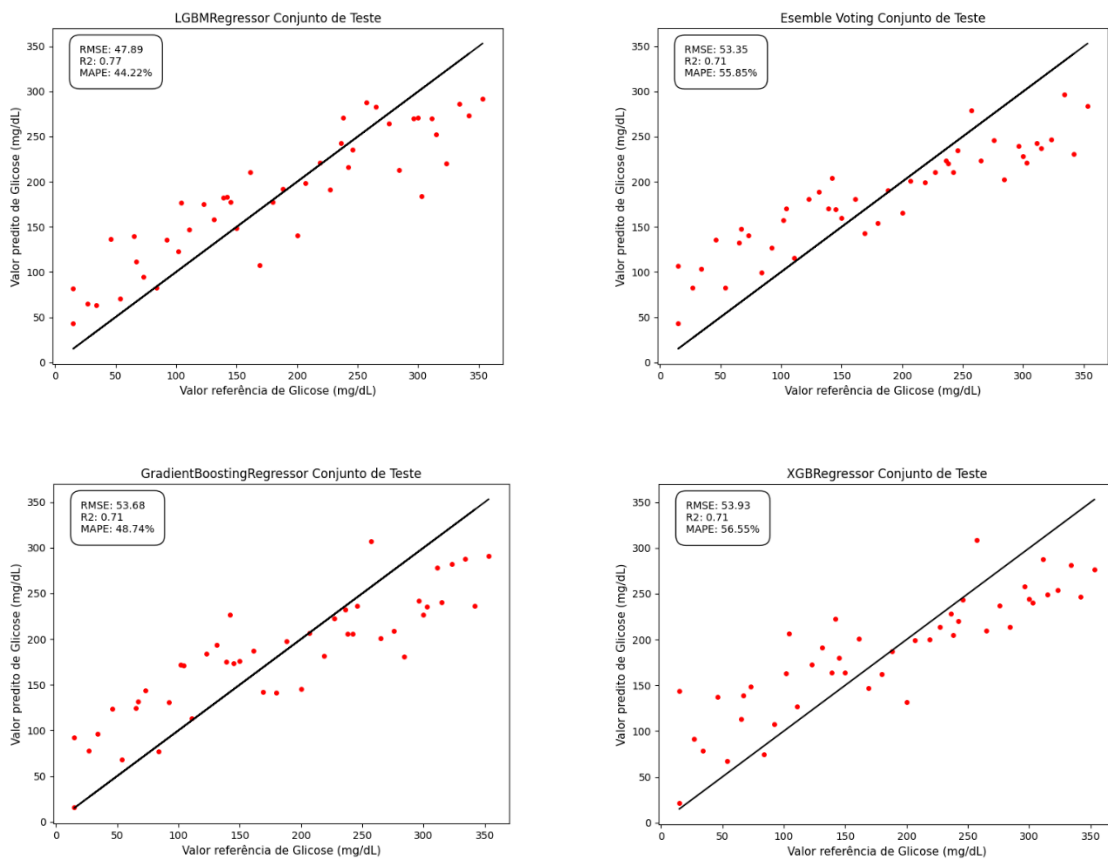
Modelo	Learning rate	Max depth	N_estimators	Min samples split	Min samples leaf	N_Neighbors	Weights
Voting	0.1	3	100	-	-	-	-
GBoosting	0.1	3	100	-	-	-	-
XGBRegressor	0.1	3	300	-	-	-	-
RF	-	None	100	5	1	-	-
LGBM	0.1	3	300	-	-	-	-
KNN	-	-	-	-	-	5	distance

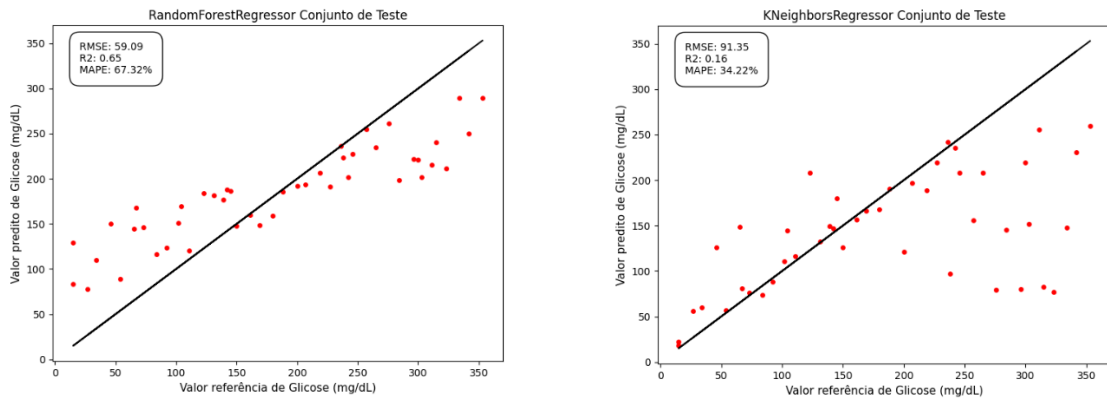
Fonte: Produção do próprio autor (2023).

7.6 TESTE DO MODELO

No contexto de avaliação da capacidade de generalização dos modelos, procede-se a um teste utilizando o conjunto de teste. O comportamento das previsões de glicose na amostra em relação aos valores de referência de glicose na amostra para o conjunto de teste é ilustrado na Figura 35. Observa-se que o RMSE obtido no conjunto de teste apresenta semelhança com o RMSE do modelo durante a fase de criação. É possível observar que o modelo que melhor generalizou os dados de treinamento, foi o modelo LGBM, sendo o modelo que obteve uma menor variação do RMSE, em comparação com os dados de validação.

Figura 35: Predição do conjunto de teste.



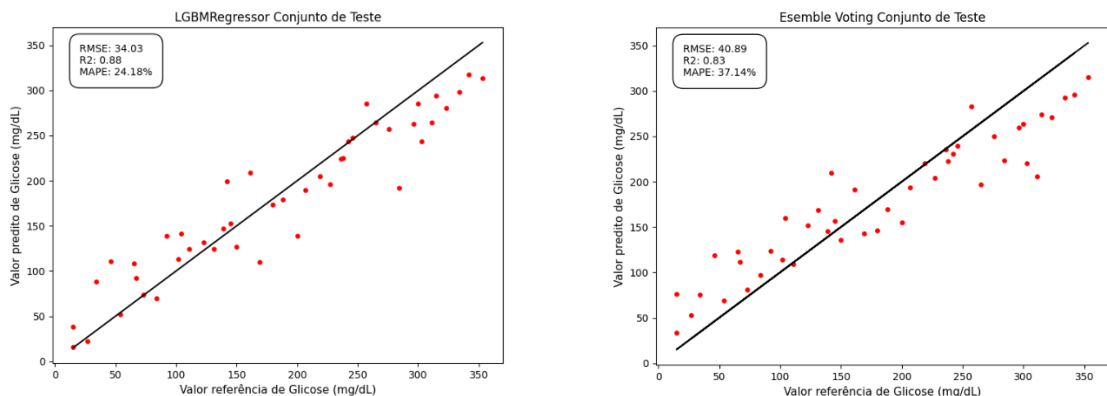


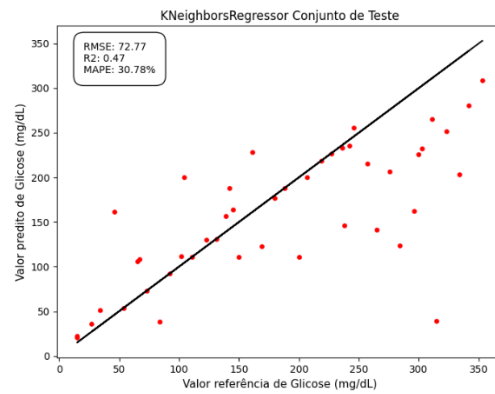
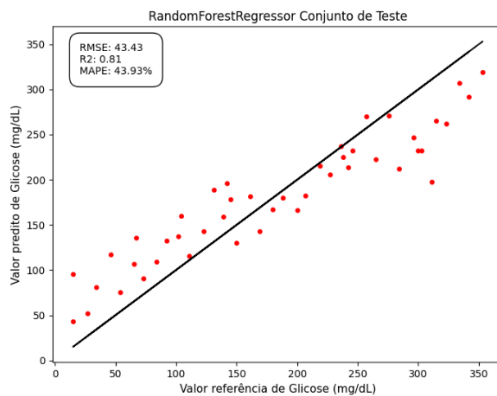
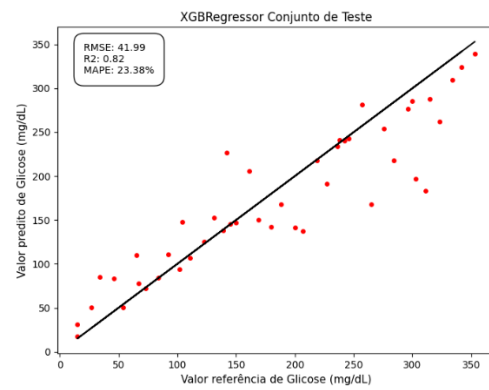
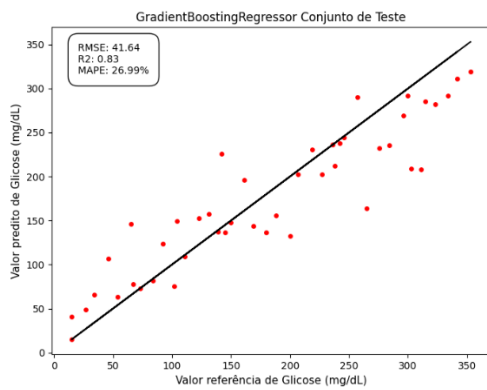
Fonte: Produção do próprio autor (2023).

LGBM RMSE na fase de validação foi de 41.74, na fase de teste foi de 47.89. Voting RMSE na fase de validação foi de 49.20, na fase de teste foi de 53.35. GB RMSE na fase de validação foi de 50.30, na fase de teste foi de 53.68. XGB RMSE na fase de validação foi de 49.20, na fase de teste foi de 53.93. RF RMSE na fase de validação foi de 54.97, na fase de teste foi de 59.09. KNN RMSE na fase de validação foi de 71.49, na fase de teste foi de 91.35. Valores para ambas as fases foram semelhantes, com ênfase para o modelo BG, o qual possui menor variação entre os dados de validação e os dados de teste.

Na figura 36 observa-se a diferença entre o RMSE da fase de criação e o RMSE da fase de teste, foi maior que em comparação com os todos os atributos. É possível observar que o modelo que melhor generalizou os dados de treinamento, foi o modelo LGBM, sendo o modelo que obteve uma menor variação do RMSE, em comparação com os dados de validação.

Figura 36: Predição do conjunto de teste com 22 atributos.





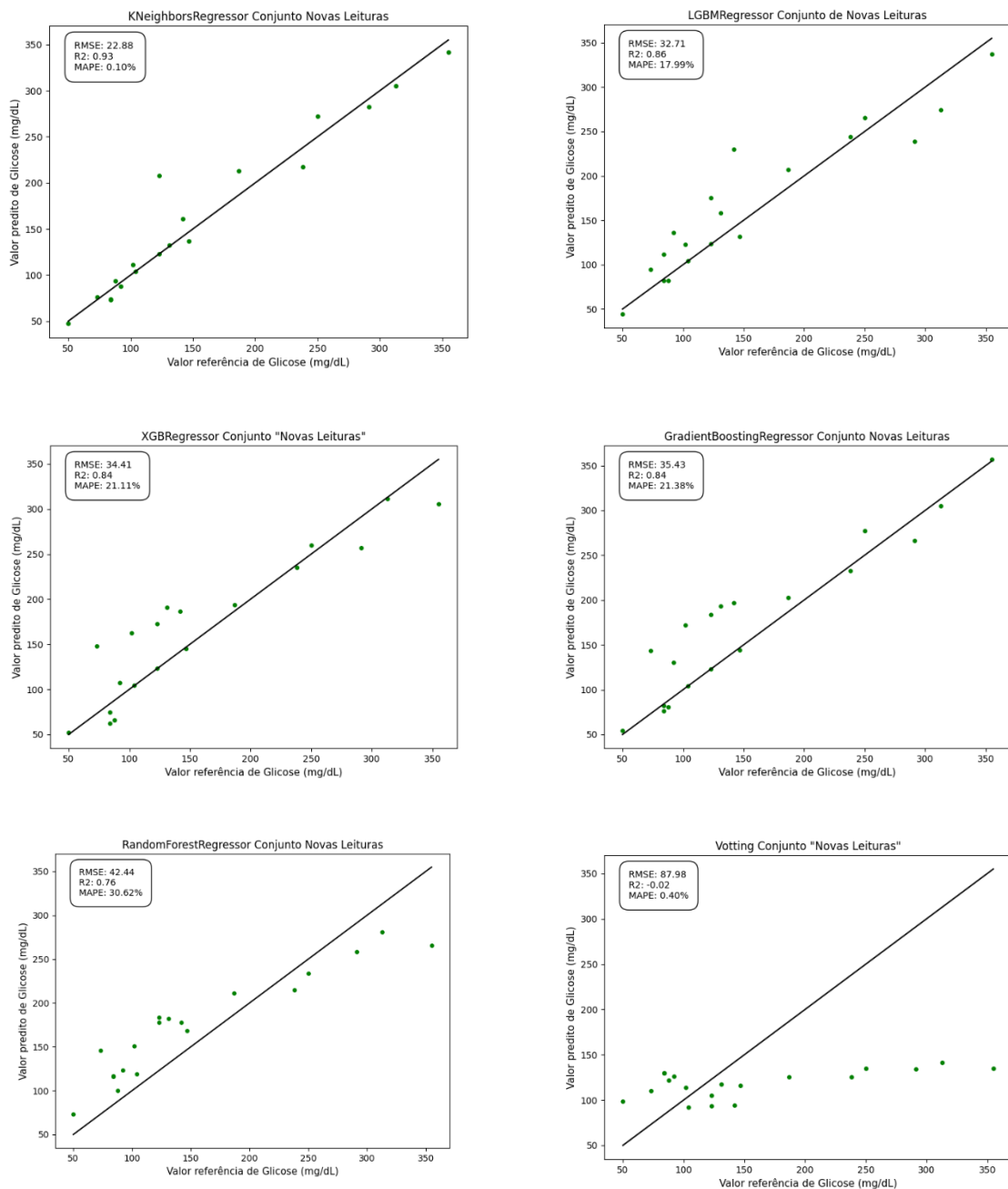
Fonte: Produção do próprio autor (2023).

LGBM RMSE na fase de validação foi de 43.84, na fase de teste foi de 34.03. Voting RMSE na fase de validação foi de 38.96, na fase de teste foi de 40.89. GB RMSE na fase de validação foi de 38.16, na fase de teste foi de 41.64. XGB RMSE na fase de validação foi de 40.11, na fase de teste foi de 41.99. RF RMSE na fase de validação foi de 43.62, na fase de teste foi de 43.43. KNN RMSE na fase de validação foi de 56.88 na fase de teste foi de 72.77.

7.7 NOVAS LEITURAS

Nesta etapa final, foram realizadas 20 leituras de amostras *in vitro*, com o intuito de apenas validar a metodologia de preparação das amostras de glicose, e comprovar a capacidade de generalização dos modelos propostos neste estudo, A Figura 37, representa as previsões do conjunto de novas leituras para a abordagem com todos os atributos, em ordem crescente da métrica de RMSE.

Figura 37: Predição do conjunto de Novas Leituras.

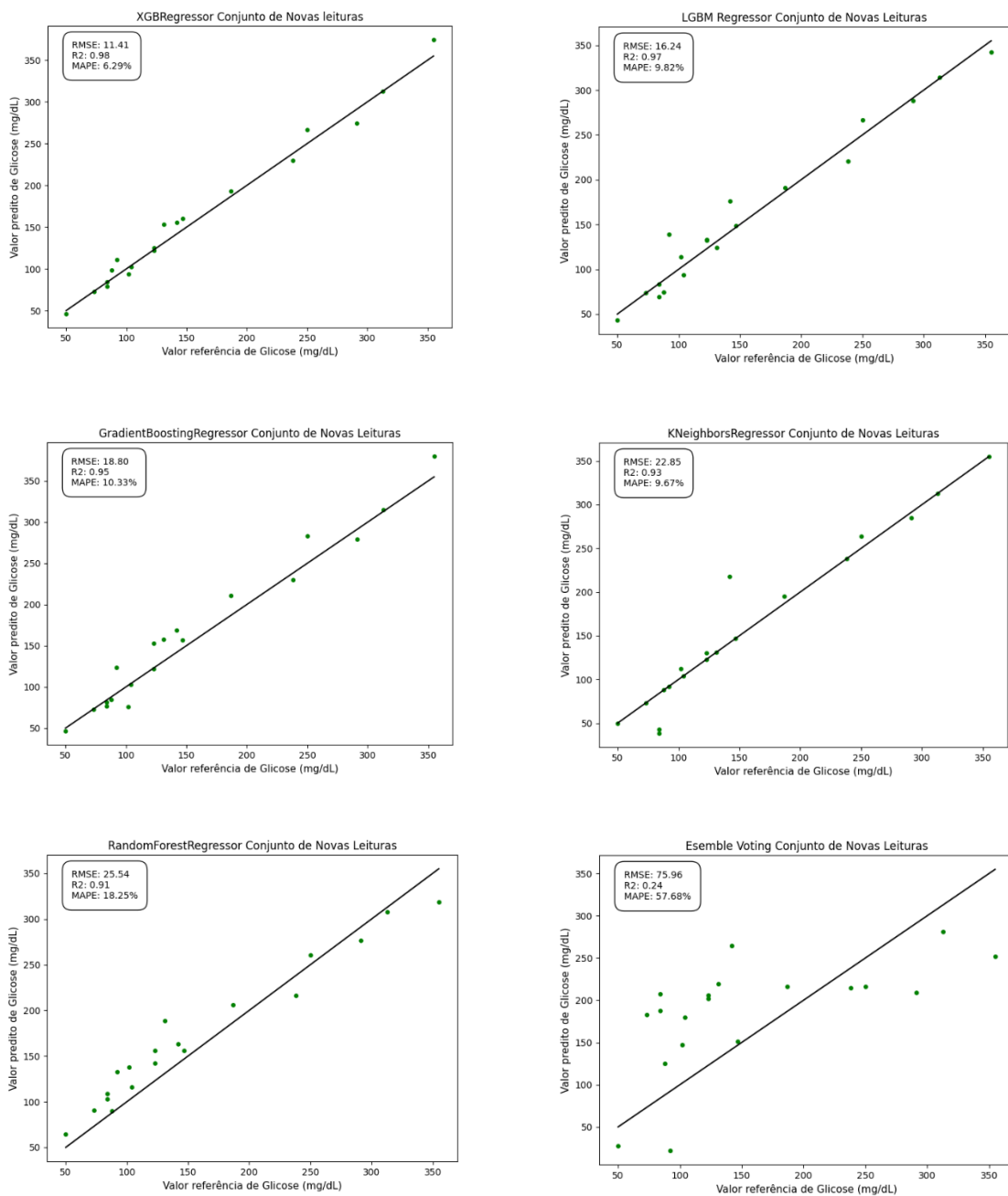


Fonte: Produção do próprio autor (2023).

O modelo KNN foi o qual obteve melhor generalização para a abordagem com todos os atributos, obtendo um RMSE de 22.88, um R2 de 0.93 e um MAPE de 10%, seguido pelo modelo LGBM o qual obteve um RMSE de 32.71, R2 de 0.86 e MAPE de 17.99%. O modelo com pior generalização para esta abordagem foi o modelo de Voting o qual obteve um RMSE de 87.98 R2 de -0.02 e MAPE de 40%.

A Figura 38, temos as previsões do conjunto de Novas leituras para a abordagem de apenas 22 atributos, em ordem crescente da métrica de RMSE.

Figura 38: Predição do conjunto de Novas Leituras abordagem de 22 atributos.



Fonte: Produção do próprio autor (2023).

Na Figura 38 temos as predições para o conjunto de novas leituras, e utilizando modelos treinados na abordagem de 22 atributos, nos quais foram obtidas as seguintes métricas, para o modelo XGB foi obtido um RMSE de 11.41, R2 de 0.98 e MAPE de 6.29%, para o modelo LGBM foi obtido um RMSE de 16.24, R2 de 0.97 e MAPE de 9.82, o modelo com pior generalização foi o modelo Voting obtendo um RMSE de 75.96, R2 de 0.24 e MAPE de 57.68%.

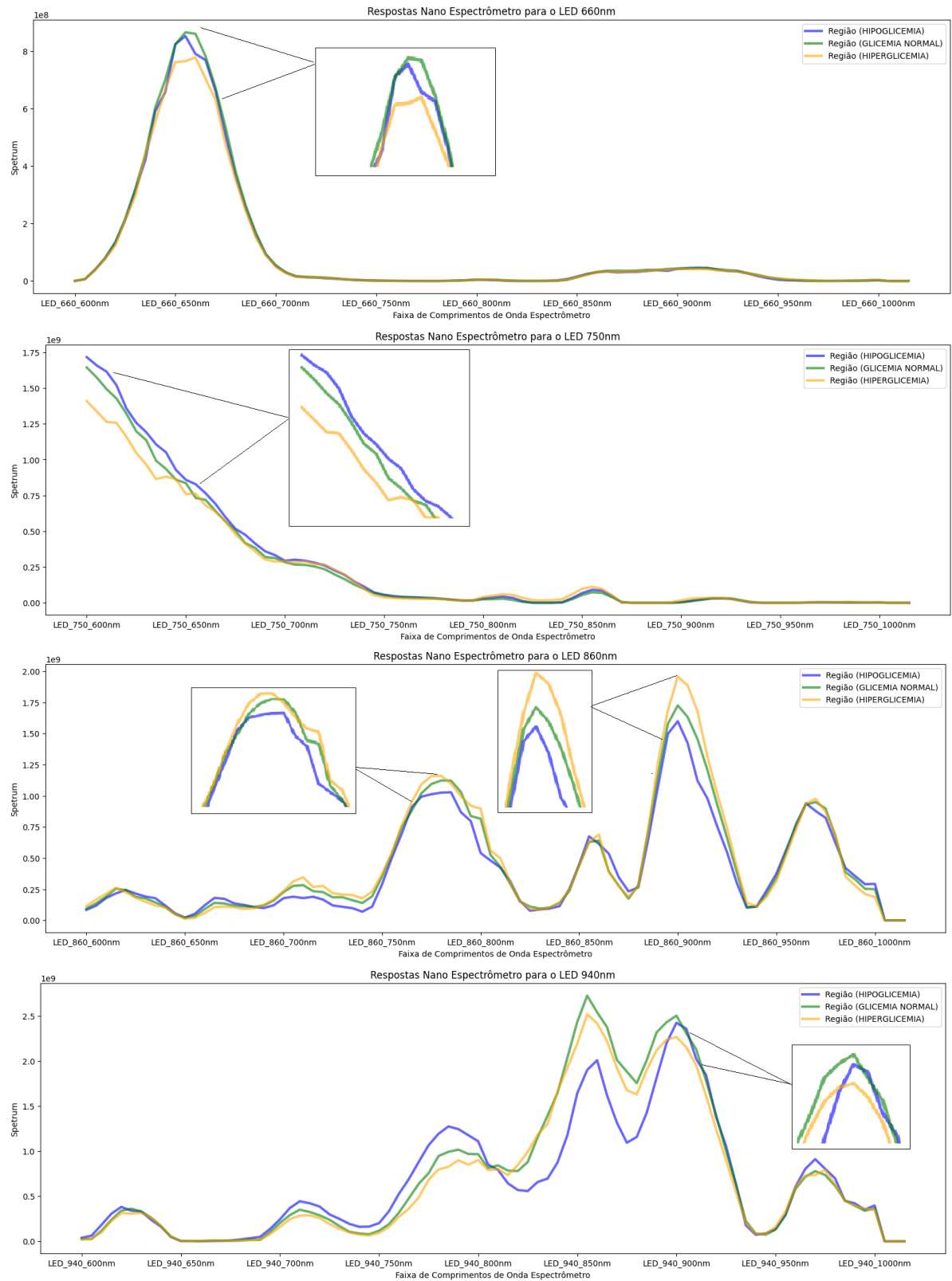
7.8 GRADE DE ERRO DE CLARKE E ANÁLISE ESTATÍSTICA DO DESEMPENHO

Apesar da validação, teste e aplicações em novas medidas já realizados nos modelos, ainda é possível realizar testes adicionais para determinar seu desempenho de forma conclusiva. As análises suplementares propostas incluem a utilização da grade de erro de Clarke e a análise estatística de desempenho. Essas análises finais são conduzidas utilizando os resultados das previsões obtidas por meio das previsões de teste e novas medidas, a fim de obter um conjunto de dados de resultados mais abrangente.

7.8.1 Assinatura Óptica

Nesta seção temos a distribuição da resposta do espectro do nano espectrômetro para cada LED utilizado no estudo, segmentados em 3 regiões, hipoglicemia, normoglicemia e hiperglicemia. É possível perceber que temos picos evidentes na região de 650 e 940 nm, regiões que segundo a literatura são regiões de resposta da glicose, como podemos ver na Figura 39.

Figura 39: Resposta espectral para LED 640, 750, 860 e 940 nm, para 3 regiões de glicemia.



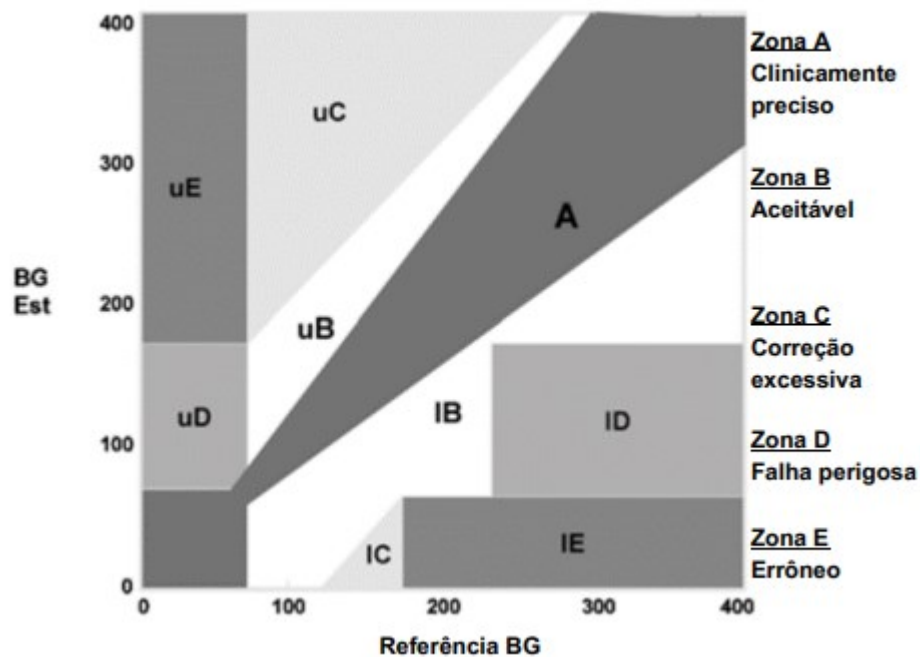
Fonte: Produção do próprio autor (2023).

7.8.2 Grade de Erro de Clarke

O Error Grid Analysis (EGA) foi desenvolvido na década de 1970 para quantificar a precisão clínica das estimativas do paciente de sua glicemia atual em comparação com o valor de glicemia obtido em seu medidor

Tendo como referência o analisador de Beckman, a grade é subdividida em cinco zonas: A, B, C, D e E (Figura 40). Os valores nas Zonas A e B representam resultados de glicose precisos ou aceitáveis. Valores na Zona C podem levar a correções desnecessárias que podem levar a um mau resultado. Os valores na Zona D representam uma falha perigosa de detectar e tratar. Os valores na Zona E representam “tratamento errôneo”. Resumindo, quanto mais valores aparecerem nas Zonas A e B, mais preciso é o dispositivo em termos de utilidade clínica.

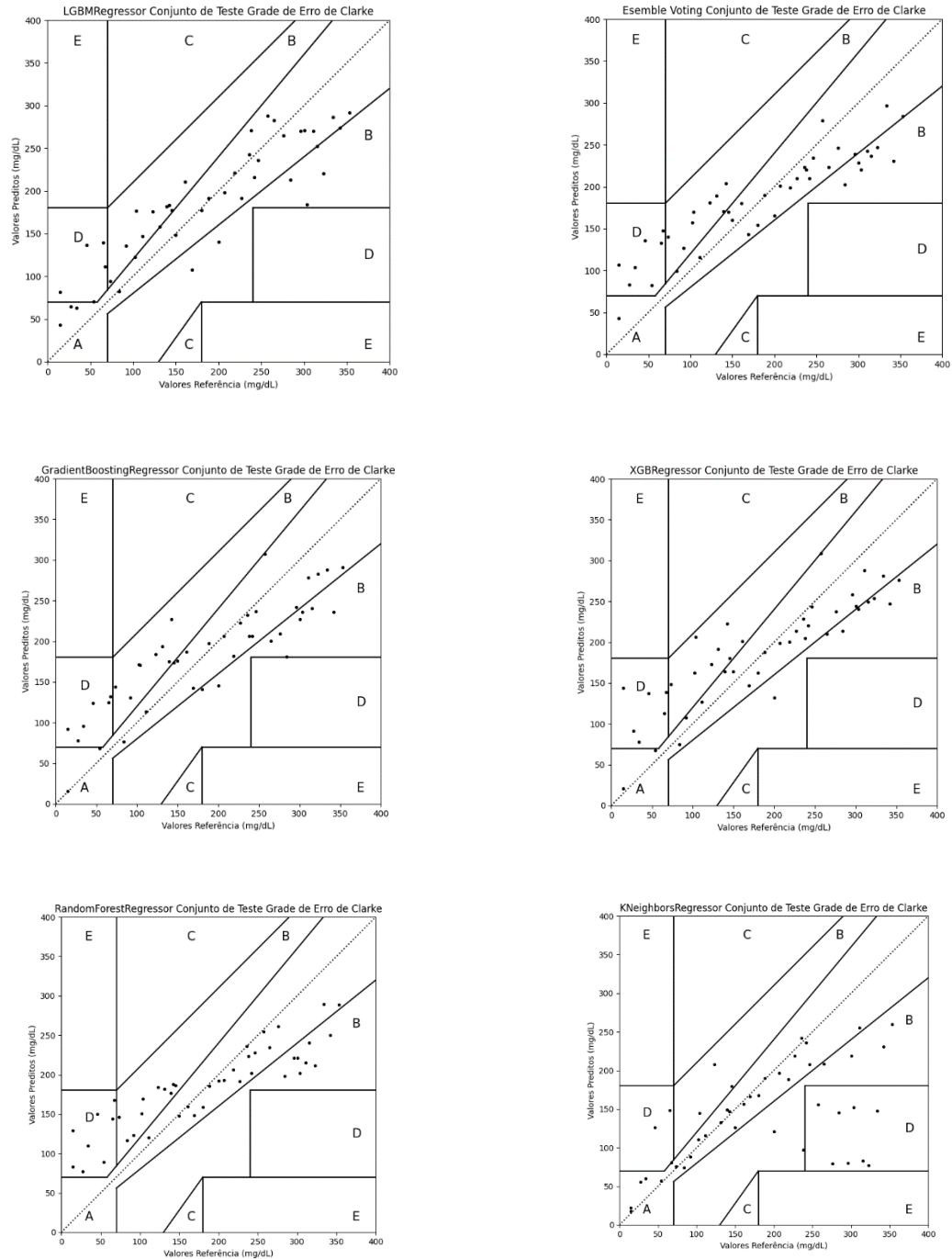
Figura 40: Grade de Erro de Clarke.



Fonte: Modificada de (CLARKE, 2005).

A Figura 41, mostra as predições da do conjunto de teste, frente os valores reais de glicose, expostos na grade de erro de Clarke. A primeira figura, é o melhor caso, até a última sendo o pior caso. O modelo com melhor generalização foi o modelo LGBM, com RMSE de 47.89 e R2 de 0.77, é possível visualizar que maioria dos pontos então na região A+B, com cerca de 89%.

Figura 41: Predições conjunto de teste, aplicados na Grade de erro de Clarke.



Fonte: Produção do próprio autor (2023).

Na tabela 10 temos a distribuição nas previsões em cada zona da grade de erro de Clarke, para cada modelo correspondente, no total são apresentadas 45 previsões e suas respectivas porcentagens estão expostas abaixo.

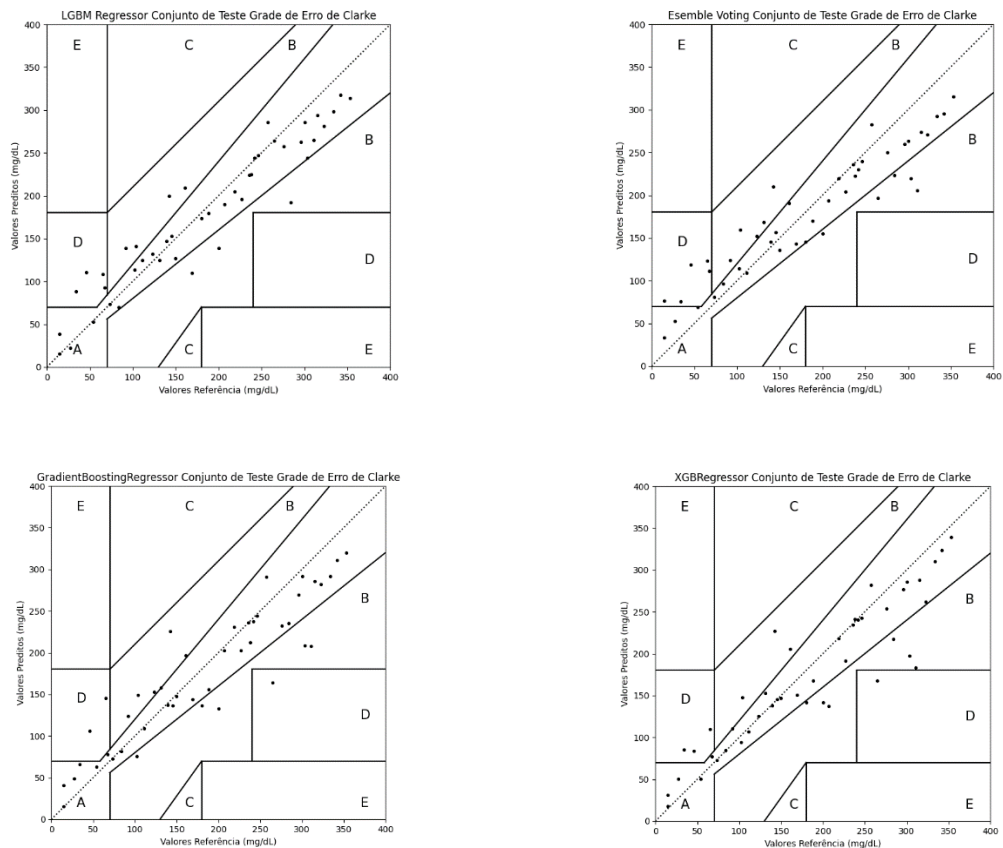
Tabela 10: Distribuição das previsões na grade de erro de Clarke.

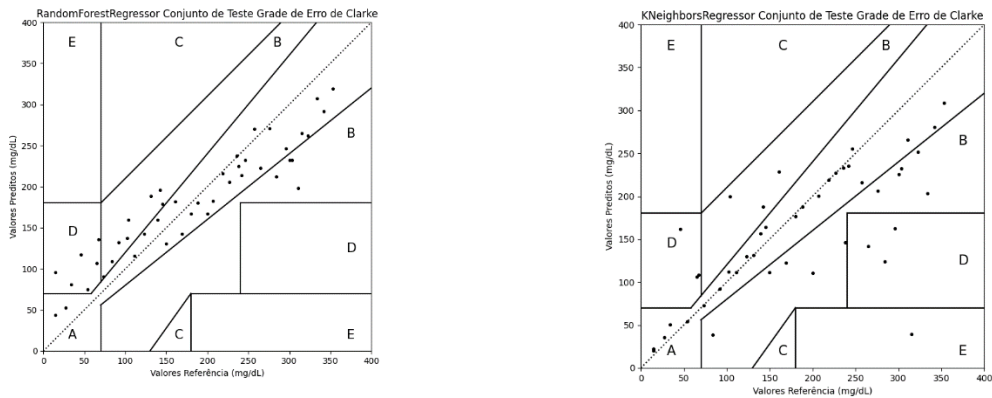
Modelo	Zona A	Zona B	Zona C	Zona D	Zona E	Zona A+B
LGBM	51%	38%	0%	11%	0%	89%
Voting	51%	33%	0%	16%	0%	84%
GB	49%	38%	0%	13%	0%	87%
XGB	49%	38%	0%	13%	0%	87%
RF	42%	40%	0%	18%	0%	82%
KNN	56%	20%	0%	24%	0%	76%

Fonte: Produção do próprio autor (2023).

É possível visualizar na Figura 42 que o modelo com menor RMSE no conjunto de teste, para apenas 22 atributos, foi o modelo LGBM, e que o mesmo apresentou cerca de 92% das previsões dentro da Região A+B.

Figura 42: Conjunto de Teste com 22 Atributos - Representação nas Grades de Erro de Clarke.





Fonte: Produção do próprio autor (2023).

Na Tabela 11 temos a distribuição nas previsões de apenas 22 atributos em cada zona da grade de erro de Clarke, para cada modelo correspondente, no total são apresentadas 45 previsões e suas respectivas porcentagens estão expostas abaixo.

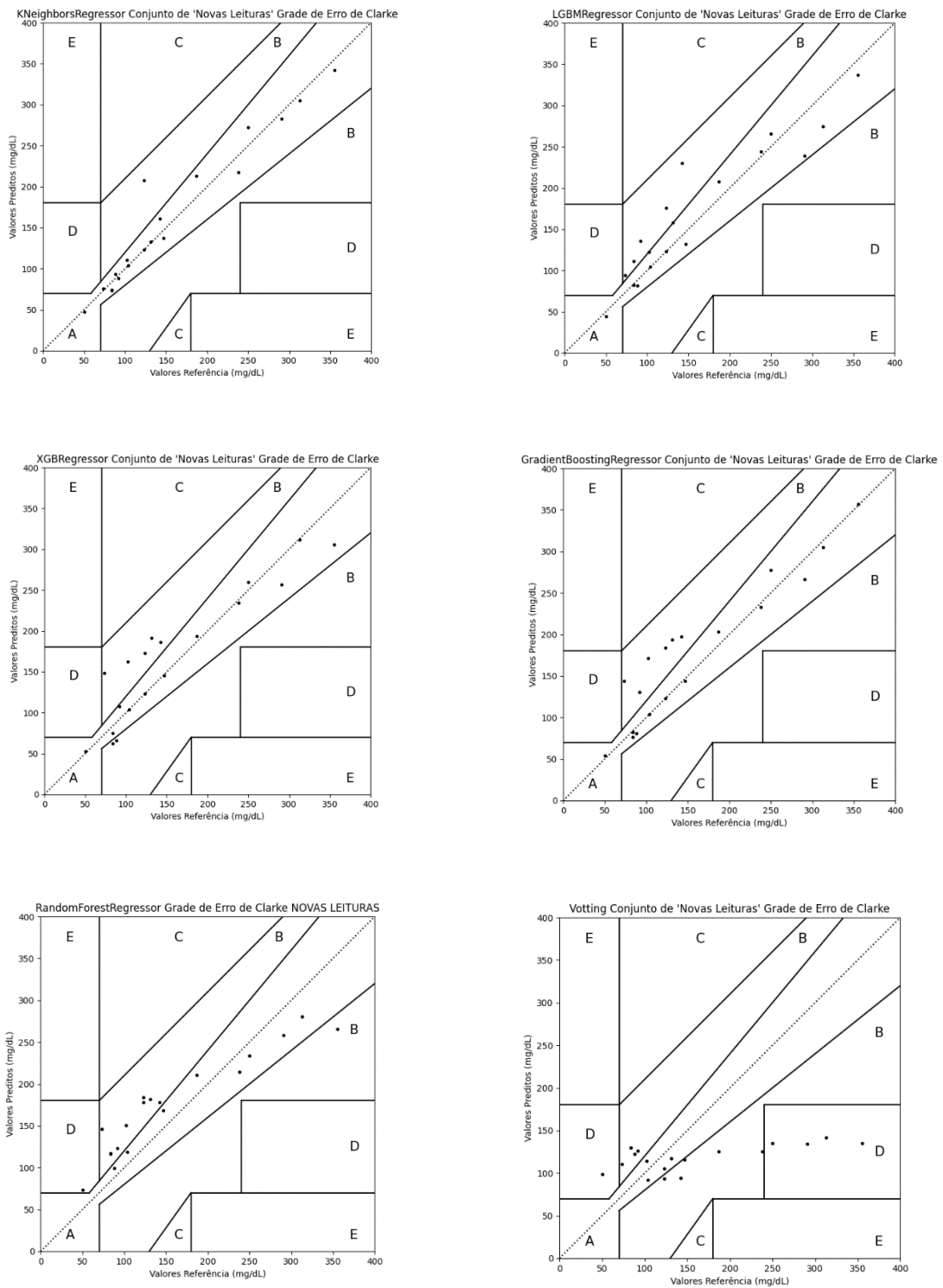
Tabela 11: Distribuição das previsões para 22 atributos na grade de erro de Clarke.

Modelo	Zona A	Zona B	Zona C	Zona D	Zona E	Zona A+B
LGBM	76%	16%	0%	8%	0%	92%
Voting	67%	22%	0%	11%	0%	89%
GB	69%	24%	0%	7%	0%	93%
XGB	69%	22%	0%	9%	0%	91%
RF	60%	27%	0%	13%	0%	87%
KNN	55%	29%	0%	13%	3%	84%

Fonte: Produção do próprio autor (2023).

Na Figura 43 temos a representação das previsões do conjunto de Novas Leituras na grande de erro de Clarke, ordenadas pelo menor RMSE obtido pelo modelo para a abordagem com todas os atributos, em seguida é exposto a tabela 12 onde temos a distribuição das previsões em cada zona.

Figura 43: Conjunto de Novas Leituras com todos os atributos: Representação nas Grades de Erro de Clarke.



Fonte: Produção do próprio autor (2023).

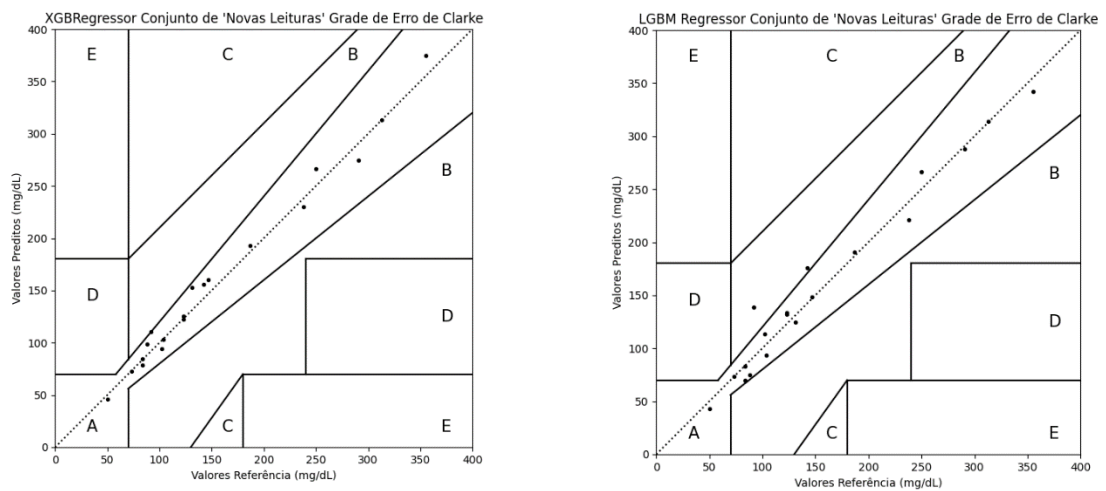
Tabela 12: Distribuição das predições do conjunto de Novas Leituras para todos os atributos na grade de erro de Clarke.

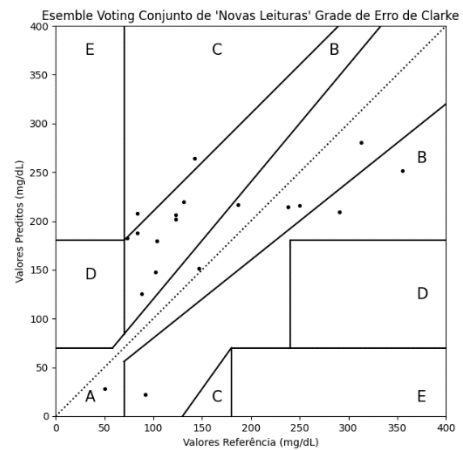
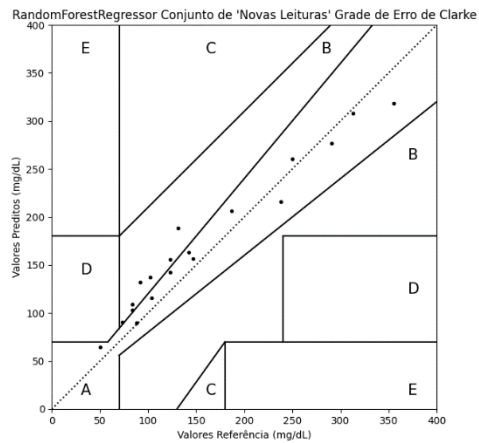
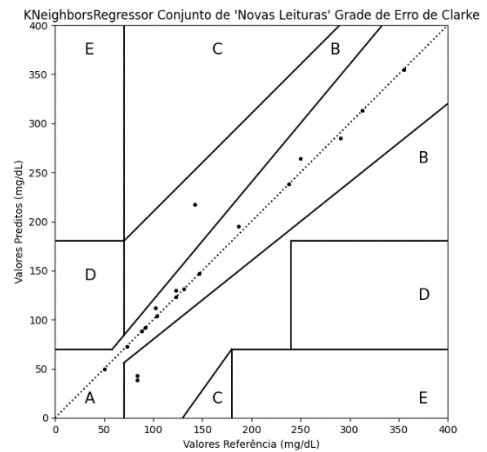
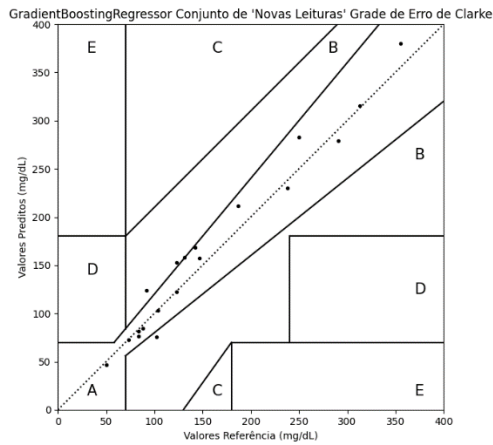
Modelo	Zona A	Zona B	Zona C	Zona D	Zona E	Zona A+B
KNN	94%	6%	0%	0%	0%	100%
LGBM	63%	37%	0%	0%	0%	100%
XGB	63%	37%	0%	0%	0%	100%
GB	68%	32%	0%	0%	0%	100%
RF	42%	53%	0%	5%	0%	95%
Voting	21%	53%	0%	26%	3%	74%

Fonte: Produção do próprio autor (2023).

Na Figura 44 temos a representação das predições do conjunto de Novas Leituras na grade de erro de Clarke, ordenadas pelo menor RMSE obtido pelo modelo para a abordagem com 22 atributos, em seguida é exposto a tabela 13 onde temos a distribuição das predições em cada zona.

Figura 44: Conjunto de Novas Leituras com 22 atributos: Representação nas Grades de Erro de Clarke.





Fonte: Produção do próprio autor (2023).

Tabela 13: Distribuição das predições do conjunto de Novas Leituras para 22 atributos na grade de erro de Clarke.

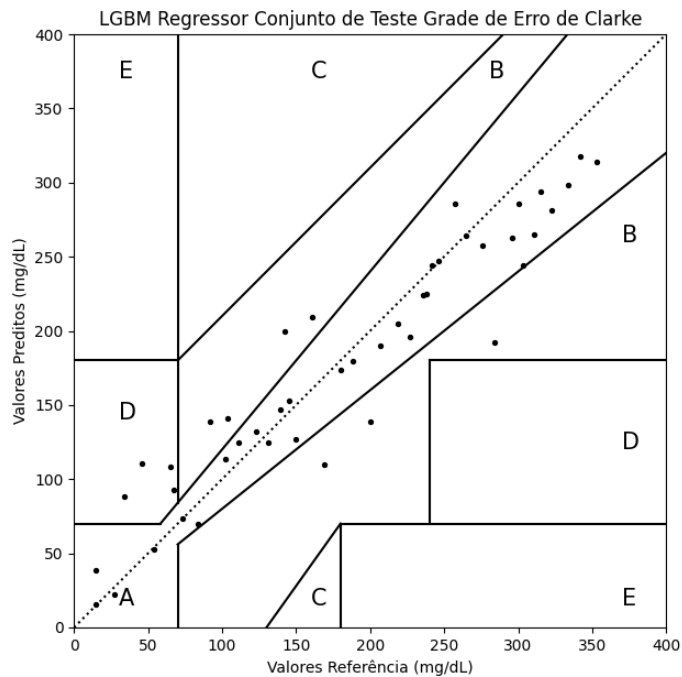
Modelo	Zona A	Zona B	Zona C	Zona D	Zona E	Zona A+B
XGB	94%	6%	0%	0%	0%	100%
LGBM	89%	11%	0%	0%	0%	100%
GB	79%	21%	0%	0%	0%	100%
KNN	84%	26%	0%	0%	0%	100%
RF	63%	37%	0%	0%	0%	100%
Voting	31%	58%	0%	11%	0%	89%

Fonte: Produção do próprio autor (2023).

7.8.3 Análise Estatística do Desempenho e Distribuição dos Erros

Na figura 45, temos a grade de erro de Clarke para o modelo de regressão com melhor desempenho no conjunto de teste (LGBM) com 22 atributos, é possível observar que 76% das predições estão na zona A, 16% na zona B e 8% na zona D.

Figura 45: Grade de Erro de Clarke Modelo de Regressão LGBM abordagem com 22 atributos.



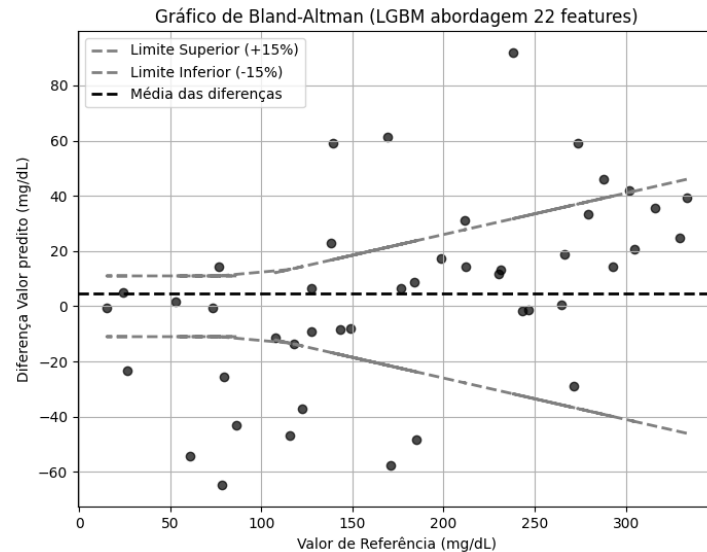
Fonte: Produção do próprio autor (2023).

Bland e Altman introduziram o gráfico de Bland-Altman (B&A) para descrever a concordância entre duas medições quantitativas (DAVIDE, 2015). Idealmente, os pontos no gráfico de Bland-Altman devem estar espalhados de forma homogênea em torno de uma linha horizontal que representa a diferença média entre os métodos. É possível identificar viés, imprecisão ou padrões sistemáticos pela disposição dos pontos no gráfico.

Esse gráfico também costuma incluir linhas de referência para mostrar limites de concordância aceitáveis, o que facilita a interpretação dos resultados. O gráfico de (B&A) é uma ferramenta valiosa para avaliar a concordância entre métodos de medição e pode ser útil em várias áreas, incluindo medicina, ciências ambientais, bioquímica e muitas outras disciplinas que envolvam medição e comparação de dados.

A Figura 41 apresenta o gráfico de Bland-Altman, que ilustra os dados previstos do conjunto de teste para o modelo LGBM, o qual obteve a métrica RMSE mais favorável. Observa-se no gráfico que uma parcela significativa dos dados possui diferenças de predição superiores aos limites superiores e inferiores.

Figura 46: Gráfico de B&A para o Modelo de Regressão LGBM no Conjunto de Teste.



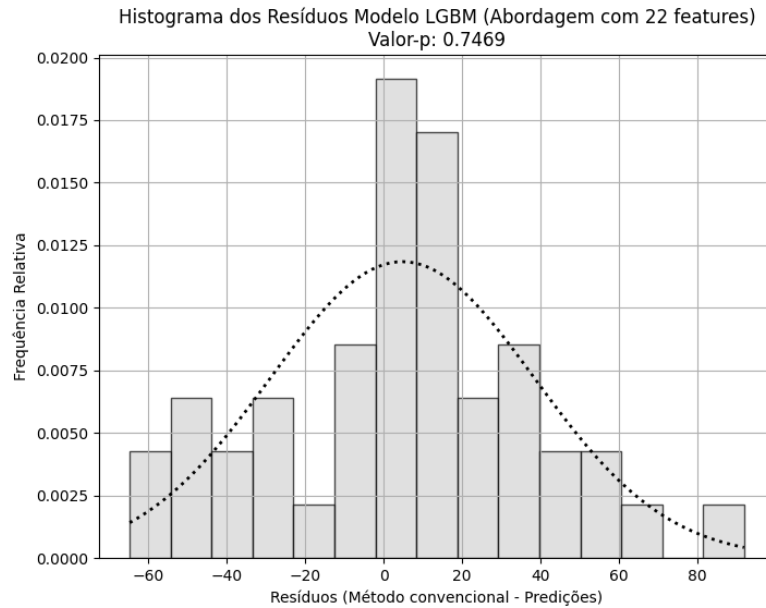
Fonte: Produção do próprio autor (2023).

A distribuição normal das diferenças deve sempre ser verificada, por exemplo, desenhando um histograma. Se isso for assimétrico ou tiver caudas muito longas, a suposição de normalidade pode não ser válida (DAVIDE, 2015).

O teste de Shapiro-Wilk (valor-p) é um teste estatístico usado para verificar se uma amostra de dados segue uma distribuição normal. A hipótese nula (H_0) do teste de Shapiro-Wilk é que os dados seguem uma distribuição normal. A hipótese alternativa (H_1) é que os dados não seguem uma distribuição normal.

Se o valor-p for maior que o nível de significância escolhido (geralmente 0,05), não temos evidências suficientes para rejeitar a hipótese nula e concluímos que os dados podem ser considerados como tendo uma distribuição normal.

Figura 47: Histograma dos Resíduos das Predições do Modelo XGB no Conjunto de Teste.



Fonte: Produção do próprio autor (2023).

No caso particular em que o valor-p é calculado como 0.7469 (Figura 47), constata-se que este valor é superior a 0.05 (nível de significância comum). Em virtude dessa constatação, não há fundamentos para rejeitar a hipótese nula, permitindo-nos concluir que os dados podem ser considerados como possuindo uma distribuição normal.

7.9 MODELO DE CLASSIFICAÇÃO

Nesta seção será exposto uma proposta de abordagem para predição de glicose em amostra *in vitro*. Afim de fundamentar a proposta da utilização de modelos de classificação para a predição da concentração de glicose, foi utilizado como base nos estudos de (ALORAYNAN *et al.*, 2022; HABIBULLAH *et al.*, 2019; SHOKREKHODAEI *et al.*, 2021).

Os estudos realizados por (ALORAYNAN *et al.*, 2022; HABIBULLAH *et al.*, 2019; SHOKREKHODAEI *et al.*, 2021) empregaram dados contínuos para categorizar classes com valores discretos, utilizando principalmente modelos de classificação baseados em árvores de decisão. Os resultados obtidos foram altamente significativos para a literatura, demonstrando uma elevada acurácia e precisão, conforme apresentado na Tabela 14.

Tabela 14: Estudos que utilizaram modelos de classificação para predição do nível de glicose em meio *in vitro*.

Artigo	Acurácia	F1-SCORE
SHOKREKHODAEI <i>et al.</i> , 2021	86%	99%
ALORAYNAN <i>et al.</i> , 2022	90.5%	94.5%
HABIBULLAH <i>et al.</i> , 2019	82%	76%

Fonte: Produção do próprio autor (2023).

Modelos de classificação são utilizados para a predição de classes, sendo problemas de apenas duas classes ou problemas com multiclasse. Desta forma se torna necessário a criação de classes, devido à baixa diversidade dos dados deste estudo, optou-se pela criação de classes em um intervalo de 15 mg/dL, sendo assim todos as leituras dos dados de treinamento que continham leituras entre 0 mg/dL e 15 mg/dL, foram inseridas em uma única classe, seguindo essa sequência até 360 mg/dL, As 5 primeiras classes (1-5) contém pontos com concentração de glicose inferiores a 70 mg/dL (hipoglicemia), as próximas 7 classes incluem as concentrações de glicose entre os valores de 70 mg/dL e 180 mg/dL (normoglicemia), e as últimas 10 classes contém valores acima de 180 mg/dL (hiperglicemia) . O valor da classe foi definido como a média entre o intervalo inferior e o intervalo superior da classe, resultando em 23 classes, um resumo das classes está exposto da Tabela 15.

Tabela 15: Definição das 23 Classes contínuas para Modelo de Classificação.

Intervalo de glicose original	Classe Definida (valor contínuo)
15 > valor >= 0	7
30 > valor >= 15	22
.	.
.	.
.	.
345 > valor >= 330	337
360 > valor >= 345	352

Fonte: Produção do próprio autor (2023).

A dimensão deste intervalo foi definida levando em consideração dois fatores, quantidade de dados, o intervalo possibilita abranger uma quantidade mínima de dados, e levando em consideração a ISO 15197:2013 a qual considera um erro aceitável de 15 mg/dL para dispositivos comerciais.

O modelo de classificação seguiu os mesmos processos dos modelos de regressão descritos no item 7.3, utilizando dados de treinamento, validação e teste, além de busca em grade, para a definição dos melhores hiperparâmetros.

A abordagem utilizada que difere dos modelos de regressão foi a técnica SMOTE, a qual tem o objetivo de balancear as classes de dados, adicionando dados sintéticos ao conjunto de treinamento. Este processo não é aplicado nos dados de validação, assim como não foi aplicado nos dados de teste.

O modelo escolhido para realização a predição, foi Extra tree classifier, pois durante os testes, foi o qual apresentou melhores resultados. O Extra Tree Classifier faz parte da família de algoritmos de "*ensemble learning*", que combina múltiplos modelos de aprendizado de máquina para melhorar o desempenho e a generalização do modelo. Especificamente, o Extra Tree é uma variação do algoritmo de Árvores de Decisão e é considerado como uma das técnicas de "*bagging*" (*bootstrap aggregating*).

Na Tabela 16, encontram-se os hiperparâmetros estabelecidos após a conclusão da busca em grade para o modelo Extra Tree. A busca em grade foi conduzida para ambas as abordagens.

Tabela 16: Hiperparâmetros definidos após busca em grade e validação cruzada com 10-Fold.

Modelo	Max depth	N_estimators	Min samples split	Min samples leaf
Extra Tree	10	200	2	1

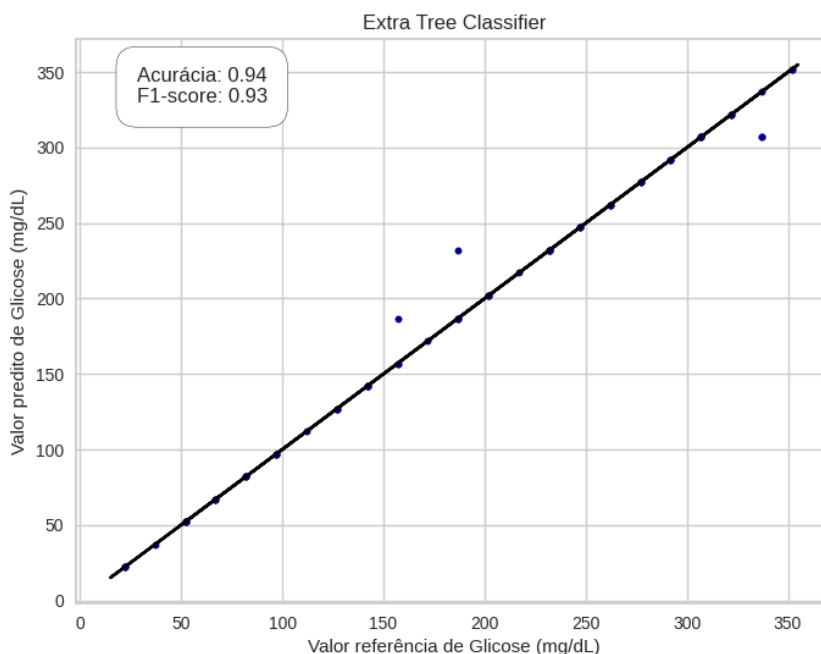
Fonte: Produção do próprio autor (2023).

7.9.1 Conjunto de Treinamento e Validação Classes Contínuas

Nesta seção são expostas as métricas de acurácia, precisão, F1-score e matriz de confusão, para os dados de validação.

Na Figura 48, é exposto as classificações de glicose *in vitro*, o eixo y são as classes previstas e o eixo x são as classes originais. É possível perceber que o valor de acurácia de 0.94 e F1-score de 0.93.

Figura 48: Predição dados de validação.

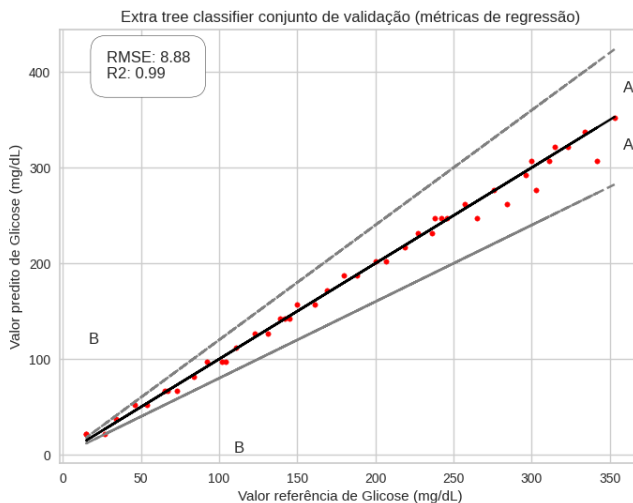


Fonte: Produção do próprio autor (2023).

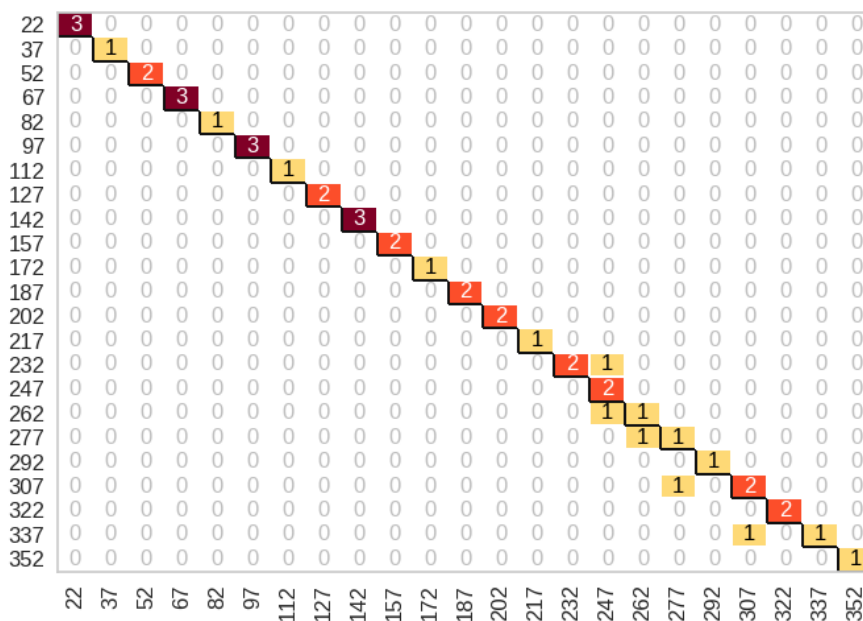
Na Figura 49, temos a matriz de confusão para a predições dos dados de validação, uma matriz de confusão é uma tabela usada para avaliar o desempenho

Figura 50: Predições conjunto de teste. (A) Predições com métricas de RMSE e R2. (B) Matriz de confusão para o conjunto de teste.

A)



B)



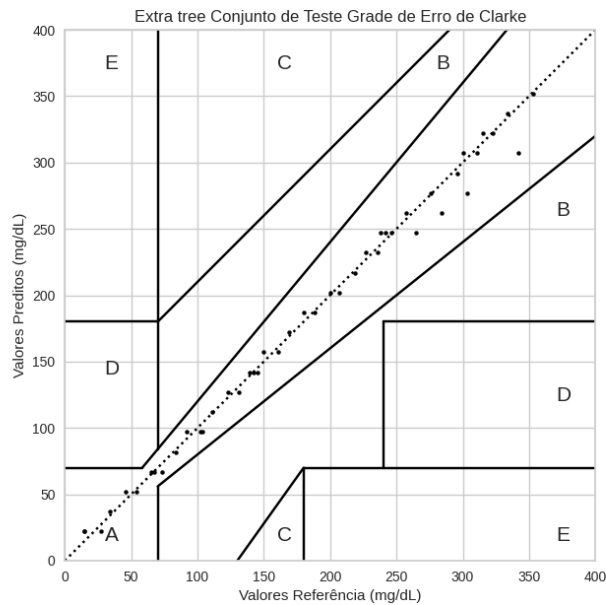
Fonte: Produção do próprio autor (2023).

É possível analisar que as predições de classes incorretas, são em predições em classes vizinhas, o que minimiza o intervalo de erro, por exemplo a classe predita em 245 mg/dL, porém seu valor real seria de 260 mg/dL, conseqüentemente temos um erro de 30 mg/dL

7.9.2.1 *Análise Estatística do Desempenho e Distribuição dos Erros*

Na Figura 51, encontra-se a grade de erro de Clarke, no qual são expostas as predições frente aos dados de referência, para o modelo de classificação utilizando valores contínuos, tivemos 100% das predições na zona A.

Figura 51: Grade de Erro de Clarke (valores contínuos).



Fonte: Produção do próprio autor (2023).

Na Tabela 17 temos a distribuição das predições na grade de erro de Clarke, lembrando que ao total são 45 predições.

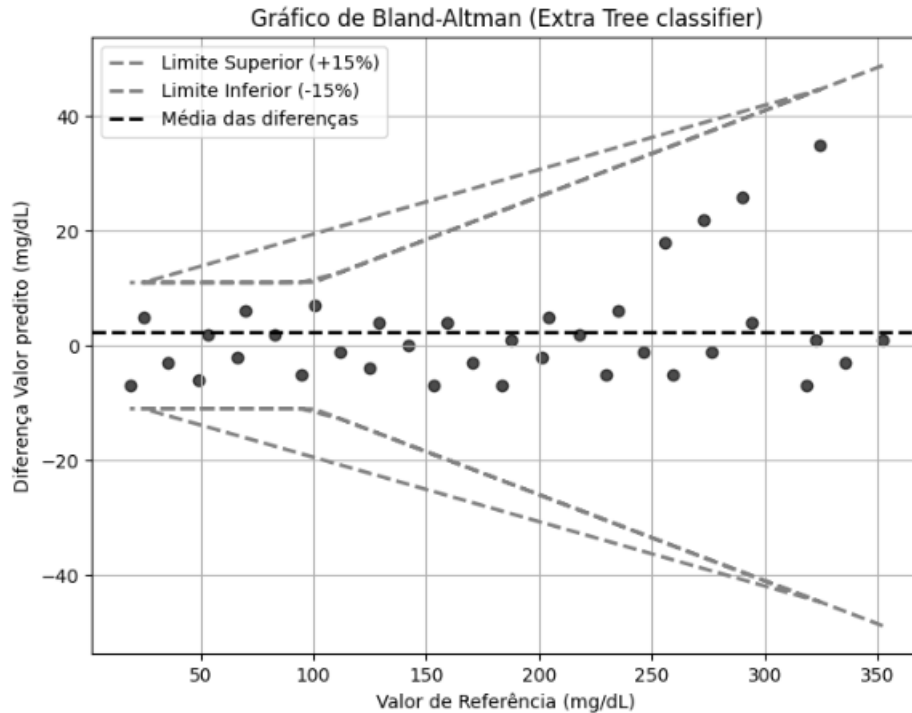
Tabela 17: Distribuição das predições para modelo Extra Tree modelo Extra Tree

Modelo	Zona A	Zona B	Zona C	Zona D	Zona E	Zona A+B
Extra Tree	100%	0%	0%	0%	0%	100%

Fonte: Produção do próprio autor (2023).

Na Figura 52, encontra-se o gráfico de Bland-Altman, no qual são representadas as predições das classes previamente definidas (valores discretos) em relação aos dados de referência (valores contínuos). Observa-se que todos os pontos se encontram dentro dos limites estabelecidos no gráfico.

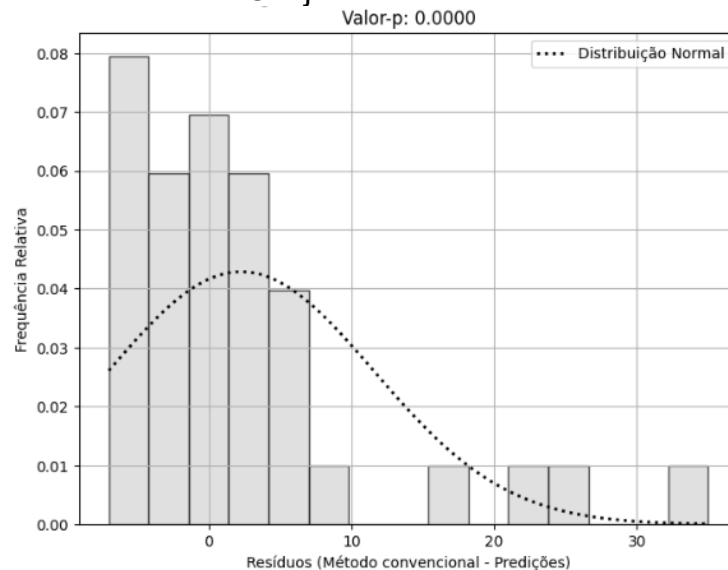
Figura 52: Gráfico de Bland-Altman para o Modelo Extra Tree no Conjunto de Teste.



Fonte: Produção do próprio autor (2023).

No caso específico em que o valor-p é calculado como 0.0125 (Figura 53), verifica-se que esse valor é inferior a 0.05 (nível de significância comum). Com base nessa constatação, temos fundamentos para rejeitar a hipótese nula, o que nos leva a concluir que os dados não podem ser considerados como possuindo uma distribuição normal.

Figura 53: Histograma dos Resíduos das Predições do Modelo Extra Tree no Conjunto de Teste.



Fonte: Produção do próprio autor (2023).

7.9.3 Utilizando 23 Classes (Valores Discretos)

A segunda abordagem segue a seguinte lógica de definição das classes, todas as leituras dos dados de treinamento que continham leituras entre 0 mg/dL e 15 mg/dL, foram inseridas em uma única classe, seguindo essa sequência até 360 mg/dL, As 5 primeiras classes (1-4) contém pontos com concentração de glicose inferiores a 70 mg/dL (hipoglicemia), as próximas 7 classes incluem as concentrações de glicose entre os valores de 70 mg/dL e 180 mg/dL (glicemia normal), e as últimas 10 classes contém valores acima de 180 mg/dL (hiperglicemia) . O valor da classe foi definido como um intervalo e.g 0 – 15 mg/dL, sendo 0 e 15 os limites inferior e superior da classe, como é possível observar na Tabela 18.

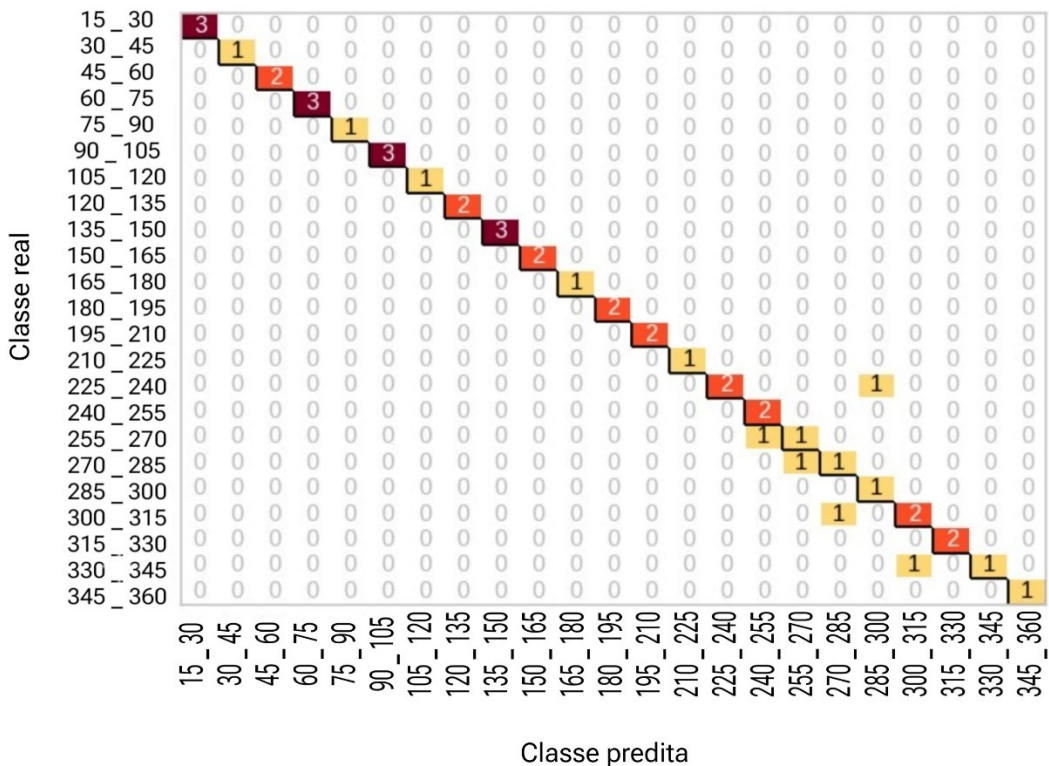
Tabela 18: Definição das 23 Classes discretas para Modelo de Classificação.

Intervalo de glicose original	Classe Definida (valor discreto)
15 > valor >= 0	0 - 15
30 > valor >= 15	15 - 30
.	.
.	.
.	.
345 > valor >= 330	330 - 345
360 > valor >= 345	345 - 360

Fonte: Produção do próprio autor (2023).

Na Figura 54 temos a matriz de confusão para os dados de teste, para as 23 classes discretas para os dados de teste. É possível visualizar que apenas 4 predições foram feitas em classes vizinhas.

Figura 54: Matriz de confusão para o conjunto de teste.



Fonte: Produção do próprio autor (2023).

Na Tabela 19 temos as métricas de acurácia, f1-score e recall, para o modelo de extra tree.

Tabela 19: Métricas modelo de classificação Extra tree valores discretos.

Modelo	Acurácia	F1-score	Recall
Extra Tree	0.89	0.89	0.86

Fonte: Produção do próprio autor (2023).

7.9.4 Utilizando Apenas 3 Classes

Uma terceira abordagem neste estudo, tem como objetivo definir apenas 3 classes, com base nos valores de glicose, dessa forma será obtido como classes alvos (Hiperglicemia, normoglicemia, Hipoglicemia). A definição de cada classe segue na Tabela 20.

Tabela 20: Definição de 3 Classes para modelo de Classificação.

Intervalo de glicose original	Classe Definida
70 >= valor >= 0	Hipoglicemia

71 >= valor >= 180
353 >= valor >= 181

Normoglicemia
Hiperglicemia

Fonte: Produção do próprio autor (2023).

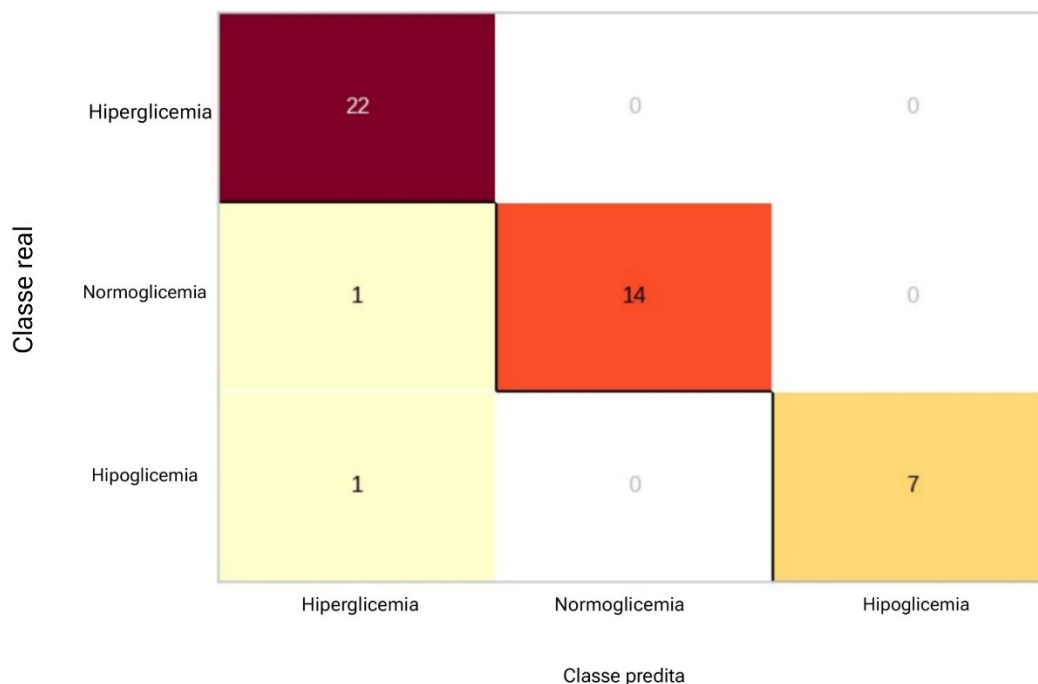
Na Figura 55 temos a matriz de confusão para os dados de teste, para apenas 3 classes definidas como (hipoglicemia, glicemia normal e hiperglicemia). É possível perceber que temos apenas 1 classificação incorreta, aonde a classe verdadeira deveria ser glicemia normal e a classe predita foi hiperglicemia. Na tabela 21 temos as métricas de acurácia, f1-score e recall, para o modelo de Extra tree e utilizando 3 classes.

Tabela 21: Métricas modelo de classificação Extra tree.

Modelo	Acurácia	F1-score	Recall
Extra Tree	0.95	0.95	0.95

Fonte: Produção do próprio autor (2023).

Figura 55: Matriz de confusão para 3 classes para o conjunto de teste.



Fonte: Produção do próprio autor (2023).

8 Discussão

O presente estudo teve como objetivo desenvolver uma metodologia para a determinação de um composto específico em uma amostra usando uma técnica espectrofotométrica. Expondo com detalhes os passos a serem seguidos para a obtenção de amostras de glicose para testes *in-vitro*, algo que se mostra faltante na literatura, não sendo possível compreender com clareza, quais foram os passos para a obtenção da amostra, e qual a qualidade do procedimento, como podemos consultar em alguns trabalhos (FOUAD; MAHMOUD; ABD EL GHANY, 2018; GENG *et al.*, 2017; HOTMARTUA *et al.*, 2015; PRIYOTI *et al.*, 2019; SONG, K. *et al.*, 2015b; WANG, Z. *et al.*, 2022; YADAV *et al.*, 2014).

O que acaba sendo um impeditivo na evolução na pesquisa de desenvolvimentos de glicosímetros não invasivos, por meio de NIRS, não sendo possível compreender a técnica e metodologia utilizada na avaliação do dispositivo, os resultados encontrados ao final da pesquisa acabam por serem de difícil reprodução por demais pesquisadores.

Durante o processo de validação do modelo, pode observar-se a importância dos atributos, e o atributo que mais afetou o modelo foi o atributo de temperatura, como é possível ver na Figura 33. O que demonstra a importância do controle ambiente e descrição das condições dos testes *in-vitro*.

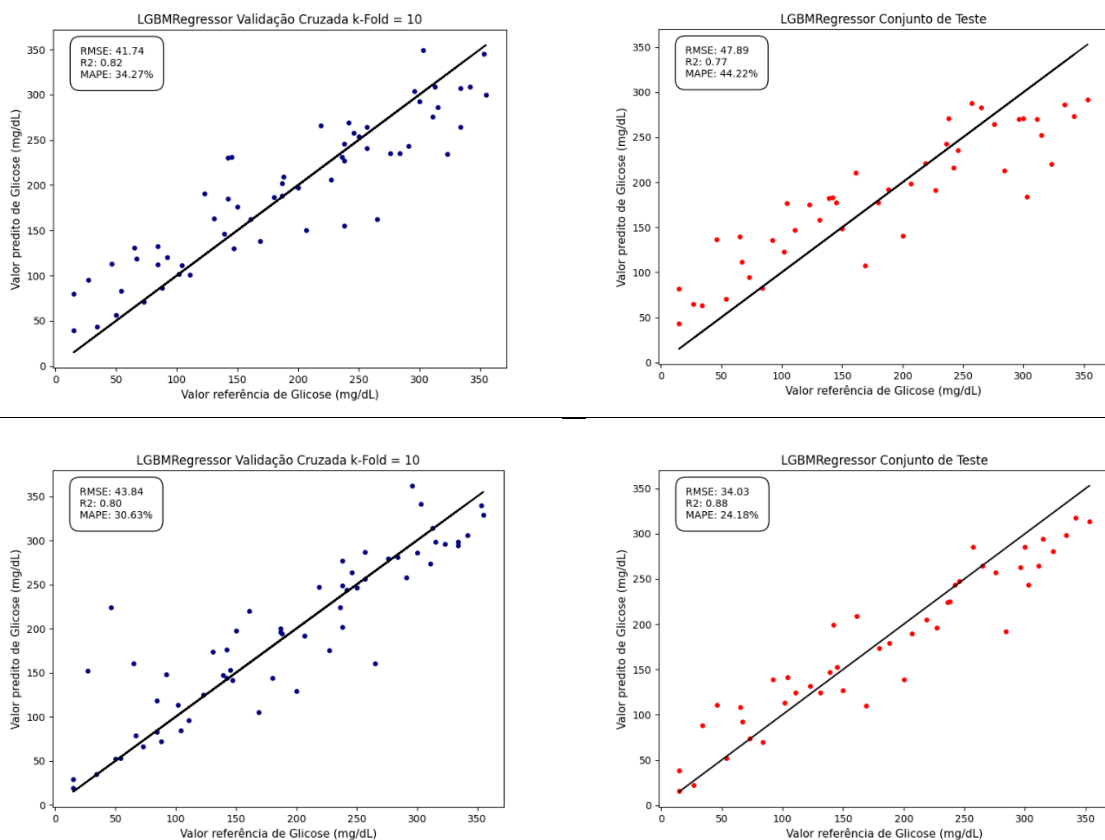
Nos primeiros testes em laboratório foi observado que as concentrações de glicose, possuíam alta variação, caso a medida fosse feita logo após a diluição da glicose em água Milliq. Partindo disto, optou-se por solubilizar uma solução padrão, com alto índice de concentração de glicose, e realizar a solubilização em vórtex mixer, para maior solubilidade do soluto (glicose), e para comprovar a não variabilidade da concentração de glicose, a solução padrão foi colocada em repouso em temperatura ambiente de 24°C por 1 hora. A concentração de glicose foi medida novamente e observou-se uma pequena variação menor que 10 mg/dL.

Partindo deste ponto, foram produzidas 3 soluções padrões em diferentes períodos do dia e diferentes dias, seguindo a metodologia desenvolvida neste trabalho. Para a realização das leituras com o protótipo proposto neste trabalho, foram realizadas diluições com água milliq Tabela 5, para cada amostra, desta forma é evitado qualquer variação de glicose, caso a amostra permaneça um longo tempo em repouso após a diluição.

No item 7.5 e 7.6, são expostas as métricas das previsões dos modelos propostos, no item 7.5 temos as métricas para o subconjunto de validação, que algum momento foi visto pelo modelo durante o treinamento, no item 8.6 temos as métricas para o conjunto de teste, conjunto o qual não foi visualizado pelo modelo, no item 7.7 temos as métricas para o conjunto de Novas Leituras, o qual foi realizado em um dia em separado com objetivo de apenas testar a metodologia proposta. O modelo com melhor ajuste aos dados de treinamento foi o modelo Voting, o qual é baseado, também foi utilizado a base de dados com apenas 22 atributos, possuindo RMSE de 38.96 após a validação, porém as métricas para o conjunto de testes foram piores, retornando um RMSE de 40.89, o que mostra que ao reduzir a quantidade de atributos, melhorou o ajuste do modelo aos dados de treinamento, mas afetou negativamente a generalização do modelo, afetando a predição de novos valores.

Os dois melhores modelos com melhores generalizações para o conjunto de teste, foram LGBM com todos os atributos e LGBM com apenas 22 atributos. Na figura 56, temos as previsões do subconjunto de validação e para o conjunto de teste, para ambos os modelos.

Figura 56: Melhores modelos ajustados aos dados de teste.



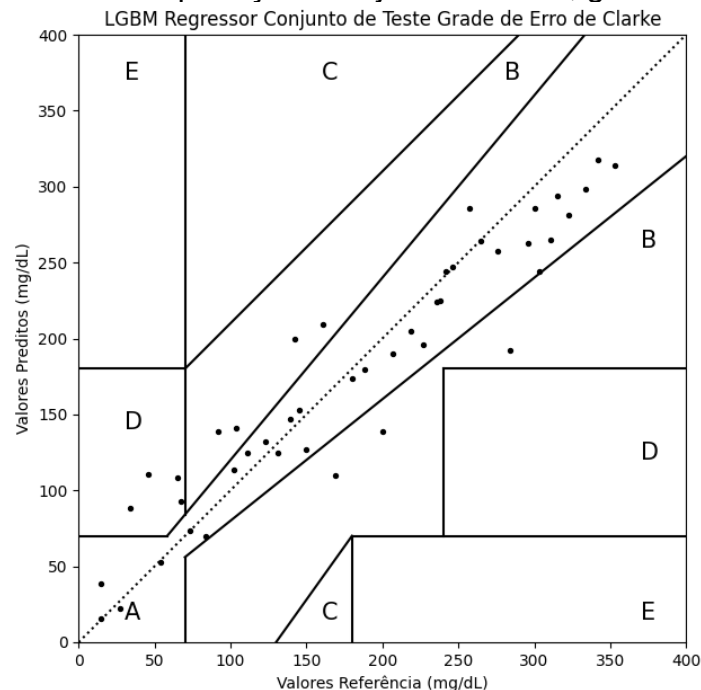
Fonte: Produção do próprio autor (2023)

LGBM utiliza abordagem “*leaf-wise*”, a divisão ocorre na direção que mais reduz a função perda. O menor RMSE obtido após a predição do conjunto de teste, foi de 34.03 e um R^2 de 0.88. Isso significa que o modelo erra em média 34.03 mg/dL.

É possível perceber que a diminuição da quantidade de atributos por meio da seleção de atributos utilizando RF e o método de importância de atributos, selecionando apenas os 22 atributos que possuíam maior peso para as predições, houve um ganho de 40,70% no valor do RMSE, indo de um RSME 47.89 (todos os atributos), para um RMSE de 34.03 (abordagem com 22 atributos)

A pouca quantidade de dados, para treinamento pode ser uma das razões para o valor alto desta métrica, além da pouca quantidade de dados, uma maior variabilidade dos dados, também contribui para melhores ajustes dos modelos aos dados. Um processo mais completo e exaustivo de diversas leituras de diferentes concentrações de glicose em amostras, em um intervalo menor, por exemplo 5 mg/dL, será benéfico para o ajuste do modelo. Na Figura 57, temos os dados previstos do modelo LGBM para o conjunto de teste.

Figura 57: Modelo LGBM predições conjunto de teste, grade de erro de Clarke.



Fonte: Produção do próprio autor (2023).

O modelo LGBM foi o modelo que obteve maior percentual de dados dentro da região A, cerca de 76%, 16% dos dados na zona B, 8% na zona D. É de suma

importância ressaltar que os dados preditos neste modelo, possuem erros em regiões críticas, como na região de hipoglicemia (glicemia menor que 80 mg/dL), onde o modelo prediz valores acima da referência. E na região de hiperglicemia (glicemia acima de 200 mg/dL), onde o modelo prediz valores abaixo da referência.

São erros que possuem potencial impactante para o indivíduo, já que prever valores acima, na região de hipoglicemia, pode acabar por trazer riscos ao paciente, não o informando de uma queda importante nos valores de sua glicemia. Da mesma forma na região de hiperglicemia, onde o indivíduo possui altos valores de glicose, e o modelo prevê valores menores, não demonstrando a importância da diminuição destes valores.

Na Tabela 22, temos as métricas RMSE e R^2 para cada região da grade de erro de Clarke, a grade foi dividida entre valores menores que 70 mg/dL (hipoglicemia), valores entre 80 e 200 mg/dL (normoglicemia) e valores acima de 200 mg/dL (hiperglicemia). Com o objetivo de entender em qual região o modelo se ajusta melhor, e demonstrar como erros nas extremidades da grade de erro de Clarke, são perigosos.

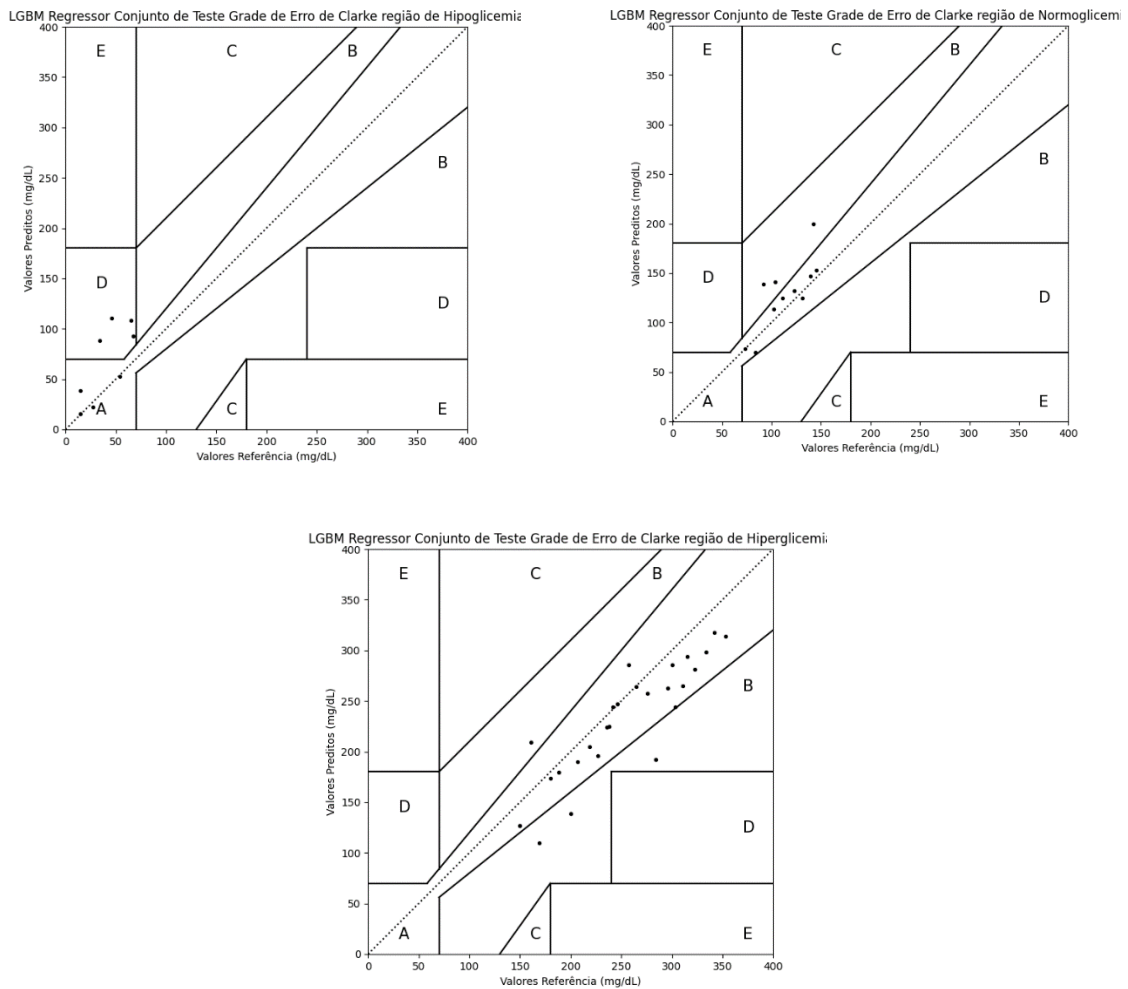
Tabela 22: Métricas para as 3 regiões definidas.

Região	RMSE	R2
Hipoglicemia	35.71	-2.36
Normoglicemia	26.41	-0,27
Hiperglicemia	36.29	0,61

Fonte: Produção do próprio autor (2023).

É possível visualizar na Figura 58, que os maiores erros nas predições do modelo, estão na região de hipoglicemia e hiperglicemia. Onde na região de hipoglicemia 50 % das predições estão localizadas a Região D. Na região de hiperglicemia cerca de 85% das predições estão dentro da região A. Na região de normoglicemia, cerca de 73% das predições estão dentro da região A.

Figura 58: Grade de Erro de Clarke dividida em 3 regiões. Esquerda superior (Hipoglicemia), direita superior (Normoglicemia) e inferior (Hiperglicemia).

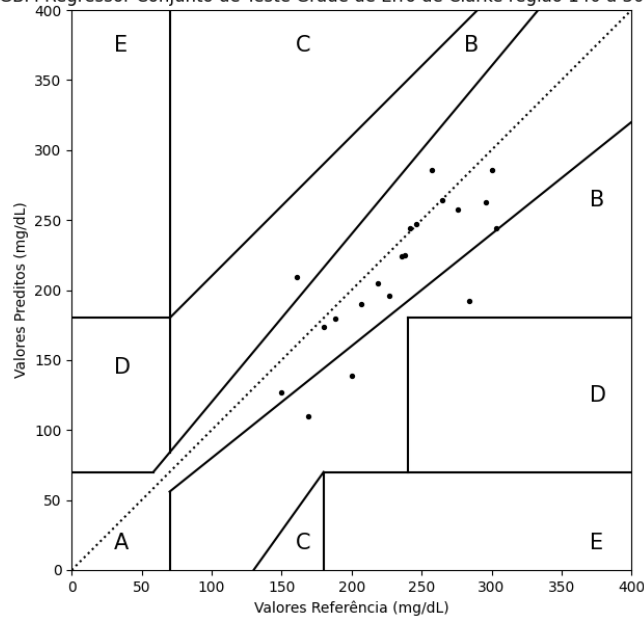


Fonte: Produção do próprio autor (2023).

Neste estudo, foram apresentadas as previsões abrangendo uma ampla faixa de glicemia, variando de 15 mg/dL até 350 mg/dL. Essas previsões evidenciaram a precisão do modelo em cada intervalo. No entanto, é importante ressaltar que em alguns estudos anteriores na literatura, os dados de teste foram limitados a um intervalo reduzido, o que pode comprometer a avaliação adequada das previsões. Isso pode ser observado na Figura 40, onde é possível notar que, se considerássemos a avaliação das previsões entre 140 mg/dL e 300 mg/dL (Figura 59), o erro quadrático médio (RMSE) seria de 35,43 e 82% das previsões estão na zona A. Por outro lado, ao avaliar apenas as previsões abaixo de 80 mg/dL, constatamos que 77% delas estão localizadas na região B da grade de erro de Clarke.

Figura 59: Glicemia entre de 140 mg/dL e 300 mg/dL.

LGBM Regressor Conjunto de Teste Grade de Erro de Clarke região 140 a 300 mg/d



Fonte: Produção do próprio autor (2023).

Conforme evidenciado por vários estudos anteriores (GUEVARA; GONZÁLEZ, 2010; PATHIRAGE et al., 2019; SANTOS, 2018; SILVA; LUIZ, 2017; SONG, K. et al., 2015b; WANG, Z. et al., 2022), os dados de teste utilizados na grade de erro de Clarke apresentam intervalos restritos (por exemplo, de 100 mg/dL a 180 mg/dL), o que limita a capacidade de generalização do modelo para toda a faixa de glicemia, incluindo valores de hipoglicemia e hiperglicemia.

No capítulo 7.8 deste estudo, é discutida a aplicação do modelo de classificação para a predição de níveis de glicose, em substituição à abordagem de regressão que prevê valores numéricos contínuos. A classificação, por sua vez, possibilita a previsão de valores numéricos categóricos, tais como intervalos definidos para os níveis de glicose. Mediante a utilização dessa abordagem, calculamos o RMSE para o conjunto de teste, cujos resultados são apresentados na tabela 23.

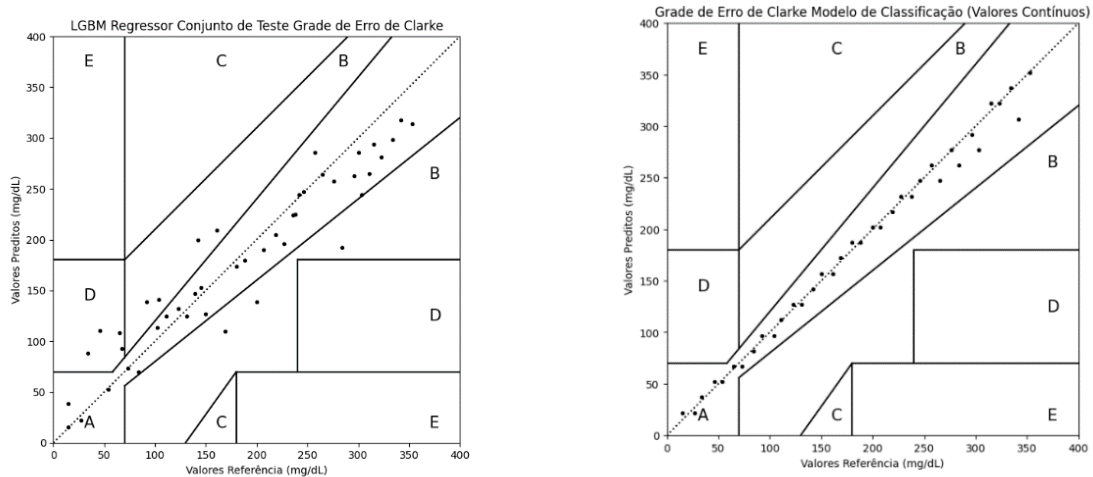
Tabela 23: RMSE abordagem de classificação vs abordagem de regressão para o conjunto de teste.

Abordagem	RMSE
Classificação	8.88
Regressão	34.03

Fonte: Produção do próprio autor (2023).

Na Figura 60(esquerda) é possível observar a grade erro de clarke para as predições com o modelo de regressão e na Figura 60(direita) temos a grade de erro de clarke para as predições com o modelo de regressão, utilizando valores numéricos contínuos.

Figura 60: Comparativo grade de erro de Clarke Modelos de Regressão(esquerda) e Classificação (Direita).



Fonte: Produção do próprio autor (2023).

É importante destacar que, embora o RMSE não seja comumente empregado como métrica para modelos de classificação, neste caso, procede-se à comparação entre o valor previsto pela abordagem e o valor original de glicemia (caracterizado como uma variável contínua). Tal procedimento viabiliza o cálculo do RMSE, possibilitando uma análise mais aprofundada do desempenho do modelo.

O modelo de classificação empregado no estudo produziu previsões com valores discretos, limitados a intervalos específicos, o que pode resultar em algumas predições situadas dentro dos limites aceitáveis, enquanto outras podem estar além desses limites, especialmente nos casos em que os valores de glicemia se aproximam dos limites dos intervalos.

Para aprimorar a precisão e a confiabilidade do modelo, tal como mencionado no texto, é recomendável adotar uma abordagem que empregue uma diversidade mais ampla de dados e reduza a amplitude dos intervalos de classes utilizados. Ao estabelecer faixas de valores menores, como por exemplo, intervalos de 10 mg/dL a 5 mg/dL, seria possível obter previsões mais aproximadas aos dados contínuos em análise, resultando em um aumento da precisão das estimativas.

Nesse sentido, a utilização de um intervalo menor de valores discretos no modelo de classificação propiciaria uma maior probabilidade de cumprimento dos critérios mínimos de acurácia estabelecidos pela norma ISO 15197. Todavia, para alcançar os níveis de acurácia requeridos, seria imprescindível coletar uma quantidade substancialmente maior de dados representativos e diversificados, visando o adequado treinamento do modelo, a fim de lidar com a complexidade e a variabilidade dos níveis de glicose. Apenas mediante uma abordagem mais abrangente e robusta seria possível obter os resultados almejados e estar em conformidade com os padrões de acurácia estipulados pelas normas vigentes para medições de glicemia.

Por fim foi aplicado ao modelo o conjunto de Novas leituras, a fim de comprovar a capacidade de generalização do modelo, no total foram feitas 20 predições, com foco na região de normoglicemia ($71 \leq \text{valor} \leq 140$ mg/dL), foi aplicado no modelo de regressão com a abordagem com todos os atributos e com 22 atributos, o melhor modelo com todos os atributos foi o KNN as métricas estão na Tabela 24, e o melhor modelo com 22 atributos foi o XGB as métricas estão na Tabela 24.

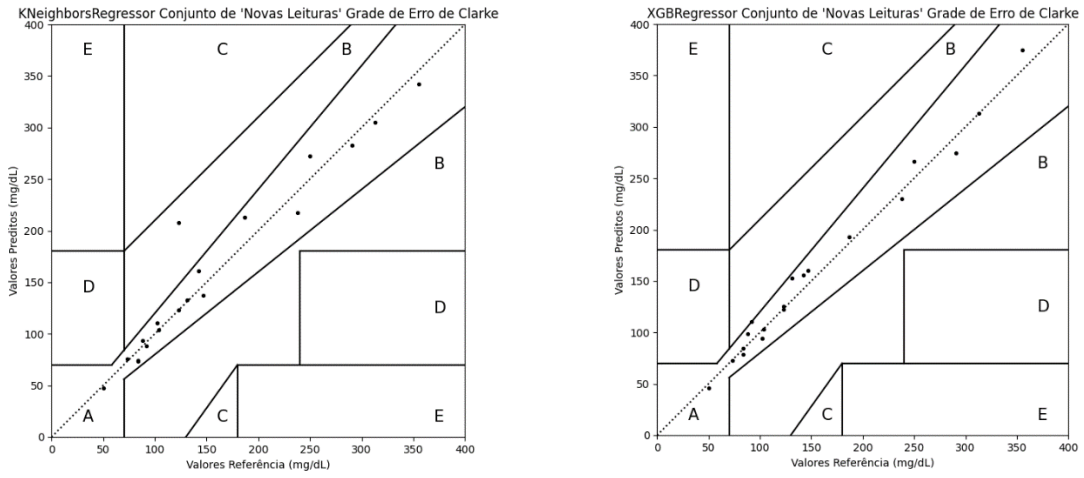
Tabela 24: Métricas para o conjunto de Novas Leituras Modelos de regressão.

MODELO	RMSE	R2	MAPE
KNN	22.88	0.93	10%
XGB	11.41	0.98	6.29%

Fonte: Produção do próprio autor (2023).

Pode-se observar que ambos os modelos apresentaram métricas superiores às obtidas com o conjunto de treinamento e teste. Isso sugere que ambos os modelos possuem uma capacidade de generalização, o que é desejável. No entanto, dado que ambas as métricas são melhores do que as métricas do conjunto de treinamento e teste, é importante considerar a possibilidade de subajuste (underfitting) durante o treinamento de ambos os modelos. A Figura 61 mostra as previsões geradas por ambos os modelos no gráfico de erro de Clarke, e é evidente que todas as previsões se encontram dentro das zonas A e B, representando 100% de precisão.

Figura 61: Grade de Erro de Clarke predições conjunto Novas Leituras. (Esquerda) Modelo KNN. (Direita) Modelo XGB abordagem de 22 atributos.



Fonte: Produção do próprio autor (2023).

9 CONCLUSÃO

A contribuição para o desenvolvimento de um sistema de medição não invasiva de glicose e metodologia de obtenção das amostras *in vitro*, capaz de fornecer medidas com acurácia, consistiu no objetivo principal deste trabalho.

A metodologia proposta foi constituída pelas etapas de obtenção da amostra *in vitro*, desenvolvimento do protótipo de NIRS e espectrometria, construção do modelo de aprendizado de máquina, validação, testes e novas medidas. A obtenção das amostras *in vitro*, foram feitas diluindo glicose em água milliq. O desenvolvimento do protótipo, foi feito em duas partes separadas, hardware e software. A criação dos modelos foi feita com foco em modelos de árvore de decisão, e a avaliação das previsões do conjunto de validação e conjunto de testes, foram feitas utilizando como métricas RMSE E R^2 .

A pesquisa deve ser destacada devido à busca por uma metodologia reprodutível e às descobertas realizadas, mesmo não tendo sido encontrados valores de RMSE inferiores. O instrumento demonstrou efetivamente a capacidade de medir glicose *in vitro* de forma não invasiva por meio transdérmico, confirmando assim o seu princípio de funcionamento básico. No entanto, em relação aos requisitos de desempenho *in vitro*, os critérios de limite de quantificação não foram completamente cumpridos. Em termos de previsão, os resultados mais satisfatórios obtidos pelo instrumento foram os valores de RMSE, R^2 e MAPE, são expostos na Tabela 25, para as etapas de teste e novas leituras.

Tabela 25: Métricas para o conjunto de teste e Novas Leituras Modelos de regressão.

MODELO	RMSE (mg/dL)	R2	MAPE
Conjunto de Teste			
LGBM	47.89	0.77	44.22%
LGBM (22 Atributos)	34.03	0.88	24.18%
Conjunto de Novas Medidas			
KNN	22.88	0.93	10%
XGB (22 atributos)	11.41	0.98	6.29%

Fonte: Produção do próprio autor (2023).

Segundo a menor de todas as estimativas do limite do branco e o melhor resultado em previsão *in vitro*, os desvios do instrumento em termos de concentração são de, ao menos, 11,0 mg/dL de glicose. Esses erros de concentração são

decorrentes, provavelmente, dos desvios intrínsecos ao próprio instrumento e, possivelmente, de fatores externos relacionados à execução dos experimentos.

De forma geral, atingiu-se o objetivo do trabalho, haja vista que foram descritos os aspectos fundamentais da técnica de espectroscopia (com potencial aplicação em outras áreas da engenharia biomédica), foi desenvolvido uma metodologia para o desenvolvimento de amostras para teste *in vitro*, para servir como suporte para futuros trabalhos, e desenvolvido um protótipo para medição de glicose por meio não invasivo, utilizando como receptor um nano espectômetro, algo que não há na literatura com para essa aplicação em específico. Entretanto, aperfeiçoamentos técnicos são ainda necessários o desempenho adequado do instrumento na medição não invasiva da glicemia. Como o controle da temperatura dos LEDs, o que pode acabar por afetar as condições da amostra, o controle de fatores ambientais, como grandes variações de temperatura e umidade.

No âmbito metodológico, é recomendável que estudos futuros adotem a metodologia empregada neste trabalho para a obtenção das amostras, utilizando um modelo simplificado, e avancem para modelos mais complexos, como fantasmas que simulem as propriedades do sangue humano, incluindo glicose em suspensão de lipídeos, proteínas e sais minerais. Somente após a demonstração de resultados em modelos *in vitro* é aconselhável prosseguir para testes em pacientes, levando em consideração que, em medições *in vivo*, as variações nas medidas são ainda mais significativas.

No desfecho deste estudo, o capítulo 7.8 apresentou a aplicação do modelo de classificação como alternativa à abordagem de regressão, com o objetivo de prever os níveis de glicose. A classificação, ao possibilitar a previsão de valores numéricos categóricos, como intervalos definidos para os níveis de glicose, revelou-se uma estratégia promissora. Ao calcular RMSE para o conjunto de teste, constatamos que a abordagem de classificação demonstrou um desempenho superior (RMSE=8.88) conforme tabela 23 em comparação com a abordagem de regressão (RMSE=11.41), conforme exibido na Tabela 24. Embora o RMSE não seja tradicionalmente empregado como métrica para modelos de classificação, sua utilização neste contexto permitiu a comparação entre os valores previstos pela abordagem e os valores originais de glicemia, caracterizados como uma variável contínua. Essa abordagem de predição com valores discretos e intervalos específicos demonstrou ser uma estratégia valiosa, porém, vislumbra-se que resultados ainda

mais precisos e confiáveis poderiam ser alcançados por meio do emprego de uma ampla diversidade de dados e redução dos intervalos de classes, o que possibilitaria uma melhor aproximação entre os valores preditos e os valores contínuos reais em análise. Como um desdobramento natural deste estudo, futuras pesquisas poderiam explorar essas recomendações para aprimorar a precisão da predição dos níveis de glicose e contribuir para avanços significativos na área médica e de análise preditiva.

A abordagem baseada na NIRS tem demonstrado potencial na literatura científica. No entanto, é recomendável explorar novas abordagens híbridas, como a combinação da NIRS com a técnica de bioimpedância. Alguns estudos na literatura utilizam exclusivamente a bioimpedância para a medição não invasiva da glicemia, enquanto outros adotam abordagens híbridas envolvendo tanto a NIRS quanto a bioimpedância, e apresentam resultados promissores (ANDERSEN *et al.*, 2019; DO AMARAL, 2008; FOUAD; MAHMOUD; ABD EL GHANY, 2018; PATHIRAGE *et al.*, 2019; SANTOS, 2018). Portanto, é encorajador investigar essas alternativas híbridas e sua viabilidade para a medição não invasiva da glicemia.

De qualquer forma, acredita-se que o desenvolvimento de dispositivos não invasivos seja uma tendência tecnológica e, com o interesse de empresas que possuam recursos financeiros e humanos adequados, essas tecnologias serão implementadas com êxito no futuro, substituindo os métodos de medição invasiva de glicose atualmente em uso. Essas tecnologias oferecem diversas vantagens, e espera-se que, em um prazo razoável, possam se tornar amplamente adotadas.

REFERÊNCIAS

ALORAYNAN, Abdulrahman *et al.* A Single Wavelength Mid-Infrared Photoacoustic Spectroscopy for Noninvasive Glucose Detection Using Machine Learning. **Biosensors**, [s. l.], v. 12, n. 3, p. 1–19, 2022. Disponível em: <https://doi.org/10.3390/bios12030166>

AMASYALI, Kadir; EL-GOHARY, Nora M. A review of data-driven building energy consumption prediction studies. **Renewable and Sustainable Energy Reviews**, [s. l.], v. 81, n. September 2017, p. 1192–1205, 2018. Disponível em: <https://doi.org/10.1016/j.rser.2017.04.095>

ANDERSEN, Jan Hugo *et al.* Bioimpedance and NIR for non-invasive assessment of blood glucose. **Journal of Electrical Bioimpedance**, [s. l.], v. 10, n. 1, p. 133–138, 2019. Disponível em: <https://doi.org/10.2478/joeb-2019-0019>

BEĆ, Krzysztof B.; GRABSKA, Justyna; HUCK, Christian W. Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. **Chemistry - A European Journal**, [s. l.], v. 27, n. 5, p. 1514–1532, 2021. Disponível em: <https://doi.org/10.1002/chem.202002838>

BERRAR, Daniel. Cross-Validation. **Data Science Laboratory**, [s. l.], n. January 2018, p. 0–8, 2019. Disponível em: <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>

BOROVČNIK, Manfred; BENTZ, Hans-Joachim; KAPADIA, Ramesh. **A Probabilistic Perspective**. [S. l.: s. n.], 2012. Disponível em: https://doi.org/10.1007/978-94-011-3532-0_2

BOUSTANY, Nada N.; BOPPART, Stephen A.; BACKMAN, Vadim. Microscopic imaging and spectroscopy with scattered light. **Annual Review of Biomedical Engineering**, [s. l.], v. 12, n. 732, p. 285–314, 2010. Disponível em: <https://doi.org/10.1146/annurev-bioeng-061008-124811>

CHANG, Cheng Chun *et al.* MW-PPG sensor: An on-chip spectrometer approach. **Sensors (Switzerland)**, [s. l.], v. 19, n. 17, 2019. Disponível em: <https://doi.org/10.3390/s19173698>

CHEN, Shao Hao; CHUANG, Yung Chi; CHANG, Cheng Chun. Development of a portable all-wavelength PPG sensing device for robust adaptive-depth measurement: A spectrometer approach with a hydrostatic measurement example. **Sensors (Switzerland)**, [s. l.], v. 20, n. 22, p. 1–12, 2020. Disponível em: <https://doi.org/10.3390/s20226556>

CHUNG, Joanne *et al.* Recent advances in noninvasive glucose monitoring. **Medical Devices: Evidence and Research**, [s. l.], p. 45, 2012. Disponível em: <https://doi.org/10.2147/mdir.s28134>

CLARKE, William. The Original Clarke Error Grid Analysis (EGA). **DIABETES TECHNOLOGY & THERAPEUTICS**, [s. l.], v. 7, n. 5, p. 776–779, 2005.

DAVIDE, Giavarina. Understanding Bland Altman Analysis. **Biochemia Medica**, [s. l.], v. 19, n. 1, p. 10–16, 2015.

DEISS, Leonardo *et al.* Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. **Geoderma**, [s. l.], v. 365, n. September 2019, p. 114227, 2020. Disponível em: <https://doi.org/10.1016/j.geoderma.2020.114227>

DIABETES, Federation International. **International Diabetes Federation**. [S. l.: s. n.], 2021. ISSN 18728227.v. 10 Disponível em: <https://doi.org/10.1016/j.diabres.2013.10.013>

DO AMARAL, Carlos Eduardo Ferrante. Multiparameter Methods for Non-invasive Measurement of Blood Glucose. [s. l.], p. 153, 2008.

DZIERGOWSKA, Katarzyna *et al.* Modern noninvasive methods for monitoring glucose levels in patients: A review. **Bio-Algorithms and Med-Systems**, [s. l.], p. 1–22, 2019. Disponível em: <https://doi.org/10.1515/bams-2019-0052>

FOUAD, Mariam; MAHMOUD, Duha; ABD EL GHANY, Mohamed. Joint NIR-BIS based Non-invasive Glucose Monitoring System. **Internacional Conference on Microelectronics**, [s. l.], v. 30, p. 1–4, 2018.

GENG, Zhanxiao *et al.* Noninvasive Continuous Glucose Monitoring Using a

Multisensor-Based Glucometer and Time Series Analysis. **Scientific Reports**, [s. l.], v. 7, n. 1, p. 1–10, 2017. Disponível em: <https://doi.org/10.1038/s41598-017-13018-7>

GÉRON, A. Mãos à obra: aprendizado de máquina com Scikit-Learn. **Keras & TensorFlow: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes.**[S. l.: sn], 2021.

GONZALES, Wilbert Villena; MOBASHSHER, Ahmed Toaha; ABBOSH, Amin. **The progress of glucose monitoring—A review of invasive to minimally and non-invasive techniques, devices and sensors.** [S. l.: s. n.], 2019. ISSN 14248220.v. 19 Disponível em: <https://doi.org/10.3390/s19040800>

GUEVARA, E.; GONZÁLEZ, F. J. Joint optical-electrical technique for noninvasive glucose monitoring. **Revista Mexicana de Física**, [s. l.], v. 56, n. 5, p. 430–434, 2010.

HABIBULLAH, Mohammad *et al.* NIR-Spectroscopic Classification of Blood Glucose Level using Machine Learning Approach. **2019 IEEE Canadian Conference of Electrical and Computer Engineering, CCECE 2019**, [s. l.], p. 1–4, 2019. Disponível em: <https://doi.org/10.1109/CCECE.2019.8861843>

HE, Diwei *et al.* A single-chip CMOS pulse oximeter with on-chip lock-in detection. **Sensors (Switzerland)**, [s. l.], v. 15, n. 7, p. 17076–17088, 2015. Disponível em: <https://doi.org/10.3390/s150717076>

HOTMARTUA, Rolamjaya *et al.* Noninvasive blood glucose detection using near infrared sensor. **Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015**, [s. l.], p. 687–692, 2015. Disponível em: <https://doi.org/10.1109/ICEEI.2015.7352586>

INJADAT, Mohammadnoor *et al.* Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection. **2018 IEEE Global Communications Conference (GLOBECOM)**, [s. l.], p. 1–6, 2018.

JOHNSON, Eric L. *et al.* American diabetes Association Standards of medical care in diabetes - 2018. **American diabetes Association**, [s. l.], v. 39, n. 1, p. 150, 2021.

Disponível em: <https://doi.org/10.2337/cd21-as01>

JONASSON, Hanna *et al.* In vivo characterization of light scattering properties of human skin in the 475- to 850-nm wavelength range in a Swedish cohort. **Journal of Biomedical Optics**, [s. l.], v. 23, n. 12, p. 1, 2018. Disponível em: <https://doi.org/10.1117/1.jbo.23.12.121608>

KOZMA, Bence *et al.* A model system and chemometrics to develop near infrared spectroscopic monitoring for Chinese hamster ovary cell cultivations. **Journal of Near Infrared Spectroscopy**, [s. l.], v. 22, n. 6, p. 401–410, 2014. Disponível em: <https://doi.org/10.1255/jnirs.1133>

LEE, Yongsu *et al.* Sticker-type ECG/PPG concurrent monitoring system hybrid integration of CMOS SoC and organic sensor device. **Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS**, [s. l.], v. 2016-Octob, p. 2014–2017, 2016. Disponível em: <https://doi.org/10.1109/EMBC.2016.7591121>

LEE, Yongsu *et al.* Sticker-Type Hybrid Photoplethysmogram Monitoring System Integrating CMOS IC with Organic Optical Sensors. **IEEE Journal on Emerging and Selected Topics in Circuits and Systems**, [s. l.], v. 7, n. 1, p. 50–59, 2017. Disponível em: <https://doi.org/10.1109/JETCAS.2016.2630301>

MALIK, Arif Jamal; SHAHZAD, Waseem; KHAN, Farrukh Aslam. Binary PSO and Random Forests Algorithm for PROBE Attacks Detection in a Network. [s. l.], p. 662–668, 2011.

MARUO, Katsuhiko; YAMADA, Yukio. Near-infrared noninvasive blood glucose prediction without using multivariate analyses: introduction of imaginary spectra due to scattering change in the skin. **Journal of Biomedical Optics**, [s. l.], v. 20, n. 4, p. 047003, 2015. Disponível em: <https://doi.org/10.1117/1.jbo.20.4.047003>

MOODLEY, Nareshni *et al.* Historical perspectives in clinical pathology: A history of glucose measurement. **Journal of Clinical Pathology**, [s. l.], v. 68, n. 4, p. 258–264, 2015. Disponível em: <https://doi.org/10.1136/jclinpath-2014-202672>

NATHAN, David M. *et al.* Impaired fasting glucose and impaired glucose tolerance:

Implications for care. **Diabetes Care**, [s. l.], v. 30, n. 3, p. 753–759, 2007. Disponível em: <https://doi.org/10.2337/dc07-9920>

NAWAZ, Asmat *et al.* Review: Non-Invasive Continuous Blood Glucose Measurement Techniques. **Bioinformatics and Diabetes**, [s. l.], v. 1, n. 3, p. 1–27, 2016. Disponível em: <https://oap-lifescience.org/jbd/article/297>

NICKERSON, Helen D.; DUTTA, Sanjoy. Diabetic complications: Current challenges and opportunities. **Journal of Cardiovascular Translational Research**, [s. l.], v. 5, n. 4, p. 375–379, 2012. Disponível em: <https://doi.org/10.1007/s12265-012-9388-1>

PATHIRAGE, Kasun D. *et al.* Removing subject dependencies on Non-Invasive Blood Glucose Measurement using Hybrid Techniques. **Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS**, [s. l.], p. 7197–7200, 2019. Disponível em: <https://doi.org/10.1109/EMBC.2019.8856391>

PFÜTZNER, Andreas *et al.* Evaluation of a New Noninvasive Glucose Monitoring Device by Means of Standardized Meal Experiments. **Journal of Diabetes Science and Technology**, [s. l.], v. 12, n. 6, p. 1178–1183, 2018. Disponível em: <https://doi.org/10.1177/1932296818758769>

PRIYOTI, Anika Tabassum *et al.* Non-invasive blood glucose measurement using near infra-red spectroscopy. **IEEE Region 10 Humanitarian Technology Conference, R10-HTC**, [s. l.], v. 2019-Novem, p. 1139–1142, 2019. Disponível em: <https://doi.org/10.1109/R10-HTC47129.2019.9042473>

PRUSTY, Sashikanta; PATNAIK, Srikanta; DASH, Sujit Kumar. SKCV : Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. [s. l.], n. August, p. 1–12, 2022. Disponível em: <https://doi.org/10.3389/fnano.2022.972421>

ROGERS, Jeremy D. *et al.* Modeling Light Scattering in Tissue as Continuous Random Media Using a Versatile Refractive Index Correlation Function. **IEEE Journal of Selected Topics in Quantum Electronics**, [s. l.], v. 20, n. 2, p. 1–13, 2014. Disponível em: <https://doi.org/10.1109/JSTQE.2013.2280999>

SANTOS, Sergio Fernandez. MEASURING AND MODELLING BLOOD GLUCOSE

LEVEL USING NEAR INFRARED AND BIOIMPEDANCE DATA. [s. l.], 2018.

SARKAR, Mukul; ASSAAD, Maher. Noninvasive Non-Contact SpO₂ Monitoring Using an Integrated Polarization-Sensing CMOS Imaging Sensor. **Sensors (Basel, Switzerland)**, [s. l.], v. 22, n. 20, p. 1–16, 2022. Disponível em: <https://doi.org/10.3390/s22207796>

SEGMAN, Yosef (Joseph). Device and Method for Noninvasive Glucose Assessment. **Journal of Diabetes Science and Technology**, [s. l.], v. 12, n. 6, p. 1159–1168, 2018. Disponível em: <https://doi.org/10.1177/1932296818763457>

SHARMA, Deepti; KUMAR, Rajneesh; JAIN, Anurag. Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. **Measurement: Sensors**, [s. l.], v. 24, n. November, 2022. Disponível em: <https://doi.org/10.1016/j.measen.2022.100560>

SHOKREKHODAEI, Maryamsadat *et al.* Non-Invasive Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications. **IEEE Access**, [s. l.], v. 9, p. 73029–73045, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2021.3079182>

SHOKREKHODAEI, Maryamsadat; QUINONES, Stella. Review of non-invasive glucose sensing techniques: Optical, electrical and breath acetone. **Sensors (Switzerland)**, [s. l.], v. 20, n. 5, 2020. Disponível em: <https://doi.org/10.3390/s20051251>

SILVA, Bruno; LUIZ, Jefferson. SISTEMA DE MEDIÇÃO NÃO INVASIVA DE GLICOSE SANGUÍNEA BASEADO EM PRINCÍPIOS DE ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO. **Dissertação**, [s. l.], p. 220, 2017.

SMITH, John. The Pursuit of Noninvasive Glucose: " Hunting the Deceitful Turkey ". [s. l.], n. August, 2020.

SONG, Hongya *et al.* Review of compact computational spectral information acquisition systems. **Frontiers of Information Technology and Electronic Engineering**, [s. l.], v. 21, n. 8, p. 1119–1133, 2020. Disponível em: <https://doi.org/10.1631/FITEE.1900266>

SONG, Kiseok *et al.* An Impedance and Multi-Wavelength Near-Infrared Spectroscopy IC for Non-Invasive Blood Glucose Estimation. **IEEE Journal of Solid-State Circuits**, [s. l.], v. 50, n. 4, p. 1025–1037, 2015a. Disponível em: <https://doi.org/10.1109/JSSC.2014.2384037>

SONG, Kiseok *et al.* An Impedance and Multi-Wavelength Near-Infrared Spectroscopy IC for Non-Invasive Blood Glucose Estimation. **IEEE Journal of Solid-State Circuits**, [s. l.], v. 50, n. 4, p. 1025–1037, 2015b. Disponível em: <https://doi.org/10.1109/JSSC.2014.2384037>

STEFAN-VAN STADEN, Raluca Ioana; MITROFAN, Grigorina. Molecular enantio recognition of l-glucose and d-glucose in whole blood samples. **Chirality**, [s. l.], v. 30, n. 5, p. 680–685, 2018. Disponível em: <https://doi.org/10.1002/chir.22843>

SUN, Hong *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. **Diabetes Research and Clinical Practice**, [s. l.], v. 183, p. 13, 2022. Disponível em: <https://doi.org/10.1016/j.diabres.2021.109119>

SVYATKOVSKIY, Alexey; KATES-HARBECK, Julian; TANG, William. **Training distributed deep recurrent neural networks with mixed precision on GPU clusters**. [S. l.: s. n.], 2017. Disponível em: <https://doi.org/10.1145/3146347.3146358>

TANG, Liu *et al.* Non-invasive blood glucose monitoring technology: A review. **Sensors (Switzerland)**, [s. l.], v. 20, n. 23, p. 1–32, 2020. Disponível em: <https://doi.org/10.3390/s20236925>

TANG, Wei *et al.* Dynamic spectrum nonlinear modeling of VIS & NIR band based on RBF neural network for noninvasive blood component analysis to consider the effects of scattering. **Infrared Physics and Technology**, [s. l.], v. 96, n. September 2018, p. 77–83, 2019. Disponível em: <https://doi.org/10.1016/j.infrared.2018.11.018>

TAO, Ziqi; SHI, Aimin; ZHAO, Jing. Epidemiological Perspectives of Diabetes. **Cell Biochemistry and Biophysics**, [s. l.], v. 73, n. 1, p. 181–185, 2015. Disponível em: <https://doi.org/10.1007/s12013-015-0598-4>

TONYUSHKINA, Ksenia; NICHOLS, James H. Glucose meters: A review of technical

challenges to obtaining accurate results. **Journal of Diabetes Science and Technology**, [s. l.], v. 3, n. 4, p. 971–980, 2009. Disponível em: <https://doi.org/10.1177/193229680900300446>

TURGUL, Volkan; KALE, Izzet. All-digital 1550 nm optical aqueous glucose solution measurement system. **ICECS 2017 - 24th IEEE International Conference on Electronics, Circuits and Systems**, [s. l.], v. 2018-Janua, n. 1, p. 462–465, 2018. Disponível em: <https://doi.org/10.1109/ICECS.2017.8292032>

VAN ENTER, Benjamin Jasha; VON HAUFF, Elizabeth. Challenges and perspectives in continuous glucose monitoring. **Chemical Communications**, [s. l.], v. 54, n. 40, p. 5032–5045, 2018. Disponível em: <https://doi.org/10.1039/c8cc01678j>

VURAL, Nuri Mert; YOLERI, Yigit; TORUN, Hamdi. On feasibility of near-infrared spectroscopy for noninvasive blood glucose measurements. [s. l.], n. February, p. 29, 2019. Disponível em: <https://doi.org/10.1117/12.2503852>

WAN, Xinghua *et al.* A review on the strategies for reducing the non-linearity caused by scattering on spectrochemical quantitative analysis of complex solutions. **Applied Spectroscopy Reviews**, [s. l.], v. 55, n. 5, p. 351–377, 2020. Disponível em: <https://doi.org/10.1080/05704928.2019.1584567>

WANG, Junzhuan; ZHENG, Binjie; WANG, Xiaomu. Strategies for high performance and scalable on-chip spectrometers. **Journal of Physics: Photonics**, [s. l.], v. 3, n. 1, p. 0–15, 2021. Disponível em: <https://doi.org/10.1088/2515-7647/abc897>

WANG, Zhuyu *et al.* Development of non-invasive blood glucose regression based on near-infrared spectroscopy combined with a deep-learning method. **Journal of Physics D: Applied Physics**, [s. l.], v. 55, n. 21, 2022. Disponível em: <https://doi.org/10.1088/1361-6463/ac4723>

WATT, Jeremy; BORHANI, Reza; KATSAGGELOS, Aggelos. **Machine Learning Refined**. [S. l.: s. n.], 2016. Disponível em: <https://doi.org/10.1017/cbo9781316402276>

YADAV, Jyoti *et al.* Near-infrared LED based non-invasive blood glucose sensor. **2014 International Conference on Signal Processing and Integrated Networks, SPIN 2014**, [s. l.], p. 591–594, 2014. Disponível em:

<https://doi.org/10.1109/spin.2014.6777023>

YAMASHITA, Yuichi. **Application of Near Infrared Spectroscopy in Mixed Spectrum Detection**. [S. l.: s. n.], 2013. ISSN 22110992.v. 98 Disponível em: https://doi.org/10.1007/978-3-030-67090-0_36

YANG, Li; SHAMI, Abdallah. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, [s. l.], v. 415, p. 295–316, 2020. Disponível em: <https://doi.org/10.1016/j.neucom.2020.07.061>

YANG, Wenya *et al.* Economic costs of diabetes in the U.S. in 2017. **Diabetes Care**, [s. l.], v. 41, n. 5, p. 917–928, 2018. Disponível em: <https://doi.org/10.2337/dci18-0007>

YANG, Zongyin *et al.* Miniaturization of optical spectrometers. **Science**, [s. l.], v. 371, n. 6528, 2021. Disponível em: <https://doi.org/10.1126/science.abe0722>

YEN, Chih-Ta *et al.* Non-Invasive Blood Glucose Estimation System Based on a Neural Network with Dual-Wavelength Photoplethysmography and Bioelectrical Impedance Measuring. **Sensors**, [s. l.], v. 22, n. 12, p. 4452, 2022. Disponível em: <https://doi.org/10.3390/s22124452>

ZHANG, Liang *et al.* A review of machine learning in building load prediction. **Applied Energy**, [s. l.], v. 285, n. January, p. 22, 2021. Disponível em: <https://doi.org/10.1016/j.apenergy.2021.116452>

ZHAO, Yifan *et al.* Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period. **Ecological Indicators**, [s. l.], v. 135, 2022. Disponível em: <https://doi.org/10.1016/j.ecolind.2021.108529>

ZIRK, K.; POETZSCHKE, H. On the suitability of refractometry for the analysis of glucose in blood-derived fluids. **Medical Engineering and Physics**, [s. l.], v. 26, n. 6, p. 473–481, 2004. Disponível em: <https://doi.org/10.1016/j.medengphy.2004.03.008>