UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Programa de Pós-Graduação em Biocombustíveis

Amanda Rocha Chaves

ANÁLISE MULTIVARIADA PARA DETERMINAÇÃO DE ATIVIDADE DE HIDROLASES COM APLICAÇÃO NO SETOR DE BIOCOMBUSTÍVEIS

Diamantina

2021

Amanda Rocha Chaves

ANÁLISE MULTIVARIADA PARA DETERMINAÇÃO DE ATIVIDADE DE HIDROLASES COM APLICAÇÃO NO SETOR DE BIOCOMBUSTÍVEIS

Tese apresentada ao programa de Pós-graduação da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, como requisito para obtenção do título de Doutor em Ciência e Tecnologia dos Biocombustíveis.

Orientador: Prof. Dr. Alexandre Soares dos Santos
Coorientadora: Profa. Drª. Lílian Araújo Pantoja

Diamantina

2021

Catalogação na fonte - Sisbi/UFVJM

C512 CHAVES, AMANDA Rocha
2021 ANÁLISE MULTIVARIADA PARA DETERMINAÇÃO DE ATIVIDADE DE HIDROLASES COM APLICAÇÃO NO SETOR DE BIOCOMBUSTÍVEIS [manuscrito] / AMANDA Rocha CHAVES. -- Diamantina, 2021.
125 p. : il.

Orientador: Prof. Alexandre Soares dos Santos.
Coorientador: Prof. Lílian de Araújo Pantoja.

Tese (Doutorado em Biocombustíveis) -- Universidade Federal dos Vales do Jequitinhonha e Mucuri, Programa de Pós-Graduação em Biocombustíveis, Diamantina, 2021.

1. análise multivariada. 2. atividade enzimática. 3. espectroscopia no infravermelho próximo. 4. quimiometria. 5. hemicelulases. I. Santos, Alexandre Soares dos. II. Pantoja, Lílian de Araújo. III. Universidade Federal dos Vales do Jequitinhonha e Mucuri. IV. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFVJM
com os dados fornecidos pelo(a) autor(a).
Bibliotecário Rodrigo Martins Cruz / CRB6-2886
Técnico em T.I. Thales Francisco Mota Carvalho

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

AMANDA ROCHA CHAVES

ANÁLISE MULTIVARIADA PARA DETERMINAÇÃO DE ATIVIDADE DE HIDROLASES COM APLICAÇÃO NO SETOR DE BIOCOMBUSTÍVEIS

Tese apresentada ao Programa de Pós-graduação em Biocombustíveis da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como requisito parcial para obtenção do título de Doutor em Ciência e Tecnologia dos Biocombustíveis.

Orientador: Prof. Dr. Alexandre Soares dos Santos

Co-orientadora: Profa. Dra. Lílian de Araújo Pantoja

Data de aprovação 25/01/2021.

Prof. Dr. Waldomiro Borges Neto **- UFU**

Prof.ª Dr.ª Verônica Ferreira Melo **- IFRJ**

Prof. Dr. Paulo Henrique Fidêncio **- UFVJM**

Prof. Dr. Paulo César de Resende Andrade **- UFVJM**

Prof. Dr. Alexandre Soares dos Santos **- UFVJM**
Orientador

Documento assinado eletronicamente por **Alexandre Soares dos Santos**, **Servidor**, em 25/03/2021, às 16:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Verônica Ferreira Melo**, **Usuário Externo**, em 25/03/2021, às 17:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Waldomiro Borges Neto**, **Usuário Externo**, em 25/03/2021, às 18:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Henrique Fidencio**, **Servidor**, em 27/08/2021, às 18:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de

8 de outubro de 2015.

Documento assinado eletronicamente por **Paulo Cesar de Resende Andrade**, **Servidor**, em 30/08/2021, às 09:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufvjm.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0259657** e o código CRC **9B5014D7**.

DEDICATÓRIA

Dedico este trabalho à querida amiga profª. Lílian Pantoja e à minha mãe.

AGRADECIMENTOS

Agradecer é relembrar e sentir gratidão por todas as pessoas e seres que estiveram presentes nessa jornada de aprendizado e compartilhamento de saberes.

Agradeço primeiramente ao universo, que em sua manifestação divina, apresentou-me às pessoas certas, que me apontariam o caminho e me sustentariam nesse percurso transformador.

Agradeço especialmente à minha querida amiga e coorientadora, profa. Lílian Pantoja, que desde o primeiro instante recebeu-me com afeto, acolheu-me e mostrou-me os primeiros passos de uma área antes totalmente desconhecida pra mim. Ela guiou-me nas tarefas laboratoriais e na escrita de todas as fases desse trabalho. Agradeço a ela por levar-me à sua casa para conseguir focar e concluir minhas tarefas. Em nenhum momento me deixou abater. E à sua maneira, deu-me forças para prosseguir. A profa. Lílian é pra mim um ser inspirador que acolhe a todos sem distinção.

Agradeço imensamente ao prof. Alexandre por, primeiramente, aceitar orientar-me e traçar comigo os rumos do meu projeto de tese. Agradeço por todos os ensinamentos na área de bioquímica, biocombustíveis, análises matemáticas e, principalmente, por todo o apoio incondicional. Agradeço pelo seu humor discreto e engraçado, levando-nos ao sentimento de "estar em casa" durante os experimentos laboratoriais, tua presença silenciosa deixou nosso trabalho mais leve e agradável. Agradeço ao professor que, juntamente com a profª. Lílian, deu todo o suporte não só a mim mas a todos que bateram à sua porta. A mim, mostrou-me como é brilhante ajudar a todos sem pedir nada em troca. Agradeço e peço desculpas por todo aperto que passou a meu favor. Sinto-me mesmo honrada de tê-los como orientadores. Eles são mesmo seres de luz.

Agradeço ao meu amigo Marcelo, que me ensinou muito sobre química e tornou a minha caminhada mais fácil. Ao prof. Paulo César pelo apoio e auxílio com estudos estatísticos.

Agradeço aos colegas que me ajudaram nos primeiros passos, como a querida Ilva de Fátima, Marina Lemos, Fernanda Monção, Fidel Aguilar, Felipe Brito e todos que estiveram no laboratório e deixaram lá uma pitadinha de saudade.

Um agradecimento muito especial aos meus amigos Ronnie e José Alberto (meu amado Zé) por me acompanharem de pertinho e tornaram os meus dias dessa fase mais

alegres e palpitantes. Agradeço-vos pelas boas risadas, abracinhos e pelos cafezinhos acolhedores e amorosos.

E não poderia me esquecer da pessoinha mais maravilhosa que apareceu e iluminou os meus dias, minha querida amiga Thaizinha. Ela chegou de mansinho, ganhou meu coração e a minha admiração. Os nossos experimentos laboratoriais foram regados a muita música, paz e concentração. Os dias de trabalho eram serenos com a sua parceria. Agradeço-vos imensamente.

Os meus agradecimentos se estendem à minha querida família, especialmente à minha mãe, a Wiviane e a Wivi, que sem vocês nada disso seria possível. Vocês movem-me sempre adiante e em momento algum reclamaram minha ausência. Agradeço e peço desculpas por todas as minhas faltas.

Agradeço, também, aos meus queridos amigos do círculo sagrado: Zeny, Pavani, Cris Silva, Ana Paula, Délcio, Paulina, José e Maria Cristina por toda vibração de luz para finalização da escrita.

Agradeço a todas as instituições que direta ou indiretamente financiam e sustentam as nossas pesquisas: FAPEMIG, CAPES e CNPQ, em especial, à UFVJM que me acolheu tanto como docente quanto discente.

Agradeço, ainda, a todos que estiveram a iluminar essa minha caminhada e que não conseguiria nomear neste momento.

RESUMO

O uso da espectrometria na região do infravermelho próximo (NIR) combinado com o método de análise multivariada tem possibilitado a determinação quantitativa, direta e não destrutiva de várias substâncias inorgânicas e orgânicas em amostras complexas. Neste trabalho foi investigada a possibilidade da determinação enzimática de forma direta para xilanases e celulases por meio de estudos de correlação entre os métodos convencionais para dosagem enzimática e a análise multivariada baseada em espectros de infravermelho próximo. Para tanto, três complexos enzimáticos, contendo as enzimas alvo, foram utilizados, sendo estes dois comerciais, o Celluclast® e o Cellic® CTec2 – Novozymes e um, obtido em laboratório. As amostras dos complexos enzimáticos foram submetidas à determinação de β-glicosidase, cmcase, fpase e xilanase e à leitura de absorbância na região de 1100 a 2500 nm em espectrofotômetro de infravermelho próximo. A metodologia para determinação de atividade enzimática foi adaptada para microplaca, propiciando uma redução no tempo de análise, uso de reagentes e produção de resíduos. Técnicas de pré-processamento foram utilizadas para remoção de amostras anômalas e melhoria do sinal espectral com o objetivo de gerar uma entrada robusta para os modelos de predição. A determinação da correlação entre os dados espectrais e as determinações analíticas foi realizada por métodos de regressão multivariados: regressão linear múltipla (MLR), regressão por mínimos quadrados parciais (PLS), regressão por componentes principais (PCR), gradiente descendente (GBoost) e quadrados mínimos lineares com e sem *Kernel* (*Kernel-Ridge*, *Ridge*). Todos os métodos foram validados por meio de validação cruzada com dez partições. Os modelos mais promissores resultaram em valores de calibração, validação e predição acima de 90%, demonstrando a possibilidade de se obter determinação direta de atividades enzimáticas de amostras provenientes de processos fermentativos. A determinação da atividade enzimática por meio de análise multivariada com base em espectros NIR se mostrou promissora, pois dispensa o processo de catálise como etapa necessária para a determinação analítica das enzimas mencionadas.

Palavras-chave: aprendizado de máquina, atividade enzimática, celulases, espectroscopia no infravermelho próximo, quimiometria, xilanases.

ABSTRACT

The use of near infrared spectrometry (NIR) combined with multivariate analysis method has made possible the quantitative, direct and non-destructive determination of several inorganic and organic substances in complex samples. In this work we investigated the possibility of direct enzyme determination for xylanases and cellulases through correlation studies between conventional methods for enzyme dosage and multivariate analysis based on near infrared spectra. To this end, three enzyme complexes containing the target enzymes were used, two commercial Celluclast® and Cellic® CTec2 - Novozymes and one obtained in the laboratory. The samples of the enzymatic complexes were submitted to the determination of β-glucosidase, cmcase, fpase and xylanase and the reading of absorbance in the region of 1100 to 2500 nm in near infrared spectrophotometer. The methodology for determining enzyme activity was adapted for microplate, providing a reduction in analysis time, reagent use, and waste production. Pre-processing techniques were used to remove anomalous samples and improve the spectral signal in order to generate a robust input for the prediction models. Correlation determination between spectral data and analytical determinations was performed by multivariate regression methods: multiple linear regression (MLR), partial least squares regression (PLS), principal component regression (PCR), gradient descent (GBoost), and linear least squares with and without Kernel (Kernel-Ridge, Ridge). All methods were validated using cross-validation with ten partitions. The most promising models resulted in calibration, validation and prediction values above 90%, demonstrating the possibility of obtaining direct determination of enzymatic activities of samples from fermentative processes. The determination of the enzymatic activity by means of multivariate analysis based on NIR spectra showed promise, since it dispenses with the catalysis process as a necessary step for the analytical determination of the enzymes mentioned.

Keywords: machine learning, enzyme activity, cellulases, near infrared spectroscopy, chemometrics, xylanases.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ART: Açúcares Redutores Totais

BDA: Ágar Batata Dextrose

CMC: Carboximetilcelulose

CMCase: Carboximetilcelulase

CV: Coeficiente de Variação

DNS: ácido 3,5-dinitrosalicílico

FPU: Unidade de Papel de Filtro

FPase: Celulase Total

GBR: Regressão por Gradiente Descendente

GOD-POD: Glicose-Oxidase-Peroxidase

GPL: Licença pública geral (do inglês *General Public Licence*)

HCA: Análise de agrupamentos hierárquicos (do inglês *Hierarchical Cluster Analysis*)

HPLC, *High Performance Liquid Chromatography* – Cromatografia Líquida de Alta Eficiência

KNN: K-ésimo vizinho mais próximo (do inglês *K-nearest neighbor*)

MLR: Regressão Linear Múltipla (do inglês *Multiple Linear Regression*)

MSC: Correção da dispersão multiplicativa (do inglês *Multiplicative Scatter Correction*)

MSE: erro quadrado médio (do inglês *Mean Square*d *Error*)

NIR: Infravermelho próximo (do inglês *Near InfraRed*)

PCA: Análise de Componentes Principais (do inglês Principal *Components Analysis*)

PCR: Regressão por Componentes Principais (do inglês *Principal Components Regression*)

PLS: Regressão dos mínimos quadrados parciais (do inglês *Partial Least Square*s)

r: Coeficiente de Correlação

R^2: Coeficiente de Determinação

RER: a razão de intervalo de erro (do inglês *range error ratio*)

RMSE: Raiz quadrada do erro quadrático médio (do inglês *Root Mean Square*s *Error*)

RMSEC: Raiz quadrada do erro quadrático médio da calibração

RMSEP: Raiz quadrada do erro quadrático médio da predição

RMSEV: Raiz quadrada do erro quadrático médio da validação

RNA: Redes neurais artificiais

RPD: razão do desempenho do desvio (do inglês *Residual Prediction Deviation*)

RPIQ: taxa de desempenho para intervalo interquartil (do inglês *Ratio of performance to interquartile rang*)

SEP: Erro padrão da predição (do inglês Standard *Error Prediction*)

SNV: Padronização Normal do Sinal (do inglês *Standard Normal Variate*)

VisNIR: Espectroscopia de refletância difusa no infravermelho próximo visível

SUMÁRIO

1 INTRODUÇÃO

O Brasil se destaca pela vasta experiência no uso de biomassas para fins energéticos, em especial no que tange seu aproveitamento para as produções de biodiesel e bioetanol. Segundo Azevedo e Lima (2016), as políticas públicas realizadas no passado e as atuais ações governamentais almejavam garantir a autossuficiência energética do Brasil, além de torná-lo um protagonista no cenário mundial, líder do setor de biocombustíveis. Neste contexto, o etanol de segunda geração, proveniente de materiais lignocelulósicos, tem se tornado relevante como uma alternativa aos combustíveis fósseis (SANTIAGO; RODRIGUES, 2017) e aos tradicionais materiais açucarados e amiláceos empregados na produção de etanol de primeira geração.

A produção de bioetanol de segunda geração possui gargalos relacionados à desestruturação da biomassa lignocelulósica para disponibilização dos açúcares necessários ao processo de fermentação. Essa desestruturação pode ser realizada por meios físicos, químicos ou biológicos, tendo este último o uso das rotas enzimáticas como destaque, como acontece em plantas industriais brasileiras (REZENDE; PASA, 2017). O uso de coquetéis enzimáticos tem sido utilizado em diversos momentos, principalmente, com o objetivo de minimizar os dispendiosos custos dos bioprocessos (FLORENCIO *et al.*, 2017; ARISMENDY *et al.*, 2018, AKRAM *et al.*, 2018). Entretanto, a obtenção desses coquetéis ainda é onerosa, tanto na etapa de produção das enzimas por processos fermentativos, quanto nas etapas analíticas.

Considerando a aplicação e uso de enzimas, a determinação da atividade enzimática se constitui em uma etapa analítica primordial, que se baseia, tradicionalmente, na quantificação indireta dos substratos ou dos produtos envolvidos na catálise, pois a enzima, como entidade molecular, não é alvo direto de análise nessas determinações analíticas. Os métodos de determinação enzimática costumam ser demorados e dispendiosos, principalmente, no que se refere aos reagentes, equipamentos utilizados, tempo e modo de execução do processo analítico. Alguns estudos têm apontado o uso da espectroscopia vibracional combinada à análise multivariada como uma alternativa para o acompanhamento da ação enzimática (BAUM *et al.*, 2013a; CHAKRABORTY *et al.*, 2014; JIN *et al.*, 2017; ARAÚJO *et al.*, 2017).

A espectroscopia no infravermelho próximo é uma técnica analítica bem estabelecida com base na absorção da energia eletromagnética na região de 780 a 2500 nm,

que permite análise de múltiplos componentes em tempo real, gerando resultados confiáveis, sem necessitar muitas etapas analíticas (SILVA *et al.*, 2014, GUTIÉRREZ DEVIA *et al.*, 2015). A espectroscopia de infravermelho próximo combinada com técnicas de análises multivariadas tais como, a regressão parcial por mínimos quadrados ganhou ampla aceitação na área cientifica por ser rápida, não destrutiva, por minimizar o uso de reagentes e, consequentemente, por não gerar derivados químicos. O uso de espectroscopia de infravermelho pode ainda, eliminar as diversas interferências causadas pelo número excessivo de etapas de reações consecutivas, utilizadas para análises quantitativas. Essa técnica tem sido aplicada com sucesso em diferentes campos, incluindo agricultura (especialmente em análise de solos), produção de alimentos, petroquímico, farmacêutico (BAPTISTA *et al.*, 2008; JIN *et al.*, 2020; ZHAO *et al.*, 2015; SKVARIL *et al.*, 2017) e biocombustíveis (CLERCQ *et al.*, 2019).

Os mesmos princípios da espectroscopia vibracional utilizados para a análise das transformações químicas dos substratos a produtos durante a catálise enzimática podem, em tese, ser aplicados para a quantificação direta das enzimas. Neste sentido, surgiu o interesse em realizar a determinação simultânea de diferentes atividades enzimáticas presentes em uma mesma amostra, valendo-se do aspecto molecular, particular de cada enzima.

Considerando que existe um leque de enzimas de interesse na área de biocombustíveis, dentre estas as celulases e xilanases, o presente trabalho investigou o uso de espectroscopia na região do infravermelho próximo, aliado a análises de correlação multiparamétrica como ferramenta para quantificação dessas enzimas. Nestes termos, o presente estudo apresentou os objetivos a seguir.

2 OBJETIVOS

2.1 Objetivo geral

O presente trabalho teve por objetivo aplicar métodos de análise multivariada para determinação simultânea de atividades hidrolíticas em preparados enzimáticos de interesse do setor de biocombustíveis.

2.2 Objetivos específicos

I- Microescalonar métodos de determinação de atividades enzimáticas de β-glicosidases, CMCases, FPases e xilanases, por determinação em microplacas;

II- Obter espectros de varredura na região de infravermelho próximo dos extratos enzimáticos comerciais e dos produzidos por processos de fermentação submersa, em escala laboratorial;

III- Empregar métodos de análise multivariada associados à espectros obtidos por meio de espectroscopia na região do infravermelho próximo para determinar simultaneamente as atividades de enzimas hemicelulolíticas e celulolíticas presentes nos extratos estudados;

IV- Utilizar algoritmos baseados em aprendizado de máquina para construção de modelos de predição da atividade enzimática;

V- Validar os modelos de predição para determinação de atividades enzimáticas hemicelulolíticas e celulolíticas.

3 REFERENCIAL TEÓRICO

3.1 Hidrolases e sua atuação na biomassa lignocelulósica

Existe uma ampla variedade de aplicações biotecnológicas cujos processos envolvem a utilização de hidrolases, biocatalizadores que promovem a cisão de material orgânico com auxílio da água. Algumas hidrolases têm sido alvo de diversas pesquisas em aplicações de biorrefinaria, com o objetivo de reaproveitar os resíduos gerados, principalmente aqueles de origem agrícola (SANTOS *et al.*, 2015; BAMDADA *et al.*, 2018; VALINHAS *et al.*, 2018, FULOP; ECKER, 2020). Esses resíduos são comumente formados de biomassa lignocelulósica, cuja estrutura química interna é complexa e formada, majoritariamente, por biopolímeros de celulose (40-60%), hemicelulose (20-40%) e lignina (15-25%), conforme ilustrado na Figura 1.

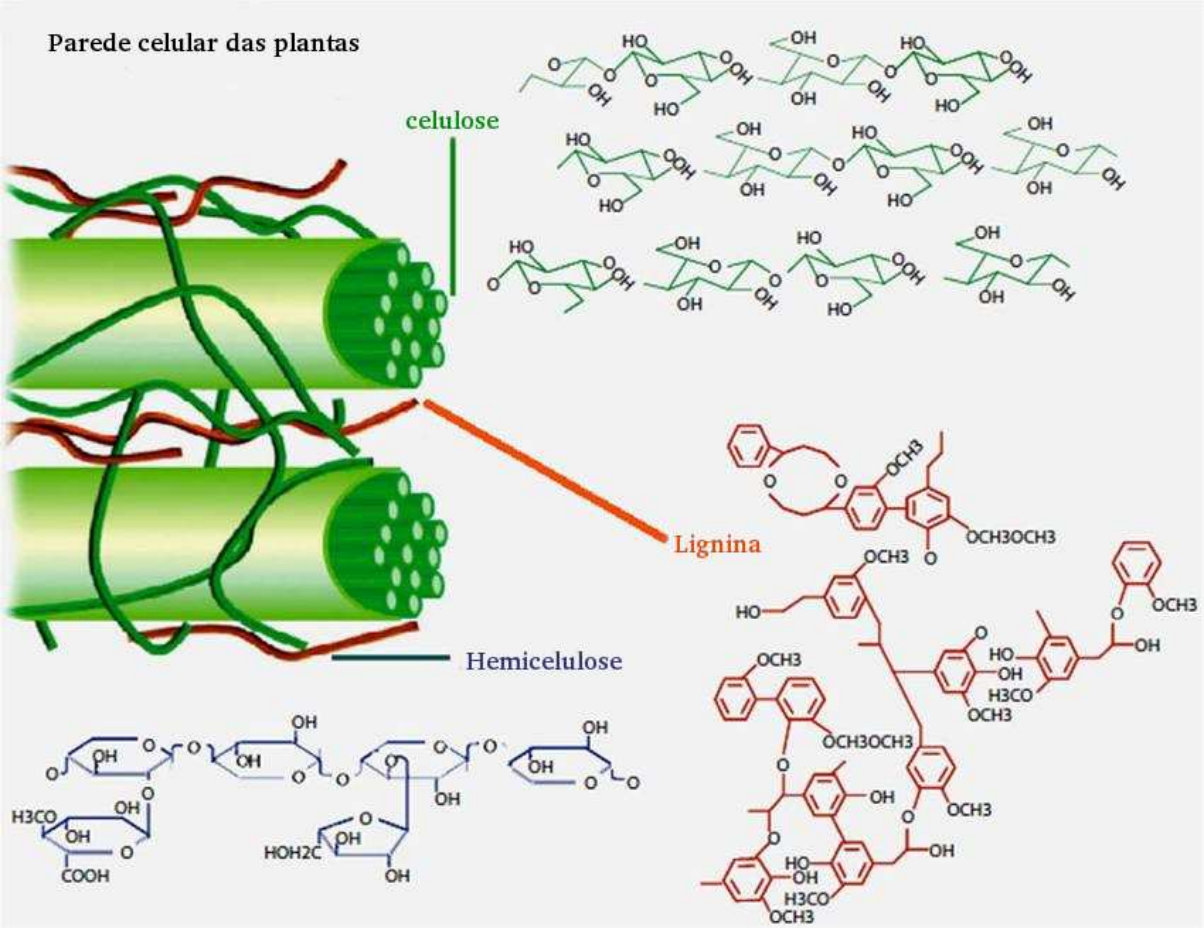

Figura 1 – Esquema representativo da composição lignocelulósica (celulose, hemicelulose e lignina) da biomassa vegetal

Fonte: BAMDADA *et al.*, 2018. Modificado.

A celulose $(C_6H_{10}O_5)_n$ é um polímero constituído por vários monômeros de glicose unidos por ligações β-1,4-glicosídicas, formando cadeias lineares que se unem através de ligações de pontes de hidrogênio entre os grupos hidroxilas das respectivas cadeias, gerando ligações fortes e difíceis de serem rompidas, o que permitem às moléculas se alinharem na forma de microfibrilas para formação da parede celular (OGEDA *et al.*, 2010; PEREIRA *et* al., 2019).

As hemiceluloses estão intimamente ligadas às microfibrilas de celulose e são constituídas de cadeias ramificadas de açúcares heterogêneos como hexoses (glicose, manose, galactose), pentoses (xiloses e arabinoses), além de ácidos urônicos, grupos acetila e outros açúcares (SANTIAGO; RODRIGUES, 2017). Dentre estes açúcares as moléculas de xiloses são predominantes e estão dispostas na macroestrutura da hemicelulose unidas por ligações β-1,4-xilosidásicas, formando estruturas de xilana (MENEZES; BARRETO, 2015). Quanto às ligninas, estas são macromoléculas amorfas de natureza aromática e complexa, difíceis de serem degradadas, sendo, portanto, preteridas em processos biocatalíticos (JIN *et al.*, 2017). As ligninas funcionam como cimento para a parede celular das plantas, conferindo a estas, maior rigidez e resistência; e se encontram intercaladas entre as microfibrilas de celulose e hemicelulose (CARVALHO *et al.*, 2009).

O aproveitamento integral da biomassa lignocelulósica aplicada à produção de bioetanol só é possível após a hidrólise das porções de celulose e hemicelulose, por disponibilizar monossacarídeos provenientes destas estruturas poliméricas, que podem ser convertidas a etanol por meio de bioprocessos. A hidrólise das frações de celulose e hemicelulose pode ser realizada por intermédio de métodos químicos ou biológicos. Os métodos biológicos podem utilizar complexos enzimáticos constituídos das seguintes hidrolases: celulases e xilanases. Essas enzimas são produzidas e secretadas por micro-organismos diversos dentre os quais se tem bactérias, leveduras e fungos filamentosos (FLORENCIO *et al.*, 2017).

As celulases promovem a quebra das moléculas de celulose e são constituídas por no mínimo três subgrupos de enzimas que trabalham de maneira sinérgica: as endo-β-1,4-glucanases, também denominadas de endoglucanases; as exoglucanases, denominadas de exo-β-1,4-glucanases ou celobiohidrolases e as ß-glicosidases (SARSAYA *et al.*, 2018). Na

Figura 2 encontra-se um esquema representativo da desestruturação da celulose por esse grupo de enzimas.

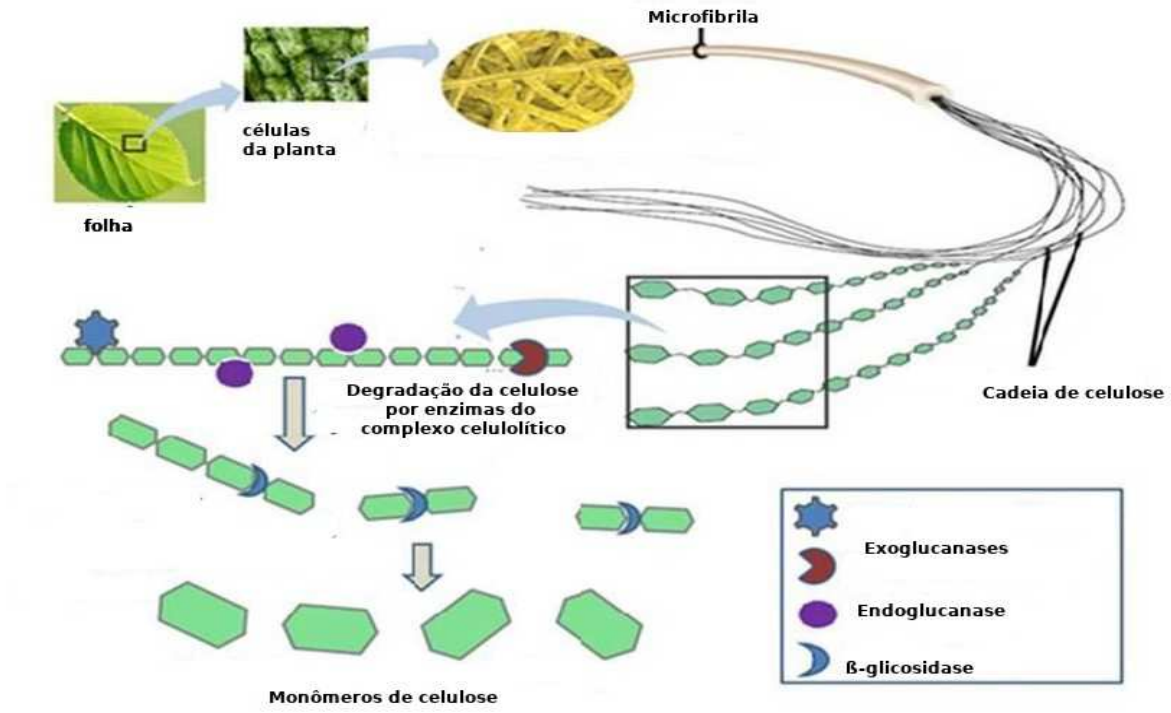

Figura 2 – Esquema representativo da ação do complexo enzimático celulolítico atuando na desestruturação da biomassa lignocelulósica

Fonte: SINGHI *et al.*, 2016. Modificado.

As endoglucanases, com destaque para carboximetilcelulase (CMCase), clivam randomicamente as ligações internas da fibra celulósica, disponibilizando oligossacarídeos com terminações redutoras e não redutoras, as quais ficam disponíveis para atuação das exoglucanases (SELVAN *et al.*, 2017). As endoglucanases, portanto, são enzimas responsáveis pela redução do grau de polimerização da celulose. As exoglucanases (celobiohidrolases e glicohidrolases) atuam na região externa das cadeias de celulose, liberando dissacarídeos de celobiose ou resíduos de glicose das extremidades redutoras e não redutoras e, então, as ß-glicosidases hidrolisam a celobiose e outros oligossacarídeos solúveis à glicose (SARSAYA *et al.*, 2018).

As hemicelulases denominadas xilanases promovem a degradação da xilana, estrutura hegemônica da porção hemicelulósica vegetal, e, semelhante às celulases, são

constituídas de diferentes enzimas que agem de forma sinérgica no processo de hidrólise da parede celular (MENEZES; BARRETO, 2015). Essas enzimas são formadas, predominantemente, de endo 1,4-β-xilanases e β-xilosidases e clivam ligações glicosídicas da cadeia principal da xilana resultando em xilo-oligossacarídeos com diferentes comprimentos, enquanto as endoxilanases degradam esses oligossacarídeos liberando xilose como produto (OGEDA; PETRI, 2010).

Diversas pesquisas têm sido desenvolvidas entorno da produção ou uso de enzimas celulolíticas e hemicelulolíticas com propósitos de contribuir para área de biocombustíveis (SANTOS; GOUVEIA, 2009; SANTOS *et al.*, 2015, LYND *et al.*, 2017). Nesse contexto, Florencio *et al.*(2017) relatam que para a produção de bioetanol de segunda geração, por exemplo, existem algumas limitações tecno econômicas como o alto custo das enzimas utilizadas na conversão da biomassa a açúcares fermentescíveis que podem inviabilizar o processo. Esses autores acrescentam, ainda, que processos mais eficientes de produção destas enzimas precisam ser desenvolvidos para que a viabilidade desse biocombustível seja possível. Para tanto, sugerem a produção de enzimas *on-site* por meio do cultivo de microrganismos potencialmente produtores de hemicelulases, utilizando como fonte de carbono resíduos lignocelulósicos.

Na direção apresentada pelos autores supracitados, Santos *et al.* (2015) conduziram um processo fermentativo, utilizando como substrato os bagaços de algodão, girassol e macaúba, tendo como agente fermentador o fungo *Aspergillus tubingensis*, e concluíram que uma linhagem específica do fungo estudado apresentou alta produção de enzimas celulolíticas e xilanolíticas, com alta expressividade destas últimas. Além disso, identificaram os resíduos de algodão como os substratos que mais induziram a produção destas enzimas. Os autores conseguiram reduzir os custos da produção das enzimas através da utilização de substratos residuais da indústria produtora de óleos vegetais.

Lynt *et al.* (2017), seguindo o propósito de produção enzimática, identificaram bactérias termofílicas do gênero *Clostridium* como melhores agentes fermentadores do que os fungos para produção de etanol. Os autores concluíram que tais bactérias são mais eficientes na desconstrução da biomassa lignocelulósica e que podem, portanto, serem utilizadas para geração de complexos enzimáticos úteis à produção de etanol celulósico.

A produção enzimática representa um dos gargalos na produção de bioetanol. Outro fator importante a ser avaliado é a eficiência de degradação da biomassa por parte dos complexos enzimático produzidos. Para avaliar essa eficiência é necessário realizar a determinação de atividade enzimática frente a cada substrato escolhido. Com foco em alguns desses gargalos, Baum *et al.* (2012, 2013) e Rodrigues *et al.* (2015) propuseram metodologias de acompanhamento de hidrólise enzimática para desconstrução de biomassas para bioenergia e Camassola e Dilon (2012) e Mansour *et al.* (2016) realizam uma melhoria de metologias tradicionais empregadas para a determinação da atividade enzimática dessas hidrolases.

3.2 Métodos analíticos para determinação enzimática

As metodologias tradicionais de determinações enzimáticas de celulases e xilanases compiladas e apresentadas por Ghose (1987) e Ghose e Bisaria (1987) se baseiam em ensaios de catálise enzimática envolvendo substratos específicos para atuação de cada conjunto de enzimas do complexo celulolítico.

As metodologias para determinação das celulases envolvem a quantificação das endoglucanases, celulases totais e β-glicosidases através da catálise envolvendo os substratos carboximetilcelulose (CMC), papel de filtro e celobiose. O substrato utilizado para quantificação das xilanases são aqueles com maior concentração de xilana. Ao finalizar o processo de catálise, os açúcares redutores totais (ART) resultantes dos ensaios são quantificados por métodos colorimétricos através de espectroscopia, utilizando faixas de absorbância específicas para cada conjunto de enzimas.

O doseamento de ART é geralmente realizado segundo métodos originalmente descritos por Bernfeld (1955) e Miller (1959), utilizando como reativo o ácido 3,5-dinitrossalicílico (DNS). O total de atividade enzimática encontrado é expresso em Unidade por mililitro (U mL^{-1}), em que uma unidade representa o total de enzima necessário para liberar 1 μmol de ART, por minuto, por volume (mililitros) de extrato enzimático.

As metodologias tradicionais de determinação enzimática envolvem a utilização de uma quantidade expressiva de reagentes e vidrarias tanto no momento da catálise quanto na quantificação dos produtos da reação. Os experimentos são realizados em tubos de ensaio com volumes de ‘substrato + enzima’ entorno de 1 mL, seguidos do acréscimo de reagentes

no momento da determinação dos açúcares redutores. Muitos desses reagentes são tóxicos, como é o caso de soluções de DNS, produzidas com adição de fenol, o qual é considerado um produto químico nocivo por inalação, ingestão e contato com pele e olhos. O fenol é considerado um composto de alta toxicidade mesmo em baixas concentrações e é destacado como poluente prioritário, o qual deve ter prioridade de remoção em processos de sanitização e limpeza de cursos d'água. Além disso, as determinações enzimáticas envolvem várias etapas e repetições, aumentando tanto o percentual de erros inerentes às manipulações como o total de resíduo tóxico produzido.

A metodologia, prioritariamente, proposta por Ghose (1987) para a determinação de celulases totais através da quantificação de unidades de papel de filtro (FPU), é utilizada nos moldes tradicionais em pesquisas que envolvem a produção de etanol de segunda geração, utilizando métodos enzimáticos como pré-tratamento para disponibilização dos açúcares fermentescíveis (INFORSATO; PORTO, 2016**)**. Contudo, inovações surgiram com o objetivo de minimizar os custos destas determinações bem como utilizar reagentes com menor toxidade.

Vasconcelos *et al.* (2013), por exemplo, realizaram um estudo sistemático sobre as metodologias de determinação de ART e relataram que

> (...) os trabalhos de Bernfeld (1955) e Miller (1959) sobre o uso do DNS para a determinação de açúcares redutores tornaram-se clássicos na literatura científica, tendo sido citados, até final de junho de 2012, respectivamente por 3.332 e 8.174 artigos listados na base *Web of Science*. (VASCONCELOS *et al.*, 2013)

Os autores, desenvolveram um protocolo bem estruturado para a determinação dos açúcares redutores pelo método DNS, seguindo o método descrito por Bernfeld (1955), sem utilizar fenol e bissulfito de sódio, ou seja, uma solução com toxidade reduzida, com a modificação no tempo de reação introduzida por Miller (1959), a qual equivale a realizar a reação por 15 minutos em água fervente, o que antes era realizado em 5 minutos. Além de promoverem uma redução nos volumes reacionais entorno de 50%.

Diferentes autores propuseram a adaptação da metodologia com base em DNS da macro escala para a microescala através do uso de microplacas de titulação de 96 poços (KING *et al*, 2009; GONÇALVES *et al.*, 2010; NEGRULESCU *et al.*, 2012; SANTOS *et al.*, 2017; SHAO; LIN, 2018) com o objetivo de reduzir ainda mais os custos e o tempo

empregados nos ensaios para determinações de ART.

A maioria dessas abordagens utilizam duas microplacas em cada ensaio, sendo a primeira para realizar a reação de catálise que resulta na obtenção de ART e a segunda para promover a reação colorimétrica entre o reagente DNS e a amostra contendo os ART de interesse. Em função do volume de DNS utilizado e da necessidade de realizar diluição para posterior análise, esta segunda microplaca é utilizada para tal empreendimento. Gonçalves *et al.* (2010) e Santos *et al.* (2017) conseguiram produzir ensaios mais econômicos, utilizando apenas uma microplaca e um valor mais reduzido do reagente (25µL de DNS), diferindo basicamente no tempo definido para conclusão da reação, que para os primeiros autores foi de 10 minutos e para os seguintes 5 minutos.

Santos *et al.* (2017), objetivando validar a abordagem proposta, comparam os resultados obtidos em microplaca com os obtidos via metodologia em cromatografia líquida de alta eficiência (HPLC, *High Performance Liquid Chromatography*). Para tanto, utilizaram 60 amostras, obtidas de cultivos laboratoriais de leveduras vinícolas, e comprovaram que o teor de ART determinados por microplaca e por HPLC diferiram em menos de 10% dos casos.

Concomitante às investigações para melhoria das determinações de ART, pesquisas, também, foram realizadas com o objetivo de diminuir os custos e tempo de ensaios envolvendo as determinações enzimáticas de celulases e xilanases (CAMASSOLA; DILLON, 2012; LUCENA *et al.*, 2013; CHAKRABORTY *et al.*, 2014; MANSOUR *et al.*, 2016; Nascimento *et al.*, 2017).

Camassola e Dillon (2012) realizaram uma adaptação do tradicional método de determinação de celulases totais, utilizando papel de filtro *Whatman* n° 1. Os autores sugeriram que a metodologia proposta, utilizando microplacas de 96 poços, poderia ser executada de maneira mais rápida e menos trabalhosa, além de produzir menos poluentes.

Lucena *et al.* (2013) foram além da microplaca e adaptaram os protocolos de determinações de açúcares redutores com DNS em ensaios enzimáticos de hidrolases como amilases e β-1,3-glucanases, utilizando termocicladores e obtiveram uma redução de 10 vezes na quantidade de reagente e no volume da amostra necessária quando comparada aos protocolos convencionais de microplacas. Os autores conseguiram realizar uma padronização das leituras de absorbância e o uso dos termocicladores resultou em calibrações de

temperatura menos demoradas e sem perda de volume por vazamento ou evaporação da microplaca. Parâmetros cinéticos foram obtidos com sucesso, e o uso do termociclador permitiu a mensuração da atividade enzimática em amostras biológicas de campo com quantidade limitada de proteína.

Chakraborty *et al.* (2014) investigaram a viabilidade do uso de espectroscopia de refletância difusa no infravermelho próximo visível (VisNIR) como um método fácil, barato e rápido com o objetivo de prever a atividade enzimática de compostos que tradicionalmente é determinada em ensaios de hidrólise utilizando diacetato de fluoresceína. Para tanto, os autores obtiveram amostras de cinco instalações de compostagem e após gerar os espectros dessas amostras utilizaram seis algoritmos multivariados distintos para realizar a previsão. Os autores identificaram que algoritmos baseados em redes neurais resultaram em melhores desempenhos por conseguir capturar as relações altamente não lineares entre a atividade enzimática do composto e os espectros gerados por refletância VisNIR. Utilizaram como pré-tratamento o alisamento de Savitzky – Golay e conseguiram demonstrar a eficiência do VisNIR DRS para predizer a atividade enzimática e microbiana do composto analisado.

Propostas mais avançadas surgiram com o objetivo de aumentar ainda mais a velocidade das determinações analíticas e não necessitar realizar o processo de catálise para as determinações enzimáticas. Para tanto, diversos autores propuseram o estudo dos resultados de espectros gerados por espectroscopia difusa ou no infravermelho próximo com o objetivo de gerar modelos multivariados para acompanhamento da ação enzimática (CÂMARA *et al*., 2014; KLIMKIEWICZ *et al.*, 2014; GUTIERREZ *et al.*, 2016; NASCIMENTO, 2016).

Baum *et al.* (2013) utilizaram um método analítico rápido e não destrutivo para a determinação da ação enzimática no farelo de milho submetido a diferentes pré-tratamentos. Para tanto, esses autores, recorreram à análise dos açúcares liberados após hidrólise enzimática por espectroscopia de reflectância difusa associada à quantificação por cromatografia líquida de alta eficiência HPLC. Associando essas duas determinações, os autores construíram, ainda, modelos de calibração multivariados incluindo regressão por mínimos quadrados parciais, os quais possibilitaram prever a liberação enzimática de diferentes níveis de arabinose, xilose e glicose de todas as amostras de farelo de milho estudadas.

A análise de amostras por espectroscopia é uma técnica de medição de comprimento de onda e intensidade de absorção de energia eletromagnética que tem por finalidade identificar a composição química da amostra, associando um espectro eletromagnético a um composto específico. Esse método de análise é consolidado, facilmente implementável e possui abrangência de aplicação, exatidão, custo reduzido e sensibilidade comparável a metodologias analíticas (TIBOLA *et al.*, 2018).

Segundo Lima e Bakker (2011) o infravermelho próximo representa a região do espectro eletromagnético imediatamente superior à região visível em termos de comprimento de onda, isto é, representa a região do infravermelho "mais próxima" da região visível (Figura 3).

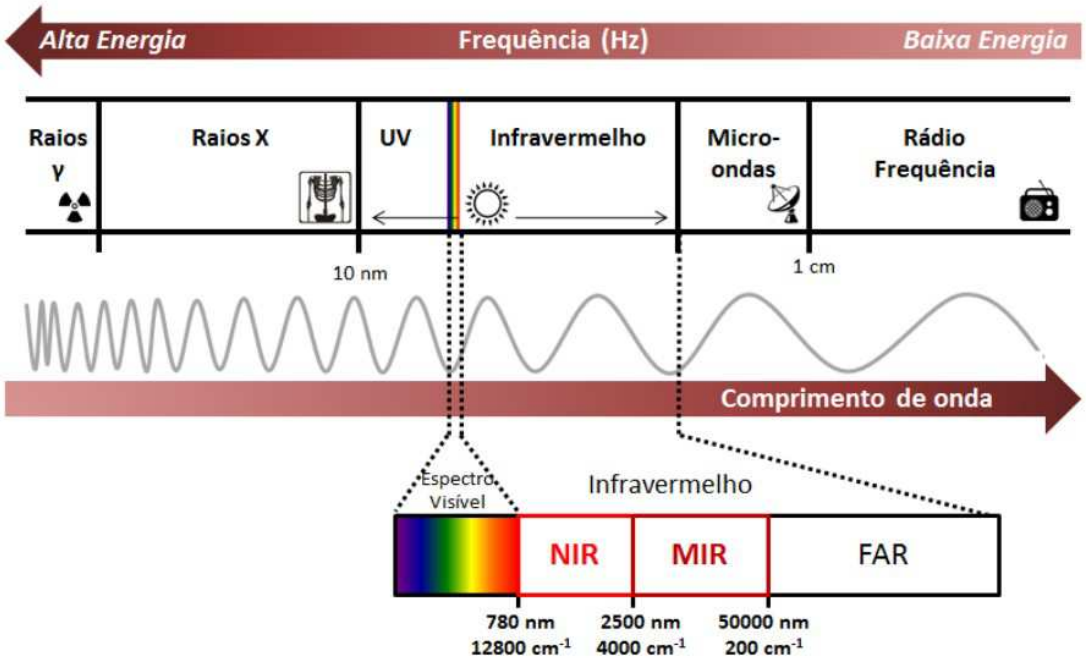

Figura 3 – Espectro eletromagnético com destaque para o infravermelho próximo

Fonte: SILVA, 2017.

A espectroscopia de infravermelho próximo (do inglês: *Near Infrared - NIR*) compreende a região do espectro eletromagnético associada aos comprimentos de onda que variam de 780 nm a 2500 nm (LEITE *et al.*, 2012). Esta faixa do espectro é tipicamente usada na determinação quantitativa de grupos funcionais orgânicos, especialmente O-H, N-H, e

C=O, cujos limites de detecção são normalmente 0,1%, podendo ser utilizada para determinações de amostras nos seus variados estados: sólido, líquido ou gasoso, com o mínimo ou nenhum preparo da amostra (SILVA *et al.*, 2014).

A espectroscopia NIR se tornou uma métrica efetiva em análise de qualidade de produtos farmacêuticos, alimentícios e na quantificação de produtos da agroindústria, bem como de bioprocessos (KUMAR *et al.*, 2010; CUNHA *et al.*, 2017; JIN *et al.*, 2017; SKVARIL *et al.*, 2017), principalmente, no que se refere à determinação da qualidade e caracterização do biodiesel.

A utilização de espectroscopia NIR, associada a algoritmos de calibração multivariada têm sido utilizadas para determinação de qualidade do biodiesel relacionando espectros de amostras de biodiesel degradado com o seu índice de acidez (OLIVEIRA, 2008). Trabalhos como o proposto por Cunha *et al.* (2017) utilizaram espectros de amostras de biodiesel produzido de diferentes fontes oleaginosas, juntamente com amostras comerciais, para predizer a qualidade do biodiesel em função de características como densidade, índice de refração e ponto de entupimento de filtro e biodiesel.

Utilizando algoritmos de calibração semelhantes, juntamente com espectros derivados de amostras de biocombustível sólido, Skavaril *et al.* (2017) demonstraram que a espectroscopia NIR pode ser eficientemente utilizada para quantificar simultaneamente propriedades diversas como teor de cinzas, umidade e taxa de aquecimento.

Neste sentido a espectroscopia NIR se mostra como uma ferramenta interessante a ser aplicada na determinação indireta do teor de celulases e xilanases de complexos enzimáticos. Essa ferramenta já vem sendo utilizada com propósito de acompanhar os resultados da ação dessas enzimas durante processo fermentativos (BAUM *et al.*, 2013a), no entanto, ainda não é empregada para predição da atividade enzimática a partir de amostras de extratos enzimáticos.

3.3 Análise multivariada

A análise multivariada é uma técnica muito utilizada na quimiometria a qual delimita qualquer abordagem analítica que permite analisar simultaneamente o comportamento de duas ou mais variáveis em uma única amostra (SOUZA; FERRÃO, 2006). Essa técnica utiliza métodos matemáticos, normalmente aplicados através de modelos de

regressão, os quais relacionam o comportamento linear ou não linear de uma variável Y com uma variável X (FERREIRA *et al.*, 1999), ou com um conjunto de N variáveis $X_{i..N}$.

De acordo com Baum *et al.* (2013) a relação entre Y e o conjunto de variáveis X pode ser utilizado para prever parâmetros analíticos como, por exemplo, o teor de atividade enzimática em uma reação de hidrólise de material lignocelulósico, bem como, determinar o teor de enzimas produzidos durante um processo fermentativo.

Na análise multivariada, dados espectrais são associados a outras determinações analíticas e organizados de forma matricial, resultando em duas matrizes distintas X e Y. A matriz X se refere às amostras espectrais, onde cada elemento da matriz contém o resultado analítico dado pelo espectrofotômetro (absorbância, transmitância ou reflectância) e Y se refere aos resultados analíticos determinados por algum método de referência (CHAKRABORTY *et al.*, 2014). Os dados matriciais são úteis quando se deseja realizar predição ou agrupamento, aplicando técnicas de mineração de dados, as quais extraem informações úteis de um conjunto expressivo de dados.

De acordo com Faceli *et al.* (2011) a técnica de mineração de dados ou de análise multivariada compreende no mínimo três etapas de processamento: a calibração ou treinamento, validação e teste, as quais são precedidas por uma etapa de pré-processamento e transformação dos dados, que aumenta a qualidade e expressividade do conjunto de dados que será minerado. Os autores reportam, ainda, que o pré-processamento pode envolver etapas de limpeza para remoção de dados inconsistentes, incompletos, redundantes ou ruidosos e que a transformação dos dados contribui para a melhoria do desempenho dos algoritmos de predição, que muitas vezes são influenciados pelo excesso de variáveis ou variabilidade dos dados.

Algoritmos como Análise de Componentes Principais (PCA - *Principal Component Analises*) e Análise de Agrupamentos Hierárquicos (HCA - *Hierarchical Cluster Analysis*) são muito utilizados na etapa de redução de dimensionalidade, e possibilitam a determinação de identidade ou diferenciação das amostras, enquanto, métodos baseados em distância, como o algoritmo *K-Nearest Neighbor* (KNN), ou seja, K-ésimo “vizinho mais próximo”, utilizam a proximidade dos dados para classificar uma nova amostra com base nos exemplos do conjunto de dados de treinamento (FACELI *et al.*, 2011).

O algoritmo KNN é considerado um dos métodos preditivos mais simplificados que existem pois não é definido a partir de premissas matemáticas mas apenas utiliza do conceito de distância entre amostras e da relação de similaridade entre as mesmas, considerando que amostras que possuem distâncias próximas sejam similares (GRUS, 2016). Normalmente essa distância pode ser calculada pela distância Euclidiana. O algoritmo KNN consegue prever uma métrica alvo para um novo objeto com base nos exemplos do conjunto de treinamento que são próximos a ele. Funciona, basicamente, memorizando os objetos de treinamento e, portanto, não generaliza um modelo sucinto para os dados. Pode ser utilizado tanto em problemas de classificação quanto de regressão. Esse tipo de algoritmo tem seu desempenho afetado pela medida de distância escolhida além de ser influenciado por variáveis com valores de atributos em escalas muito díspares. Portanto, para utilização desse tipo de algoritmo é necessário realizar uma etapa de pré-processamento que permita normalizar os dados.

Além do algoritmo KNN, diversos outros algoritmos podem ser utilizados nas etapas de calibração, validação e predição multivariada. Dentre estes, os mais utilizados em quimiometria são: regressão linear múltipla (MLR – *Multiple Linear Regression*), regressão por componentes principais (PCR – *Principal Component Regression*), a regressão por mínimos quadrados parciais (PLS – *Partial Linear Regression*); máquina de vetor de suporte (SVM – Support Vector Machine) e redes neurais artificiais (RNA) (FICHER *et al.*, 2017, BRAGA *et al.*, 2011). Novos algoritmos têm surgido com propósitos semelhantes como regressão por *kernel*, *ridge*, gradiente descendente, dentre outros.

O algoritmo por regressão linear multivariada (MLR) procura encontrar uma equação matemática, representada graficamente por uma linha de regressão em um espaço n-dimensional no qual uma resposta y é dada por uma função de k variáveis, em que cada variável k representa uma relação linear com a variável resposta y. A resposta y que se busca é dada pela equação: $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + e$, em que e equivale ao erro do modelo.

A Análise de Componentes Principais (PCA) é considerada o algoritmo mais popular aplicado à redução de dimensionalidade (GÉRON, 2019). Esse algoritmo encontra um hiperplano que mais se aproxima dos dados analisados e realiza uma projeção dos mesmos sobre esse hiperplano, preservando a maior quantidade de variância no conjunto de

treinamento dos dados. Uma informação importante determinada pelo PCA é a taxa de variância explicada de cada componente principal. A soma das taxas de variância de cada componente leva a decidir quantas componentes são necessárias para se obter variância máxima (que se aproxima de 100%) do conjunto de dados. A escolha do número correto de dimensões, portanto, não é arbitrária, mas sim determinada pelas dimensões que adicionam a maior variância. A regressão por componentes principais (PCR) é realizada em duas etapas: inicialmente é realizada a redução da dimensionalidade através do PCA e em seguida é aplicado um algoritmo de regressão linear.

A regressão por mínimos quadrados parciais (PLS) visa relacionar uma ou mais variáveis resposta y com variáveis independentes x, baseada no uso de fatores (ou componentes). No contexto desse trabalho, uma matriz X seria formada por todos os valores de absorbância nos diversos comprimentos de onda, variando de 1100 nm a 2500 nm, com intervalos de 2 nm, e uma matriz Y, constituída de todos os valores de determinações enzimáticas. O algoritmo PLS consegue identificar combinações lineares das variáveis x que melhor modelam as variáveis y. O método PLS vem sendo utilizado em diversos trabalhos na área de biocombustíveis.

O método de regressão com base em gradiente descendente tem como fundamento um algoritmo de otimização genérico que consegue buscar soluções ótimas para variados problemas. Esse algoritmo realiza um ajuste interativo dos seus parâmetros com o intuito de minimizar uma função de custo. O que esse algoritmo faz é medir o gradiente local da função de erro em relação a um vetor de parâmetros, indo em direção ao gradiente descendente zero ou mínimo, melhorando gradualmente à medida que diminui a função de custo. Outra medida importante para esse algoritmo é a taxa de aprendizado, sendo que esta nem pode ser muito pequena (pois demoraria muito a atingir o custo mínimo), nem pode ser muito alta para que não se ultrapasse o vale da função de custo mínimo.

A escolha de um método adequado de regressão para a modelagem é importante, todavia para que os modelos de predição gerados sejam robustos estes necessitam apresentar uma boa relação entre as medidas *bias* e a variância (BELLON-MAUREL *et al*., 2010), métricas que auxiliarão na verificação se a técnica utilizada está gerando um sub ou super ajustamento, permitindo que seja avaliado o seu poder de generalização.

O viés, também denominado erro sistemático, quando elevado demonstra que a

técnica utilizada não consegue prever valores alvo que se aproximam dos valores reais. A variância, denominada erro aleatório, permite identificar diferenças sutis entre as amostras. Para melhorar a relação *bias*-variância o número de amostras deve ser aumentado, mas em casos em que essa estratégia não seja possível, e para evitar que ocorra *overfitting*, pode-se utilizar algoritmos de aprendizado que utilizem regularização, que pode ser encontrada nos modelos *Ridge* e *Kernel Ridge*. A regularização funciona como a inserção de *bias* no modelo, evitando o excesso de ajuste nos dados com o objetivo de diminuir a variância.

Independente do método de regressão escolhido, para construção dos modelos preditivos, cada método é aplicada sob um conjunto de dados que tenha sido previamente subdividido de maneira que as amostras que foram utilizadas na etapa de calibração não sejam utilizadas na validação, e as amostras de teste não sejam utilizadas nas duas etapas precedentes. A subdivisão do conjunto de dados é necessária para evitar a geração de modelos otimistas ou pessimistas e pode ser determinada por diferentes metodologias de amostragem como aleatória, validação cruzada e *bootstrap* (FACELI *et al.*, 2011). Dentre estas metodologias, a validação cruzada é uma das técnicas mais utilizadas e consiste em particionar o conjunto total de dados em n partições de tamanho aproximado, em que um subconjunto será utilizado para teste e as amostras das n-1 partições restantes são usadas no treinamento do modelo. Este processo é repetido n vezes e, em cada ciclo, uma partição diferente será utilizada para teste. O desempenho final do modelo é calculado em função da média de desempenho de cada ciclo realizado.

A avaliação dos algoritmos de regressão normalmente é realizada por meio da análise de desempenho do modelo durante o processo de predição sobre novas amostras, que não foram submetidas previamente às etapas de calibração ou treinamento. O erro da hipótese de predição pode ser calculado pela distância entre o valor conhecido (y_i) e o predito pelo modelo ($\hat{y}_i$.). A medida mais utilizada para calcular este erro é o “erro quadrático médio” (MSE – *Mean Squared Error*) ou uma derivação deste, a “raiz do erro quadrático médio” (RMSE – *Root Mean Squared Error*). Esse último acrescido das letras C, V ou P (RMSEC, RMSEV, RMSEP) se refere à “raiz do erro quadrático médio” na calibração, validação e predição, respectivamente. A obtenção de valores reduzidos destas medidas representam melhores modelos, ou seja, as predições de cada modelo se aproximam dos padrões verdadeiros.

Considerando as amostras do conjunto de teste e o valor do erro RMSEP obtido a partir do modelo induzido, o dobro do valor determinado para RMSEP representa um intervalo de confiança de 95% do valor predito em relação ao valor conhecido. Por exemplo, se o modelo presume para uma amostra do conjunto de teste um valor de atividade enzimática para xilanase no valor de 100 U mL^{-1} e o RMSEP é de 5 U mL^{-1} , então há 95% de probabilidade de que a atividade enzimática determinada para essa amostra, através do método laboratorial em microplaca, esteja compreendido entre 95 e 105 U mL^{-1}.

Os seguintes parâmetros são utilizados como métricas auxiliares para escolha dos modelos mais eficientes: coeficiente de determinação (R^2); o número de componentes principais, fatores ou variáveis latentes para alguns modelos; viés (*bias*); a razão do desempenho do desvio (RPD, *Residual Prediction Deviation*); o erro padrão da performance (SEP, *Standard Error Prediction*); a razão de intervalo de erro (RER) e a taxa de desempenho para intervalo interquartil RPIQ (*Ratio of performance to interquartile rang*). A seguir são apresentadas as definições para cada uma dessas métricas.

- R^2: a métrica R^2 representa o coeficiente de determinação múltipla (no caso de regressão múltipla), é uma medida estatística que descreve o ajuste dos dados ao modelo estatístico escolhido ou à linha de regressão determinada para este modelo. Valores de R^2 variam de 0 a 1, sendo que os valores próximos a 1 estão superajustados, enquanto valores próximos a zero não apresentam nenhum ajuste. Um modelo é considerado 100% preciso quando $R^2 = 1$, ou seja, todas as amostras apresentadas estão ajustadas à linha de regressão do modelo.

- *Bias*: é definido como a diferença média entre o valor predito a partir do espectro e o valor real do atributo alvo. Um *bias* positivo significa que, em média, o modelo está superestimando, enquanto um valor negativo representa uma subestimação. É desejado que o *bias* esteja o mais próximo de zero.

- SEP: erro padrão de predição - O SEP ao quadrado é aproximadamente igual ao RMSEP ao quadrado menos o Viés ao quadrado. Portanto, se o bias for baixo, os valores de RMSEP e SEP serão semelhantes.

- RPD: relação do erro padrão do desempenho em relação ao desvio padrão. O RPD é calculado dividindo-se o valor do desvio padrão do atributo alvo pelo valor de RMSE. Enquanto o RMSEP, SEP e *Bias* usam as mesmas unidades de medição, os valores de R^2,

RPD e RER são adimensionais, o que significa que eles podem ser comparados se calculados para o mesmo conjunto de dados entre modelos multivariados distintos e para determinações também variadas. (FEARN, 2002). Segundo Fearn (2002) “Se o RPD for igual a um, então o SEP é igual ao desvio padrão dos dados de referência, o que significa que o modelo não está prevendo os valores de referência. Valores mais altos para o RPD sugere modelos cada vez mais precisos” (FEARN, 2002).

- RER: a razão de intervalo de erro equivale ao quociente entre o intervalo predito (ou seja, o valor máximo menos o valor mínimo) e o valor de SEP. Os valores calculados para RER serão, geralmente, de quatro a cinco vezes maiores do que para a métrica RPD e a relação entre estas métricas depende da distribuição das amostras no conjunto de teste (RAMBO *et al*., 2018). Os seguintes limites são considerados para as métricas RPD e RER:

- RER> 4 e RER <=10, significa que o modelo de calibração induzido é considerado aceitável;

- RER>10 e RER <=15, significa que o modelo de calibração induzido é aceitável para controle de qualidade

- RER >15, significa que o modelo de calibração é bom para quantificar analitos.

O valor de RER é considerado um teste de melhor qualidade para indução de modelos, contudo sofre bastante interferência da presença de *outlier*s.

Apesar de muitos autores utilizarem as métricas RPD e RER para avaliar modelos preditivos, Minasny e McBratney (2013) sugerem a utilização apenas de R^2 e RPIQ pelas seguintes razões: R^2 e RPD são consideradas métricas equivalentes, portanto uma ou outra pode ser utilizada para avaliação dos modelos preditivos. RPD equivale a $(1-R^2)^{-0.5}$. Pode-se afirmar, portanto, que se $R^2 > 0,75$ equivale a afirmar que RPD > 2. Esses valores indicam que o modelo apresenta um bom ajuste. Diferentemente, caso $R^2 <0.5$ equivale a RPD < 1.4. Neste caso o modelo não prevê de forma satisfatória.

4 MATERIAL E MÉTODOS

Os complexos enzimáticos utilizados no presente trabalho foram provenientes de empresas produtoras de enzimas (comercial) e de ensaios fermentativos realizados no laboratório de pesquisa em biorreator.

O substrato utilizado no processo fermentativo foi a torta de caroço de algodão, submetida à moagem em moinho de facas (Moinho tipo *Willey*). A torta de algodão foi obtida da Indústria de Óleo, Rações e Plásticos Montes Claros LTDA, estado de Minas Gerais. A linhagem de *Aspergillus tubigensis* AN1257, utilizada como agente do bioprocesso, foi proveniente do banco de culturas do Laboratório de Bioprocessos e Biotransformação da UFVJM. A linhagem foi mantida por cultivos periódicos, pela técnica de repique tubo a tubo, em meios de cultura ágar batata dextrose (BDA), conservados sob refrigeração a 4±1°C.

No fluxograma apresentado na Figura 4 encontra-se uma visão global da metodologia desenvolvida no presente trabalho, a qual se constituiu em quatro etapas: obtenção de extratos enzimáticos; construção da base de dados para modelagem computacional por intermédio da determinação de atividades enzimáticas e geração de espectros NIR; pré-processamento dos dados e construção dos modelos de predição.

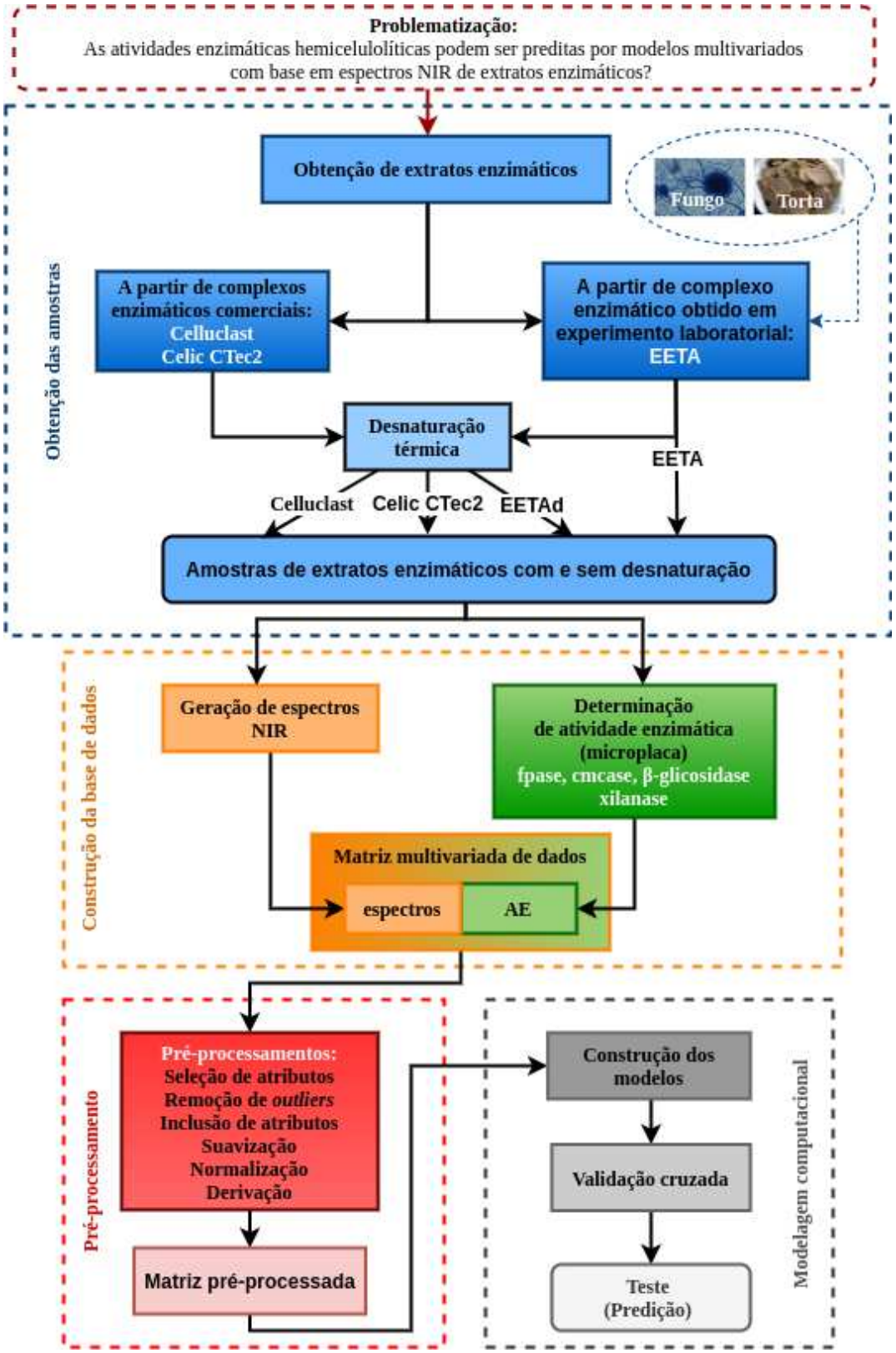

Figura 4 – Visão geral das etapas utilizadas no estudo para modelagem computacional preditiva de atividades enzimáticas de β-glicosidase, CMCases, FPases, e xilanases.

AE: Atividade Enzimática; NIR: Infravermelho Próximo

As etapas definidas na Figura 4 encontram-se detalhadas a seguir.

4.1 Obtenção dos extratos enzimáticos

Os complexos comerciais Celluclast® e Cellic®CTec2 da Novozymes A/S (Basgsvaerd, Dianamarca), com o objetivo de produzir uma variabilidade de amostras, foram submetidos ao processo de desnaturação térmica através de ensaios com variação de temperatura de 60°C a 100°C, por intervalos de tempos distintos, definidos experimentalmente. Em cada ensaio, alíquotas de 2 mL de solução foram coletadas, em triplicata, centrifugadas e o sobrenadante reservado para a determinação das atividades enzimáticas e geração de espectros NIR.

O complexo enzimático não comercial, denominado EETA, foi obtido em laboratório a partir de um processo de fermentação submersa em biorreator instrumentado (Tec-Bio-V 4,5 L) com controle de pH, aeração e temperatura. O bioprocesso, em batelada simples, foi realizado utilizando como substrato a torta de caroço de algodão em meio composto por 1,25% da torta, diluído em 1,5 L de água, adicionado de 0,1g L^{-1} de NaCl, 0,2 g L^{-1} de $MgSO_4.7H_2O$, 0,4 g L^{-1} de KH_2PO_4 , 0,1 g L^{-1} de K_2HPO_4 e 0,9 g L^{-1} de ureia. A fermentação foi realizada com pH 5,0, aeração de 2 Lpm (L min^{-1}) sob agitação a 200 rpm e temperatura a 30°C, sendo esta conduzida por um período de nove dias, empregando a linhagem de *Aspergillus tubigensis* AN1257 como agente do bioprocesso. A fermentação foi monitorada a cada 24 horas por meio da retirada de alíquotas de 6 mL, em triplicata, as quais foram submetidas à centrifugação a 10.000 rpm por 10 minutos, seguida da coleta do sobrenadante para posteriores análises das atividades das enzimas β-glicosidasese, CMCase, FPase e xilanases.

Visando a obtenção de um conjunto maior de amostras, parte do extrato enzimático coletado, ao final do processo, foi submetido à desnaturação térmica a 70ºC durante 30 minutos. A seguir alíquotas de 2 mL também foram coletadas seguindo o procedimento descrito para as enzimas comerciais.

4.2 Determinação das atividades enzimáticas

A atividade enzimática foi determinada de duas maneiras distintas: uma seguindo o protocolo tradicional e outra, seguindo o método adaptado para microplaca, o qual foi

empregado para fins de validação. A metodologia em microplaca foi escolhida para a determinação das atividades enzimáticas que foram utilizadas como atributos alvo da matriz de dados para a construção dos modelos preditivos.

4.2.1 Determinação das atividades enzimáticas (método tradicional)

As análises enzimáticas, apresentadas a seguir, foram determinadas apenas nos extratos enzimáticos comerciais. Os ensaios foram realizados em microtubos tipo Eppendorfs® em sextuplicata. O comprimento de onda utilizado variou em função do tipo de enzima aferida, sendo 540 nm para a determinação de endoglucanases, celulases totais e xilanases e 510 nm para a determinação de β-glicosidases.

4.2.1.1 Endoglucanase - CMCase

A atividade de endo-1,4-β-glucanase foi quantificada segundo método descrito por Ghose (1987). Este método consiste na hidrólise do substrato carboximetilcelulose (CMC) 2% em solução tampão citrato 50 mmol L^{-1}, pH 4,8. O meio reacional foi constituído de uma mistura 1:1 de extrato enzimático:substrato e a hidrólise foi conduzida à temperatura de 50°C. A determinação da atividade foi feita com base na quantificação dos açúcares redutores totais (ART) liberados após 30 minutos de reação. O doseamento desses açúcares foi determinado utilizando como reativo o ácido 3,5-dinitrosalicílico (DNS), elaborado sem o uso de fenol e bissulfito de sódio, de acordo com Bernfeld (1955), modificado por Miller (1959) e consolidadas por Vasconcelos *et* al. (2013).

O ensaio foi realizado a partir de 100 µL de amostra adicionado de 100 µL de DNS, seguido de homogeneização, aquecimento em banho-de-água fervente por 15 minutos e resfriamento em banho de água com gelo. Em seguida, a amostra foi diluída 10 vezes com água destilada e submetida à leitura em espectofotômetro a 540 nm.

A quantificação da concentração dos ART foi determinada com base em curvas de calibração utilizando glicose como solução padrão nas concentrações de 0 a 1 g L^{-1}, seguindo os mesmos procedimentos analíticos adotados para quantificação das amostras. A concentração dos açúcares redutores foi expressa em g L^{-1}.

Uma unidade de CMCase determinada representa a quantidade de enzima capaz de liberar 1 μmol de ART, expresso em glicose, por minuto, por volume (mililitros) de extrato enzimático (U mL^{-1}).

4.2.1.2 Celulases Totais - FPase

A atividade de celulases totais (FPase) foi quantificada segundo a técnica descrita por Ghose (1987) a qual consiste na hidrólise de papel de filtro *Whatman* nº 1 (1x6 cm), equivalente a 50 mg, na presença do extrato enzimático diluído (1:1) em tampão citrato 50 mmol L^{-1}, pH 4,8. A hidrólise foi conduzida à temperatura de 50ºC durante 60 minutos. O teor de açúcares redutores liberado também foi determinado pelo método DNS (MILLER, 1959), de acordo com adaptação feita por Vasconcelos *et al*. (2013). A quantificação da concentração dos ART foi realizada nas mesmas condições descritas no item 4.2.1.1.

Uma unidade de FPase determinada representa a quantidade de enzima capaz de liberar 1 μmol de ART, expresso em glicose, por minuto, por volume (mililitros) de extrato enzimático (U mL^{-1}).

4.2.1.3 β-glicosidase

A atividade de β-glicosidase foi quantificada segundo método descrito por Ghose (1987), o qual consiste na hidrólise de celobiose a 1% em solução tampão citrato 100 mmol L^{-1}, pH 4,8. O meio reacional foi constituído de uma mistura 1:1 de extrato enzimático:substrato e a hidrólise foi conduzida à temperatura de 50°C, por 30 minutos, seguida de banho de água fervente por 5 minutos e resfriamento em banho de água com gelo. A determinação da atividade foi feita com base na quantificação da glicose liberada segundo método descrito por Lloyd e Whelan (1969), procedimento padrão enzimático Glicose-oxidase/peroxidase GOD-POD. A determinação foi realizada a partir de 10 μL de amostra adicionada de 1 mL do reagente GOD-POD, seguida de homogeneização, aquecimento em banho-maria a 37º por 10 minutos e leitura da absorbância em espectofotômetro a 510 nm.

A quantificação da concentração da glicose foi determinada com base em curva de calibração utilizando glicose como solução padrão nas concentrações de 0 a 1 g L^{-1}, seguindo os mesmos procedimentos analíticos adotados para quantificação das amostras. A concentração de glicose foi expressa em g L^{-1}.

Uma unidade de β-glicosidase determinada representa a quantidade de enzima capaz de liberar 1 μmol de glicose por minuto, por volume (mililitros) de extrato enzimático (U mL^{-1}).

4.2.1.4 Xilanase

A atividade xilanolítica foi quantificada a partir da hidrólise de xilana de Birchwood (Sigma) 1,4% em tampão citrato 100 mmol L^{-1}, pH 5,0. O meio reacional foi constituído de uma mistura 1:1 de extrato enzimático:substrato e a hidrólise foi conduzida à temperatura de 50°C. A determinação da atividade foi feita com base na quantificação dos açúcares redutores totais (ART) liberados após 30 minutos de reação. O doseamento desses açúcares foi determinado como descrito no item 4.2.1.1

A quantificação da concentração dos ART foi determinada com base em curvas de calibração utilizando xilose como solução padrão nas concentrações de 0 a 1 g L^{-1}, seguindo os mesmos procedimentos analíticos adotados para quantificação das amostras. A concentração dos açúcares redutores foi expressa em g L^{-1}.

Uma unidade de xilanase determinada representa a quantidade de enzima capaz de liberar 1 μmol de ART, expresso em xilose, por minuto, por volume (mililitros) de extrato enzimático (U mL^{-1}).

4.2.2 Microescalonamento do processo de determinação de atividade enzimática

As atividades fpásica, cmcásica (endoglucanásica), β-glicosidásica e xilanásica foram determinadas em microplacas de 96 poços de fundo chato, com capacidade máxima de 350 μL. A determinação analítica para quantificação dos açúcares foi adaptada para microescala, com modificações no tempo de incubação e na temperatura, bem como eliminação da etapa de adição de água no final da reação. Os ensaios foram realizados em 14 repetições e o branco foi feito em triplicata, elaborado substituindo a enzima por água.

4.2.2.1 Fpase

A atividade de fpase foi determinada utilizando como substrato uma tira de papel de filtro *Whatman* nº 1 de tamanho 1 x 0,3 cm, adicionado de uma alíquota de 50 μL de

tampão acetato de sódio 50 mmol L^{-1}, pH 4,8 e 50 µL de extrato enzimático, seguido de homogeneização e incubação a 50ºC em estufa, por 60 minutos.

A atividade enzimática de fpases foi determinada por meio da quantificação de glicose, liberada após hidrólise no ensaio enzimático. Este açúcar foi quantificados pelo método colorimétrico de DNS, de acordo com método consolidado por Vasconcelos *et al.* (2013). O ensaio consistiu na adição de 100 µmol do reagente DNS logo após o período de incubação da enzima/substrato contidos na microplaca, a qual foi recoberta com parafilme e, posteriormente, submetida ao aquecimento de 70°C, em banho-maria, por 25 minutos. Ao final da reação, o parafilme foi removido e a microplaca submetida ao resfriamento em gelo tipo escama, seguida da leitura de absorbância a 540 nm em leitor de microplacas (ASYS UVM 340).

A quantificação da concentração de glicose foi determinada com base em curvas de calibração utilizando glicose como solução padrão nas concentrações de 0 a 0,9 g L^{-1}, seguindo os mesmos procedimentos analíticos adotados para quantificação das amostras. A concentração dos açúcares redutores foi expressa em g L^{-1}.

Uma unidade de FPase foi definida como a quantidade de enzima que libera 1 µmol de glicose por minuto por mL de solução.

4.2.2.2 Cmcase

A determinação da atividade de cmcase foi quantificada utilizando uma alíquota de 50 µL de uma solução de carboximetilcelulose (CMC 1% em tampão acetato de sódio 50 mmol L^{-1}, pH 4,8) como substrato, e 50 µL de extrato enzimático, seguido de homogeneização e incubação a 50ºC em estufa, por 30 minutos.

A atividade enzimática de cmcases foi determinada por meio da quantificação da glicose pelo método colorimétrico de DNS conforme descrito no item 4.2.2.1.

Uma unidade de cmcase foi determinada como o total de enzima capaz de liberar 1 µmol de glicose por minuto por mL de solução do ensaio.

4.2.2.3 β-glicosidase

A atividade de β-glicosidase foi determinada utilizando como substrato uma solução de celobiose 15 mM preparada em tampão acetato de sódio 50 mmol L^{-1}, pH 4.8. O

ensaio enzimático foi conduzido com a adição de 50 µL de substrato adicionado a 50 µL de extrato enzimático contidos na microplaca. Logo após procedeu-se à homogeneização e a amostra foi incubada em estufa a 50ºC, por 10 minutos, seguida de banho-de-água fervente durante 1 minuto e posterior resfriamento em gelo tipo escama. A glicose liberada foi quantificada pelo método colorimétrico proposto por Lloyd e Whelan (1969), utilizando procedimento padrão enzimático GOD-POD, adaptado para microplaca. A determinação da glicose proveniente da hidrólise foi realizada em microplaca pela adição de 10 µL da amostra proveniente da reação enzimática e 300 µL do reagente GOD-POD.

Uma curva de calibração de glicose com concentração variando de 0 a 0,7 g L^{-1} foi construída com intervalos de 0,1 g L^{-1} e submetida, simultaneamente, aos mesmos procedimentos que a amostra. Os resultados da concentração de glicose foram expressos em g L^{-1}.

Uma unidade de β-glicosidase representa o total de enzima capaz de liberar 1 µmol de glicose por minuto por mL de solução do ensaio.

4.2.2.4 Xilanase

A determinação da atividade xilanolítica foi quantificada utilizando uma alíquota de 50 µL de uma solução de xilana de Birchwood (Sigma) 1,4% em tampão acetato de sódio 50 mmol L^{-1}, pH 4.8 como substrato, e 50 µL de extrato enzimático, seguido de homogeneização e incubação a 50ºC em estufa, por 5 minutos.

A atividade enzimática de xilanases também foi determinada pelo método colorimétrico de DNS. Os ensaios foram realizados conforme descrito no item 4.2.2.1. com curva de calibração tendo xilose como solução padrão nas concentrações de 0 a 0,9 g L^{-1}. Os resultados da quantificação dos açúcares redutores foram expressos em g L^{-1}**.**

Uma unidade de xilanase representa o total de enzima capaz de liberar 1 µmol de xilose por minuto por mL de solução do ensaio.

4.3 Validação do método de determinação de ART em microplaca

O método de determinação de ART em microplaca foi previamente validado para posterior utilização na quantificação das atividades enzimáticas. Para tanto, os ensaios consistiram na adição de 100 µL de solução de glicose a 0,5 g L^{-1} adicionados a 100 µL de

DNS contidos em microplaca de 96 poços, seguida de incubação em banho-maria em tempos e temperaturas distintas conforme apresentados na Tabela 1. As microplacas foram recobertas com parafilme (pvc) com a finalidade de evitar a entrada de água e após a incubação procedeu-se o resfriamento em gelo tipo escama, seguida da leitura de absorbância a 540 nm em leitor de microplacas (ASYS UVM 340).

Tabela 1 – Condições de temperatura e tempo de incubação para avaliar as melhores condições de reação para quantificação dos ART a serem empregados nos ensaios enzimáticos

Ensaio	Temperatura (ºC)	Tempo de incubação (minutos)
A	100	5
B	70	5
C	70	10
D	70	15
E	70	20
F	70	25
G	70	30
H	70	35
I	70	40
J	70	45
K	70	50

O ensaio consistiu de 11 diferentes condições envolvendo temperatura e tempo de incubação e 16 repetições. Os dados obtidos foram submetidos à análise de variância a 5% de significância.

4.4 Ensaios preliminares: correlação dos métodos de determinação enzimática tradicional e microplaca

Visando adquirir informações sobre a relação existente entre os resultados obtidos nos ensaios enzimáticos provenientes dos métodos tradicional e em microplaca procedeu-se como descrito a seguir:

Os complexos enzimáticos comerciais, Celluclast® e Cellic®CTec2 foram submetidos a cinco diluições diferentes para posteriores determinações enzimáticas. As amostras provenientes do complexo Celluclast® foram preparadas considerando diluições de

500 a 1000, com intervalos de 100 vezes. Diferentemente, as amostras advindas do complexo Cellic®CTec2 foram diluídas de 1000 a 2000 vezes, com intervalos de 200 vezes.

A metodologia determinada em microplaca foi avaliada segundo as métricas de linearidade e sensibilidade. A linearidade foi medida considerando o coeficiente de correlação de Pearson; a sensibilidade, determinada pelo coeficiente angular das respectivas retas de calibração.

4.5 Geração das amostras por espectroscopia na região do infravermelho próximo

As amostras provenientes dos três complexos enzimáticos (Celluclast®, Cellic®CTec2 e EETA) também foram analisadas quanto aos espectros na região do infravermelho próximo. Os complexos enzimáticos foram utilizados como fontes das celulases e xilanases e, objetivando produzir uma variabilidade nos valores dessas enzimas, cada um desses preparados foi submetido a tratamentos térmicos variados, conforme Tabela 2, em banho-maria termostático.

Tabela 2 – Espectros gerados a partir dos diferentes complexos enzimáticos (Celluclast®, Cellic®CTec2 e EETA) utilizados como fontes das celulases (β-glicosidases, cmcases e fpases) e xilanases submetidos a diferentes temperaturas de desnaturação e diferentes intervalos de tempo.

EE	Exp	TD	Tt/Ir (min)	Enzimas	Total de espectros
Cellic® CTec2	1	70ºC	120/10	β-glicosidases	105
	2	65ºC	120/10		117
	3	60ºC	180/10	xilanases	171
	4	60ºC	120/10		117
	5	70ºC	55/10		63
	6	60ºC	57/10		63
	7	60ºC	53/10		63
	8	85ºC	30/15		27
	9	100ºC	15/15	β-glicosidases, cmcases, fpases, xilanase	9
	10	70ºC	30/3		33
	11	70ºC	33/3	β-glicosidases, xilanase	33
EETA	12	SD	---	β-glicosidases, cmcases, fpases, xilanase	15
	13	SD	10.080/1.440		81
	14	70ºC	30/1 e 5 *		33

EE	Exp	TD	Tt/Ir (min)	Enzimas	Total de espectros
celluclast®	15	100ºC	15/15	β-glicosidases, cmcases, fpases, xilanase	9
	16	70ºC	33/3		33
	17	70ºC	30/5		93
	18	70ºC	15/1		144
	19	70ºC	15/1		144
	20	80	10/1		98
	21	85	20/2		99
	22	75	10/1		132

EETA: extrato enzimático produzido em biorreator neste trabalho; EE: extrato enzimático; Exp: experimento; SD: sem desnaturação; * coleta realizada a cada um minuto durante cinco minutos e sequencialmente a cada cinco minutos; Tt/Ir=Tempo total/Intervalo regular de coleta em minutos; TD = Temperatura de Desnaturação.

Os ensaios foram realizados em intervalos pré-definidos experimentalmente, onde alíquotas contendo 2 mL de amostras tratadas termicamente foram recuperadas em microtubos tipo *Eppendorf*, em triplicatas, e submetidas a resfriamento em banho de gelo por 30 segundos, seguido de centrifugação a 10.000 rpm por 10 minutos. Os sobrenadantes foram recuperados e reservados para determinações analíticas em infravermelho próximo em espectrofotômetro modelo NIR 9000 (Femto), ajustado para varredura na região de 1100 a 2500 nm, com resolução de 2 nm.

4.6 Construção da base de dados para modelagem computacional

Os resultados obtidos dos espectros NIR e das médias das determinações analíticas das atividades de β-glicosidaseses, cmcases, fpases e xilanases foram utilizados para construção de uma matriz de dados para posterior aplicação de técnicas multivariadas.

A matriz foi organizada de forma que as médias das atividades enzimáticas representassem os atributos dependentes (atributo alvo) e os valores de absorbância, advindos dos espectros, atributos independentes. Alguns atributos adicionais foram incluídos na matriz de dados com o objetivo de auxiliar no pré-processamento e interpretação. Os atributos foram nomeados como: a) experimento (contendo informação sobre o tipo de complexo enzimático utilizado, e se resultou de fermentação ou desnaturação), b) descrição da amostra (considerando o nome da amostra e dados da triplicata), c) intervalo de obtenção da amostra (variando de zero a 192 minutos), d) duração (tempo total do experimento), e) intervalo (tempo de retirada de cada amostra) e f) temperatura (temperatura utilizada nos ensaios de

fermentação ou desnaturação). Essa matriz representa a entrada do sistema da modelagem computacional empregada neste estudo.

4.7 Pré-processamento dos dados

A limpeza e a transformação dos dados necessários para aplicação das técnicas multivariadas bem como a modelagem computacional foram implementadas na linguagem de programação *Python* utilizando várias bibliotecas, tais como:

a) *Pandas*, para manipulação e análise dos dados (MCKINNEY, 2018);

b) *Numpy*, para análise numérica, utilizada para cálculos matriciais (OLIPHANT, 2006);

c) *Sklearn*, para aplicação dos algoritmos de aprendizado de máquina (SZYMAŃNSKI; KAJDANOWICZ, 2019);

d) *Scipy*, para aplicação do filtro digital de Savitzky-Golay, utilizado para suavização dos espectros NIR (JONES *et al.*, 2001);

e) *Matplotlib*, para criação de gráficos de alta qualidade (HUNTER *et al.*, 2020);

f) *Seaborn*, para geração de gráficos estatísticos (TAYLOR-MORGAN, 2020);

g) *Stadmodels*, para realização de testes estatísticos (SEABOLD; PERKTOLD, 2010).

A plataforma utilizada foi a *Jupyter Notebook*, com a versão 3.7 da linguagem *Python* (PERKEL, 2018). Todas as bibliotecas citadas são distribuídas com licença GNU GPL (*General Public Licence*), podendo ser livremente utilizadas.

Visando melhorar o desempenho das técnicas multivariadas utilizadas neste estudo, inicialmente, foi feita a seleção de atributos por meio do algoritmo de regressão de informação mútua. Após a seleção, os atributos considerados relevantes foram submetidos à análise das componentes principais com o objetivo de verificar se o conjunto de amostras poderia ser utilizado em sua totalidade na construção dos modelos preditivos ou, ainda, se as mesmas apresentavam algum tipo de agrupamento que fosse determinante para escolha de um grupo de amostras.

O processo de limpeza de dados consistiu na identificação de valores ausentes e na remoção de amostras atípicas. Essa remoção foi auxiliada por técnicas de identificação de

*outlier*s. Após a limpeza dos dados, para eliminar os efeitos da dispersão dos espectros foram aplicadas as técnicas de pré-processamento abaixo:

a) Alisamento de Savitzky–Golay,

b) Padronização e,

c) Derivação.

Essas técnicas foram aplicadas individualmente e ou associadamente, como apresentadas a seguir. A técnica de suavização foi ajustada nos parâmetros janela, grau do polinômio e derivada. A janela variou entre 3 e 5 pontos; o grau do polinômio variou entre, 1 e 2 e, a deriva foi de 1ª ou 2ª ordem. Os parâmetros foram apresentados de forma abreviada acompanhada por números que representam a variação destes, como por exemplo, Par: 3, 2, 1', onde: Par = parâmetro; 3 = representa janela de 3 pontos; 2 = polinômio de segundo grau e 1 = primeira derivada.

1: Padronização Normal do Sinal (SNV – *Standard Normal Variate*) (individual)

2: Suavização (SavGol) - Par:3,1,1 (associada)

3: Suavização (SavGol) - Par:3,2,1 (associada)

4: Suavização (SavGol) - Par:5,1,1 (associada)

5: Suavização (SavGol) - Par:5,2,1 (associada)

6: Suavização (SavGol) - Par:3,2,2 (associada)

7: Suavização (SavGol) - Par:5,2,2 (associada)

8: Suavização (SavGol) - Par:3,1,1 seguida por Padronização (associada)

9: Padronização seguida por Suavização (SavGol) - Par:3,1,1 (associada)

10: Correção da dispersão multiplicativa (MSC – *Multiplicative Scatter Correction*) (individual)

Visando comprovar que essas técnicas foram relevantes para a melhoria dos resultados, a modelagem, também, foi realizada sobre os dados sem nenhuma aplicação das técnicas supracitadas.

4.8 Modelagem computacional: construção e avaliação dos modelos

A construção dos modelos de calibração, validação e predição foi realizada sobre todo o conjunto de dados e sobre subconjuntos de dados, separados em função do experimento ou do tipo de extrato enzimático utilizado. A modelagem foi realizada com o

auxílio de duas ferramentas distintas: uma proprietária (*software Unscrambler)* e uma livre (*jupyter-notebook*). Na primeira foi realizada calibração e validação sobre um pequeno conjunto de dados, utilizando o algoritmo PLS. Na segunda ferramenta, foram realizadas todas as etapas: calibração, validação e teste, com um número maior de amostras e testados oito diferentes algoritmos, inclusive o PLS.

4.8.1 Modelagem com software proprietário: calibração e validação

Uma modelagem inicial foi realizada utilizando o s*oftware Unscrambler* X (*Camo Software*, Oslo, Noruega), com o objetivo de verificar a existência de correlação entre os espectros e a atividade enzimática. Para tanto, o conjunto de dados representado pelo Experimento 16 da Tabela 2, obtido a partir do extrato enzimático comercial Celluclast®, foi utilizado como fonte de celulases e xilanases.

Os dados espectrais foram submetidos a pré-processamentos com base na suavização de Savitzky-Golay (SAVITZKY; GOLAY, 1964) e padronização SNV (BARNES *et al.*, 1989). A regressão por mínimos quadrados parciais (PLS) foi o algoritmo usado para análise multivariada de dados. A qualidade dos modelos de regressão PLS gerados foi avaliada através de RMSEC (erro quadrático médio da calibração), RMSECV (erro quadrático médio da validação cruzada), RPD (Razão de Desempenho para o Desvio) e análise RER (Razão de intervalo de erro). Para os cálculos procedeu-se como descrito a seguir:

a) RPD - Foi calculado como a razão entre o desvio padrão dos dados de referência para o conjunto de validação e o erro padrão de previsão (de validação cruzada).

b) RER - Foi calculado como a razão entre o intervalo de dados de referência de validação e o erro padrão de predição (de validação cruzada).

4.8.2 Modelagem com software livre: calibração, validação e predição

Uma modelagem posterior foi realizada, utilizando uma plataforma de programação livre, na qual foi possível implementar todas as tarefas de indução dos modelos multivariados (calibração, validação e teste), bem como todas as métricas de validação dos mesmos. Para tanto, o conjunto de dados escolhido, após o pré-processamento, foi

automaticamente particionado em dados de treino e teste, sendo 25% destes selecionados para teste.

O grau de concordância entre o valor estimado pelo modelo multivariado e o valor de referência (média de atividade enzimática determinada em microplaca) foi verificado por medidas de precisão como a raiz quadrada do erro quadrático médio da calibração/validação/predição (RMSEC/RMSEV-/RMSEP *Root Mean Square Error CalibRation*/*Validation*/*Prediction*)

O desempenho de previsão dos algoritmos foi determinado através da comparação do coeficiente de determinação encontrado para cada modelo. A qualidade dos modelos de calibração e predição foi determinada através das seguintes métricas:

a) *Bias*,

b) SEP,

c) Relação do desempenho para o desvio RPD (*Residual Prediction Deviation*),

d) Razão do desempenho interquartil RPIQ (*Ratio of Performance to IQ*) e

e) Razão de intervalo de erro RER (*Range Error Ratio*).

As métricas foram calculadas de acordo com Fearn (2002), Santos *et al.* (2019) e (Jin *et al.*, 2020), conforme apresentadas pelas equações de 1 a 6.

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

$$bias = \frac{\sum y_i - \hat{y}_i}{n} \quad (2)$$

$$SEP = \sqrt{\frac{\sum ((y_i - \hat{y}_i) - bias)^2}{n-1}} \quad (3)$$

$$RPD = \frac{SD}{RMSE} \quad (4)$$

$$RPIQ = \frac{(Q1 - Q3)}{RMSE} \quad (5)$$

$$RER = \frac{y_{max} - y_{min}}{SEP} \quad (6)$$

em que:

n = número de amostras,

y = valor obtido pelo método de referência,

$\hat{y}$ = valor predito,

Q1 = interquartil 1 (25%)

Q3 = interquartil 3 (75%)

Um variado número de algoritmos de aprendizado de máquina foi treinado e testado com o objetivo de determinar qual modelo apresenta a melhor taxa de calibração/validação/predição, sendo estes os seguintes métodos:

a) regressão linear múltipla (MLR);

b) regressão por mínimos quadrados parciais (PLSR);

c) regressão por componentes principais (PCR);

d) regressão por aumento de gradiente (GBR), e

e) regressão por quadrados mínimos lineares com e sem *Kernel* (*KenelRidge e Ridge*).

A determinação simultânea das atividades enzimáticas de β-glicosidaseses, CMCases, FPases e xilanases foi realizada com auxilio do algoritmo *MultiOutputRegressor*, presente no pacote *Sklearn.multioutput*.

A implementação dos algoritmos MLR e Ridge foram importados do pacote Sklearn.linear_model, o qual, também, serviu para implementação da PCR, juntamente com a PCA, obtida do pacote sklearn.decoposition. Os algoritimos PLS e GBR foram importados dos pacotes sklearn.cross_decomposition e sklearn.ensemble, respectivamente. Algumas métricas de validação foram importadas dos pacotes sklearn.model_selection e sklearn.metrics; as demais métricas foram implementadas.

5 RESULTADOS E DISCUSSÃO

Os resultados obtidos neste trabalho estão apresentados em subseções abordando os seguintes tópicos: determinações enzimáticas, validação da metodologia de determinação enzimática em microescala, geração de espectros NIR, construção da base de dados, pré-processamento e modelagem computacional.

5.1 Determinações enzimáticas

Os valores médios da atividade enzimática de β-glicosidade, cmcase, fpase e xilanase provenientes dos complexos enzimáticos Celluclast®, Cellic® CTec2 e EETA, bem como os valores mínimo, máximo e o coeficiente de variação (CV) encontram-se dispostos na Tabela 3.

Tabela 3 – Valores mínimo, máximo e médio de atividade enzimática de β-glucosidade, CMCase, FPase e xilanase nos complexos Celluclast®, Cellic® CTec2 e EETA

Extrato enzimático	Atividade enzimática (U mL^{-1})				
	Medidas	**β-glucosidade**	**CMCase**	**FPase**	**Xilanase**
EETA	**Min**	3,44	0,052	0,0542	14,19
	Max	3,74	0,056	0,0567	14,86
	Média ± Desvio	3,58±0,12	0,054±0,002	0,055±0,001	14,53 ±0,47
	CV %	**3,39**	**3,76**	**1,88**	**1,87**
Celluclast®	**Min**	248,61	144,64	67,13	1018,64
	Max	273,58	162,12	69,85	1060,04
	Média ± Desvio	257,74±7,95	155,06±5,32	68,88±0,89	1043,77±12,74
	CV %	**5,34**	**5,94**	**2,20**	**2,12**
Cellic® CTec2	**Min**	2302,27	195,81	99,59	2637,05
	Max	2609,64	207,14	107,24	2723,30
	Média ± Desvio	2469,01±126,83	200,68±4,76	104,38±3,40	2683,76±25,2
	CV %	**5,14**	**2,37**	**3,26**	**0,94**

U mL^{-1}: uma Unidade de atividade enzimática representa o total de enzima capaz de liberar 1 µmol de açúcar$_{redutor}$ min^{-1} mL^{-1} ; CV: coeficiente de variação da determinação enzimática em microplaca em 14 repetições.

Os resultados de atividade enzimática obtidos para os três extratos enzimáticos antes da etapa de desnaturação permitiram observar que o complexo enzimático Celluclast® contém um maior número de enzimas xilanolíticas (~1043 UI), seguido de β-glicosidases (~257 UI), cmcases (~155 UI) e fpases (~68). O extrato Cellic® CTec2 apresentou valores 2,5 vezes maiores de xilanases e aproximadamente 10 vezes de β-glicosidases em comparação com o Celluclast®, enquanto o teor de cmcases e fpases foi equivalente. Além disso, foi observado que o extrato Cellic® CTec2 possui uma concentração de β-glicosidases aproximadamente 15 vezes superior às outras enzimas, podendo ser caracterizado como um concentrado desta enzima. Apesar do extrato EETA possuir concentrações maiores de xilanases e β-glicosidases do que CMCase e FPase, os teores destas enzimas neste complexo enzimático são inferiores quando comparadas às concentrações obtidas para os complexos enzimáticos comerciais. Essa diferença está relacionada ao fato destes extratos serem puros e concentrados e o EETA ser um extrato obtido de um processo de fermentação sem ter sido submetido à purificação.

As xilanases apresentaram menor coeficiente de variação em todos os complexos enzimáticos analisados, enquanto o maior foi registrado para a enzima CMCase no complexo Celluclast®, com valores em torno de 6%. Apesar do CV se aproximar de 6% na determinação de atividade para esta enzima, em todos os outros ensaios a média não ultrapassou 4%. O coeficiente de variação registrado nos ensaios colorimétricos adaptados para microplaca permaneceu majoritariamente (cerca de 75% dos ensaios) abaixo de 4%. A variação observada neste trabalho foi comparável aos valores obtidos por König *et al.* (2002), Decker *et al.* (2003), Lai *et al.* (2006) e Yu *et al.* (2016), que também obtiveram coeficientes de variação inferiores a 10%. König *et al.* (2002) relataram que valores de CV abaixo de 10% são considerados aceitáveis para determinações enzimáticas.

Conforme apresentado nas Tabela 4 e 5, os processos de desnaturação dos três extratos enzimáticos estudados produziram valores de atividade enzimática variados, apresentando médias distintas daquelas encontradas nas análises iniciais para cada complexo enzimático. Esta assertiva pode ser melhor observada ao comparar os dados da Tabela 4 com os apresentados na Tabela 3, com foco para o coeficiente de variação, representado na Tabela 4, o qual foi calculado em relação a todas as determinações enzimáticas em cada intervalo do processo de desnaturação. Portanto, foi observado que a desnaturação teve um efeito positivo

em produzir uma variabilidade no teor de atividade enzimática das amostras, refletindo no aumento no valor do coeficiente de variação.

Na Tabela 4 são apresentados os valores mínimo, máximo e o coeficiente de variação para cada extrato enzimático durante as etapas de desnaturação térmica, demonstrando o aumento da variabilidade dos teores de atividade enzimática.

Tabela 4 – Valores mínimo e máximo de atividade enzimática de β-glucosidade, CMCase, FPase e xilanase provenientes dos extratos enzimáticos Celluclast®, Cellic® CTec2 durante as etapas de desnaturação

Extrato enzimático	**Atividade enzimática (U mL^{-1})**				
	Medidas	**β-glucosidade**	**Cmcase**	**Fpase**	**Xilanase**
EETA	Min	1,94	0,02	0,005	0,3
	Max	3,74	0,05	0,03	10,2
	CV %	**21,55**	**31,52**	**64,12**	**96,92**
Celluclast®	Min	2,38	50,73	2,65	19,41
	Max	929,79	147,03	58,55	1102,15
	CV %	**87,51**	**20,65**	**47,38**	**58,70**
Cellic® CTec2	Min	122,12	88,78	20,8	359,49
	Max	2918,57	287,96	163,33	3186,5
	CV %	**37,75**	**33,84**	**56,98**	**62,39**

CV: coeficiente de variação em relação aos valores de atividade enzimática obtidos durante a desnaturação térmica.

No total de 470 determinações enzimáticas realizadas foi observada uma variabilidade (CV) nos valores de atividade enzimática entorno de 21 a 97%. Dentres essas medições, 309 foram obtidas para as quatro atividades enzimáticas simultaneamente, sendo 14 medições geradas a partir do complexo enzimático Celic® CTec2, 44 a partir do EETA e 251 a partir do Celluclast®.

Tabela 5 – Total de determinações enzimáticas geradas a partir dos complexos enzimáticos Celluclast®, Cellic® CTec2 durante as etapas de desnaturação

Extrato enzimático	**Enzimas**	**Total de determinações enzimáticas**
Cellic® CTec2	β-glicosidases	39
	xilanases	111
	β-glicosidases, CMCcases, FPases, xilanase	14
EETA	β-glicosidases, xilanase	11
	β-glicosidases, CMCcases, FPases, xilanase	44
Celluclast®	β-glicosidases, CMCcases, Fpases, xilanase	251

No processo de desnaturação térmica dos complexos enzimáticos Celluclast®, Celic®, CTec2 e EETA também foi observado que o tempo resultou em uma redução na atividade enzimática de forma seletiva, conforme ilustrado nas Figuras de 5 a 7.

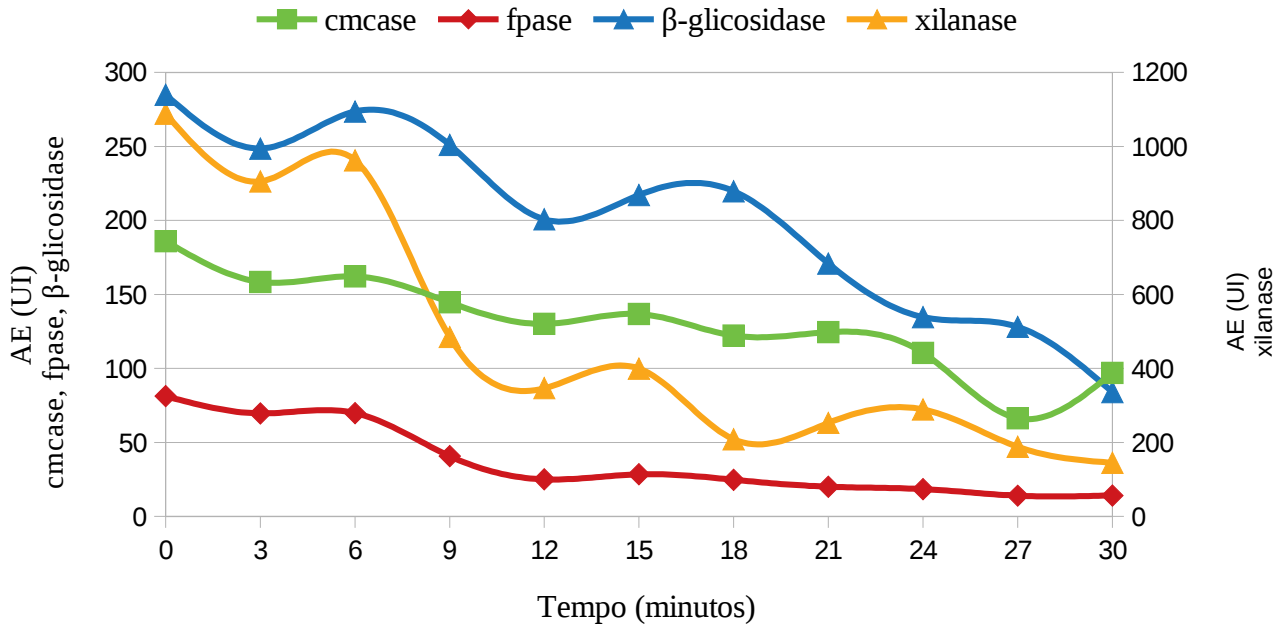

Figura 5 – Determinações enzimáticas de β-glicosidase, CMCase, FPase, e xilanase durante a desnaturação térmica do extrato enzimático Celluclast® a 70ºC, durante 30 minutos

Na Figura 5 encontra-se ilustrado o processo de desnaturação de um dos ensaios utilizando o complexo enzimático Celluclast®, desnaturado a 70°C, durante 30 minutos, com amostras sendo retiradas a cada três minutos. O decaimento rápido de atividade enzimática para todas as enzimas analisadas foi observado após os primeiros 10 minutos de desnaturação térmica, resultando, portanto, em uma variabilidade nos valores de atividade enzimática, o que possibilitou, também, a geração de espectros NIR com valores de absorbância variados, permitindo aumentar o número de amostras do conjunto de dados.

Uma estabilidade maior da atividade enzimática da CMCase foi observada, resultando em um conjunto de amostras com menor variabilidade para esta atividade. Os coeficientes de variação para as atividades β-glicosidase, CMCase, FPase e xilanase nesse ensaio foram de 87%, 21%, 47%, e 59%, respectivamente, o que revela uma maior variabilidade para β-glicosidase e xilanase.

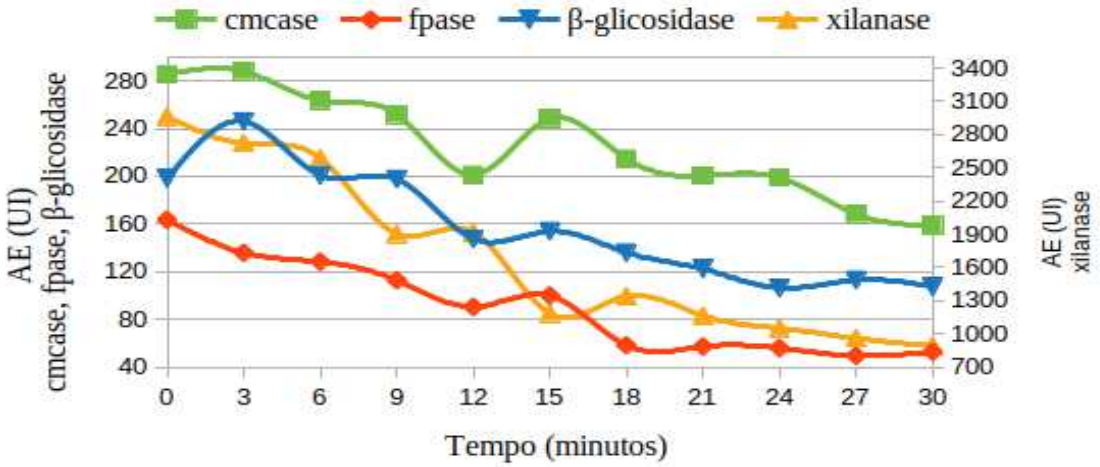

Figura 6 – Determinações enzimáticas para β-glicosidase, CMCase, FPase, e xilanase durante a desnaturação térmica do extrato enzimático Cellic® CTec2 a 70ºC, durante 30 minutos.

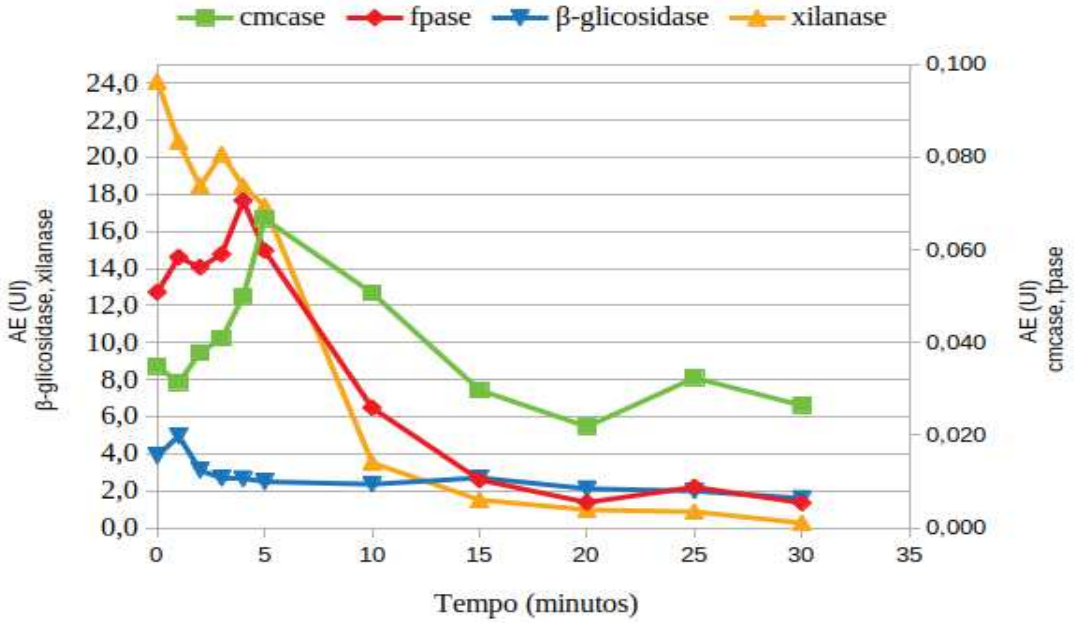

Figura 7 – Determinações enzimáticas para β-glicosidase, CMCase, FPase, e xilanase durante a desnaturação térmica do extrato enzimático EETA a 70ºC, durante 30 minutos.

Nas Figuras 6 e 7 são apresentadas representações gráficas da desnaturação térmica dos complexos enzimáticos Cellic® CTec2 e EETA, respectivamente. Foi observado um resultado equivalente ao verificado na desnaturação do complexo enzimático Celluclast®

(Figura 5), no qual, após 10 minutos de desnaturação térmica, os valores de atividade enzimática para todas as enzimas decaiu rapidamente. As atividades de CMCase e β-glicosidase para o extrato EETA desnaturado (Figura 7) apresentaram uma variabilidade inferior a 32%, enquanto para as enzimas fpase e xilanase a variabilidade foi superior a 64%.

Os resultados obtidos a partir do processo fermentativo, representado na Figura 8, permitiram a obtenção de uma variabilidade superior àquela obtida nos processos de desnaturação térmica parra a maioria das enzimas, pois gerou um coeficiente de variação mínimo de 50% (para enzima FPase), indo até 81% no caso da enzima β-glicosidase. Os resultados das atividades enzimáticas da fermentação foram utilizados na geração dos modelos para o complexo EETA, pois produziram a variabilidade que se desejava para a construção dos modelos de predição.

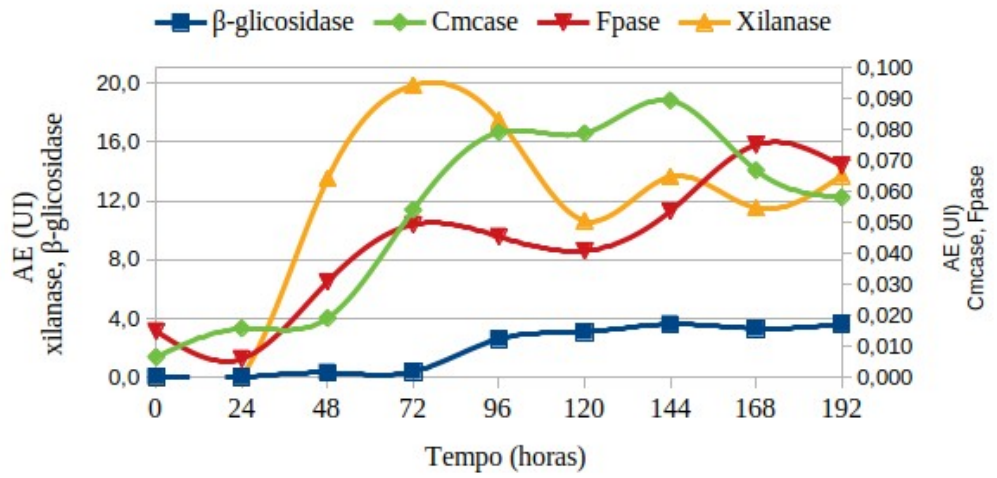

Figura 8 – Determinações enzimáticas para β-glicosidase, CMCase, FPase, e xilanase durante fermentação de torta de caroço de algodão por *Aspergilus turbigensis* em biorreator, a 30ºC, durante 192 horas, a 200 rpm e 2 Lpm.

Durante o processo fermentativo (Figura 8), que resultou na produção do extrato enzimático EETA, foi observado que o teor das enzimas CMCases e FPases encontrado foi inferior às demais enzimas estudadas, xilanase e β-glicosidase, visto que à medida que se realizava os ensaios para as determinações enzimáticas, a atividade enzimática média máxima obtida não ultrapassou 0,1 U mL^{-1}, no intervalo de 144 horas de fermentação. A AE máxima para a xilanase, por outro lado, ocorreu no período de 72 horas, com aproximadamente 20 U mL^{-1}, ou seja teores 200 vezes maior de xilanase do que CMCases e

FPases. A enzima β-glicosidase apresentou picos de atividade a partir das 96 horas de fermentação, chegando a uma concentração máxima de 4 U mL^{-1}, aproximadamente. O comportamento observado permitiu obter um conjunto de espectros com atividades enzimáticas distintas com variabilidade superior às obtidas no processo de desnaturação térmica.

5.2 Validação da metodologia de determinação de ART utilizando DNS sem fenol e bissulfito de sódio

Os ensaios para a determinação dos açúcares redutores em microplaca, seguindo as adaptações de produção de reagente DNS propostas por Vasconcelos *et al*. (2013), utilizando como reagente o DNS sem formol e bissulfito de sódio, conforme descrito na seção 4.3., resultaram em concentrações de glicose e desvios-padrão conforme apresentados nas Figuras 9 e 10.

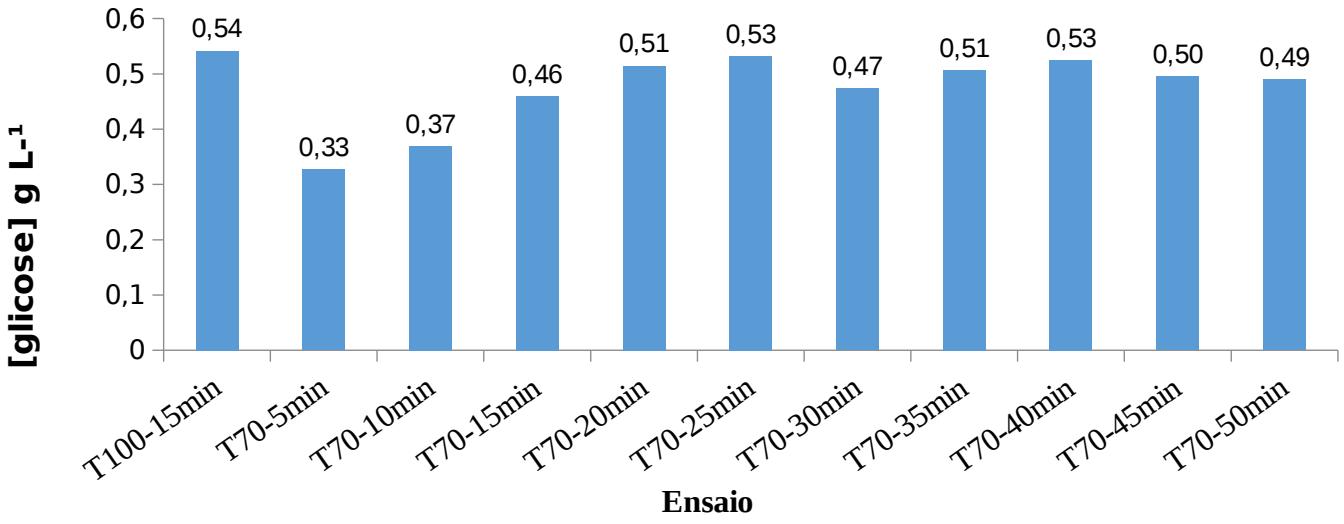

Figura 9 – Concentração de glicose de ensaios com variação de temperatura e tempo de incubação

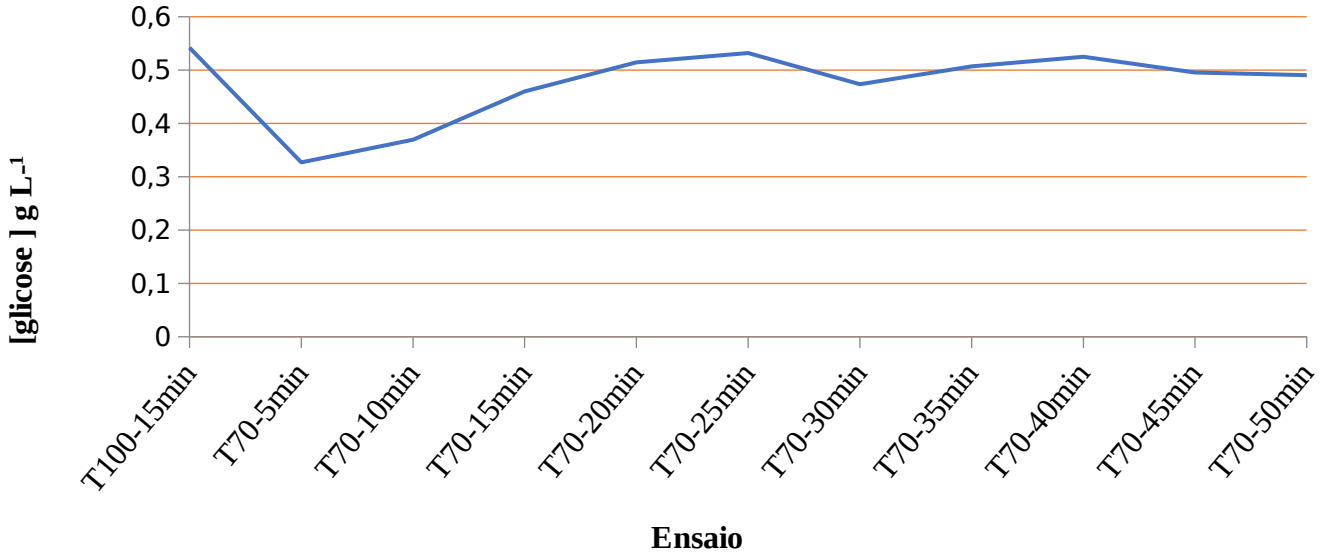

Figura 10 – Desvios-padrão de ensaios com variação de temperatura e tempo de incubação

Os resultados obtidos revelaram que todos os ensaios provenientes de aquecimento tanto a 70ºC quanto a 100ºC, com tempos de incubação superiores a 15 minutos, apresentaram concentrações de glicose próximas ao esperado, 0,5 g L^{-1}. Este resultado corrobora com aqueles encontrados por Vasconcelos *et al.* (2013), os quais constataram que as reações colorimétricas utilizando o reativo DNS, sem fenol e bissulfito em sua formulação, necessitam de no mínimo 15 minutos para serem concluídas. Além disso, foi possível constatar que apesar do tempo da reação colorimétrica com o DNS ser acrescida, isso não promoveu um aumento no tempo final de análise devido à supressão da etapa de diluição realizada neste trabalho. Esta supressão é também defendida por Wood *et al*. (2012), que afirmaram que as diluições interferem na precisão do método.

Para além do tempo de reação, os experimentos em microplaca, realizados neste trabalho, resultaram também em uma redução de resíduos tóxicos gerados, pois o volume do reagente DNS foi reduzido de 500 µL para 100 µL e a precisão do método foi mantida, conforme pode ser constatado nos resultados apresentados na Tabela 6. Nesta tabela está representada a análise de variância (ANOVA) executada sobre as médias de glicose obtidas dos ensaios de determinação de ART com DNS sem formol e bissulfito em microplaca que teve por objetivo verificar se havia diferença estatística significativa entre os ensaios bem

como determinar se, havendo diferença, qual o percentual de variação estaria associado à temperatura, ao tempo de incubação ou aos resíduos ou fatores não controlados.

Tabela 6 – Resultados de ANOVA para ensaios de determinação de ART com DNS sem formol e bissulfito

Origem de variações	**Soma dos quadrados (SQ)**	**Graus de liberdade (GL)**	**Quadrado Médio (MQ)**	**Valor de F**	**Valor P**	**F crítico**
Entre grupos (tratamentos)	0,328	10	0,033	13,433	**<0,001***	1,977
Dentro dos grupos (resíduos)	0,161	66	0,002			
Total	0,489	76				

*$p=1,73E^{-12}$

Conforme ilustrado na Tabela 6, p-valor foi menor que 0,05, mostrando que existe diferença significativa entre os ensaios, dependendo da mudança na temperatura e tempo de incubação, com 95% de confiança. Ademais, esses mesmos fatores explicam 67% da variação dos resultados e são suficientes para determinar a diferença entre os ensaios.

Os resultados encontrados para a concentração média de glicose dos ensaios de determinação de ART com duração de 5 e 10 minutos, que foram de 0,33 g L^{-1} e 0,37 g L^{-1}, sugerem que estes tempos não são suficientes para que a reação colorimétrica ocorra e, portanto, não determinam com precisão o teor de ART para a adaptação da metodologia em microplaca. Além disso, o coeficiente de variação para o tempo de 5 minutos foi de 41,8%, revelando que esses tempos de incubação devam ser desconsiderados. Os resultados do teste de Tukey para verificação da existência de diferença entre os ensaios e confirmação desses pressupostos, bem como as médias de concentração de glicose obtidas, o desvio padrão das médias e coeficiente de variação são apresentados na Tabela 7. Foi possível constatar, portanto, que não existe diferença estatística, a 5% de significância, entre todos os ensaios realizados com no mínimo 15 minutos de incubação (A), independentemente da temperatura utilizada. Foi possível, portanto, confirmar, da mesma maneira que o fizerem Vasconcelos *et al.*(2013), que o tempo mínimo para a reação ocorrer deveria ser de 15 minutos e que, portanto, para os ensaios apresentados existe diferença estatística significativa entre aqueles realizados durante 5 e 10 minutos (B) e todos os demais ensaios (A).

Tabela 7 – Comparação entre médias de concentração de glicose (g L^{-1}) considerando as temperaturas de 70°C e 100°C e o tempo de reação de 5 a 50 minutos, utilizando o teste de Tukey

Ensaio	Média ±DP (gL^{-1})	CV (%)
T100-15min	0,54 ±0,01 A	1,80
T70-05min	**0,33 ± 0,14 B**	**41,61**
T70-10min	**0,37 ± 0,01 B**	**2,00**
T70-15min	0,46 ± 0,04 A	9,13
T70-20min	0,51 ± 0,01 A	2,60
T70-25min	0,53 ± 0,03 A	4,83
T70-30min	0,47 ± 0,02 A	4,49
T70-35min	0,51 ± 0,01 A	2,58
T70-40min	0,53 ± 0,00 A	0,92
T70-45min	0,50 ± 0,03 A	5,51
T70-50min	0,49 ± 0,02 A	4,06

DP: Desvio Padrão; CV: coeficiente de variação.

Observou-se que os ensaios para determinação dos ART com tempo mínimo de incubação de 15 minutos tanto a 100ºC quanto a 70ºC não diferiam estatisticamente entre si, porém os ensaios realizados sob temperaturas de 100°C provocavam imediata deformação nas microplacas a partir de 5 minutos de incubação, conforme pode ser observado na Figura 11, o que impossibilitou a reutilização das mesmas.

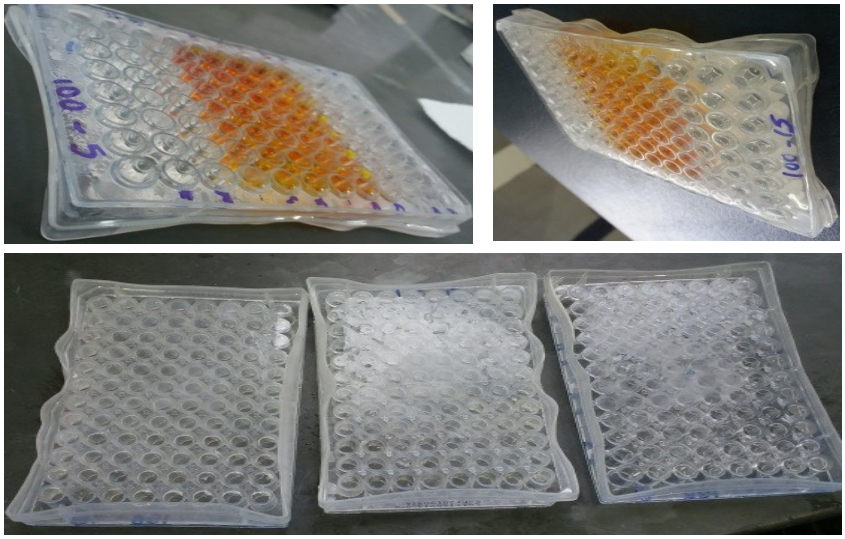

Figura 11 – Microplacas submetidas a banho-maria a 100 ºC, partes frontal e posterior, durantes no mínimo 5 minutos de incubação.

Considerando que não houve diferença estatística significativa entre os ensaios realizados a 70ºC e a 100ºC submetidos a incubação por tempo superior a 15 minutos (Tabela 7), que os ensaios submetidos a 70°C, independente do tempo de incubação resultaram em concentrações de glicose satisfatórias, próximos a 0,5 g L^{-1} e que, os ensaios com duração de

no mínimo 20 minutos apresentaram os menores coeficientes de variação, conclui-se que o uso de temperatura de 70°C foi considerada a mais adequada uma vez que não ocasionou danos físicos às microplacas. Em relação ao tempo de incubação, optou-se de forma aleatória pelo tempo de 25 minutos a ser aplicado nas determinações das atividades enzimáticas de CMCase, FPases e xilanases.

Os resultados obtidos através da análise de correlação entre as determinações enzimáticas realizadas do modo tradicional e adaptadas para microplaca, conforme decisões definidas acima, são apresentados na próxima seção.

5.3 Correlação entre as determinações enzimáticas realizadas pelo método tradicional e em microescala

Os coeficientes de correlação de Pearson (r) calculados para as curvas padrão de glicose (uma utilizando DNS e a outra GOD-POD como reagentes) e xilose obtidas tanto na metodologia tradicional quanto em microplaca para as determinações enzimáticas resultaram em valores superiores a 0,99 em todos os ensaios, representando, portanto, que em ambas as metodologias foi possível determinar uma correlação positiva forte entre concentração e absorbância. Observou-se nos ensaios para determinação das celulases e hemicelulases com os extratos enzimáticos Cellic® CTec2 e Celluclast® sob cinco diluições seriadas que, tanto na metodologia tradicional quanto em microplaca, os valores de atividade enzimática calculados apresentaram redução, acompanhando um aumento da diluição, conforme pode ser visualizado nas Figuras de 12 a 15 (Celluclast®) e nas Figuras de 16 a 19 (Cellic® CTec2).

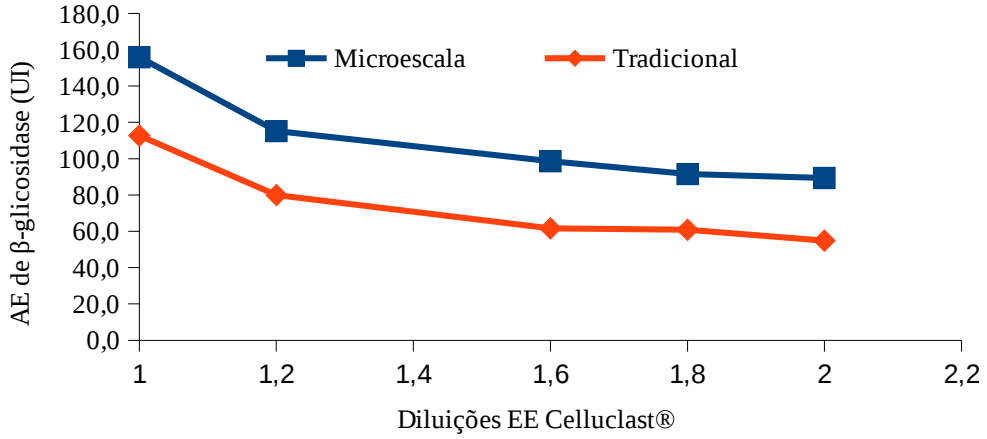

Figura 12 – Atividade de β-glicosidase obtida com base em diluições do extrato Celluclast®

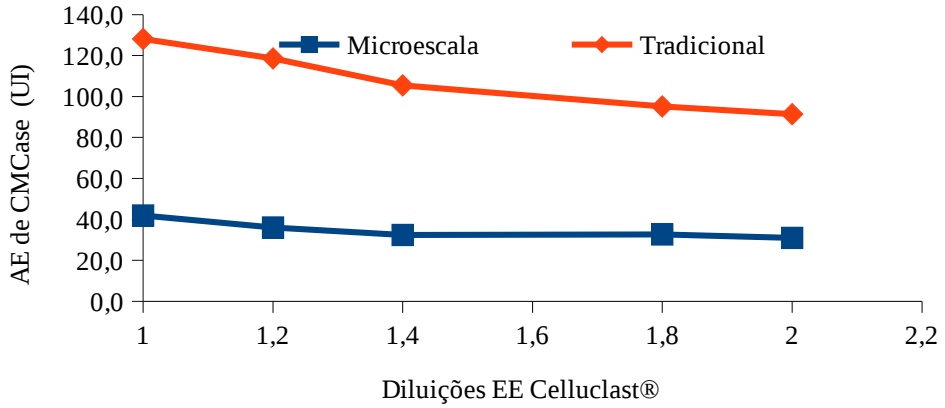

Figura 13 – Atividade de CMCase obtida com base em diluições do extrato Celluclast®

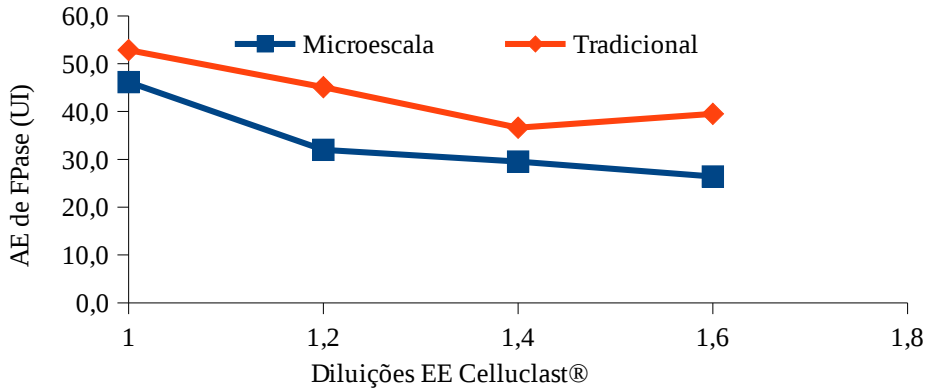

Figura 14 – Atividade de FPase obtida com base em diluições do extrato Celluclast®

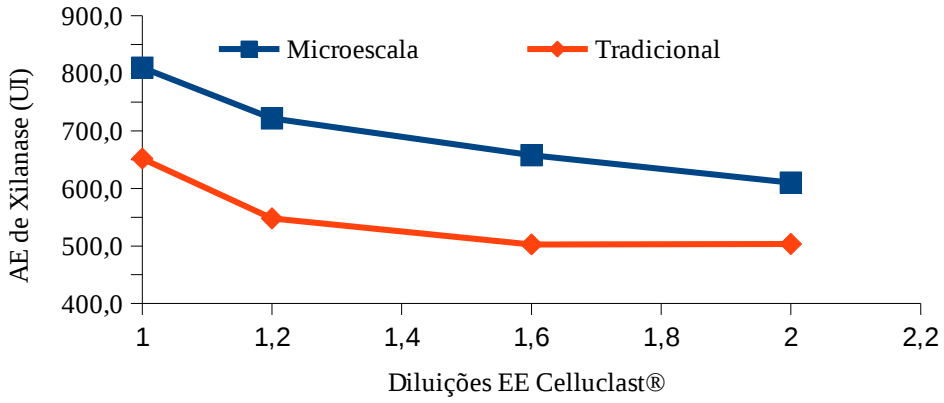

Figura 15 – Atividade de xilanase obtida com base em diluições do extrato Celluclast®

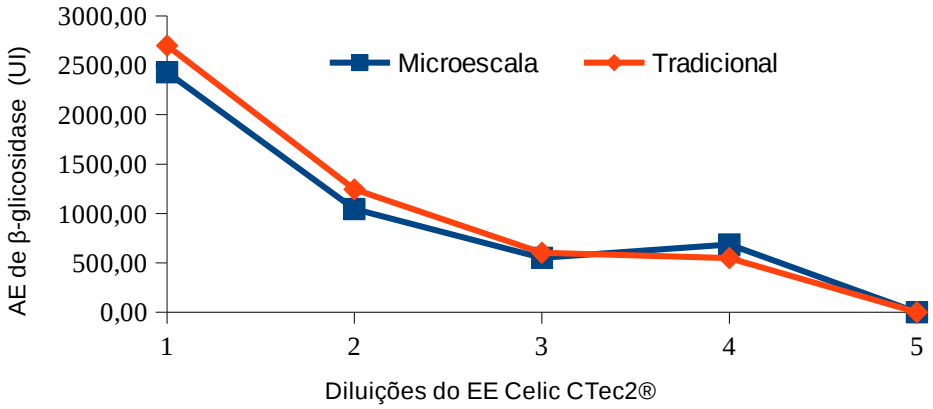

Figura 16 – Atividade de β-glicosidase obtida com base em diluições do extrato Celic CTec2®

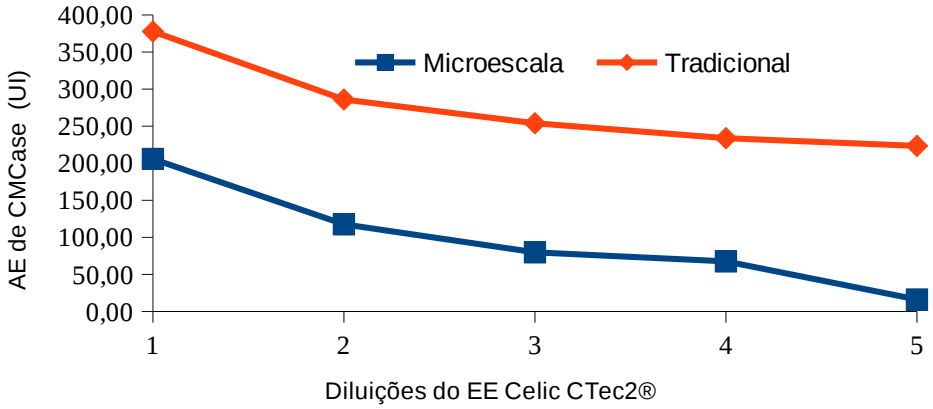

Figura 17 – Atividade de CMCase obtida com base em diluições do extrato Celic CTec2®

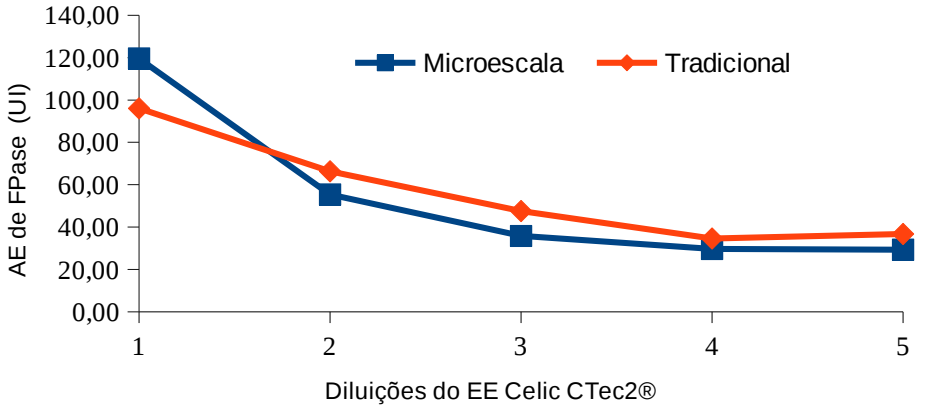

Figura 18 – Atividade de FPase obtida com base em diluições do extrato Celic CTec2®

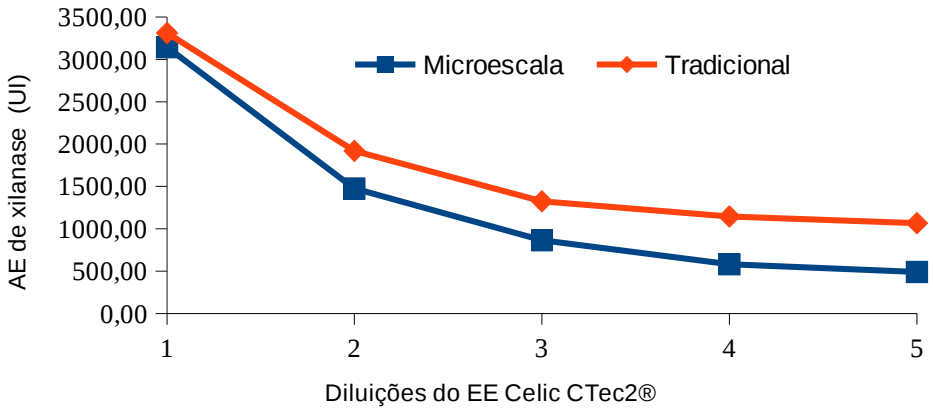

Figura 19 – Atividade de xilamase obtida com base em diluições do extrato Celic CTec2®

O comportamento decrescente de atividade enzimática foi observado tanto na metodologia de determinação enzimática tradicional (em tubo) quanto na metodologia adaptada (em microescala), conforme ilustrado nas figuras de 12 a 19, em ambos os complexos enzimáticos Celluclast® e Cellic CTec2®. Uma correlação positiva entre todas as determinações, superior a 0,9 foi determinada, conforme apresentada na Tabela 8.

Tabela 8 – Correlação entre os valores de atividades enzimáticas dos complexos Celluclast® e Cellic® CTec2

Atividade enzimática	Celluclast®	Cellic® CTec2
Fpase	0,92	0,96
Cmcase	0,93	0,97
β-glicosidase	0,99	0,95
Xilanase	0,95	0,99

Os resultados da análise estatística realizada sobre o conjunto de dados das médias de atividade enzimática para as enzimas β-glicosidase, CMCase, FPase e xilanase para os complexos enzimáticos Celluclast® e Cellic® CTec2 demonstraram que os dados seguem uma distribuição normal, segundo o teste de *Shapiro-Wilk*, e, pelo teste-t de *Student*, as médias do método tradicional frente o método em microescala não se diferem estatisticamente, com 95% de confiança.

5.4 Espectros gerados a partir dos complexos enzimáticos

Um total de 1391 espectros NIR foram gerados a partir dos 22 experimentos realizados. O total de espectros obtidos a partir dos complexos enzimáticos Celluclast®, Cellic CTec2® e EETA foram, nesta ordem, 752, 492 e 147. Da totalidade dos 1391 espectros, 908 continham valores de atividade enzimática para as quatro enzimas. Os 483 espectros restantes apresentaram valores de atividade para xilanases ou β-glicosidases. O perfil espectroscópico das amostras pode ser visualizado através dos gráficos apresentados na Figura 20, com destaque para as áreas de ruído (Figura 14 b e c). As amostras apresentaram duas intensas áreas de variância, caracterizadas como sendo áreas de ruído. Estas áreas abrangem as faixas de 1.862 a 2.032 nm e de 2.288 a 2.500 nm, do espectro eletromagnético.

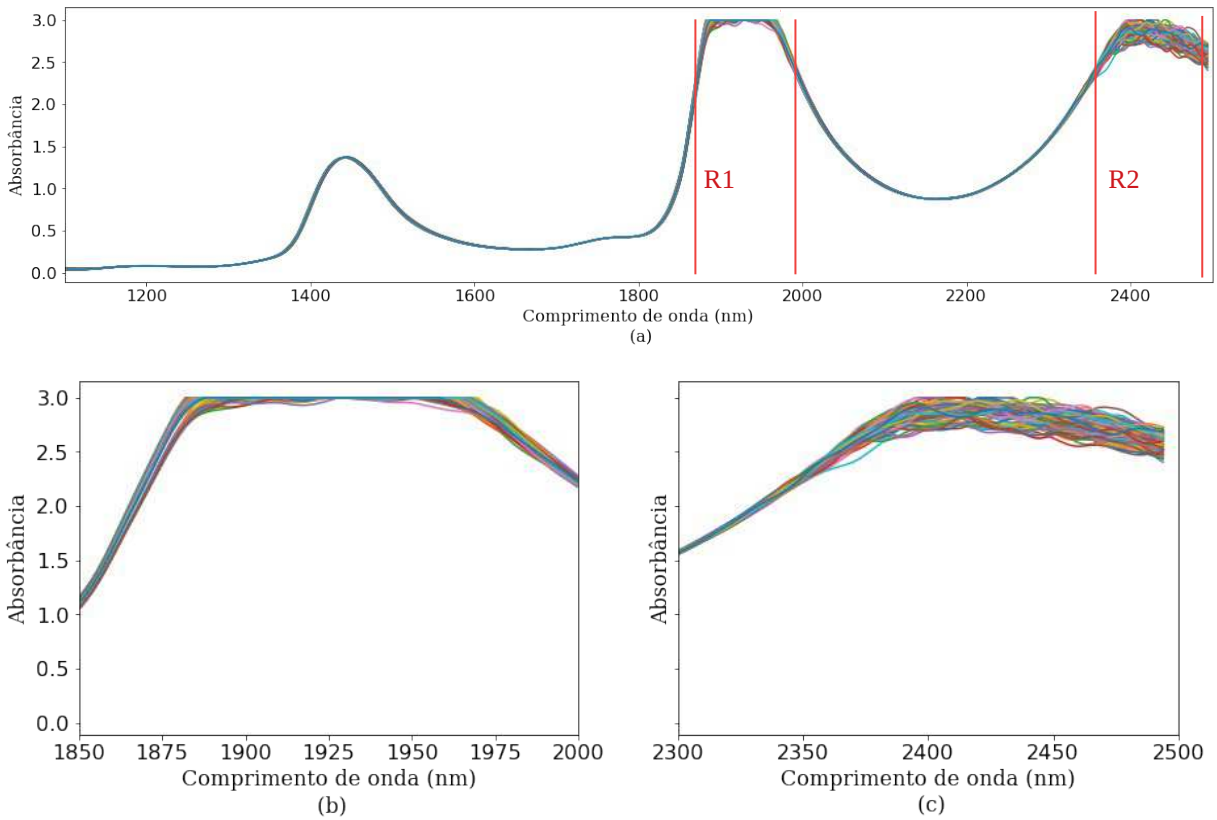

Figura 20 – Espectros de varredura no região do infravermelho próximo das amostras obtidas dos ensaios para determinação da atividade enzimática de β-glicosidases, CMCases, FPases e xilanases, contidas nos complexos Celluclast®, Cellic CTeC® e EETA; a) espectros evidenciando duas regiões de ruídos com destaque para as regiões R1 (b) e R2 (c);

Foi possível determinar, após utilização do algoritmo de seleção de atributos *mutual information*, as regiões dos espectros que apresentavam maior correlação com as atividades de cada enzima em estudo, β-glicosidases, CMCases, FPases e xilanases. Na

Tabela 9 são apresentadas as faixas de absorbância mais e menos significativas para a determinação de β-glicosidase, CMCase, FPase e xilanase.

Tabela 9 – Identificação das faixas de absorbâncias com alta ou baixa correlação frente às determinações enzimáticas

Enzima	Faixas de absorbância menos significativas (nm)*	Faixas de absorbância mais significativas (nm)**
β-glicosidase	1918 a 1976 e 2374	1396 a 1416 e 1842 a 1858
CMCase	1916 a1978 e 2366	1502, 1856 e 1858
FPase	1914 a 1976	1460 a 1494, 1856 a 1860
xilanase	1460 a 1498	1374, 1604 a 1636, 1810, 2136

* as faixas de absorbância menos significativas foram as que apresentaram correlação inferior a 0.3 para todas as enzimas;
**as faixas de absorbância mais significativas foram as que apresentaram correlação superior a 0, 75 para β-glicosidase e xilanase e 0,65 para CMCase e FPase.

De maneira geral, as faixas dos espectros com comprimento de onda entre 1914 nm e 1980 nm foram consideradas as menos relevantes para a determinação das celulases, enquanto as faixas de 1460 nm a 1498 nm foram as menos significativas para as xilanases.

Os modelos de calibração, validação e predição que resultaram nos melhores ajustes, ou seja, cujo coeficiente de determinação tenha sido o mais próximo de 1 e cujos erros (RMSEC, RMSEV e RMSEP) tenham sido os menores possíveis, foram gerados após a seleção das faixas de absorbância, combinando as mais significativas para a determinação de cada atividade enzimática. Ao comparar os resultados obtidos sem a seleção de atributos, isto é, utilizando todo o espectro do infravermelho, e após a seleção de atributos foi possível verificar que a seleção de atributos realizada resultou em um desempenho superior em todos os modelos testados.

Além da seleção de atributos também foi avaliada a utilização de Análise de Componentes Principais (PCA), a qual permitiu verificar o grau de variabilidade contido no conjunto total das amostras (1391 espectros) e em subconjuntos como, por exemplo, amostras oriundas do processo fermentativo. Foi possível, através dessa técnica, perceber além da variabilidade das amostras uma formação de *clusters* tanto em função do complexo enzimático utilizado quanto em relação ao experimento com suas variações de tempo de incubação e temperatura. Na Figura 21 são apresentadas a variância explicada e acumulada pelas componentes principais para o conjunto de 1391 amostras de espectros NIR.

A junção de três componentes principais (PC1, PC2 e PC3) foi capaz de explicar a variabilidade da maioria das amostras, sendo que a primeira componente (PC1) foi responsável por 47,08 % da variabilidade, a componente 2 (PC2) por contribuir com mais 22,9% e a componente 3 (PC3) por mais 10,1%, apresentando uma contribuição cumulativa na variabilidade dos dados de 81%.

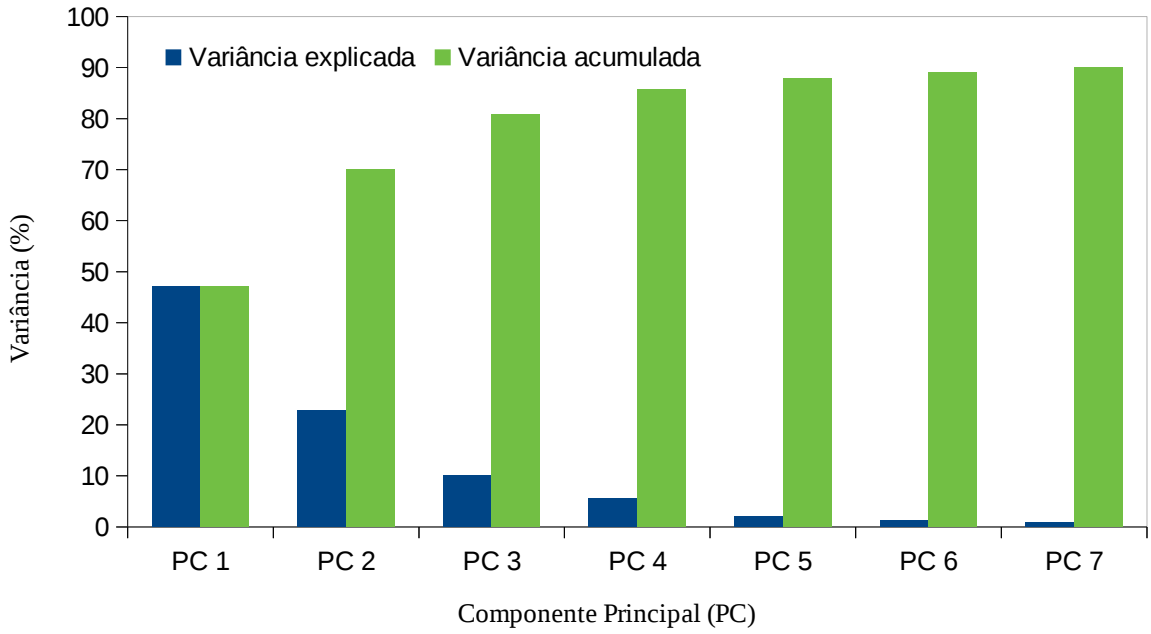

Figura 21 – Variância explicada e acumulada pelas componentes principais sobre o conjunto de 1391 espectros NIR

Observou-se a partir da análise dos gráficos dos *scores* das componentes PC1 com a PC2 e PC1 com PC3, ilustrados nas Figuras 22 e 23, a formação de três *clusters*, demonstrando que o uso da PCA permitiu identificar a separação do conjunto total dos espectros em subgrupos em função do complexo enzimático utilizado. Logo tanto a PC1 associada à PC2, quanto a PC1 associada à PC3 permitiram distinguir a origem das amostras, evidenciando que experimentos realizados com diferentes complexos enzimáticos promovem um distanciamento das amostras e evidencia a heterogeneidade do conjunto de dados. Verificou-se, ainda, que o conjunto de amostras derivadas do extrato EETA apresentou um distanciamento mais evidente das demais amostras, provavelmente em função do teor de enzimas contido neste extrato ser inferior aos demais complexos e não ser um extrato concentrado.

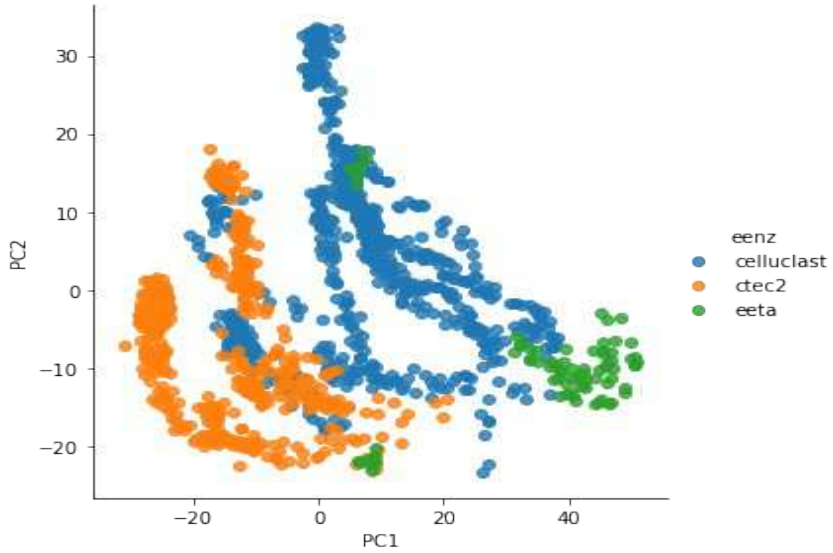

Figura 22 – Gráfico de *scores* da primeira componente principal (PC1) versus segunda componente principal (PC2) para todas as amostras utilizadas neste estudo.
eenz: extrato enzimático.

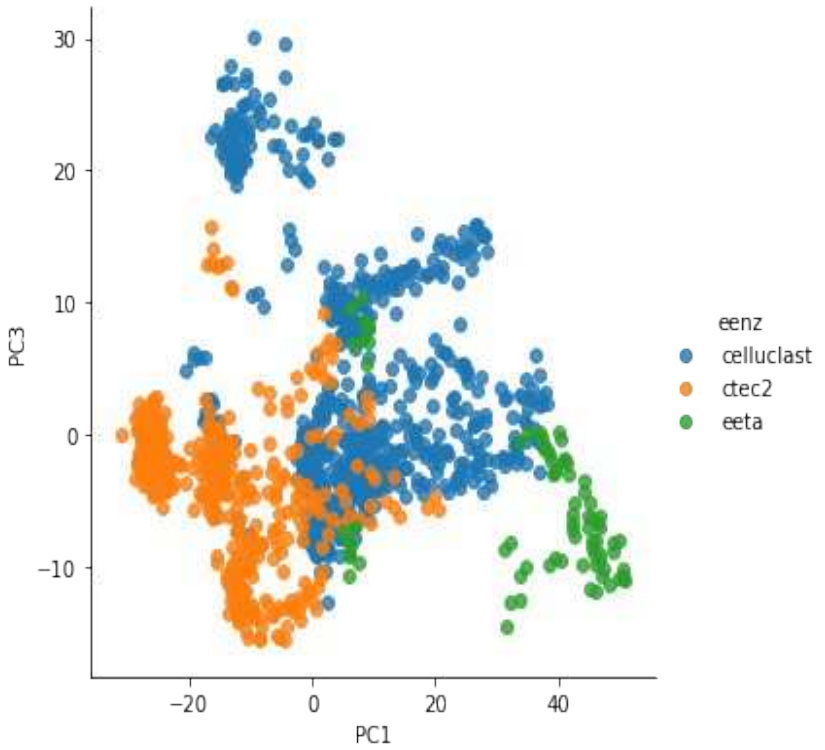

Figura 23 – Gráfico de *scores* da primeira componente principal (PC1) versus terceira componente principal(PC3) para todas as amostras utilizadas neste estudo.
eenz: extrato enzimático.

Outro tipo de agrupamento foi observado após aplicação de PCA ao conjunto dos 1391 espectros NIR, com foco nas temperaturas de desnaturação térmica utilizadas em cada experimento. Observou-se novamente uma separação das amostras em função da temperatura utilizada conforme ilustrado no gráfico da Figura 24.

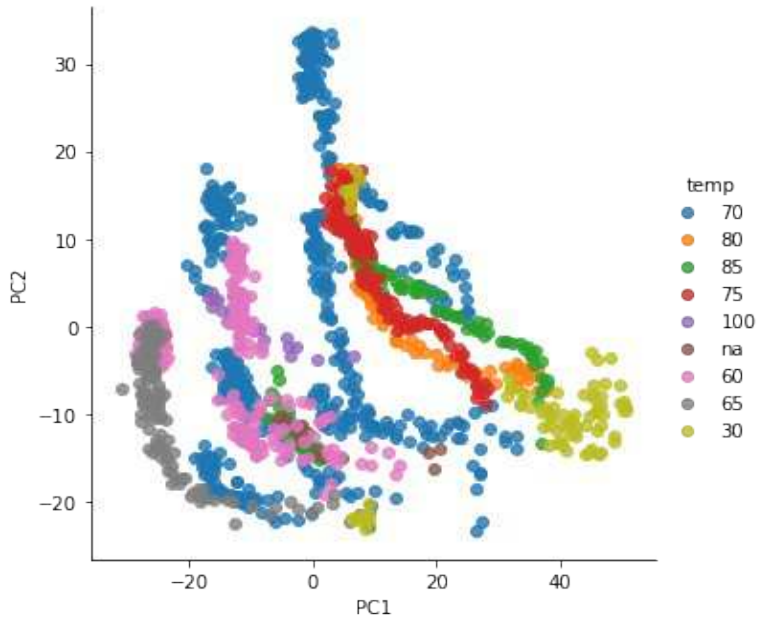

Figura 24 – Gráfico de *scores* da primeira componente principal (PC1) versus segunda componente principal (PC2) para todas as amostras utilizadas neste estudo, com foco na temperatura de desnaturação térmica
temp: temperatura

A temperatura utilizada durante os ensaios parece ter sido determinante para subdividir o conjunto de dados em subgrupos menores (Figura 24) bem definidos. Verificou-se, ainda, que o conjunto de amostras desnaturadas a 70°C apresentou um espalhamento maior das amostras. A formação de *clusters* tanto em função do extrato enzimático quanto da temperatura utilizada nos ensaios indicou que o conjunto completo das amostras poderia ser utilizado para construção de modelos preditivos para classificação de amostras em função de sua origem origem, ou seja, modelos que indiquem qual o tipo de complexo enzimático foi utilizado para a determinação enzimática.

Os resultados obtidos com a geração dos espectros NIR e as medições das atividades enzimáticas foram unidos para composição do conjunto de dados que norteou a geração de todos os modelos preditos, conforme apresentado na seção 4.8.2, e se encontram

disponíveis na plataforma *github*[1], juntamente com as implementações desenvolvidas, em linguagem *python*.

Os processos de calibração, validação e predição desenvolvidos neste trabalho são descritos na próxima seção.

5.5 Modelagem: calibração, validação e predição

5.5.1 Modelagem com auxílio da ferramenta proprietária: calibração e validação

A modelagem inicial realizada com o coquetel enzimático comercial Celluclast® como fonte das enzimas celulases e xilanases permitiu verificar a existência de correlação entre atividade enzimática e espectros NIR para um conjunto de 11 espectros em triplicata.

Os valores de atividade enzimática registrados para β-glicosidase, CMCase, FPase e xilanase no momento inicial da desnaturação estão apresentados na Tabela 10. Os dados coletados permitiram observar que o coquetel enzimático apresentou a maior quantidade de atividade de xilanase (1043 U mL^{-1}), seguida por β-glicosidase (254 U mL^{-1}), CMCase (155 U mL^{-1}) e FPase (68 U mL^{-1}). As taxas do coeficiente de variação (CV%) registradas nas determinações analíticas colorimétricas adaptadas à microplaca variaram de 2,1 a 5,9. A variância observada foi comparável àquelas observadas em outras propostas metodológicas e considerada aceitável para determinações da atividade enzimática.

Tabela 10 – Médias de atividade enzimática para β-glicosidase, CMCase, FPase e xilanase registrados para para o extrato enzimático Celluclast®

	Atividade enzimática (U mL^{-1})			
Valores	**β-glicosidase**	**CMCase**	**FPase**	**xilanase**
Min	248,61	144,64	67,13	1018,64
Max	273,58	162,12	69,85	1060,04
Média	257,74	155,06	68,88	1043,77
CV%	5,34	5,94	2,20	2,12

U: quantidade de enzima capaz de formar 1 µmol do produto avaliado por minuto nas condições de ensaio; CV: Coeficiente de variação

O processo de desnaturação térmica gradual permitiu a obtenção de amostras com valores de atividade enzimática diferentes dos apresentados na Tabela 10, produzindo uma variabilidade necessária para a construção dos modelos de calibração, com atividade enzimática mínima e máxima e coeficientes de variação (CV) de: β-glicosidase (83,9 a 285,9

[1] A base de dados utilizada neste trabalho pode ser acessada no sítio: https://github.com/amandarochac/tese

U mL^{-1}, CV = 0,32), CMCase (97,1 a 186,1 U mL^{-1}, CV = 0,25), FPase (14,2 a 81,3 U mL^{-1}, CV = 0,67) e xilanase (144,9 a 1088,5 U mL^{-1}, CV = 0,71). O CV para estas determinação representou a variação dos valores de atividade em todas as amostras durante todo o processo de desnaturação.

As mesmas amostras obtidas com o tratamento térmico produziram 33 espectros na faixa de 1500 a 2500 nm, quando foram submetidas à espectroscopia no infravermelho próximo, incluindo as repetições realizadas em triplicata. Para esse subconjunto de dados foi possível destacar duas áreas espectrais apresentando grande variação irregular de sinal em regiões que variam de 1862 a 2032 nm e de 2288 a 2500 nm. Essas duas regiões apresentam comportamento espectral típico de ruído e dispersão de sinal.

A modelagem inicial, utilizando o algoritmo dos quadrados mínimos parciais iterativos não lineares (NIPALS, *Non Linear Iterative Partial Leasts Quares*), levando em consideração todos o conjunto espectral, compreendendo todas as faixas de absorbância, de 1500 a 2500 nm resultou em coeficientes de determinação (R^2) muito baixos e os valores de erro quadrático médio (RMSE) muito altos (Tabela 11).

Tabela 11 – Resultados da calibração e validação para os modelos de predição de atividade enzimática de β-glicosidase, CMCase, FPase e xilanase por espectroscopia NIR utilizando o algoritmo PLS

PLS	**Medidas**	**Enzimas**			
		β-glicosidase	**FPase**	**CMCase**	**Xilanase**
Com todas as faixas de absorbância	**Amostras**	33	33	33	33
	Fatores	1	1	1	1
	R^2_{Cal}	0,534	0,489	0.518	0,455
	R^2_{Val}	0,508	0,453	0,463	0,424
	RMSEC	42,19	16,93	21,80	240,52
	RMSECV	47,04	12,37	24,26	266,67
	$Bias_{Val}$	0,42	-0,04	-0,003	-0,27
	RPD	1,31	1,26	1,29	1,22
	RER	4,29	5,42	3,67	3,54
Sem pré-processamento, com redimensionamento das faixas de absorbância	**Amostras**	33	33	33	33
	Fatores	4	4	4	4
	R^2_{Cal}	0,824	0,926	0,847	0,942
	R^2_{Val}	0,783	0,751	0,771	0,729
	RMSEC	25,93	6,42	12,27	78,28
	RMSECV	29,73	12,37	15,83	172,28
	$Bias_{Val}$	-0,91	-0,63	-0,55	-9,43
	RPD	2,07	1,91	1,98	1,89
	RER	6,66	5,35	5,53	5,40

Com pré-processamento e com redimensionamento de faixas de absorbância	**Amostras**	30	30	25	27
	Fatores	4	7	6	6
	R^2_{Cal}	0,935	0,959	0,987	0,963
	R^2_{Val}	0,857	0,871	0,920	0,839
	RMSEC	13,25	3,01	2,34	57,23
	RMSECV	21,46	8,06	6,63	124,47
	$Bias_{Val}$	-0,83	-0,32	0,07	-7,60
	RPD	2,88	2,61	4,72	2,61
	RER	7,04	8,19	11,91	7,45

R^2_{Cal} = coeficiente de determinação para a calibração; R^2_{Val} = coeficiente de determinação para a validação; PLS: *Partial Least Square*; RMSEC= *root mean square error of calibration*; RMSECV= *root mean square error of cross validation*; RPD= *Ratio of Performance to Deviation*; RER= *Range Error Ratio*.

A primeira estratégia usada para tornar o conjunto de preditores mais apropriado para a modelagem residia na remoção da região de ruído. Essa região ficou evidente devido à aplicação da transformação dos dados por derivada de segunda ordem, que, por sua vez, levou ao aumento da relação ruído-sinal. Os dados espectrais foram dimensionados para a faixa de 1100 a 1830 nm, levando a uma melhora significativa nos valores registrados para R^2 e RMSE quando o algoritmo PLS (NIPALS) foi executado com o conjunto de dados redimensionado, embora sem qualquer processamento adicional.

O pré-processamento de dados espectrais é a etapa mais importante antes da modelagem bi-linear quimiométrica (RINNAN *et al*, 2009), portanto, além do redimensionamento dos dados espectrais, as combinações de pré-processamento matemático de dados e remoção de *outliers* realizados resultaram em uma melhoria dos modelos de regressão tanto nas etapas de calibração quanto validação. Fenômenos como mudanças de linha de base entre amostras, efeitos de dispersão e outros ruídos inespecíficos e aleatórios reduzem a relação sinal-ruído (SNR), afetam a resolução dos espectros e perturba a exatidão e precisão de um modelo de calibração (XU *et al*., 2008; KOONJAH *et al*., 2019). Neste estudo, foram combinados tratamentos de suavização (transformada de Savitzky-Golay), normalização por desvio padrão dos sinais (SNV) e remoção de deslocamentos sistemáticos inerentes ao equipamento utilizado ou à matriz da amostra (*Detrend*). Os *outliers* foram identificados pela análise do gráfico de pontuação (elipse *Hoetlling* T^2) a partir da modelagem de cada atividade enzimática, gerada com o número ótimo de fatores, e nível de significância de 5%. Uma combinação da transformada de Savitzky-Golay, alisamento polinomial de

segunda ordem com janela de 5 pontos, seguida da transformada de variável normal padrão (SNV) resultou no melhor ajuste do modelo de regressão por PLS para a predição da atividade de β-glicosidase. A sequência SNV e a transformada polinomial *Detrend* de quarta ordem foi a combinação mais adequada para modelar a regressão PLS das atividades de CMCase, Fpase e xilanase.

Todos os modelos de regressão foram gerados variando o número de fatores de 1 a 7. O aumento do número de fatores (número de variáveis latentes) tende aumentar o coeficiente de determinação. No entanto, esse aparente aumento na qualidade do modelo de regressão também significou um aumento nos valores de erros, algo que só é útil para ajustar as observações do conjunto de treinamento e não é útil para ajustar novas observações (SOUZA E FERRÃO, 2006; ABDI, 2010). A determinação do número ótimo de fatores para cada modelagem foi realizada através de análise da variância residual versus a curva de erros (RMSE) gerado pelo *software Unscrambler*, observando a quebra da variância residual monotonicamente decrescente. Após redimensionamento, pré-processamento de preditores (dados NIR) e remoção de *outliers*, foi observado que o número de fatores ótimos para modelagem de regressões aumentou de 4 para 6 e de 4 para 7 para atividades CMCase e xilanase e para a atividade FPase, respectivamente. Mesmo assim, os valores dos erros RMSE diminuíram e os valores RPD e RER aumentaram sensivelmente. Foi possível manter em todos os casos o valor de no mínimo quatro vezes o número de amostras em relação ao número de fatores selecionados para as previsões de validação cruzada. De acordo com Baum *et al*. (2013b), para obter bons ajustes no modelo de regressão, é desejável que o número de amostras seja de no mínimo 3 vezes maior que o número escolhido de variáveis latentes.

Os modelos PLS ajustados para as previsões das atividades enzimáticas avaliadas no presente trabalho por validação cruzada alcançaram valores de RPD> 2,5. De acordo com Viscarra Rossel *et al* (2006) valores de RPD> 2,5 indicam excelentes modelo, úteis para previsões quantitativas. Além disso, os modelos de previsão das atividades de FPase e CMCase obtiveram valores de RER de 8,2 e 11,9, respectivamente. De acordo com os limiares fornecidos por Malley *et al*. (2004) valores de RER compreendidos entre 10 (inclusive) e 15 são considerados modelos moderadamente bem-sucedidos e que 8 ≤ RER <10 são considerados modelos de previsão moderadamente úteis.

5.5.2 Modelagem com auxílio de software livre: calibração, validação e predição

Os resultados encontrados, após a realização da modelagem preliminar, descrita na seção anterior, demonstraram a existência de forte correlação entre os espectros NIR (variáveis preditoras - valores de absorbância) e as variáveis preditas (atividade enzimática de β-glicosidase, CMCase, FPase e xilanase). Como o conjunto de dados utilizados na modelagem inicial foi extremamente pequeno, apenas 11 amostras, não foi possível realizar a etapa de predição. O conjunto maior dos dados foi utilizado, portanto, para geração de modelos diversos em uma segunda etapa, cujos resultados são apresentados nesta seção.

O mecanismo de calibração para geração dos modelos de regressão, utilizando os diferentes algoritmos de regressão, GBR, PLSR, Ridge, Kernel-Ridge, PLS e PCR, envolveu a definição de parâmetros de entrada específicos para cada algoritmo utilizado. Alguns parâmetros como tipo de pré-processamento e número de componentes principais foram aplicados na maioria dos modelos. Os modelos que resultaram em boa capacidade preditiva produziram baixos valores de erros RMSEC, RMSEV e RMSEP e um coeficiente de determinação elevado, próximo a 1.

As maiores taxas de desempenho dos modelos foram obtidas após o pré-processamento padronização e suavização, utilizando janela 3, grau do polinômio 2 e 1ª derivada. Os outros pré-processamentos produziram um *overfitting* dos modelos pois resultaram em valores altos de R^2 para calibração e valores próximos a zero para validação e predição, levando a modelos com baixo valor preditivo, o que significa que os modelos gerados com esses pré-processamentos se especializaram nas amostras usadas para o treinamento, resultando em modelos com baixo poder de generalização.

Os resultados obtidos para o coeficiente de determinação R^2, ao aplicar os seis algoritmos descritos anteriormente ao conjunto de amostras total sem prévio pré-processamento matemático, revelaram alta correlação entre espectro e atividade enzimática para todas as enzimas avaliadas, no entanto quando esses mesmos algoritmos foram aplicados validação cruzada e nos dados de teste, a correlação se aproximou de zero, o que revela um superajustamento dos modelos aos dados durante a calibração, tornando o uso dessas técnicas inviáveis sem o devido pré-processamento. Em contrapartida, os resultados se tornaram relevantes quando o pré-processamento foi realizado. Além disso, a precisão aumentou à

medida que foram feitas a exclusão das faixas de ruído dos espectros ou a seleção das faixas mais correlacionadas às determinações enzimáticas através do algoritmo de seleção de atributo *mutual information*, conforme pode ser visualizado no Anexo I, no qual são apresentados os resultados de desempenho para todos os modelos gerados para as etapas de calibração, validação e predição, após a remoção de faixas de ruído (Tabelas 13 a 18) e após seleção de atributos (Tabelas 19 a 24).

A construção de modelos de predição com base apenas no conjunto de dados derivados do processo fermentativo resultou em modelos com alto desempenho. Na Tabela 12 são apresentados os pré-processamentos utilizados para esse conjunto de dados que apresentaram os resultados mais otimizados para calibração, validação e predição para cada modelo escolhido.

Tabela 12 – Pré-processamentos utilizados no conjunto de dados da fermentação que otimizaram os modelos nas etapas de calibração, validação e predição

Modelo	**Pré-processamento**
GBR	MSC
PLS	Padronização seguida por suavização(SavGol), com janela de tamanho 3, polinômio de 1º grau e 1ª derivada, seguida de padronização
Ridge	Padronização
Kernel-Ridge	Suavização(SavGol), com janela de tamanho 3, polinômio de 1º grau e 1ª derivada, seguida de padronização
PCR	Padronização
MLR	MSC

GBR (Gradient Boosting Regressor); PLSR (Partial Least Squared Regression); Ridge e Kernel Ridge: quadrados mínimos lineares com e sem *Kernel;* PCR:Principal Component Regression; MLR (Multiple Linear Regression); MSC (correção do espalhamento multiplicativo de sinal)

Ocorreu uma predominância do pré-processamento padronização, seguido ou não de suavização como a alternativa que permitiu elevar ao máximo o coeficiente de determinação quando esse tipo de pré-processamento é aplicado na maioria dos algoritmos de regressão. A suavização consistiu em utilizar uma janela de tamanho três, polinômio de garu 1 e primeira derivada. Os pré-processamentos restantes produziram *overfitting* dos modelos pois resultaram em valores altos de R^2 para calibração e valores próximos a zero para validação e predição, gerando modelos com baixo valor preditivo.

Os modelos de calibração, validação e predição para a determinação múltipla das atividades enzimáticas de β-glicosidases, cmcase, fpases e xilanases, utilizando os dados do complexo EETA, após pré-processamento otimizado, resultaram em valores de desempenho que podem ser visualizados no ANEXO II (Tabelas 25 a 30). Dentre as métricas utilizadas

para avaliação dos modelos gerados, as métricas R^2 e RMSE são ilustradas nas Figuras de 25 a 36 nas etapas de calibração, validação e predição para a determinação de β-glicosidase (Figuras 25, 26 e 27), CMCases (Figuras 28, 29 e 30), Fpases (Figuras 31, 32 e 33) e xilanases (Figuras 34, 35 e 36).

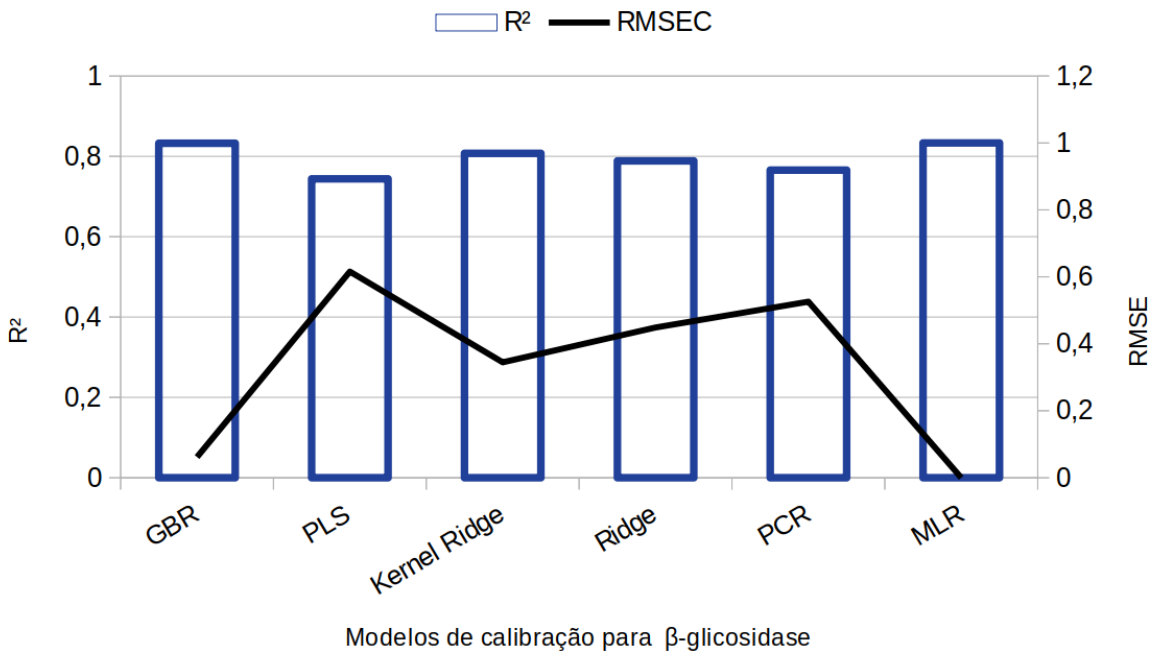

Figura 25 – Resultados de calibração para determinação de β-glicosidase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

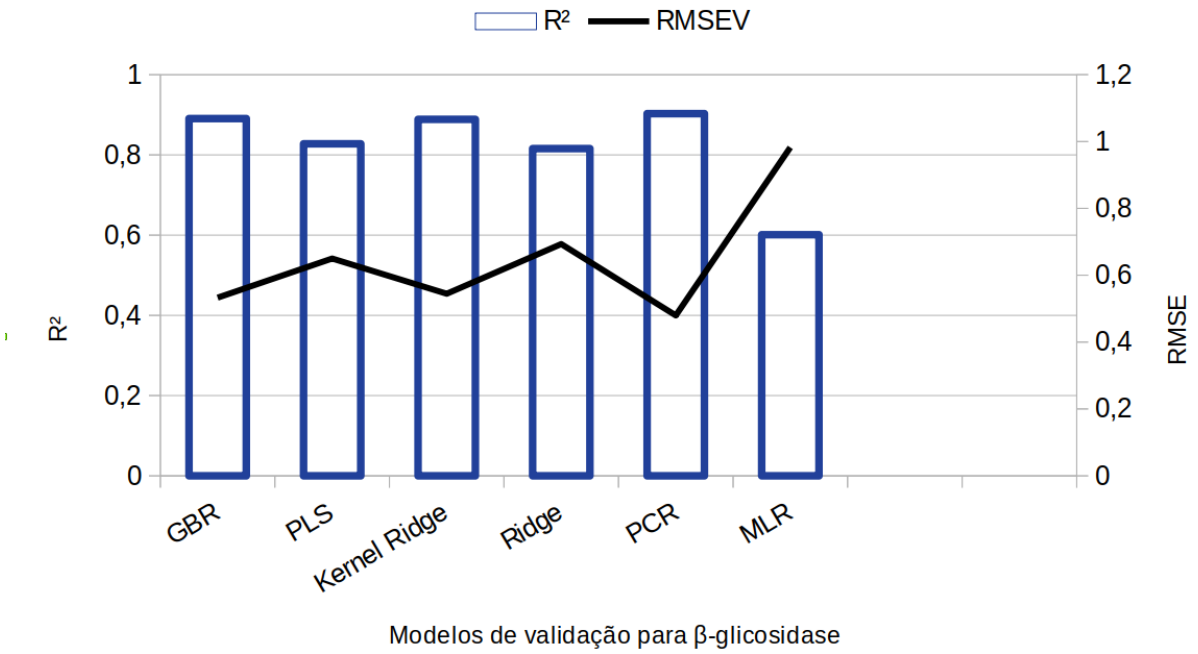

Figura 26 – Resultados de validação para determinação de β-glicosidase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

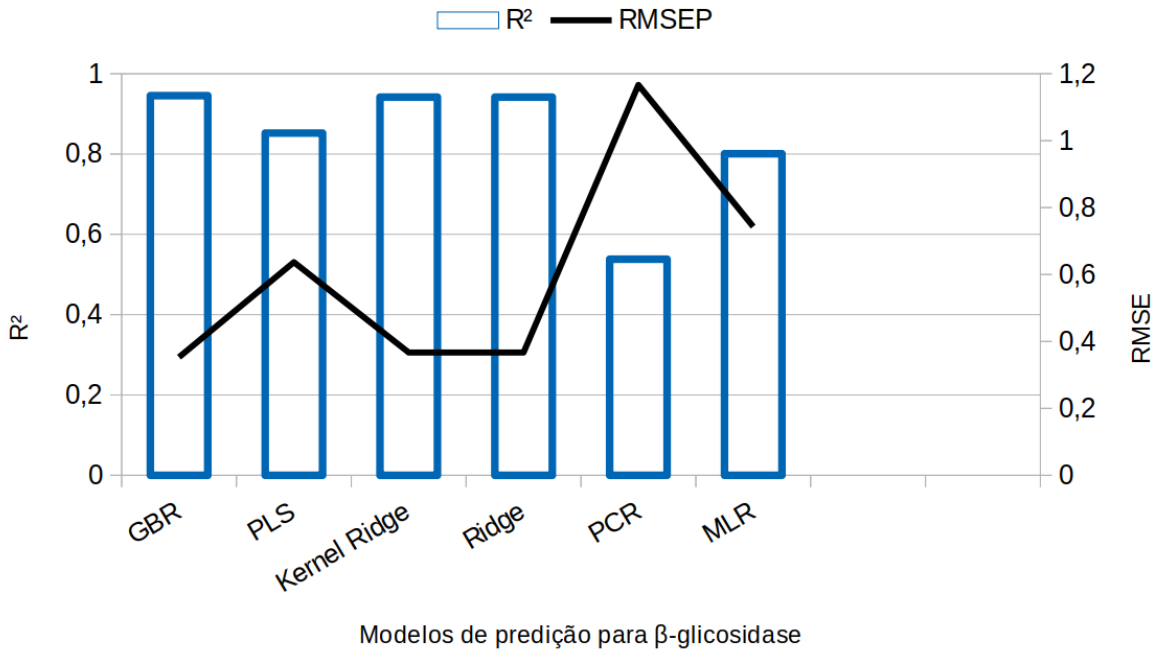

Figura 27 – Resultados de predição para determinação de β-glicosidase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

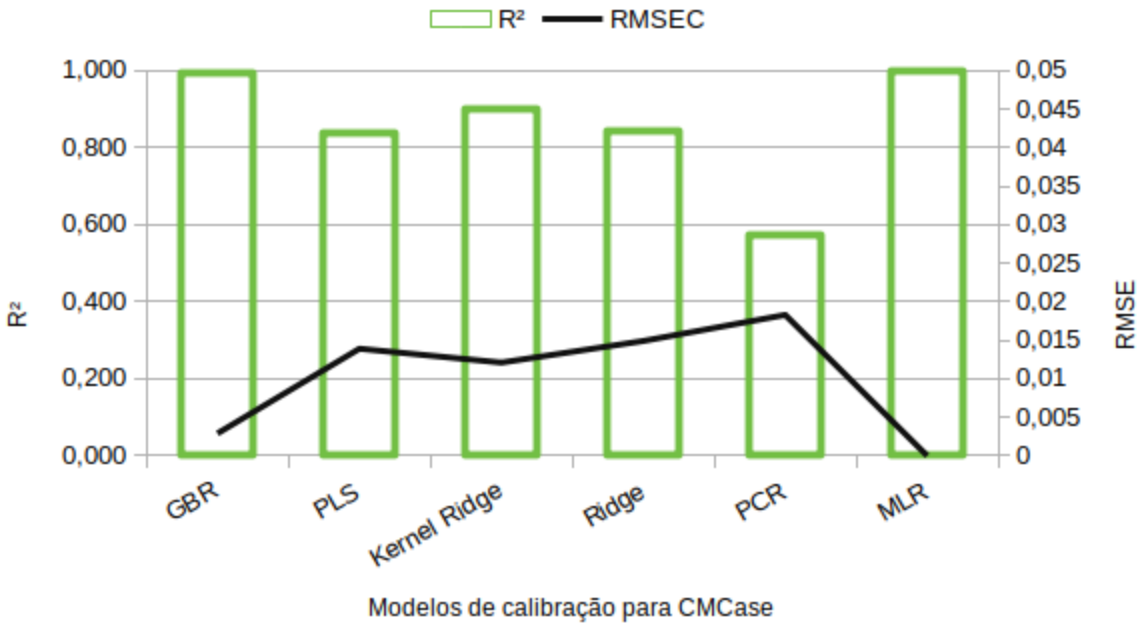

Figura 28 – Resultados de calibração para determinação de CMCase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

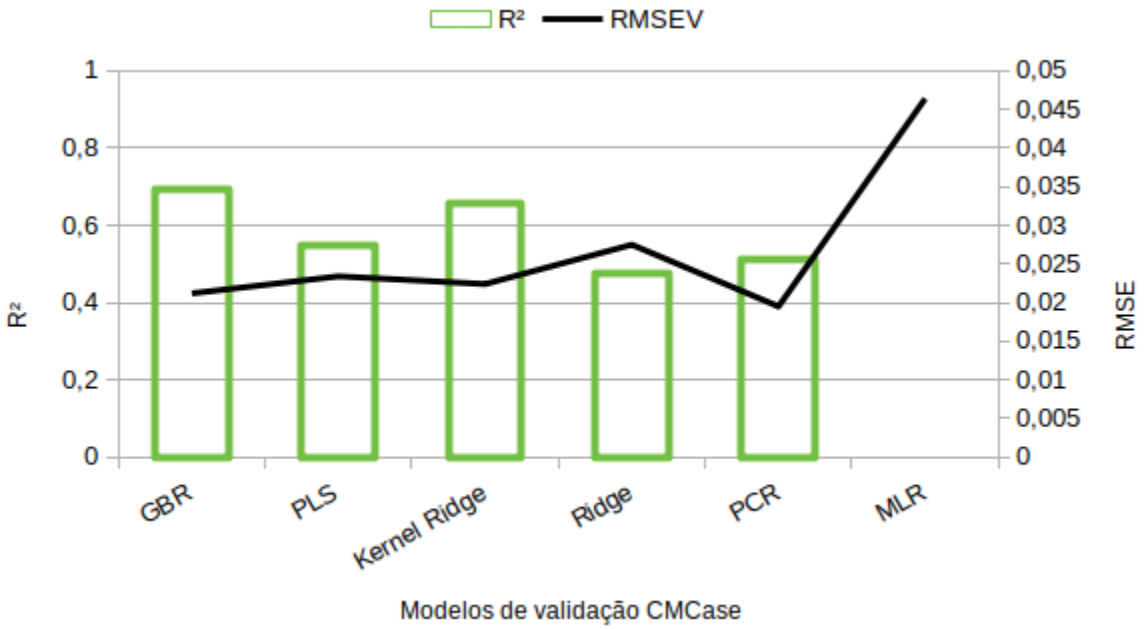

Figura 29 – Resultados de validação para determinação de CMCase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

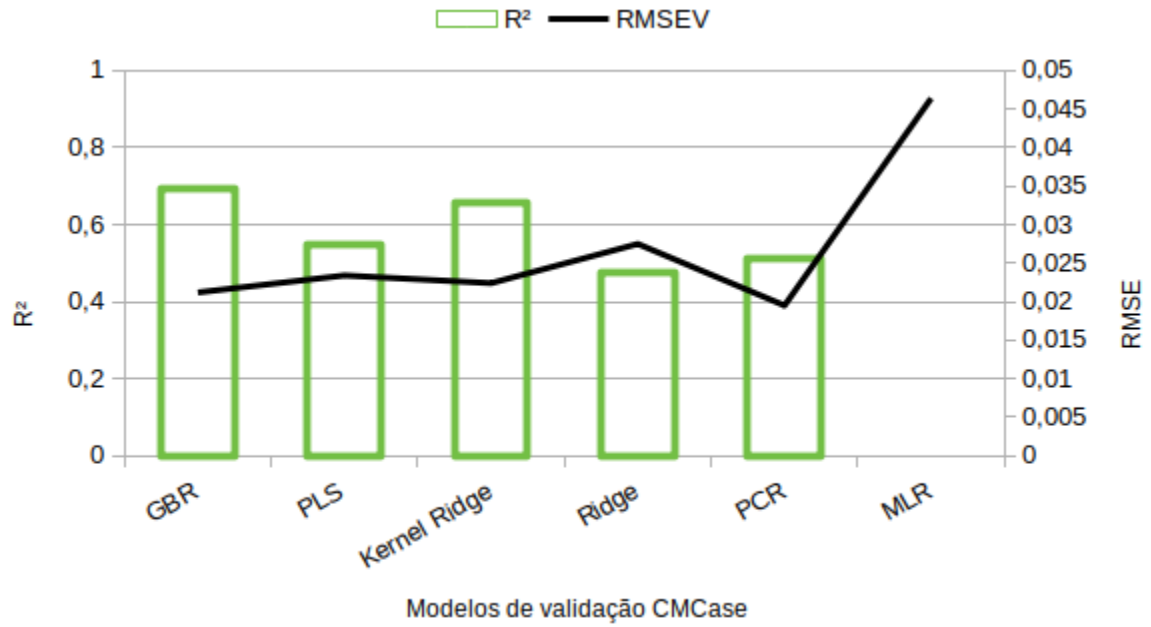

Figura 30 – Resultados de predição para determinação de CMCase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

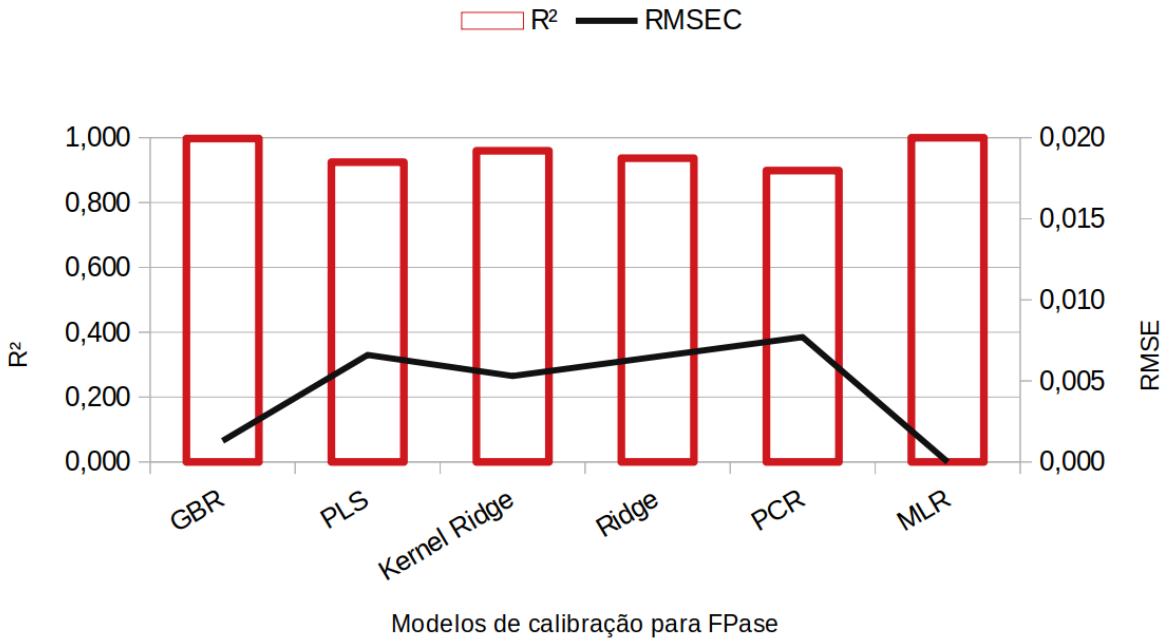

Figura 31 – Resultados de calibração para determinação de FPase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

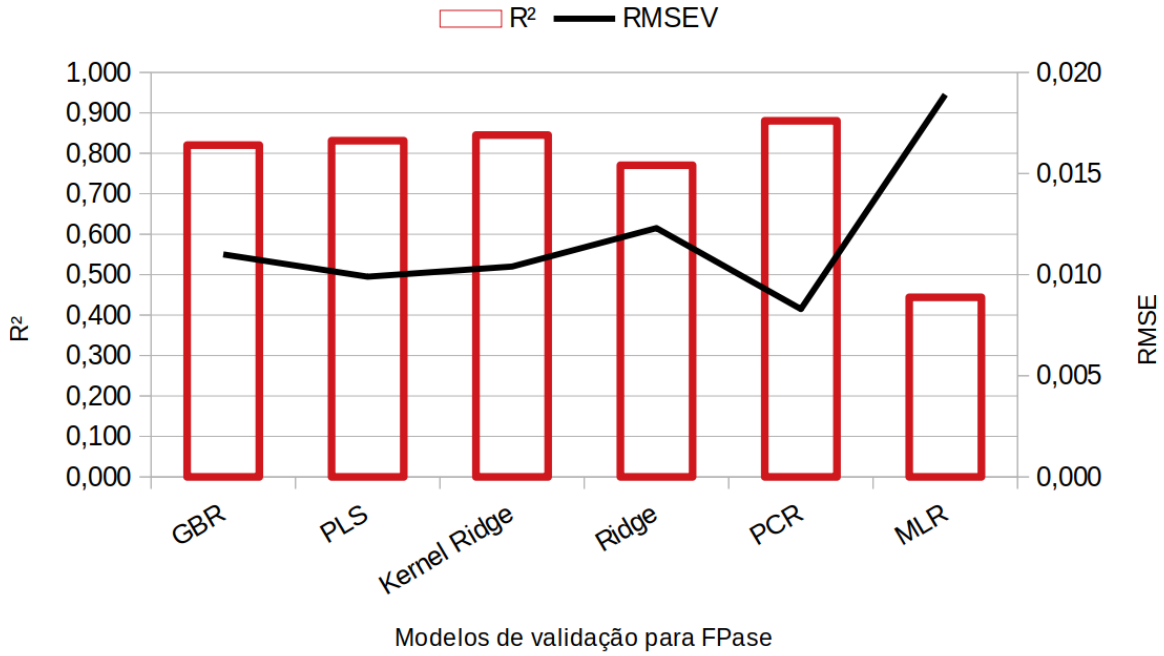

Figura 32 – Resultados de validação para determinação de FPase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

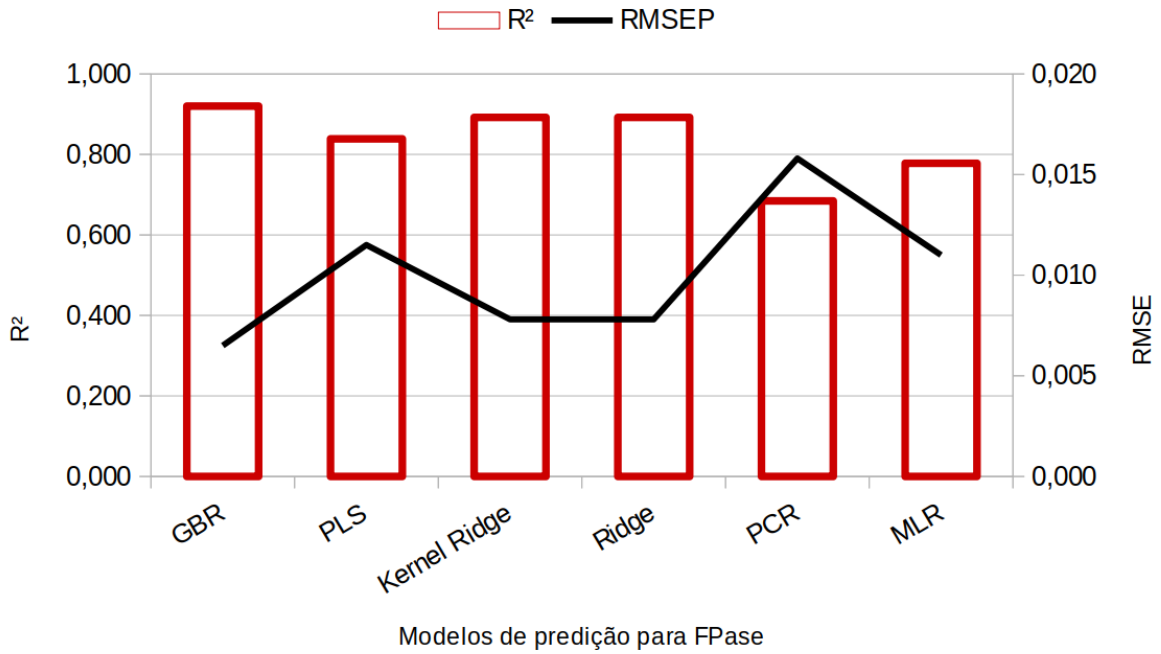

Figura 33 – Resultados de predição para determinação de FPase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

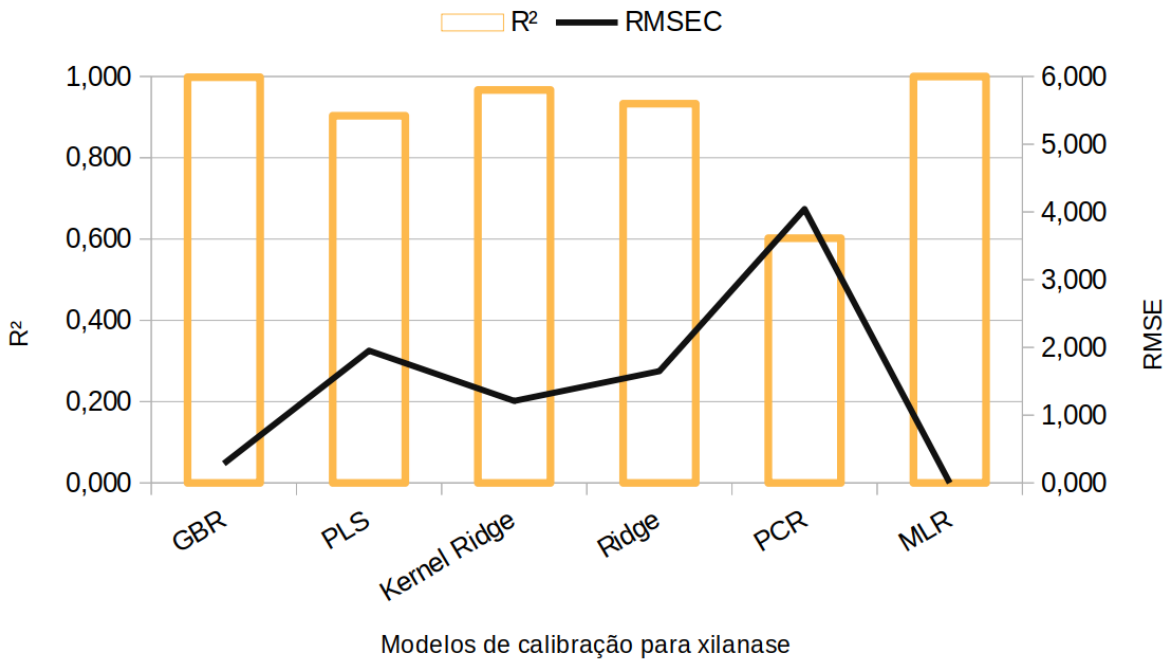

Figura 34 – Resultados de calibração para determinação de Xilanase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

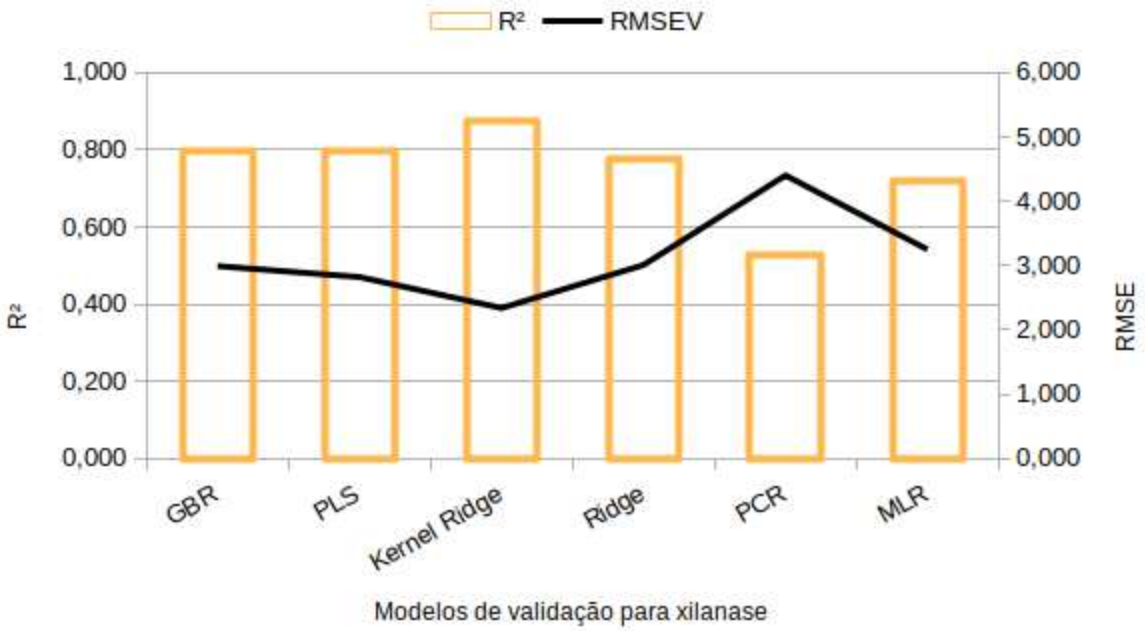

Figura 35 – Resultados de validação para determinação de Xilanase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

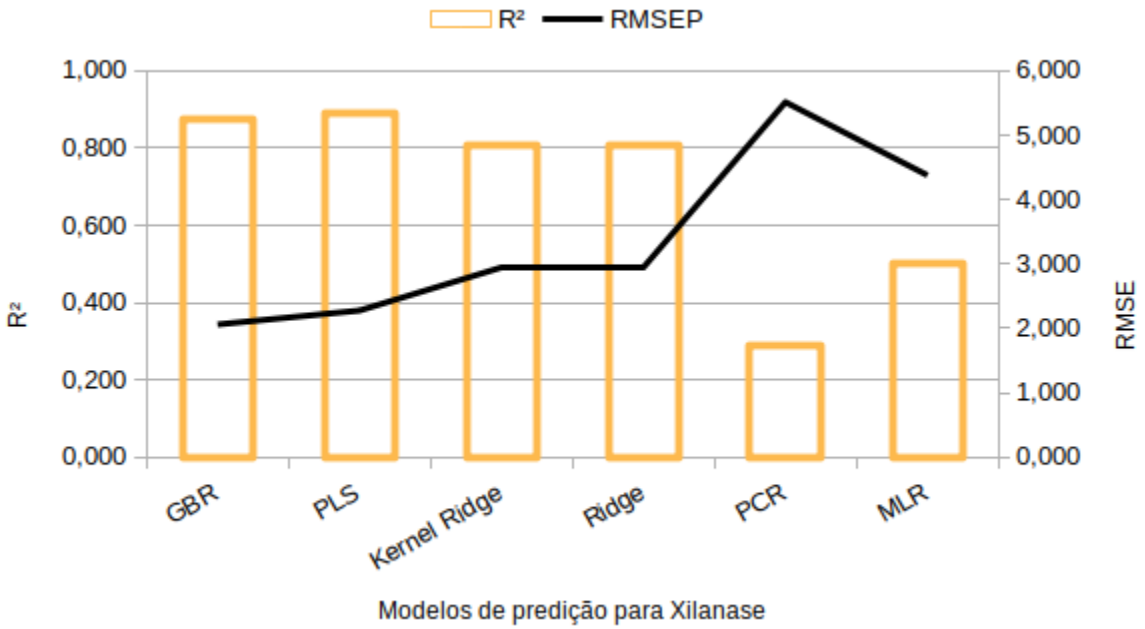

Figura 36 – Resultados de predição para determinação de Xilanase - R^2 e RMSEC obtidos dos modelos Gradiente Descendente (GBR), regressão por mínimos quadrados parciais (PLS), regressão por *Kernel Ridge*, regressão por *Kernel*, regressão por componentes principais (PCR) e regressão Linear Múltipla (MLR)

Os coeficientes de determinação para a calibração (R^2_{Cal}) foram elevados na grande maioria dos modelos, com valores superiores a 0,95, evidenciando a existência de uma forte correlação entre os dados de referência (atividade enzimática), obtidos por metodologia baseada em DNS e GOD-POD, e os espectros gerados com espectrofotômetro no infravermelho próximo.

O melhor desempenho geral da calibração foi facultado ao modelo GBR seguido pelos modelos MLR, contudo modelos de regressão MLR resultaram no menor desempenho na etapa de predição, possivelmente por ter ocorrido um superajustamento aos dados, levando a pouco ou nenhum poder de generalização. O modelo PCR foi o que resultou no menor desempenho para a determinação da atividade xilanásica, e na maior taxa de erro padrão RMSE para todas as outras atividades enzimáticas.

Os modelos obtidos na validação cruzada, excetuando-se a determinação de atividade enzimática CMCásica, a qual foi a menor para todas as enzimas, apresentaram coeficiente de correlação (R^2_{Val}) entre os valores 0,6 e 0,89, evidenciando uma redução na correlação, o que implicou em um aumento nos erros relacionados às estimativas dos modelos, conforme demonstrado nos valores de RMSEC(raiz quadrada do erro quadrático

médio da calibração) comparados aos valores do RMSEV (raiz quadrada do erro quadrático médio da validação).

A validação externa ou predição para os modelo considerados de melhor desempenho, o GBR, PLS, Kernel-Ridge e Ridge, na qual o modelo gerado foi testado com amostras que não participaram do treinamento e da validação cruzada, apresentou coeficiente de determinação (R^2_{Val}) na faixa de 0,85 a 0,94 para β-glicosidases, 0,68 a 0,87 para CMCase, 0,83 a 0,92 para FPases e 0,80 0,92 para xilanases. A predição, utilizando o modelo GBR, apresentou uma redução de R^2 inferior a 0,6 para a enzima β-glicosidase se comparada ao valor encontrado para a calibração. Essa redução do desempenho pode ser confirmada com o aumento nos erros relacionados às estimativas dos modelos, conforme demonstrado nos valores de RMSEC (raiz quadrada do erro quadrático médio da calibração) comparados aos valores do RMSEV (raiz quadrada do erro quadrático médio da validação) e do RMSEP (raiz quadrada do erro quadrático médio da predição).

Os valores de desempenho para o modelo MLR para o conjunto de dados da fermentação foram considerados insatisfatórios para determinação de CMCases e FPases, pois esse algoritmo apresentou um superajustamento dos dados, com R^2 elevado na calibração e validações reduzidas, abaixo de 0.5. Vale ressaltar que para o extrato EETA, o qual foi avaliado no processo fermentativo, o teor dessas enzimas foi considerado muito baixo se comparado ao teor de β-glicosidases e xilanases.

O modelo PCR resultou no pior desempenho, com R^2 abaixo de 0,5 na etapa de predição, para a determinação de três das atividades enzimáticas.

A capacidade de predição dos modelos gerados também foi determinada através da medida RPD (Relação do desempenho do desvio). Esta medida permitiu avaliar o quanto o modelo utilizado pode distinguir entre amostras diferentes, sendo que valores de RPD acima de 1,5 indicam que o erro do valor estimado é menor que o desvio padrão do valor real da amostra (LAZZAROTTO *et al*., 2016). Essa medida, portanto, é considerada como determinação do padrão de qualidade de um modelo e, segundo Lazzarotto *et al*. (2016), quando o RPD está acima de 2,5 significa que o modelo tem boa capacidade de predição, quando ultrapassa a razão 5, o modelo que o gerou poderá ser usado para controle de qualidade e, ultrapassando o valor 8, o modelo poderá ser usado para qualquer tipo de aplicação.

Os valores calculados para a métrica RPD no modelo GBR na etapa de validação cruzada foram 3,02, 1,8, 2,35 e 2,29 para as atividades enzimáticas de β-glicosidase, CMCase, FPase e xilanase, nesta ordem. Na predição, os valores determinados para RPD para essas mesmas atividades enzimáticas foram: 4,28, 2,8, 3,53 e 2,8. Observou-se, portanto, que para a determinação das enzimas β-glicosidase, CMCase, FPase e xilanase, os modelos GBR e Kernel Ridge são considerados bons preditores. Os modelos Ridge e PLS poderiam ser utilizados para as determinações de β-glicosidase, FPase e xilanase e os modelos PCR e MLR apenas para triagem.

Todos os modelos gerados possibilitaram, também, realizar o acompanhamento do processo fermentativo, conforme apresentado nas Figuras 37 a 48. Nessas figuras é apresentado o comportamento dos modelo GBR (Figuras 37 e 38), *Kernel Ridge* (Figuras 39 e 40), *Ridge* (Figuras 41 e 42), PLS (Figuras 43 e 44), PCR (Figuras 45 e 46) e MLR (Figuras 47 e 48) para determinação das atividades enzimáticas de β-glicosidase, CMCcase, FPase e xilanase durante a calibração, validação e predição. Os gráficos apresentam os valores preditos versus os valores reais.

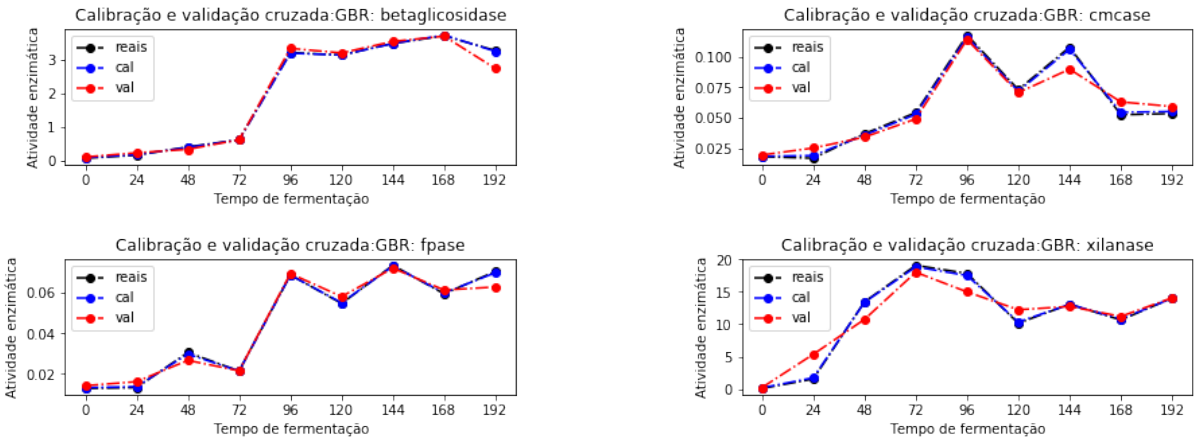

Figura 37 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo GBR

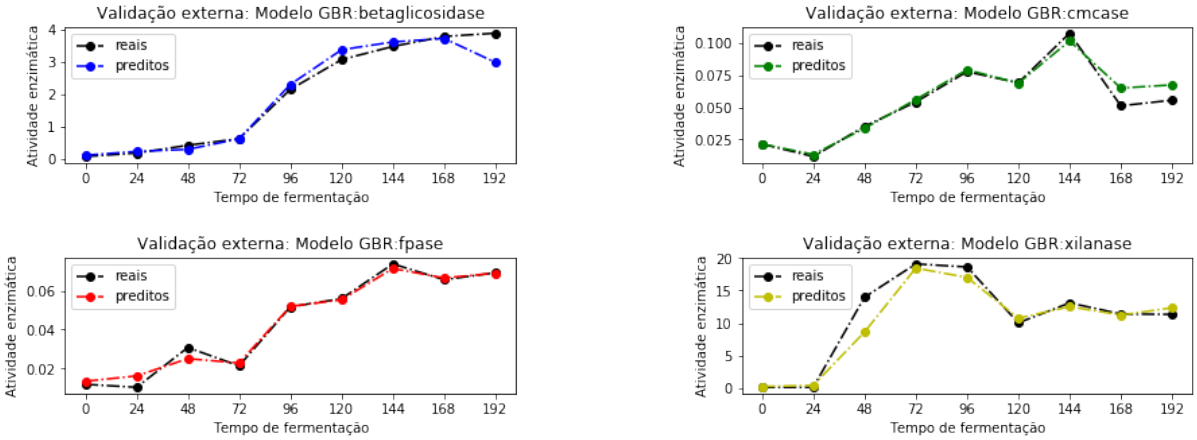

Figura 38 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo GBR

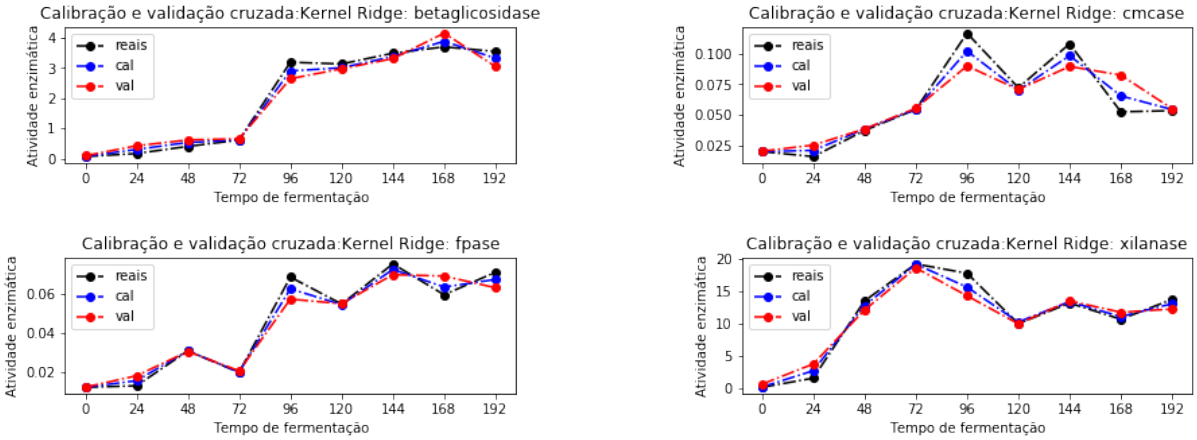

Figura 39 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo *Kernel-Ridge*

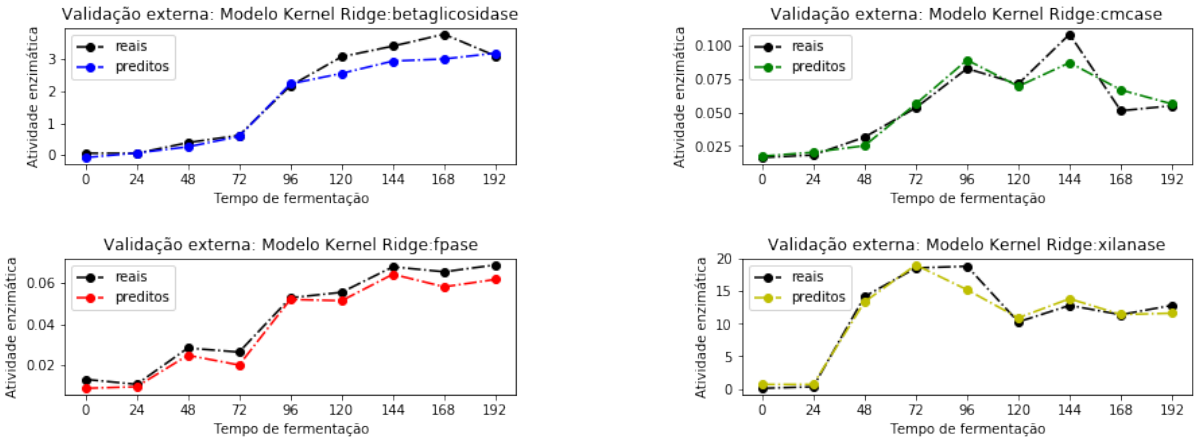

Figura 40 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo *Kernel-Ridge*

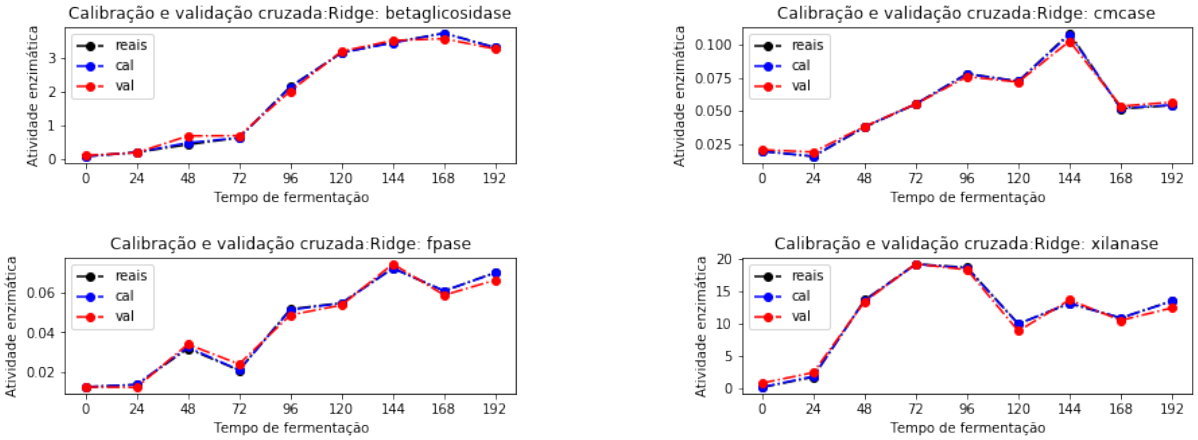

Figura 41 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo *Ridge*

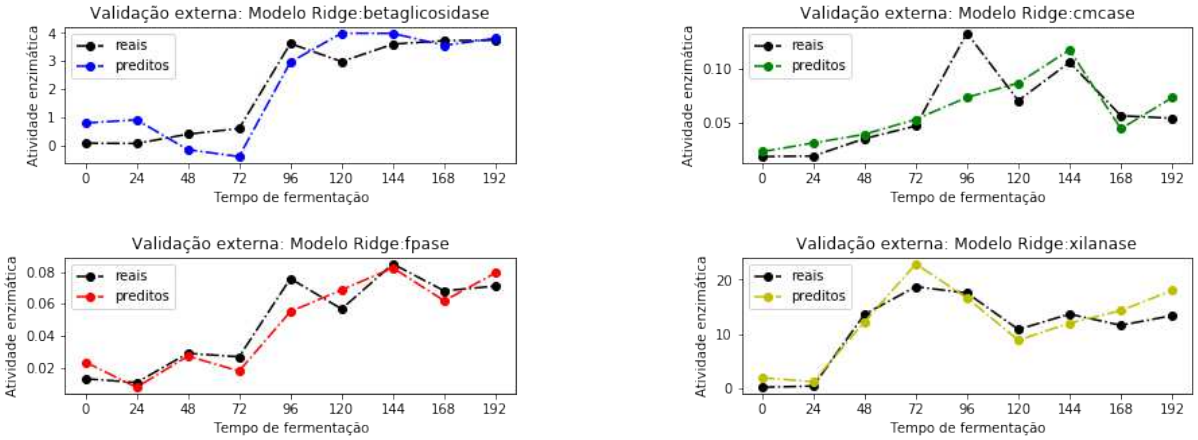

Figura 42 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo *Ridge*

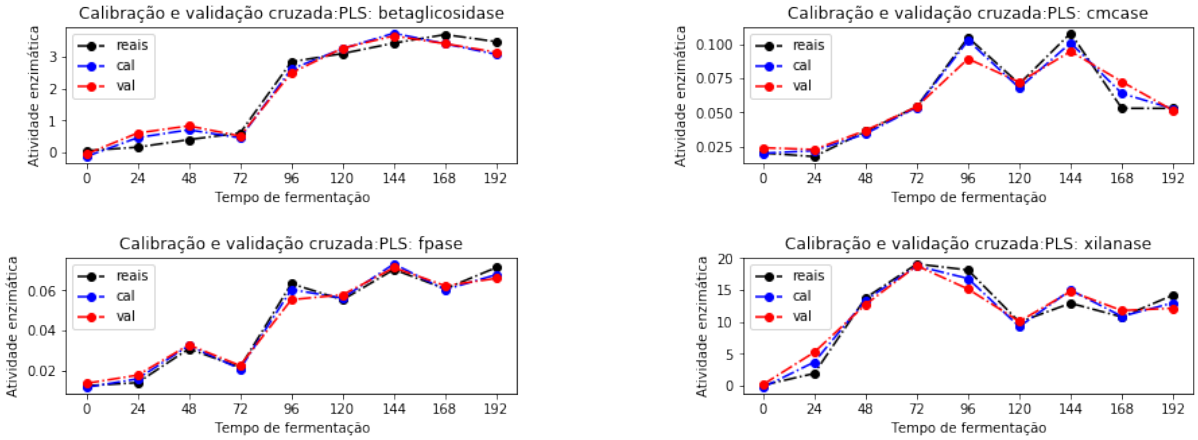

Figura 43 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo PLS

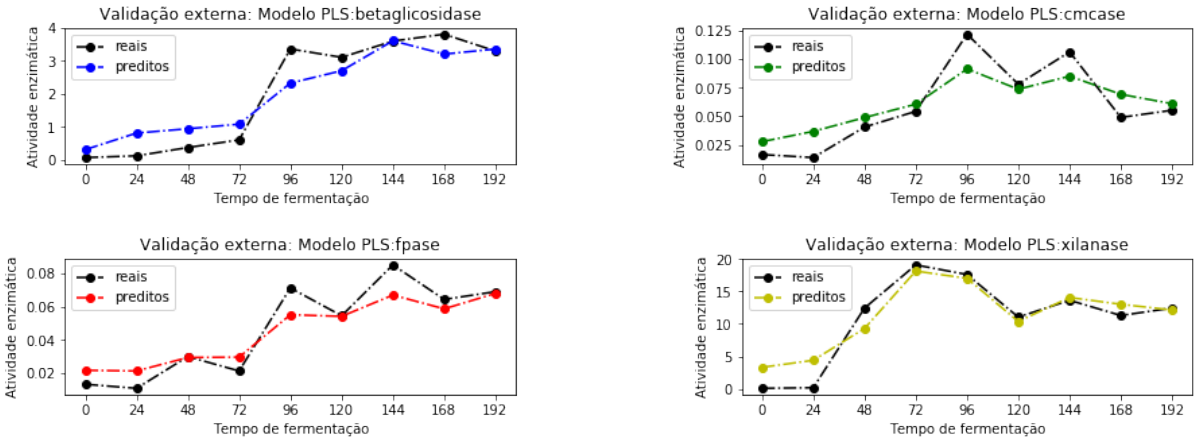

Figura 44 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo PLS

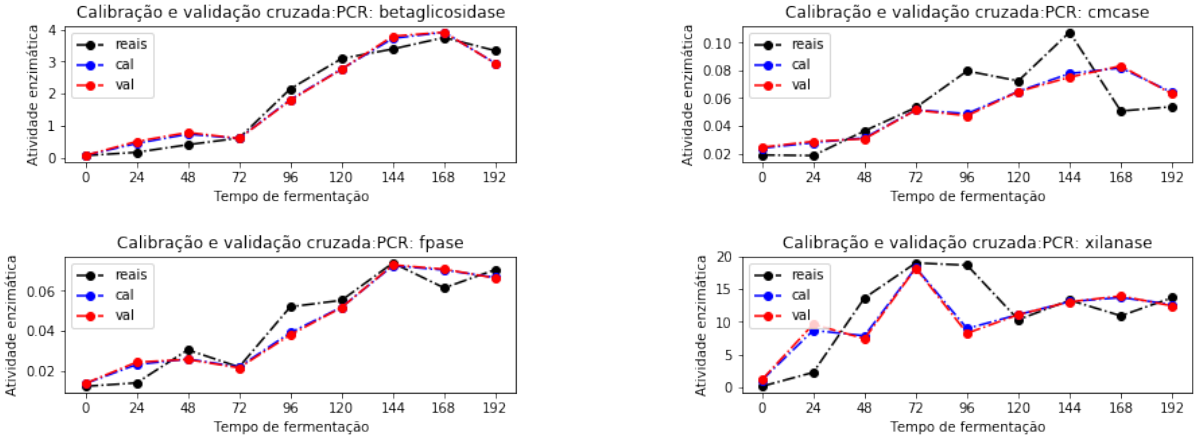

Figura 45 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo PCR

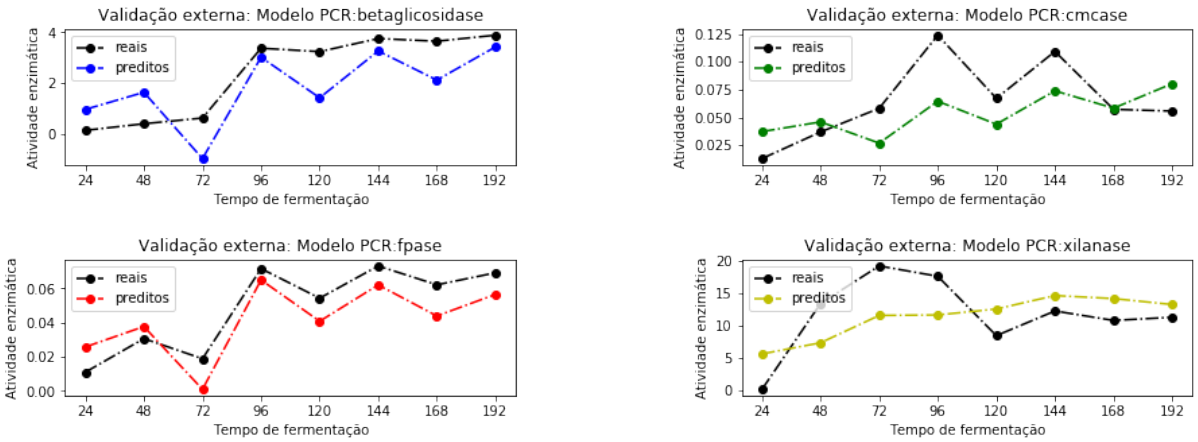

Figura 46 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo PCR

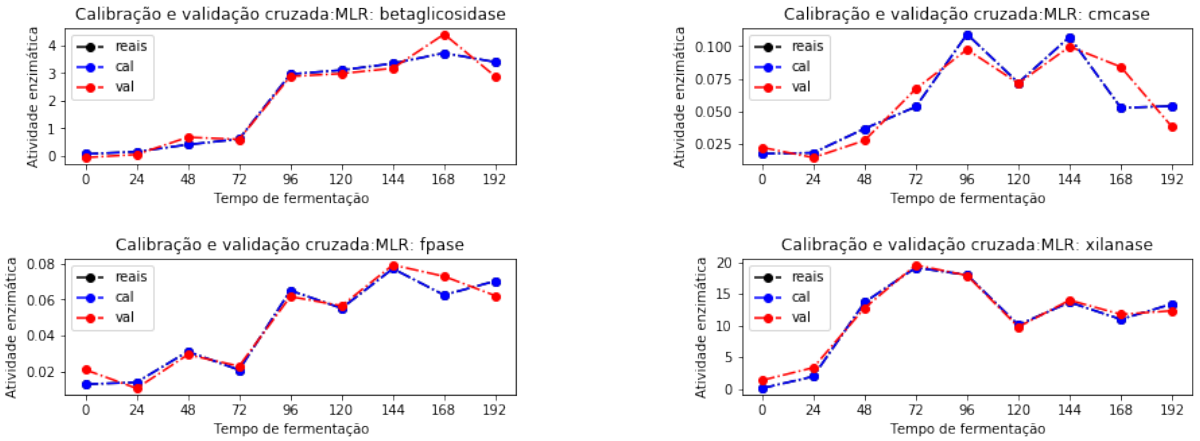

Figura 47 – Curva de treinamento e validação cruzada durante o processo de fermentação para as quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo MLR

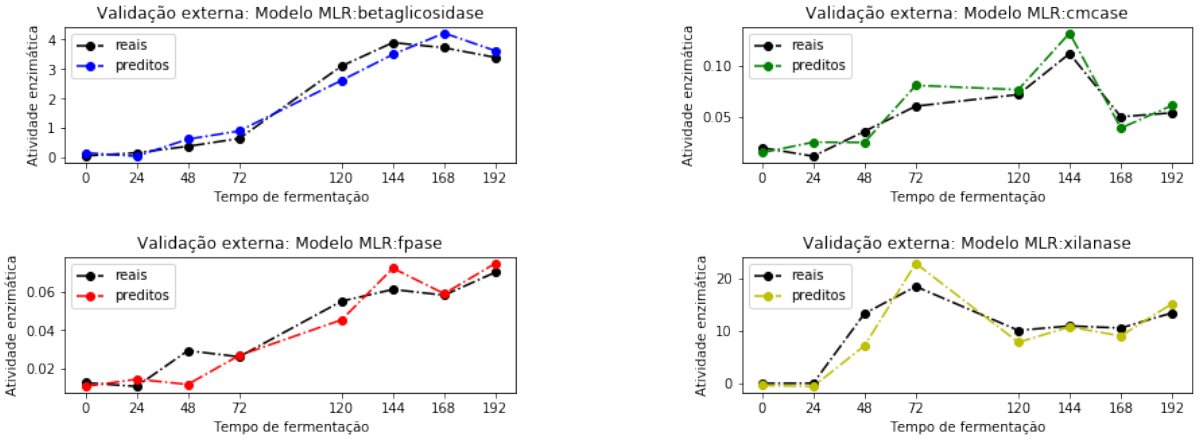

Figura 48 – Curva de predição durante o processo de fermentação paras quatro atividades enzimáticas: β-glicosidase, CMCase, FPase e xilanase versus dados reais para o modelo MLR

Ao observar o comportamento dos modelos apresentados nas Figuras 37 a 48, foi possível constatar que é possível acompanhar o processo fermentativo para o extrato enzimático EETA mesmo para os modelos com RPD inferior a 3, como é o caso do PCR e MLR. Além disso, todos os modelos construídos que resultaram em um bom coeficiente de variação e valores de RPD superiores a 2,5, poderiam ser utilizados para acompanhar processos fermentativos gerais com o objetivo de reduzir os custos experimentais, demonstrando assim, que as ferramentas de análise multivariada associadas à espectroscopia NIR representam uma eficiente metodologia para determinação analítica de atividade enzimática, pois produz uma redução do tempo de análise e do quantitativo de resíduos químicos gerados.

Os resultados obtidos para a grande maioria dos modelos gerados evidenciaram que todos os modelos de calibração, validação e predição, excetuando-se os modelos PCR e MLR resultaram em uma forte correlação entre o espectro eletromagnético no infravermelho próximo e a atividade enzimática determinada em microescala por espectroscopia. As atividades CMCásicas e FPásicas foram consideradas muito baixas se comparadas às duas outras para o extrato enzimático EETA, o que, possivelmente, interferiu na redução de desempenho para determinação destas enzimas.

O presente trabalho gerou como resultados um conjunto de dados constituído de 1391 amostras de espectros NIR associadas a determinações enzimáticas de β-glicosidase, CMCcase, FPase e xilanase, bem como uma coleção de arquivos de programação, desenvolvidos em linguagem python, para a realização de análises estatísticas, seleção de atributos, remoção de *outliers* e modelagem computacional para calibração, validação e predição. Esses arquivos poderão ser reutilizados tanto para entendimento da pesquisa realizada quanto para efetuar determinações enzimáticas de outros complexos enzimáticos e, ainda, na construção de outros modelos computacionais de análises quimiométricas diversas que envolvam a utilização de espectros NIR. No Anexo III são listados os conteúdos gerados bem como o sítio web onde todos esses documentos podem ser acessados. Além disso, um artigo contendo os resultados da modelagem preliminar realizada foi aceito para publicação na Revista Virtual de Química.

6 CONSIDERAÇÕES FINAIS

Os resultados do presente estudo permitiram concluir que a adaptação da metodologia de determinação enzimática de macro para microescala permitiu promover uma determinação enzimática rápida e eficiente com redução significativa do uso de reagentes e preservação das microplacas. Permitiu também constatar que a aplicação da espectroscopia na região do infravermelho próximo para quantificar quimiometricamente a atividade das enzimas aqui investigadas foi capaz de produzir uma correlação matemática aceitável e promissora para o desenvolvimento da metodologia original. Foi possível predizer a atividade enzimática com precisão para amostras que não foram submetidas ao treinamento.

Esta abordagem abre espaço para o desenvolvimento de métodos diretos de determinação da atividade enzimática com base na observação da enzima como entidade molecular, ao invés de substratos ou produtos de reações catalisadas por enzimas. A possível adoção desse tipo de abordagem metodológica implica em menor tempo de análise, redução da quantidade ou dispensação de reagentes, menor geração de resíduos químicos por processos analíticos e provável redução do custo de análise.

Os resultados encontrados permitiram concluir que existe uma correlação linear entre os espectros NIR gerados de complexos enzimáticos comerciais e/ou oriundos de processos experimentais, utilizando torta de caroço de algodão como substrato e fungo *Apergilus niger*, como agente fermentativo e o teor de enzimas contidas nestes extratos.
Concluiu-se também que algumas faixas de comprimento de onda do espectro eletromagnético estão diretamente relacionadas à presença das enzimas celulolíticas e hemicelulolíticas e podem ser utilizadas para a determinação direta das mesmas.

REFERÊNCIAS

ABDI, H. Partial least squares regression and projection on latent structure regression (PLS Regression). **WIREs Computational Statatistics**, 2010, v.2, p.97-106. Doi: 0.1002/wics.51.

AKRAM, F.; HAQ, I.; IMRAN, W.; MUKHTAR, H. Insight perspectives of thermostable endoglucanases for bioethanol production: A review. Renewable Energy, 2018, v. 122, p. 225-238 Doi: 10.1016/j.renene.2018.01.095.

ARAÚJO, C. K. C. de ; OLIVEIRA, C. A. de; ARAÚJO, C. E. P. de ; SOUSA JÚNIOR, F. C. de; NASCIMENTO, R. J. A. do; MACEDO, G. R. de ; SANTOS, E. S. dos. Enhancing enzymatic hydrolysis of coconut husk through Pseudomonas aeruginosa AP 029/GLVIIA rhamnolipid prepa*Ratio*n. **Bioresource Technology**, 2017, v. 237, p. 20-26.

ARISMENDY, A. M.; SEQUEIRA, M. J. FELISSIA, F. E.; AREA, M. C; CHAMORRO, E. R. Evaluación de Cepas Fermentativas en la Hidrólisis y Fermentación Simultáneas (SSF) de Cascarilla de Arroz para la Producción de Bioetanol. In: **Revista Tecnología y Ciencia**, 2018, v. 30, p:357-363.

AZEVEDO, A. N. G. DE; LIMA, B. G. DE A. Biocombustíveis: desenvolvimento e inserção internacional. **Revista Direito Ambiental e sociedade**, 2016, v. 6, n. 1, p. 77-100.

BAMDADA, H.; HAWBOLDTA, K.; MACQUARRIEB, S. A review on common adsorbents for acid gases removal: Focus on biochar. **Renewable and Sustainable Energy Reviews**, 2018, v. 81, n. 2, p. 1705-1720. ISSN 1364-0321. Doi: https://doi.org/10.1016/j.rser.2017.05.261.

BAPTISTA, P; FELIZARDO, P.; MENEZES, J. C.; CORREIA, M. J. N. Multivariate near infrared spectroscopy models for predicting the methyl esters content in biodiesel. **Analytica Chimica Acta**, 2008, v. 607, n. 2, p. 153-159.

BARNES, R. J.; DHANOA, M. S.; LISTER, S. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. **Applied Spectroscopy**, 1989, v. 43, n. 5, p. 772–777.

BAUM, A., MIKKELSEN, J. D., HANSEN, P. W, EGEBO, M. **Descriptive and predictive assessment of enzyme activity and enzyme related processes in biorefinery using IR spectroscopy and chemometrics**. Tese. Center for BioProcess Engineering, Department of Chemical and Biochemical Engineering. Dinamarca, 2013, p.124.

BAUM, A.; AGGER, J.; MEYER, A. S.; EGEBO, M. ;MIKKELSEN, J. D. Rapid near infrared spectroscopy for *Prediction* of enzymatic hydrolysis of corn bran after various pretreatments. **New Biotechnology,** 2012, v. 29, p. 293-301.

BERNFELD, P. Amylases, α and β. **Methods in Enzymology**,1955, v. 1, p. 149-157.

BELLON-MAUREL, V.; FERNANDEZ-AHUMADA,E.; PALAGOS, B.; ROGER, J.M.; MCBRATNEY, A. Prediction of soil attributes by NIR spectroscopy. A critical review of chemometric indicators commonly used for assessing the quality of the prediction. **TrAC Trends in Analytical Chemistry**, 2010, v.29 , n.9, p. 1073-1081.

BRAGA, A. de P.; CARVALHO, A. P. de L. F. de.; Ludrmir, T. B. **Redes neurais artificias: teorias e aplicações**, 2011, 2ª ed. [reimp] Rio de Janeiro: LTC.

CÂMARA, A. B.F. CARVALHO, L. DE; DE MORAIS, C. L.M.; DE LIMA, L.A.S.; DE ARAÚJO, H. O.M.; DE OLIVEIRA, F. M.; DE LIMA, K.M.G. MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends, **Fuel**, 2014, v. 210, p. 497-506.

CAMASSOLA, M.; DILLON, A. J. P. **Cellulase determination: modifications to make the filter paper assay easy, fast, practical and efficient,** 2012, v.1: 125. Doi:10.4172/scientificreports.125

CARVALHO, W.; CAMILHA, L.; FERRAZ, A. Milagres, a. M. F. Uma visão sobre a estrutura, composição e biodegradação da madeira. **Quim. Nova**, 2009, v. 32: 8, p. 2191-2195.

CHAKRABORTY, S.; DAS, B.S.; ALI ; LI, M.M.; SARATHJITH, M.C.; MAJUMDAR, K.; RAY, D.P. Rapid estimation of compost enzymatic activity by spectral analysis method combined with machine learning. **Waste management**. 2014, v. 34.3, p. 623-631.

CLERCQ, D. de; JALOTA, D.; SHANG, K. N.; ZHANG, Z.; KHAN, A.; WEN, Z.; CAIACEDO, L.; YUAN, K. Machine learning powered *Software* for accurate *Prediction* of biogas production: A case study on industrial-scale Chinese production data. **Journal of cleaner production**, 2019, v. 218, p. 390-399. Doi: 10.1016/j.jclepro.2019.01.031.

CUNHA, C. L.; LUNA, A. S.; OLIVEIRA, R. C.G. ; XAVIER, G. M. ; PAREDES, M. L.L.. TORRES, A. R. Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate *CalibRation*. **Fuel,** 2017, v. 204 p. 185–194.

DECKER S. R.; ADNEY W. S.; JENNINGS E.; VINZANT, T. B.; HIMMEL, M. E. Automated filter paper assay for determination of cellulase activity. **Applied Biochemistry and Biotechnology**, 2003, v.107, p.689-703.

FACELI, K.; LORENA, A. C; GAMA, J.; CARVALHO, A. C. P. L. F. DE. **Inteligência Artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FEARN, T. Assessing *CalibRation*s: SEP, RPD, RER and R². **NIR News**, 2002, v. 13, n. 6, p. 12-13. Doi:10.1255/nirn.689.

FERREIRA, M. C.; ANTUNES, A. M.; MELGO, M.S.; VOLPE, P. L.O. Quimiometria I: calibração multivariada, um tutorial. **Quím. Nova**, 1999, v. 22,n. 5, p. 724-731. Doi: 10.1590/ S0100-40421999000500016.

FLORENCIO, C, BADINO , A. C. ; FARINAS , C. S. Desafios relacionados à produção e aplicação das enzimas celulolíticas na hidrólise da biomassa lignocelulósica. **Química Nova**, 2017, v. 40, n. 9, p.1082-1093. Doi: 10.21577/0100-4042.20170104.

FULOP, L.; ECKER, J. An overview of biomass conversion: exploring new opportunities. **PeerJ 8**, 2020, PubMed 32765969. Doi: 10.7717/peerj.9586

GHOSE, T.K. Measurement of cellulase activities. **Pure and Applied Chemistry**, 2009, v. 59, n. 2, p. 257-268. Doi:10.1351/pac198759020257.

GHOSE T.K; BISARIA, V.S. Measurement of hemicellulase activities part I: Xylanases. **Pure and Applied Chemistry,** 1987, v. 59, n.12, p. 1739–1751. Doi:10.1351/pac198759121739.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Alta Books, 2019.

GONÇALVES, C.; RODRIGUEZ-JASSO, R. M.; GOMES, N.; TEIXEIRA, J. A.; BELO, I. Adaptation of dinitrosalicylic acid method to microtiter plates. **Analytical Methods London**, 2010, v. 2, p. 2046-2048. Doi: 10.1039/C0AY00525H.

GRUS , JOEL. **Data science do zero: primeiras regras com** *Python***.** Tradução de Welington Nascimeento. Rio de janeiro: Alta Books, 2016.

GUTIERREZ DEVIA, FLÓREZ , J. E. M.; MARTÍNEZ, J. E. B.; ESPINOSA, D. M. I.; ARANA, A. C.; CAETANO, C. M. Application of NIR spectroscopy for the *Prediction* of soluble solids in guava pulp. **Acta Agron. Palmira**, 2015, v.64,n. 2, p. 103-109, Doi:.org/10.15446/acag.v64n2.40107.

HUNTER, J.; DALE, D.; FIRIGN, E. ; DROETTBOOM,M. **The** *Matplotlib* **User's Guide**, 2020. Disponível em <https://*Matplotlib*.org/*Matplotlib*.pdf>. Acesso em 25 de agosto de 2020.

INFORSATO , F. J; PORTO, A. L. M. Atividade enzimática de celulases pelo método DNS de fungos isolados de sementes em germinação. **Revista Brasileira de Energias Renováveis**, 2016, v. 5, n.4, p. 444-465.

JIN, X.; CHEN, X;, SHI, C.; LI, M.; GUAN, Y.; YU, C. Y.; YANANDA, T.; SACKS, E. PENG, J. Determination of hemicellulose, cellulose and lignin content using visible and near infrared spectroscopy in Miscanthus sinensis. **Bioresource Technology**, 2017**,** v. 241, p. 603-609.

JIN, X.; LI, S.; ZHANG, W.; ZHU, J.; SUN, J. *Prediction* of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. **Applied Sciences**, 2020, v.10, 1520. Doi:10.3390/app10041520.

JONES, E.; OLIPHANT, T.; PETERSON, P. *Scipy*: Open Source Scientific Tools for *Python*. *Scipy* **is a** *Python*-**based ecosystem of open-source** *Software* **for mathematics, science, and engineering**, 2001. Disponível em<https://www.*Scipy*.org/> Acesso em 25 de agosto de 2020.

KING B.C; DONNELLY, M.K; BERGSTROM, G.C.; WALKER, L.P.; GIBSON, D.M. An optimized microplate assay system for quantitative evaluation of plant cell wall-degrading enzyme activity of fungal culture extracts. **Biotechnol Bioeng**, 2009; v.102, n.4, p.1033-1044. Doi:10.1002/bit.22151.

KLIMKIEWICZ, A. ; MORTENSEN, P. P. ; ZACHARIASSEN, C. B.; VAN den BERG, F. W. J. Monitoring an enzyme purification process using on-line and in-line NIR measurements. **Chemometrics and Intelligent Laboratory Systems**, 2014, v. 32, p.30-38.

KÖNIG, J.; GRASSER, R.; PIKOR, H. VOGEL, K. Determination of xylanase, β-glucanase, and cellulase activity. **Analytical and Bioanalytical Chemistri**, 2002, v. 374, p.80–87. Doi: 10.1007/s00216-002-1379-7.

KOONJAH, S. S.; BEEKHARRY, A.; BADALOO, M. G. H.; HENDERSON, C.; DOOKUN SAUMTALLY, A. Evaluation of Near Infrared Spectroscopy for the Direct Analysis of Cane Quality Characters. **Universal Journal of Agricultural Research**, 2019, v.7, p.169-176. Doi:10.13189/ujar.2019.070501.

KUMAR, S.; BARTH, A. Following Enzyme Activity with Infrared Spectroscopy. **Sensors**, 2010 . Doi: 10.3390/s100402626.

LAI, T. E.; PULLAMMANAPPALLIL, P. C.; CLARKE, W. P. Quantification of cellulase activity using cellulose-azure. **Talanta**. 2006, v.69,, n.1, p.68-72.

LAZZAROTTO, M.;NETIPANYJ, R. R.; MAGALHAES, W. L. E.; AGUIAR, A. V. de.Espectroscopia no infravermelho próximo para estimativa da densidade básica de madeiras de Pinus, 2016, v. 7, n. 3, p. 119-126. Doi: 10.12953/2177-6830/rcm.v7n3p119-126.

LIMA, A.; BAKKER, J. Espectroscopia no infravermelho próximo para a monitorização da perfusão tecidual. **Rev Bras Ter Intensiva**, 2011, v.23, n.3, p. 341-351.

LLOYD, J. B.; WHELAN, W. J. An improved method for enzymic determination of glucose in the presence of maltose. **Analytical Biochemistry**, 1969, v. 30, n.3, p. 467-470. Doi:10.1016/0003-2697(69)90143-2.

LUCENA S. A. MORAES, C.S; COSTA, S. G.; SOUZA, W.; AZAMBUJA, P.; GARCIA, E. S.; GENTA, F. A. Miniaturization of hydrolase assays in thermocyclers. **Anal Biochem**, 2013, v. 434, n.1, p. 39-43. Doi: 10.1016/j.ab.2013.10.032.

LYND, L. R. LIANG, X., BIDDY, M. J., ALLEE, A., CAI, H., FOUST, T.,HIMMEL, M. E. LASER, M. S., WANG, M. WYMAN, C. E. Cellulosic ethanol: status and innovation. **Current Opinion in Biotechnology**, 2017, v. 45, p. 202 – 211. Doi:10.1016/j.copbio.2017.03.008.

LEITE, D. de O.; PRADO, R. J. Espectroscopia no infravermelho: uma apresentação para o Ensino Médio. **Rev. Bras. Ensino Fís**., 2012, v. 34, n. 2, p. 1-9. Doi: 10.1590/S1806-11172012000200015.

MANSOUR, A.A.; COSTA, A. da; ARNAUD, T.; LU-CHAU, T.A. FDZ-POLANCO, M.; MOREIRA, M.T.; CACHO RIVERO, J.A. Review of lignocellulolytic enzyme activity analyses and scale-down to microplate-based assays. **Talanta**, 2016, v. 150, p. 629-637. Doi: 10.1016/j.talanta.2015.12.073.

MALLEY, D. F.; MARTIN, P. D.; BEN-DOR, E. **Application in analysis of soils. In: Near-infrared spectroscopy in agriculture**. Roberts, C. A.; Workman J.; Reeves, J. B. (Eds.). Agronomy, v.44. ASA-CSSA-SSSA, Madison, WI, USA. p.729-784. 2004. Doi: 10.2134/agronmonogr44.c26

MCKINNEY, W. W. **Python para análise de dados: tratamento de dados com pandas numpy e ipython**. 2018, Novatec Editora Ltda.

MENEZES, C. R. DE; BARRETO, A. R. Biodegradação de resíduos lignocelulósicos por fungos basidiomicetos: caracterização dos resíduos e estudo do complexo enzimático fúngico. **Revista Eletrônica em Gestão, Educação e Tecnologia Ambiental**, e-ISSN 2236 1170, 2015, v. 19, n. 2, p. 1365-1391. Doi: 105902/2236117016853

MILLER, G. L. Use of dinitrosalicylic acid reagent for determination of reducing sugar. **Anal. Chem**, 1959, v.31, p.426-428.

NASCIMENTO, R. J. A. **Monitoramento em tempo real da hidrólise enzimática do bagaço da casca de coco verde por espectroscopia no infravermelho próximo (nirs).** Tese (Doutorado), 2016 . Natal, 136 p.

NASCIMENTO, R. J. A. DO; MACEDO, G. R. DE; SANTOS, E. S. DOS; OLIVEIRA, J. A. de. Real time and in situ near-infrared spectroscopy (nirs) for quantitative monitoring of biomass, glucose, ethanol and glycerine concent*Ratio*ns in an alcoholic fermentation, **Braz. J. Chem. Eng.**, 2017, v. 34, n. 2, p. 459-468. Doi: 10.1590/0104-6632.20170342s20150347.

NEGRULESCU, A., PATRULEA, V, MINCEA, M M., IONASCU, C, VLAD-OROS, B. A.; OSTAFE, V. Adapting the reducing sugars method with dinitrosalicylic acid to microtiter plates and microwave heating. **Journal of the Brazilian Chemical Society**, 2012, v. 23, n. 12, p. 2176-2182. Doi:10.1590/S0103-50532013005000003

OGEDA, T. L.; PETRI, D. F. S.. Hidrólise Enzimática de Biomassa. **Quím. Nova**, São Paulo , 2010, v. 33, n. 7, p. 1549-1558. Doi:10.1590/S0100-40422010000700023.

OLIPHANT, T. E. **A guide to *Numpy*** (Vol. 1). Trelgol Publishing USA, 2006.

OLIVEIRA, I.K. P. H. **Aplicação de quimiometria e espectroscopia no infravermelho no controle de qualidade do biodiesel e mistura biodiesel/diesel**. Dissertação (Mestrado)- Campinas - SP, 2008.

PEREIRA, N. R. L.; ANJOS, F. E.; MAGNAGO, R. F. Resíduos lignocelulósicos da bananicultura: uma revisão sobre os processos químicos de extração da celulose. **Revista Virtual de Química**, 2019, v. 11 n. 4, p. 1165-1179. Doi: 10.21577/1984-6835.20190080.

PERKEL, J. M. "Why Jupyter is data scientists' computational notebook of choice." **Nature**, 2018, v. 563, n. 7732, p. 145+.

RAMBO, M. K. D.; FERREIRA, M. M. C. Análise de Resíduos Lignocelulósicos por Espectroscopia NIR Associada a Pré-tratamentos Multivariados Dentro do Contexto de Química Verde**. Rev. Virtual Quim**., 2018, v.10, n. 2, p. 421-431.

REZENDE, D. B. DE, PASA, V. M. D. Tendências e oportunidades nas pesquisas em biocombustíveis.**The Journal of Engineering and Exact Sciences**, 3, abr. 2017. Doi: 10.18540/jcecvl3iss3pp561-572.

RINNAN, A.; BERG, F.; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. ***TrAC Trends in Analytical Chemistry,*** 2009, v.28, p.1201-1222. Doi: 10.1016/j.trac.2009.07.007

RODRIGUES, A.C.; HAVEN, M. Ø.; lINDEDAM, J.; FELBY, C.; GAMA, M. Celluclast and Cellic® CTec2: Saccharification/fermentation of wheat: straw, solid–liquid partition and potential of enzyme recycling by alcaline washing. **Enzyme and Microbial Technology**, 2015. Doi:10.1016/j.enzmictec.2015.06.019.

SANTIAGO, B. L. S; RODRIGUES, F. A. Processamento de biomassa lignocelulósica para produação de etanol: uma revisão. **The Journal of Engineering and Exact Sciences – JCEC**. 2017, v. 3, n. 7, p. 1011-1022. Doi: 10.18540/jcecvl3iss7pp1011-1022.

SANTOS, J. R. A.; GOUVEIA, E. R. **Produção de bioetanol de bagaço de cana-de-açúcar.** Revista Brasileira de Produtos Agroindustriais, Campina Grande. 2009, v. 11:1, p. 27-33

SANTOS, R. S.; ALVS, F.K.P.; VANZELA, A.P.F.C.; PANTOJA, L.A.; SANTOS, A.S. Holocellulolytic hydrolases production by filamentous fungi using oil cakes as substrate. **Brazilian Journal of Microbiology**, 2015.

SANTOS, S. A. ; JUNIOR, R. D. MÜLLER, G.; GOULART, D.; STAMBUK, M. BORIS; ALVES, S. J. Dosagem de açúcares redutores com o reativo DNS em microplaca. **Brazilian Journal of Food Technology.** Campinas, 2017a, v. 20, e2015113.

SANTOS, D. A. DOS, LIMA, K. P. DE, CONSOLIN, M. F. B., CONSOLIN FILHO, N., MARÇO, P. H., & VALDERRAMA, P. Multi-product multivariate *CalibRation*: determination of quality parameters in soybean industrialized juices. **Acta Scientiarum. Technology**, 2019, v.41, n.1, e37382. Doi: 10.4025/actascitechnol.v41i2.37382.

SARSAIYA, S.; AWASTHI, S. K.; AWASTHI, M. K.; AWASTHI, A. K., MISHRA, S, CHEN, J. (2018). The dynamic of cellulase activity of fungi inhabiting organic municipal solid waste. **Bioresource Technology**, 2018: v. 251, p. 411-415.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least *Square*s procedures. **Analytical Chemistry**. 1964, v.36, p.1627–1639. Doi: 10.1021/ac60214a047.

SEABOLD, S.; PERKTOLD, J. Statsmodels: Econometric and statistical modeling with *Python*. **Proceedings of the 9th** *Python* **in Science Conference**. 2010. Disponível em <https://conference.*Scipy*.org/proceedings/*Scipy*2010/pdfs/seabold.pdf> Acesso em 25 de agosto de 2020.

SELVAM, K; SENBAGAM, D.; SELVANKUMAR, T.; SUDHAKAR,C.; KAMALA-KANNAN, S. SENTHILKUMAR, B., GOVARTHANAN, M. Cellulase enzyme: homology modeling, binding site dentification and molecular docking, **Journal of Molecular Structure**, 2017, v. 1150, p. 61-67.

SHAO, Y.; LIN, A. H. Improvement in the quantification of reducing sugars by miniaturizing the Somogyi-Nelson assay using a microtiter plate. **Food Chemistry**, 2018, v. 240, p. 898-903. Doi: 10.1016/j.foodchem.2017.07.083.

SILVA, D. A. ; ALMEIDA, V. C.; VIANA, L.C.; KLOCK, U.; DE MUÑIZ, G. I . B. Avaliação das propriedades energéticas de resíduos de madeiras tropicais com uso da espectroscopia nir. **Floresta e Ambiente**, 2014, v. 21, n.4, p. 561-568. Doi: *10.1590/2179-8087.043414*.

SILVA, C. S. **Espectroscopia no infravermelho para aplicações forenses: documentoscopia e identificação de sêmen em tecidos**. Tese (Doutorado), 2017.

SINGH, G.; VERMA, A. K; KUMAR. Catalytic properties, functional attributes and industrialapplications ofb-glucosidases. **Gopal Biotech**, 2016; v. 6:3. Doi 10.1007/s13205-015-0328-z.

SKVARIL, J.; KYPRIANIDIS, K; AVELIN, A.; ODLARE, M. DAHLQUIST, E. Fast determination of fuel properties in solid biofuel mixtures by near infrared spectroscopy. **Energy Procedia**, 2017, v. 105, p. 1309-1317.

SOUZA, J. S.; FERRÃO, M. F. Aplicações da espectroscopia no infravermelho no controle de qualidade de medicamentos contendo diclofenaco de potássio. Parte I: Dosagem por regressão multivariada. **Revista Brasileira de Ciências Farmacêuticas**, 2006, v.42, p.437-445. Doi: 10.1590/S1516-93322006000300013.

SZYMAŃNSKI, P. ; KAJDANOWICZ, T. Scikit-multilearn: A scikit-based *Python* environment for performing multi-label classification. **Journal of Machine Learning Research**, 2019, v. 20, p. 1-22.

TAYLOR-MORGAN, S. **Data visualization made simple in** *Python* **with** *Seaborn*. 27 de maio de 2020. Disponível em <https://opensource.com/article/20/5/*Seaborn-Python*>. Acesso em 25 de agosto de 2020.

TIBOLA, C, S; MEDEIROS, E. P. de; SIMEONE, M. L. F.; OLIVEIRA, M. A. de. **Espectroscopia no Infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos** / Casiane Salete Tibola... [*et al.*], editores técnicos. – Brasília, DF : Embrapa, 2018. 200 p. ISBN: 978-85-7035-839-4 1.

VALINHAS, R. V.; PANTOJA, L. A.; MAIA, 2, A. C. F.; MIGUEL, M. G. C. P.; VANZELA, A. P. F.C; NELSON, D. L.; SANTOS, A. S. Xylose fermentation to ethanol by new Galactomyces geotrichum and Candida akabanensis strains. **PeerJ 6**:e4673, 2018. Doi:10.7717/peerj.4673.

VASCONCELOS, N. M.; PINTO, G. A. S.; ARAGAO, F. A. S. de. Determinação de açúcares redutores pelo ácido 3,5-dinitrosalicílico: histórico do desenvolvimento do método e estabelecimento de um protocolo para o laboratório de bioprocessos. **Boletim de pesquisa e desenvolvimento. Embrapa Agroindústria Tropical,** 2013, v. 88, 29p.

VISCARRA ROSSEL, R. A.; WALVOORT, D. J. J.; MCBRATNEY, A. B.; JANIK, L. J.; SKJEMSTAD, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. **Geoderma**, 2006, v.131, p.59-75. Doi: 10.1016/j.geoderma.2005.03.007

WOOD, I. P.; ELLISTON, A.; RYDEN, P.; BANCROFT, I.; ROBERTS, I. N.; WALDRON, K. W. Rapid quantification of reducing sugars in biomass hydrolysates: improving the speed and precision of the dinitrosalicylic acid assay. **Biomass and Bioenergy**, Oxford, 2012, v. 44, p. 117-121.

XU, L.; ZHOU, Y-P.; TANG, L-J.; WU, H-L.; JIANG, J-H.; SHEN, G-L.; YU, R-Q. Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. **Analytica Chimica Acta**, 2008, 616, 138-143. Doi: 10.1016/j.aca.2008.04.031

YU, X.; LIU, Y.; CUI, Y.; CHENG, Q.; ZHANG, Z.; LU, J. H.; MENG, Q.; TENG, L.; REN, X. Measurement of filter paper activities of cellulase with microplate-based assay. **Saudi Journal of Biological Sciences**, 2016, v.23, p.S93-S98.

ZHAO, N.;WU, Z.; ZHANG, Q.; XHI, X.; QIAO, Y. Optimization of parameter selection for partial least *Square*s model development. **Scientific Reports**, 2015. Doi:10.1038/srep11647.

ZAREEF, M., CHEN, Q., HASSAN, M.M.;ARSLAN, M.; HASHIN, M. M.; KUTSANEDZIE, W. A.F.; AGYEKUM,A.A. An overview on the applications of typical non-linear algorithms coupled with nir spectroscopy in food analysis. **Food Eng Rev**, 2020, v. 12, p. 173–190. Doi:10.1007/s12393-020-09210-7.

ANEXO I - TABELAS DE RESULTADOS DA MODELAGEM COMPUTACIONAL PARA O CONJUNTO DE DADOS COMPLETO

O presente anexo contém tabelas com resultados de métricas de avaliação para os modelos de calibração, validação e predição construídos a partir do conjunto de dados incluindo os complexos enzimáticos EETA, Celluclast® e Celic CTec2® com valores de atividade enzimática para as quatro atividades com faixas de absorbância selecionadas de duas maneiras: 1) com a remoção de duas faixas de absorbância consideradas ruidosas; 2) com a aplicação de seleção de atributos, que permitiu selecionar faixas com maior correlação entre os valores de atividade enzimática:

1) Resultados com a remoção de faixas com ruídos

Tabela 13 – Resultados de desempenho da calibração, validação e predição do modelo GBR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo:		**Atividade enzimática**			
GBR *	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	13,8276	0,955	1,0702	25,2427
	MSE	17561,953	208,3741	119,1047	46680,777
	R^2	**0,9056**	**0,9048**	**0,7642**	**0,7657**
	RER	22,128	19,977	15,0268	14,6383
	RMSE	132,5215	14,4352	10,9135	216,0573
	RPD	3,2547	3,2403	2,0594	2,066
	RPIQ	1,116	2,0472	1,9402	2,1827
	SEP	131,8935	14,414	10,8688	214,7329
Validação	**BIAS**	14,536	1,0854	1,2331	31,9693
	MSE	56364,5942	519,514	272,9737	106652,6828
	R^2	**0,697**	**0,7626**	**0,4596**	**0,4647**
	RER	12,3073	12,6385	9,9057	9,6645
	RMSE	237,4123	22,7928	16,5219	326,5772
	RPD	1,8168	2,0522	1,3603	1,3668
	RPIQ	0,6229	1,2965	1,2816	1,444
	SEP	237,1383	22,7835	16,4878	325,2438
Predição	**BIAS**	-48,5522	-4,2994	-1,8663	-22,6099
	MSE	21920,1157	477,6676	157,3469	55108,5848
	R^2	**0,6947**	**0,6817**	**0,4971**	**0,5742**
	RER	13,2398	9,3627	7,2658	8,1876
	RMSE	148,0544	21,8556	12,5438	234,7522
	RPD	1,8097	1,7725	1,4101	1,5325
	RPIQ	0,9096	1,0682	1,8015	1,9587
	SEP	140,1708	21,4751	12,4311	234,1682

***Parâmetros utilizados:** Pré-processamento →0, loss:huber, semente-→ 0

Tabela 14 – Resultados de desempenho da calibração, validação e predição do modelo PLS construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo:		**Atividade enzimática**			
PLS	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0	0	0	0
	MSE	55664,3701	505,775	172,3352	66169,2312
	R2	**0,595**	**0,73**	**0,5497**	**0,5474**
	RER	12,3611	12,7943	10,3277	10,5653
	RMSE	235,933	22,4894	13,1277	257,2338
	RPD	1,5713	1,9244	1,4902	1,4865
	RPIQ	0,6586	1,1788	1,6538	1,8746
	SEP	236,1064	22,506	13,1373	257,4229
Validação	**BIAS**	2,4464	0,1085	0,0696	0,1133
	MSE	98575,3851	930,1086	314,0211	120893,6063
	R2	**0,2827**	**0,5034**	**0,1795**	**0,1732**
	RER	9,2891	9,4348	7,6509	7,8164
	RMSE	313,9672	30,4977	17,7206	347,6976
	RPD	1,1807	1,4191	1,104	1,0997
	RPIQ	0,4949	0,8692	1,2252	1,3869
	SEP	314,1884	30,5199	17,7335	347,9531
Predição	**BIAS**	32,2373	3,6366	2,228	17,8807
	MSE	94298,5003	904,6878	398,4538	130443,19
	R2	**0,4133**	**0,6135**	**0,3147**	**0,3131**
	RER	7,8521	9,5342	8,2151	8,1905
	RMSE	307,0806	30,078	19,9613	361,1692
	RPD	1,3056	1,6085	1,208	1,2066
	RPIQ	0,5274	1,0072	1,0921	1,1726
	SEP	306,0587	29,9234	19,8804	361,5235

Tabela 15 – Resultados de desempenho da calibração, validação e predição do modelo Kernel-Ridge construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo:		**Atividade enzimática**			
Kernel R	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,1354	0,1046	0,0334	0,6222
	MSE	2271,1922	53,2744	24,7991	9698,3445
	R2	**0,9831**	**0,9723**	**0,9428**	**0,9371**
	RER	61,1957	39,4259	32,7733	30,0462
	RMSE	**47,657**	**7,2989**	**4,9799**	**98,4802**
	RPD	7,6964	6,0052	4,1798	3,9861
	RPIQ	3,1014	3,8116	4,3018	4,8725
	SEP	47,6919	7,3035	4,9834	98,5506
Validação	**BIAS**	3,4934	0,2217	0,1946	3,1558
	MSE	45464,6121	469,6669	210,7006	73763,5707
	R2	**0,6621**	**0,7555**	**0,5137**	**0,5213**
	RER	13,6794	13,2777	11,2444	10,8953
	RMSE	**213,2243**	**21,6718**	**14,5155**	**271,5945**
	RPD	1,7202	2,0225	1,434	1,4453
	RPIQ	0,6932	1,2837	1,4758	1,7668
	SEP	213,3524	21,6866	14,5249	271,7758
Predição	**BIAS**	3,2337	-3,539	-1,7446	-33,0267
	MSE	91814,1738	715,5575	317,9767	119021,4068
	R2	**0,4602**	**0,6737**	**0,2605**	**0,2827**
	RER	9,6111	10,836	7,6283	7,9023
	RMSE	**303,0085**	**26,7499**	**17,8319**	**344,9948**
	RPD	1,3611	1,7505	1,1629	1,1807
	RPIQ	0,5141	1,028	1,2375	1,1714
	SEP	303,6609	26,5734	17,7856	344,1692

Tabela 16 – Resultados de desempenho da calibração, validação e predição do modelo Ridge construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo:		**Atividade enzimática**			
Ridge*	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	48876,747	477,216	176,236	63273,412
	R2	**0,647**	**0,752**	**0,568**	**0,577**
	RER	13,191	13,172	12,294	11,763
	RMSE	**221,081**	**21,845**	**13,275**	**251,542**
	RPD	**1,682**	**2,009**	**1,521**	**1,538**
	RPIQ	0,630	1,287	1,623	2,010
	SEP	221,243	21,861	13,285	251,727
Validação	**BIAS**	2,561	0,077	-0,066	-0,704
	MSE	85666,466	831,085	306,782	109106,275
	R2	**0,381**	**0,568**	**0,248**	**0,271**
	RER	9,965	9,981	9,318	8,958
	RMSE	**292,688**	**28,829**	**17,515**	**330,312**
	RPD	**1,271**	**1,522**	**1,153**	**1,171**
	RPIQ	0,476	0,976	1,230	1,531
	SEP	292,892	28,850	17,528	330,554
Predição	**BIAS**	1,458	-2,292	-0,689	-30,601
	MSE	**89646,293**	**824,408**	**325,965**	**123434,485**
	R2	0,436	0,622	0,357	0,310
	RER	**9,726**	**10,039**	**9,033**	**8,442**
	RMSE	**299,410**	**28,713**	**18,055**	**351,332**
	RPD	1,331	1,626	1,247	1,204
	RPIQ	0,672	0,866	1,078	1,080
	SEP	0,375	0,015	0,006	2,931

***Parâmetros:** Pré-processamento→1, , semente-→ 16

Tabela 17 – Resultados de desempenho da calibração, validação e predição do modelo PCR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo: PCR	Medida	Atividade enzimática β-glicosidase	Cmcase	Fpase	Xilanase
Calibração	BIAS	0,000	0,000	0,000	0,000
	MSE	114797,701	1624,657	409,796	145212,657
	R2	**0,224**	**0,244**	**0,125**	**0,123**
	RER	8,608	7,139	8,062	7,765
	RMSE	**338,818**	**40,307**	**20,243**	**381,068**
	RPD	**1,135**	**1,150**	**1,069**	**1,068**
	RPIQ	0,493	0,709	1,098	1,305
	SEP	339,067	40,337	20,258	381,348
Validação	BIAS	-0,044	-0,079	-0,013	-0,358
	MSE	117917,572	1670,009	419,368	148754,083
	R2	**0,203**	**0,223**	**0,105**	**0,101**
	RER	8,493	7,041	7,970	7,672
	RMSE	**343,391**	**40,866**	**20,479**	**385,687**
	RPD	**1,120**	**1,134**	**1,057**	**1,055**
	RPIQ	0,486	0,700	1,085	1,289
	SEP	343,644	40,896	20,494	385,969
Predição	BIAS	0,220	0,017	0,004	1,427
	MSE	1,360	0,002	0,000	30,409
	R2	**0,538**	**0,253**	**0,684**	**0,287**
	RER	3,875	4,002	5,195	3,673
	RMSE	**1,166**	**0,043**	**0,016**	**5,514**
	RPD	**1,472**	**1,157**	**1,780**	**1,184**
	RPIQ	2,859	1,030	2,671	1,218
	SEP	1,174	0,040	0,016	5,458

Parâmetros: Pré-processamento→1 , NPC:4, semente-→ 28

Tabela 18 – Resultados de desempenho da calibração, validação e predição do modelo MLR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast® e Celic CTec2®, 923 amostras e 577 faixas de absorbância

Modelo:		**Atividade enzimática**			
MLR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	-3,529	0,9118	0,0604	-1,0015
	MSE	3753,6506	**45,5003**	17,2557	5974,2456
	R2	0,9718	**0,9762**	0,9597	0,9612
	RER	47,6805	**43,052**	39,2924	38,2847
	RMSE	61,267	**6,7454**	4,154	77,2932
	RPD	5,9548	**6,4804**	4,9785	5,0764
	RPIQ	2,5083	**4,049**	5,0223	5,814
	SEP	61,2103	**6,6884**	4,1566	77,3436
Validação	**BIAS**	-13,3215	-3,1624	-2,3914	-56,4905
	MSE	547961,6671	6511,8343	2389,3744	906125,6438
	R2	-3,1168	-2,4078	-4,5868	-4,8857
	RER	3,9404	3,5684	3,3428	3,1139
	RMSE	740,2443	80,6959	48,8812	951,9063
	RPD	0,4929	0,5417	0,4231	0,4122
	RPIQ	0,2076	0,3385	0,4268	0,4721
	SEP	740,6685	80,6932	48,8586	950,9271
Predição	**BIAS**	-3,44E+12	8,27E+11	8,86E+11	5,29E+12
	MSE	1,19E+25	6,84E+23	7,86E+23	2,80E+25
	R2	-6,80E+19	-3,07E+20	-1,76E+21	-1,68E+20
	RER	0,00E+00	0,00E+00	0,00E+00	0,00E+00
	RMSE	3,44E+12	8,27E+11	8,87E+11	5,29E+12
	RPD	0,00E+00	0,00E+00	0,00E+00	0,00E+00
	RPIQ	0,00E+00	0,00E+00	0,00E+00	0,00E+00
	SEP	7,09E+10	1,70E+10	1,83E+10	1,09E+11

Parâmetros: Pré-processamento→10 , , semente-→ 60

2) Resultados de desempenho dos modelos multivariados após a seleção de faixas de absorbância por algoritmos de seleção de atributos

Tabela 19 – Resultados de desempenho da calibração, validação e predição do modelo GBR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
GBR*	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0	0	0	0
	MSE	3595,9098	122,2953	70,6541	21802,7577
	R^2	**0,975**	**0,9388**	**0,8386**	**0,8606**
	RER	48,6342	26,019	19,416	20,0389
	RMSE	59,9659	11,0587	8,4056	147,6576
	RPD	6,328	4,0417	2,4892	2,6782
	RPIQ	2,572	2,6606	2,5337	2,7972
	SEP	60,01	11,0669	8,4118	147,7661
Validação	**BIAS**	0,8042	-0,1032	-0,1411	2,4824
	MSE	61670,0077	446,8982	232,2731	81705,6254
	R^2	**0,5717**	**0,7763**	**0,4694**	**0,4775**
	RER	11,7439	13,6112	10,709	10,3519
	RMSE	248,3345	21,14	15,2405	285,842
	RPD	1,528	2,1143	1,3729	1,3835
	RPIQ	0,6211	1,3918	1,3974	1,445
	SEP	248,5157	21,1553	15,2511	286,0413
Predição	**BIAS**	-13,5634	-0,4533	0,8662	9,9973
	MSE	33733,9164	494,3923	177,1488	70236,46
	R^2	**0,7614**	**0,7478**	**0,5755**	**0,5593**
	RER	15,8987	12,9244	10,193	10,247
	RMSE	183,668	22,2349	13,3097	265,0216
	RPD	2,0471	1,9914	1,5347	1,5063
	RPIQ	0,7677	1,034	1,6706	1,9883
	SEP	183,5712	22,2794	13,3109	265,4183

***Parâmetros utilizados:** Pré-processamento→1, loss:ls, semente-→0

Tabela 20 – Resultados de desempenho da calibração, validação e predição do modelo Kernel-Ridge construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
Kernel Ridge*	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,095	0,1061	0,0334	0,6352
	MSE	6914,1962	113,3153	56,8473	21637,0737
	R2	**0,9486**	**0,941**	**0,8688**	**0,8596**
	RER	35,0732	27,0317	21,646	20,1157
	RMSE	**83,1516**	**10,645**	**7,5397**	**147,0955**
	RPD	4,411	4,1176	2,7607	2,6687
	RPIQ	1,7775	2,6135	2,8413	3,2621
	SEP	83,2127	10,6523	7,5452	147,2022
Validação	**BIAS**	4,9365	0,3258	0,2595	3,812
	MSE	50460,5795	498,324	224,7601	82158,9205
	R2	**0,6249**	**0,7406**	**0,4812**	**0,4668**
	RER	12,986	12,891	10,8876	10,3238
	RMSE	**224,6343**	**22,3232**	**14,992**	**286,6338**
	RPD	1,6328	1,9635	1,3884	1,3695
	RPIQ	0,658	1,2463	1,4289	1,6741
	SEP	224,7451	22,3372	15,0008	286,8191
Predição	**BIAS**	5,6756	-2,7781	-1,5555	-31,0355
	MSE	77170,7026	626,9696	289,0106	113984,1814
	R2	**0,5463**	**0,7141**	**0,3279**	**0,313**
	RER	10,485	11,5458	7,9966	8,0721
	RMSE	**277,7962**	**25,0394**	**17,0003**	**337,6154**
	RPD	1,4846	1,8701	1,2198	1,2065
	RPIQ	0,5608	1,0983	1,2981	1,197
	SEP	278,352	24,9398	16,9664	336,9289

***Parâmetros utilizados:** Pré-processamento→8, k=poly, semente-→80

Tabela 21 – Resultados de desempenho da calibração, validação e predição do modelo Ridge construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
Ridge*	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	60690,232	508,492	200,269	76146,549
	R2	**0,587**	**0,743**	**0,544**	**0,533**
	RER	11,838	12,760	11,532	10,723
	RMSE	**246,354**	**22,550**	**14,152**	**275,947**
	RPD	**1,556**	**1,974**	**1,481**	**1,464**
	RPIQ	0,632	1,215	1,544	1,787
	SEP	246,535	22,566	14,162	276,150
Validação	**BIAS**	0,835	-0,009	-0,050	-0,540
	MSE	83193,168	718,213	291,618	106728,064
	R2	**0,422**	**0,641**	**0,334**	**0,318**
	RER	10,111	10,737	9,557	9,057
	RMSE	**288,432**	**26,800**	**17,077**	**326,693**
	RPD	**1,316**	**1,668**	**1,225**	**1,211**
	RPIQ	0,535	1,098	1,247	1,264
	SEP	288,643	26,819	17,089	326,932
Predição	**BIAS**	-22,514	-2,924	-0,027	7,096
	MSE	**84783,180**	**867,788**	**266,746**	**101906,283**
	R2	0,400	0,557	0,361	0,361
	RER	**10,031**	**9,802**	**8,289**	**8,503**
	RMSE	**291,176**	**29,458**	**16,332**	**319,228**
	RPD	1,291	1,503	1,251	1,251
	RPIQ	0,484	0,781	1,361	1,651

Parâmetros: Pré-processamento→1, , semente-→ 16

Tabela 22 – Resultados de desempenho da calibração, validação e predição do modelo PLS construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
PLS*	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0	0	0	0
	MSE	64334,1745	578,1126	221,5679	81838,9412
	R2	**0,5943**	**0,7292**	**0,5327**	**0,518**
	RER	11,4981	11,9671	10,9641	10,3431
	RMSE	**253,6418**	**24,044**	**14,8852**	**286,0751**
	RPD	1,57	1,9218	1,4628	1,4403
	RPIQ	0,5906	1,1565	1,4454	1,7353
	SEP	253,8283	24,0616	14,8961	286,2853
Validação	**BIAS**	0,1814	-0,0436	-0,0259	-0,6673
	MSE	113144,0054	1457,2356	381,7219	133129,1067
	R2	**0,2142**	**0,2705**	**0,128**	**0,1487**
	RER	8,6702	7,5376	8,3532	8,1095
	RMSE	**336,3689**	**38,1738**	**19,5377**	**364,8686**
	RPD	1,1281	1,1708	1,0709	1,0838
	RPIQ	0,4585	0,7708	1,0901	1,132
	SEP	336,616	38,2018	19,552	365,1362
Predição	**BIAS**	-14,7596	-1,7363	0,3054	9,8948
	MSE	111849,1605	1388,2157	346,4233	133170,2767
	R2	**0,2088**	**0,2919**	**0,1698**	**0,1644**
	RER	8,7159	7,7197	7,2745	7,4392
	RMSE	**334,4386**	**37,2588**	**18,6125**	**364,925**
	RPD	1,1242	1,1884	1,0975	1,0939
	RPIQ	0,4216	0,6171	1,1946	1,444
	SEP	334,8511	37,3005	18,6511	365,597

***Parâmetros:** Pré-processamento→2, VL-→ 15, semente-→ 24

Tabela 23 – Resultados de desempenho da calibração, validação e predição do modelo PCR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
PCR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	117925,834	1732,245	417,096	147634,389
	R2	**0,256**	**0,189**	**0,120**	**0,130**
	RER	8,493	6,913	7,991	7,701
	RMSE	**343,403**	**41,620**	**20,423**	**384,232**
	RPD	**1,160**	**1,110**	**1,066**	**1,072**
	RPIQ	0,436	0,668	1,054	1,292
	SEP	343,656	41,651	20,438	384,515
Validação	**BIAS**	-0,026	-0,015	-0,036	-0,693
	MSE	121582,060	1788,863	432,268	153089,835
	R2	**0,233**	**0,162**	**0,088**	**0,098**
	RER	8,364	6,803	7,850	7,562
	RMSE	**348,686**	**42,295**	**20,791**	**391,267**
	RPD	**1,142**	**1,093**	**1,047**	**1,053**
	RPIQ	0,430	0,658	1,035	1,269
	SEP	348,943	42,326	20,806	391,554
Predição	**BIAS**	-33,884	-2,315	-1,535	-11,412
	MSE	92442,850	1233,090	299,441	149348,563
	R2	**0,048**	**0,205**	**0,023**	**-0,254**
	RER	9,638	8,200	7,854	7,025
	RMSE	**304,044**	**35,115**	**17,304**	**386,456**
	RPD	**1,025**	**1,121**	**1,012**	**0,893**
	RPIQ	0,526	0,725	1,270	1,112
	SEP	302,818	35,117	17,274	387,142

Tabela 24 – Resultados de desempenho da calibração, validação e predição do modelo MLR construído para o conjunto de dados com os três complexos enzimáticos: EETA, Celluclast e Celic CTec2 e faixas de absorbância após seleção de atributos

Modelo:		**Atividade enzimática**			
MLR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0	0	0	0
	MSE	19358,8536	196,062	92,7952	33753,6994
	R2	**0,8656**	**0,9019**	**0,788**	**0,7842**
	RER	20,9607	20,5494	16,942	16,1053
	RMSE	139,1361	14,0022	9,633	183,7218
	RPD	2,7273	3,192	2,172	2,1525
	RPIQ	1,1085	2,1013	2,2109	2,2481
	SEP	139,2384	14,0125	9,6401	183,8568
Validação	**BIAS**	0,0033	-0,4453	-0,1785	-6,5058
	MSE	122776,9167	1119,7873	530,8182	197688,1478
	R2	**0,1473**	**0,4395**	**0**	**0**
	RER	8,3232	8,5994	7,0838	6,6556
	RMSE	350,3954	33,4632	23,0395	444,6214
	RPD	1,083	1,3357	0,9081	0,8894
	RPIQ	0,4402	0,8793	0,9244	0,9289
	SEP	350,6529	33,4849	23,0557	444,9005
Predição	**BIAS**	-2,25E+01	-2,92E+00	-2,74E-02	7,10E+00
	MSE	1,85E+05	1,42E+03	5,65E+02	2,16E+05
	R2	**0,00E+00**	**0,00E+00**	**0,00E+00**	**0,00E+00**
	RER	6,78E+00	7,65E+00	5,70E+00	5,84E+00
	RMSE	4,30E+02	3,77E+01	2,38E+01	4,65E+02
	RPD	8,74E-01	1,17E+00	8,60E-01	8,59E-01
	RPIQ	3,28E-01	6,10E-01	9,36E-01	1,13E+00
	SEP	4,31E+02	3,77E+01	2,38E+01	4,66E+02

ANEXO II - TABELAS DE RESULTADOS DA MODELAGEM COMPUTACIONAL PARA O CONJUNTO DE DADOS EETA

O presente anexo contém as tabelas de resultados para os modelos de calibração, validação e predição construídos a partir do conjunto de dados oriundos do complexo enzimático EETA:

Tabela 25 – Resultados de desempenho da calibração, validação e predição do modelo GBR construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
GBR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	0,003	0,000	0,000	0,081
	R^2	**0,999**	**0,994**	**0,998**	**0,998**
	RER	88,137	54,733	62,722	69,877
	RMSE	**0,051**	**0,003**	**0,001**	**0,285**
	RPD	**31,293**	**13,179**	**20,142**	**23,412**
	RPIQ	62,795	15,144	38,171	31,459
	SEP	0,052	0,003	0,001	0,287
Validação	**BIAS**	0,029	0,000	0,000	-0,111
	MSE	0,284	0,000	0,000	8,949
	R^2	**0,891**	**0,693**	**0,820**	**0,799**
	RER	8,523	7,493	7,346	6,656
	RMSE	**0,533**	**0,021**	**0,011**	**2,991**
	RPD	**3,022**	**1,804**	**2,359**	**2,229**
	RPIQ	6,064	2,073	4,470	2,994
	SEP	0,537	0,021	0,011	3,015
Predição	**BIAS**	0,024	-0,003	0,000	0,636
	MSE	0,125	0,000	0,000	4,243
	R^2	**0,946**	**0,873**	**0,920**	**0,873**
	RER	10,719	10,642	11,141	9,981
	RMSE	**0,353**	**0,010**	**0,007**	**2,060**
	RPD	**4,285**	**2,808**	**3,533**	**2,804**
	RPIQ	8,877	3,001	6,337	1,884
	SEP	0,361	0,009	0,007	2,008

Parâmetros utilizados: Pré-processamento→10, loss: ls, semente-→ 0

Tabela 26 – Resultados de desempenho da calibração, validação e predição do modelo PLS construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
PLS	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	0,263	0,000	0,000	3,808
	R2	**0,893**	**0,839**	**0,924**	**0,903**
	RER	8,840	11,378	12,221	10,196
	RMSE	**0,513**	**0,014**	**0,007**	**1,951**
	RPD	**3,052**	**2,494**	**3,636**	**3,217**
	RPIQ	6,181	2,567	5,349	2,924
	SEP	0,517	0,014	0,007	1,968
Validação	**BIAS**	-0,029	0,000	0,000	-0,076
	MSE	0,422	0,001	0,000	7,962
	R2	**0,828**	**0,547**	**0,831**	**0,798**
	RER	6,985	6,777	8,181	7,054
	RMSE	**0,650**	**0,023**	**0,010**	**2,822**
	RPD	**2,409**	**1,485**	**2,434**	**2,225**
	RPIQ	4,879	1,529	3,580	2,022
	SEP	0,655	0,024	0,010	2,845
Predição	**BIAS**	-0,066	-0,004	0,000	-0,840
	MSE	0,405	0,001	0,000	5,187
	R2	**0,852**	**0,682**	**0,839**	**0,893**
	RER	7,008	7,487	6,942	9,249
	RMSE	**0,637**	**0,021**	**0,012**	**2,277**
	RPD	**2,601**	**1,774**	**2,490**	**3,054**
	RPIQ	5,378	1,827	4,799	5,849
	SEP	0,649	0,021	0,012	2,169

Parâmetros: Pré-processamento→9, VL: 6, semente-→ 88

Tabela 27 – Resultados de desempenho da calibração, validação e predição do modelo *Kernel-Ridge* construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
Ridge	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,025	0,001	0,001	0,141
	MSE	0,083	0,000	0,000	1,465
	R2	**0,969**	**0,900**	**0,960**	**0,967**
	RER	15,843	13,164	15,384	16,554
	RMSE	**0,287**	**0,012**	**0,005**	**1,210**
	RPD	**5,678**	**3,162**	**4,982**	**5,479**
	RPIQ	11,188	3,433	9,521	5,718
	SEP	0,289	0,012	0,005	1,212
Validação	**BIAS**	0,031	0,001	0,001	0,269
	MSE	0,297	0,001	0,000	5,472
	R2	**0,889**	**0,657**	**0,845**	**0,876**
	RER	8,339	7,093	7,823	8,563
	RMSE	**0,545**	**0,022**	**0,010**	**2,339**
	RPD	**2,995**	**1,706**	**2,541**	**2,835**
	RPIQ	5,901	1,852	4,855	2,958
	SEP	0,548	0,023	0,010	2,343
Predição	**BIAS**	0,265	-0,001	0,005	0,187
	MSE	0,202	0,000	0,000	2,609
	R2	**0,904**	**0,835**	**0,923**	**0,927**
	RER	10,400	8,786	16,667	11,412
	RMSE	**0,449**	**0,011**	**0,006**	**1,615**
	RPD	**3,219**	**2,460**	**3,592**	**3,707**
	RPIQ	5,884	2,615	6,868	2,091
	SEP	0,372	0,011	0,004	1,644

Parâmetros: Pré-processamento→8, kernel= poly, degree=3, semente-→ 56

Tabela 28 – Resultados de desempenho da calibração, validação e predição do modelo *Ridge* construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
Ridge	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	0,140	0,000	0,000	2,720
	R2	**0,946**	**0,846**	**0,937**	**0,933**
	RER	12,130	10,646	12,507	12,065
	RMSE	**0,374**	**0,015**	**0,007**	**1,649**
	RPD	**4,319**	**2,545**	**3,975**	**3,863**
	RPIQ	8,598	2,488	6,636	3,214
	SEP	0,377	0,015	0,007	1,663
Validação	**BIAS**	-0,072	-0,002	-0,001	-0,183
	MSE	0,481	0,001	0,000	9,052
	R2	**0,816**	**0,474**	**0,771**	**0,777**
	RER	6,576	5,789	6,601	6,625
	RMSE	**0,693**	**0,028**	**0,012**	**3,009**
	RPD	**2,329**	**1,379**	**2,087**	**2,117**
	RPIQ	4,636	1,348	3,484	1,761
	SEP	0,695	0,028	0,012	3,029
Predição	**BIAS**	-0,026	-0,004	-0,005	-0,673
	MSE	0,135	0,000	0,000	8,634
	R2	**0,942**	**0,697**	**0,892**	**0,809**
	RER	10,318	6,719	12,347	6,839
	RMSE	**0,367**	**0,015**	**0,008**	**2,938**
	RPD	**4,136**	**1,818**	**3,043**	**2,288**
	RPIQ	7,754	2,555	5,730	1,940
	SEP	0,375	0,015	0,006	2,931

Parâmetros: Pré-processamento→1, , semente-→ 96

Tabela 29 – Resultados de desempenho da calibração, validação e predição do modelo PCR construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
PCR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	0,192	0,000	0,000	16,342
	R2	**0,919**	**0,572**	**0,899**	**0,602**
	RER	8,755	5,443	9,624	4,922
	RMSE	**0,438**	**0,018**	**0,008**	**4,043**
	RPD	**3,512**	**1,528**	**3,141**	**1,585**
	RPIQ	7,253	2,122	6,399	1,535
	SEP	0,442	0,019	0,008	4,077
Validação	**BIAS**	-0,016	0,000	0,000	-0,016
	MSE	0,230	0,000	0,000	19,382
	R2	**0,903**	**0,514**	**0,880**	**0,528**
	RER	8,004	5,109	8,857	4,519
	RMSE	**0,480**	**0,020**	**0,008**	**4,403**
	RPD	**3,209**	**1,434**	**2,891**	**1,456**
	RPIQ	6,627	1,991	5,889	1,409
	SEP	0,483	0,020	0,008	4,440
Predição	**BIAS**	0,220	0,017	0,004	1,427
	MSE	1,360	0,002	0,000	30,409
	R2	**0,538**	**0,253**	**0,684**	**0,287**
	RER	3,875	4,002	5,195	3,673
	RMSE	**1,166**	**0,043**	**0,016**	**5,514**
	RPD	**1,472**	**1,157**	**1,780**	**1,184**
	RPIQ	2,859	1,030	2,671	1,218
	SEP	1,174	0,040	0,016	5,458

Parâmetros: Pré-processamento→1 , NPC:4, semente-→ 24

Tabela 30 – Resultados de desempenho da calibração, validação e predição do modelo MLR construído para o conjunto de dados do complexo enzimático EETA

Modelo:		**Atividade enzimática**			
MLR	**Medida**	**β-glicosidase**	**Cmcase**	**Fpase**	**Xilanase**
Calibração	**BIAS**	0,000	0,000	0,000	0,000
	MSE	0,000	0,000	0,000	0,000
	R2	**1,000**	**1,000**	**1,000**	**1,000**
	RER	--	--	--	--
	RMSE	**0,000**	**0,000**	**0,000**	**0,000**
	RPD	--	--	--	--
	RPIQ	--	--	--	--
	SEP	0,000	0,000	0,000	0,000
Validação	**BIAS**	0,013	0,000	-0,001	-0,189
	MSE	0,965	0,002	0,000	10,568
	R2	**0,601**	**0,000**	**0,444**	**0,721**
	RER	4,617	3,424	4,267	6,131
	RMSE	**0,982**	**0,046**	**0,019**	**3,251**
	RPD	**1,583**	**0,790**	**1,341**	**1,891**
	RPIQ	3,226	0,736	2,209	1,849
	SEP	0,991	0,047	0,019	3,273
Predição	**BIAS**	-0,015	-0,003	0,001	0,784
	MSE	0,552	0,001	0,000	19,170
	R2	**0,801**	**-0,327**	**0,778**	**0,502**
	RER	5,081	2,868	5,494	4,159
	RMSE	**0,743**	**0,034**	**0,011**	**4,378**
	RPD	**2,240**	**0,868**	**2,122**	**1,416**
	RPIQ	4,567	1,145	4,276	2,561
	SEP	0,761	0,035	0,011	4,414

Parâmetros: Pré-processamento→10 , , semente-→ 60

ANEXO III – IMPLEMENTAÇÃO DA MODELAGEM COMPUTACIONAL

Todos os algoritmos implementados para a modelagem computacional, em linguagem python, bem como a base de dados utilizada, estão disponíveis no sítio do Github através do endereço eletrônico https://github.com/amandarochac/tese/ e podem ser utilizados livremente para fins de validação.