



Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Uma Estratégia de Ensemble aplicada sobre Modelos de Predição de Evasão do Ensino Superior com janelas semestrais

Maurílio Lacerda Leonel Júnior

Teresina-PI, Abril de 2024

Maurílio Lacerda Leonel Júnior

Uma Estratégia de Ensemble aplicada sobre Modelos de Predição de Evasão do Ensino Superior com janelas semestrais

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Computação Aplicada), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: André Macedo Santana

Teresina-PI

Abril de 2024

FICHA CATALOGRÁFICA

Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Divisão de Representação da Informação

L583e Leonel Júnior, Maurílio Lacerda.

Uma Estratégia de Ensemble aplicada sobre Modelos de
Predição de Evasão do Ensino Superior com janelas semestrais
/ Maurílio Lacerda Leonel Júnior. – – 2024.

121 f. : il.

Dissertação (Mestrado) – Universidade Federal do Piauí,
Centro de Ciências da Natureza, Programa de Pós-Graduação
em Ciência da Computação, Teresina, 2024.

Orientação: Prof. Dr. André Macedo Santana.

1. Mineração de Dados Educacionais (MDE). 2. Ensino
superior - Evasão. 3. Janelas semestrais. 4. Métodos
Ensemble. I. Santana, André Macedo. III. Título.

CDD - 006.32

Maurílio Lacerda Leonel Júnior

Uma Estratégia de Ensemble aplicada sobre Modelos de Predição de Evasão do Ensino Superior com janelas semestrais

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Computação Aplicada), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 22 de abril de 2024:

Documento assinado digitalmente
 **ANDRE MACEDO SANTANA**
Data: 19/06/2024 11:25:50-0300
Verifique em <https://validar.iti.gov.br>

André Macedo Santana

Orientador
Documento assinado digitalmente
 **VINICIUS PONTE MACHADO**
Data: 13/06/2024 21:49:38-0300
Verifique em <https://validar.iti.gov.br>

Vinícius Ponte Machado

Coorientador
Documento assinado digitalmente
 **RAIMUNDO SANTOS MOURA**
Data: 13/06/2024 10:29:22-0300
Verifique em <https://validar.iti.gov.br>

Raimundo Santos Moura

Convidado 2
Documento assinado digitalmente
 **IALIS CAVALCANTE DE PAULA JUNIOR**
Data: 13/06/2024 10:53:37-0300
Verifique em <https://validar.iti.gov.br>

Iális Cavalcante de Paula Júnior
Avaliador Externo

Teresina-PI
Abril de 2024

*Aos meus pais Teresinha e Leonel,
por sempre estarem comigo em todos os momentos e me educarem para o bem,
incondicionalmente;*

*À minha esposa, Nayara,
pelo amor, pelo apoio, pela compreensão, pelo zelo com nossa família e pela dádiva de sua
companhia diária há tantos anos;*

*E às criaturas mais importantes e valiosas de minha vida, meus filhos Clarisse e Rodrigo,
que diariamente me dão significado à vida e força para travar as batalhas, por eles e para
eles!*

Agradecimentos

Agradeço a Deus.

Agradeço à minha mãe, Teresinha Leonel, por toda uma vida de doação e exemplos, dedicação ao crescimento, desenvolvimento e independência de todas as pessoas que a cercam, em especial à mim, às minha irmãs, meus filhos e sobrinhos; Ao meu pai, Maurílio Leonel, por todos os ensinamentos e por fomentar em mim o desejo de fazer parte desta instituição, a UFPI, como estudante e como profissional.

Agradeço às minhas irmãs, e à todos os meus familiares e amigos pelo incentivo e compreensão.

Agradeço ao meu orientador, André Macedo Santana, pela confiança de sempre, pelos conselhos, pela paciência, condução e parceria.

Aos meus amigos Cledjan Torres, Euclydes Melo, Taison Almeida, Matheus Campanhã e Ricardo Lira por todo apoio técnico, incentivo e parceria antes, durante e depois do desenvolvimento dos trabalhos do Mestrado.

Aos professores do PPGCC Vinícius Machado, Rodrigo Melo, Laurindo Neto e Raimundo Moura, com os quais pude aprender muito além de conteúdo, bem como o professor Jonas Alves, da UFDPAr, pelo apoio, solicitude e generosidade no decorrer do Estágio Docente.

À UFPI e UFDPAr, instituições que marcaram minha vida profissional e pessoal por inúmeros motivos, e pelo apoio e compreensão quando necessitei me dedicar por mais tempo à conclusão de minhas Pós-Graduações e à realização deste trabalho de pesquisa.

*“A tecnologia não é boa nem má;
depende de como a usamos.”
(William Gibson)*

Resumo

Com o aumento no volume de dados gerados pelos sistemas de informação educacionais, a Mineração de Dados se tornou uma abordagem promissora de investigação sobre importantes causas educacionais, como a evasão nos mais diversos níveis de ensino, em especial o ensino superior. Mais de 2,3 milhões de alunos aptos a matricular-se em disciplinas em 2021 abandonaram o ensino superior, de acordo com o censo mais recente da educação superior, resultando em um prejuízo financeiro superior a R\$ 23 bilhões. Neste estudo de Mineração de Dados Educacionais (MDE), desenvolvemos modelos de classificação com aprendizado supervisionado e cinco abordagens com métodos Ensembles aplicadas a dados reais coletados de uma Instituição Federal de Ensino Superior. No primeiro cenário, criamos modelos do tipo Bagging utilizando a estratégia de Janelas Semestrais na construção dos conjuntos de dados envolvidos. No segundo cenário, além da abordagem de Janelas Semestrais (Nível Período), criamos conjuntos de dados em outros dois níveis adicionais (Curso e Instituição) e combinamos os modelos gerados (cada um associado a um conjunto de dados de nível distinto) por meio de técnicas de Ensemble Stacking. A combinação das abordagens formando modelos Stacking com as previsões dos modelos Bagging resultou em um maior nível de robustez e na maximização dos resultados (mensurados por Acurácia e MCC) mesmo com o uso de um único algoritmo por modelo, ao explorar a diversidade nos diferentes níveis de dados.

Palavras-chaves: MDE. Evasão. Janelas Semestrais. Métodos Ensemble.

Abstract

With the increasing volume of data generated by educational information systems, Data Mining has become a promising approach to investigate important educational issues, such as dropout rates across various educational levels, particularly in higher education. Over 2.3 million eligible students dropped out of higher education in 2021, according to the latest census on higher education, resulting in a financial loss exceeding R\$ 23 billion. In this study of Educational Data Mining (EDM), we developed classification models using supervised learning and five approaches with Ensemble methods applied to real data collected from a Federal Higher Education Institution. In the first scenario, we created Bagging models using the Academic Periods strategy in building the datasets involved. In the second scenario, in addition to the Academic Periods approach (Period Level), we created datasets at two additional levels (Course and Institution) and combined the generated models (each associated with a dataset of a distinct level) through Stacking Ensemble techniques. Combining the approaches to form Stacking models with the predictions of Bagging models resulted in a higher level of robustness and maximization of results (measured by Accuracy and MCC) even with the use of a single algorithm per model, by exploring diversity across different data levels.

Keywords: Educacional Data Mining (EDM). Dropout. Academic Periods. Ensemble Methods.

Lista de ilustrações

Figura 1 – Indicadores de trajetória dos ingressantes nas 2017 em IES brasileiras. Fonte: (SEMESP, 2023)	2
Figura 2 – Visão Geral da Proposta	5
Figura 3 – Estrutura do Trabalho	7
Figura 4 – Fases do processo de KDD. Fonte: Adaptado por Rigo et al. (2014) . .	10
Figura 5 – Exemplo de matriz de confusão. Fonte: (RUBACK; AVILA; CANTERO, 2021)	14
Figura 6 – Aprendizado de Máquina. Fonte: Adaptado de Araujo (2014)	17
Figura 7 – Árvore de decisão traduzida e adaptada a partir da fonte: (SACRAMENTO, 2021)	19
Figura 8 – Três razões fundamentais pelas quais um ensemble pode trabalhar melhor que um classificador individual. Fonte: (DIETTERICH, 2000) .	22
Figura 9 – Exemplo de classificadores independentes (A) combinados por ensemble (B). Fonte: Adaptado de Sagi e Rokach (2018)	23
Figura 10 – Características dos Ensembles Homogêneos e Heterogêneos.	24
Figura 11 – O Processo de Ensemble Bagging. Fonte: (KALIRANE, 2024)	27
Figura 12 – O Processo de Ensemble Boosting. Fonte: (KALIRANE, 2024)	28
Figura 13 – O Processo de Ensemble Stacking. Fonte: (KALIRANE, 2024)	28
Figura 14 – O Processo de Ensemble por Votação Majoritária	29
Figura 15 – Visão detalhada da metodologia proposta para EC1 e EC2	37
Figura 16 – Detalhamento da Etapa Coleta e Pré-Processamento	39
Figura 17 – Detalhamento da Etapa Separação dos Dados de Teste e Criação dos Níveis.	42
Figura 18 – Natureza dos dados de cada nível.	43
Figura 19 – Número de Amostras x Número de Evadidos - MAT e PED	44
Figura 20 – Número de Amostras x Número de Evadidos - Demais Cursos	45
Figura 21 – Etapa de Separação dos Dados de Teste e Criação dos Níveis, ampliada em EC3 com a criação dos novos Conjuntos de Features e dos Datasets .	48
Figura 22 – Ampliação da etapa de Criação dos modelos com foco no quantitativo de modelos criados ao longo de EC1, EC2 e EC3.	49
Figura 23 – Detalhamento da etapa de Criação dos Modelos de Treinamento do EC1.	51
Figura 24 – Média de Acurácia por Nível estratificadas por curso e tipo do modelo	52
Figura 25 – Comparativo entre ACC de DT e BAG por curso e período	53
Figura 26 – Médias de Acurácia por Nível	55
Figura 27 – Médias de Acurácia por Nível estratificadas por Período	56
Figura 28 – Detalhamento do processo de montagem do modelo Stack	60

Figura 29 – Médias de Acurácia por Tipo do Modelo (DT, BAG e STACK.)	61
Figura 30 – Médias de Acurácia por Tipo do Modelo estratificadas por período . .	61
Figura 31 – Médias de Acurácia por Tipo do Modelo (IND e STACK) estratificadas por curso e período.	62
Figura 32 – Detalhamento da inserção de RF e ET na fase de criação dos modelos .	63
Figura 33 – Detalhamento do processo de montagem específico dos modelos VOT, VOT BAG e STACK BAG.	65
Figura 34 – Comparação por curso e período do MCC de STACK BAG com DT e STACK BAG com ET	69
Figura 35 – Gráfico Boxplot das Acurácias de cada estratégia agrupadas por período, contendo dados de todas as combinações.	71
Figura 36 – Gráfico Boxplot dos valores de MCC de cada estratégia agrupados por período, contendo dados da combinação com ORIG e ET.	72

Lista de tabelas

Tabela 1	– Features originalmente presentes em cada um dos níveis de <i>Datasets</i> .	47
Tabela 2	– Resultados da comparação entre as abordagens DT e BAG utilizando Teste de Hipótese T pareado.	55
Tabela 3	– Resultados da comparação entre as abordagens IND (DT) e STACK (DT) utilizando Teste de Hipótese T Pareado.	63
Tabela 4	– Features presentes em cada um dos níveis de cada conjunto de <i>Datasets</i> da proposta	64
Tabela 5	– Percentual de Melhoria e Aplicabilidade de BAG sobre IND mensurados pela ACC para o conjunto de 66 modelos com estimador-base DT usando FILT.	66
Tabela 6	– Combinações relacionadas aos modelos Bagging.	67
Tabela 7	– Percentual de Melhoria e Aplicabilidade de STACK BAG sobre IND mensurados por MCC para o conjunto de 66 modelos com estimador-base ET usando ORIG.	70
Tabela 8	– Maiores percentuais de Melhoria por período para cada métrica, entre todos os tipos de modelo.	70

Lista de abreviaturas e siglas

ACC	Acurácia Geral
ACO	Algoritmos de Otimização por Colônias de Formiga
AD	Árvore de Decisão
ADM	Curso de Administração
AM	Aprendizado de Máquina
APX	Percentual de Aplicabilidade calculado para cada período baseado no teste de hipótese T
AUC	Área Sob a Curva ROC
BAG	Estratégia de criação de modelos utilizando métodos Ensemble Bagging
BIO	Curso de Ciências Biológicas
BIOM	Curso de Biomedicina
CON	Curso de Ciências Contábeis
CV	Cross-validation
DCBD	Descoberta de Conhecimento em Bases de Dados
DT	Decision Tree
EC	Estudo de Caso
EC1	Estudo de Caso 1
EC2	Estudo de Caso 2
EC3	Estudo de Caso 3
ECO	Curso de Ciências Econômicas
ENEM	Exame Nacional do Ensino Médio
ET	Extra-Trees
ETL	(Processo de) Extração, Transformação e Carga
FILT	Novo conjunto de datasets criados a partir dos datasets originais

FIS	Curso de Fisioterapia
FN	Falso Negativo
FP	Falso Positivo
IES	Instituições de Ensino Superior
IFES	Instituições Federais de Ensino Superior
IND	Algoritmo classificador individual
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	Knowledge Discovery in Databases
MAT	Curso de Matemática
MCC	Coeficiente de Correlação de Matthews
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação
MPX	Percentual de Melhoria calculado para cada período baseado no teste de hipótese T
NC	Nível Curso
NI	Nível Instituição
NP	Nível Período
ORIG	Conjunto de datasets originalmente criados na proposta
P1	1º Período cronológico do discente na instituição
P2	2º Período cronológico do discente na instituição
P3	3º Período cronológico do discente na instituição
P4	4º Período cronológico do discente na instituição
P5	5º Período cronológico do discente na instituição
P6	6º Período cronológico do discente na instituição
PED	Curso de Pedagogia

PES	Curso de Engenharia de Pesca
PSI	Curso de Psicologia
PSO	Algoritmo por Enxame de Partículas
RF	Random Forest
SEMESP	Sindicato das mantenedoras de instituições de ensino superior do Brasil
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SISU	Sistema de Seleção Unificada
STACK BAG	Estratégia de criação de modelos utilizando métodos Ensemble Stacking sobre as predições da estratégia de BAG
STACK	Estratégia de criação de modelos utilizando métodos Ensemble Stacking
SVM	Support Vector Machines
TUR	Curso de Turismo
UFDFar	Universidade Federal do Delta do Parnaíba
UFPI	Universidade Federal do Piauí
UFRN	Universidade Federal do Rio Grande do Norte
VC	Validação Cruzada
VN	Verdadeiro Negativo
VOT BAG	Estratégia de criação de modelos utilizando métodos Ensemble Voting sobre as predições da estratégia de BAG
VOT	Estratégia de criação de modelos utilizando métodos Ensemble Voting (Votação Majoritária)
VP	Verdadeiro Positivo

Sumário

1	INTRODUÇÃO	1
1.1	Contexto e Motivação	1
1.2	Objetivos e Contribuição	4
1.3	Visão Geral da Proposta	5
1.4	Organização do Trabalho	6
2	REFERENCIAL TEÓRICO	9
2.1	Contextualização	9
2.2	Descoberta de Conhecimento em Bases de Dados	9
2.2.1	Coleta	10
2.2.2	Pré-Processamento	11
2.2.3	Transformação	12
2.2.4	Mineração de Dados	13
2.2.5	Análise e Interpretação dos Resultados	13
2.2.5.1	Acurácia	14
2.2.5.2	MCC	15
2.3	Aprendizado de Máquina	16
2.3.1	Aprendizado Supervisionado	18
2.3.1.1	Técnica de Classificação	18
2.3.2	Algoritmos	19
2.3.2.1	Algoritmo Decision Tree	19
2.3.2.2	Algoritmo Random Forest	20
2.3.2.3	Algoritmo Extra Trees	21
2.3.3	Aprendizado Ensemble (Ensemble Learning)	21
2.3.3.1	Diversidade dos Classificadores-base	23
2.3.3.2	Desafios para os Métodos Ensemble	25
2.3.3.3	Técnica Ensemble Bagging	26
2.3.3.4	Técnica Ensemble Boosting	27
2.3.3.5	Técnica Ensemble Stacking	27
2.3.3.6	Técnica Ensemble Voting	28
2.4	Mineração de Dados Educacionais	29
2.4.1	Trabalhos Relacionados	30
2.4.2	Considerações do Capítulo	34
3	METODOLOGIA PROPOSTA E RESULTADOS ENCONTRADOS	35
3.1	Contextualização	35

3.2	Ferramentas	36
3.3	Compreensão dos Dados e Criação dos Datasets	38
3.3.1	Etapa: Coleta e Pré-Processamento	39
3.3.2	Etapa: Separação dos Dados de Teste e Criação dos Níveis	41
3.4	Construção e Treinamento dos Modelos	48
3.4.1	Estudo de Caso 1 (EC1)	50
3.4.1.1	Resultados do Estudo de Caso 1	51
3.4.1.2	Observações complementares de EC1	55
3.4.1.3	Estudo de Caso 2	58
3.4.1.4	Resultados do Estudo de Caso 2	60
3.4.1.5	Estudo de Caso 3	63
3.4.1.6	Resultados do Estudo de Caso 3	66
3.5	Considerações do Capítulo	71
4	CONCLUSÕES E TRABALHOS FUTUROS	73
	REFERÊNCIAS	77
	APÊNDICES	83
	APÊNDICE A – RESULTADOS DOS TESTES ESTATÍSTICOS	85
	APÊNDICE B – RESULTADOS DAS MÉTRICAS	93

1 Introdução

1.1 Contexto e Motivação

A evasão no ensino superior brasileiro é um fenômeno multifatorial que afeta o desenvolvimento do país, influenciando tanto o aspecto educacional quanto o econômico e social. As Instituições de Ensino Superior do Brasil, públicas e privadas, enfrentam desafios significativos relacionados à evasão, que impactam diretamente no melhor uso de seus recursos e suas finanças (SILVA; SANTOS, 2017). Bem como no progresso acadêmico dos estudantes, podendo atrasar ou cancelar sonhos, limitar oportunidades de emprego e crescimento pessoal, além de afetar a renda (TONTINI; WALTER, 2014).

Para o Estado, o abandono dos estudantes representa um desperdício de recursos, com estruturas acadêmicas subutilizadas devido à saída de estudantes. Essa situação também pode indicar falhas no processo de ensino (GILIOLI, 2016), afetando a qualidade do ensino e a sustentabilidade das instituições. Isso resulta em ociosidade de espaço físico, professores, funcionários e equipamentos, levando a desperdício de investimentos no caso de instituições públicas e à redução de receita em instituições privadas (FILHO et al., 2007).

Esse é um problema que afeta todos os níveis de ensino e é particularmente desafiador no ensino superior brasileiro, tanto público quanto privado, como destacado por Lobo (2012). Existem várias formas de evasão, incluindo a evasão do curso, a da instituição de ensino superior (IES), a do sistema e a do aluno. Estudos em várias áreas enfatizam a importância de compreender a evasão do aluno, pois ela é a raiz de todos os outros tipos de evasão, impactando o curso, a IES e o sistema educacional como um todo.

No Ensino Superior, ela pode estar associada a fatores que antecedem e sucedem o ingresso dos estudantes na instituição. Aspectos internos e externos, demográficos, acadêmicos, pessoais e familiares, influenciam a decisão dos estudantes de permanecer ou deixar o curso, como destacado por Fritsch, Rocha e Vitelli (2015). Coulon (2017) ressalta a mudança fundamental que ocorre quando os estudantes ingressam na universidade, envolvendo a adaptação às regras, aos conhecimentos e às redes de sociabilidade desse novo ambiente. O fracasso acadêmico se torna uma ameaça quando os estudantes não conseguem se integrar adequadamente a esse ambiente, cumprir as demandas acadêmicas e se adaptar à organização institucional.

Estudos indicam que a maioria das evasões ocorre entre o primeiro e o segundo período acadêmico (TINTO, 2006) (BITENCOURT; SILVA; XAVIER, 2021)(PAZ; CAZELLA, 2017)(GILIOLI, 2016)(SALES; BALBY; CAJUEIRO, 2016), destacando os desa-

fios enfrentados pelos estudantes ao ingressar no ensino superior, incluindo a resolução de problemas, os relacionamentos interpessoais e a adaptação a um nível mais elevado de exigência e carga de trabalho, como apontado por [Sahão e Kienen \(2021\)](#). Segundo [Tinto \(2006\)](#), é fundamental envolver os estudantes desde o primeiro ano da educação superior, por meio da implementação de ações que enriqueçam sua experiência acadêmica.

Para ilustrar o cenário, dados dos Indicadores de Trajetória do [INEP \(2023\)](#) indicam que 54% dos estudantes que ingressaram em cursos de graduação em 2015 abandonaram suas formações após cinco anos, resultando em uma perda financeira significativa para o sistema de ensino. Segundo o censo mais recente da educação superior, mais de 2,3 milhões de alunos elegíveis para matrícula em 2021 abandonaram o ensino superior, resultando em um prejuízo financeiro de mais de R\$ 23 bilhões ([HIPOLITO, 2023](#)).

De acordo com [SEMESP \(2023\)](#), conforme mostra a Figura 1, a taxa de conclusão para os ingressantes de 2017 entre instituições públicas e privadas é de apenas 26,3%, com 55,5% de taxa de desistência acumulada. Os percentuais de desistência e conclusão são maiores nas instituições privadas. Isso por que, de acordo com os indicadores do INEP, a quantidade de alunos que não se formaram e não tem registro de evasão nas públicas é três vezes maior do que nas privadas.



Figura 1 – Indicadores de trajetória dos ingressantes nas 2017 em IES brasileiras. Fonte: ([SEMESP, 2023](#))

Em dados mais recentes, referentes aos ingressantes de 2018 e utilizando a mesma metodologia empregada por ([HIPOLITO, 2023](#)), constatou-se que o mesmo percentual de 54% observado para os ingressantes de 2015 foi atingido com um ano a menos de curso. No entanto, esses novos dados podem ser ainda mais preocupantes, pois este coorte corresponde a discentes que atravessaram o período pandêmico da Covid-19 durante o curso natural de seus estudos. Vale ressaltar que a coleta dos dados pelo INEP ocorreu em 2022,

enquanto a pandemia persistia, o que pode não refletir completamente os desligamentos de alunos que já não frequentavam mais as aulas, mas que ainda não haviam sido oficialmente desligados da instituição devido à expectativa de retorno, por exemplo.

A Universidade Federal do Delta do Parnaíba (UFDPAr), selecionada para este estudo, foi criada recentemente por meio do desmembramento de um campus avançado da Universidade Federal do Piauí (UFPI) e está em processo de estabelecer sua gestão independente. No entanto, muitos de seus processos ainda estão vinculados à sede da instituição original. Assim como em muitas outras universidades, a UFDPAr enfrenta o desafio da evasão acadêmica, refletido em seus indicadores individuais, que são tão preocupantes quanto os números globais mencionados anteriormente.

Por exemplo, nos cursos de Engenharia de Pesca e Matemática, a Taxa de Desistência Acumulada até 2021 atinge respectivamente 75% e 79,5% para a coorte de ingressantes em 2012 (INEP, 2023). Além disso, para a coorte de ingressantes de 2018 no curso de Ciências Econômicas, por exemplo, 42% já haviam formalmente evadido até o final de 2022. No entanto, para este último grupo, há preocupações com a qualidade dos dados devido ao período pandêmico.

A UFDPAr foi criada em 2019, e os cursos existentes no antigo campus da UFPI permaneceram inalterados, alguns datando da década de 1970. A maioria desses cursos foi criada em 2007 como parte do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (MEC, 2010) (MEC, 2020) (PORTAL..., 2018).

São 12 cursos com características distintas, não organizados em Departamentos, Centros de Ensino ou Faculdades, como é comum em instituições maiores, estando todos ligados à Reitoria. Vale mencionar que o curso de Medicina, o mais recente de todos, foi criado em 2014 e formou sua primeira turma durante a pandemia, em conformidade com a Portaria 374/2020 do MEC (2020), que permitiu que os alunos antecipassem suas formaturas para contribuir no combate à pandemia (UFDPAR, 2020). Devido a essas peculiaridades, juntamente com a disparidade no número de períodos e as baixíssimas taxas de evasão, decidimos não incluir este curso em nossa análise. Portanto, os 11 cursos analisados são: Psicologia (PSI), Pedagogia (PED), Ciências Contábeis (CON), Ciências Econômicas (ECON), Administração (ADM), Turismo (TUR), Biomedicina (BIOM), Fisioterapia (FIS), Engenharia de Pesca (PES), Ciências Biológicas (BIO) e Matemática (MAT).

É importante observar que, devido ao processo de implantação do sistema de gestão acadêmica iniciado em 2011 na instituição originária, os dados disponíveis com a consistência e estrutura necessárias para a mineração só estão disponíveis a partir de 2012.

Sendo a evasão um desafio crítico especialmente no ensino superior, várias pesquisas têm utilizado técnicas de Mineração de Dados Educacionais (MDE) para compreender o

comportamento dos alunos e fatores relacionados à aprendizagem (BAKER; ISOTANI; CARVALHO, 2011). O uso efetivo de recursos, especialmente em universidades públicas, torna-se essencial, e investir em programas de assistência estudantil eficazes pode reduzir significativamente o custo por aluno formado, desde que o aluno em risco de evasão seja identificado adequadamente. Nesse contexto, as técnicas de mineração de dados desempenham um papel fundamental na identificação de características associadas a uma maior probabilidade de evasão (BRITO; MELLO; ALVES, 2020).

Iniciando seu caminho, a UFDPPar precisa conduzir estudos próprios que apoiem a prevenção da evasão. Isso permitirá que a instituição identifique os fatores específicos de evasão em seu contexto, desenvolva estratégias personalizadas para retê-los e, ao mesmo tempo, promova a qualidade do ensino, assegurando o sucesso dos estudantes e a sustentabilidade de sua missão educacional, fortalecendo sua reputação no cenário educacional.

1.2 Objetivos e Contribuição

Um importante destaque deste trabalho é a aplicação dos métodos propostos em um cenário real, com dados devidamente tratados e anonimizados. Nossa proposta está estruturada em 3 Estudos de Caso (EC) com objetivos distintos envolvendo Métodos Ensemble:

Estudo de Caso 1 (EC1): Este EC tem como objetivo avaliar se é possível desenvolver modelos de predição de evasão mais eficazes por meio de Métodos Ensemble em cenários de janelas semestrais. Nele, comparamos os resultados de métodos Ensemble Bagging com os de algoritmos aplicados individualmente na predição de evasão, utilizando a abordagem de janelas semestrais.

Estudo de Caso 2 (EC2): Já este EC tem como objetivo criar modelos através de bases de dados distintas e um mesmo classificador, e avaliar a produção de predições de evasão mais eficazes ao combinar esses modelos através de métodos Ensemble Stacking. Cada base de dados corresponde a um nível de agregação dos dados, incluindo o de janelas semestrais.

Estudo de Caso 3 (EC3): Por fim, este EC objetiva experimentar novas técnicas, algoritmos e conjuntos de datasets, evoluindo a metodologia de EC1 e EC2 e buscando melhores resultados em termos de métricas e de percentual de modelos ensemble superiores aos treinados individualmente.

Nossas pesquisas não revelaram trabalhos que tenham abordado a combinação de diferentes níveis de organização de dados de estudantes, incluindo a dimensão temporal, com o objetivo de explorar a diversidade inerente a esses conjuntos de dados. Esse enfoque

se apresenta como uma vantagem significativa na aplicação de modelos ensemble para a predição de evasão em instituições de ensino superior. Portanto, as principais contribuições deste estudo residem na demonstração da eficácia dessa abordagem. Nossos resultados indicam uma melhoria significativa, segundo nossas métricas, em comparação com modelos de predição tradicionais. Embora ultrapasse o escopo deste estudo, a aplicação dessa abordagem com o uso de classificadores e parâmetros ótimos tem o potencial de auxiliar na identificação precoce de alunos em risco de evasão, permitindo que os administradores desenvolvam políticas de retenção mais direcionadas e eficazes.

1.3 Visão Geral da Proposta

Este trabalho visa explorar o uso de técnicas de Mineração de Dados Educacionais como uma ferramenta no combate à evasão no ensino superior público federal. A análise desses dados nos permite criar modelos preditivos para auxiliar as instituições de ensino na implementação de estratégias de intervenção direcionadas. Nossa abordagem envolve a aplicação de Métodos Ensemble em três níveis de agregação de dados, ou seja, três níveis de modelos (Instituição, Curso e Período), voltados para diferentes níveis de gestores acadêmicos de uma Instituição Federal de Ensino Superior (IFES) recentemente estabelecida, com dados reais. A Figura 2 exibe uma visão geral das etapas aplicadas em nosso processo de descoberta de conhecimento em bases de dados.

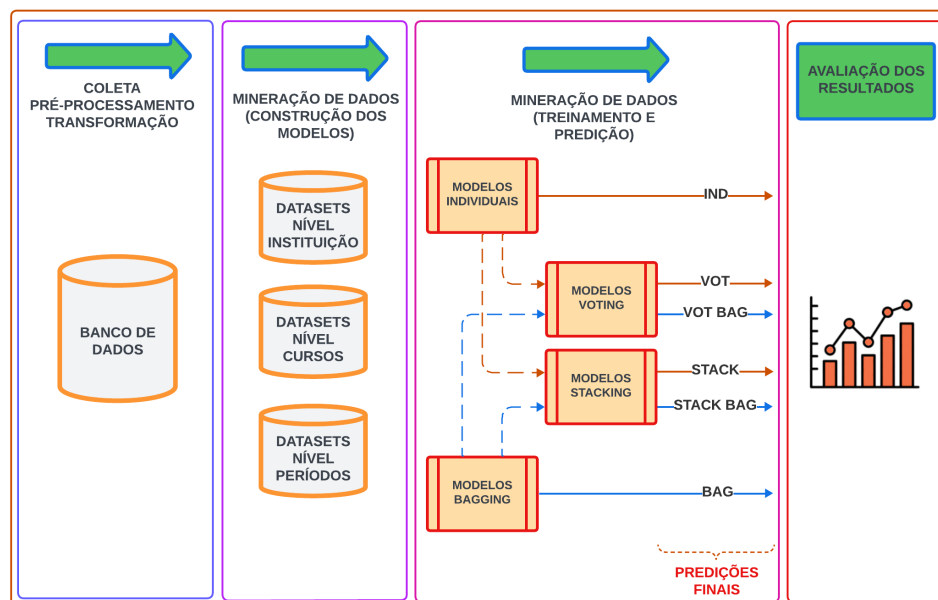


Figura 2 – Visão Geral da Proposta

Após a criação dos datasets em três níveis distintos a partir dos dados coletados, construímos os modelos com um classificador individual e um ensemble bagging (para cada modelo). As previsões de ambos serão avaliadas mas também servirão como entrada para

outras abordagens ensemble. No total os modelos produzirão previsões finais por meio de seis tipos distintos de modelos que serão avaliados pelas métricas e testes estatísticos escolhidos.

1.4 Organização do Trabalho

Este trabalho está organizado em 4 capítulos. Neste primeiro, intitulado Introdução, apresentamos um contexto sobre evasão nas Instituições de Ensino Superior (IES) brasileiras, e a instituição sobre a qual os dados foram coletados; os objetivos de nosso estudo juntamente com uma visão simplificada de nossa proposta; e por fim esta explicação sobre como está organizado o trabalho.

No capítulo 2, intitulado ‘Referencial Teórico’, serão apresentados os conceitos, técnicas, métodos e tecnologias utilizadas ao longo do processo, em especial referentes à Descoberta de Conhecimento em Bases de Dados e Aprendizado de Máquina; e os trabalhos de Mineração de Dados Educacionais que influenciaram este estudo.

No Capítulo 3, abordaremos as ferramentas e sua utilização durante a Descoberta de Conhecimento em Banco de Dados, desde a Coleta até a fase de Mineração de Dados, abordada junto à explanação do processo específico de cada Estudo de Caso e acompanhada da Análise dos Resultados de cada caso.

Por fim, no Capítulo 4, como um resumo do processo, sintetizamos os resultados e encerramos este documento, apontando também insights para trabalhos futuros relacionados a este estudo. A Figura 3 oferece uma visão mais detalhada e, ao mesmo tempo, concisa da organização deste trabalho em termos de capítulos e seções.

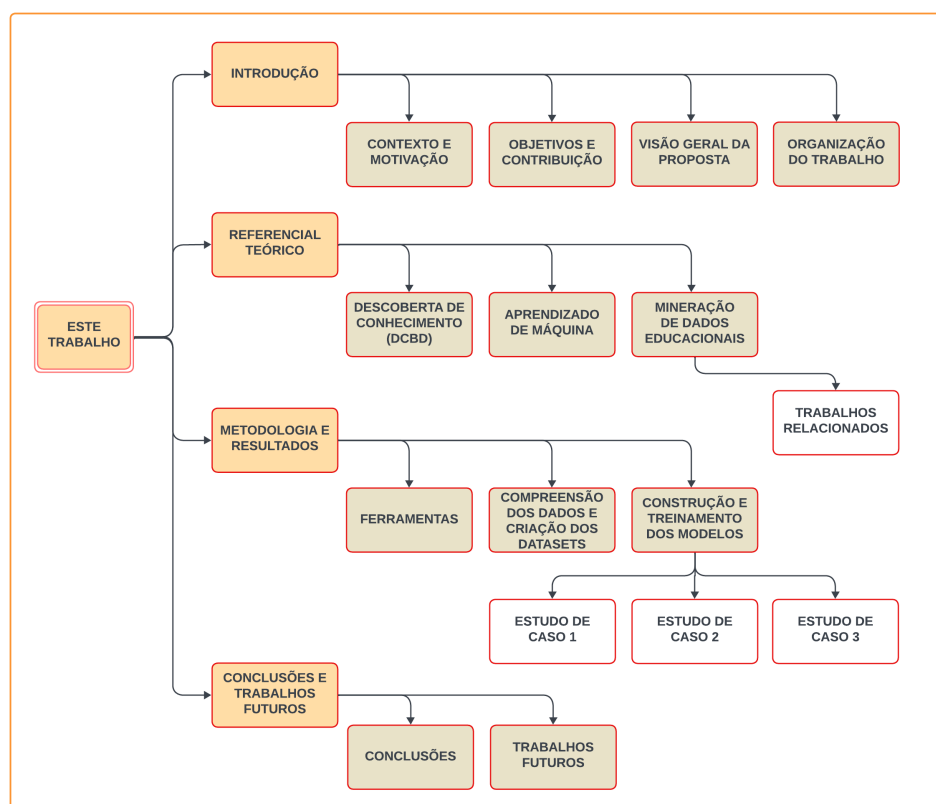


Figura 3 – Estrutura do Trabalho

2 Referencial Teórico

Neste capítulo, exploraremos o processo de Descoberta de Conhecimento em Bases de Dados, destacando seus conceitos e fases. Além disso, apresentaremos os fundamentos da Aprendizagem de Máquina, abordando tipos, métodos, técnicas e algoritmos, com foco especial no Aprendizado Supervisionado e no Aprendizado Ensemble. Por fim, discutiremos a relevância da Mineração de Dados Educacionais, especialmente através das influências em nosso estudo.

2.1 Contextualização

A rápida evolução tecnológica das últimas duas décadas, com a produção massiva de dados e a diversificação das formas de coleta e análise de informações, tem criado uma demanda crescente por algoritmos, métodos e aplicações capazes de lidar eficazmente com esse cenário em constante transformação. Nesse contexto, a Inteligência Artificial, em particular o Aprendizado de Máquina, tem se desenvolvido em paralelo à Mineração de Dados, possibilitando soluções tecnológicas inovadoras (SILVA et al., 2019). Soluções estas aplicáveis aos mais diversos temas, entre eles, o combate à evasão na educação superior. Neste referencial teórico, abordaremos alguns conceitos e temas envolvidos nesse combate.

2.2 Descoberta de Conhecimento em Bases de Dados

A Descoberta de Conhecimento em Bases de Dados (DCBD ou KDD, do inglês Knowledge Discovery in Databases), emergiu como uma abordagem que envolve não apenas a ciência da computação, mas também estatística, matemática, banco de dados e outras áreas, em virtude da necessidade de uma abordagem interdisciplinar para lidar com a nova realidade no universo de dados. Em um clássico e importante trabalho a respeito, Fayyad, Piatetsky-Shapiro e Smyth (1996) definiram o termo KDD como um processo complexo que visa identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados.

Durante este processo faz-se necessário o envolvimento de três atores/papéis principais: o Analista de Dados, condutor do processo de análise de dados e conhecedor das técnicas envolvidas no processo de DCBD; o Especialista de Domínio, conhecedor do campo/área de aplicação do estudo; e o Usuário, interessado nos resultados da análise de dados que subsidiarão a tomada de decisões e a implementação de ações relacionadas. Estes papéis são distintos, porém complementares e podem ser acumulados por um mesmo agente, bem como compartilhado entre mais de uma pessoa ou departamento (VIANA;

SANTANA; RABÊLO, 2022). O importante é a garantia de que a análise seja municiada pelos conhecimentos especializados do domínio e seja útil para as partes interessadas finais.

O modelo apresentado na Figura 4, uma adaptação do modelo proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996), sugere ainda que o processo de DCBD é composto por várias fases interconectadas (porém sem uma estrutura rígida, sendo possível transitar entre elas) cada uma com objetivos e técnicas específicas. Ela sugere o processo estruturado em cinco principais etapas: Seleção, Pré-processamento, Transformação, Mineração de Dados, Avaliação e Interpretação dos Resultados.

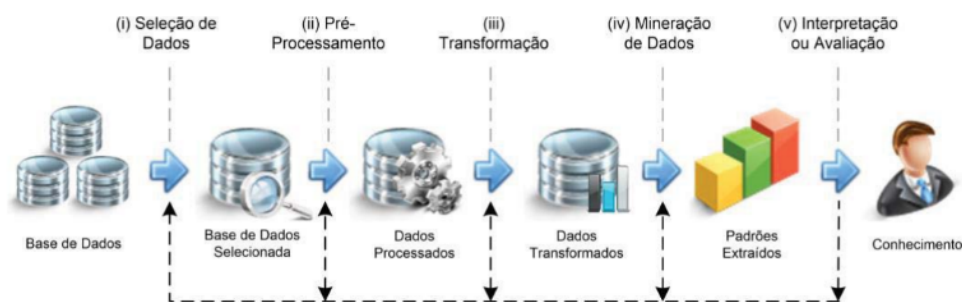


Figura 4 – Fases do processo de KDD. Fonte: Adaptado por Rigo et al. (2014)

2.2.1 Coleta

A fase de Coleta, também conhecida como fase de Seleção de Dados, é a primeira etapa fundamental do processo de DCBD. Ela envolve a compreensão do domínio da aplicação e do conhecimento prévio relevante, identificando o objetivo do processo de DCBD do ponto de vista do cliente, bem como a seleção de um conjunto de dados sobre os quais a descoberta será realizada (BRACHMAN; ANAND, 1994). O envolvimento dos especialistas de domínio é então primordial para o início do processo.

A seleção adequada dos dados é um pré-requisito fundamental para o sucesso de todo o processo de DCBD, pois a qualidade e a representatividade dos mesmos influenciam diretamente a eficácia das etapas subsequentes. A literatura apresenta três principais tipos de entraves que podem dificultar este processo (ARAUJO, 2014): problemas decorrentes de questões éticas e legais, como o uso inadequado dos resultados e a manipulação de dados sensíveis e privados dos indivíduos constantes na base analisada; aqueles relacionados a questões estratégicas que podem impactar de alguma forma na reputação, estratégias comerciais ou estratégias competitivas da instituição estudada; e problemas decorrentes de dificuldades de conectividade e obtenção dos dados, geralmente devido a questões tecnológicas, como a ausência de permissões de acesso ao banco de dados ou formatos de dados discrepantes entre as fontes envolvidas. A mitigação desses entraves será fundamental para o sucesso da etapa.

2.2.2 Pré-Processamento

A fase subsequente, conhecida como Pré-Processamento, abrange uma série de técnicas e operações destinadas a preparar os dados e é considerada uma das mais onerosas, podendo consumir cerca de 80% de todo o processo (PARMEZAN et al., 2012). Essas práticas englobam estratégias para lidar com campos de dados ausentes, o tratamento de dados ruidosos e diversas outras tarefas relacionadas à Limpeza dos Dados. Dados ausentes referem-se àqueles em que uma ou mais tuplas não contêm informações relacionadas a uma determinada característica. Nessas situações, a tupla (ou até mesmo a feature) pode ser excluída ou submetida a um processo de imputação de informação, utilizando média, mediana, valores das tuplas mais próximas ou outras abordagens de tratamento.

Dados ruidosos, por sua vez, são aqueles que, de alguma forma, fogem à regra, como um discente ativo com ano de nascimento em 1880 ou um produto com preço inferior a uma dezena monetária cadastrado com valor na casa dos milhões. A Análise Exploratória de Dados, uma abordagem estatística e visual para examinar, resumir e entender as principais características de um conjunto de dados, pode ser uma grande aliada na identificação de outliers ou dados incoerentes. As estratégias de mitigação também envolvem a exclusão de registros (ou features) e imputação de novo valor por medida estatística ou outras abordagens, ou ainda a simples adoção de modelos com algoritmos mais robustos a ruídos. Esses e outros ajustes são necessários, pois os dados pós-coleta frequentemente apresentam as anomalias citadas, tornando-os inadequados para a aplicação de algoritmos de mineração. Os métodos de limpeza e correção dos dados são essenciais para um treinamento rápido e eficaz (WANG; BAH; HAMMAD, 2019).

Outro desafio comum durante a fase de Pré-Processamento é o desbalanceamento dos dados, que ocorre quando uma classe é significativamente mais representada do que outra nos dados coletados. Esse desbalanceamento pode comprometer e enviesar os resultados do treinamento, pois a classe majoritária tende a predominar devido ao seu maior volume de dados. Para lidar com esse problema, existem diversas abordagens, desde o uso de pesos distintos para cada classe até estratégias de amostragem, como Undersampling (selecionando menos amostras da classe majoritária) ou Oversampling (gerando novas amostras sintéticas para a classe minoritária). Adicionalmente temos o uso de algoritmos e métricas mais adequados para lidar com dados desbalanceados, bem como as estratégias que envolvem aprendizado Ensemble (FUJIWARA et al., 2020). Essas estratégias, ao utilizar diferentes subconjuntos de dados na construção de modelos, permitem que cada um se especialize em certos aspectos do problema, incluindo o reconhecimento da classe minoritária, possibilitando uma classificação mais equilibrada através da combinação das predições.

Por fim, uma tarefa fundamental desta fase é a seleção dos atributos que serão utilizados na montagem/estruturação dos datasets. Um processo criterioso de escolha desses

atributos é essencial para garantir a eficiência e a qualidade do modelo. A seleção pode ser realizada por meio de diferentes métodos, cada um com suas vantagens e desvantagens. Esses métodos geralmente podem ser categorizados como: Filter, correspondendo aos métodos utilizados antes do treinamento, portanto independentes do algoritmo, e que selecionam um subconjunto de atributos ‘filtrando’ aqueles com menor relevância segundo algum critério; Wrapper, métodos mais exigentes computacionalmente, que utilizam o próprio algoritmo de classificação para avaliar os subconjuntos de atributos de acordo com a sua capacidade preditiva; e Embedded, que incorporam a seleção diretamente no processo de treinamento do modelo (PARMEZAN et al., 2012), sendo parte intrínseca do algoritmo de aprendizado. É importante considerar a natureza dos dados e os objetivos do modelo ao escolher como lidar com os dados desbalanceados e qual abordagem de seleção mais adequada, uma vez que o desempenho e a interpretabilidade do modelo final serão impactados diretamente por estes processos.

2.2.3 Transformação

Na fase de transformação, são executadas tarefas para reduzir, ampliar e organizar os dados, com o objetivo de encontrar as características, formatos e agregações úteis para representar os dados, dependendo do objetivo da tarefa (BRACHMAN; ANAND, 1994). Essa etapa envolve a manipulação e reestruturação dos dados brutos para atender aos requisitos específicos dos algoritmos, por meio de operações como Engenharia de Atributos, Normalização e Discretização dos dados.

A Engenharia de Atributos consiste na criação de novas variáveis para representar características de forma mais significativa para o domínio do problema. Isso pode incluir a combinação de variáveis existentes, a extração de características específicas dos dados, a criação de indicadores ou índices importantes para o contexto do estudo, e até mesmo a geração de atributos derivados de operações matemáticas ou estatísticas. Essa prática pode aprimorar significativamente o desempenho dos modelos de aprendizado de máquina, fornecendo-lhes informações mais discriminativas e relevantes.

O tratamento dos Atributos de Tipos de Dados Complexos é frequentemente considerado uma subárea da Engenharia de Atributos (embora possa ser visto como um processo separado), e tem como objetivo a conversão para formatos mais adequados ou a criação de novas variáveis derivadas desses atributos. Um exemplo comum é a conversão de um atributo do tipo data para outro do tipo inteiro, como criar o atributo Idade a partir da Data de Nascimento.

A Normalização visa ajustar as escalas dos atributos, trazendo-os para um intervalo específico. Essa operação é especialmente útil quando atributos distintos estão preenchidos com valores em escalas diferentes (por exemplo, atributo1: de 0 a 10; atributo2: de 0 a 1.000.000). Com ambos os atributos trabalhando na mesma escala, eles contribuirão

igualmente para o processo de aprendizado, evitando a predominância de um atributo com uma escala maior.

Por outro lado, as operações de Discretização visam simplificar a representação de dados contínuos, transformando-os em categorias discretas mais fáceis de analisar. Por exemplo, podemos categorizar a idade em fases como 'Jovem', 'Adulto' e 'Idoso'. Muitos algoritmos de aprendizado de máquina requerem dados numéricos como entrada, o que pode exigir a transformação de variáveis categóricas para formatos numéricos adequados por meio da Codificação de Variáveis Categóricas.

2.2.4 Mineração de Dados

Dentre essas etapas, a Mineração de Dados (MD) é considerada a mais importante, conforme destacado por [Castro e Ferrari \(2016\)](#). Nessa fase, ocorre a análise de grandes volumes de dados com o propósito de extrair conhecimento, atribuir significado e proporcionar utilidade às informações contidas nos dados. Durante essa etapa, os modelos de aprendizado são construídos com base nos padrões identificados durante a Análise Exploratória e na preparação dos dados. Os algoritmos são então empregados para identificar padrões ocultos nos dados, muitas vezes imperceptíveis a olho nu.

A MD atua em conjunto com o Aprendizado de Máquina neste processo, formando um ciclo iterativo. À medida que novos padrões são identificados, a compreensão do domínio é aprimorada, levando a ajustes nos dados e nas implementações das técnicas e algoritmos utilizados. Esse ciclo iterativo permite a melhoria contínua dos modelos e a obtenção de conhecimento mais refinado ao longo do tempo.

Os métodos e algoritmos empregados são escolhidos com base na compreensão do problema e na definição dos objetivos estabelecidos na fase da Coleta, por meio dos conhecimentos do Especialista de Domínio e do Analista de Dados. Essas técnicas podem ser organizadas, divididas ou categorizadas de diversas formas. Por exemplo, em tarefas Descritivas (técnicas de identificação de padrões, associações e clusterização) e tarefas Preditivas (englobando as técnicas Classificação e Regressão). Enquanto as tarefas de predição buscam prever o comportamento de dados futuros, as tarefas de descrição revelam padrões e propriedades presentes nos dados avaliados ([ARAUJO, 2014](#)).

2.2.5 Análise e Interpretação dos Resultados

Por fim, na fase de Avaliação e Interpretação, os resultados são avaliados quanto à qualidade e relevância prática em relação aos objetivos. Os modelos e padrões identificados são testados e validados, incluindo o teste de modelos preditivos em conjuntos de testes e a análise de segmentação (identificação de grupos ou segmentos distintos dentro de um conjunto de dados) ([BRACHMAN; ANAND, 1994](#)).

A análise dos resultados de uma classificação geralmente começa com o uso de uma matriz de confusão. Esta matriz destaca o número de classificações corretas e incorretas para cada uma das categorias envolvidas (VIANA; SANTANA; RABÊLO, 2022). Ela é construída a partir de quatro valores principais: Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN). Em nosso contexto esses quatro valores correspondem, respectivamente, à quantidade de casos em que uma evasão foi predita corretamente (VP), à quantidade de predições de evasão sobre casos que na verdade não eram de discente evasor (FP), à quantidade de discentes concluintes que tiveram predições apontando-os como não evadidos (VN) e à quantidade de discentes cuja predição apontava como não evasor um discente que na verdade viria a se evadir (FN). A figura 5 traz um exemplo de matriz de confusão. Em nosso contexto o exemplo aponta para 70 acertos (VP) e 30 erros (FP) na predição de evadidos, além de 50 acertos (VN) e 10 erros (FN) na predição de não evadidos (VN).

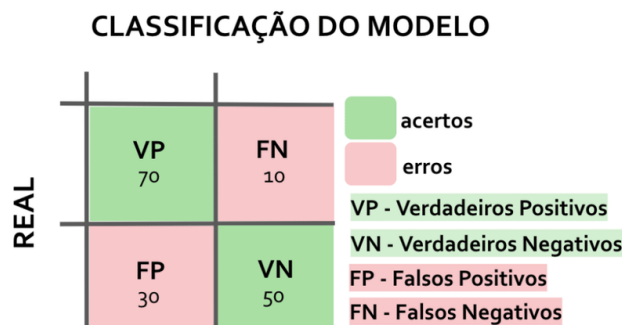


Figura 5 – Exemplo de matriz de confusão. Fonte: (RUBACK; AVILA; CANTERO, 2021)

2.2.5.1 Acurácia

Com base na matriz de confusão, diversas métricas são utilizadas para mensurar diferentes aspectos do modelo. O desempenho geral de um modelo de predição pode ser avaliado por sua exatidão, também conhecida como acurácia geral, que é calculada pela quantidade de classificações corretas dividida pelo número total de casos na amostra de teste. Sua equação, demonstrada abaixo, resultará em valores entre 0 e 1, onde 1 representa 100% de acurácia (todas as predições estão corretas) e 0 representa que nenhuma predição é correta. Comumente esse resultado é multiplicado por 100 para representação em formato de percentual.

$$\text{Acurácia} = \frac{\text{Número de predições corretas}}{\text{Total de predições}} = \frac{VP + VN}{VP + FN + VN + FP} \quad (2.1)$$

Por sua simplicidade, a Acurácia (ACC) é frequentemente mencionada e compreendida mesmo por aqueles que não são especialistas em aprendizado de máquina.

2.2.5.2 MCC

O Coeficiente de Correlação de Matthews é uma medida estatística usada para avaliar a precisão das previsões em modelos de classificação, levando em conta os quatro quadrantes da matriz de confusão: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN). Essa característica torna o MCC mais robusto em cenários de conjuntos de dados desbalanceados (CASANOVA, 2011). Ao considerar todos os quadrantes da matriz de confusão, o MCC oferece uma avaliação mais completa do desempenho dos modelos, especialmente em situações em que a minimização dos falsos negativos é crítica. Sua fórmula está representada na equação abaixo:

$$\text{MCC} = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VN + FN)}}. \quad (2.2)$$

Essa equação resultará em valores entre -1 e 1, onde: a proximidade de -1, indica uma predição inversa (o modelo está fazendo previsões incorretas na maioria das vezes); a proximidade de 0, indica que o modelo está fazendo previsões similares a uma escolha aleatória, sem habilidade discriminativa; e a proximidade de 1 indica uma predição muito boa, onde o modelo está fazendo previsões corretas na maioria das vezes.

Além dessas métricas, existem outras medidas de desempenho comumente utilizadas em modelos de classificação. A *Precisão*, por exemplo, avalia a proporção de verdadeiros positivos em relação a todas as previsões positivas. O *Recall*, por sua vez, reflete a proporção de verdadeiros positivos em relação às suas previsões positivas e negativas incorretas. A *Área Sob a Curva ROC (AUC)* representa a sensibilidade do classificador, demonstrando a taxa de verdadeiros positivos em relação à taxa de falsos positivos. Quanto maior a AUC, mais preciso é o classificador. O *F1-Score* é uma medida que combina precisão e recall, fornecendo um equilíbrio entre as duas métricas. O *Índice Kappa* é um coeficiente de concordância utilizado para escalas nominais, medindo a relação entre a concordância e a discordância esperada. Embora essas métricas não sejam utilizadas neste estudo, são importantes ferramentas para avaliar o desempenho de modelos de classificação em diferentes contextos.

Escolhidas as métricas e aferidos os resultados, os mesmos podem ser visualizados de maneira intuitiva por meio de gráficos, mapas ou outras representações visuais, facilitando a compreensão e a comunicação das descobertas para diferentes públicos. Eles devem ser interpretados à luz do domínio da aplicação e as descobertas podem impulsionar ações práticas e decisões.

2.3 Aprendizado de Máquina

A necessidade de analisar grandes volumes de dados de maneira automatizada é abordada por meio do uso de técnicas de Aprendizado de Máquina (AM), também conhecida como Machine Learning. As técnicas de AM capacitam programas de computador a ajustarem seu comportamento sem depender de programação explícita. Em vez disso, eles se baseiam em resultados anteriores, permitindo flexibilidade na tomada de decisões, sem a necessidade de seguir regras e instruções predefinidas, como destacado por [Oliveira, Medeiros e Andrade \(2022\)](#).

Conforme observado por [Jesus, Rodriguez e Junior \(2021\)](#), o Aprendizado de Máquina é responsável por gerar novos conhecimentos a partir dos dados contidos em uma base de dados, enquanto a Mineração de Dados extrai padrões desses dados usando técnicas de Aprendizado de Máquina ([WITTEN; FRANK, 2002](#)). Ou seja, o Aprendizado de Máquina se concentra em criar modelos preditivos capazes de aprender a partir dos dados, com o objetivo de fazer previsões ou tomar decisões com base em novas informações, enquanto a Mineração de Dados emprega técnicas do Aprendizado de Máquina para identificar padrões ocultos e informações úteis dentro de conjuntos de dados extensos. Ambas são fundamentais e trabalham em conjunto para a análise de dados.

Durante o processo de mineração, diversos algoritmos, técnicas e ferramentas de aprendizado de máquina podem ser aplicados para identificar padrões. Entre essas técnicas, destacam-se os Classificadores. [Rodrigues et al. \(2014\)](#) apontaram que os métodos de Classificação, Associação e Agrupamento eram os mais comuns nas pesquisas de MDE à época. A pesquisa também revelou que a técnica de classificação predominava significativamente, representando 35,13% das técnicas utilizadas nos trabalhos de MDE publicados no Brasil desde 2006. Além disso, a Detecção ou Predição de Evasão foi identificada como um dos principais objetivos educacionais abordados.

O Aprendizado de Máquina é uma vasta área composta por várias abordagens e técnicas de aprendizado. A Figura 6 ilustra a trajetória adotada em nosso projeto dentre algumas das abordagens disponíveis, bem como algumas técnicas que merecem menção. Dentre as diversas estratégias e paradigmas de aprendizado, destacam-se ([ARAÚJO, 2014](#)):

- **Aprendizado por Hábito:** onde, todo conhecimento é simplesmente transmitido do instrutor para o aprendiz, sem que o aprendiz faça inferências sobre as informações fornecidas, sem uma compreensão profunda ou análise crítica, apenas memorizando-as.
- **Aprendizado por Instrução:** onde o aprendiz adquire conceitos de suas fontes, mas não armazena diretamente a informação, passando-a por um processo de identificação dos conceitos mais importantes e relevantes da informação recebida e adaptando-a

para uma forma que seja mais adequada e compreensível para ele.

- **Aprendizado por Dedução:** que parte de premissas gerais para inferir conclusões específicas. O aprendiz forma conceitos através da aplicação de raciocínio lógico sobre o conhecimento prévio. Isso envolve a transformação e interpretação do conhecimento existente para chegar a novas conclusões ou entender novos conceitos.
- **Aprendizado por Analogia:** onde o aprendiz adapta conceitos prévios para entender novos conceitos. Em vez de criar novas regras, ele ajusta regras existentes para descrever o novo conhecimento.
- **Aprendizado por Indução:** onde o raciocínio parte de exemplos específicos para chegar a generalizações sobre um conceito. Essa abordagem envolve inferências lógicas que buscam extrair padrões dos dados de treinamento para induzir um conceito mais amplo, embora essas generalizações possam não ser sempre verdadeiras em todos os casos.

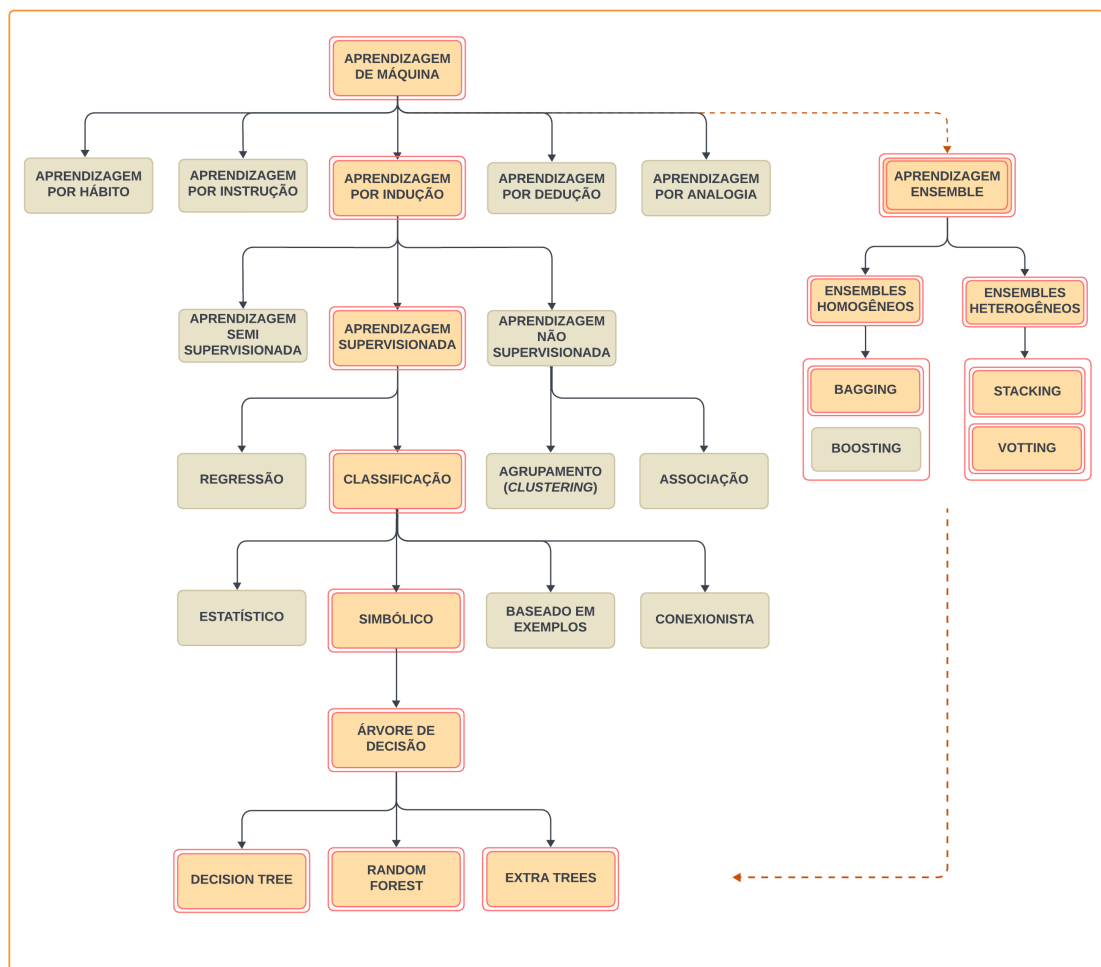


Figura 6 – Aprendizado de Máquina. Fonte: Adaptado de Araujo (2014)

O **Aprendizado Ensemble** não é necessariamente um tipo de aprendizado distinto, e sim uma técnica que pode ser aplicada em conjunto com qualquer um dos tipos mencionados acima. Ela realiza a combinação de múltiplos modelos para melhorar o desempenho geral do sistema e será abordada mais adiante.

2.3.1 Aprendizado Supervisionado

O Aprendizado por Indução presente neste estudo pode ser subdividido em diversas estratégias de aprendizado, tais como: Aprendizado Supervisionado, em que o modelo é treinado em um conjunto de dados rotulados; Aprendizado Semi Supervisionado, que envolve a utilização de um pequeno conjunto de dados rotulados juntamente com um grande conjunto de dados não rotulados; e Aprendizado Não Supervisionado, que se concentra na identificação de padrões intrínsecos nos dados não rotulados.

2.3.1.1 Técnica de Classificação

Especificamente, o Aprendizado Supervisionado engloba essencialmente as técnicas de Classificação e Regressão. Neste estudo, utilizamos a técnica de Classificação, cujo objetivo é induzir um classificador (ou hipótese), por meio de conjuntos de dados previamente rotulados, para classificar novos exemplos ainda não rotulados (SILVA et al., 2017). Na técnica de Classificação os rótulos das classes possuem valores discretos (como 0 ou 1, True ou False). Por outro lado, na técnica de Regressão, os valores dos rótulos são contínuos (como peso e altura de uma pessoa por exemplo), sendo essa a principal diferença entre as duas técnicas.

Aliado à estratégia de Classificação, utilizamos o paradigma Simbólico, que foca na interpretação explícita de regras e símbolos, em contraste com o paradigma Estatístico, que utiliza modelos probabilísticos para realizar inferências; o Baseado em Exemplos, que generaliza a partir de exemplos específicos; e o Conexcionista, que se baseia em redes neurais artificiais para modelar relações complexas.

Por meio do Aprendizado Supervisionado, este trabalho buscou a predição de classes com base no conhecimento adquirido a partir de conjuntos de dados previamente rotulados. No nosso caso temos um rótulo binário, representado pelo atributo 'Evadido', que determina se um discente evadiu/evadirá (True) ou concluiu/concluirá (False) seu curso. Para abordar nossos dados e o tipo de rótulo em questão, empregamos métodos de Classificação, com a criação de Classificadores baseados no Decision Tree (DT); posteriormente, expandimos os experimentos com o uso de Aprendizado Ensemble e outros algoritmos, como veremos mais adiante.

2.3.2 Algoritmos

Os paradigmas de aprendizado supervisionado são abordagens metodológicas distintas para tratar de questões de aprendizagem cujos dados de treinamento estão rotulados. A família de algoritmos de Árvore de Decisão (AD) é a principal representante do paradigma Simbólico.

A Figura 7 traz um exemplo didático do funcionamento desses algoritmos. Cada nó interno da árvore corresponde a uma decisão tomada baseada em um atributo (na imagem representado por um questionamento), enquanto os nós folha representam as classes ou previsões (recusar ou aceitar a oferta). Com base nas características dos dados, novas tomadas de decisões vão sendo tomadas ao longo da estrutura da árvore.

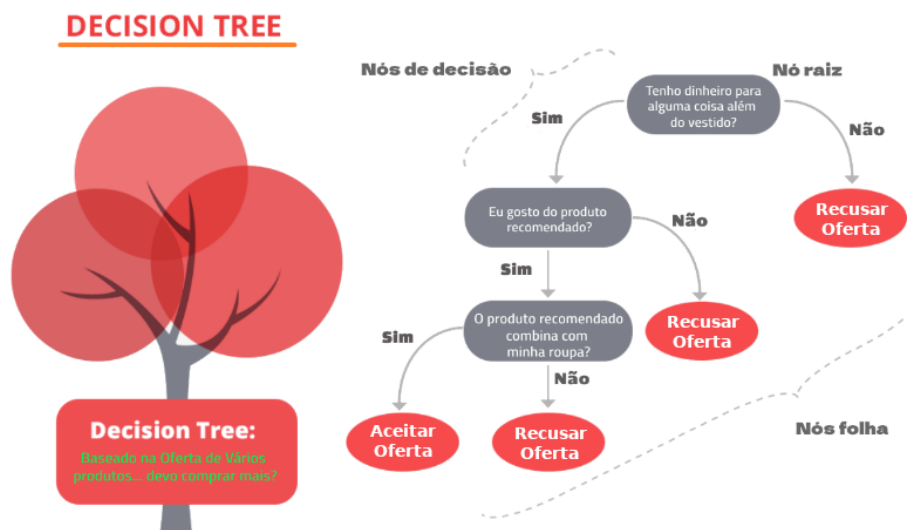


Figura 7 – Árvore de decisão traduzida e adaptada a partir da fonte: (SACRAMENTO, 2021)

As Árvores de Decisão são facilmente interpretáveis e podem ser visualizadas de forma intuitiva, potencializando a explicabilidade. Elas não exigem muita preparação de dados em comparação com alguns outros algoritmos de aprendizado de máquina. Como desvantagem, podem ser propensas ao overfitting, especialmente em conjuntos de dados complexos ou com muitos atributos, exigindo contramedidas. Pertencem a essa família os três algoritmos utilizados ao longo de nosso trabalho, que serão apresentados a seguir.

2.3.2.1 Algoritmo Decision Tree

As árvores de decisão são ferramentas importantes e amplamente adotadas para tarefas de classificação e predição. Elas representam conjuntos de regras que podem ser compreendidas por humanos e implementadas em sistemas de conhecimento, como bancos de dados (PRAJWALA, 2015).

O algoritmo Decision Tree (DT) é um classificador não paramétrico, o que significa que é altamente flexível em relação aos dados. Ele realiza divisões contínuas nos dados com base em regras de decisão simples. A escolha dos atributos a serem usados na divisão dos dados em cada nó é feita com o objetivo de maximizar o ganho de informação, como descrito por [Dib et al. \(2020\)](#).

Em termos simples, a classificação é realizada com base nas características que melhor separam os dados. Os itens de dados são agrupados com base nos valores dessas características, e esse processo é repetido recursivamente em cada subconjunto. A recursão é encerrada quando todos os itens de dados no subconjunto atual pertencem à mesma classe ([ALI et al., 2012](#)), ou quando um critério de parada é acionado para evitar o overfitting.

O DT é reconhecido por sua simplicidade, ampla utilização e bom desempenho ([SANTOS; SARAIVA; OLIVEIRA, 2021](#)). Ele está presente no Scikit-Learn (biblioteca de aprendizado de máquina em Python que utilizamos neste projeto), sendo implementado por meio da classe `DecisionTreeClassifier`. Essa classe também serve como a espinha dorsal na construção das árvores de decisão empregadas nos algoritmos Random Forest e Extra-Trees, exemplos de abordagens ensemble que também utilizamos no EC3.

2.3.2.2 Algoritmo Random Forest

O Random Forest (RF) é uma variante específica da técnica Ensemble Bagging (que veremos mais adiante) e uma extensão do algoritmo Decision Tree. Enquanto o DT cria uma única árvore de decisão, o RF constrói uma coleção (ou floresta) de árvores de decisão. Cada árvore é treinada em uma amostra aleatória dos dados e, durante a classificação, a previsão final é obtida por agregação (voto majoritário para problemas de classificação ou pela média para problemas de regressão) das previsões de todas as árvores ([ALI et al., 2012](#)). Esse processo reduz a variância do modelo e melhora sua generalização para dados não vistos.

A principal ideia por trás do RF é introduzir aleatoriedade no processo de construção da árvore, tanto na escolha (aleatória) dos atributos quanto nas amostras dos dados. Isso ajuda a evitar o overfitting e aumenta a estabilidade e robustez do modelo. O RF também permite medir a importância de cada atributo na classificação, fornecendo uma visão mais detalhada sobre a relevância de cada característica para o modelo final.

Por outro lado, uma desvantagem do RF é que o tempo de execução pode ser maior em comparação com modelos que utilizam apenas uma árvore de decisão. Ele está implementado no Scikit-Learn por meio da classe `RandomForestClassifier`, e utiliza como estimador-base o DT. A classe oferece uma série de parâmetros que permitem personalização para, por exemplo, reduzir o tempo de execução. Entre eles a quantidade de árvores na floresta (`n_estimators`), a profundidade máxima de cada árvore (`max_depth`) e a quantidade mínima de amostras necessárias para dividir um nó (`min_samples_split`).

2.3.2.3 Algoritmo Extra Trees

O algoritmo Extra Trees (ET), ou Extremely Randomized Trees, é outra variante da técnica Ensemble Bagging e uma alternativa ao Random Forest. Assim como o RF, o ET é um método ensemble que constrói uma coleção de árvores de decisão. No entanto, o ET introduz uma camada adicional de aleatoriedade durante o processo de divisão de nós nas árvores.

Enquanto o RF seleciona a melhor divisão entre um subconjunto aleatório de atributos, o ET vai além, optando por escolher seus limiares de divisão de forma totalmente aleatória (VIANA; SANTANA; RABÊLO, 2022). Essa abordagem torna o ET ainda mais robusto ao overfitting e mais eficiente computacionalmente. Ao dispensar a busca pelo melhor limiar de divisão, ele se torna significativamente mais rápido para treinar em comparação com o RF. O ET é particularmente eficaz em cenários com dados ruidosos ou conjuntos de dados com muitos atributos.

O Scikit-Learn implementa o algoritmo ET por meio da classe `ExtraTreesClassifier`, que disponibiliza parâmetros similares aos da classe `RandomForestClassifier` para personalização e ajustes visando melhores resultados. Embora normalmente mais eficiente que o RF, seu tempo de execução também é relativamente elevado e pode ser ajustado por meio desses parâmetros.

2.3.3 Aprendizado Ensemble (Ensemble Learning)

Para além do uso da técnica de Aprendizado por Indução, uma outra importante área foi principalmente explorada em nosso trabalho: o Aprendizado Ensemble (ou Ensemble Learning). Foram exploradas estratégias que utilizam técnicas de Ensemble, ambas empregando inicialmente o algoritmo DT como estimador-base, e posteriormente Extra-Trees (ET) e Random Forest (RF).

Métodos Ensemble, também conhecidos como técnicas de combinação de classificadores, desempenham um papel essencial na melhoria da acurácia da classificação. Eles operam como um comitê de especialistas que unem os resultados de diversos modelos preditivos para alcançar um desempenho superior, conforme Piao et al. (2015) salienta.

Dietterich (2000) elenca três principais deficiências dos algoritmos de aprendizado padrão, às quais segundo ele os métodos ensemble têm a promessa de reduzir (ou até eliminar):

- **Estatística:** a questão estatística surge quando há poucos dados de treinamento em relação ao espaço de hipóteses, levando a múltiplas hipóteses igualmente precisas. Um ensemble com diferentes classificadores pode realizar uma agregação de seus resultados e reduzir o risco de escolher o classificador errado.

- **Computacional:** problemas computacionais surgem quando os algoritmos ficam presos em ótimos locais durante a busca local. Um ensemble, formado por várias buscas locais a partir de pontos diferentes, pode superar essa limitação.
- **Representacional:** a limitação representacional ocorre quando o espaço de hipóteses não pode representar a função verdadeira. Os ensembles podem expandir esse espaço, combinando hipóteses ponderadas para representar funções mais complexas e abrangentes.

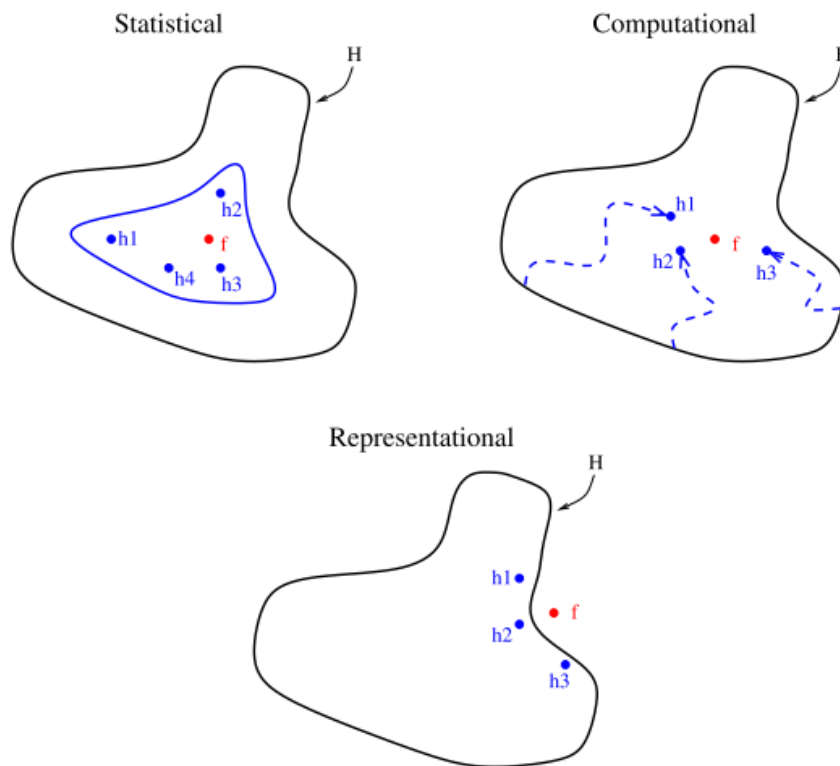


Figura 8 – Três razões fundamentais pelas quais um ensemble pode trabalhar melhor que um classificador individual. Fonte: (DIETTERICH, 2000)

Na Figura 8 a curva externa denota o espaço de hipóteses H . A curva interna (na imagem do problema estatístico) denota o conjunto de hipóteses que fornecem boa precisão nos dados de treinamento. O ponto rotulado f é a verdadeira hipótese, e podemos ver que, ao fazer a média das hipóteses precisas ($h1, h2, h3$ e $h4$), podemos encontrar uma boa aproximação para f .

O processo de Ensemble é dividido em duas fases distintas: Na primeira fase, ocorre a criação de um conjunto de classificadores-base a partir dos dados de treinamento. Na segunda fase, a classificação efetiva é executada pela combinação dos resultados dos classificadores base. Para garantir a precisão do conjunto combinado de múltiplos classificadores, é importante que cada classificador base seja diversificado e independente, como ressaltado por Piao et al. (2015), e que esses componentes individuais forneçam

informações complementares sobre um padrão de entrada, com o objetivo de aprimorar a eficácia do processo de reconhecimento, como argumentado por [Santana e Canuto \(2014\)](#).

Ao encarar a predição de cada classificador individual como uma hipótese independente ([DIETTERICH, 2000](#)), a abordagem dos métodos Ensemble consiste em empregar vários modelos para combinar essas hipóteses em um resultado unificado destinado a resolver tarefas específicas de aprendizado de máquina. A premissa central é que ponderar e agregar múltiplas opiniões individuais (hipóteses) será mais eficaz do que depender de um único classificador, conforme enfatizado por [Sagi e Rokach \(2018\)](#).

Essa abordagem é exemplificada na Figura 9, que ilustra três hipóteses (predições) provenientes de três classificadores individuais em (a), e o resultado combinado pelo ensemble, proporcionando uma predição mais precisa para cada rótulo/valor real em (b). Cada classificador individual tem uma habilidade distinta e classifica os 'emojis felizes' de um mesmo conjunto de dados à sua maneira, cometendo eventuais erros associados à sua forma de classificar. Esses erros podem ser corrigidos por uma forma de classificação diferente de outros modelos, quando combinados em um modelo ensemble. Por exemplo, os dois 'emojis' do canto superior direito seriam classificados incorretamente como 'felizes' pelo classificador individual ao centro da figura, mas os demais classificadores identificaram-nos como 'emojis tristes' e suas hipóteses se sobressaíram no modelo combinado.

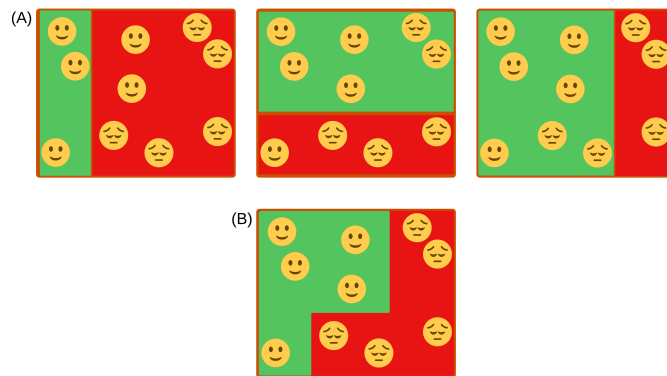


Figura 9 – Exemplo de classificadores independentes (A) combinados por ensemble (B).
Fonte: Adaptado de [Sagi e Rokach \(2018\)](#)

2.3.3.1 Diversidade dos Classificadores-base

O treinamento dos classificadores-base em um ensemble tem um objetivo claro: produzir classificadores que sejam precisos e que, ao mesmo tempo, diferem entre si em relação aos erros cometidos em padrões idênticos. Esse conceito é conhecido como Diversidade ([RAHMAN; VERMA, 2013](#)).

A combinação de classificadores tem duas principais categorias: ensembles homogêneos e ensembles heterogêneos ([ALSHDAIFAT; AL-HASSAN; ALOQAILY, 2021](#)). A figura 10 apresenta algumas características dessas categorias.

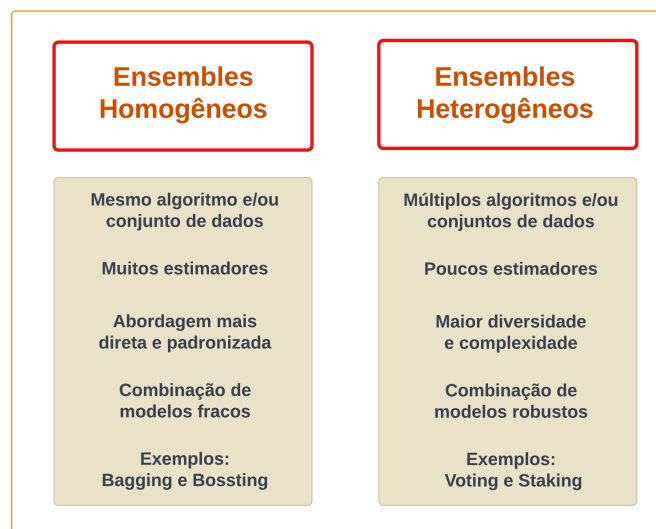


Figura 10 – Características dos Ensembles Homogêneos e Heterogêneos.

Os ensembles homogêneos são caracterizados pela construção de um grande número de classificadores fracos, que combinados promovem uma espécie de "sabedoria das multidões". Ou seja, através da opinião (predição) de muitos indivíduos (classificadores), a coletividade poderá chegar a decisões mais confiáveis e precisas do que um único indivíduo. Eles geralmente utilizam um mesmo algoritmo e um mesmo conjunto de dados inicial, que em geral é particionado por amostragem reduzindo a robustez dos modelos, porém incrementando a diversidade pelos 'pontos de vista' distintos criados.

Por outro lado, os ensembles heterogêneos tendem a ser mais complexos, combinando classificadores robustos com o objetivo de maximizar a precisão das predições. Segundo [Silva et al. \(2019\)](#), eles devem constituir modelos com diversidade, permitindo a manipulação do processo de modelagem por meio de diferentes conjuntos de dados, técnicas ou parâmetros. A diversidade de dados permite a geração de muitos conjuntos de dados a partir do conjunto de dados original para treinar diferentes preditores. Esses devem ser diferentes entre si para que haja várias decisões a partir dos resultados dos preditores treinados.

Já a construção de classificadores-base nos métodos Ensemble pode ser subdividida em quatro grupos principais, conforme [Rahman e Verma \(2013\)](#): (1) Selecionar diferentes subconjuntos de instâncias do conjunto de treinamento para construir cada classificador-base; (2) Escolher diferentes subconjuntos de recursos dentre os recursos de entrada disponíveis para construir cada classificador-base; (3) Basear-se em diferentes categorias dos rótulos de classe para construir cada classificador-base; (4) Manipular o algoritmo de aprendizagem.

É importante destacar que a seleção dos dados que são fornecidos aos classificadores frequentemente têm um impacto maior nos ganhos obtidos com a combinação dos classificadores do que a escolha do método de combinação, como evidenciado por [Tumer](#)

e Ghosh (1996). Deste modo, nosso trabalho estará focado em explorar a diversidade advinda dos dados, tanto através da seleção de subconjuntos de instâncias quanto no uso de conjuntos (e não subconjuntos) de recursos de entrada distintos para a construção de cada classificador. Essas abordagens serão exploradas nos estudos de caso 1 e 2 deste trabalho, respectivamente.

Conforme indicado por Fletcher, Verma e Zhang (2020), a diversidade nos dados tem sido o foco de estudos que se concentram no uso de vários algoritmos de classificação para treinar os classificadores-base. Alguns estudos realizam comparações explícitas entre os classificadores base para avaliar suas semelhanças e diferenças. Além disso, a diversidade pode ser obtida através da variação do espaço de atributos ou características, onde cada classificador é treinado com um subconjunto distinto de características entre as disponíveis, ampliando a diversidade do que cada classificador aprende e prevê. A seleção de características pode ser realizada de maneira aleatória, similar ao método de Random Forest, ou de maneira inteligente, baseada na utilidade ou diversidade de cada característica.

Quando se trata do design de classificadores individuais para compor sistemas Ensemble, Santana e Canuto (2014) ressaltam que a seleção de subconjuntos de recursos é considerada um problema de busca em alguns estudos, envolvendo abordagens exaustivas e heurísticas, como o uso de Algoritmos Genéticos, de Otimização por Colônias de Formiga (ACO) e de Enxame de Partículas (PSO) para a seleção de atributos.

Entretanto, nossa abordagem não se trata apenas da seleção de um subconjunto de recursos. Ela abrange a criação de modelos completos, distintos e eficazes, a partir de datasets com dados estruturalmente distintos. Cada tipo de estrutura corresponde a um ‘nível’, sendo a existência de características temporais em um deles a diferença estrutural mais significativa. Esses níveis de dataset já foram experimentados em outros estudos de análise de predição de evasão (VIANA; SANTANA; RABÊLO, 2022)(BITENCOURT; SILVA; XAVIER, 2021)(JESUS; RODRIGUEZ; JUNIOR, 2021)(PAZ; CAZELLA, 2017)(SANTOS et al., 2019) com resultados satisfatórios em seus casos específicos isolados.

Criamos modelos em três níveis de organização/estruturação dos dados, a partir da perspectiva de uma IFES específica. Sobre os modelos cujo nível se mostrou com melhores resultados de predição, foram aplicadas técnicas de Ensemble individualmente, formando novos modelos (EC1). Adicionalmente, combinamos modelos dos três níveis em um único modelo agregado visando maximizar a diversidade dos modelos finais (EC2). E, por fim, incorporamos outros algoritmos, métodos e conjuntos de atributos (EC3).

2.3.3.2 Desafios para os Métodos Ensemble

Os métodos ensemble apresentam uma abordagem para superar os desafios técnicos enfrentados na construção de um único estimador, como delineado por Alhamid (2021):

- Alta variância: Este desafio se manifesta quando o modelo demonstra alta sensibilidade às entradas dos recursos de treinamento.
- Baixa precisão: Há situações em que um único modelo ou algoritmo, ajustando-se a todos os dados de treinamento, pode não ser suficientemente preciso para atender às expectativas.
- Presença de ruído e tendência (Viés): Alguns modelos dependem excessivamente de um ou alguns recursos ao realizar previsões, o que pode resultar em previsões tendenciosas.

Vale ressaltar que a alta variância está intrinsecamente ligada ao fenômeno de Overfitting, um ponto que merece destaque. O overfitting é um problema crítico na área de aprendizado de máquina (CHUNG; LEE, 2019) e ocorre quando um modelo se ajusta em excesso aos dados de treinamento, capturando ruídos nos dados e resultando em alta variância. Isso, por sua vez, leva a um desempenho insatisfatório ao lidar com conjuntos de dados futuros, ou seja, uma baixa capacidade de generalização.

Felizmente, várias técnicas bem estabelecidas foram desenvolvidas para criar classificadores ensemble, algumas das quais se destacam como soluções eficazes para os desafios mencionados anteriormente. Entre essas técnicas, destacam-se (YANG et al., 2010):

- Bagging: Essa técnica visa reduzir a variância do modelo.
- Boosting: O foco do Boosting é reduzir o viés do modelo.
- Stacking: Esta abordagem visa maximizar a generalização do modelo.
- Técnicas baseadas em votação: Estas técnicas buscam melhorar os resultados sem se concentrar em uma causa específica.

2.3.3.3 Técnica Ensemble Bagging

Os métodos (ou técnicas) ensemble do tipo Bagging (Bootstrap Aggregating) trazem benefícios tanto para Regressão quanto para Classificação. São frequentemente empregados em conjunto com árvores de decisão e têm um impacto significativo na estabilidade dos modelos, reduzindo a variância, o que, por sua vez, aprimora a precisão e elimina o overfitting (YANG et al., 2010). Essa técnica é um exemplo da categoria de ensembles homogêneos, onde todos os classificadores-base que formam o ensemble são produzidos usando um mesmo algoritmo. (ALSHDAIFAT; AL-HASSAN; ALOQAILY, 2021)

Bagging é uma abordagem eficaz que cria vários modelos independentes por meio da técnica de amostragem com reposição, chamada Bootstrap. Essa técnica gera diversas amostras a partir dos dados originais, sendo que cada uma contém instâncias selecionadas

aleatoriamente, com a possibilidade de que algumas instâncias apareçam em múltiplas amostras devido à reposição. Em problemas de Classificação, a predição final para uma instância ausente nos dados de treinamento é obtida por meio da votação majoritária das predições dos modelos (PIAO et al., 2015). A figura 11 ilustra essa técnica.

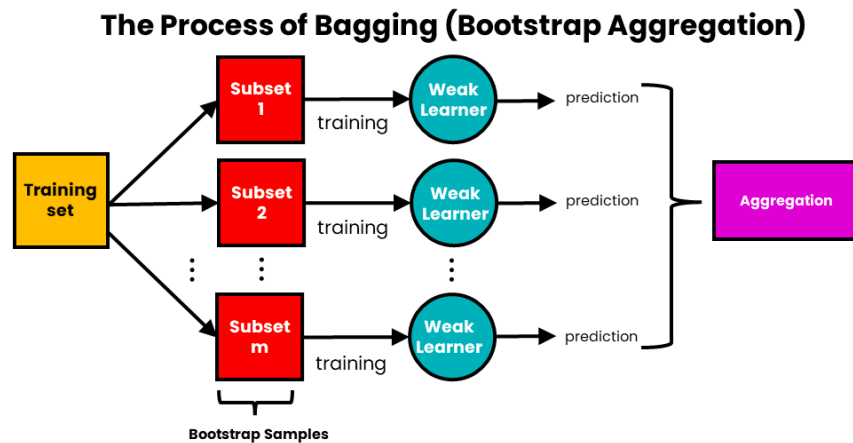


Figura 11 – O Processo de Ensemble Bagging. Fonte: (KALIRANE, 2024)

2.3.3.4 Técnica Ensemble Boosting

Os métodos Boosting referem-se a outra técnica ensemble amplamente utilizada, com o objetivo de aprimorar o desempenho de algoritmos de aprendizado que geram diversos classificadores e combinam suas predições. Ao contrário do Bagging, o Boosting atribui pesos a cada instância de treinamento e ajusta esses pesos ao final de cada rodada de Boosting (PIAO et al., 2015). Inicialmente, são criados modelos fracos que são treinados com o conjunto de dados em várias rodadas, com informações dos modelos anteriores influenciando o aprendizado dos modelos futuros por meio dos pesos. No final, as predições de todos os modelos são combinadas para obter a predição final (SANTOS, 2023). Os métodos baseados em Boosting aprimoram a generalização, mas, em vez de atuar na variância, tendem a diminuir o viés (YANG et al., 2010). A figura 12 ilustra essa técnica.

2.3.3.5 Técnica Ensemble Stacking

A abordagem de Stacking é uma técnica abrangente que utiliza pelo menos duas etapas de classificação e combina os classificadores de base com um meta-classificador para otimizar a generalização (YANG et al., 2010)(AGGARWAL et al., 2015). Na primeira fase, cada classificador de base é treinado e, em seguida, suas predições são usadas como características de entrada para treinar um meta-classificador (meta-modelo, ou ainda meta-learner). Em vez de fazer predições diretamente com base nos dados de entrada, o meta-classificador utiliza as predições individuais dos classificadores de base como características de entrada e, em seguida, produz uma predição final. Essa abordagem

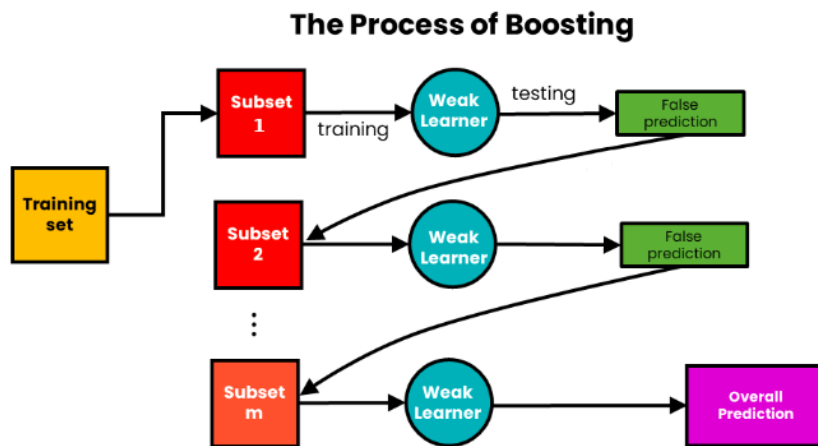


Figura 12 – O Processo de Ensemble Boosting. Fonte: (KALIRANE, 2024)

visa reduzir a variância do erro, tornando o meta-classificador menos suscetível a erros e, portanto, com melhor capacidade de generalização no conjunto de testes (DIB et al., 2020). A figura 13 ilustra essa técnica.

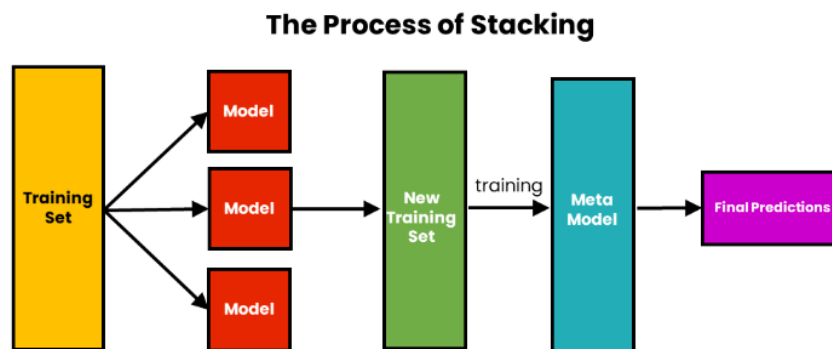


Figura 13 – O Processo de Ensemble Stacking. Fonte: (KALIRANE, 2024)

2.3.3.6 Técnica Ensemble Voting

Os métodos baseados em votação são métodos que recebem as previsões de seus estimadores-base como uma espécie de voto, visando “eleger” a previsão mais adequada. Dentro dos métodos baseados em votação, destaca-se o amplamente utilizado Majority Voting (Votação Majoritária) (ALSHDAIFAT; AL-HASSAN; ALOQAILY, 2021). Nesse método, a classe mais frequentemente predita pelos classificadores de base é escolhida como a previsão final na segunda fase. Espera-se que esse modelo alcance maior precisão do que os melhores modelos individuais, pois corrige erros não correlacionados por meio da votação. Além disso, há variações, como a Votação Ponderada, na qual os modelos de classificação

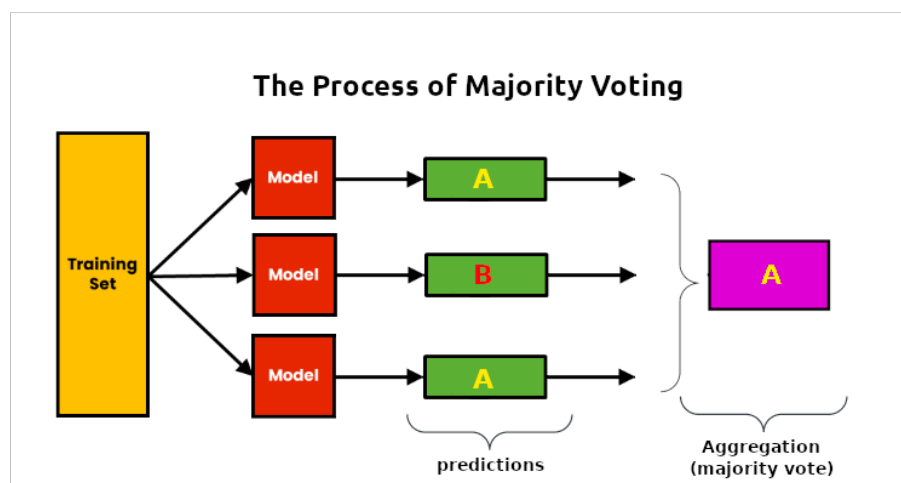


Figura 14 – O Processo de Ensemble por Votação Majoritária

recebem pesos diferentes, associados à sua precisão de classificação (TSOUMAKAS; ANGELIS; VLAHAVAS, 2005). Em resumo, cada modelo emite um voto para classificar uma observação, mas, diferentemente do Majority Voting, esses votos são ponderados (ALSHDAIFAT; AL-HASSAN; ALOQAILY, 2021). A figura 14 ilustra essa técnica.

2.4 Mineração de Dados Educacionais

A Mineração de Dados desempenha um papel central no processo de DCBD (Descoberta de Conhecimento em Base de Dados), conforme definido por Han, Kamber e Pei (2012). Essa etapa é essencial, pois é nela que algoritmos são aplicados para extrair padrões e conhecimento valioso a partir de grandes volumes de dados, que podem ser provenientes de diversos tipos de fontes, como bancos de dados, data warehouses, a Web e outros repositórios de informações. Transformar dados brutos em informações relevantes e acionáveis é o cerne desse processo.

No contexto da pesquisa atual, a Mineração de Dados se mostra uma abordagem muito importante para a tomada de decisões e a previsão de tendências futuras em várias áreas, como vendas, bioinformática e ações contra o terrorismo, como destacado por Baker, Isotani e Carvalho (2011). Uma investigação breve sobre estudos relacionados à Educação que aplicam a Mineração de Dados revela a importância crescente desse campo.

Nesse sentido, a Mineração de Dados Educacionais (MDE) envolve a aplicação de técnicas de Mineração de Dados a conjuntos de dados provenientes de ambientes educacionais diversos. Essa área de pesquisa se concentra em tópicos como predição, agrupamento, mineração de relações, desenvolvimento de modelos e tratamento de dados para apoio à tomada de decisões, como mencionado por Rigo et al. (2014). Independentemente do foco, essas abordagens podem ser aplicadas de maneira variada às questões relacionadas à evasão escolar e à aquisição de modelos para a antecipação de diagnósticos.

Baker, Isotani e Carvalho (2011) definem ainda que a MDE é uma área de pesquisa focada no desenvolvimento de métodos para explorar dados coletados em ambientes educacionais. Técnicas tradicionais de mineração de dados têm sido aplicadas com sucesso nesse domínio, mas, devido às características únicas dos sistemas educacionais, há uma necessidade de abordagens personalizadas para lidar com dados relacionados a alunos.

2.4.1 Trabalhos Relacionados

À medida que a MDE se torna uma área emergente, novas pesquisas contribuem para aprimorar as técnicas adotadas e desenvolver abordagens personalizadas, como destacado por Romero e Ventura (2020). Nesta seção exploraremos trabalhos relacionados e pesquisas anteriores que se dedicaram a investigar esse fenômeno complexo, utilizando abordagens baseadas em MDE, e que influenciaram nossa pesquisa.

Em um desses trabalhos, Teodoro e Kappel (2020) utilizaram uma base de dados pública disponibilizada pelo INEP para investigar quais as variáveis mais importantes para a previsão de evasão em Instituições Públicas de Ensino Superior do Brasil. Foram utilizados cinco algoritmos de Classificação na construção dos Modelos, com destaque para os resultados obtidos pelo modelo que utilizou Random Forest (Uma abordagem Ensemble de Árvores de Decisão). Os modelos abordaram uma base de dados única com dados de estudantes de Instituições de todo o país na busca das variáveis mais importantes para a previsão de evasão, sendo a idade, a participação em atividades extracurriculares e a carga horária total do curso as principais características influentes apontadas.

Outro artigo Jesus, Rodriguez e Junior (2021) apresenta um estudo sobre MDE, cujo objetivo é a previsão de alunos com risco de evasão em um curso de Licenciatura em Computação de uma universidade pública. Esta pesquisa, dentre outras questões, utilizou um modelo preditivo composto de rede neural de múltiplas camadas e uma base de dados contendo os históricos acadêmicos de 136 alunos do curso com matrículas ativas na instituição até o primeiro semestre letivo de 2019. Nele atingiu-se acurácia de cerca de 98% em relação a determinadas situações cujo artigo considera como Risco de Evasão.

Em um nível que envolveu dados de estudantes de vários cursos de uma mesma IFES, (MELO; SOUZA, 2023) realizaram um trabalho que envolveu aprendizado semi-supervisionado como suporte à predição da evasão. Utilizando um dataset com 12.657 instâncias, o aprendizado supervisionado auxiliou na rotulação dos 4.646 registros de discentes que se encontravam como “matriculados”, situação esta que pode se referir a estudantes que estão de fato cursando as disciplinas, ou que na prática já evadiram ou se formaram mas oficialmente ainda estão vinculados. Com um rótulo mais adequado, os autores chegaram a uma melhoria na predição utilizando aprendizado supervisionado com Random Forest.

Em outro nível de análise de dados, [Carneiro et al. \(2022\)](#) buscaram identificar e prevenir a retenção e a evasão dos alunos através de seus desempenhos em um componente do primeiro período do curso de Bacharelado em Sistemas de Informação de uma IFES brasileira. Este componente possuía um altíssimo índice de 78% de reprovação. A investigação se baseou em um dataset de 98 estudantes matriculados na disciplina entre 2018 e 2019, e com base nos resultados sobre estudantes propensos a evadir-se, ações de combate foram tomadas, reduzindo substancialmente os índices de evasão nos anos em questão em relação aos anos anteriores (2013 a 2017). As características ‘distância entre a moradia e a casa dos pais’, ‘renda per capita da família’, ‘tipo de escola frequentada no ensino médio’, ‘média em matemática’ e ‘distância de casa até a universidade’ foram os fatores-chave apontados para a identificação dos estudantes em risco.

Na análise de 22 trabalhos nacionais e 28 trabalhos internacionais que abordam a área de mineração de dados educacionais com temática central voltada para o problema da evasão escolar, [Santos, Saraiva e Oliveira \(2021\)](#) apontam o Decision Tree (DT) como o algoritmo de aprendizado de máquina mais utilizado nos trabalhos nacionais, aparecendo em 13 dos 22 trabalhos, bem como também se destaca como um dos algoritmos com melhor desempenho em quatro trabalhos. No cenário nacional, o Support Vector Machines (SVM) e o DT aparecem como os algoritmos com melhores desempenhos. O estudo apontou ainda que, dentre os trabalhos publicados nacional e internacionalmente, a maioria das bases de dados avaliadas é composta por tipos de dados acadêmicos (43 trabalhos) e socioeconômicos (15 trabalhos).

[Silva et al. \(2019\)](#) se propuseram a realizar a predição de evasão em IFES através dos dados do Censo e dos indicadores de Fluxo da Educação disponibilizados pelo INEP e de três modelos de regressão ensemble baseados em métodos de bagging. As informações utilizadas tinham teor demográfico, acadêmico e socioeconômico. Os principais fatores associados ao processo de evasão no estudo foram ‘indicador de permanência acumulada’, ‘indicador de conclusão acumulada’, ‘Estudo no Período Noturno’ e ‘Número de estudantes que permanecem no curso’. Quanto aos demais achados, os modelos propostos apresentaram melhores resultados do que os da literatura abordada pelo artigo.

A aplicação de técnicas de Mineração de Dados para tentar identificar alunos com comportamento semelhante a alunos que se evadiram no passado também foi o foco do trabalho de [Santos, Martins e Plastino \(2021\)](#). Foram utilizadas técnicas de classificação por indução de árvores de decisão para gerar modelos capazes de prever se um aluno irá evadir ou irá se formar. Os dados utilizados foram basicamente os dados de desempenho dos alunos nas disciplinas de cada período do curso. O estudo utilizou abordagem de construção de modelos por período que o aluno se encontrava no curso. Os modelos foram capazes de atingir acurácia preditiva entre 79,31% e 98,25%.

Já o estudo de [Sales, Balby e Cajueiro \(2016\)](#) utilizou dados de 76 cursos de uma

IFES, gerando dois níveis de modelos para predição de evasão de alunos: um por período englobando todos os cursos e outro mais específico de períodos dentro de cada curso. Os modelos com todos os cursos obtiveram os melhores valores nas métricas Precision e Recall do que os modelos específicos, devido principalmente ao desbalanceamento das classes encontrado nas bases de dados usadas no modelo específico. Mesmo usando técnicas de balanceamento, observou-se que alguns semestres apresentam uma taxa de recall e precisão baixa, o que foi interpretado como uma indicação de que os atributos utilizados não eram suficientes para prever evasões e que outras abordagens e alternativas precisavam ser consideradas.

[Oliveira e Medeiros \(2024\)](#) utilizaram os algoritmos Decision Tree, Random Forest e SVM na identificação de alunos com alta probabilidade de evasão. O trabalho foi realizado a nível de instituição com foco em um contexto de dados relacionados à Autoavaliação Institucional, processo exigido pelo Ministério da Educação (MEC) para todas as Instituições de Ensino Superior (IES). Nesse processo, os discentes realizam uma autoavaliação de sua participação em cada disciplina, além de avaliarem o desempenho do docente, a disciplina em si e o curso. Eles também expressam seu interesse em evadir do curso atual em uma escala de zero a dez pontos. Os autores sugerem como trabalhos futuros a investigação de modelos específicos por curso e a incorporação de dados de outras naturezas como socioeconômicos, acadêmicos e culturais.

O trabalho de [Viana, Santana e Rabêlo \(2022\)](#) também adotou uma abordagem de criação de modelos para cada período cursado pelos alunos, chamada de 'Janelas Semestrais'. Estratégias de classificação foram aplicadas a dois cursos de graduação específicos na área de Tecnologia da Informação, com modelos construídos do 1º ao 6º período, devido à escassez de dados de alunos evadidos a partir do 7º período. Os alunos presentes nas amostras dos modelos propostos por Viana não são exclusivos de cada modelo, o que significa que um mesmo aluno pode estar presente em todos os modelos de um curso, embora com dados acadêmicos cumulativamente registrados até o período específico do modelo. Esses dados acadêmicos se mostraram os mais relevantes, com ênfase na 'média de notas em cada período' e no 'número de disciplinas concluídas no prazo', demonstrando forte influência na predição. Os resultados também revelaram uma acurácia crescente com o passar dos períodos, variando de 86% a 95,7% nos modelos dos quatro principais algoritmos utilizados, sugerindo que um algoritmo pode ser mais adequado para os dados disponíveis em um período, enquanto outro algoritmo pode ser melhor para outro período.

A abordagem adotada no estudo de [Costa, Júnior e Santana \(2023\)](#), aplicada à estratégia de janelas semestrais, incorpora técnicas de aprendizado de máquina supervisionado e não supervisionado. Inicialmente, o trabalho visa ampliar o volume de dados por modelo de maneira não sintética, utilizando o algoritmo de clusterização K-means para identificar cursos com comportamentos semelhantes. Posteriormente, são realizadas

predições com base nos dados dos alunos agrupados em cursos similares, em contraposição à utilização exclusiva dos dados de um único curso. Essa abordagem tornou viável a criação de modelos de predição específicos para cada período semestral pretendido, superando a limitação anterior de dados insuficientes. Com a disponibilidade de mais dados, obtida através do agrupamento de alunos em grupos similares, os modelos foram capazes de capturar mais informações e padrões, resultando em predições mais confiáveis.

Os estudos de influência abordam uma variedade de algoritmos e técnicas, todos relacionados à evasão no ensino superior (incluindo uma revisão sistemática que também considera outros níveis de ensino). Em oito deles, foram utilizados métodos de Aprendizado Supervisionado com técnicas de classificação, enquanto em um deles foram aplicadas técnicas de regressão. Este último foi o único que utilizou técnicas de ensemble além do simples uso de algoritmos dessa natureza, como o Random Forest (RF). Aliás, o RF é um dos algoritmos mais utilizados nesses estudos, juntamente com Decision Tree. Uma variedade de níveis de organização dos dados pode ser observada nesses estudos, variando entre dados de todo o país (muitas instituições), uma única instituição, um ou mais cursos de uma instituição, um ou mais períodos, um ou mais cursos de uma instituição, ou até mesmo de uma disciplina de um determinado curso. Assim como a natureza dos dados, que também é bastante diversa, entre dados socioeconômicos, acadêmicos, demográficos, geográficos e dados provenientes de questionários, como a autoavaliação discente.

Neste trabalho, utilizamos algumas das estratégias empregadas em trabalhos relacionados, com destaque para os estudos de [Costa, Júnior e Santana \(2023\)](#) e [Viana, Santana e Rabêlo \(2022\)](#). Ambos os estudos utilizaram a mesma base de dados, uma vez que as instituições envolvidas fazem uso do mesmo sistema acadêmico (SIGAA-UFPI), desenvolvido pela UFRN. No entanto, nosso estudo se diferencia dos demais devido à aplicação de técnicas de Ensemble sobre modelos construídos de acordo com as especificidades acadêmicas de cada período, e sobre datasets com estruturas distintas. Buscamos aproveitar a diversidade entre os classificadores criados a partir deles, combinando suas predições.

Para isso, criamos datasets em diferentes níveis (Instituição, Cursos e Períodos), testando cada um separadamente. Os resultados individuais foram promissores, demonstrando utilidade em situações específicas (assim como nos trabalhos de influência). Em seguida, aplicamos métodos de ensemble Bagging sobre os modelos de Períodos, o que resultou em melhorias nos níveis de predição em relação aos modelos originais. Posteriormente, combinamos modelos dos três níveis usando a técnica de Stacking, alcançando também resultados expressivos. Por fim, combinamos as duas estratégias (Stacking com Bagging) para obter os melhores resultados do estudo, demonstrando um grande potencial para uma investigação mais aprofundada dessa abordagem.

2.4.2 Considerações do Capítulo

Neste capítulo, apresentamos o processo de Descoberta de Conhecimento em Bases de Dados, uma introdução ao Aprendizado de Máquina e os trabalhos de influência relacionados à Mineração de Dados Educacionais, como embasamento teórico para auxílio à compreensão do tema e dos conceitos associados a este estudo. Destacamos os conceitos e fases envolvidas no DCBD e as diversas abordagens, técnicas e paradigmas envolvidos na área de Aprendizado de Máquina. Além disso, apresentamos os fundamentos e métodos do Aprendizado Ensemble, foco da metodologia apresentada a seguir.

3 Metodologia Proposta e Resultados Encontrados

Neste capítulo vamos abordar as ferramentas e metodologia empregadas na Descoberta de Conhecimento em Bases de Dados, destacando a construção dos conjuntos de atributos e dos datasets utilizados em nossa proposta. Dividimos as abordagens em três estudos de caso, precedidos por uma explicação geral das situações comuns a todos eles, culminando na apresentação dos resultados específicos de cada abordagem.

3.1 Contextualização

A Metodologia proposta para este trabalho de Mineração de Dados Educacionais emprega todas as fases do processo de DCBD, desde a Coleta à Análise e Interpretação de Resultados. Com apoio de diversas ferramentas de análise e manipulação, os dados foram preparados para atender a cada modelo de aprendizado de máquina conforme explanado ao longo das seções 'Compreensão dos Dados e Criação dos Datasets' e 'Construção e Treinamento dos Modelos'.

Em todos esses modelos, tanto os treinados com um algoritmo classificador individual (IND) quanto aqueles que utilizaram classificadores ensembles, empregamos a técnica de validação cruzada (cross-validation, em inglês) para avaliar a capacidade de generalização desses modelos. Nesta técnica utilizada para avaliar modelos de AM, o conjunto de dados original é dividido aleatoriamente em K subconjuntos de tamanho igual. Um dos subconjuntos é escolhido como conjunto de validação para testar o modelo, enquanto os outros $K-1$ subconjuntos são usados para treinamento. Isso é repetido K vezes, com cada subconjunto sendo usado uma única vez para validação. Após esse processo, calcula-se a média dos K resultados para obter uma única estimativa (ARAUJO, 2014). Em nosso caso, utilizamos Validação cruzada com cinco subdivisões (ou $K=5$), utilizando o método conhecido como K -fold.

Inicialmente, todas as predições foram avaliadas com base na métrica Acurácia (ACC), visando garantir a consistência das análises e facilitar comparações com outros estudos, devido à sua simplicidade e uso comum. No entanto, é importante complementar a avaliação do modelo com outra métrica mais apropriada aos nossos dados e objetivos de negócio. Considerando nossa premissa de combater a evasão e criar políticas públicas para enfrentar suas causas, o Recall poderia ser uma métrica interessante por ser mais sensível aos casos de Falsos Negativos. Em nosso contexto, esses Falsos Negativos ocorrem quando o modelo erroneamente prevê que um estudante não está propenso à evasão, o

que impossibilita a aplicação de políticas de mitigação dos fatores de risco específicos. Isso pode resultar na evasão efetiva, comprometendo nossos objetivos de intervenção eficaz. Porém, em outros contextos que exigissem por exemplo maior sensibilidade sobre os Falsos Positivos, o Recall não seria a métrica ideal.

O MCC, no entanto, é uma métrica que considera todas as possibilidades em um problema de classificação binária, incluindo os casos Positivos (verdadeiros e falsos) e Negativos (verdadeiros e falsos). Embora forneça uma visão geral do desempenho do modelo, o uso da ACC (ou mesmo do Recall) pode não ser suficiente, especialmente em situações com desequilíbrio de classes que, como veremos mais adiante, é o nosso caso. O MCC é uma métrica robusta para conjuntos de dados desbalanceados, e, portanto, avaliamos como a métrica ideal para complementar nossas análises juntamente com a ACC. Os resultados foram registrados em uma nova estrutura de dados, especificamente projetada para comparar os valores das métricas, a partir dos quais foram gerados gráficos para visualização e análise.

Em relação aos algoritmos, inicialmente adotamos DT como estimador, tanto para os modelos IND quanto como estimadores-base nos modelos ensemble. Isso com o propósito de padronizar as comparações neste aspecto, potencializando o foco na análise da diversidade presente nos conjuntos de dados e simplificando a avaliação da quantidade de modelos. Posteriormente outros algoritmos da família de Árvores de Decisão também foram avaliados. A Figura 15 apresenta todas as fases do processo de DCBD aplicadas nos Estudos de Caso 1 e 2 deste trabalho, organizadas em quatro etapas (retângulos) e será melhor detalhada ao longo das seções seguintes.

3.2 Ferramentas

As ferramentas adotadas neste estudo desempenham um papel fundamental na condução da pesquisa. Elas não apenas são amplamente estabelecidas nesse campo, mas também foram selecionadas com o propósito de alinhar nosso projeto com as abordagens estabelecidas por [Viana, Santana e Rabêlo \(2022\)](#), que serviram como referência para nossa pesquisa. A seguir, apresentamos uma visão detalhada de cada uma dessas ferramentas.

Python

Python¹ é uma linguagem de programação amplamente utilizada em aprendizado de máquina devido à sua simplicidade e eficiência. Sua versatilidade e vasta comunidade de desenvolvedores a tornam uma escolha sólida para nossas necessidades.

¹ Disponível em: <<https://www.python.org/>>

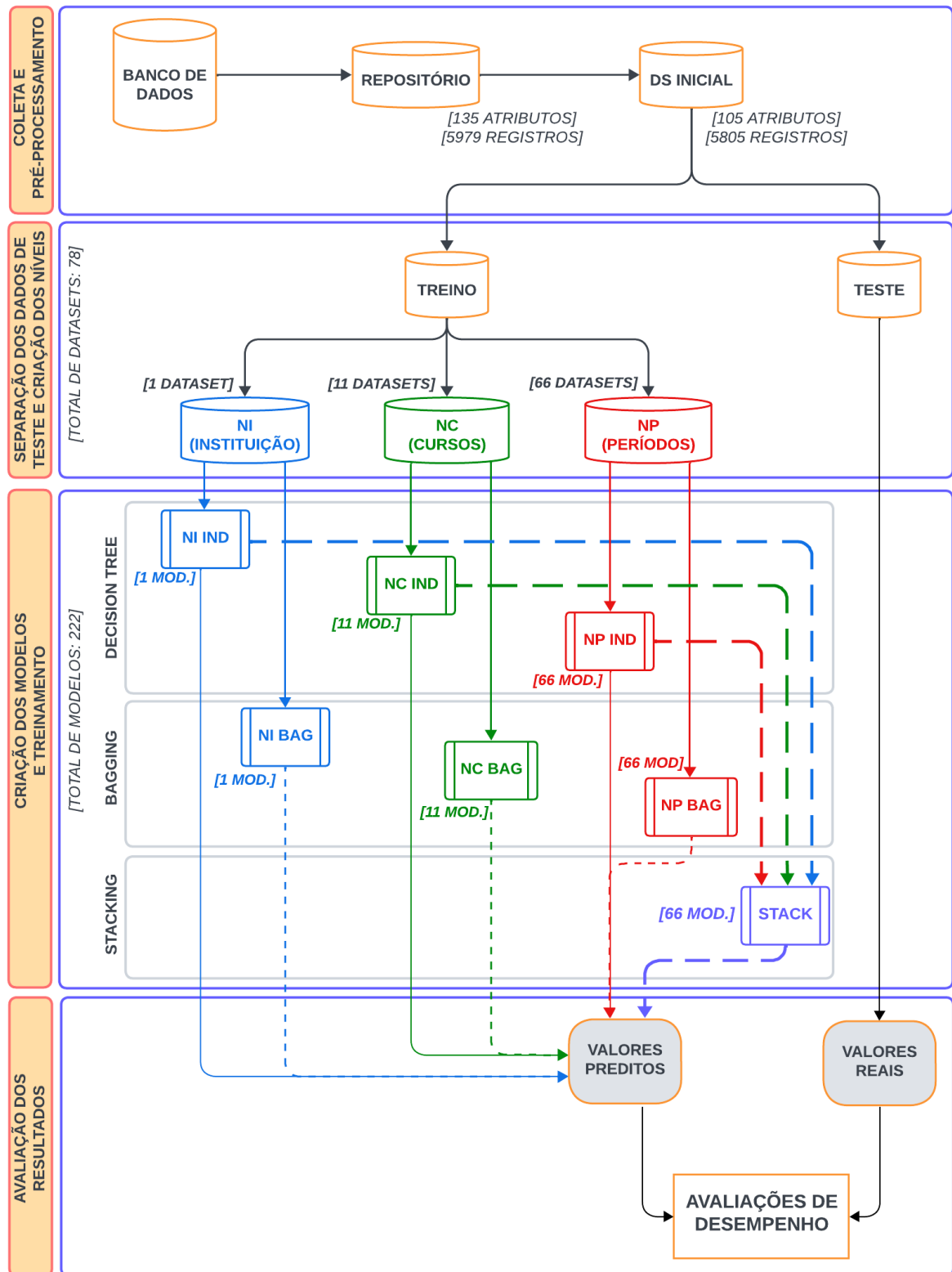


Figura 15 – Visão detalhada da metodologia proposta para EC1 e EC2

Scikit-learn

Já o Scikit-learn² é uma biblioteca de aprendizado de máquina em Python. Ela oferece uma ampla gama de algoritmos e ferramentas para tarefas de classificação, regressão e agrupamento. É uma escolha sólida para construir e avaliar modelos de aprendizado de máquina.

Pandas

O Pandas³ é uma biblioteca Python fundamental para manipulação e análise de dados. Com suas estruturas de dados flexíveis, como DataFrames, fomos capazes de lidar com os dados de forma eficiente e eficaz, manipulando-os desde a planilha (‘.csv’) inicial contendo os dados brutos, passando pela limpeza, transformações, geração dos novos datasets etc.

Google Colab

Para otimizar nossa pesquisa, também fizemos uso do Google Colab⁴. Essa plataforma em nuvem possibilita a execução de código Python e a colaboração em projetos de ciência de dados, tornando nosso trabalho mais eficaz e disponível em qualquer local. Além disso, o Google Colab dispensa a necessidade de possuir computadores de alto desempenho ou configurações locais, pois oferece um ambiente pronto para uso na nuvem.

Visualização de Dados com Seaborn e Matplotlib

Finalmente, depois de estabelecer nossas bases em Python e no processamento de dados com Pandas, mergulhamos na visualização de dados. Para isso, usamos duas bibliotecas poderosas: Seaborn⁵ e Matplotlib⁶. Essas ferramentas simplificam a criação de gráficos informativos e esteticamente agradáveis, permitindo-nos comunicar nossos resultados de forma eficaz.

3.3 Compreensão dos Dados e Criação dos Datasets

A fonte original dos dados utilizados neste estudo é a base de dados do sistema acadêmico utilizado pela instituição que deu origem à UFDPAr, o SIGAA, da Universidade Federal do Piauí (UFPI). Esse sistema resulta de um acordo de Cooperação Técnica entre a UFPI e a Universidade Federal do Rio Grande do Norte (UFRN), que iniciou sua

² Disponível em: <<https://scikit-learn.org/>>

³ Disponível em: <<https://pandas.pydata.org/>>

⁴ Disponível em: <<https://colab.google/>>

⁵ Disponível em: <<https://seaborn.pydata.org/>>

⁶ Disponível em: <<https://matplotlib.org/>>

implementação em 2003. Através deste tipo de acordo, a UFRN, detentora dos direitos autorais do sistema, permitiu que diversas outras IFES do país adotassem o sistema. Sua implantação na UFPI teve início em 2011, o mesmo ano em que a UFPI passou a adotar predominantemente a nota do Exame Nacional do Ensino Médio (Enem) como critério de seleção de candidatos à graduação, e o processo seletivo passou a ocorrer por meio do Sistema de Seleção Unificada (SISU) (UFPI, 2019).

Antes disso, a Instituição utilizava um sistema próprio, bem menos robusto. E utilizava-se de seus próprios processos seletivos para o ingresso dos alunos. Com a aquisição do novo sistema e a mudança na forma de ingresso, bem como na estrutura dos currículos e outras situações ocorridas em consequência da migração de dados/sistema, muitas das informações advindas do sistema legado não possuem hoje nível de organização suficiente para trabalhos relacionados à Aprendizagem de Máquina. Apenas para exemplificar, alguns tipos e componentes curriculares não possuíam o registro de notas por avaliação no sistema, sendo de responsabilidade do docente o lançamento exclusivamente da nota final (média) do discente no componente. Outro exemplo é o registro de Turmas, que era realizado de forma distinta do SIGAA e não pôde ser migrado, ocasionando ausência de dados legados que praticamente inviabilizam o processo de análise dos mesmos.

3.3.1 Etapa: Coleta e Pré-Processamento

A primeira etapa de nosso processo de descoberta de conhecimento engloba as atividades relacionadas a Coleta, Análise Exploratória, Limpeza e Pré-Processamento, conforme Figura 16 que expande a visualização desta fase a partir da imagem com o processo completo. A fase (e os procedimentos) da Coleta foi realizada em maio de 2023 e restringiu-se ao nível de graduação na modalidade presencial regular, abrangendo discentes que ingressaram entre 2012.1 e 2023.1. Em um processo simplificado de ETL (Extração, Transformação e Carga), um repositório inicial foi criado com apoio dos Analistas de Dados e Especialistas de Domínio da UFPI e da UFDPAr.

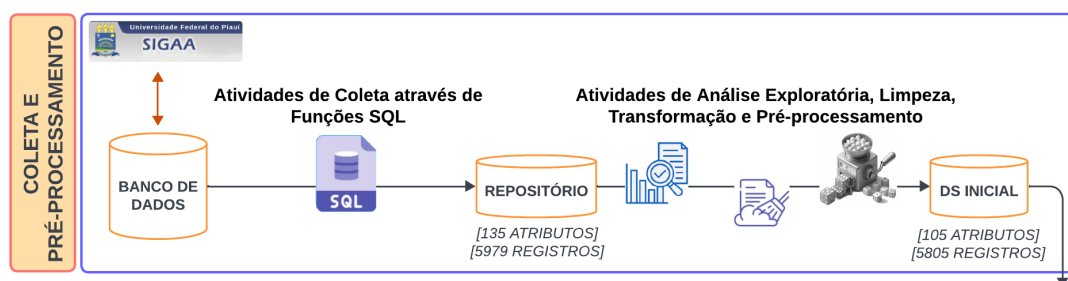


Figura 16 – Detalhamento da Etapa Coleta e Pré-Processamento

A extração ocorreu por meio de consultas diretas ao banco de dados do sistema, as quais já realizavam de antemão uma Limpeza Prévia e algumas pequenas transformações nos dados, em especial quando da criação dos atributos relacionados às informações

acadêmicas cumulativas (por semestre) dos discentes. Através da contabilização de registros de algumas características ocorridas de forma cumulativa ao longo do curso, criamos para cada um dos períodos, atributos referentes ao total de trancamento de curso, total de trancamento e cancelamento de turmas/componentes, total de reprovações, total de carga horária integralizada, quantidade de monitorias, quantidade de pesquisa, quantidade de projetos de ensino, quantidade de projetos de extensão, quantidade de reprovações de disciplinas obrigatórias por semestre e quantidade de disciplinas (obrigatórias ou não) aprovadas no prazo ideal ou antecipadas. Os atributos *evadido*, *idade_ingresso*, *intervalo_EM_ingresso* e *curvou_na_pandemia* também foram criados ainda durante a consulta às bases de dados através de outras características existentes, como o status do discente, a presença ou ausência de matrículas em disciplinas, a data de nascimento, a data de ingresso na instituição e a data de conclusão do ensino médio. Além de outros atributos retirados durante as etapas seguintes.

Na sequência, no processo de Pré-Processamento, passamos por etapas mais detalhadas de preparação, limpeza e transformação dos dados. Nesta etapa realizamos uma Análise Exploratória para conhecer melhor os dados e identificar ruídos, dados ausentes e necessidades de intervenções. Foram realizadas atividades para tratamento de dados ruidosos, como discentes com data de nascimento ou ano de conclusão do segundo grau inferior ao ano 1000, ou ainda conclusão no mesmo ano (ou inferior) ao ano de nascimento. Imputações baseadas em médias de registros similares ou, em último caso, exclusão de registros foram soluções adotadas de acordo com as especificidades de cada caso.

As ausências de informações relacionadas a conclusão em escola pública (ou não), em grande parte puderam ser resolvidas através de imputações baseadas na Ação Afirmativa/Cota concorrida pelo discente. As colunas *id_convocacao_processo_seletivo* e *id_processo_seletivo*, identificadores/chaves das tabelas referentes aos processos seletivos e às convocações de alunos aprovados, foram transformadas em um novo atributo numérico chamado *convocacao*, responsável por identificar se o discente adentrou entre os primeiros alunos convocados daquele processo seletivo ou em convocações posteriores, sendo os atributos originais descartados por irrelevância em seu formato inicial para nossos treinamentos. Da mesma forma, assim como outras colunas, a coluna *ano_conclusao_segundo_grau* foi excluída após dar origem a coluna *intervalo_EM_ingresso*.

Percebemos ainda durante a Análise Exploratória o desbalanceamento de nossos dados, ou seja, a distribuição das classes é significativamente desigual, sendo nossos dados compostos, a priori, em grande maioria por registros de discentes não evadidos. Com a criação de outros níveis, essa realidade se inverte em grande parte dos novos modelos. Optamos então por não realizar atividades de balanceamento diretamente sobre a construção dos dados, mantendo-os mais realísticos, e deixando a questão a cargo das estratégias de mineração e avaliação.

Percebemos também baixíssima quantidade de dados positivos na classe alvo (target) para alguns cursos no sétimo e oitavo períodos. Embora tratemos de cursos totalmente distintos, essa percepção confirma a verificação de [Viana, Santana e Rabêlo \(2022\)](#) que não utilizou dados desses períodos por este mesmo motivo. Optamos então por manter estes períodos fora do escopo de nosso estudo, focando nas características acadêmicas cumulativas apenas até o sexto período cronológico dos discentes.

Realizamos ainda uma cuidadosa seleção manual de atributos, apoiada pela Matriz de Correlação de Pearson, a qual é a técnica mais difundida para avaliação da correlação entre duas variáveis quantitativas, e amplamente utilizada em análises de dados ([MIOT, 2018](#)). Esta matriz de correlação de atributos nos permitiu avaliar as relações lineares entre eles, identificar e eliminar alguns deles de acordo com a relevância das correlações, fossem elas em relação ao target ou a outros atributos. Assim, como resultado do Pré-Processamento, o repositório inicial criado, que continha 120 atributos e 5979 registros, gerou um primeiro Dataset de Trabalho, com 105 atributos e 5805 registros.

Na sequência, novos processos de Transformação continuaram a ser realizados, tendo em vista a necessidade de criação de outros conjuntos de dados derivados e de modificações nos atributos. Transformamos alguns atributos categóricos em numéricos, bem como alguns dados em formatos complexos como datas, que foram convertidos em atributos numéricos como o intervalo de tempo entre a conclusão do ensino médio e o ingresso na universidade (*intervalo_EM_ingresso*) e a idade de ingresso.

3.3.2 Etapa: Separação dos Dados de Teste e Criação dos Níveis

A partir daí, visando construir as estruturas dos datasets tal qual serão consumidas pelos modelos de aprendizagem, adentramos em uma das etapas mais delicadas deste processo, que refere-se às estratégias de separação dos dados, a fim de evitar o 'vazamento' de dados de teste para algum dos conjuntos de treinamento. Esse tipo de vazamento pode levar ao overfitting e à superestimação do desempenho do modelo, uma vez que o modelo pode se ajustar a esses exemplos, mas não conseguir generalizar bem para novos casos não vistos.

Para evitar esse vazamento, dado o grande número de datasets que vamos utilizar, conceitualmente nesta etapa selecionamos aleatoriamente uma amostra 'representativa' de 30% dos discentes do universo total, garantindo que a seleção fosse estratificada. Essa seleção ocorreu como um primeiro processo de separação dos dados, com essa faixa de discentes correspondendo a partir de então ao nosso conjunto de dados de 'Teste', conforme mostrado na Figura 17, que oferece uma visão ampliada da etapa de Separação dos Dados de Teste e Criação dos Níveis.

Esses dados não são utilizados em nenhum dos modelos durante a fase de Treina-

mento. Em vez disso, são reservados exclusivamente para a realização de validações em cada modelo, seja na sua totalidade ou em subconjuntos criados para manter a proporção de 70% para Treinamento e 30% para Teste em cada modelo específico. Na prática, no entanto, essa separação inicial ocorreu no início da criação dos modelos de EC1, como detalharemos mais adiante.

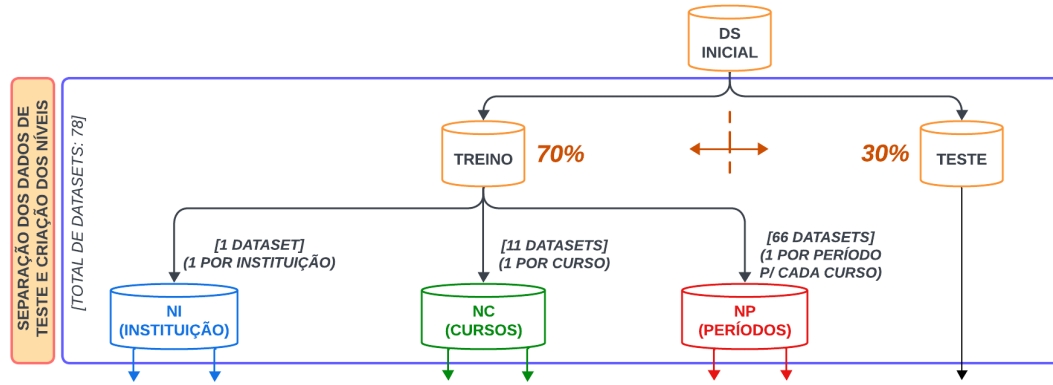


Figura 17 – Detalhamento da Etapa Separação dos Dados de Teste e Criação dos Níveis.

Assim, desenvolvemos três níveis de datasets, aptos a compor modelos de níveis distintos já existentes na literatura. A decisão de escolher os atributos de forma manual, em detrimento de uma estratégia de seleção automática, foi tomada pelos especialistas do domínio, visando possibilitar a criação de níveis distintos com características que se presumia capazes de introduzir diversidade entre os três modelos. Cada componente individual do ensemble realiza o mesmo trabalho de rotulação dos padrões de entrada. Porém cada um deles precisa ser implementado sob diferentes visões do conjunto de dados (SANTANA; CANUTO, 2014).

Faz-se necessário então abdicar do melhor desempenho dos classificadores individuais em benefício do resultado combinado. Enquanto a seleção automática se concentraria na otimização do desempenho dos modelos-base, ao optar pela seleção manual de atributos, os especialistas fortaleceram a robustez do conjunto, permitindo a incorporação de diversas nuances e perspectivas para aproveitar-se da natureza combinatória do ensemble. A intenção é que cada nível contribua de maneira única, seja por meio do volume, especialização dos dados ou características específicas, resultando em um ensemble capaz de lidar com maior variedade de cenários e maximizando a eficácia da abordagem adotada. Agora, exploraremos os níveis criados e suas características:

Nível Instituição (NI): O Modelo gerado teve suas features resumidas de 105 para 24 atributos, mantendo no entanto os 5805 registros resultantes dos processos de limpeza e transformação aplicados aos dados coletados. Desses registros, 3298 correspondem a discentes considerados evadidos e 2507 concluíram seus cursos. Para este nível, a grande maioria das características está associada a questões/situações definidas anteriormente ao início efetivo do curso, como dados demográficos e informações referentes à transição

do ensino médio à Universidade. Um dos atributos identifica a qual curso o discente está vinculado. Esta feature é a única deste nível que não se encontra no nível NC.

Nível Curso (NC): Nesse nível de especialização, agrupamos os discentes por seus cursos, através do atributo identificador do curso do discente, mencionado anteriormente, formando 11 datasets distintos. Este atributo não está presente no NC, pois, sendo uma constante dentro de cada curso/dataset, foi descartado durante a fase de Transformação. Salvo essa exceção, os demais atributos são os mesmos do nível NI. Quanto ao número de registros, os datasets contêm entre 280 e 653 registros, dependendo do curso (em média 528 registros por dataset NC).

Nível Período (NP): Ampliando o nível de ‘especialização’ dos datasets, neste último nível, equivalente às Janelas Semestrais, utilizamos como atributo para agrupar os registros dos discentes com informações em um determinado período o *‘período_cronologico_do_discente’*.

Além disso, selecionamos as features adequadas para a montagem de cada dataset deste nível, que são chamados de P1, P2, P3, P4, P5 e P6, em referência aos seis primeiros períodos do discente no curso. Esse não é um processo de separação de discentes em Datasets de Períodos, mas sim o agrupamento das informações acadêmicas cumulativas dos discentes. Por exemplo, se um discente cursou três períodos, haverá um registro com dados cumulativos sobre seu desempenho acadêmico nos datasets P1, P2 e P3. As features em cada dataset são distintas e possuem valores referentes à época em que o discente concluiu aquele período específico. Além das features presentes no nível NC. No total, temos seis datasets para cada um dos onze cursos, totalizando (inicialmente) 66 datasets NP. A figura 18 demonstra a adição dos dados de natureza acadêmica no nível NP.

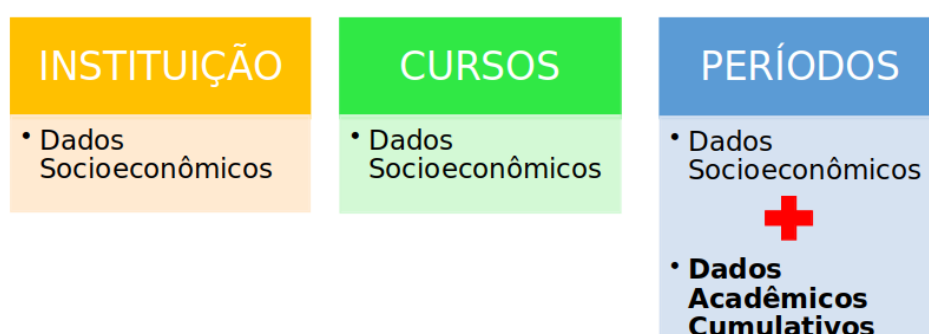


Figura 18 – Natureza dos dados de cada nível.

Para ilustrar, em médias bastante simplificadas, temos cerca de 5805 registros para NI, aproximadamente 528 registros por dataset no NC, e cerca de 378 registros por dataset no nível NP. Com base nessa ilustração, podemos melhor elucidar que a principal diferença entre NI e NC está no universo amostral, enquanto de NC para NP está nas diferentes features relacionadas a dados acadêmicos temporalmente cumulativos. Ressaltamos ainda,

novamente, que para o NP não se trata de uma divisão ou agrupamento de registros, mas sim das informações acadêmicas cumulativas dos discentes. A Figura ?? exibe as distribuições reais de registros por dataset NP dos cursos de MAT e PED, destacando as discrepâncias nas quantidades de registros por dataset e a proporção visual de registros correspondentes a alunos considerados evadidos. As barras representam o total de registros naquele dataset, enquanto os valores sobre a linha vermelha representam quantos destes registros são referentes a discentes que evadiram.

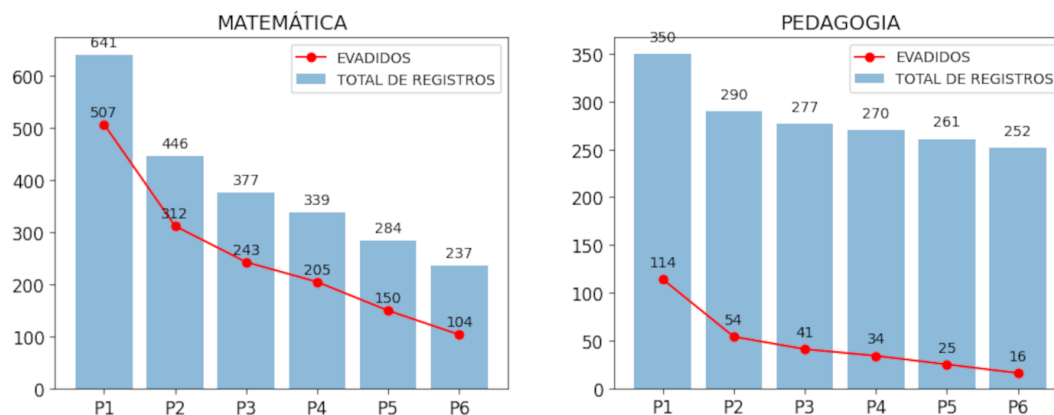


Figura 19 – Número de Amostras x Número de Evadidos - MAT e PED

Entre um curso e outro, observa-se uma variação significativa na proporção de registros de evadidos em relação ao total de registros, sendo que em MAT os registros correspondentes a evadidos no P1 são maioria, enquanto em PED, para o mesmo período, são a minoria (cerca de um terço). Além disso, em ambos os cursos, o número de registros de evadidos diminui proporcionalmente à redução do total de registros. Na Figura ??, que ilustra as distribuições de registros dos demais cursos, percebemos que esse padrão não se repete uniformemente em todos os casos. Por exemplo, em Ciências Biológicas (BIO), a maioria dos registros em P1 corresponde a evadidos (407 evadidos de um total de 629 registros), enquanto no P6 essa proporção passa a ser minoritária (apenas 104 dos 325 registros correspondem a evadidos).

Percebe-se então um claro desbalanceamento dos dados, onde a proporção de registros relacionados a discentes evadidos varia tanto entre os cursos em um mesmo período, quanto entre períodos de um mesmo curso. Ou seja, há uma distribuição desigual, com variações entre os modelos de cada período e entre os diferentes cursos, com uma alternância da classe minoritária entre os modelos.

Essa subdivisão em nível de Períodos é similar à efetuada por Viana, Santana e Rabêlo (2022), mas com atributos e estratégias de criação ligeiramente distintas. Enquanto Viana, Santana e Rabêlo (2022) utilizaram um total de 23 atributos, cada dataset NP usado em EC1 e EC2, por exemplo, possui 33 atributos. Destes, 23 são 'iguais' e incluem o target *evadido*, enquanto os outros 10 representam dados acadêmicos cumulativos e identificam o

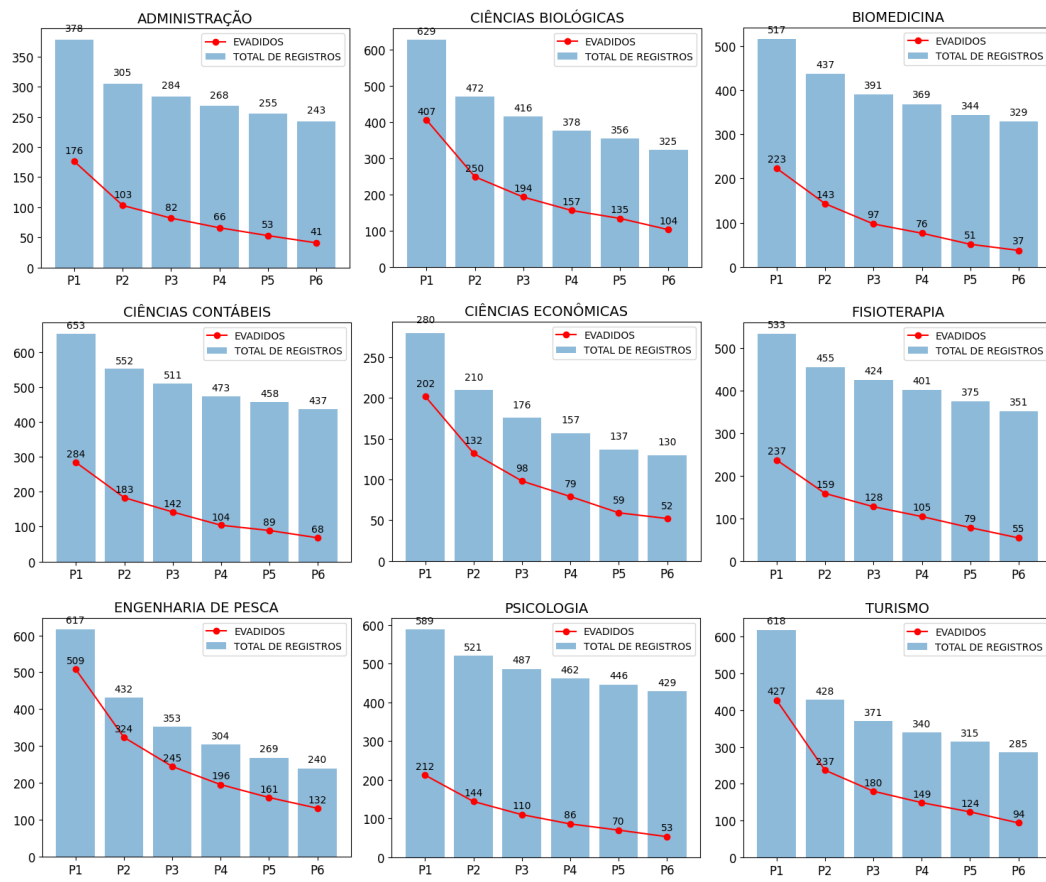


Figura 20 – Número de Amostras x Número de Evadidos - Demais Cursos

semestre em que o discente se encontra no curso. Em outras palavras, para um dataset referente ao P3 de um curso, teremos 23 atributos 'imutáveis' ao longo dos períodos, como *idade_de_ingresso*, *nota_enem_humanas*, *uf_naturalidade* e *tipoacaoafirmativa*, bem como 10 atributos referentes a dados cumulativos do discente até aquele período, como *total_trancamento_turma_P3*, que indica quantos trancamentos de turma foram efetuados pelo discente até o 3º período cronológico após seu ingresso na instituição.

Já o atributo classe, igual para todos os níveis, foi criado com base no status dos discentes: 'CONCLUÍDO' foi classificado como 'não evadido', e 'CANCELADO' como 'evadido'. Aos 'não evadidos', foram adicionados discentes ativos que cumpriram os requisitos curriculares, faltando apenas a colação de grau. Aos 'evadidos', foram adicionados discentes ativos que não efetuaram matrícula nos últimos três semestres ou mais até a coleta de dados (2023.1), situação em que o discente já é, seguindo as normas de graduação da instituição, passível de Cancelamento do Vínculo por Abandono de Curso (adicionando-se um período, como margem). Consideramos como discentes ativos (supramencionados) aqueles que possuíam o status ATIVO, CADASTRADO, FORMANDO, GRADUANDO ou TRANCADO. Salvo as exceções acima, discentes com estes status não participaram do estudo.

Além disso, como já mencionamos, removemos o atributo identificador de curso no NC, uma vez que é constante dentro de cada dataset (um mesmo número para todos os registros de um dataset). Esse identificador é, aliás, uma das duas únicas diferenças entre NI e NC, juntamente com a quantidade/variedade de registros, que, no nível Instituição, engloba discentes de todos os cursos.

A Tabela 1 apresenta a lista de atributos integrantes de cada Nível de Dataset, onde os registros terminados em “PX” (destacados ao final da tabela) representam, na verdade, os seis períodos analisados, sendo X uma representação dos Algarismos de 1 a 6. Em outras palavras, cada atributo terminado em ‘PX’ representa, na verdade, seis atributos, um para cada período. Por exemplo, no dataset NP referente ao P3, *qtde_trancamento_PX* é, na verdade, *qtde_trancamento_P3*, com seus respectivos dados cumulativos. Também em destaque (negrito) no início da tabela está nosso atributo *evadido*, o target (alvo, ou ainda rótulo) de nosso estudo. Baseado nos registros rotulados como evadidos (true) ou não evadidos (false), buscamos aprender padrões e prever outros registros não rotulados. Para a manipulação desses registros, criamos o atributo *id_discente*, um identificador do registro do discente criado na montagem do primeiro repositório, ainda em nosso processo de coleta, distinto do identificador da base de dados.

Doravante intitulado como ORIG (em referência a "original"), este primeiro conjunto de datasets abrange todos os datasets utilizados em EC1 e EC2. A partir dele, construímos outro conjunto de datasets para os demais experimentos realizados no EC3. Para isso utilizamos uma abordagem baseada no método de seleção de atributos *Filter*, cujos atributos são ranqueados através de determinada métrica e um limiar (*threshold*) é estabelecido, sob o qual todos os atributos que estiverem abaixo serão descartados da seleção (Araújo, 2014). Os datasets do conjunto ORIG foram formados praticamente pelos atributos disponíveis após os processos mencionados nas fases anteriores. Após o término das atividades de remoção de atributos de baixíssima relevância, como constantes, identificadores e atributos transformados em outros mais relevantes, e a seleção com base na Matriz de Correlação de Pearson, praticamente não restaram outras opções de atributos que não os utilizados no ORIG. Deste modo, é uma tendência natural que os atributos de um novo conjunto de datasets estejam contidos nos datasets do conjunto ORIG.

No contexto da análise de features, os algoritmos pertencentes à família de Árvore de Decisão, como o DT, possuem a capacidade nativa de extrair de cada modelo um conjunto ranqueado de features quanto à sua influência em relação aos objetivos do modelo. Ou seja, eles são capazes de apontar quais características possuem maior influência sobre o modelo durante a classificação do discente como ‘Evadido’ ou ‘Não Evadido’.

Após o treinamento dos modelos presentes no Estudo de Caso 1 por meio do algoritmo DT, foi possível extrair a importância de cada uma das features para cada um dos 156 modelos envolvidos, dos tipos Individual e Bagging criados até então, e que

Tabela 1 – Features originalmente presentes em cada um dos níveis de *Datasets*.

FEATURE	NI	NC	NP	TIPO	DESCRIÇÃO
evadido	x	x	x	Booleano	Target do estudo, classifica o discente como evadido ou não evadido (true or false)
id_discente	x	x	x	Numérico	Índice para os registros dos discentes nos datasets (não influencia nos treinamentos).
concluiu_escola_publica	x	x	x	Booleano	Indicador de conclusão do ensino médio em escola pública ou não
convocacao	x	x	x	Numérico	Identificador da Convocação do Processo Seletivo na qual o discente foi convocado
curso	x			Numérico	Identificador do curso
cursou_na_pandemia	x	x	x	Booleano	Indicador de matrículas em turmas de 2020 e 2021 (durante a pandemia)
estadocivil	x	x	x	Numérico	Estado civil declarado pelo discente
idade_ingresso	x	x	x	Numérico	Idade do discente quando ingressou na instituição
intervalo_EM_ingresso	x	x	x	Numérico	Intervalo entre a conclusão do Ensino Médio e o ingresso na instituição (anos)
nota_ciencias_humanas	x	x	x	Numérico	Nota na prova de Ciências Humanas no Enem
nota_ciencias_natureza	x	x	x	Numérico	Nota na prova de Ciências da Natureza no Enem
nota_linguagem	x	x	x	Numérico	Nota na prova de Linguagem no Enem
nota_matematica	x	x	x	Numérico	Nota na prova de Matemática no Enem
nota_redacao	x	x	x	Numérico	Nota na prova de Redação no Enem
opcao_inscricao	x	x	x	Numérico	Opção de Inscrição no Processo Seletivo (Ordem de preferência de curso)
periodo_ingresso	x	x	x	Numérico	Período de ingresso na instituição
qtde_membros_familia	x	x	x	Numérico	Quantidade de membros no núcleo familiar do discente
raca	x	x	x	Numérico	Raça declarada pelo discente
sexo	x	x	x	Numérico	Sexo do discente
tipoacaoafirmativa	x	x	x	Numérico	Ação afirmativa pela qual o discente ingressou
temcredito_cc_ca	x	x	x	Booleano	Indicador de uso de créditos no histórico por disciplinas já cursadas anteriormente
tem_ensino	x	x	x	Booleano	Indicador de participação de atividades de ensino no histórico
turno	x	x	x	Numérico	Identificador do turno
uf_naturalidade	x	x	x	Numérico	Unidade da Federação em que o discente nasceu
ap_ob_semestre_PX			x	Numérico	Quantidade de registros de aprovações em componentes/disciplinas obrigatórias até o período específico
ap_no_prazo_PX			x	Numérico	Quantidade de registros de aprovações em componentes/disciplinas até o período específico
media_acumulada_PX			x	Numérico	Média do rendimento acadêmico até o período específico
qtd_reprova_ft_nt_PX			x	Numérico	Quantidade de registros de reprovação em disciplinas por falta e nota até o período específico
qtd_reprova_ft_PX			x	Numérico	Quantidade de registros de reprovação em disciplinas por falta e nota até o período específico
qtd_reprova_nt_PX			x	Numérico	Quantidade de registros de reprovação em disciplinas exclusivamente por nota até o período específico
qtd_tranca_turma_PX			x	Numérico	Quantidade de registros de trancamentos de turmas até o período específico
qtd_cancela_turma_PX			x	Numérico	Quantidade de registros de cancelamentos de turmas até o período específico
rep_ob_semestre_PX			x	Numérico	Quantidade de registros de reprovações em componentes/disciplinas obrigatórias até o período específico
tranca_programa_PX			x	Numérico	Quantidade de registros de trancamento do curso até o período específico

abordaremos na próxima seção. Essa importância foi obtida através da análise do atributo 'feature_importances_', presente nos algoritmos de árvores de decisão, que fornece um ranking das features com base em sua contribuição para a precisão do modelo. Optamos então por contabilizar, como nossa 'métrica' no processo de seleção, a quantidade de vezes que as features compuseram o ranking das cinco features mais influentes de um modelo. Em outras palavras, decidimos armazenar em uma nova estrutura de dados cada ocorrência para cada feature presente no Top-5 das features mais influentes de um modelo. E por fim, somamos os números totais de cada feature.

Entretanto, os 156 modelos são construídos com base em oito conjuntos de atributos distintos (um para NI, um para NC e seis para NP). Essa divisão dificulta a escolha dos

atributos mais frequentes, especialmente porque os atributos acadêmicos cumulativos estão presentes apenas nos datasets NP, e são distintos de um período para outro. Como contorno, registramos esses atributos conforme constam na Tabela 1. Utilizamos um sufixo '_PX' para representar o período, consolidando todos os atributos cujo nome difere apenas no sufixo (atributos terminados em 01, 02, 03, 04, 05 e 06). Isso nos permite identificar, independentemente do período, o grupo de atributos que, em tese, melhor representa o nível como um todo. Por exemplo, se *qtde_trancamento_P1* e *qtde_trancamento_P3* aparecerem uma vez cada entre as Top-5 features mais importantes de qualquer modelo, contabilizamos *qtde_trancamento_PX* duas vezes.

Assim, ao ranquearmos os atributos que mais frequentemente se destacam como relevantes para cada nível (e período), obtivemos um novo conjunto de datasets (os mesmos dados reorganizados sob oito novos conjuntos de features) intitulado FILT (em referência à "filtrado"). Os níveis NI e NC foram compostos por todos os atributos que figuraram entre os cinco mais influentes em pelo menos um modelo, resultando em seis e doze atributos para cada um, respectivamente. Os demais, dos mais de vinte atributos desses níveis no ORIG, não alcançaram o Top-5 em nenhum modelo. No nível NP, apenas quatro atributos não figuraram entre os cinco mais influentes em nenhum modelo. No entanto, a simples retirada desses quatro atributos resultaria em datasets muito semelhantes aos NP do conjunto ORIG.

Optamos, portanto, por reduzir o número de atributos excluindo aqueles que figuraram entre os cinco melhores em menos de 10% do total de modelos. Isso resultou em um total de quinze atributos, aproximadamente metade do número de atributos do conjunto ORIG. A Figura 21 mostra uma nova versão expandida da etapa de Separação dos Dados de Teste e Criação dos Níveis a ser utilizada no Estudo de Caso 3, onde detalharemos os atributos que compõem o conjunto FILT.

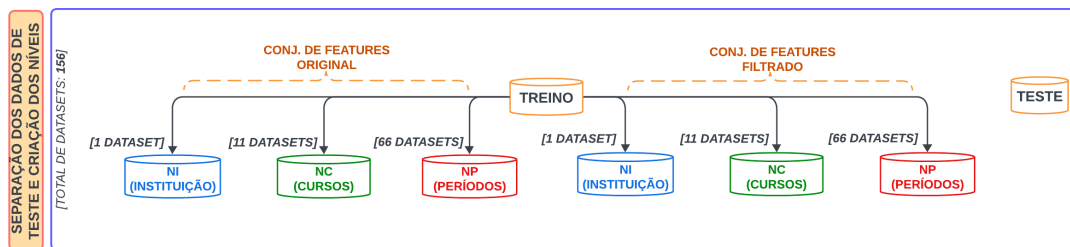


Figura 21 – Etapa de Separação dos Dados de Teste e Criação dos Níveis, ampliada em EC3 com a criação dos novos Conjuntos de Features e dos Datasets

3.4 Construção e Treinamento dos Modelos

Na fase de Mineração, dividimos nossos experimentos em três Estudos de Caso, visando abordar de forma diferenciada a construção dos modelos para cada tipo de análise,

bem como os seus resultados. Ao todo, foram criados 2520 modelos, conforme ampliação da segunda e terceira fases de nosso processo presente na Figura 22.

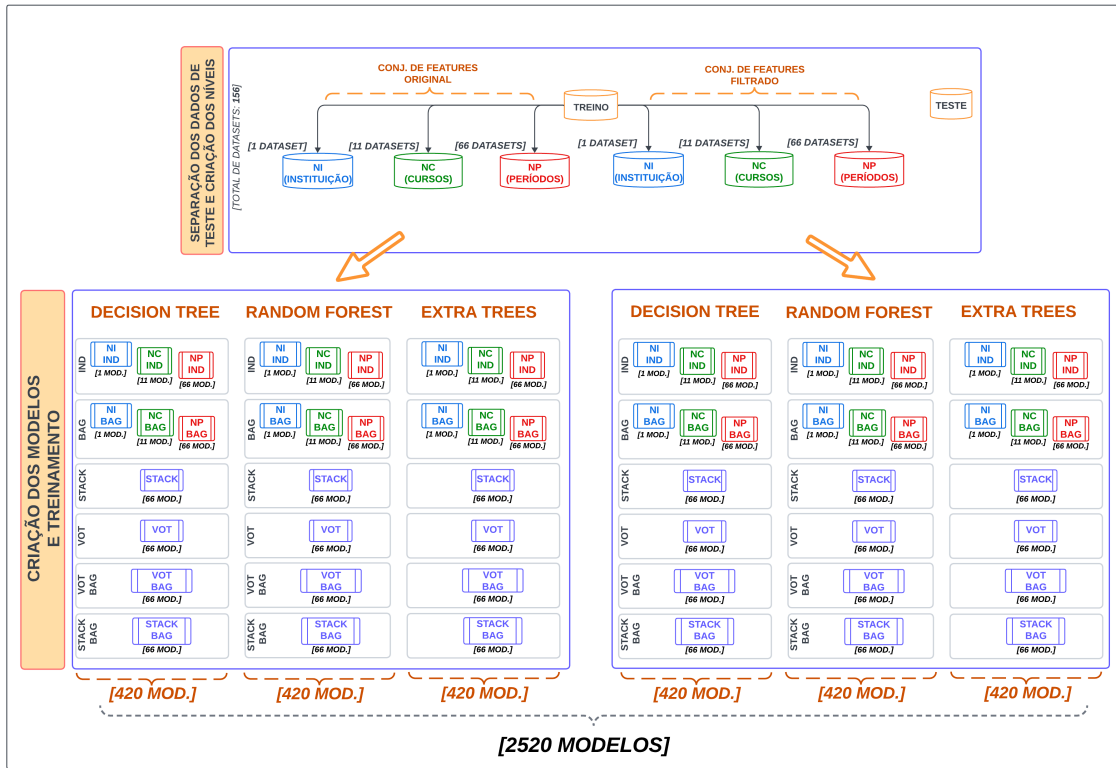


Figura 22 – Ampliação da etapa de Criação dos modelos com foco no quantitativo de modelos criados ao longo de EC1, EC2 e EC3.

Para abordá-los ao longo dos Estudos de Caso, por diversos momentos os modelos serão tratados como um grupo formado por todos aqueles associados a um mesmo conjunto de datasets (dentre os dois no total), um mesmo algoritmo (dentre os três no total) e um mesmo tipo de modelo (dentre os seis no total). Complementando os fatores que utilizaremos para agregar e tratar de forma combinada esses modelos, a mesma métrica de avaliação utilizada (dentre as duas no total) integra o que chamaremos de Combinações, as quais teremos 72 no total.

Como forma de avaliar estas combinações, ao analisar estatisticamente as diferenças entre os resultados das métricas de cada combinação, trataremos como Percentual de Melhoria a razão entre a quantidade de modelos ensemble de uma combinação que se mostrou significativamente superior a respectivo modelo da combinação do tipo IND que utiliza o mesmo conjunto de datasets, o mesmo algoritmo base e a mesma métrica para a comparação. E como forma de medir se nossas estratégias são viáveis para aplicação prática, trataremos como Percentual de Aplicabilidade a razão entre os modelos que se mostraram sem diferença estatística significativa ou superiores estatisticamente (ou seja, que não pioraram com a aplicação dos ensembles) segundo a mesma metodologia de comparação do percentual de melhoria.

3.4.1 Estudo de Caso 1 (EC1)

Considerando que precisaremos avaliar comparativamente os modelos, decidimos utilizar inicialmente na criação de todos eles o algoritmo Decision Tree (DT), amplamente utilizado nesta área (SANTOS; SARAIVA; OLIVEIRA, 2021).

Por amostragem, verificamos que a acurácia de muitos de nossos modelos, ao ser aferida nos dados de treinamento sem o uso de validação cruzada, frequentemente se aproximava do valor máximo, oscilando entre 97% e 100%. No entanto, ao avaliarmos os mesmos modelos nos dados de teste, raramente a acurácia ultrapassava 87%. Isso sugere fortemente que os modelos estavam se ajustando excessivamente aos dados de treinamento (overfitting). Ao realizarmos a mesma avaliação utilizando a validação cruzada, observamos uma redução significativa na acurácia dos dados de treinamento, aproximando-se mais dos valores obtidos nos dados de teste. Isso indica que o uso da validação cruzada está contribuindo para mitigar o overfitting.

Como discutido anteriormente, o uso de métodos ensemble bagging pode ser uma estratégia eficaz para reduzir a variância do modelo e lidar com o overfitting, especialmente quando combinado com árvores de decisão ou outros modelos propensos ao overfitting. Esses métodos tendem a aumentar a estabilidade e melhorar a generalização dos modelos. Portanto, neste primeiro estudo de caso, nosso objetivo é criar modelos de predição de evasão mais eficazes usando Métodos Ensemble Bagging em um cenário de janelas semestrais.

Se conceitualmente os processos de separação dos dados de Treinamento e Testes de todos os modelos e abordagens estão vinculados à etapa anterior, na prática foi no início desta etapa que realizamos os processos práticos dessa separação. Quando da execução pela primeira vez de nosso método de criação e treinamento dos modelos, a função *train_test_split* se encarrega de separar do conjunto total de exemplos trinta por cento dos registros para formação dos dados de Teste, com o parâmetro *stratify* apontando a necessidade de que esses dados mantenham a representatividade de diferentes grupos ou estratos dentro do conjunto de dados. Esses conjuntos iniciais de Treinamento e Teste servirão como base para a formação de todos os outros modelos, sendo repassados como parâmetros aos métodos responsáveis pelos treinamentos, sejam em sua totalidade ou calculados os percentuais exatos para a especificidade de cada modelo.

Assim, utilizando DT, geramos todos os modelos iniciais, incluindo um modelo para toda a Instituição, um modelo por Curso (totalizando 11) e um modelo para cada período de cada curso (totalizando 66). Em seguida, geramos a mesma quantidade de modelos por nível, tendo o DT como o estimador-base de um classificador BaggingClassifier (estratégia que doravante trataremos como BAG), criando mais 78 modelos do tipo Bagging (1 NI, 11 NC e 66 NP). Ou seja, para cada modelo individual criado nos três níveis estudados, um

novo modelo foi criado, no entanto substituindo IND por BAG (ambos utilizando DT), conforme Figura 23.

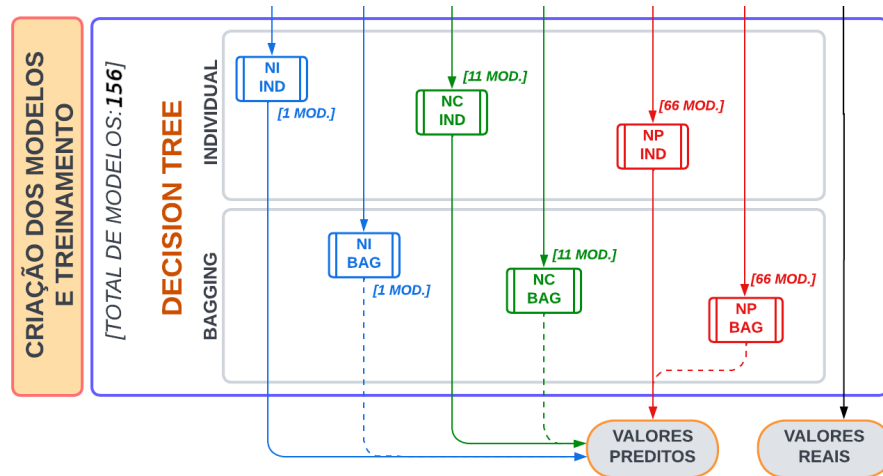


Figura 23 – Detalhamento da etapa de Criação dos Modelos de Treinamento do EC1.

Internamente, para a criação dos modelos BAG, cada modelo foi submetido à estratégia de amostragem com reposição (característica dos métodos Bagging), para a criação de múltiplas Árvores de Decisão (DT) e a posterior combinação de seus resultados. Como resultados do treinamento dos modelos de nosso EC1, armazenamos todos os valores extraídos pelas métricas e retornamos as previsões para o processo visando a reutilização das mesmas por outros modelos abordados em EC2 e EC3.

3.4.1.1 Resultados do Estudo de Caso 1

Ao adentrar na fase de Avaliação e Interpretação dos Resultados, utilizamos as bibliotecas Seaborn e Matplotlib para apresentar nossos resultados através de gráficos, melhorando a visualização e facilitando a captura de insights sobre eles. Para o EC1, visando comparar os modelos BAG com os modelos IND, ambos utilizando DT, observamos inicialmente na Figura 24 a comparação das médias destes dois tipos, agrupadas por níveis (NI, NC e NP) para cada curso, e mensurada por meio da métrica ACC.

Para cada um dos níveis, há duas barras agrupadas, sendo a primeira (em azul) correspondente à média das ACCs dos modelos IND e a segunda à média das ACCs dos modelos BAG (no caso de NI, existe apenas um modelo por Tipo, sendo então seus valores repetidos para todos os gráficos).

Por meio das médias é possível vislumbrar a superioridade dos modelos BAG sobre IND, tendo em vista que em apenas um caso (NC do curso CON) destoa dos demais sendo a ACC aferida individualmente com DT superior à ACC aferida com BAG. Este mesmo padrão de comportamento, inclusive no caso de exceção citado, ocorre quando a métrica utilizada é o MCC.

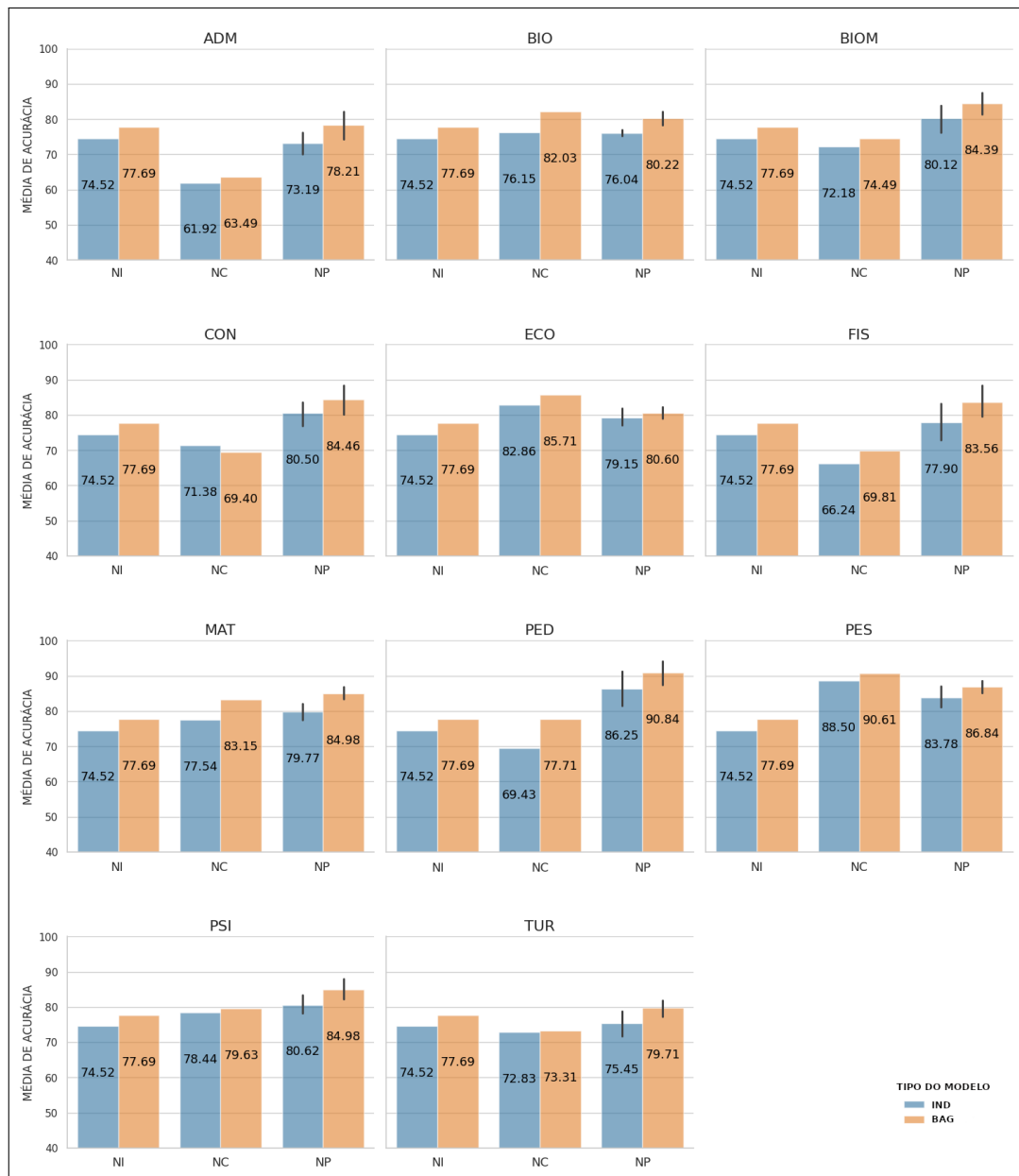


Figura 24 – Média de Acurácia por Nível estratificadas por curso e tipo do modelo

A Figura 25, contém gráficos de linha que comparam curso a curso, período a período os modelos NP das abordagens IND e BAG. Ao usar os modelos do NP para comparar as ACCs aferidas pelo IND (DT) individualmente (linha vermelha tracejada) e por BAG (linha verde contínua), podemos observar graficamente a vantagem dos modelos com ensemble em praticamente todos os períodos de todos os cursos. Utilizando o MCC, o comportamento vantajoso é praticamente idêntico para cada curso período (respeitadas as escalas e peculiaridades de cada métrica), sendo os modelos em que BAG está graficamente abaixo de IND pertencentes aos mesmos cursos e períodos para as duas métricas (basicamente CON P2, ECO P1 e ECO P5).

No gráfico, incluímos uma linha reta sólida (preta) para simular uma linha de tendência de percentual de ACC ao longo dos períodos para o modelo IND. Com base nisso,

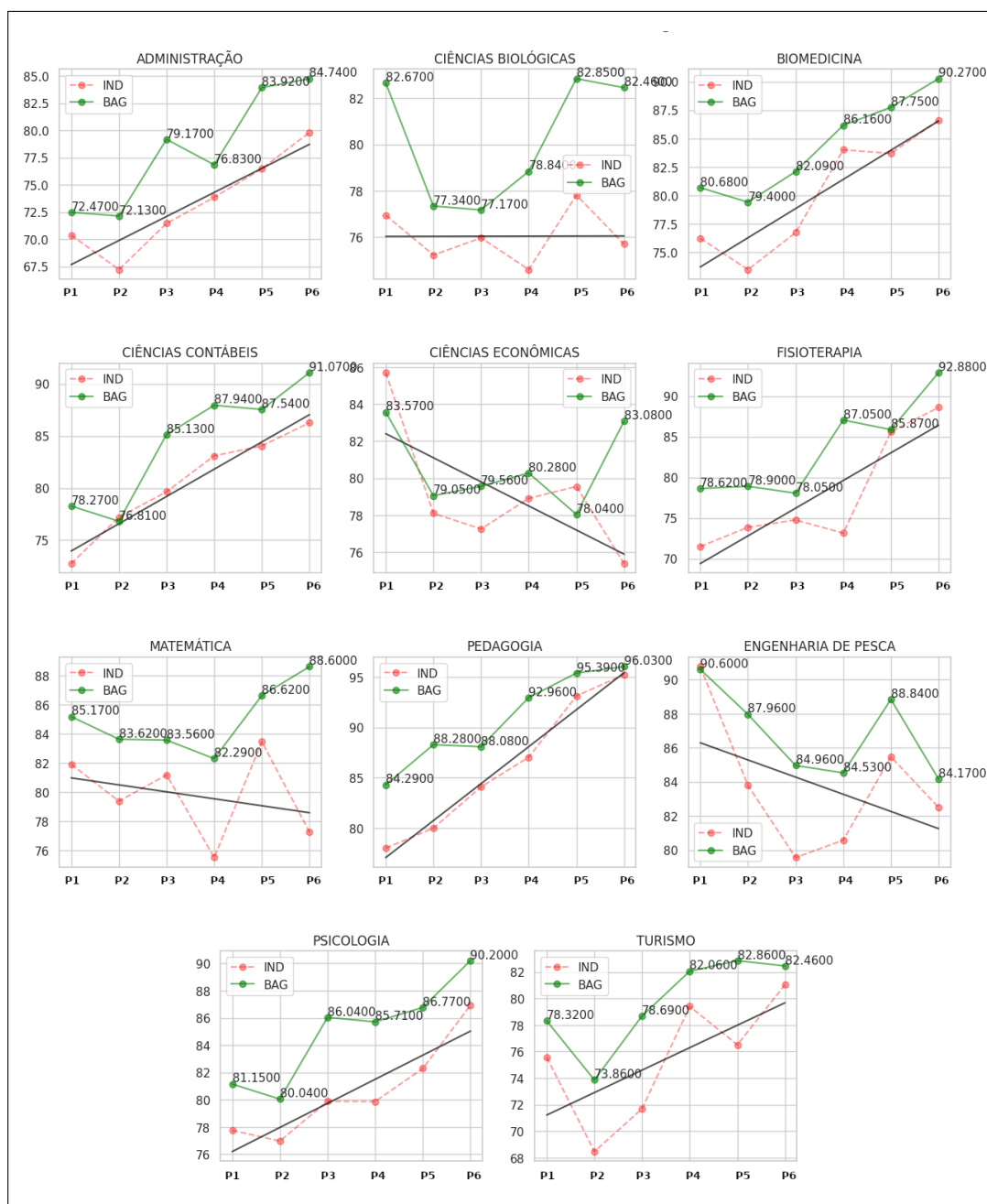


Figura 25 – Comparativo entre ACC de DT e BAG por curso e período

podemos observar cursos que não apresentam um padrão claro de variação ascendente das ACCs dos modelos ao longo dos períodos: ECO, PES, MAT e BIO. Esses cursos também possuem clara melhoria nas suas predições utilizando BAG e uma linha de tendência sobre eles teria clara melhora em sua orientação em relação à traçada sobre os modelos IND.

Acreditamos que esse comportamento irregular (percentual de ACCs sem uma tendência clara de crescimento nos modelos dos períodos mais avançados) seja causado por características não abordadas em nossos conjuntos de dados, embora seja necessária uma investigação mais granular. Essas características não abordadas teriam uma influência maior em determinados períodos do que em outros, não estando evidentes nos dados

acadêmicos cumulativos específicos dos modelos de nível NP.

É importante esclarecer que, embora o gráfico de linha e a linha de tendência sugiram uma 'continuidade', eles são apenas uma forma de visualização dos dados e não devem ser interpretados como 'contínuos'. Em outras palavras, para cada Curso/Período, há um conjunto de dados distinto, com características distintas, resultando em modelos totalmente independentes um do outro, não sendo real a ideia de continuidade.

Em seguida, prosseguimos realizando uma análise mais detalhada sobre as acurácias de ambos os tipos de modelos (IND e BAG). Embora os gráficos nos permitam perceber algumas diferenças visualmente, precisamos de uma medida estatística para determinar de fato se um nível ou modelo é estatisticamente inferior, igual ou superior a outro. Por esse motivo, aplicamos o Teste T de Student para Amostras Pareadas. Esse teste é frequentemente utilizado para comparar as médias de duas amostras relacionadas, geralmente antes e depois de uma intervenção. Nosso objetivo ao aplicar esse teste foi avaliar se existem diferenças estatisticamente significativas entre esses resultados.

Visando identificar padrões e situações merecedoras de análises futuras mais aprofundadas, utilizamos um índice de significância de 10%, assumindo uma probabilidade de erro um pouco maior neste momento, considerando que para o nosso nicho de pesquisa, um índice de significância inferior pode ser demasiadamente rigoroso em situações onde, na prática, observamos diferenças relevantes. Além disso, esses são apenas os primeiros resultados de nossa pesquisa, voltados à identificação de padrões e situações que exigem análises mais detalhadas no futuro. Um índice de significância mais elevado aumenta o risco de aceitar resultados como 'diferentes', mesmo que essas diferenças não sejam estatisticamente significativas. No entanto, o teste torna-se mais sensível a efeitos e diferenças menores entre os resultados.

A Tabela 2 apresenta os resultados do Teste de Hipóteses T sobre nossos resultados. Ela demonstra que nossa proposta é aplicável em 100% dos modelos por cursos/períodos analisados, pois os resultados para BAG são melhores (+1) ou estatisticamente equivalentes (0) aos modelos em que as predições foram feitas apenas com DT individualmente. Em quase metade dos casos (48,48%), constatamos que houve uma melhoria estatisticamente significativa. Assim como não houve nenhum caso em que a abordagem com BAG foi estatisticamente pior (-1). Nessa tabela (assim como nas próximas) a coluna MPX e APX correspondem ao percentual de melhoria e percentual de aplicabilidade calculados para cada período, respectivamente.

Com isso, acreditamos ter alcançado nosso objetivo neste primeiro EC, pois conseguimos desenvolver modelos com maior capacidade de predição de evasão por meio de abordagens com janelas semestrais utilizando Métodos Ensemble.

Tabela 2 – Resultados da comparação entre as abordagens DT e BAG utilizando Teste de Hipótese T pareado.

PX	ADM	BIO	BIOM	CON	ECO	FIS	MAT	PED	PES	PSI	TUR	MPX	APX
P1	0	+1	0	+1	0	0	+1	+1	+1	+1	+1	63,64%	100%
P2	0	+1	0	0	0	+1	+1	0	+1	0	0	36,36%	100%
P3	+1	0	0	+1	0	+1	0	0	0	+1	+1	45,45%	100%
P4	0	0	+1	+1	0	0	+1	+1	+1	+1	0	54,55%	100%
P5	+1	0	+1	+1	0	0	0	0	0	+1	+1	45,45%	100%
P6	+1	0	+1	0	+1	0	+1	0	0	+1	0	45,45%	100%

3.4.1.2 Observações complementares de EC1

Ainda analisando os dados gerados em nosso primeiro EC, identificamos uma aparente vantagem dos modelos criados com base nos datasets do nível NP em comparação aos outros níveis. O gráfico apresentado na Figura 26 evidencia essa vantagem ao analisar as médias de ACC dos modelos para cada nível, utilizando BAG com DT e o conjunto de datasets ORIG. A barra correspondente ao nível NI representa a ACC do modelo (único) NI; a correspondente a NC representa a médias das ACCs de cada um dos onze (11) cursos, e NP representa a média das ACCs dos sessenta e seis (66) modelos correspondentes aos seis (6) períodos dos onze (11) cursos.

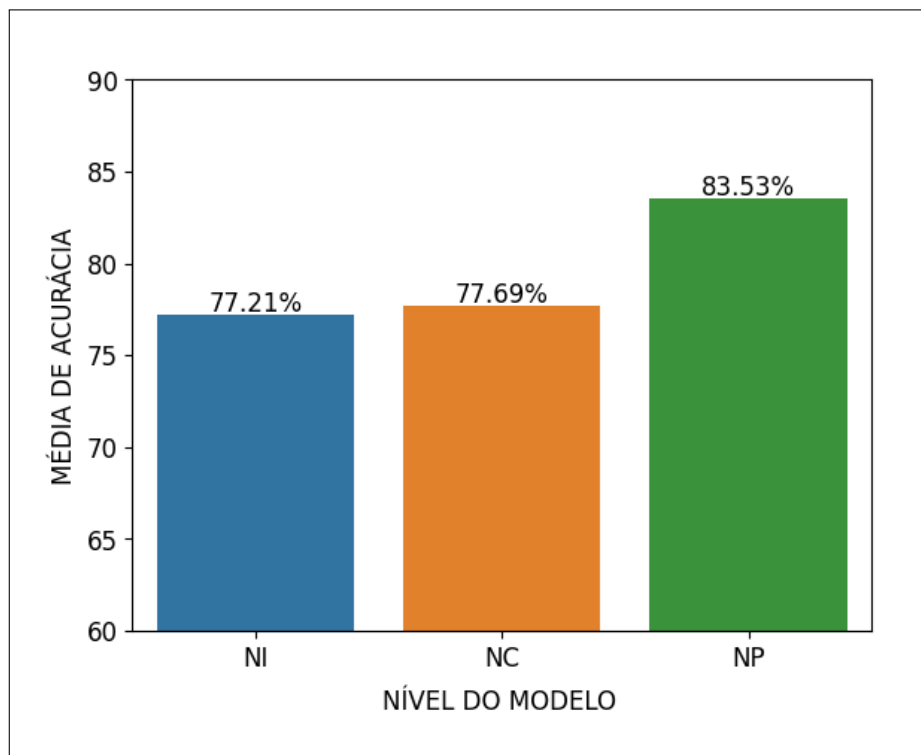


Figura 26 – Médias de Acurácia por Nível

Em todos os outros cenários e combinações de algoritmos, tipos de estratégia, métricas e conjuntos de dados utilizados ao longo deste trabalho, o comportamento é similar, com a média das ACCs de NP superior às demais, embora a diferença e os percentuais sejam variados.

Este comportamento evidencia que o valor médio da ACC é substancialmente superior no NP, mesmo com a menor quantidade de amostras dos modelos deste nível. Isso destaca a influência das características relacionadas aos dados acadêmicos cumulativos presentes em seus modelos. Para além das médias de todos os modelos, ao analisarmos os mesmos dados de forma mais detalhada no gráfico da Figura 27, podemos observar as ACCs agrupadas por períodos em um comportamento coerente com o do gráfico anterior.

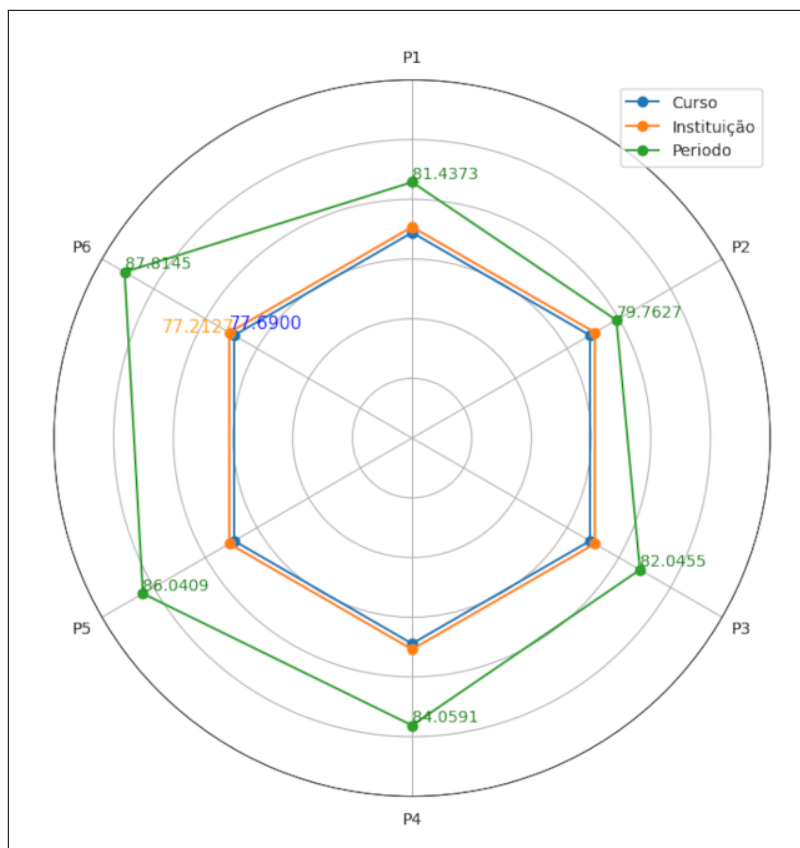


Figura 27 – Médias de Acurácia por Nível estratificadas por Período

Observa-se que a pequena superioridade do nível NP sobre os outros dois níveis observada em P2 evolui ao longo dos demais períodos, chegando a atingir uma diferença superior a dez pontos percentuais para NI e NC no último período analisado (P6). É importante lembrar que os modelos NI e NC não variam ao longo dos períodos, figurando no gráfico para fins de comparação. E frisar que também por este prisma, todos os outros cenários de combinações de algoritmos, tipos de estratégia, métricas e conjuntos de dados, o comportamento é similar a este gráfico.

Já analisando curso a curso, há muitas variações dentre as combinações possíveis, porém o comportamento dos cursos ECO e PES chama atenção pois, na maioria das combinações (de algoritmos, estratégias, datasets e métricas), encontram no nível NC seus melhores resultados em relação à médias de ACC, diferentemente dos demais cujo NP costuma ser superior na maioria dos casos (novamente podemos recorrer à Figura 24 para ilustração deste cenário).

Não por acaso, os supramencionados ECO e PES são dois dos quatro cursos salientados anteriormente como não tendo uma tendência clara de ascensão de ACC em seus modelos de Períodos ao longo dos períodos. Corroborando com a ideia de que, apesar de o nível NP possuir as maiores médias para acurácia na maioria dos casos, NC e NI podem ser muito úteis quando os modelos NP não conseguem identificar adequadamente os fatores de influência chave, ocasionando em falhas na predição. Com mais registros, outros padrões podem ser identificados auxiliando NC e NI em suas predições.

Para confirmar nossas impressões, aplicamos o Teste Estatístico T de Welch Independente, adequado para comparar diferenças de médias de dois grupos independentes, onde as distribuições possuem variâncias desiguais e quando o número de amostras em cada grupo é diferente. Aplicamos esse teste para verificar em que situações os modelos do NP se sobressaem ou se equivalem aos dos demais níveis. Ele avalia se existem diferenças significativas nas médias de ACC entre dois grupos. Como temos três grupos, realizamos comparações em pares entre os conjuntos de dados. Os resultados, que também foram aferidos com índice de significância de 10%, demonstram que apesar de conter dados sobrepostos, ou seja, dados de um mesmo discente podem estar em vários datasets distintos, os conjuntos de dados presentes em cada nível são estatisticamente diferentes.

Comparando os dados de forma agregada similarmente ao que foi feito na geração do gráfico da Figura 26, ou seja, com todos os dados de cada nível juntos durante uma mesma comparação, os resultados dos testes de Hipótese condizem com o que verificamos naquele gráfico. O NP mostra-se significantemente superior aos demais. Já para os valores de NI e NC, o teste informou que não há diferença estatística significativa entre os dados analisados. Porém analisando de forma mais granular, comparando os valores de NI com os de cada curso, NI mostra-se estatisticamente superior a NC em 5 cursos (ADM, CON, FIS, PED e TUR), equivalente em 2 (BIOM e BIO) e inferior em 4 (ECO, MAT, PES e PSI).

Embora não seja um objetivo claro de nossa pesquisa, esta análise é importante para atestar a distinção entre os níveis do estudo. Nossa hipótese é que essa distinção entre os resultados seja reflexo de dados também distintos o suficiente para gerar erros independentes e não correlacionados em nossos classificadores, possibilitando que a combinação dos modelos individuais resulte em modelos combinados mais eficazes. Através dessa análise podemos atestar a superioridade dos modelos do nível NP, e a influência dos dados acadêmicos cumulativos, ausentes nos demais níveis. Através destas observações, também verificamos que foi acertada nossa decisão por utilizar como um dos níveis do estudo o modelo de janelas semestrais (NP).

3.4.1.3 Estudo de Caso 2

Neste EC objetivamos implementar modelos através da combinação de bases de dados dos diferentes níveis criados, a serem treinados por um mesmo algoritmo-base, por meio do método Ensemble Stacking (STACK). Ainda durante o mesmo processo de criação, treinamento e teste dos modelos descrito no EC1, também introduzimos a abordagem STACK para combinação dos modelos IND.

As abordagens que utilizam ensembles para combinar conjuntos de dados distintos são pouco exploradas na área de MDE para predição de Evasão. O Ensemble Stacking é uma técnica amplamente difundida e encontrada em diversas pesquisas, de diversas áreas (DŽEROSKI; ŽENKO, 2004)(DIVINA et al., 2018)(DIB et al., 2020)(SANTOS, 2023) porém criando modelos através de algoritmos-base distintos, como uma forma de confrontar paradigmas de predição visando maior robustez na escolha. No entanto, não encontramos trabalhos sobre o tema que utilizem conjuntos de dados ‘funcionais’ e distintos para a criação de modelos visando um mesmo resultado de predição combinado.

Ao citar os conjuntos de dados como funcionais, estamos nos referindo à capacidade dos mesmos de individualmente gerarem modelos de classificação eficazes visando atender seus próprios objetivos. Em nosso caso, temos a combinação de modelos que por sua natureza podem ser úteis para públicos-alvo específicos de gestores educacionais e educadores. Os modelos de nível Instituição podem ser úteis, por exemplo, para a Administração Acadêmica da Instituição, fornecendo insights para ações a um nível mais ‘global’ de combate à evasão. Os modelos do nível Curso podem indicar a um Coordenador de Curso ou Colegiado de curso específico quais ações podem ser mais efetivas no combate à evasão diante das peculiaridades daquele grupo de estudantes. Do mesmo modo, o grupo de Docentes responsável pelos componentes ofertados em um determinado período, apoiados pela Coordenação e/ou Colegiado do curso podem propor ações mais específicas para o combate à evasão baseado nas descobertas evidenciadas pelo modelo daquele curso/período (ou Janela Semestral).

Como exemplos hipotéticos, o modelo da Instituição pode gerar à Pró-Reitoria de Graduação a informação que a Renda per capita e a Distância da cidade natal são os atributos mais influentes para evasão medida sobre o universo total de alunos, e permitirá direcionar ações como bolsas de estudo específicas para a permanência de alunos carentes que moram longe da cidade natal; Os modelos de nível Curso poderiam evidenciar ao Coordenador Acadêmico de PED que o atributo sexo é altamente influente na evasão e que a esmagadora maioria dos homens evadem, podendo direcionar ações de maior acolhimento aos homens nesse curso; Já nos modelos por Períodos do Curso, ciente que a quantidade de trancamento de disciplinas no 3º período seria o atributo mais influente e que 100% dos alunos que trancam a disciplina X quando lecionada pelo professor Y evadem, o colegiado do curso poderia sugerir a troca do docente ou da abordagem desta disciplina.

Como dito, são exemplos hipotéticos de possíveis utilidades de cada um dos níveis de modelo já abordados pela literatura, cada um deles úteis individualmente para determinados fins e por isso tratados aqui como ‘funcionais’. Com o uso deste termo também estamos diferenciando essa abordagem daquelas que utilizam subconjuntos de atributos, ou seja, pequenos conjuntos de características de um mesmo conjunto de dados. Este é o caso inclusive do funcionamento de variações específicas de Métodos Ensemble Bagging, o algoritmo Random Forest e o Extra Trees, que como já foi dito, criam várias árvores de Decisão a partir de subconjuntos aleatórios de características de um dataset original, e combinam os seus resultados ao final.

Retornando ao fluxo de criação dos nossos modelos, a partir dos modelos individuais (DT), realizamos a junção de um modelo de cada nível em um novo modelo combinado. Isso foi feito para cada período de cada curso, resultando em um total de 66 modelos STACK que, por sua vez, geraram 66 novas predições combinadas. Para esta junção, respeitamos obviamente as relações de cada discente com os respectivos modelos. Ou seja, um modelo NP não pode ser combinado a um modelo NC cujo curso dos discentes não seja o mesmo nos dois modelos.

Na prática, resumidamente, em uma estrutura de repetição aninhada, combinamos a predição de NI com a predição de NC de um primeiro curso avaliado, e com a predição de NP de P1 para o curso em questão. Em seguida, iteramos sobre os modelos de períodos e usamos o mesmo modelo de NI e NC para formarmos a próxima combinação com P2 do mesmo curso. Após 6 iterações, 6 modelos combinados para os 6 períodos de um determinado curso foram montados. Deste modo, iteramos o curso e passamos a montar modelos relacionados aos 6 períodos deste próximo curso. E assim sucessivamente até que obtivemos os 66 modelos STACK.

A título de exemplo, podemos mencionar a criação de um modelo STACK que resultou da combinação do modelo NI (presente em todas as combinações), um modelo NC referente ao curso de FIS e um modelo NP relacionado ao P1 do curso de FIS. Estes três modelos são capazes de individualmente predizer com acurácia satisfatória o target para um ingressante no curso de FIS. Ao combiná-los em um novo modelo, nosso objetivo foi superar os desafios comuns associados aos modelos ensemble (por sua capacidade de superá-los) e ampliar a eficácia das predições em comparação com os modelos individuais.

Para esta segunda fase da montagem dos modelos STACK (combinação), visando a redução de ruídos nos novos dados de entrada (ou seja, as predições da primeira fase e o target real), procedemos com a exclusão dos registros do novo conjunto de dados, cujo os três classificadores-base realizaram juntos erroneamente a mesma predição, ou seja, diferente do valor real do caso de Teste em questão (Figura 28). Posteriormente, particionamos esses dados em conjuntos de Treinamento e Teste na mesma proporção (70% e 30% respectivamente), treinamos os modelos e realizamos as predições.

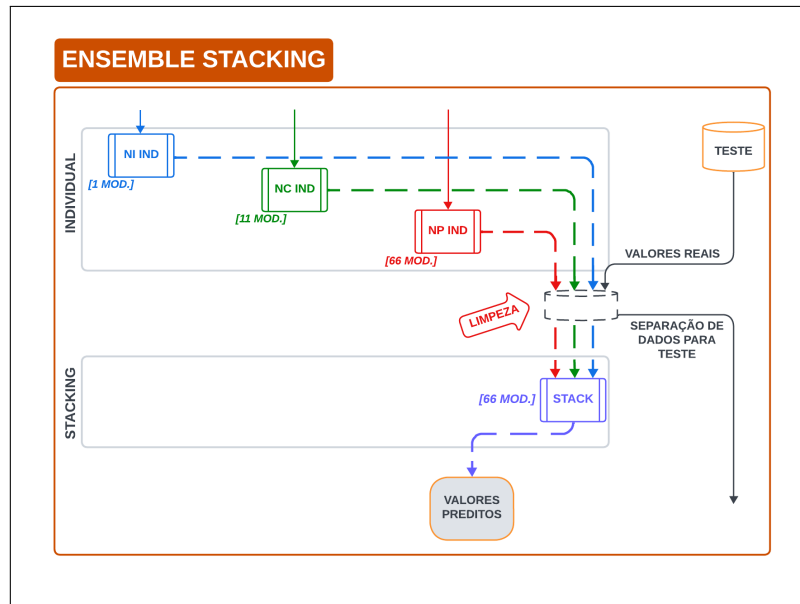


Figura 28 – Detalhamento do processo de montagem do modelo Stack

3.4.1.4 Resultados do Estudo de Caso 2

Como já explicado, os métodos ensemble geralmente envolvem o treinamento de vários preditores e a combinação das várias previsões em uma única. Eles empregam um conjunto de hipóteses e utilizam algum critério de escolha ou votação para determinar o resultado final, aumentando assim a probabilidade de que a decisão seja precisa (DIETTERICH, 2000). Logo, a ideia do aprendizado ensemble é empregar múltiplos modelos para combinar suas previsões. Nesse contexto, nossa proposta considera ainda a combinação das previsões de modelos gerados por múltiplos conjuntos de dados independentes, cada um capaz de compor modelos eficazes, ao invés de apenas variar os estimadores em um mesmo conjunto de treinamento. Também conforme já mencionado, quanto mais diversificados e diferentes forem os nossos preditores, maiores são as chances de melhorar a acurácia do nosso modelo combinado. Essa diversidade nos dados de entrada e nos estimadores de base, sempre que aplicável, pode ajudar a reduzir o overfitting e melhorar o desempenho geral do ensemble. Nossa estratégia visa aproveitar a diversidade implícita nas características temporais, no universo amostral e na independência entre os modelos NP, NC e NI.

Novamente, para a fase de Avaliação recorreremos à construção de gráficos, visando analisar os dados obtidos na fase de Mineração, onde as previsões dos modelos NI, NC e NP foram utilizadas para a formação de uma nova previsão combinada. O gráfico da Figura 29 agrupa os dados de acordo com a estratégia utilizada e exibe a média das ACCs obtidas em cada uma delas em barras no gráfico. Para as estratégias IND e BAG, filtramos apenas os resultados obtidos no NP, pois este nível demonstrou ser o mais eficaz ao longo de nossa pesquisa, e é também o mais equivalente ao STACK. Isso porque embora se utilize da previsão de outros níveis (NI e NC) para sua formação, o STACK é composto de uma nova combinação (modelo) por período, podendo ser assim melhor comparado com o NP.

Sendo assim, o gráfico compara 66 ACCs de cada estratégia, uma por período. As médias do resultados de BAG e STACK são muito próximas, ambas bem superiores a IND.

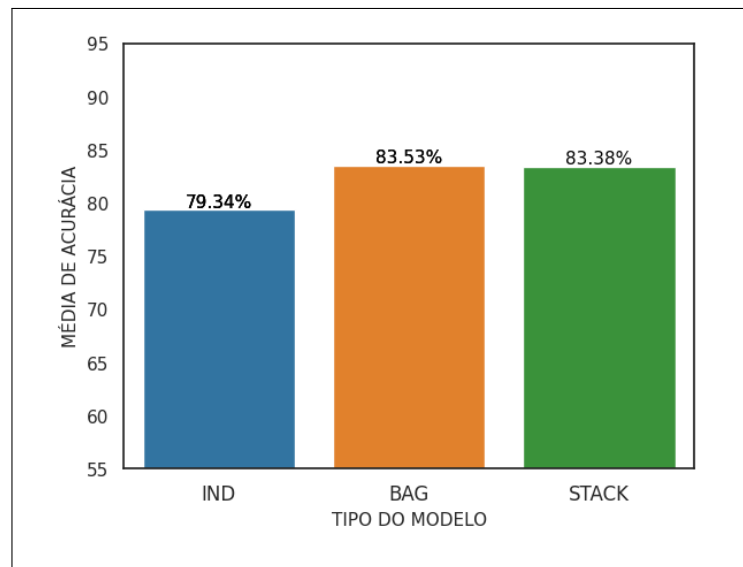


Figura 29 – Médias de Acurácia por Tipo do Modelo (DT, BAG e STACK.)

Já na Figura 30, os valores de acurácia foram estratificados por períodos, com cada barra de cada gráfico correspondendo à média das acurácias de todos os cursos para aquele período específico, em sua respectiva estratégia. Graficamente, podemos avaliar que as médias das acurácias dos modelos STACK ficaram muito próximas (algumas acima e algumas abaixo) às médias de BAG quando avaliamos com o conjunto de datasets ORIG e o algoritmo DT.

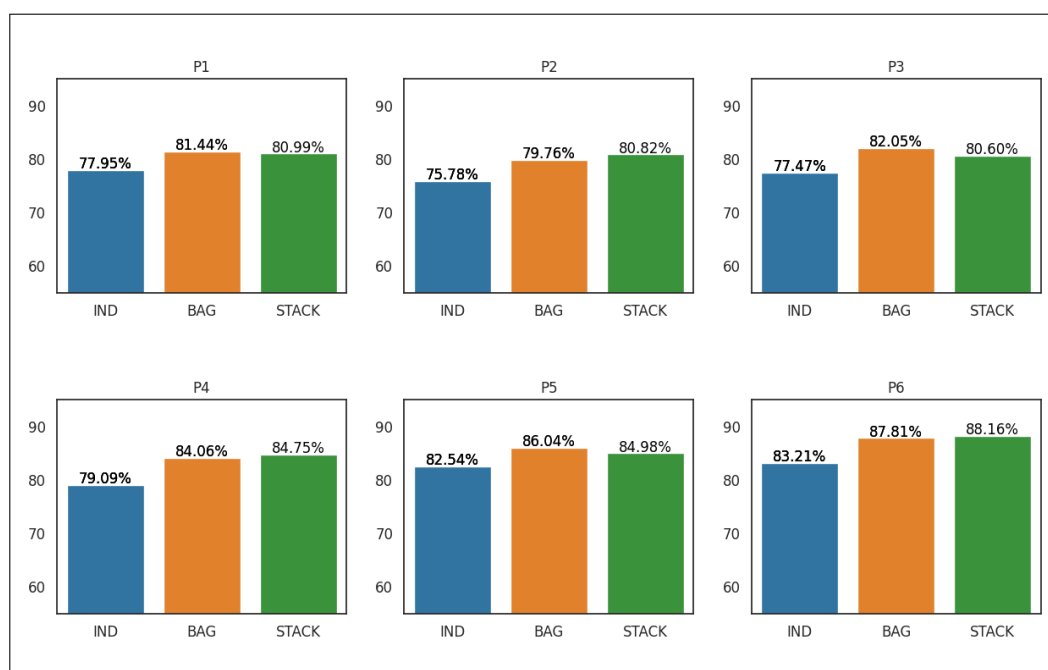


Figura 30 – Médias de Acurácia por Tipo do Modelo estratificadas por período

Em seguida, a Figura 31 exibe a estratificação por curso, apresentando os valores exatos das acurácias medidas para as abordagens IND e STACK. Para melhor legibilidade, optamos por plotar somente os dados da linha que, em geral, obteve melhores resultados (STACK) ao longo dos pontos da linha.

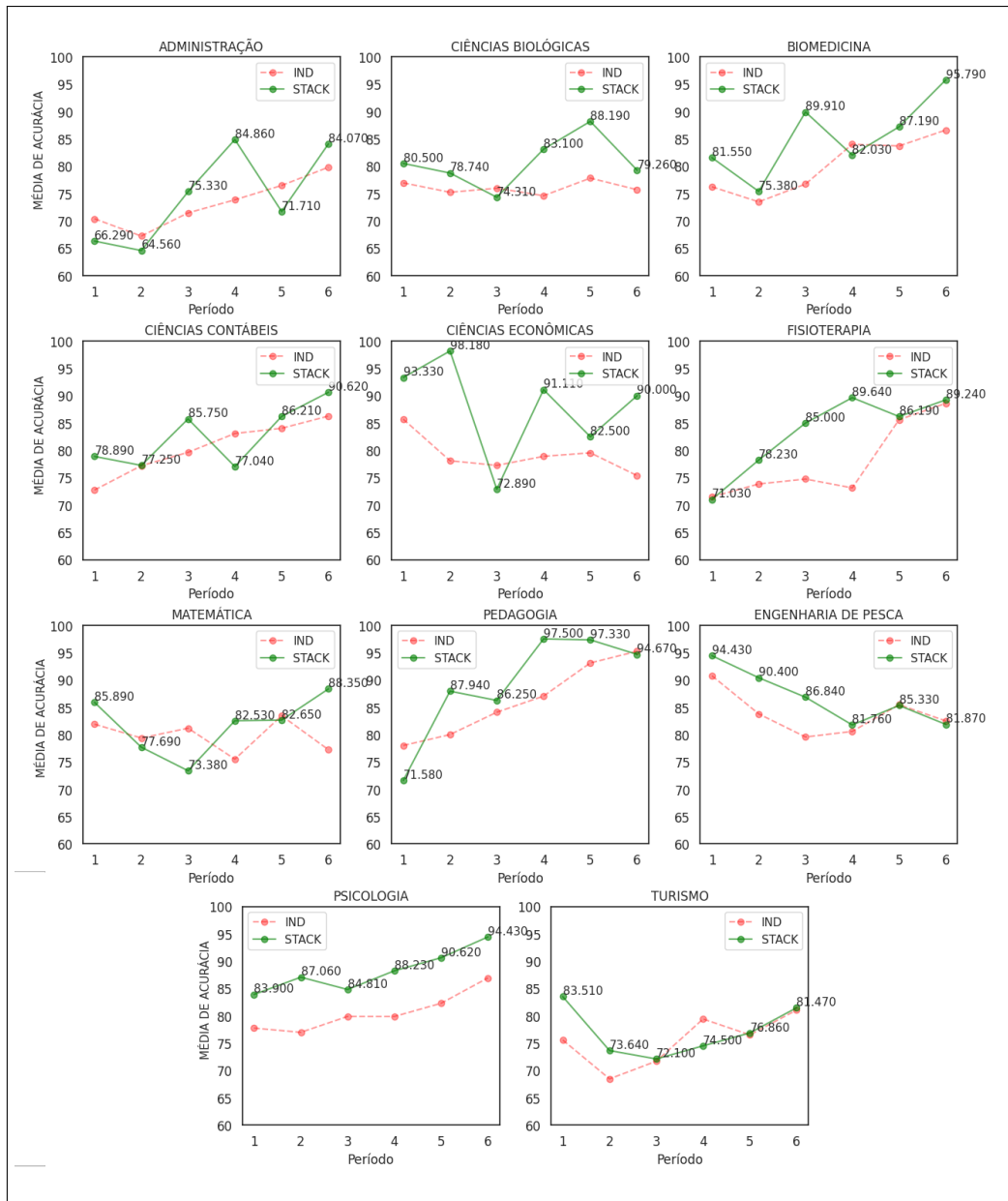


Figura 31 – Médias de Acurácia por Tipo do Modelo (IND e STACK) estratificadas por curso e período.

Estes resultados estão alinhados com nosso objetivo em EC2, onde conseguimos implementar a estratégia de ensemble utilizando modelos formados por conjuntos de dados independentes e um mesmo algoritmo classificador. Esta abordagem se mostrou viável e mais eficaz na predição de evasão do que os modelos isolados. Os resultados do teste de hipótese apresentados na Tabela 3 confirmam esse entendimento ao demonstrar estatisticamente (com um nível de significância de 10%) que 92,43% dos modelos preditos através

do Ensemble Stacking, utilizando DT e o conjunto de dados ORIG, são estatisticamente melhores ou equivalentes aos modelos IND correspondentes. Em 33,33% dos casos, houve melhoria estatisticamente significativa. Apenas em cinco das sessenta e seis comparações, o STACK mostrou-se estatisticamente inferior em termos de acurácia (e em 7 modelos em termos de MCC). Dentre eles, modelos cujas predições já apresentavam dificuldades no EC1. Além disso, os cursos de ECO e PES apresentaram resultados idênticos tanto ao realizar o teste T sobre os dados de acurácia quanto de MCC, reforçando a suspeita de que os resultados inferiores não são necessariamente atribuíveis à estratégia avaliada. Isso nos leva a considerar que outros elementos podem ter impactado o desempenho dos modelos.

Tabela 3 – Resultados da comparação entre as abordagens IND (DT) e STACK (DT) utilizando Teste de Hipótese T Pareado.

PX	ADM	BIO	BIOM	CON	ECO	FIS	MAT	PED	PES	PSI	TUR	MPX	APX
P1	0	0	0	0	+1	0	0	+1	0	+1	0	27,27%	100%
P2	0	0	0	0	+1	+1	0	0	+1	+1	0	36,36%	100%
P3	0	+1	0	0	0	+1	0	-1	0	+1	0	27,27%	90,91%
P4	+1	0	+1	0	+1	0	0	+1	+1	+1	0	54,55%	100%
P5	-1	0	+1	0	-1	0	0	0	+1	+1	0	27,27%	81,82%
P6	0	+1	0	0	+1	-1	0	0	-1	+1	0	27,27%	81,82%

3.4.1.5 Estudo de Caso 3

Em nosso terceiro Estudo de Caso, objetivamos expandir os experimentos anteriores por meio da introdução de novos conjuntos de features, técnicas de ensemble adicionais e do uso de diferentes algoritmos. Essas expansões resultaram em um aumento considerável no número de modelos criados através da combinação desses métodos e algoritmos, aplicados sobre dados distintos e avaliados pelas métricas de ACC e MCC. A análise dessa metodologia nos revela quais dessas combinações resultam em um quantitativo maior de modelos estatisticamente superiores aos modelos individuais.

Assim como nas tabelas 2 e 3, que nos permitem extrair o percentual de modelos superiores dentro das estratégias BAG e STACK, aprimoramos nosso processo de construção de modelos para que fosse possível gerar dados similares para todas as combinações utilizadas ao longo deste trabalho. Em uma nova rodada de treinamentos, incrementamos a estrutura utilizada nos EC1 e EC2 com o conjunto de datasets FILT (cujo atributos de cada nível constam na Tabela 4) e a adição dos algoritmos RF e ET (Figura 32).

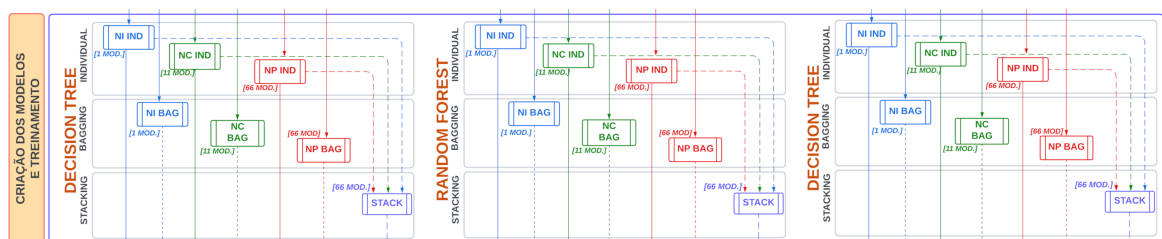


Figura 32 – Detalhamento da inserção de RF e ET na fase de criação dos modelos

Tabela 4 – Features presentes em cada um dos níveis de cada conjunto de *Datasets* da proposta

FEATURE	ORIG			FILT			TIPO
	NI	NC	NP	NI	NC	NP	
evadido	x	x	x	x	x	x	Booleano
id_discente	x	x	x	x	x	x	Numérico
concluiu_escola_publica	x	x	x				Booleano
convocacao	x	x	x				Numérico
curso	x						Numérico
cursou_na_pandemia	x	x	x		x		Booleano
estadocivil	x	x	x				Numérico
idade_ingresso	x	x	x		x	x	Numérico
intervalo_EM_ingresso	x	x	x		x	x	Numérico
nota_ciencias_humanas	x	x	x	x	x	x	Numérico
nota_ciencias_natureza	x	x	x	x	x	x	Numérico
nota_linguagem	x	x	x	x	x	x	Numérico
nota_matematica	x	x	x	x	x	x	Numérico
nota_redacao	x	x	x	x	x	x	Numérico
opcao_inscricao	x	x	x				Numérico
periodo_ingresso	x	x	x				Numérico
qtde_membros_familia	x	x	x		x		Numérico
raca	x	x	x				Numérico
sexo	x	x	x				Numérico
tipoacaoafirmativa	x	x	x		x		Numérico
temcredito_cc_ca	x	x	x		x		Booleano
tem_ensino	x	x	x	x	x	x	Booleano
turno	x	x	x				Numérico
uf_naturalidade	x	x	x				Numérico
tranca_programa_PX			x				Numérico
qtd_reprova_ft_nt_PX			x			x	Numérico
qtd_reprova_ft_PX			x				Numérico
qtd_reprova_nt_PX			x			x	Numérico
qtd_tranca_turma_PX			x			x	Numérico
qtd_cancela_turma_PX			x				Numérico
media_acumulada_PX			x			x	Numérico
rep_ob_semestre_PX			x			x	Numérico
ap_ob_semestre_PX			x			x	Numérico
ap_no_prazo_PX			x			x	Numérico

Enquanto nos experimentos anteriores realizamos as etapas delineadas na Figura 15 uma única vez, com as atualizações, agora duplicamos as bases e triplicamos os algoritmos envolvidos. Como esperado, observamos um alto impacto no tempo de máquina, ou seja, o tempo de execução de nosso processo aumentou exponencialmente. Para mitigar esse efeito, optamos por ajustar os parâmetros exclusivamente para os algoritmos RF e ET. Reduzimos o número de estimadores (`n_estimators`) de 100 para 10 e limitamos a profundidade máxima das árvores (`max_depth`) a 10, ao invés de deixá-la não limitada como é o padrão.

Para fins de registro, os modelos com FILT reduziram seu tempo de execução em aproximadamente 10% em relação aos correspondentes com ORIG. Já a alteração dos parâmetros de RF e ET tornou possível o treinamento dos modelos que anteriormente extrapolavam o uso de recursos disponíveis no ambiente do Google Colab.

Adicionalmente, incorporamos mais três estratégias aos tipos de ensembles já testados: Votação Majoritária (VOT), Votação Majoritária baseada nas previsões de Bagging (VOT BAG) e Stacking com Bagging como modelo-base (STACK BAG). A última estratégia consiste em aplicar a abordagem do EC2 (STACK), utilizando para seu treinamento, modelos e previsões do EC1 (BAG). Da mesma forma, VOT e VOT BAG realizam as previsões por meio da estratégia de Votação Majoritária sobre as previsões de IND e BAG respectivamente. A Figura 33 detalha essas estratégias. Com essas adições, temos agora um total de seis tipos de modelos para comparações (IND, BAG, STACK, VOT, VOT BAG e STACK BAG).

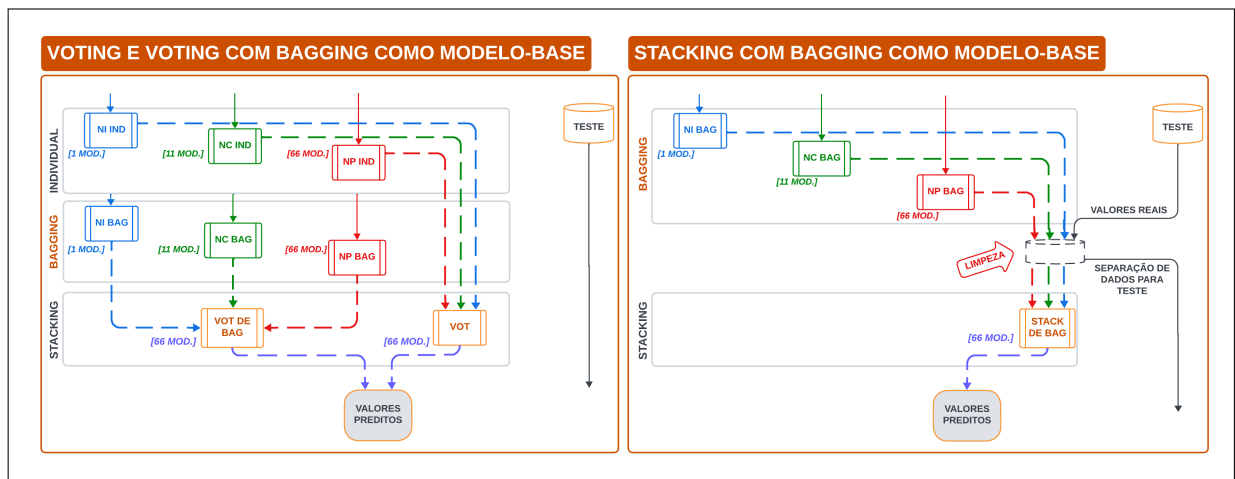


Figura 33 – Detalhamento do processo de montagem específico dos modelos VOT, VOT BAG e STACK BAG.

Embora não necessariamente tragam os melhores resultados em termos de métricas, os novos algoritmos introduzidos no EC3 foram selecionados, entre outros motivos, pela facilidade de operação, pertencendo ao mesmo paradigma (simbólico) e à família de algoritmos de árvore de decisão. Isso garante que a preparação dos dados realizada com base em Decision Trees esteja adequada também aos novos algoritmos. Essa uniformidade simplificou ainda a construção de um conjunto de datasets adicional (FILT) uma vez que esses modelos possuem nativamente a característica de apontar as features mais influentes na predição. O que corrobora ainda com a compreensão dos resultados. Além disso, a diversificação proporcionada pelo fato de esses algoritmos fazerem parte da família de métodos ensemble, juntamente com as estratégias ensemble adotadas ao longo do trabalho, tende a conferir maior robustez e potencializar o desempenho dos modelos, resultando em ganhos adicionais.

Assim, em uma única execução, essa nova abordagem é capaz de empregar as métricas de ACC e MCC na mensuração dos resultados das predições para os modelos dos tipos IND, BAG, VOT, STACK, VOT BAG e STACK BAG, para cada um dos conjuntos de datasets criados (ORIG e FILT) e para cada um dos algoritmos listados (DT, ET e RF), totalizando 72 combinações. É relevante ressaltar que neste documento, os modelos IND identificaram até então apenas os modelos treinados individualmente por DT. Doravante, o termo IND também identificará os modelos treinados por ET e RF quando treinados individualmente, ou quando servirem como base para as outras estratégias (por exemplo, um modelo IND treinado com DT ou um modelo STACK cujo estimador-base IND utiliza RF).

Por fim, utilizando a mesma metodologia de compilação dos resultados dos Testes T dos estudos de caso anteriores, calculamos o Percentual de Melhoria de cada Combinação referente a modelos ensemble sobre a correspondente Combinação de modelos IND (percentual de "+1" sobre o total de registros comparados), bem como o Percentual de Aplicabilidade (percentual de "+1" ou "0" sobre o total). A Tabela 5 exemplifica esses resultados para os índices resultantes da comparação entre ACCs de BAG e IND, combinados com FILT e DT.

Tabela 5 – Percentual de Melhoria e Aplicabilidade de BAG sobre IND mensurados pela ACC para o conjunto de 66 modelos com estimador-base DT usando FILT.

PX	ADM	BIO	BIOM	CON	ECO	FIS	MAT	PED	PES	PSI	TUR	MPX	APX
P1	0	+1	0	0	0	0	+1	+1	+1	+1	0	45,45%	100%
P2	0	+1	+1	+1	0	0	+1	+1	+1	+1	0	63,64%	100%
P3	+1	0	+1	+1	0	+1	+1	+1	+1	+1	0	72,73%	100%
P4	+1	+1	+1	+1	0	0	0	0	0	+1	0	45,45%	100%
P5	+1	+1	+1	+1	0	0	+1	0	0	+1	+1	63,64%	100%
P6	0	0	0	+1	0	0	+1	0	0	+1	0	27,27%	100%

3.4.1.6 Resultados do Estudo de Caso 3

Calculados os Percentuais de Melhoria e Aplicabilidade para todos os modelos, podemos observar algumas situações interessantes, como o fato de que todos os modelos que resultam em 100% de Aplicabilidade, ou seja, aqueles que são estatisticamente melhores ou pelo menos equivalentes aos respectivos modelos IND, são do tipo BAG com DT ou BAG com ET. Esses casos envolvem quatro combinações, destacadas na Tabela 6), cada uma correspondendo a 66 modelos. Repare ainda que a primeira linha desta tabela corresponde à toda Tabela 5, que exibe período a período da combinação composta por FILT, ACC, DT e BAG.

Outra observação interessante relacionada aos modelos BAG é que todas as combinações envolvendo DT apresentaram melhores índices de melhoria (53,03%, 42,42%, 48,48%, 36,36%) em relação às combinações envolvendo os demais algoritmos, independentemente do conjunto de atributos e da métrica analisada. Além disso, todas as demais

Tabela 6 – Combinações relacionadas aos modelos Bagging.

DATASETS	MÉTRICA	CLASSIFICADOR	COMPARAÇÃO	% MELHORIA	% APLICABIL.
FILT	ACC	DT	BAG x IND	53,03%	100%
FILT	ACC	ET	BAG x IND	10,60%	100%
FILT	ACC	RF	BAG x IND	19,69%	96,97%
FILT	MCC	DT	BAG x IND	42,42%	96,97%
FILT	MCC	ET	BAG x IND	13,63%	100%
FILT	MCC	RF	BAG x IND	19,69%	95,45%
ORIG	ACC	DT	BAG x IND	48,48%	100%
ORIG	ACC	ET	BAG x IND	13,63%	95,45%
ORIG	ACC	RF	BAG x IND	16,66%	96,97%
ORIG	MCC	DT	BAG x IND	36,36%	98,48%
ORIG	MCC	ET	BAG x IND	15,15%	96,97%
ORIG	MCC	RF	BAG x IND	16,66%	96,97%

combinações tiveram menos de 20% de melhoria. Isso significa que ao contrário de DT, os algoritmos ET e RF construíram modelos BAG que, em sua grande maioria, não foram significativamente superiores aos IND. A inserção destes algoritmos não superou então nossos experimentos em EC1. Já em relação ao FILT, verificamos que ele resultou nos dois maiores percentuais dentre os DT mencionados em relação aos modelos que utilizaram ORIG (53,03% contra 48,48% para ACC e 42,42% contra 36,36% para MCC). A tabela completa com todos os valores pode ser acessada no Apêndice A.

Para os modelos VOT, a combinação mais bem-sucedida em relação à Melhoria teve menos de 23% dos modelos do grupo, utilizando DT, o conjunto FILT e a métrica MCC. Não houve clara superioridade entre algoritmos ou conjuntos de datasets deste tipo de ensemble, assim como para as combinações envolvendo VOT BAG, apesar de um percentual de Melhoria superior (40,91%) na combinação que também utilizou DT, FILT e MCC.

A propósito, todas as combinações deste estudo cujo Percentual de Melhoria foi inferior a 25% pertencem às combinações que envolveram os métodos mencionados até aqui nesta seção (BAG, VOT e VOT BAG); Em outras palavras, embora possuam algumas combinações interessantes, esses três tipos estão associados aos piores Percentuais de Melhoria do estudo.

Dentre os modelos STACK, a combinação mais bem sucedida (57,58% de melhoria) foi aferida com MCC, utilizando FILT e ET. Observamos que o conjunto de datasets FILT teve um desempenho melhor (em relação ao Percentual de Melhoria) em todas as comparações com seus equivalentes que utilizaram o ORIG. Em relação aos algoritmos, o ET produziu os melhores índices de melhoria, independentemente do conjunto de atributos e da métrica analisada, variando de 46,97% a 57,58% de melhoria e de 86,36% a 90,91% de aplicabilidade.

Quanto aos modelos do tipo STACK BAG, observamos inicialmente que esses são os modelos com os Percentuais de Melhoria mais elevados entre todos os outros tipos (de 43,93% a 68,18%), com Percentual de Aplicabilidade variando de 86,36% a 96,97%. Os

algoritmos DT e ET estão associados aos percentuais de melhoria mais altos em duas combinações cada um, sendo o DT melhor nas combinações aferidas por ACC e ET sendo melhor nas que foram aferidas por MCC. Em relação aos datasets, o ORIG produziu os melhores percentuais em relação às combinações semelhantes que utilizaram o FILT, tanto em ACC quanto em MCC, sendo o melhor de todos esses o de 68%, obtido da combinação do conjunto de atributos ORIG, da métrica MCC e do estimador-base ET.

Em suma, a aplicação da técnica STACK BAG sobre modelos formados a partir do conjunto de atributos ORIG e utilizando ET como classificador-base dos modelos BAG, que por sua vez serviram como modelos-base para um meta-classificador Stacking, produziu a maior quantidade de modelos estatisticamente superiores aos de nível IND em nosso estudo, segundo a métrica MCC.

Em termos de percentual de MCC, a Figura 34 (na página seguinte) exibe então uma comparação entre os modelos IND com ET (linha pontilhada vermelha) e os modelos pertencentes à melhor combinação, no tocante ao Percentual de Melhoria, identificada em nosso estudo (linha sólida azul). Em uma linha sólida verde com marcações (triângulos) exibimos ainda, como parâmetro, os valores de MCC para STACK BAG com DT (ambos utilizando ORIG). Observa-se os valores para STACK BAG (azul) substancialmente superiores para a maioria dos cursos/períodos sobre os valores dos modelos IND (vermelho). Em apenas 2 modelos o MCC indicou uma predição próxima de uma escolha aleatória (ADM P6 e BIOM P5) e em 5 modelos indicou índice próximo a 1, apontando que os modelos estão realizando predições corretamente na maioria das vezes.

Ao analisar a nível de Períodos (Tabela 7), verificamos que o percentual de aplicabilidade para P1, P2 e P3 é de 100%, ou seja, nenhum dos modelos STACK BAG ET desta combinação é estatisticamente inferior aos IND correspondentes. Esse percentual no entanto decai em P4 e P5 (72,73% em ambos), períodos onde pode-se observar as maiores oscilações entre os valores das predições em todo o estudo, bem como a maior quantidade de modelos ensemble inferiores aos respectivos IND. Destes, o P5 possui modelos de apenas cinco cursos que são estatisticamente superiores aos IND (com outros três equivalentes e três inferiores), tendo o menor Percentual de Melhoria desta combinação (45,45%). Porém os três primeiros períodos apresentam excelentes Percentuais de Melhoria (72,73%, 81,82% e 90,91% respectivamente). Ficam nítidos os ganhos dessa estratégia, quando constatamos por exemplo que para 27 modelos os resultados foram significativamente melhores dentre os 33 correspondentes aos 3 primeiros períodos (mais de 80% de melhoria neste recorte). E ainda sem nenhuma perda (100% de aplicabilidade).

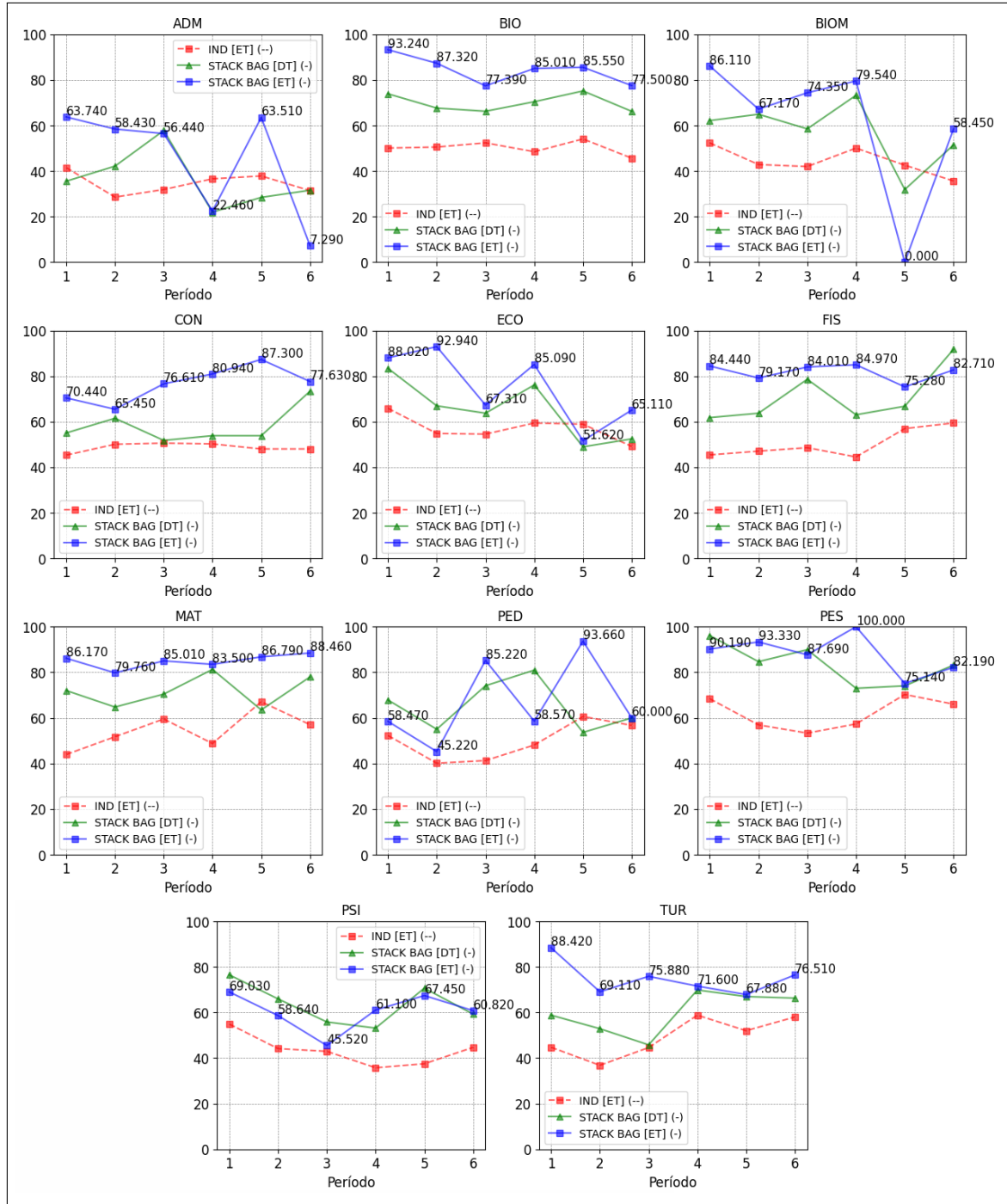


Figura 34 – Comparação por curso e período do MCC de STACK BAG com DT e STACK BAG com ET

Por fim, ao realizar uma análise final focada nos períodos para obter outras possíveis conclusões, verificamos que a abordagem STACK BAG obteve ainda os melhores resultados em termos de Percentual de Melhoria, tanto aferido por ACC quanto por MCC, em todos os períodos (com uma ocorrência de empate com BAG em P3). A Tabela 8 elenca as combinações com os maiores percentuais para ambas as métricas, agrupadas por período, corroborando com essa informação (para os casos de empate em uma mesma métrica ambas as combinações foram exibidas).

Tabela 7 – Percentual de Melhoria e Aplicabilidade de STACK BAG sobre IND mensurados por MCC para o conjunto de 66 modelos com estimador-base ET usando ORIG.

PX	ADM	BIO	BIOM	CON	ECO	FIS	MAT	PED	PES	PSI	TUR	MPX	APX
P1	0	+1	+1	+1	+1	+1	+1	0	+1	0	+1	72,73%	100%
P2	0	+1	+1	+1	+1	+1	+1	+1	+1	0	+1	81,82%	100%
P3	0	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	90,91%	100%
P4	-1	+1	+1	+1	+1	+1	+1	-1	+1	-1	0	63,64%	72,73%
P5	0	+1	+1	+1	-1	+1	+1	-1	0	0	-1	45,45%	72,73%
P6	-1	0	+1	+1	0	+1	+1	+1	0	+1	0	54,55%	90,91%

Tabela 8 – Maiores percentuais de Melhoria por período para cada métrica, entre todos os tipos de modelo.

PX	DATASETS	MÉTRICA	CLASSIFICADOR	COMPARAÇÃO	% MELHORIA
P1	FILT	ACC	RF	STACK BAG	81,82%
	FILT	MCC	RF	STACK BAG	72,73%
	ORIG	MCC	ET	STACK BAG	72,73%
P2	FILT	ACC	DT	STACK BAG	72,73%
	ORIG	ACC	DT	STACK BAG	72,73%
	ORIG	MCC	ET	STACK BAG	81,82%
P3	ORIG	ACC	ET	STACK BAG	72,73%
	FILT	ACC	DT	BAG	72,73%
	ORIG	MCC	ET	STACK BAG	90,91%
P4	ORIG	ACC	DT	STACK BAG	81,82%
	ORIG	MCC	DT	STACK BAG	90,91%
P5	ORIG	ACC	RF	STACK BAG	81,82%
	FILT	MCC	ET	STACK BAG	81,82%
P6	FILT	ACC	DT	STACK BAG	72,73%
	FILT	MCC	DT	STACK BAG	81,82%

Além disso, a Figura 35 exibe, na forma de um gráfico de BoxPlot agrupado, todas as ACCs aferidas (de todos os cursos com todos os algoritmos e conjuntos de datasets) para cada período e agrupadas por tipo de modelo, sugerindo um comportamento muito similar entre os períodos em relação à superioridade dos percentuais de ACC entre um tipo e outro, com destaque para o tipo STACK BAG (caixa mais à direita). Em P4 e P6 podemos inclusive observar que um ou mais modelos, em uma determinada combinação de STACK BAG, chegaram a atingir 100% de acurácia.

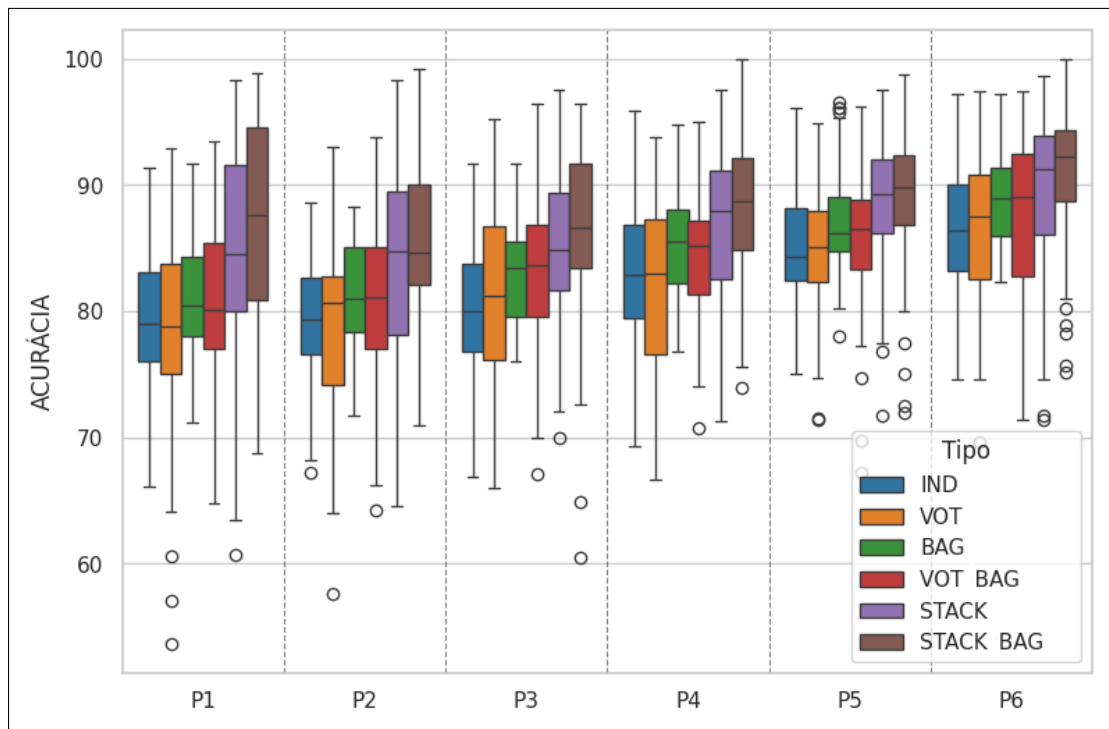


Figura 35 – Gráfico Boxplot das Acurácias de cada estratégia agrupadas por período, contendo dados de todas as combinações.

Este mesmo padrão de destaque para STACK BAG se mantém quando analisamos as combinações de forma individualizada, embora haja alternâncias de superioridade entre as outras métricas (com VOT passando BAG em quatro períodos para a combinação ACC+ET+ORIG, por exemplo). A Figura 36 exibe o mesmo gráfico, porém com os dados específicos da combinação com maior percentual de melhoria do estudo (MCC+ET+ORIG).

3.5 Considerações do Capítulo

Neste capítulo, detalhamos as estratégias empregadas na construção de modelos de aprendizado de máquina, com foco no uso de métodos ensemble. Exploramos o processo de Descoberta de Conhecimento em Bases de Dados (DCBD) desde a obtenção inicial dos dados até a sua preparação e transformação em conjuntos de datasets, organizados em diferentes níveis de granularidade, adequados para predições em variados contextos.

Realizamos três estudos de caso distintos, cada um abordando cenários específicos de investigação. Iniciamos com a confirmação da eficácia dos métodos ensemble, usando o bagging, na maximização dos resultados de predição de evasão, especialmente quando aplicados em uma abordagem de janelas semestrais. Em seguida, exploramos a combinação de modelos com heterogeneidade nos dados, representando diferentes níveis de dados, por meio da técnica de stacking, resultando em um modelo único e robusto.

Por fim, buscamos identificar quais combinações entre algoritmos, métodos e tipos

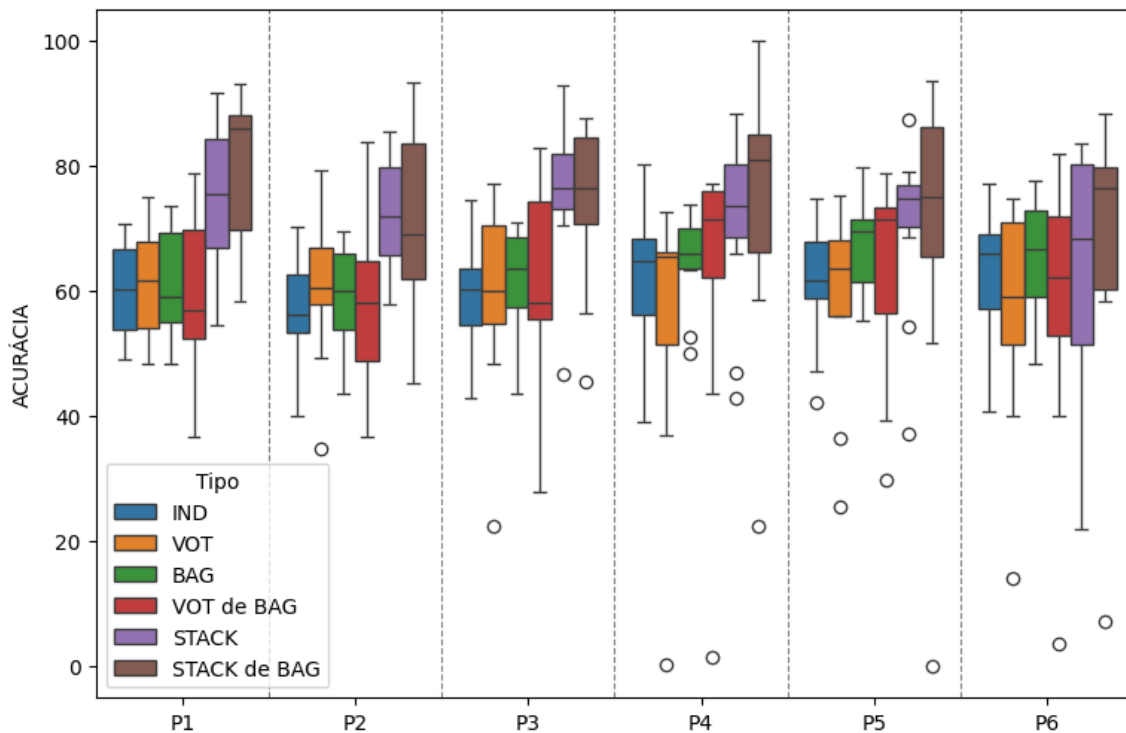


Figura 36 – Gráfico Boxplot dos valores de MCC de cada estratégia agrupados por período, contendo dados da combinação com ORIG e ET.

de ensemble produziram a maior quantidade de modelos significativamente superiores aos modelos treinados individualmente (Percentual de Melhoria). Destacamos que a combinação dos métodos Stacking e Bagging (STACK BAG) resultou nos maiores ganhos percentuais, aproveitando a capacidade do Bagging de lidar com dados desbalanceados, reduzir a variância e eliminar o overfitting; e do stacking de potencializar os resultados ao permitir que modelos diferentes aprendam com as fraquezas uns dos outros, capturando relações mais complexas nos dados, aumentando a capacidade preditiva do ensemble.

4 Conclusões e Trabalhos Futuros

Conclusões

Neste estudo, exploramos a aplicação de métodos ensemble na construção de modelos de predição de evasão em uma instituição federal de ensino superior, situando nossa pesquisa na área de Mineração de Dados Educacionais. Ao longo deste trabalho, desenvolvemos um total de 2520 modelos com base em 156 datasets distintos, derivados da base de dados original da instituição.

Adotamos um processo completo de Descoberta de Conhecimento em Bancos de Dados (KDD), abrangendo as etapas de Coleta, Limpeza, Transformação, Mineração e Análise dos Resultados. O foco principal recaiu na etapa de Mineração, em que empregamos um algoritmo/estimador-base uniforme para todos os modelos, em todos os casos de uso.

Estruturamos os conjuntos de dados em 3 níveis de especialização: 1 dataset para o Nível Instituição, 11 para o Nível Curso (1 para cada curso) e 66 para o Nível Período (referentes aos seis primeiros períodos de cada um dos 11 cursos). A partir deles, criamos 78 modelos baseados no algoritmo Decision Tree e outros 78 utilizando a estratégia Ensemble Bagging. Posteriormente, para cada modelo do Nível Período baseado em Decision Tree, criamos um novo modelo combinando-o com um modelo do Nível Curso e o modelo do Nível Instituição, culminando em outros 66 modelos da estratégia Ensemble Stacking. Por fim, maximizamos essa estratégia com a adição de mais um conjunto de 78 datasets, outras 3 estratégias ensemble e outros 2 algoritmos.

Observamos que, apesar do menor volume de dados, o Nível Período (Janelas Semestrais) geralmente resultou em métricas mais altas. Os modelos desenvolvidos a partir deste nível, utilizando técnicas de Bagging, demonstraram uma superioridade significativa em relação aos modelos baseados no mesmo algoritmo, porém aplicados individualmente (Decision Tree), em praticamente todos os cursos e períodos analisados.

Em seguida, constatamos que a combinação de modelos de níveis distintos (Instituição, Curso e Período) aptos a apoiar diferentes níveis de gestão acadêmica (Administração Superior, Coordenações de Curso e Docentes do Período, por exemplo), através da técnica de Ensemble Stacking, resultou em melhores níveis de Acurácia e MCC em relação aos modelos originais (treinados individualmente com o mesmo algoritmo usado pelos ensembles, Decision Tree). Ou seja, o uso de modelos ensemble com heterogeneidade advinda dos dados (e não dos algoritmos utilizados) agrega valor aos resultados ao aproveitar-se de modelos distintos para um mesmo fim.

Ao combinar essa estratégia com a estratégia de Bagging, conseguimos utilizar algumas das qualidades de ambas as técnicas. Combinando (Stacking) modelos heterogêneos fortes criados pelo método de Bagging, foi possível potencializar os resultados, aumentando a capacidade de predição do ensemble e alcançando melhores índices em nossas métricas Acurácia e MCC.

Esta pesquisa contribui então para a compreensão das vantagens dos Métodos Ensemble na predição de evasão educacional, destacando sua aplicabilidade em diferentes níveis de granularidade dos dados. Os resultados obtidos sugerem que essa abordagem, potencializada pela diversidade dos dados, pode melhorar significativamente a precisão dos modelos de predição, aproveitando-se de modelos capazes de oferecer insights valiosos para diferentes níveis de gestores educacionais na identificação precoce de alunos em risco de evasão. Embora não seja o objetivo de nosso estudo, cada nível de modelo pode identificar, de forma nativa (no caso do uso de árvores de decisão), as características mais influentes na predição em relação à Instituição, ao Curso ou ao Período. Características estas que podem ser utilizadas como foco na construção de políticas públicas. E a combinação desses modelos proporciona maior precisão no resultado final.

Vale frisar que todos os métodos utilizados neste trabalho empregaram, propositalmente, um único algoritmo simultaneamente. Isso se deve ao fato de que o propósito deste estudo foi comparar e combinar estratégias que explorassem a diversidade dos dados, em vez de focar na maximização dos valores das métricas. Dessa forma, optamos por padronizar os modelos do ponto de vista dos algoritmos, a fim de tornar mais clara a avaliação em relação aos ganhos provenientes da diversidade dos dados introduzida em nossos modelos. Isso significa que nossos resultados foram alcançados sem a necessidade de utilizar um dos principais recursos desses métodos: a combinação de modelos criados com algoritmos distintos. Além de não termos necessitado efetuar ajustes nos hiperparâmetros. Isso nos leva a acreditar que o uso desses recursos poderia proporcionar resultados ainda melhores, com o uso de árvores de decisão para auxiliar no processo de explicabilidade, e de outros algoritmos para proporcionar melhores resultados nas métricas.

Trabalhos Futuros

Como direcionamentos para trabalhos futuros, os métodos de Boosting são ainda uma alternativa não abordada neste processo e potencialmente poderiam agregar maior valor ao processo, além do uso de outros paradigmas e famílias de algoritmos.

Outro ponto complementar seria a obtenção de novos atributos que possam agregar mais valor aos modelos de todos os níveis. Por exemplo, a posição ou colocação do aluno entre os aprovados no processo seletivo e as notas dos alunos em cada disciplina do currículo para modelos de um determinado período. Outros atributos que não constam na base de dados poderiam ser interessantes, como o meio de transporte utilizado e a existência de emprego concomitante ao estudo. Outros níveis dentre os existentes na literatura também poderiam ser muito úteis, agregando ainda mais diversidade, como o abordado em [Oliveira e Medeiros \(2024\)](#) construído com base nos dados da Autoavaliação Institucional.

Uma estratégia interessante a ser implementada é a construção de uma arquitetura capaz de identificar quais algoritmos são mais adequados a cada nível e período em tempo de execução, como um passo anterior à construção dos modelos ensemble finais. Isso permitiria potencializar os resultados usando um algoritmo ideal específico para cada nível de dataset/período, combinando-os em modelos ensemble stacking e agregando ainda mais diversidade aos modelos com conjuntos de dados/níveis distintos. Por exemplo, se para os níveis de Instituição, Curso e Período for identificado que os algoritmos mais bem sucedidos são Random Forest, MLP e Gaussian Naive Bayes respectivamente, estes seriam os algoritmos utilizados na construção de um modelo stacking junto aos dados de cada nível. Ou ainda uma estratégia de seleção de classificadores ótimos (SCO) que visa melhorar o desempenho geral do ensemble, eliminando classificadores que contribuem pouco ou prejudicam a precisão das predições ([ALSHDAIFAT; AL-HASSAN; ALOQAILY, 2021](#)). A presença de um classificador de árvore pode ser muito útil para a explicabilidade.

Adicionalmente, o trabalho de ([COSTA; JÚNIOR; SANTANA, 2023](#)) aborda uma forma de agrupamento de dados sob a abordagem de janelas semestrais que pode potencializar os resultados especialmente para os períodos não abordados em nosso estudo e para alguns cujo resultado pode ter sido influenciado negativamente devido a baixa quantidade de amostras. A combinação das abordagens utilizadas com o uso de métodos ensemble provavelmente trarão melhores resultados de predição.

Por fim, este trabalho, tendo um propósito conceitual de comparação de técnicas no contexto da predição de evasão, não abordou a busca pelos melhores parâmetros de cada algoritmo utilizado, tendo suas únicas intervenções realizadas propositalmente para proporcionar melhor tempo de máquina nos algoritmos Random Forest e Extra Trees. O ajuste de hiperparâmetros é uma etapa essencial para a obtenção de melhores resultados em situações práticas e representa uma oportunidade importante para trabalhos futuros.

Referências

AGGARWAL, C. C. et al. *Data mining: the textbook*. [S.l.]: Springer, 2015. v. 1. Citado na página 27.

ALHAMID, M. *Ensemble Models. A guide to learning ensemble techniques in a simple walkthrough*. 2021. (Accessed on 10/27/2023). Disponível em: <<https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c>>. Citado na página 25.

ALI, J. et al. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, International Journal of Computer Science Issues (IJCSI), v. 9, n. 5, p. 272, 2012. Citado na página 20.

ALSHDAIFAT, E.; AL-HASSAN, M.; ALOQAILY, A. Effective heterogeneous ensemble classification: An alternative approach for selecting base classifiers. *ICT Express*, Elsevier, v. 7, n. 3, p. 342–349, 2021. Citado 5 vezes nas páginas 23, 26, 28, 29 e 75.

ARAUJO, F. H. D. d. *Descoberta de Conhecimento em Base de Dados para o aprendizado da Regulação Médica/Odontológica em Operadora de Plano de Saúde*. Dissertação (Mestrado) — Universidade Federal do Piauí, 2014. Dissertação de Mestrado em Ciência da Computação. Citado 5 vezes nas páginas 15, 10, 13, 16 e 17.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de informática na educação*, v. 19, n. 02, p. 03, 2011. Citado 3 vezes nas páginas 4, 29 e 30.

BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. d. C. May artificial intelligence support actions against school dropout? *Ensaio: Avaliação e Políticas Públicas em Educação*, SciELO Brasil, v. 30, p. 669–694, 2021. Citado 2 vezes nas páginas 1 e 25.

BRACHMAN, R. J.; ANAND, T. *The process of knowledge discovery in a first sketch*. [S.l.], 1994. Citado 3 vezes nas páginas 10, 12 e 13.

BRITO, B. C. P. de; MELLO, R. F. L. de; ALVES, G. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1032–1041. Citado na página 4.

CARNEIRO, M. G. et al. Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. *Revista Brasileira de Informática na Educação*, v. 30, p. 379–395, 2022. Citado na página 31.

CASANOVA, M. A. *Utilizando aprendizado de máquina para construção de uma ferramenta de apoio a moderação de comentários*. Tese (Doutorado) — PUC-Rio, 2011. Citado na página 15.

CASTRO, L. N. D.; FERRARI, D. G. *Introdução à mineração de dados: Conceitos básicos, algoritmos e aplicações*. 1ª edição. ed. [S.l.]: Saraiva Educação SA, 2016. ISBN 978-8547200985. Citado na página 13.

CHUNG, J. Y.; LEE, S. Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, Elsevier, v. 96, p. 346–353, 2019. Citado na página 26.

COSTA, C. T. da; JÚNIOR, M. L. L.; SANTANA, A. M. União de dados por clusterização para construção de modelos de predição de evasão. *Revista Novas Tecnologias na Educação*, v. 21, n. 2, p. 393–402, 2023. Citado 3 vezes nas páginas 32, 33 e 75.

COULON, A. I. *The student's craft: The entrance to university life*. 2017. 1239-1250 p. Citado na página 1.

DIB, M. et al. Single and ensemble classifiers for defect prediction in sheet metal forming under variability. *Neural Computing and Applications*, Springer, v. 32, p. 12335–12349, 2020. Citado 3 vezes nas páginas 20, 28 e 58.

DIETTERICH, T. G. Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*. [S.l.]: Springer, 2000. p. 1–15. Citado 5 vezes nas páginas 15, 21, 22, 23 e 60.

DIVINA, F. et al. Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, MDPI, v. 11, n. 4, p. 949, 2018. Citado na página 58.

DŽEROSKI, S.; ŽENKO, B. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, Springer, v. 54, p. 255–273, 2004. Citado na página 58.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996. Citado 2 vezes nas páginas 9 e 10.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, p. 641–659, 2007. Citado na página 1.

FLETCHER, S.; VERMA, B.; ZHANG, M. A non-specialized ensemble classifier using multi-objective optimization. *Neurocomputing*, Elsevier, v. 409, p. 93–102, 2020. Citado na página 25.

FRITSCH, R.; ROCHA, C. S. d.; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. *Revista Educação em Questão*, v. 52, n. 38, p. 81–108, 2015. Citado na página 1.

FUJIWARA, K. et al. Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. *Frontiers in public health*, Frontiers Media SA, v. 8, p. 178, 2020. Citado na página 11.

GILIOLI, R. d. S. P. *Evasão em instituições federais de ensino superior no Brasil: expansão da rede, SISU e desafios*. [S.l.], 2016. v. 49, 1–55 p. Disponível em: <https://bd.camara.leg.br/bd/bitstream/handle/bdcamara/28239/evasao_instituicoes_gilioli.pdf?sequence=1>. Citado na página 1.

HAN, J.; KAMBER, M.; PEI, J. *Data mining concepts and techniques third edition*. 2012. Citado na página 29.

HIPOLITO, O. *A falta de engajamento dos estudantes no ensino superior - 01/07/2023 - Opinião - Folha*. 2023. (Accessed on 10/27/2023). Disponível em: <<https://www1.folha.uol.com.br/opiniao/2023/07/a-falta-de-engajamento-dos-estudantes-no-ensino-superior.shtml>>. Citado na página 2.

INEP. *Indicadores de Fluxo da Educação Superior — Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira | Inep*. 2023. (Accessed on 10/27/2023). Disponível em: <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior>>. Citado 2 vezes nas páginas 2 e 3.

JESUS, H. O. de; RODRIGUEZ, L. C.; JUNIOR, A. d. O. C. Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação*, v. 29, p. 255–272, 2021. Citado 3 vezes nas páginas 16, 25 e 30.

KALIRANE, M. *Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking*. 2024. (Accessed on 02/15/2024). Disponível em: <<https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/>>. Citado 3 vezes nas páginas 15, 27 e 28.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, p. 14, 2012. Citado na página 1.

MEC. *Reuni - Reestruturação e Expansão das Universidades Federais - REUNI*. 2010. <<https://reuni.mec.gov.br/o-que-e-o-reuni>>. (Accessed on 10/27/2023). Citado na página 3.

MEC. *Portaria MEC Nº 374 de 30 de abril lde 2020: Dispõe sobre a antecipação da colação de grau para os alunos dos cursos de Medicina, Enfermagem, Farmácia e Fisioterapia, exclusivamente para atuação nas ações de combate à pandemia do novo coronavírus - Covid-19*. 2020. (Accessed on 10/27/2023). Disponível em: <<https://abmes.org.br/arquivos/legislacoes/Portaria-mec-374-2020-04-03.pdf>>. Citado na página 3.

MELO, E. C.; SOUZA, F. S. H. de. Improving the prediction of school dropout with the support of the semi-supervised learning approach. *iSys-Brazilian Journal of Information Systems*, v. 16, n. 1, p. 10–1, 2023. Citado na página 30.

MIOT, H. A. Correlation analysis in clinical and experimental studies. *Jornal vascular brasileiro*, SciELO Brasil, v. 17, p. 275–279, 2018. Citado na página 41.

OLIVEIRA, I. S.; MEDEIROS, F. P. A.; ANDRADE, F. G. Seleção de atributos para classificadores de evasão escolar com dados da plataforma nilo peçanha. In: SBC. *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*. [S.l.], 2022. p. 30–39. Citado na página 16.

OLIVEIRA, R. dos S.; MEDEIROS, F. P. A. de. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. *Revista Brasileira de Informática na Educação*, v. 32, p. 1–21, 2024. Citado 2 vezes nas páginas 32 e 75.

PARMEZAN, A. R. S. et al. *Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados*. [S.l.]: dez, 2012. Citado 2 vezes nas páginas 11 e 12.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 624. Citado 2 vezes nas páginas 1 e 25.

PIAO, Y. et al. A new ensemble method with feature space partitioning for high-dimensional data classification. *Mathematical Problems in Engineering*, Hindawi, v. 2015, 2015. Citado 3 vezes nas páginas 21, 22 e 27.

PORTAL da Câmara dos Deputados. 2018. <<https://www2.camara.leg.br/legin/fed/lei/2018/lei-13651-11-abril-2018-786547-norma-pl.html>>. (Accessed on 10/27/2023). Citado na página 3.

PRAJWALA, T. A comparative study on decision tree and random forest using r tool. *International journal of advanced research in computer and communication engineering*, v. 4, n. 1, p. 196–199, 2015. Citado na página 19.

RAHMAN, A.; VERMA, B. Ensemble classifier generation using non-uniform layered clustering and genetic algorithm. *Knowledge-Based Systems*, Elsevier, v. 43, p. 30–42, 2013. Citado 2 vezes nas páginas 23 e 24.

RIGO, S. J. et al. Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, v. 22, n. 01, p. 132, 2014. Citado 3 vezes nas páginas 15, 10 e 29.

RODRIGUES, R. L. et al. A literatura brasileira sobre mineração de dados educacionais. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO DOURADOS, MS. *Congresso Brasileiro de Informática na Educação*. [S.l.], 2014. p. 621–630. Citado na página 16.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, Wiley Online Library, v. 10, n. 3, p. e1355, 2020. Citado na página 30.

RUBACK, L.; AVILA, S.; CANTERO, L. Vieses no aprendizado de máquina e suas implicações sociais: Um estudo de caso no reconhecimento facial. In: SBC. *Anais do II Workshop sobre as Implicações da Computação na Sociedade*. [S.l.], 2021. p. 90–101. Citado 2 vezes nas páginas 15 e 14.

SACRAMENTO, G. *Árvore de Decisão: Entenda Esse Algoritmo de Machine Learning*. 2021. (Accessed on 15/10/2021). Disponível em: <<https://blog.somostera.com/data-science/arvores-de-decisao>>. Citado 2 vezes nas páginas 15 e 19.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. Citado 2 vezes nas páginas 15 e 23.

SAHÃO, F. T.; KIENEN, N. University student adaptation and mental health: A systematic review of literature. *Psicologia Escolar e Educacional*, SciELO Brasil, v. 25, 2021. Citado na página 2.

- SALES, A.; BALBY, L.; CAJUEIRO, A. Exploiting academic records for predicting student drop out: a case study in brazilian higher education. *Journal of Information and Data Management*, v. 7, n. 2, p. 166–166, 2016. Citado 2 vezes nas páginas 1 e 31.
- SANTANA, L. E. A. d. S.; CANUTO, A. M. de P. Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications*, Elsevier, v. 41, n. 4, p. 1622–1631, 2014. Citado 3 vezes nas páginas 23, 25 e 42.
- SANTOS, C. H. D.; MARTINS, S. de L.; PLASTINO, A. É possível prever evasão com base apenas no desempenho acadêmico? In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 792–802. Citado na página 31.
- SANTOS, D. H. d. *Aplicação de técnicas de ensemble learning na detecção estática e dinâmica de malwares*. 2023. Trabalho de Conclusão de Curso (Bacharelado) - Universidade Estadual Paulista (Unesp). Citado 2 vezes nas páginas 27 e 58.
- SANTOS, K. J. d. O. et al. Supervised learning in the context of educational data mining to avoid university students dropout. In: IEEE. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2019. v. 2161, p. 207–208. Citado na página 25.
- SANTOS, V. H. B. dos; SARAIVA, D. V.; OLIVEIRA, C. T. de. Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 1196–1210. Citado 3 vezes nas páginas 20, 31 e 50.
- SEMESP, I. *Mapa do Ensino Superior. 13ª Edição*. 2023. Disponível em: <<https://www.semesp.org.br/mapa/edicao-13/brasil/indicadores-de-trajetoria/>> (Accessed on 20/10/2023). Disponível em: <<https://www.semesp.org.br/mapa/edicao-13/brasil/indicadores-de-trajetoria/>>. Citado 2 vezes nas páginas 15 e 2.
- SILVA, A. M. d.; SANTOS, B. C. S. Eficácia de políticas de acesso ao ensino superior privado na contenção da evasão. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, SciELO Brasil, v. 22, p. 741–757, 2017. Citado na página 1.
- SILVA, A. M. L. et al. Descoberta de conhecimento através de métodos de aprendizagem de máquina supervisionados aplicados ao sigaa/ufpi. *Revista de Sistemas e Computação-RSC*, v. 7, n. 1, 2017. Citado na página 18.
- SILVA, P. M. D. et al. Ensemble regression models applied to dropout in higher education. In: IEEE. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2019. p. 120–125. Citado 3 vezes nas páginas 9, 24 e 31.
- TEODORO, L. de A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, v. 28, p. 838–863, 2020. Citado na página 30.
- TINTO, V. *Research and practice of student retention: What next?* [S.l.]: SAGE Publications Sage CA: Los Angeles, CA, 2006. 1–19 p. Citado 2 vezes nas páginas 1 e 2.

- TONTINI, G.; WALTER, S. A. Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, SciELO Brasil, v. 19, p. 89–110, 2014. Citado na página 1.
- TSOUMAKAS, G.; ANGELIS, L.; VLAHAVAS, I. Selective fusion of heterogeneous classifiers. *Intelligent Data Analysis*, IOS Press, v. 9, n. 6, p. 511–525, 2005. Citado na página 29.
- TUMER, K.; GHOSH, J. Error correlation and error reduction in ensemble classifiers. *Connection science*, Taylor & Francis, v. 8, n. 3-4, p. 385–404, 1996. Citado na página 25.
- UFDPAR. *Em dia histórico, UFDPAR realizou formatura de 21 novos médicos em Parnaíba*. 2020. <<https://ufpi.br/ultimas-noticias-parnaiba/36056-em-dia-historico-ufdpar-realizou-formatura-de-21-novos-medicos-em-parnaiba>>. (Accessed on 10/27/2023). Citado na página 3.
- UFPI. *Ingresso e Seleção*. 2019. <<https://ufpi.br/ingresso-e-selecao>>. (Accessed on 10/27/2023). Citado na página 39.
- VIANA, F. S.; SANTANA, A. M.; RABÊLO, R. d. A. L. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In: SBC. *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2022. p. 908–919. Citado 9 vezes nas páginas 10, 14, 21, 25, 32, 33, 36, 41 e 44.
- WANG, H.; BAH, M.; HAMMAD, M. Progress in outlier detection techniques: A survey. *IEEE Access*, v. 7, p. 1–1, 08 2019. Citado na página 11.
- WITTEN, I. H.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, ACM New York, NY, USA, v. 31, n. 1, p. 76–77, 2002. Citado na página 16.
- YANG, P. et al. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, Bentham Science Publishers, v. 5, n. 4, p. 296–308, 2010. Citado 2 vezes nas páginas 26 e 27.

Apêndices

APÊNDICE A – Resultados dos Testes Estatísticos

Resultados completos da aplicação do Teste T de Student entre os modelos com Ensemble e o modelo IND com a respectiva combinação (Comparação entre os modelos usando o mesmo Conjunto de Atributos, a mesma métrica, e o mesmo algoritmo)

CONJ. ATRIBUTOS	MÉTRICA	CLASSIFICA DOR	TIPO ENSEMBLE	PX	ADM	BIOM	BIO	CON	ECO	PES	FIS	MAT	PED	PSI	TUR	% MELHORIA POR PERÍODO	% MELHORIA POR COMBINAÇÃO	% APLICABILIDADE POR PERÍODO	% APLICABILIDADE POR COMBINAÇÃO
FILT	ACC	DT	BAG	1	0	1	0	0	0	0	1	1	1	1	0	45,45%	53,03%	100,00%	100%
FILT	ACC	DT	BAG	2	0	1	1	1	0	0	1	1	1	1	0	63,64%		100,00%	
FILT	ACC	DT	BAG	3	1	0	1	1	0	1	1	1	1	1	0	72,73%		100,00%	
FILT	ACC	DT	BAG	4	1	1	1	1	0	0	0	0	0	1	0	45,45%		100,00%	
FILT	ACC	DT	BAG	5	1	1	1	1	0	0	1	0	0	1	1	63,64%		100,00%	
FILT	ACC	DT	BAG	6	0	0	0	1	0	0	1	0	0	1	0	27,27%		100,00%	
FILT	ACC	DT	STACK	1	0	1	0	1	1	1	0	1	0	0	0	45,45%	46,97%	100,00%	93,94%
FILT	ACC	DT	STACK	2	1	0	0	1	1	1	1	0	1	1	0	63,64%		100,00%	
FILT	ACC	DT	STACK	3	0	1	0	1	0	0	0	0	1	1	1	45,45%		100,00%	
FILT	ACC	DT	STACK	4	0	1	1	0	0	0	1	0	1	0	0	36,36%		100,00%	
FILT	ACC	DT	STACK	5	0	0	1	1	1	0	0	-1	-1	1	0	36,36%		81,82%	
FILT	ACC	DT	STACK	6	0	1	0	0	1	-1	1	1	-1	1	1	54,55%		81,82%	
FILT	ACC	DT	STACK BAG	1	1	1	1	0	1	1	0	1	0	1	0	63,64%	57,58%	100,00%	95,46%
FILT	ACC	DT	STACK BAG	2	0	0	1	1	1	1	1	1	1	1	0	72,73%		100,00%	
FILT	ACC	DT	STACK BAG	3	0	0	0	1	0	1	1	0	1	1	0	45,45%		100,00%	
FILT	ACC	DT	STACK BAG	4	0	0	1	1	0	1	1	0	0	1	0	45,45%		100,00%	
FILT	ACC	DT	STACK BAG	5	1	0	1	1	-1	0	1	0	-1	0	1	45,45%		81,82%	
FILT	ACC	DT	STACK BAG	6	1	1	0	1	1	1	1	1	-1	1	0	72,73%		90,91%	
FILT	ACC	DT	VOT	1	-1	1	0	0	0	0	0	1	0	0	0	18,18%	16,67%	90,91%	87,88%
FILT	ACC	DT	VOT	2	0	-1	0	0	1	1	0	0	0	0	0	18,18%		90,91%	
FILT	ACC	DT	VOT	3	0	0	0	0	0	0	0	-1	1	0	0	9,09%		90,91%	
FILT	ACC	DT	VOT	4	0	0	0	0	0	0	0	0	0	1	0	9,09%		100,00%	
FILT	ACC	DT	VOT	5	0	0	1	-1	1	-1	0	0	-1	1	0	27,27%		72,73%	
FILT	ACC	DT	VOT	6	0	0	0	-1	1	0	1	0	-1	0	0	18,18%		81,82%	
FILT	ACC	DT	VOT BAG	1	0	1	0	1	0	1	0	1	0	1	-1	45,45%	36,36%	90,91%	90,91%
FILT	ACC	DT	VOT BAG	2	0	0	0	1	0	1	0	0	1	0	0	27,27%		100,00%	
FILT	ACC	DT	VOT BAG	3	1	0	0	1	-1	1	1	0	1	0	0	45,45%		90,91%	
FILT	ACC	DT	VOT BAG	4	1	0	1	0	-1	0	1	0	0	1	0	36,36%		90,91%	
FILT	ACC	DT	VOT BAG	5	0	0	1	0	-1	0	0	0	-1	1	0	18,18%		81,82%	
FILT	ACC	DT	VOT BAG	6	0	1	0	0	1	0	1	1	-1	1	0	45,45%		90,91%	
FILT	ACC	ET	BAG	1	0	0	0	0	0	0	0	0	0	0	0	0,00%	10,61%	100,00%	100%
FILT	ACC	ET	BAG	2	0	0	0	0	0	0	1	0	0	0	1	18,18%		100,00%	
FILT	ACC	ET	BAG	3	0	0	0	0	1	0	1	0	0	0	0	18,18%		100,00%	
FILT	ACC	ET	BAG	4	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
FILT	ACC	ET	BAG	5	1	0	1	0	0	0	0	0	0	0	0	18,18%		100,00%	
FILT	ACC	ET	BAG	6	0	0	0	0	0	1	0	0	0	0	0	9,09%		100,00%	
FILT	ACC	ET	STACK	1	-1	1	1	0	1	1	0	1	0	1	1	63,64%	50%	90,91%	90,91%
FILT	ACC	ET	STACK	2	0	0	1	0	1	1	0	0	1	0	0	36,36%		100,00%	
FILT	ACC	ET	STACK	3	0	1	1	1	1	1	0	0	1	0	0	54,55%		100,00%	
FILT	ACC	ET	STACK	4	-1	1	1	0	-1	1	0	0	0	1	0	36,36%		81,82%	
FILT	ACC	ET	STACK	5	1	1	1	1	1	0	0	1	-1	0	0	54,55%		90,91%	
FILT	ACC	ET	STACK	6	1	1	0	0	1	1	1	-1	-1	0	1	54,55%		81,82%	
FILT	ACC	ET	STACK BAG	1	0	1	1	1	0	1	0	1	0	0	1	54,55%	53,03%	100,00%	90,91%
FILT	ACC	ET	STACK BAG	2	0	1	1	0	1	1	1	1	0	0	0	54,55%		100,00%	
FILT	ACC	ET	STACK BAG	3	1	1	1	0	1	1	0	0	1	0	1	63,64%		100,00%	
FILT	ACC	ET	STACK BAG	4	0	1	1	0	0	1	0	0	0	0	0	27,27%		100,00%	
FILT	ACC	ET	STACK BAG	5	1	0	1	1	-1	1	1	1	-1	-1	1	63,64%		72,73%	
FILT	ACC	ET	STACK BAG	6	0	1	1	1	1	1	-1	-1	-1	0	1	54,55%		72,73%	
FILT	ACC	ET	VOT	1	-1	0	0	0	0	1	0	0	0	0	-1	9,09%	13,64%	81,82%	78,79%
FILT	ACC	ET	VOT	2	0	0	0	0	0	1	0	0	0	0	-1	9,09%		90,91%	
FILT	ACC	ET	VOT	3	0	0	0	0	-1	0	0	0	1	0	0	9,09%		90,91%	
FILT	ACC	ET	VOT	4	-1	1	0	0	-1	0	0	0	-1	0	0	9,09%		72,73%	
FILT	ACC	ET	VOT	5	0	0	1	0	-1	0	0	0	-1	-1	0	9,09%		72,73%	
FILT	ACC	ET	VOT	6	-1	1	0	0	1	1	1	-1	-1	0	-1	36,36%		63,64%	

FILT	ACC	ET	VOT BAG	1	0	0	0	0	0	1	0	1	0	0	0	18,18%	12,12%	100,00%	80,31%	
FILT	ACC	ET	VOT BAG	2	0	0	0	0	-1	1	0	0	0	0	-1	9,09%		81,82%		
FILT	ACC	ET	VOT BAG	3	0	0	0	0	0	-1	0	0	0	1	0	9,09%		90,91%		
FILT	ACC	ET	VOT BAG	4	0	1	0	0	-1	0	0	0	-1	0	-1	9,09%		72,73%		
FILT	ACC	ET	VOT BAG	5	0	0	-1	0	-1	0	1	0	-1	-1	0	9,09%		63,64%		
FILT	ACC	ET	VOT BAG	6	-1	0	0	0	1	0	0	-1	-1	0	1	18,18%		72,73%		
FILT	ACC	RF	BAG	1	0	0	1	0	1	1	0	1	0	0	0	36,36%	19,7%	100,00%	96,97%	
FILT	ACC	RF	BAG	2	0	0	0	1	0	0	0	0	0	0	0	9,09%		100,00%		
FILT	ACC	RF	BAG	3	0	1	1	0	0	0	0	0	0	0	0	18,18%		100,00%		
FILT	ACC	RF	BAG	4	1	0	1	0	-1	0	0	0	0	1	-1	27,27%		81,82%		
FILT	ACC	RF	BAG	5	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%		
FILT	ACC	RF	BAG	6	0	0	1	1	0	0	0	0	1	0	0	27,27%		100,00%		
FILT	ACC	RF	STACK	1	0	0	1	0	1	1	0	1	0	0	0	36,36%	31,82%	100,00%	83,34%	
FILT	ACC	RF	STACK	2	0	0	1	1	0	1	0	1	0	0	0	36,36%		100,00%		
FILT	ACC	RF	STACK	3	0	0	0	0	-1	0	1	0	1	0	0	18,18%		90,91%		
FILT	ACC	RF	STACK	4	0	1	1	1	-1	1	0	0	-1	0	0	36,36%		81,82%		
FILT	ACC	RF	STACK	5	0	0	1	0	-1	-1	0	1	-1	1	1	36,36%		72,73%		
FILT	ACC	RF	STACK	6	0	-1	1	0	-1	1	-1	-1	-1	0	1	27,27%		54,55%		
FILT	ACC	RF	STACK BAG	1	0	1	1	1	1	1	0	1	1	1	1	81,82%	43,94%	100,00%	89,4%	
FILT	ACC	RF	STACK BAG	2	0	1	1	0	1	1	0	1	0	1	0	54,55%		100,00%		
FILT	ACC	RF	STACK BAG	3	0	0	0	0	-1	1	0	1	1	0	0	27,27%		90,91%		
FILT	ACC	RF	STACK BAG	4	0	0	0	0	1	1	0	1	-1	0	0	27,27%		90,91%		
FILT	ACC	RF	STACK BAG	5	0	0	1	1	-1	0	1	1	-1	0	1	45,45%		81,82%		
FILT	ACC	RF	STACK BAG	6	0	1	0	0	-1	1	0	-1	-1	0	1	27,27%		72,73%		
FILT	ACC	RF	VOT	1	0	0	0	1	0	0	0	0	0	0	0	9,09%	9,09%	100,00%	86,37%	
FILT	ACC	RF	VOT	2	0	0	0	0	0	1	0	0	0	0	0	9,09%		100,00%		
FILT	ACC	RF	VOT	3	0	0	0	0	0	-1	1	0	0	1	0	0		90,91%		
FILT	ACC	RF	VOT	4	0	-1	0	0	-1	0	0	0	0	0	0	-1		0,00%		
FILT	ACC	RF	VOT	5	0	0	0	0	-1	0	0	0	-1	0	0	0,00%		81,82%		
FILT	ACC	RF	VOT	6	0	-1	0	0	0	-1	0	0	1	-1	0	1		18,18%		
FILT	ACC	RF	VOT BAG	1	0	0	0	1	0	0	0	0	0	0	0	-1	9,09%	7,58%	90,91%	86,37%
FILT	ACC	RF	VOT BAG	2	0	0	0	0	0	1	0	0	0	0	0	-1	9,09%		90,91%	
FILT	ACC	RF	VOT BAG	3	0	0	0	0	0	0	0	0	0	1	0	0	9,09%		100,00%	
FILT	ACC	RF	VOT BAG	4	0	-1	0	0	0	0	0	0	0	0	0	-1	0,00%			
FILT	ACC	RF	VOT BAG	5	0	0	0	0	-1	0	1	0	-1	0	0	9,09%	81,82%			
FILT	ACC	RF	VOT BAG	6	0	1	0	0	0	-1	0	0	-1	-1	0	0	9,09%		72,73%	
FILT	MCC	DT	BAG	1	0	0	0	0	0	1	1	1	0	1	0	36,36%	42,42%	100,00%	96,97%	
FILT	MCC	DT	BAG	2	0	0	1	0	0	0	1	0	1	0	1	36,36%		100,00%		
FILT	MCC	DT	BAG	3	1	0	0	1	0	1	1	1	0	1	0	54,55%		100,00%		
FILT	MCC	DT	BAG	4	1	1	1	1	0	1	1	0	0	1	0	63,64%		100,00%		
FILT	MCC	DT	BAG	5	1	0	1	1	0	0	0	0	0	0	1	36,36%		100,00%		
FILT	MCC	DT	BAG	6	-1	1	0	0	0	0	0	-1	0	1	1	0		27,27%		81,82%
FILT	MCC	DT	STACK	1	0	1	0	1	1	1	0	0	0	0	0	36,36%	42,42%	100,00%	90,91%	
FILT	MCC	DT	STACK	2	0	0	1	0	1	1	1	0	1	0	0	45,45%		100,00%		
FILT	MCC	DT	STACK	3	1	0	0	0	0	0	0	0	0	1	0	1		27,27%		100,00%
FILT	MCC	DT	STACK	4	-1	1	1	-1	0	0	1	0	1	0	0	36,36%		81,82%		
FILT	MCC	DT	STACK	5	1	-1	1	0	1	0	0	-1	1	0	0	36,36%		81,82%		
FILT	MCC	DT	STACK	6	1	1	0	0	-1	1	-1	1	1	1	1	72,73%		81,82%		
FILT	MCC	DT	STACK BAG	1	1	0	1	0	1	1	0	1	0	1	0	54,55%	60,61%	100,00%	96,97%	
FILT	MCC	DT	STACK BAG	2	0	0	1	1	1	1	1	1	1	0	0	63,64%		100,00%		
FILT	MCC	DT	STACK BAG	3	0	0	0	1	0	1	1	1	1	1	0	0		45,45%		100,00%
FILT	MCC	DT	STACK BAG	4	1	0	1	0	0	1	1	0	1	1	0	54,55%		100,00%		
FILT	MCC	DT	STACK BAG	5	1	-1	1	1	-1	0	0	1	1	1	1	63,64%		81,82%		
FILT	MCC	DT	STACK BAG	6	1	1	0	1	1	1	1	1	1	1	0	81,82%		100,00%		
FILT	MCC	DT	VOT	1	-1	1	0	0	0	0	0	0	-1	0	0	9,09%	22,73%	81,82%	78,79%	
FILT	MCC	DT	VOT	2	-1	-1	0	-1	1	1	0	0	-1	0	0	18,18%		63,64%		
FILT	MCC	DT	VOT	3	1	-1	0	0	0	0	0	1	1	0	0	27,27%		90,91%		
FILT	MCC	DT	VOT	4	-1	0	0	-1	0	0	0	0	1	0	0	9,09%		81,82%		
FILT	MCC	DT	VOT	5	1	-1	1	-1	1	-1	0	0	1	0	0	36,36%		72,73%		
FILT	MCC	DT	VOT	6	-1	1	0	-1	1	0	1	0	1	0	0	36,36%		81,82%		

FILT	MCC	DT	VOT BAG	1	0	1	0	1	0	1	0	1	0	0	-1	36,36%	40,91%	90,91%	87,88%
FILT	MCC	DT	VOT BAG	2	-1	0	0	1	0	1	0	1	1	0	0	36,36%		90,91%	
FILT	MCC	DT	VOT BAG	3	0	0	0	1	-1	1	1	1	1	0	0	45,45%		90,91%	
FILT	MCC	DT	VOT BAG	4	1	0	1	0	-1	0	1	0	-1	1	0	36,36%		81,82%	
FILT	MCC	DT	VOT BAG	5	1	-1	1	0	-1	0	0	0	1	0	0	27,27%		81,82%	
FILT	MCC	DT	VOT BAG	6	1	1	0	0	1	1	-1	1	1	1	0	63,64%		90,91%	
FILT	MCC	ET	BAG	1	0	0	0	0	0	0	0	1	0	0	0	9,09%	13,64%	100,00%	100%
FILT	MCC	ET	BAG	2	0	0	0	0	0	0	1	0	0	0	1	18,18%		100,00%	
FILT	MCC	ET	BAG	3	0	0	0	0	0	1	0	1	0	0	0	18,18%		100,00%	
FILT	MCC	ET	BAG	4	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
FILT	MCC	ET	BAG	5	1	1	1	0	0	0	0	0	0	0	0	27,27%		100,00%	
FILT	MCC	ET	BAG	6	0	0	0	0	0	0	1	0	0	0	0	9,09%		100,00%	
FILT	MCC	ET	STACK	1	-1	1	1	0	1	1	0	1	0	1	1	63,64%	57,58%	90,91%	90,91%
FILT	MCC	ET	STACK	2	1	0	1	0	1	1	0	1	1	0	0	54,55%		100,00%	
FILT	MCC	ET	STACK	3	0	1	1	0	1	1	0	1	1	0	0	54,55%		100,00%	
FILT	MCC	ET	STACK	4	-1	1	1	0	-1	1	0	0	-1	1	0	36,36%		72,73%	
FILT	MCC	ET	STACK	5	1	1	1	1	1	1	0	1	-1	-1	0	63,64%		81,82%	
FILT	MCC	ET	STACK	6	1	1	0	0	1	1	1	1	1	0	1	72,73%		100,00%	
FILT	MCC	ET	STACK BAG	1	0	1	1	1	0	1	0	1	0	1	1	63,64%	65,15%	100,00%	89,4%
FILT	MCC	ET	STACK BAG	2	0	1	1	0	1	1	1	1	0	0	0	54,55%		100,00%	
FILT	MCC	ET	STACK BAG	3	1	1	1	1	1	1	0	1	1	-1	1	81,82%		90,91%	
FILT	MCC	ET	STACK BAG	4	1	1	1	0	0	1	0	0	-1	1	0	45,45%		90,91%	
FILT	MCC	ET	STACK BAG	5	1	1	1	1	1	1	1	1	-1	-1	1	81,82%		81,82%	
FILT	MCC	ET	STACK BAG	6	-1	1	1	1	1	1	-1	-1	1	0	1	63,64%		72,73%	
FILT	MCC	ET	VOT	1	-1	0	0	0	0	1	0	0	0	0	-1	9,09%	19,7%	81,82%	78,79%
FILT	MCC	ET	VOT	2	1	0	0	-1	0	1	0	0	1	0	-1	27,27%		81,82%	
FILT	MCC	ET	VOT	3	0	0	0	0	-1	0	0	0	0	1	0	9,09%		90,91%	
FILT	MCC	ET	VOT	4	-1	1	0	0	-1	0	0	0	-1	0	0	9,09%		72,73%	
FILT	MCC	ET	VOT	5	1	0	1	0	-1	0	0	0	-1	-1	0	18,18%		72,73%	
FILT	MCC	ET	VOT	6	-1	1	0	0	1	1	1	-1	1	0	-1	45,45%		72,73%	
FILT	MCC	ET	VOT BAG	1	0	0	0	0	0	1	0	0	0	0	0	9,09%	19,7%	100,00%	80,31%
FILT	MCC	ET	VOT BAG	2	0	0	0	0	-1	1	0	0	0	0	0	9,09%		90,91%	
FILT	MCC	ET	VOT BAG	3	0	0	0	0	-1	0	0	0	0	1	-1	9,09%		81,82%	
FILT	MCC	ET	VOT BAG	4	1	1	0	0	-1	0	0	0	-1	-1	-1	18,18%		63,64%	
FILT	MCC	ET	VOT BAG	5	1	1	-1	0	-1	0	1	0	-1	-1	0	27,27%		63,64%	
FILT	MCC	ET	VOT BAG	6	-1	1	0	0	1	0	0	1	1	-1	1	45,45%		81,82%	
FILT	MCC	RF	BAG	1	0	0	1	0	1	1	0	1	0	0	0	36,36%	19,7%	100,00%	95,46%
FILT	MCC	RF	BAG	2	0	0	0	1	0	0	0	0	0	0	0	9,09%		100,00%	
FILT	MCC	RF	BAG	3	0	1	0	0	0	1	0	0	0	0	0	18,18%		100,00%	
FILT	MCC	RF	BAG	4	1	0	1	0	-1	0	0	0	0	1	-1	27,27%		81,82%	
FILT	MCC	RF	BAG	5	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
FILT	MCC	RF	BAG	6	0	-1	1	1	0	0	0	0	0	1	0	27,27%		90,91%	
FILT	MCC	RF	STACK	1	0	0	1	0	1	1	0	1	0	0	0	36,36%	42,42%	100,00%	81,82%
FILT	MCC	RF	STACK	2	1	1	1	1	0	1	0	0	1	0	0	54,55%		100,00%	
FILT	MCC	RF	STACK	3	0	0	0	0	-1	0	1	0	1	0	0	18,18%		90,91%	
FILT	MCC	RF	STACK	4	-1	1	1	1	-1	1	0	0	-1	1	0	45,45%		72,73%	
FILT	MCC	RF	STACK	5	1	-1	1	0	-1	-1	0	1	1	-1	1	45,45%		63,64%	
FILT	MCC	RF	STACK	6	1	-1	1	1	-1	0	-1	-1	1	1	1	54,55%		63,64%	
FILT	MCC	RF	STACK BAG	1	0	1	1	1	1	1	0	1	0	1	1	72,73%	50%	100,00%	90,91%
FILT	MCC	RF	STACK BAG	2	0	1	1	0	1	1	0	1	0	1	0	54,55%		100,00%	
FILT	MCC	RF	STACK BAG	3	0	0	0	0	-1	1	0	1	1	0	0	27,27%		90,91%	
FILT	MCC	RF	STACK BAG	4	0	0	0	0	1	1	0	1	-1	1	0	36,36%		90,91%	
FILT	MCC	RF	STACK BAG	5	-1	1	1	1	-1	0	1	1	1	0	1	63,64%		81,82%	
FILT	MCC	RF	STACK BAG	6	1	1	0	0	-1	1	0	-1	1	0	1	45,45%		81,82%	
FILT	MCC	RF	VOT	1	0	0	0	1	0	0	0	0	1	0	0	18,18%	19,7%	100,00%	81,82%
FILT	MCC	RF	VOT	2	0	-1	0	0	0	1	0	0	1	0	0	18,18%		90,91%	
FILT	MCC	RF	VOT	3	0	0	0	0	-1	1	0	0	1	0	0	18,18%		90,91%	
FILT	MCC	RF	VOT	4	-1	-1	0	0	-1	0	0	0	-1	0	-1	0,00%		54,55%	
FILT	MCC	RF	VOT	5	1	-1	0	0	-1	0	0	0	1	0	0	18,18%		81,82%	
FILT	MCC	RF	VOT	6	-1	-1	0	1	-1	0	0	1	1	1	1	45,45%		72,73%	

FILT	MCC	RF	VOT BAG	1	0	0	0	1	0	0	0	0	0	0	-1	9,09%	16,67%	90,91%	84,85%	
FILT	MCC	RF	VOT BAG	2	0	0	0	0	0	1	0	0	0	0	-1	9,09%		90,91%		
FILT	MCC	RF	VOT BAG	3	0	0	0	0	0	0	0	0	1	0	0	9,09%		100,00%		
FILT	MCC	RF	VOT BAG	4	1	-1	0	0	0	0	0	0	-1	-1	-1	9,09%		63,64%		
FILT	MCC	RF	VOT BAG	5	1	-1	0	0	0	-1	0	1	0	1	0	0		27,27%		81,82%
FILT	MCC	RF	VOT BAG	6	-1	1	0	0	0	-1	0	0	1	1	1	0		36,36%		81,82%
ORIG	ACC	DT	BAG	1	0	1	0	1	0	0	1	1	1	1	1	63,64%	48,48%	100,00%	100%	
ORIG	ACC	DT	BAG	2	0	1	0	0	0	1	1	0	1	0	0	36,36%		100,00%		
ORIG	ACC	DT	BAG	3	1	0	0	1	0	1	0	0	0	1	1	45,45%		100,00%		
ORIG	ACC	DT	BAG	4	0	0	1	1	0	0	1	1	1	1	0	54,55%		100,00%		
ORIG	ACC	DT	BAG	5	1	0	1	1	0	0	0	0	0	1	1	45,45%		100,00%		
ORIG	ACC	DT	BAG	6	1	0	1	0	1	0	1	0	0	1	0	45,45%		100,00%		
ORIG	ACC	DT	STACK	1	0	0	0	0	1	0	0	1	0	1	0	27,27%	33,33%	100,00%	92,43%	
ORIG	ACC	DT	STACK	2	0	0	0	0	0	1	1	0	0	1	1	0		36,36%		100,00%
ORIG	ACC	DT	STACK	3	0	1	0	0	0	1	0	-1	0	1	0	27,27%		90,91%		
ORIG	ACC	DT	STACK	4	1	0	1	0	1	0	0	1	1	1	0	54,55%		100,00%		
ORIG	ACC	DT	STACK	5	-1	0	1	0	-1	0	0	0	1	1	0	27,27%		81,82%		
ORIG	ACC	DT	STACK	6	0	1	0	0	1	-1	0	0	-1	1	0	27,27%		81,82%		
ORIG	ACC	DT	STACK BAG	1	0	0	1	0	1	1	0	1	1	1	1	63,64%	62,12%	100,00%	96,97%	
ORIG	ACC	DT	STACK BAG	2	1	1	1	0	1	1	0	0	1	1	1	72,73%		100,00%		
ORIG	ACC	DT	STACK BAG	3	1	0	1	0	-1	1	1	0	1	1	0	54,55%		90,91%		
ORIG	ACC	DT	STACK BAG	4	1	1	1	0	-1	1	1	1	1	1	1	81,82%		90,91%		
ORIG	ACC	DT	STACK BAG	5	0	0	1	0	0	0	0	0	1	1	1	36,36%		100,00%		
ORIG	ACC	DT	STACK BAG	6	0	0	1	1	0	1	1	1	1	1	0	63,64%		100,00%		
ORIG	ACC	DT	VOT	1	-1	0	0	0	0	0	0	0	0	0	0	0,00%	9,09%	90,91%	90,91%	
ORIG	ACC	DT	VOT	2	-1	0	0	0	0	-1	0	0	0	0	0	0,00%		81,82%		
ORIG	ACC	DT	VOT	3	0	0	0	0	0	0	0	0	0	0	-1	0		0,00%		90,91%
ORIG	ACC	DT	VOT	4	0	0	0	0	0	1	0	0	0	1	1	-1		27,27%		90,91%
ORIG	ACC	DT	VOT	5	0	0	0	0	0	0	0	0	0	-1	0	0		0,00%		90,91%
ORIG	ACC	DT	VOT	6	0	1	0	0	0	1	0	0	0	0	1	0		27,27%		100,00%
ORIG	ACC	DT	VOT BAG	1	0	0	0	0	0	-1	0	0	0	0	0	0	0,00%	18,18%	90,91%	93,94%
ORIG	ACC	DT	VOT BAG	2	0	0	0	1	0	0	0	0	0	0	0	0	9,09%		100,00%	
ORIG	ACC	DT	VOT BAG	3	1	0	0	0	0	0	1	0	0	0	-1	0	18,18%		90,91%	
ORIG	ACC	DT	VOT BAG	4	0	-1	0	0	1	0	0	1	1	1	0	36,36%	90,91%			
ORIG	ACC	DT	VOT BAG	5	0	0	1	0	0	0	0	0	0	1	1	27,27%	100,00%			
ORIG	ACC	DT	VOT BAG	6	0	1	0	0	0	0	0	1	0	-1	0	0	18,18%		90,91%	
ORIG	ACC	ET	BAG	1	0	0	0	0	0	0	0	0	-1	0	0	0	0,00%	13,64%	90,91%	95,46%
ORIG	ACC	ET	BAG	2	0	1	0	0	0	0	0	0	0	0	1	0	18,18%		100,00%	
ORIG	ACC	ET	BAG	3	0	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
ORIG	ACC	ET	BAG	4	1	0	0	1	0	0	0	0	0	-1	0	0	18,18%		90,91%	
ORIG	ACC	ET	BAG	5	1	1	0	1	0	0	1	1	-1	0	0	45,45%	90,91%			
ORIG	ACC	ET	BAG	6	0	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
ORIG	ACC	ET	STACK	1	0	0	1	1	1	1	1	1	1	0	0	1	63,64%	46,97%	100,00%	86,37%
ORIG	ACC	ET	STACK	2	0	0	1	1	1	1	0	0	1	1	0	54,55%	100,00%			
ORIG	ACC	ET	STACK	3	1	1	0	1	1	1	0	1	1	1	0	63,64%	100,00%			
ORIG	ACC	ET	STACK	4	0	-1	1	1	1	0	-1	1	-1	0	0	36,36%	72,73%			
ORIG	ACC	ET	STACK	5	1	0	0	1	-1	-1	1	0	-1	0	0	27,27%	72,73%			
ORIG	ACC	ET	STACK	6	0	1	0	1	0	0	-1	0	-1	-1	1	1	36,36%		72,73%	
ORIG	ACC	ET	STACK BAG	1	0	1	1	0	1	1	1	1	1	0	1	1	72,73%	57,58%	100,00%	90,91%
ORIG	ACC	ET	STACK BAG	2	0	1	1	1	1	1	0	1	0	0	0	54,55%	100,00%			
ORIG	ACC	ET	STACK BAG	3	0	1	1	1	1	-1	1	1	1	0	1	72,73%	90,91%			
ORIG	ACC	ET	STACK BAG	4	0	1	1	1	1	1	1	1	-1	1	0	72,73%	90,91%			
ORIG	ACC	ET	STACK BAG	5	0	0	1	1	-1	0	0	1	-1	0	-1	27,27%	72,73%			
ORIG	ACC	ET	STACK BAG	6	0	1	0	1	0	0	1	-1	1	1	0	45,45%	90,91%			
ORIG	ACC	ET	VOT	1	0	0	0	0	0	-1	0	0	0	0	0	0,00%	16,67%	90,91%	81,82%	
ORIG	ACC	ET	VOT	2	0	0	0	1	1	1	0	0	0	1	0	36,36%		100,00%		
ORIG	ACC	ET	VOT	3	0	0	0	0	0	0	0	0	1	-1	-1	0		9,09%		81,82%
ORIG	ACC	ET	VOT	4	0	1	0	0	1	0	0	0	-1	0	-1	18,18%		81,82%		
ORIG	ACC	ET	VOT	5	0	0	-1	1	-1	-1	1	0	-1	-1	0	18,18%		54,55%		
ORIG	ACC	ET	VOT	6	0	1	0	1	0	0	0	-1	-1	0	0	18,18%		81,82%		

ORIG	ACC	ET	VOT BAG	1	0	0	0	0	-1	1	1	0	0	0	0	18,18%	21,21%	90,91%	86,37%
ORIG	ACC	ET	VOT BAG	2	0	0	0	0	0	1	0	0	0	0	0	9,09%		100,00%	
ORIG	ACC	ET	VOT BAG	3	0	0	0	0	0	1	0	0	1	-1	0	18,18%		90,91%	
ORIG	ACC	ET	VOT BAG	4	0	1	0	1	1	0	0	0	-1	-1	-1	27,27%		72,73%	
ORIG	ACC	ET	VOT BAG	5	0	0	0	1	0	0	1	-1	-1	0	0	18,18%		81,82%	
ORIG	ACC	ET	VOT BAG	6	0	1	0	1	0	0	1	-1	-1	0	1	36,36%	16,67%	81,82%	96,97%
ORIG	ACC	RF	BAG	1	0	0	0	0	0	1	0	0	0	0	-1	9,09%		90,91%	
ORIG	ACC	RF	BAG	2	0	0	0	1	1	0	0	0	0	0	0	18,18%		100,00%	
ORIG	ACC	RF	BAG	3	1	0	0	0	0	0	0	1	0	1	0	27,27%		100,00%	
ORIG	ACC	RF	BAG	4	0	0	0	1	0	0	0	0	0	0	0	9,09%		100,00%	
ORIG	ACC	RF	BAG	5	0	0	1	1	0	0	0	0	-1	1	0	27,27%	27,27%	90,91%	89,4%
ORIG	ACC	RF	BAG	6	1	0	0	0	0	0	0	0	0	0	0	9,09%		100,00%	
ORIG	ACC	RF	STACK	1	0	0	1	0	1	1	0	1	0	0	1	45,45%		100,00%	
ORIG	ACC	RF	STACK	2	0	0	0	0	1	1	0	-1	0	0	0	18,18%		90,91%	
ORIG	ACC	RF	STACK	3	0	0	0	0	0	1	1	1	1	1	0	36,36%		100,00%	
ORIG	ACC	RF	STACK	4	0	1	0	0	0	1	1	0	-1	0	0	27,27%	48,48%	90,91%	86,37%
ORIG	ACC	RF	STACK	5	0	0	0	0	-1	0	0	0	-1	1	1	18,18%		81,82%	
ORIG	ACC	RF	STACK	6	0	0	0	0	0	-1	0	0	-1	-1	1	18,18%		72,73%	
ORIG	ACC	RF	STACK BAG	1	0	1	1	0	-1	1	0	1	0	1	1	54,55%		90,91%	
ORIG	ACC	RF	STACK BAG	2	0	1	1	0	1	1	0	1	-1	0	0	45,45%		90,91%	
ORIG	ACC	RF	STACK BAG	3	0	1	0	1	1	1	0	1	0	0	0	45,45%	9,09%	100,00%	84,85%
ORIG	ACC	RF	STACK BAG	4	0	1	0	1	-1	1	-1	1	-1	0	0	36,36%		81,82%	
ORIG	ACC	RF	STACK BAG	5	1	0	1	1	-1	1	1	1	-1	0	1	63,64%		72,73%	
ORIG	ACC	RF	STACK BAG	6	0	1	0	1	-1	1	0	-1	1	0	1	45,45%		81,82%	
ORIG	ACC	RF	VOT	1	0	0	0	0	-1	1	-1	1	0	0	0	18,18%	9,09%	81,82%	84,85%
ORIG	ACC	RF	VOT	2	0	0	0	0	-1	1	0	0	0	0	0	9,09%		90,91%	
ORIG	ACC	RF	VOT	3	0	0	0	0	0	1	0	0	0	1	0	18,18%		100,00%	
ORIG	ACC	RF	VOT	4	-1	1	0	0	0	0	0	0	-1	0	0	9,09%		81,82%	
ORIG	ACC	RF	VOT	5	0	0	0	0	-1	0	0	0	-1	0	0	0,00%		81,82%	
ORIG	ACC	RF	VOT	6	0	0	0	0	0	-1	0	0	-1	-1	0	0,00%	9,09%	72,73%	84,85%
ORIG	ACC	RF	VOT BAG	1	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
ORIG	ACC	RF	VOT BAG	2	0	0	0	0	-1	0	0	0	0	0	-1	0,00%		81,82%	
ORIG	ACC	RF	VOT BAG	3	0	0	0	0	1	1	0	0	0	0	0	18,18%		100,00%	
ORIG	ACC	RF	VOT BAG	4	0	1	0	1	-1	0	0	0	0	0	0	18,18%	36,36%	90,91%	98,49%
ORIG	ACC	RF	VOT BAG	5	-1	0	0	0	-1	-1	0	0	-1	0	0	0,00%		63,64%	
ORIG	ACC	RF	VOT BAG	6	0	1	0	1	-1	0	0	-1	-1	0	0	18,18%		72,73%	
ORIG	MCC	DT	BAG	1	0	1	0	1	0	0	1	1	0	0	1	45,45%		100,00%	
ORIG	MCC	DT	BAG	2	0	1	0	0	0	1	1	0	0	0	1	36,36%	39,39%	100,00%	89,4%
ORIG	MCC	DT	BAG	3	0	0	0	1	0	1	0	0	0	1	1	36,36%		100,00%	
ORIG	MCC	DT	BAG	4	0	0	1	1	0	0	1	1	1	1	0	54,55%		100,00%	
ORIG	MCC	DT	BAG	5	0	0	1	0	0	0	0	0	0	0	1	18,18%		100,00%	
ORIG	MCC	DT	BAG	6	0	0	0	0	1	0	1	0	-1	1	0	27,27%		90,91%	
ORIG	MCC	DT	STACK	1	0	0	0	0	1	0	0	0	0	1	0	18,18%	63,64%	100,00%	95,46%
ORIG	MCC	DT	STACK	2	0	0	1	0	1	1	0	0	1	1	0	45,45%		100,00%	
ORIG	MCC	DT	STACK	3	1	1	0	1	0	1	0	-1	-1	0	0	36,36%		81,82%	
ORIG	MCC	DT	STACK	4	1	0	1	-1	1	0	0	1	1	1	0	54,55%		90,91%	
ORIG	MCC	DT	STACK	5	-1	-1	1	0	-1	0	1	0	1	1	0	36,36%		72,73%	
ORIG	MCC	DT	STACK	6	1	1	0	0	1	-1	0	0	1	1	0	45,45%	12,12%	90,91%	77,28%
ORIG	MCC	DT	STACK BAG	1	0	0	1	0	1	1	0	1	1	1	1	63,64%		100,00%	
ORIG	MCC	DT	STACK BAG	2	0	1	1	0	1	1	0	0	1	1	1	63,64%		100,00%	
ORIG	MCC	DT	STACK BAG	3	1	0	1	0	1	1	1	0	1	0	0	54,55%		100,00%	
ORIG	MCC	DT	STACK BAG	4	1	1	1	0	1	1	1	1	1	1	1	90,91%		100,00%	
ORIG	MCC	DT	STACK BAG	5	-1	-1	1	0	0	0	0	0	1	1	1	36,36%	12,12%	81,82%	77,28%
ORIG	MCC	DT	STACK BAG	6	-1	1	1	1	0	1	1	1	1	1	0	72,73%		90,91%	
ORIG	MCC	DT	VOT	1	-1	0	0	0	0	0	0	0	0	0	0	0,00%		90,91%	
ORIG	MCC	DT	VOT	2	-1	0	0	-1	-1	0	0	0	1	0	0	9,09%		72,73%	
ORIG	MCC	DT	VOT	3	-1	0	0	0	0	0	0	0	1	-1	0	9,09%		81,82%	
ORIG	MCC	DT	VOT	4	-1	0	0	-1	1	0	0	0	1	1	-1	27,27%	12,12%	72,73%	77,28%
ORIG	MCC	DT	VOT	5	-1	0	0	0	-1	0	-1	0	1	-1	0	9,09%		63,64%	
ORIG	MCC	DT	VOT	6	-1	0	0	0	1	0	0	0	-1	1	0	18,18%		81,82%	

ORIG	MCC	DT	VOT BAG	1	0	0	0	0	-1	0	0	0	1	0	0	9,09%	28,79%	90,91%	90,91%
ORIG	MCC	DT	VOT BAG	2	0	0	0	1	0	0	0	1	1	0	0	27,27%		100,00%	
ORIG	MCC	DT	VOT BAG	3	1	0	0	0	0	1	0	0	1	-1	0	27,27%		90,91%	
ORIG	MCC	DT	VOT BAG	4	1	-1	0	0	1	0	0	1	1	1	0	45,45%		90,91%	
ORIG	MCC	DT	VOT BAG	5	-1	-1	1	0	-1	0	0	0	1	1	1	36,36%		72,73%	
ORIG	MCC	DT	VOT BAG	6	0	1	0	0	0	0	0	1	0	1	0	27,27%		100,00%	
ORIG	MCC	ET	BAG	1	0	0	0	0	0	0	0	-1	1	0	0	9,09%	15,15%	90,91%	96,97%
ORIG	MCC	ET	BAG	2	0	1	0	0	0	0	0	0	0	1	0	18,18%		100,00%	
ORIG	MCC	ET	BAG	3	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
ORIG	MCC	ET	BAG	4	0	0	0	1	0	0	0	0	-1	0	0	9,09%		90,91%	
ORIG	MCC	ET	BAG	5	1	1	0	1	0	0	1	1	1	0	0	54,55%		100,00%	
ORIG	MCC	ET	BAG	6	0	0	0	0	0	0	0	0	0	0	0	0,00%		100,00%	
ORIG	MCC	ET	STACK	1	0	0	1	1	1	1	1	1	0	0	1	63,64%	51,52%	100,00%	86,37%
ORIG	MCC	ET	STACK	2	0	0	1	1	1	1	0	0	0	1	0	45,45%		100,00%	
ORIG	MCC	ET	STACK	3	1	1	0	1	1	1	0	1	1	0	1	72,73%		100,00%	
ORIG	MCC	ET	STACK	4	1	-1	1	0	1	0	-1	1	-1	0	0	36,36%		72,73%	
ORIG	MCC	ET	STACK	5	1	1	0	1	-1	-1	1	0	-1	-1	0	36,36%		63,64%	
ORIG	MCC	ET	STACK	6	-1	1	0	1	0	1	0	1	1	-1	1	54,55%		81,82%	
ORIG	MCC	ET	STACK BAG	1	0	1	1	1	1	1	1	1	0	0	1	72,73%	68,18%	100,00%	89,4%
ORIG	MCC	ET	STACK BAG	2	0	1	1	1	1	1	1	1	1	0	1	81,82%		100,00%	
ORIG	MCC	ET	STACK BAG	3	0	1	1	1	1	1	1	1	1	1	1	90,91%		100,00%	
ORIG	MCC	ET	STACK BAG	4	-1	1	1	1	1	1	1	1	-1	-1	0	63,64%		72,73%	
ORIG	MCC	ET	STACK BAG	5	0	1	1	1	-1	0	1	1	-1	0	-1	45,45%		72,73%	
ORIG	MCC	ET	STACK BAG	6	-1	1	0	1	0	0	1	1	1	1	0	54,55%		90,91%	
ORIG	MCC	ET	VOT	1	0	0	0	0	0	-1	0	0	0	0	0	0,00%	18,18%	90,91%	77,27%
ORIG	MCC	ET	VOT	2	0	0	0	1	1	1	0	0	0	1	0	36,36%		100,00%	
ORIG	MCC	ET	VOT	3	0	0	0	0	0	0	0	0	1	-1	-1	9,09%		81,82%	
ORIG	MCC	ET	VOT	4	-1	1	0	0	1	0	0	0	-1	-1	-1	18,18%		63,64%	
ORIG	MCC	ET	VOT	5	0	-1	-1	1	-1	-1	1	0	-1	-1	0	18,18%		45,45%	
ORIG	MCC	ET	VOT	6	-1	1	0	1	0	0	0	-1	1	0	0	27,27%		81,82%	
ORIG	MCC	ET	VOT BAG	1	0	0	0	0	-1	1	1	0	0	0	-1	18,18%	27,27%	81,82%	84,85%
ORIG	MCC	ET	VOT BAG	2	0	0	0	0	0	1	0	0	1	0	0	18,18%		100,00%	
ORIG	MCC	ET	VOT BAG	3	0	0	0	0	0	0	1	0	0	1	0	18,18%		100,00%	
ORIG	MCC	ET	VOT BAG	4	-1	1	0	1	1	0	0	0	-1	-1	-1	27,27%		63,64%	
ORIG	MCC	ET	VOT BAG	5	0	1	0	1	-1	0	1	1	-1	0	0	36,36%		81,82%	
ORIG	MCC	ET	VOT BAG	6	-1	1	0	1	0	0	1	-1	1	0	1	45,45%		81,82%	
ORIG	MCC	RF	BAG	1	0	0	0	0	0	1	0	0	0	0	-1	9,09%	16,67%	90,91%	96,97%
ORIG	MCC	RF	BAG	2	0	0	0	1	1	0	0	0	0	0	0	18,18%		100,00%	
ORIG	MCC	RF	BAG	3	1	0	0	0	0	0	0	0	1	0	1	27,27%		100,00%	
ORIG	MCC	RF	BAG	4	0	0	0	1	0	0	0	0	0	0	0	9,09%		100,00%	
ORIG	MCC	RF	BAG	5	0	0	1	1	1	0	0	0	-1	0	0	27,27%		90,91%	
ORIG	MCC	RF	BAG	6	1	0	0	0	0	0	0	0	0	0	0	9,09%		100,00%	
ORIG	MCC	RF	STACK	1	0	1	1	0	1	1	0	1	0	0	1	54,55%	39,39%	100,00%	87,88%
ORIG	MCC	RF	STACK	2	0	0	0	0	0	1	-1	0	0	1	0	1		90,91%	
ORIG	MCC	RF	STACK	3	0	0	0	0	0	0	1	1	1	0	0	36,36%		100,00%	
ORIG	MCC	RF	STACK	4	-1	1	0	-1	0	1	1	0	-1	0	0	27,27%		72,73%	
ORIG	MCC	RF	STACK	5	1	1	0	0	1	0	0	0	-1	1	1	45,45%		90,91%	
ORIG	MCC	RF	STACK	6	-1	1	0	1	-1	0	0	-1	1	1	1	45,45%		72,73%	
ORIG	MCC	RF	STACK BAG	1	0	1	1	1	-1	1	0	1	0	1	1	63,64%	59,09%	90,91%	86,37%
ORIG	MCC	RF	STACK BAG	2	0	1	1	1	1	1	0	1	0	0	0	54,55%		100,00%	
ORIG	MCC	RF	STACK BAG	3	0	1	1	1	1	1	0	1	1	1	0	72,73%		100,00%	
ORIG	MCC	RF	STACK BAG	4	1	-1	0	1	-1	1	-1	1	-1	0	0	36,36%		63,64%	
ORIG	MCC	RF	STACK BAG	5	1	-1	1	0	1	1	1	1	-1	0	1	63,64%		81,82%	
ORIG	MCC	RF	STACK BAG	6	-1	1	0	1	1	1	0	-1	1	1	1	63,64%		81,82%	
ORIG	MCC	RF	VOT	1	0	0	0	0	-1	1	-1	1	0	0	0	18,18%	15,15%	81,82%	80,31%
ORIG	MCC	RF	VOT	2	-1	0	0	0	-1	1	0	0	1	0	0	18,18%		81,82%	
ORIG	MCC	RF	VOT	3	0	0	0	0	0	0	1	0	0	1	0	100,00%		100,00%	
ORIG	MCC	RF	VOT	4	-1	1	0	0	0	0	0	0	-1	0	0	9,09%		81,82%	
ORIG	MCC	RF	VOT	5	-1	-1	0	0	1	0	0	0	-1	0	0	9,09%		72,73%	
ORIG	MCC	RF	VOT	6	-1	1	0	-1	-1	0	0	-1	1	0	0	18,18%		63,64%	

ORIG	MCC	RF	VOT BAG	1	0	0	0	0	0	0	0	0	0	0	0	0,00%	21,21%	100,00%	86,37%
ORIG	MCC	RF	VOT BAG	2	0	0	0	1	-1	0	0	0	1	0	0	18,18%		90,91%	
ORIG	MCC	RF	VOT BAG	3	1	0	0	0	1	1	0	0	0	0	0	27,27%		100,00%	
ORIG	MCC	RF	VOT BAG	4	1	1	0	1	1	0	0	1	-1	-1	0	45,45%		81,82%	
ORIG	MCC	RF	VOT BAG	5	0	-1	0	0	1	-1	0	0	-1	0	0	9,09%		72,73%	
ORIG	MCC	RF	VOT BAG	6	-1	1	0	1	-1	0	0	-1	1	0	0	27,27%		72,73%	

APÊNDICE B – Resultados das Métricas

TABELAS COM OS VALORES DAS MÉTRICAS AFERIDAS DURANTE A ETAPA DE AVALIAÇÃO DO PROCESSO DE DCBD

CONJUNTO DE DATASETS: ORIGINAL

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: INDIVIDUAL (ORIG)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	70.37	41.51	76.22	52.33	76.94	50.09	72.75	45.31	85.71	65.87	90.76	68.40	71.49	45.42	81.90	43.92	78.00	52.27	77.75	54.91	75.57	44.75
	2	67.21	28.56	73.47	42.86	75.22	50.54	77.19	50.09	78.10	54.83	83.79	56.92	73.85	47.07	79.37	51.72	80.00	40.11	76.97	44.16	68.46	36.83
	3	71.45	31.90	76.72	41.98	75.96	52.37	79.64	50.58	77.25	54.54	79.56	53.31	74.75	48.59	81.17	59.51	84.09	41.27	79.88	42.96	71.69	44.62
	4	73.84	36.59	84.01	50.07	74.59	48.48	83.09	50.25	78.91	59.41	80.58	57.34	73.13	44.52	75.50	48.76	87.04	48.09	79.87	35.75	79.41	58.84
	5	76.47	37.84	83.69	42.57	77.81	54.08	84.04	48.03	79.55	58.90	85.47	70.28	85.60	56.97	83.45	67.08	93.11	60.59	82.29	37.51	76.51	52.02
	6	79.80	31.42	86.61	35.63	75.69	45.65	86.27	48.03	75.38	49.13	82.50	66.03	88.61	59.43	77.24	57.04	95.24	56.73	86.93	44.70	81.05	58.04
ET	1	75.91	53.09	80.10	60.37	84.42	66.22	76.27	52.21	88.57	70.79	91.42	70.05	83.31	67.01	85.95	54.46	84.00	62.68	81.33	59.17	78.32	49.11
	2	73.77	40.15	80.09	53.74	81.36	63.09	80.80	56.18	81.43	60.15	88.65	70.38	84.40	65.90	84.53	62.45	87.59	54.39	80.81	48.45	76.17	53.07
	3	77.41	42.94	83.62	55.85	80.28	60.59	84.73	60.31	79.00	57.78	88.94	74.63	87.03	68.52	84.59	66.24	90.98	60.86	84.39	51.89	76.00	53.23
	4	78.71	39.02	86.45	56.47	82.79	64.90	86.47	58.94	75.77	52.45	86.16	70.55	89.52	72.22	83.48	65.21	95.93	80.33	87.67	55.86	82.94	66.35
	5	82.75	42.24	87.77	47.24	84.25	66.52	88.42	60.46	81.01	61.79	87.34	74.80	87.47	59.84	83.09	66.16	95.78	72.84	90.13	57.78	85.08	69.46
	6	84.72	40.81	91.79	58.28	85.23	66.03	89.68	55.98	83.85	67.33	87.50	75.30	90.32	60.30	85.64	70.98	97.22	77.10	89.27	46.14	84.56	67.10
RF	1	73.00	47.81	76.81	53.74	79.96	56.86	76.88	53.66	88.57	71.07	89.79	63.73	81.24	63.78	82.37	42.73	80.86	57.40	80.65	59.19	80.42	53.76
	2	72.79	38.62	78.73	51.04	79.45	59.73	78.62	51.62	79.05	55.80	87.96	68.80	80.44	59.20	83.18	59.34	86.90	51.02	81.76	53.11	77.11	54.63
	3	77.74	43.82	82.60	51.07	79.81	60.18	82.99	58.24	77.30	55.18	84.40	62.75	83.69	63.63	83.02	62.63	90.24	59.12	83.78	50.05	75.47	52.82
	4	80.97	44.52	86.44	56.96	80.13	58.79	86.04	57.77	78.93	58.85	84.52	67.55	88.78	70.99	81.72	61.92	94.07	71.48	87.66	59.06	78.24	56.07
	5	85.88	50.47	88.36	49.24	82.28	62.03	89.29	63.13	78.07	55.12	87.74	75.23	86.67	60.61	83.08	66.14	96.17	78.87	89.24	55.60	80.63	59.96
	6	85.99	41.77	91.49	53.85	84.31	64.29	89.01	53.47	82.31	64.12	83.75	68.10	89.18	54.58	88.15	76.02	96.42	64.86	90.21	49.17	84.91	66.16

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: BAGGING (ORIG)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	72.47	46.43	80.68	61.83	82.67	62.18	78.27	56.51	83.57	60.95	90.60	68.37	78.62	58.54	85.17	58.02	84.29	63.82	81.15	60.65	78.32	51.21
	2	72.13	37.21	79.40	52.55	77.34	54.85	76.81	46.75	79.05	56.25	87.96	69.18	78.90	57.76	83.62	62.05	88.28	57.92	80.04	49.66	73.86	49.62
	3	79.17	46.88	82.09	49.56	77.17	54.73	85.13	61.48	79.56	58.76	84.96	64.96	78.05	55.75	83.56	64.03	88.08	48.44	86.04	57.29	78.69	58.53
	4	76.83	36.15	86.16	58.24	78.84	56.67	87.94	63.58	80.28	61.52	84.53	66.54	87.05	66.66	82.29	63.49	92.96	66.50	85.71	50.07	82.06	64.35
	5	83.92	44.23	87.75	52.00	82.85	63.57	87.54	55.94	78.04	55.08	88.84	76.94	85.87	58.39	86.62	73.20	95.39	74.09	86.77	48.25	82.86	64.13
	6	84.74	40.19	90.27	51.51	82.46	58.95	91.07	63.60	83.08	64.84	84.17	68.73	92.88	71.52	88.60	76.99	96.03	65.23	90.20	56.00	82.46	62.20
ET	1	72.99	48.47	78.16	56.81	85.05	67.60	78.11	55.79	89.64	73.69	91.74	70.75	84.99	70.62	84.55	49.22	86.00	68.07	80.99	59.23	81.07	54.21
	2	74.75	43.61	82.38	59.39	81.14	63.12	82.78	59.98	85.71	69.55	87.96	67.54	85.49	68.23	85.42	64.51	87.59	53.84	82.54	54.10	75.94	53.28
	3	77.75	43.58	85.16	58.74	83.16	67.05	85.70	62.54	81.83	63.67	87.53	70.29	87.97	71.02	86.73	70.88	91.70	64.49	84.60	52.29	77.09	55.99
	4	82.79	50.01	88.61	63.30	83.33	65.95	89.21	66.80	81.51	63.88	87.18	72.16	89.28	71.14	85.24	69.13	94.44	73.86	86.58	52.77	82.06	64.84
	5	87.06	55.21	90.40	57.33	85.65	69.52	91.69	71.75	83.17	66.11	89.97	79.71	88.80	64.06	85.56	71.12	96.15	78.40	90.14	58.97	85.40	69.64
	6	88.87	53.53	92.10	57.06	85.85	66.72	91.07	62.43	85.38	70.31	87.50	75.48	90.60	61.22	88.99	77.79	97.22	77.10	90.44	48.47	85.26	68.45
RF	1	72.46	48.29	76.41	52.66	80.28	56.87	77.19	54.24	90.00	74.70	91.09	67.61	80.30	62.54	83.15	42.53	84.00	63.31	80.31	58.71	78.48	49.22
	2	71.80	37.67	80.08	54.11	79.46	60.05	80.98	57.41	84.76	68.09	87.73	66.34	79.12	57.91	84.08	61.40	87.93	55.78	83.50	56.84	76.87	55.00
	3	79.15	47.81	83.62	55.18	79.81	60.39	85.32	62.48	82.41	64.79	86.67	68.55	86.31	68.35	86.72	70.62	91.32	63.13	85.43	55.39	76.02	54.24
	4	81.67	47.81	88.61	64.13	83.59	66.71	88.79	65.90	78.97	58.77	86.85	71.30	88.77	71.26	82.89	63.96	94.81	74.60	87.88	58.25	80.88	62.34
	5	85.49	52.28	90.39	57.49	85.37	69.12	91.03	68.95	82.43	64.73	86.25	71.87	88.00	62.78	85.56	71.07	96.54	79.60	90.14	60.16	84.44	68.08
	6	88.45	51.48	92.10	57.06	85.23	65.60	91.53	64.33	82.31	63.70	87.08	74.46	90.03	60.27	89.00	77.72	96.82	72.25	90.91	50.15	86.67	70.22

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: STACKING (ORIG)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	66.29	35.51	81.55	62.40	80.50	57.10	78.89	57.54	93.33	82.51	94.43	79.62	71.03	42.26	85.89	49.78	71.58	39.21	83.90	64.87	83.51	59.66
	2	64.56	17.73	75.38	43.76	78.74	60.33	77.25	47.12	98.18	95.40	90.40	75.07	78.23	50.77	77.69	44.70	87.94	55.99	87.06	64.91	73.64	45.94
	3	75.33	24.28	89.91	71.15	74.31	50.43	85.75	65.37	72.89	49.99	86.84	71.58	85.00	61.23	73.38	45.96	86.25	14.62	84.81	40.33	72.10	45.65
	4	84.86	51.57	82.03	47.74	83.10	66.91	77.04	18.84	91.11	84.46	81.76	64.12	89.64	69.83	82.53	65.82	97.50	87.33	88.23	49.97	74.50	49.91
	5	71.71	-0.43	87.19	28.18	88.19	76.60	86.21	50.46	82.50	66.16	85.33	74.10	86.19	52.33	82.65	66.87	97.33	60.00	90.62	42.31	76.86	55.55
	6	84.07	0.00	95.79	68.82	79.26	50.86	90.62	56.96	90.00	80.74	81.87	62.71	89.24	58.67	88.35	78.54	94.67	0.00	94.43	63.97	81.47	54.73
ET	1	77.43	54.70	85.53	71.81	95.76	90.61	86.86	74.93	94.92	85.91	97.75	91.73	91.42	82.84	93.90	79.11	85.00	61.94	82.31	60.21	90.78	75.52
	2	81.37	57.81	84.60	64.71	91.20	83.48	88.41	73.78	93.18	85.59	93.57	82.28	90.40	77.25	85.78	66.92	91.76	68.93	89.89	71.95	81.27	64.16
	3	90.00	73.17	90.08	70.63	88.18	77.98	90.67	76.53	89.33	80.90	96.95	92.97	89.60	73.39	94.37	88.19	96.32	83.10	86.11	46.84	86.53	75.28
	4	87.24	42.86	91.65	71.43	94.10	88.45	91.35	73.12	91.11	84.46	88.09	76.27	87.83	73.59	93.39	87.41	97.42	73.66	86.58	46.93	82.75	65.91
	5	92.00	71.67	92.00	54.32	87.68	74.92	93.99	79.23	86.79	74.79	87.71	76.61	93.68	77.28	82.50	68.73	97.50	87.33	89.69	37.27	87.58	74.45
	6	85.90	45313	93.89	49.63	88.13	72.97	96.06	83.23	74.64	53.38	89.45	80.29	92.24	68.44	91.21	83.68	98.67	40.00	93.60	64.85	91.25	80.47
RF	1	74.00	49.79	82.76	67.07	92.91	84.87	84.02	69.34	91.25	79.95	95.52	83.73	81.24	63.96	94.87	75.83	88.58	70.33	81.36	56.51	89.20	72.77
	2	81.18	56.93	81.17	58.75	87.97	76.69	77.84	47.67	85.91	69.34	97.60	93.11	85.97	70.80	80.00	57.52	88.09	51.39	84.78	59.70	82.86	67.68
	3	87.17	64.61	84.55	52.42	82.25	65.59	82.34	54.50	77.33	57.92	92.05	79.94	90.93	78.93	88.96	77.88	96.47	84.44	83.39	43.32	81.52	66.10
	4	76.76	3.43	94.55	82.48	89.00	78.42	84.68	43.70	78.00	55.79	88.01	75.83	93.83	80.27	86.53	74.25	92.50	55.90	89.57	61.72	80.64	62.72
	5	78.67	45.34	87.14	0.00	82.68	66.22	92.39	78.02	90.00	82.47	88.00	78.17	90.74	71.05	79.04	62.69	92.33	33.63	95.32	78.98	92.22	84.53
	6	84.57	11.32	90.89	26.79	83.10	64.54	85.38	43.15	71.43	49.60	83.96	68.68	92.24	64.60	91.21	83.13	97.33	0.00	93.57	63.58	92.94	84.04

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: VOTING (ORIG)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	57.04	14.41	75.58	50.79	76.20	47.67	77.05	54.03	84.49	61.31	91.39	70.16	70.62	41.41	85.45	53.49	70.48	33.79	80.24	56.53	77.99	48.83
	2	57.60	-11.51	74.10	39.48	73.20	49.68	68.66	26.62	81.28	59.75	86.92	64.58	83.97	64.55	74.53	44.36	86.27	44.54	77.74	42.09	70.52	40.46
	3	68.63	-6.24	81.34	44.22	74.40	49.71	79.91	47.93	75.64	51.22	81.17	58.47	81.23	54.64	79.96	58.73	88.16	44.30	76.18	45342	66.05	32.96
	4	75.22	28.72	85.61	51.06	74.55	48.13	80.32	17.99	87.78	77.23	74.09	42.65	80.17	50.19	76.52	50.94	93.82	63.10	88.49	52.16	66.67	32.25
	5	75.42	-8.13	87.43	45.33	79.39	55.36	80.48	33.94	71.39	43.19	83.97	67.82	84.90	44.81	79.15	58.67	93.75	53.66	85.81	29.91	74.74	47.99
	6	82.29	-2.67	93.95	68.07	79.68	50.91	88.63	51.48	90.00	80.74	77.71	56.18	86.80	47.45	87.52	76.36	92.08	-1.43	92.22	53.42	80.20	54.55
ET	1	74.58	51.17	76.81	53.96	83.60	63.47	81.10	61.74	89.34	73.71	92.43	72.30	87.50	75.07	88.04	62.92	80.95	55.57	76.84	48.45	81.19	54.27
	2	72.81	34.79	81.85	57.64	80.94	62.36	87.38	71.40	88.85	77.35	92.31	79.40	82.49	60.46	82.02	59.85	88.50	58.39	85.38	61.64	72.83	49.40
	3	83.66	59.51	86.38	60.03	77.60	57.46	88.28	70.77	72.00	48.34	87.75	71.57	87.60	70.12	89.49	77.25	92.87	62.33	78.28	22.54	75.77	52.00
	4	72.79	0.19	91.90	72.59	83.32	66.44	90.17	71.01	81.56	65.91	80.47	59.31	87.67	65.86	83.33	65.57	92.50	55.90	84.87	36.91	73.52	47.07
	5	87.00	56.11	88.52	36.56	85.15	68.97	91.24	69.84	81.39	63.73	77.79	56.13	92.02	75.21	77.84	58.55	94.92	67.29	85.07	25.51	83.16	65.76
	6	83.62	14.22	93.89	49.63	82.63	59.17	93.90	74.74	84.64	70.26	81.81	64.68	89.61	56.33	86.19	72.70	96.00	40.00	90.74	53.38	87.12	71.90
RF	1	64.11	28.63	73.71	48.43	83.58	64.25	82.18	63.82	88.16	73.17	92.45	74.37	75.00	50.52	86.50	53.51	82.86	60.01	78.52	52.29	79.06	49.11
	2	64.04	14.69	74.96	45.31	79.53	60.51	81.34	56.54	81.41	62.11	91.54	76.83	82.49	60.94	81.28	61.43	86.27	45.46	77.74	38.56	70.49	41.99
	3	81.44	51.77	80.51	42.58	75.20	51.39	83.14	56.54	75.64	50.91	88.74	73.15	89.08	74.29	86.88	72.50	92.87	60.24	78.97	35.83	78.54	57.28
	4	75.29	10.79	93.68	80.31	83.36	66.05	84.53	49.91	73.11	50.09	82.51	62.87	84.37	59.45	80.48	61.47	93.75	62.23	87.04	49.74	79.48	59.03
	5	83.08	29.67	84.62	-1.83	79.52	56.08	91.32	72.07	84.17	65.91	86.40	73.50	86.68	55.14	82.61	67.40	93.67	60.95	88.09	47.76	86.32	71.60
	6	82.19	-2.26	91.89	40.56	84.74	63.89	87.86	42.07	69.64	41.26	82.00	64.08	90.56	61.71	86.19	74.20	93.42	0.00	89.94	50.03	87.19	71.05

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: VOTING BAGGING (ORIG)

	PX	ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
		ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	64.78	45471	74.33	48.96	79.89	54.94	76.04	52.21	87.06	69.29	92.45	74.94	77.50	56.27	86.03	54.82	76.19	41.98	80.75	57.80	77.97	49.06
	2	70.70	30.85	77.18	50.92	73.87	48.45	81.35	56.63	79.74	59.86	89.23	72.70	78.12	51.91	82.79	64.20	84.12	35.13	82.78	55.04	64.25	28.37
	3	82.55	55.73	81.34	43.80	73.60	47.34	81.78	51.43	70.00	41.74	88.70	73.53	86.74	68.58	83.40	67.50	89.34	44.35	75.54	18.44	72.25	44.64
	4	78.97	45618	86.48	56.01	81.70	62.85	80.96	40.70	83.56	70.24	82.63	64.58	85.20	60.70	86.29	75.11	92.50	55.90	88.47	51.16	75.57	52.02
	5	80.58	45436	87.57	31.62	85.11	69.11	87.70	57.52	67.22	34.96	83.97	66.14	89.41	66.33	82.55	66.30	91.17	27.29	88.83	50.52	78.95	56.55
	6	86.19	32.74	90.89	54.87	77.53	48.49	92.42	70.23	71.43	52.23	83.33	66.94	95.24	82.00	86.00	73.33	97.33	60.00	93.05	63.49	81.31	55.97
ET	1	72.77	46.73	81.33	61.94	85.18	67.46	78.58	57.07	89.26	72.20	93.51	78.97	86.88	74.17	85.45	50.78	80.95	55.57	79.10	54.10	75.28	36.71
	2	78.13	48.07	77.29	45.96	78.15	58.27	83.16	61.33	87.18	72.86	93.85	83.96	85.37	68.19	82.76	60.84	86.14	36.79	81.55	49.48	74.40	50.47
	3	82.48	55.39	84.75	55.89	77.60	56.65	87.66	68.84	77.45	55.06	92.51	82.89	90.68	78.27	86.01	70.63	95.22	78.11	79.61	27.84	78.46	58.17
	4	74.04	1.58	92.77	76.71	85.14	69.32	92.27	77.13	87.78	76.16	85.85	71.52	89.30	71.52	85.24	69.96	95.00	75.90	87.01	43.52	77.48	54.89
	5	88.33	60.91	88.52	29.94	86.06	71.44	93.47	78.87	69.72	39.31	86.47	72.64	92.92	78.51	86.08	73.24	96.25	73.66	88.83	51.99	83.16	65.28
	6	82.19	3.62	93.89	58.45	81.68	56.91	92.39	67.85	82.14	65.11	80.48	62.19	95.28	82.02	87.62	75.92	96.08	40.00	89.94	48.89	89.54	77.39
RF	1	72.77	46.90	80.08	60.33	85.16	67.29	81.12	62.51	83.38	62.05	91.37	70.12	80.00	60.46	85.47	50.20	81.90	57.66	79.67	55.21	80.11	52.54
	2	79.24	51.65	76.47	50.38	81.65	64.73	84.31	64.17	81.15	59.11	89.23	70.01	81.01	57.16	86.55	70.39	87.39	49.32	82.88	53.20	68.92	38.06
	3	81.50	47.74	86.41	60.23	80.80	62.34	88.30	70.78	83.45	67.01	88.70	73.33	85.17	63.71	85.14	68.63	92.87	67.57	79.63	37.57	75.85	52.75
	4	77.79	45548	89.17	67.25	84.23	67.43	92.29	76.96	81.78	67.76	83.68	65.92	87.67	66.98	86.24	72.13	91.32	49.59	83.44	32.05	75.48	50.60
	5	83.08	49.42	83.67	-3.65	87.92	76.38	91.32	71.36	77.22	54.15	86.62	72.06	90.24	69.09	82.61	66.64	94.92	67.29	86.58	47.98	84.21	68.18
	6	77.90	0.72	94.89	65.50	81.63	56.39	92.45	69.24	77.50	54.20	83.33	68.03	91.52	66.65	88.95	79.04	96.08	40.00	89.20	43.56	87.19	71.55

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: STACKING BAGGING (ORIG)

	PX	ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
		ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	68.74	35.54	80.50	62.15	87.65	73.89	77.84	55.04	93.50	83.40	98.86	96.14	80.92	61.79	91.67	72.00	87.63	67.85	89.74	76.52	80.96	58.83
	2	77.94	42.15	83.60	64.92	83.63	67.64	83.69	61.49	85.00	66.95	94.37	84.68	84.37	63.71	84.49	64.76	89.19	55.01	88.18	65.97	74.58	52.90
	3	85.15	57.78	85.37	58.51	83.08	66.24	81.75	51.79	80.89	63.72	95.84	89.92	91.73	78.60	84.62	70.42	93.82	74.11	86.44	55.86	72.62	45.81
	4	86.10	21.96	90.65	73.24	84.98	70.42	85.40	53.86	87.33	76.12	86.76	72.99	87.83	63.03	89.79	81.27	96.17	80.95	88.77	53.16	84.80	69.88
	5	77.43	28.44	89.14	31.89	88.00	75.15	88.72	53.85	72.50	48.96	86.75	74.06	90.91	66.77	81.40	63.50	97.33	53.63	95.20	70.69	83.86	67.01
	6	80.29	31.57	90.84	51.46	85.56	66.20	93.85	73.35	75.71	52.46	90.99	83.16	98.05	91.76	88.57	77.97	100.00	60.00	94.40	59.30	85.51	66.33
ET	1	81.38	63.74	92.67	86.11	96.97	93.24	84.26	70.44	96.08	88.02	97.16	90.19	92.00	84.44	96.49	86.17	82.19	58.47	86.31	69.03	96.13	88.42
	2	81.44	58.43	85.27	67.17	93.33	87.32	84.68	65.45	96.52	92.94	97.60	93.33	90.77	79.17	90.40	79.76	89.34	45.22	84.00	58.64	82.61	69.11
	3	83.75	56.44	91.03	74.35	87.51	77.39	90.60	76.61	83.11	67.31	94.14	87.69	93.57	84.01	92.42	85.01	96.40	85.22	88.89	45.52	86.14	75.88
	4	81.33	22.46	93.55	79.54	92.33	85.01	93.57	80.94	91.56	85.09	100.00	100.00	94.78	84.97	91.58	83.50	94.83	58.57	90.28	61.10	84.91	71.60
	5	89.50	63.51	87.14	0.00	92.89	85.55	96.27	87.30	75.00	51.62	86.67	75.14	90.91	75.28	92.50	86.79	98.75	93.66	93.60	67.45	83.99	67.88
	6	78.86	7.29	93.89	58.45	89.93	77.50	94.52	77.63	82.14	65.11	90.77	82.19	96.10	82.71	94.07	88.46	100.00	60.00	92.80	60.82	89.19	76.51
RF	1	76.00	52.22	87.96	76.16	98.14	95.86	83.72	70.00	87.50	70.71	96.08	84.90	81.89	67.20	95.44	81.05	79.37	57.32	90.62	79.96	92.59	79.02
	2	70.98	31.06	89.20	75.85	92.67	85.21	84.56	66.97	90.00	79.44	93.60	82.64	82.25	64.20	89.69	75.14	82.13	35.06	86.30	62.93	76.88	56.76
	3	85.37	52.79	91.94	76.24	86.41	75.75	90.60	76.61	89.78	80.72	92.89	82.57	90.20	74.87	91.43	82.85	92.65	56.79	87.35	48.77	76.90	54.81
	4	84.00	15.69	89.83	65.42	87.75	75.21	94.26	83.33	78.67	60.75	92.94	85.62	88.62	75.61	92.63	86.22	94.83	47.29	87.01	48.79	82.75	66.15
	5	90.67	53.17	88.00	26.83	92.84	86.48	94.05	79.47	86.11	70.40	93.50	86.76	91.73	68.87	88.58	79.99	97.42	87.29	91.42	62.32	89.80	79.94
	6	83.24	30.90	95.89	69.63	87.02	70.20	94.58	76.59	84.64	72.25	89.34	80.63	91.24	62.63	94.07	88.46	100.00	60.00	92.80	41.82	92.57	83.23

CONJUNTO DE DATASETS: FILTRADO

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: INDIVIDUAL (FILT)

	PX	ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
		ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	66.14	33.69	69.06	39.04	75.04	46.40	69.08	38.12	85.36	64.10	86.55	51.99	71.49	45.42	77.53	31.36	76.57	49.62	74.03	46.10	76.21	48.53
	2	68.20	30.20	73.46	42.46	71.83	43.53	70.29	37.36	76.67	52.18	82.87	54.03	73.85	47.07	76.89	43.38	79.66	33.07	76.60	42.61	69.89	38.98
	3	66.84	25.36	77.50	41.11	71.39	43.74	71.84	37.75	76.11	52.63	77.04	46.42	74.75	48.59	79.56	55.34	84.45	44.46	78.46	38.71	73.85	48.69
	4	69.36	26.53	82.39	48.78	71.41	41.52	82.25	50.31	79.60	62.13	80.59	57.14	73.13	44.52	76.96	51.78	90.00	58.55	81.17	40.59	77.06	53.63
	5	75.69	33.08	86.02	44.48	75.00	47.38	82.96	42.91	80.24	60.31	83.28	65.40	85.60	56.97	81.33	62.54	94.25	65.86	82.51	41.37	79.68	57.27
	6	80.63	34.27	88.14	29.61	76.00	45.69	86.03	48.72	74.62	46.79	80.83	61.98	88.61	59.43	77.68	57.77	94.83	57.50	82.26	35.39	82.81	62.45
ET	1	75.64	52.88	76.03	51.95	82.03	60.17	76.25	52.27	87.86	68.87	89.64	63.02	83.31	67.01	82.21	40.80	84.00	63.41	79.29	56.24	81.55	56.68
	2	73.11	39.74	78.47	50.32	79.87	59.68	78.80	51.84	83.33	64.47	85.41	60.31	84.40	65.90	84.29	62.15	87.59	53.20	84.07	57.63	78.03	56.90
	3	77.43	41.46	83.37	53.51	76.69	53.87	82.01	55.13	78.43	56.53	85.25	64.59	87.03	68.52	84.06	65.04	91.68	64.28	85.01	53.29	76.28	53.81
	4	80.22	43.10	87.24	58.82	81.21	61.31	87.31	60.63	80.87	62.42	85.21	68.26	89.52	72.22	82.89	64.08	92.59	61.95	87.45	53.64	83.53	66.89
	5	84.31	46.71	88.36	50.28	82.56	62.72	86.45	53.23	83.12	65.69	83.98	67.74	87.47	59.84	84.15	68.20	95.39	72.00	90.58	60.81	85.71	70.22
	6	86.82	47.59	91.49	54.10	85.23	65.01	90.84	62.09	84.62	68.40	80.00	61.15	90.32	60.30	85.20	70.08	96.42	69.00	90.92	53.20	87.72	71.89
RF	1	69.82	41.99	71.58	43.16	78.85	54.45	72.43	44.87	84.29	60.73	87.68	56.87	81.24	63.78	80.96	39.63	80.57	56.22	78.95	55.93	78.96	49.72
	2	71.80	36.70	78.28	50.49	79.25	58.56	77.18	49.43	81.90	62.90	85.88	61.86	80.44	59.20	83.18	58.96	86.21	52.05	81.20	50.25	77.81	56.47
	3	77.40	44.47	80.81	48.03	76.19	53.09	81.62	53.80	83.54	67.19	82.98	60.37	83.69	63.63	80.11	55.14	90.25	58.31	83.38	50.40	75.75	53.81
	4	78.71	39.23	88.61	62.93	79.35	57.42	85.63	55.91	84.05	69.20	83.87	65.43	88.78	70.99	81.41	61.98	93.33	67.04	87.66	57.32	84.71	70.01
	5	83.92	44.57	89.22	49.12	83.13	64.12	85.38	51.77	82.43	64.64	86.62	72.44	86.67	60.61	81.67	63.28	95.77	74.30	88.56	55.37	84.44	67.83
	6	88.04	50.31	91.18	46.06	83.08	60.30	90.37	58.83	83.08	65.78	84.17	69.00	89.18	54.58	86.44	72.52	96.41	67.50	87.64	43.84	86.32	68.99

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: BAGGING (FILT)

	PX	ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
		ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	71.95	45.85	72.93	45.82	77.11	50.95	73.96	47.47	86.07	65.86	88.66	59.54	77.50	57.76	80.80	42.05	79.71	54.70	78.95	54.80	78.00	51.36
	2	71.80	36.63	76.89	46.84	77.76	55.52	75.90	45.36	78.57	55.25	85.18	59.84	77.14	53.99	82.72	58.93	87.93	55.49	80.81	51.25	74.77	50.69
	3	77.04	44.28	81.06	47.58	77.65	55.58	81.60	53.69	80.14	60.22	82.70	59.64	78.05	55.63	82.75	62.64	88.07	49.80	85.43	54.96	77.65	56.85
	4	77.95	38.71	88.07	63.00	80.15	59.34	85.62	56.89	80.89	62.49	82.56	63.34	86.29	66.34	82.30	64.39	92.22	64.81	85.92	49.59	81.47	63.67
	5	83.53	41.92	88.34	48.44	83.96	65.64	86.24	53.39	80.26	59.68	84.01	67.25	84.53	54.36	85.56	71.35	94.24	65.52	85.43	44.52	83.17	65.57
	6	85.58	42.24	90.28	48.28	82.46	59.26	90.84	62.41	82.31	64.18	82.92	66.89	92.60	69.31	89.00	77.76	95.62	52.97	88.56	48.51	84.21	64.48
ET	1	72.46	47.19	73.90	48.11	82.98	62.48	73.65	47.34	88.57	70.82	90.60	66.49	79.73	61.32	83.30	45.61	83.14	61.78	79.63	56.62	81.88	57.78
	2	75.08	44.80	81.69	57.73	78.60	57.33	80.07	54.62	85.24	68.37	86.33	63.19	84.84	67.00	84.53	62.83	87.59	55.26	82.92	55.08	81.54	63.87
	3	78.46	45.34	82.85	52.10	81.23	62.43	83.38	59.40	83.54	66.94	84.69	63.89	88.68	72.81	84.86	66.92	90.97	62.27	85.22	55.09	77.64	57.29
	4	82.44	48.72	87.52	59.09	82.79	64.42	88.79	65.58	82.18	64.87	85.52	68.53	88.53	70.00	83.18	64.58	93.33	67.34	88.31	58.17	85.29	71.18
	5	87.45	56.90	89.81	51.80	85.65	69.54	89.07	62.54	80.26	60.13	85.12	69.73	88.80	64.10	84.15	68.23	95.78	74.83	89.91	57.19	86.67	72.19
	6	88.86	53.24	91.79	55.98	86.15	67.92	92.21	67.48	86.15	72.28	87.08	74.55	91.45	65.56	88.13	76.00	95.62	60.77	91.38	53.73	88.77	74.31
RF	1	71.14	44.81	72.54	44.92	81.71	59.78	75.19	49.96	90.00	74.48	90.60	66.18	78.80	59.98	83.77	46.32	84.29	63.62	79.62	57.34	78.48	49.18
	2	72.13	39.09	79.86	53.99	78.61	57.67	80.25	56.17	84.29	67.30	86.34	62.90	79.34	57.19	85.65	65.34	88.28	56.33	83.11	56.73	78.28	57.55
	3	78.82	46.66	83.62	54.88	81.00	62.31	82.01	56.20	82.98	65.78	85.54	65.93	83.69	64.56	86.19	69.71	90.60	59.88	85.21	55.65	77.10	56.44
	4	82.07	48.10	87.80	61.84	84.91	68.94	87.31	61.46	80.91	62.50	86.52	70.68	87.53	68.25	82.88	64.17	93.70	68.57	89.40	63.04	80.88	62.53
	5	84.71	47.26	88.64	50.30	85.65	69.54	88.86	60.80	82.41	64.57	85.51	70.28	88.27	64.35	84.86	69.73	96.15	77.13	90.58	61.56	85.40	69.62
	6	87.63	49.06	91.18	51.92	86.77	68.84	91.98	66.31	84.62	68.16	84.58	69.67	92.31	69.66	87.29	74.29	96.42	67.40	91.15	53.42	89.12	75.23

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: STACKING (FILT)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	63.43	28.34	83.45	66.13	77.52	53.64	78.57	58.85	88.90	65.72	93.17	75.53	73.77	48.42	86.19	43.37	80.64	47.32	80.76	54.67	79.26	50.46
	2	76.91	44.30	76.38	50.41	80.77	66.32	74.84	41.59	92.73	84.07	97.53	93.00	86.15	68.32	70.77	31.85	86.84	33.04	84.46	54.44	74.13	50.51
	3	76.54	29.82	84.55	58.36	74.51	51.46	81.63	47.08	77.56	53.83	82.68	57.77	79.17	49.05	80.91	61.72	93.82	61.02	84.44	43.74	82.48	65.77
	4	72.48	3.48	92.42	75.78	82.52	65.86	80.71	27.58	78.22	59.00	84.71	68.54	92.09	77.21	76.37	48.38	95.00	63.06	84.13	48.63	80.29	63.16
	5	81.58	25.62	87.00	33.01	89.74	79.50	87.57	48.24	92.50	84.83	77.67	52.38	89.83	64.12	81.25	63.45	97.33	60.00	92.06	58.52	79.67	58.62
	6	82.64	0.00	93.74	59.20	79.77	54.49	85.38	30.87	83.57	67.22	76.15	54.68	95.32	84.67	86.81	75.75	98.67	60.00	94.30	55.82	91.25	81.48
ET	1	64.11	30.97	80.45	64.02	93.33	85.25	77.98	57.01	93.67	82.96	98.29	94.10	85.22	70.91	92.08	68.00	82.79	55.33	89.86	77.63	91.61	71.15
	2	77.39	46.58	78.17	56.30	94.27	88.55	75.97	45.60	89.55	76.37	98.33	95.99	81.00	61.29	87.57	74.38	91.76	67.95	84.99	58.23	72.96	44.92
	3	82.72	57.41	89.41	70.59	84.70	72.32	86.55	62.89	88.00	76.74	96.84	93.21	83.17	63.81	88.83	77.61	94.04	76.30	87.14	62.99	79.74	61.02
	4	71.25	3.20	93.55	79.71	89.00	77.93	87.14	61.19	88.89	77.32	91.69	83.56	89.57	69.60	85.32	71.33	88.75	11.71	90.28	63.36	85.91	73.08
	5	89.58	57.07	94.05	61.40	95.84	91.36	90.23	67.97	94.64	90.49	90.29	82.22	91.86	72.79	88.75	80.17	91.08	13.63	90.62	47.52	89.41	79.65
	6	90.10	44.91	93.89	58.45	89.18	75.36	91.45	64.78	91.79	84.70	92.31	85.64	96.23	86.13	94.07	88.46	94.67	0.00	91.26	45.52	90.29	78.38
RF	1	60.69	45562	79.89	61.13	90.30	76.76	73.24	49.61	92.50	80.87	97.13	91.24	74.17	48.97	88.95	62.83	83.98	57.20	81.11	57.38	85.16	55.01
	2	75.56	42.36	82.50	64.37	91.97	84.34	84.84	66.31	84.70	65.98	97.50	93.26	87.07	70.29	85.97	63.17	89.34	51.41	83.33	53.14	71.49	46.91
	3	77.50	44.80	84.78	56.80	82.61	68.31	82.00	54.20	70.00	44.74	87.00	71.17	91.93	79.69	84.42	69.41	97.57	89.44	86.08	52.21	80.79	64.28
	4	79.58	16.65	93.55	79.54	90.00	80.02	92.51	73.16	88.89	81.37	88.63	77.19	87.83	64.50	85.26	73.24	91.08	40.81	85.73	45.96	79.28	61.46
	5	84.86	47.58	89.10	37.30	87.79	74.55	88.89	59.44	77.50	53.97	84.25	67.97	89.91	66.08	88.82	79.00	92.42	33.63	90.55	43.45	92.94	86.34
	6	90.00	42.81	93.89	54.34	91.29	80.48	91.54	58.66	71.79	48.20	89.45	79.93	87.62	62.14	94.07	88.46	96.00	0.00	92.03	44.29	95.07	88.40

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: VOTING (FILT)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	53.68	7.79	79.50	58.48	74.59	43.25	72.46	44.74	82.21	56.04	90.28	64.83	71.25	41.94	82.39	38.80	70.48	45311	76.76	49.76	76.34	43.91
	2	66.32	21.84	64.44	21.95	78.10	57.35	68.11	23.38	82.82	62.51	92.31	79.73	81.80	59.26	71.60	28.97	82.75	25.53	80.28	45.89	72.03	45.90
	3	74.44	23.99	76.27	27.21	72.00	43.74	74.71	30.23	70.00	39.85	79.35	51.36	78.92	47.84	82.61	66.86	91.62	53.96	76.90	25.57	79.41	59.23
	4	74.04	-2.98	88.26	64.35	78.97	57.69	78.89	26.53	75.11	53.16	73.92	42.78	86.80	64.05	70.67	41.02	91.40	43.06	84.15	38.96	73.43	48.79
	5	83.17	30.63	86.52	17.58	84.07	67.68	76.83	5.21	88.33	77.10	71.54	48.01	84.94	50.42	75.56	53.64	93.75	53.66	86.58	40.18	74.74	45.93
	6	82.29	-2.67	90.89	43.54	74.63	40.13	81.08	13.81	85.00	71.81	74.95	52.92	95.32	84.67	87.52	75.85	97.42	60.00	89.94	46.72	83.66	64.72
ET	1	60.59	21.50	74.33	52.41	82.53	61.29	75.03	50.63	88.16	70.30	92.97	76.54	76.25	54.24	84.95	51.43	75.24	40.94	80.21	56.12	76.87	38.58
	2	75.96	44.56	80.34	54.77	78.79	58.96	72.26	38.18	81.15	58.84	93.08	82.21	86.93	71.55	82.76	63.05	86.21	42.63	81.53	50.29	65.11	33.05
	3	80.33	50.73	85.62	57.57	76.00	54.41	84.43	59.72	83.45	66.68	87.71	70.94	86.80	68.15	85.10	70.34	94.04	76.30	85.06	50.40	75.89	51.93
	4	74.12	-2.77	91.90	73.83	76.32	51.48	87.39	59.34	83.56	69.74	83.68	67.41	85.17	60.45	81.29	63.65	90.15	31.71	87.04	52.94	78.48	56.94
	5	88.33	53.71	91.38	56.79	87.97	75.57	87.67	58.05	86.11	74.95	82.72	66.59	90.24	68.39	83.66	69.71	91.17	27.29	86.58	41.07	80.00	60.60
	6	82.19	-2.26	93.89	58.45	85.74	65.28	89.40	56.69	89.64	80.50	86.10	73.56	96.23	86.13	87.62	75.92	96.08	40.00	89.17	38.45	86.08	69.27
RF	1	64.15	31.18	74.31	49.50	83.09	62.86	77.56	54.85	86.99	67.94	90.83	67.97	83.75	67.88	81.84	40.49	77.14	45.17	77.37	49.33	74.72	35.54
	2	75.85	41.95	75.04	41.56	80.17	62.86	77.70	50.76	77.95	51.50	92.31	79.73	81.75	60.81	81.97	59.89	86.27	45.21	79.62	45.58	72.06	44.01
	3	81.44	51.11	83.84	53.39	72.00	45.46	81.85	52.51	76.18	51.97	88.70	73.13	89.08	73.85	84.31	69.01	95.22	78.11	80.34	40.85	69.64	40.44
	4	76.62	45641	90.08	68.89	86.92	73.62	87.34	60.13	81.56	65.49	81.46	62.64	82.70	50.87	80.38	60.05	88.82	33.98	85.63	39.45	69.62	40.68
	5	85.75	47.35	85.57	15.75	82.29	62.85	87.70	57.12	79.72	58.65	82.72	66.69	89.37	67.28	84.97	71.65	92.42	47.29	85.01	33.35	83.16	67.68
	6	82.19	-2.26	93.89	54.34	82.58	61.42	90.14	53.39	77.14	58.47	82.00	66.81	91.47	65.15	90.29	81.85	93.42	0.00	89.14	37.06	89.48	76.61

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: VOTING BAGGING (FILT)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	64.98	30.94	77.52	55.86	79.39	54.06	76.58	53.65	88.16	70.30	91.35	69.97	78.75	57.51	85.48	53.42	78.10	48.42	79.08	53.43	70.43	24.20
	2	66.26	11.39	73.42	39.18	76.03	53.54	78.93	53.54	78.08	53.64	93.85	83.51	74.50	43.16	80.54	60.22	83.92	35.79	75.14	33.59	69.72	38.96
	3	81.44	50.67	81.34	40.87	71.20	43.01	83.78	57.95	79.82	59.42	87.71	72.21	85.97	66.56	86.05	71.74	94.04	73.93	82.34	46.51	67.08	34.29
	4	77.79	17.89	87.31	59.74	86.92	73.57	83.82	48.02	81.56	64.26	80.47	60.22	86.87	63.66	81.29	64.33	88.90	22.23	88.47	57.04	70.76	42.86
	5	83.08	25.60	86.57	45432	81.34	60.49	88.41	58.89	78.89	59.61	84.12	68.72	86.72	56.60	83.73	70.20	93.75	53.66	86.58	41.07	74.74	45.87
	6	86.29	24.32	95.89	71.74	77.58	47.63	90.14	59.79	84.64	71.46	86.10	73.56	91.43	63.35	88.86	79.06	96.08	40.00	94.58	73.31	87.12	72.17
ET	1	73.68	50.67	74.38	51.63	80.97	57.39	78.08	56.30	86.91	66.53	93.51	78.97	80.62	62.08	85.98	53.28	80.00	51.69	82.48	61.49	80.11	50.27
	2	76.96	43.57	80.31	54.04	80.17	61.38	80.12	53.09	85.90	69.72	93.85	83.96	86.11	69.88	85.78	68.33	88.50	58.39	82.18	50.51	69.72	42.19
	3	84.90	61.99	83.88	52.92	73.60	48.90	84.43	60.75	81.64	62.63	89.70	76.31	89.11	74.49	86.92	74.55	96.40	83.13	81.61	34.38	79.37	59.43
	4	77.79	17.65	92.77	76.55	84.31	67.92	86.65	60.31	81.33	64.09	80.41	59.01	86.83	64.71	84.29	69.26	86.40	-1.95	85.63	40.19	74.52	49.24
	5	89.67	58.71	90.43	42.72	85.19	68.51	89.15	64.46	77.22	54.15	86.47	72.72	93.79	81.29	84.90	71.73	91.17	27.29	86.58	41.07	83.16	66.38
	6	82.19	-2.26	93.89	59.27	84.74	64.29	91.65	67.06	92.50	85.89	80.48	62.43	90.52	62.83	87.62	76.81	97.42	60.00	89.14	36.00	90.72	79.56
RF	1	73.79	47.73	75.62	51.15	82.01	60.02	76.53	53.13	85.74	63.72	92.45	75.19	76.88	53.70	86.50	54.36	82.86	60.31	80.22	56.94	75.29	35.94
	2	75.96	40.18	78.77	50.74	78.08	58.78	75.85	45.75	84.23	66.05	92.31	79.75	81.75	60.09	87.32	71.00	89.67	63.37	84.07	58.10	69.75	40.65
	3	79.15	47.80	83.88	50.51	76.00	51.93	85.08	62.78	76.00	51.34	84.94	65.42	86.00	66.01	87.79	74.36	94.04	73.88	81.68	41.92	75.81	52.72
	4	80.29	35.54	90.08	69.08	86.92	73.25	84.56	52.10	77.11	58.09	84.74	67.91	85.20	58.91	83.29	65.78	88.90	25.37	86.32	40.95	74.48	49.23
	5	88.25	54.75	86.57	45432	86.06	71.97	88.44	61.17	81.67	64.45	87.72	75.31	92.89	77.85	82.61	66.21	92.42	47.29	87.32	45.25	85.26	70.51
	6	82.19	-2.26	94.89	65.50	82.58	59.53	89.40	55.37	82.50	68.21	81.90	64.76	90.52	62.83	88.95	78.16	97.42	60.00	91.51	51.37	87.19	71.55

TIPO DE ESTRATÉGIA USADA NA CONSTRUÇÃO DOS MODELOS: STACKING BAGGING (FILT)

		ADM		BIOM		BIO		CON		ECO		PES		FIS		MAT		PED		PSI		TUR	
	PX	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC	ACC	MCC
DT	1	78.89	58.01	77.59	55.03	90.55	78.42	75.66	51.22	92.50	81.30	97.70	91.52	80.00	60.64	88.51	66.29	82.11	50.92	90.46	79.43	78.59	49.02
	2	73.90	42.30	73.08	44.81	82.12	65.90	82.58	59.01	85.00	69.03	98.37	95.79	84.40	63.67	86.21	67.36	86.76	60.32	83.56	51.74	75.14	53.68
	3	78.24	38.84	79.57	50.56	80.00	60.95	83.33	58.42	60.55	23.28	93.74	85.33	87.90	69.87	84.74	72.39	96.25	80.99	86.96	51.35	73.48	50.28
	4	78.25	0.00	89.09	62.06	94.29	89.14	88.25	63.36	80.67	62.49	86.84	74.54	93.91	81.56	83.42	67.75	89.92	13.63	92.51	68.28	73.98	49.08
	5	81.71	54.20	85.29	-3.46	86.00	70.93	90.94	68.44	86.43	73.73	83.50	67.69	91.65	68.18	85.44	72.65	98.67	60.00	87.48	42.91	88.24	77.36
	6	89.89	57.75	96.89	75.87	75.16	42.05	93.85	75.20	91.43	85.56	90.77	83.31	93.24	69.99	92.64	85.50	96.00	0.00	96.83	81.77	80.96	61.48
ET	1	69.09	42.95	82.86	68.08	95.68	89.95	80.53	62.41	91.25	76.35	98.86	96.14	79.66	60.13	94.90	80.99	86.14	67.20	85.45	68.93	96.11	87.01
	2	81.37	51.88	86.10	69.54	95.87	91.62	72.63	40.71	91.67	82.04	99.17	97.77	90.06	77.54	92.28	83.63	81.18	38.19	79.68	44.82	78.66	62.59
	3	86.84	64.74	88.50	68.01	85.57	72.67	85.33	62.45	85.78	69.92	92.84	84.05	88.95	73.12	89.09	79.03	94.04	76.30	83.79	37.32	88.10	79.09
	4	80.00	23.84	94.46	82.44	93.29	87.25	89.81	68.43	75.56	54.02	91.54	82.22	90.43	71.50	81.68	61.91	91.25	39.04	88.15	48.78	87.71	76.94
	5	90.67	51.38	92.19	45.53	90.14	80.59	92.48	75.10	80.00	61.95	90.83	81.67	95.45	85.52	88.75	79.28	89.92	13.63	91.45	55.02	89.61	79.80
	6	83.43	5.88	94.84	63.38	92.22	82.62	93.08	72.79	94.64	90.10	92.31	85.88	88.57	65.42	92.64	85.92	98.67	60.00	90.46	41.36	91.76	81.00
RF	1	74.52	50.84	80.00	61.29	96.27	91.14	81.17	63.42	91.17	76.29	97.19	90.10	80.00	60.79	94.92	80.19	87.58	67.95	89.55	76.26	87.34	60.57
	2	77.52	51.39	83.63	63.42	87.80	75.73	79.74	56.49	91.36	80.59	96.80	91.86	82.18	60.84	89.66	77.51	91.69	66.75	84.55	57.17	72.69	43.92
	3	77.57	55.69	85.14	55.97	84.78	71.84	77.57	48.06	64.89	36.91	94.84	88.97	85.60	64.72	92.60	85.89	96.32	83.10	79.44	44.35	81.47	66.84
	4	78.92	44.48	83.12	60.78	88.74	79.16	87.59	66.17	93.78	89.09	89.41	78.72	91.23	74.45	89.47	78.79	81.08	1.42	84.90	46.38	83.33	68.27
	5	72.00	11.88	89.14	15.87	92.00	83.74	92.42	71.54	87.50	76.37	87.00	73.72	95.45	85.52	88.58	78.92	92.42	47.29	89.78	52.65	90.92	82.09
	6	88.57	27.17	95.84	69.61	84.68	65.23	92.25	65.34	89.29	79.32	93.85	88.62	91.43	66.80	78.24	63.43	98.67	60.00	91.26	50.72	93.75	85.75