Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Bacharelado em Estatística

Algoritmos de Aprendizado de Máquina como Preditor do Desempenho Acadêmico: Uma Análise Comparativa de Modelos

Carlos Alberto Alves de Meneses

João Pessoa - PB
2025

Carlos Alberto Alves de Meneses

Algoritmos de Aprendizado de Máquina como Preditor do Desempenho Acadêmico: Uma Análise Comparativa de Modelos

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Marcelo Rodrigo Portela Ferreira

João Pessoa - PB
2025

Catalogação na publicação
Seção de Catalogação e Classificação

M543a Meneses, Carlos Alberto Alves de.
Algoritmos de aprendizado de máquina como preditor do desempenho acadêmico : uma análise comparativa de modelos / Carlos Alberto Alves de Meneses. - João Pessoa, 2025.
47 p. : il.

Orientação: Marcelo Rodrigo Portela Ferreira.
TCC (Curso de Bacharelado em Estatística) - UFPB/ccen.

1. Aprendizado de máquina. 2. Regressão supervisionada. 3. Random Forest. 4. Predição de desempenho acadêmico. 5. Inteligência artificial. 6. Análise comparativa de modelos. I. Ferreira, Marcelo Rodrigo Portela. II. Título.

UFPB/CCEN CDU 311(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

Carlos Alberto Alves de Meneses

Algoritmos de Aprendizado de Máquina como Preditor do Desempenho Acadêmico: Uma Análise Comparativa de Modelos

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Estatística.

Trabalho aprovado. João Pessoa - PB, 09 de outubro de 2025:

Prof. Dr. Marcelo Rodrigo Portela Ferreira
Orientador

Profa. Dra. Tarciana Liberal Pereira de Araújo
Examinador

Prof. Dr. Luiz Medeiros de Araújo Lima Filho
Examinador

João Pessoa - PB
2025

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
COORDENAÇÃO DO CURSO DE ESTATÍSTICA

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

"Algoritmos de Aprendizagem de Máquina Como Preditores do Desempenho Acadêmico: Uma Análise Comparativa de Modelos"

Carlos Alberto Alves de Meneses

Aos nove dias do mês de Outubro de 2025 às 08h00, de modo presencial, no Laboratório Rui Barbosa do Departamento de Estatística, realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) do discente Carlos Alberto Alves de Meneses, matrícula 20180003202, com a Banca Examinadora composta pelos professores: Dr. Marcelo Rodrigo Portela Ferreira, Presidente/Orientador (Departamento de Estatística - UFPB), Dra. Tarciana Liberal Pereira de Araújo (Departamento de Estatística - UFPB), Dr. Luiz Medeiros de Araújo Lima Filho, Examinador (Departamento de Estatística - UFPB) e Dra. Maria Lídia Côco Terra, Examinadora Suplente (Departamento de Estatística - UFPB). Iniciando-se os trabalhos, o presidente da Banca Examinadora cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que se fizesse, oralmente, a exposição do TCC intitulado *"Algoritmos de Aprendizagem de Máquina Como Preditores do Desempenho Acadêmico: Uma Análise Comparativa de Modelos*". Concluída a apresentação, a Banca Examinadora iniciou a arguição do candidato. Encerrados os trabalhos de arguição, os examinadores reuniram-se para avaliação e deram o parecer final sobre a apresentação e defesa oral do candidato, tendo sido atribuída à sua apresentação a nota oito e meio (8,5) na disciplina de TCC II, resultante da média aritmética das notas atribuídas pelos membros da Banca Examinadora. A aprovação do discente está condicionada à entrega da versão final do TCC com a inserção da ficha catalográfica e as alterações sugeridas pelos examinadores à Coordenação do Curso de Bacharelado em Estatística no prazo de 4 dias corridos após a defesa.

gov.br Documento assinado digitalmente
MARCELO RODRIGO PORTELA FERREIRA
Data: 12/10/2025 11:12:53-0300
Verifique em https://validar.iti.gov.br

Dr. Marcelo Rodrigo Portela Ferreira
(Professor Orientador)

gov.br Documento assinado digitalmente
TARCIANA LIBERAL PEREIRA DE ARAUJO
Data: 13/10/2025 09:45:53-0300
Verifique em https://validar.iti.gov.br

Dra. Tarciana Liberal Pereira de Araújo
(Professora Examinadora)

gov.br Documento assinado digitalmente
LUIZ MEDEIROS DE ARAUJO LIMA FILHO
Data: 12/10/2025 11:23:07-0300
Verifique em https://validar.iti.gov.br

Dr. Luiz Medeiros de Araújo Lima Filho
(Professor Examinador)

gov.br Documento assinado digitalmente
CARLOS ALBERTO ALVES DE MENESES
Data: 13/10/2025 14:15:16-0300
Verifique em https://validar.iti.gov.br

Carlos Alberto Alves de Meneses
(Discente)

João Pessoa, 09 de Outubro de 2025

Agradecimentos

Gostaria de expressar minha sincera gratidão a todas as pessoas que, de alguma forma, contribuíram para minha jornada acadêmica na busca pela formação em Bacharelado em Estatística.

À minha família, pela paciência e compreensão diante dos muitos dias de ausência em razão das longas horas dedicadas aos estudos.

Registro, em especial, meus agradecimentos ao meu professor e orientador pelos ensinamentos, pela orientação segura e, sobretudo, pela compreensão e paciência durante o processo de elaboração deste Trabalho de Conclusão de Curso.

Aos professores do Departamento de Estatística, com quem aprendi muito durante a realização do curso.

Resumo

As pesquisas na área da Inteligência Artificial (IA) têm avançado de forma acelerada, refletindo o crescente interesse e investimento nesse campo. Diversos setores, como educação, medicina e economia, vêm se beneficiando dessa tecnologia. Este trabalho discute a utilização de modelos de aprendizado de máquina, com o objetivo de realizar uma análise comparativa entre diferentes modelos disponíveis para essa finalidade.

A aprendizagem de máquina, uma subárea da inteligência artificial, tornou-se uma ferramenta importante com o avanço tecnológico. Essa abordagem computacional permite que sistemas aprendam a partir dos dados, identifiquem padrões complexos e realizem previsões. No contexto educacional, compreender os fatores que influenciam o desempenho acadêmico é fundamental para o desenvolvimento de estratégias eficazes de ensino e apoio aos estudantes.

Este estudo aplica diferentes algoritmos de **aprendizado de máquina supervisionado**, especificamente modelos de **regressão supervisionada**, tais como Regressão Linear, Regressão Elastic Net, Random Forest, Máquinas de Vetores de Suporte (SVM) e Redes Neurais, para prever o desempenho de estudantes do ensino médio. A escolha pela regressão supervisionada fundamenta-se na natureza do problema abordado, onde se busca predizer valores contínuos (notas finais) a partir de variáveis preditoras conhecidas, utilizando um conjunto de dados rotulados para treinamento dos modelos.

A metodologia de regressão supervisionada é particularmente adequada para este contexto, pois permite estabelecer relações quantitativas entre as características dos estudantes e suas notas finais, oferecendo não apenas predições pontuais, mas também insights sobre a magnitude e direção das associações entre as variáveis preditoras e o desempenho acadêmico.

O trabalho utiliza dados que contemplam variáveis demográficas, sociais e escolares, possibilitando uma análise ampla dos fatores preditivos do rendimento acadêmico. A abordagem de regressão supervisionada permite que os algoritmos aprendam a relação entre as características dos estudantes e suas notas finais, estabelecendo padrões que podem ser generalizados para novos casos.

Os resultados indicam que o modelo Random Forest apresentou o melhor desempenho entre os algoritmos de regressão supervisionada testados, destacando os fatores mais relevantes para o sucesso acadêmico. Espera-se que esses resultados contribuam para o

desenvolvimento de estratégias educacionais mais adequadas e eficazes, demonstrando a aplicabilidade da regressão supervisionada na predição de desempenho acadêmico.

Palavras-chave: aprendizado de máquina, regressão supervisionada, predição de desempenho acadêmico, Random Forest, educação, inteligência artificial, análise comparativa de modelos

Abstract

Research in the field of Artificial Intelligence (AI) has advanced rapidly, reflecting the growing interest and investment in this field. Various sectors, such as education, medicine, and economics, have been benefiting from this technology. This work discusses the use of machine learning models, with the objective of conducting a comparative analysis between different models available for this purpose.

Machine learning, a subarea of artificial intelligence, has become an important tool with technological advancement. This computational approach allows systems to learn from data, identify complex patterns, and make predictions. In the educational context, understanding the factors that influence academic performance is fundamental for the development of effective teaching strategies and student support.

This study applies different **supervised machine learning** algorithms, specifically **supervised regression** models, such as Linear Regression, Elastic Net Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks, to predict the performance of high school students. The choice of supervised regression is based on the nature of the problem addressed, where we seek to predict continuous values (final grades) from known predictor variables, using a labeled dataset for model training.

The supervised regression methodology is particularly suitable for this context, as it allows establishing quantitative relationships between student characteristics and their final grades, offering not only point predictions but also insights about the magnitude and direction of associations between predictor variables and academic performance.

The work uses data that includes demographic, social, and school variables, enabling a comprehensive analysis of predictive factors of academic performance. The supervised regression approach allows algorithms to learn the relationship between student characteristics and their final grades, establishing patterns that can be generalized to new cases.

The results indicate that the Random Forest model presented the best performance among the tested supervised regression algorithms, highlighting the most relevant factors for academic success. It is expected that these results will contribute to the development of more appropriate and effective educational strategies, demonstrating the applicability of supervised regression in predicting academic performance.

Keywords: machine learning, supervised regression, academic performance prediction, Random Forest, education, artificial intelligence, comparative model analysis

Lista de tabelas

Tabela 1 – Procedimento para Avaliação de Métodos 26
Tabela 2 – Descrição das variáveis dos conjuntos de dados Maths.csv e Portuguese.csv 28
Tabela 3 – Variáveis de notas relacionadas com a disciplina do curso 29
Tabela 4 – Resultados das métricas de avaliação do modelo Random Forest 30
Tabela 5 – Descrição das métricas utilizadas 30
Tabela 6 – Resultados das métricas de avaliação do melhor modelo para Português 39
Tabela 7 – Comparação dos resultados com benchmarks educacionais 41
Tabela 8 – Resultados do modelo Random Forest para Matemática e Português . 43

Lista de ilustrações

Figura 1 – Aprendizado supervisionado 17
Figura 2 – Aprendizado não supervisionado 19
Figura 3 – Categorias das variáveis discretas 29
Figura 4 – Correlação entre as variáveis - Matemática 30
Figura 5 – Análise abrangente do desempenho dos modelos: (a) RMSE e (b) R^2 em função do Workflow Rank 31
Figura 6 – Desempenho dos seis melhores workflows em termos de RMSE 32
Figura 7 – Dados preditos e observados para o modelo Random Forest 34
Figura 8 – Categorias das variáveis discretas - Português 35
Figura 9 – Correlação entre as variáveis - Português 35
Figura 10 – Análise abrangente do desempenho dos modelos para Português: (a) RMSE e (b) R^2 em função do Workflow Rank 36
Figura 11 – Desempenho dos seis melhores workflows para Português em termos de RMSE 37
Figura 12 – Dados preditos e observados para o modelo Random Forest - Português 42

Sumário

1 INTRODUÇÃO 14
1.1 Objetivo Geral 15
1.2 Objetivos Específicos 15

2 REFERENCIAL TEÓRICO 16
2.1 Aprendizado de Máquina 16
2.2 Tipos de Aprendizado de Máquina 16
2.2.1 Aprendizado Supervisionado 17
2.2.2 Aprendizado Não Supervisionado 19
2.2.3 Aprendizado por Reforço 20
2.3 Fluxo de Tarefas de um Projeto de Aprendizado de Máquina 20
2.4 Modelos de Aprendizado de Máquina 21
2.4.1 Regressão Linear 21
2.4.2 Random Forest (Floresta Aleatória) 22
2.4.3 Máquinas de Vetores de Suporte (SVM) 22
2.4.4 Redes Neurais Artificiais 23

3 METODOLOGIA 24
3.1 Origem dos Dados e Reprodutibilidade 24
3.2 Pré-processamento dos Dados 24
3.3 Modelos Implementados 25
3.4 Métricas de Avaliação 25
3.5 Divisão dos Dados e Validação 26

4 RESULTADOS 27
4.1 Conjunto de Dados 27
4.2 Resultados para Matemática 30
4.3 Análise Comparativa do Desempenho dos Modelos 31
4.3.1 Análise Abrangente de Workflows 31
4.3.2 Análise Focada nos Melhores Modelos 32
4.3.3 Implicações Metodológicas 32
4.3.4 Interpretação das Métricas 33
4.3.5 Conclusões da Análise Comparativa 33
4.4 Resultados para Português 34
4.4.1 Correlação 35

4.5 Análise Comparativa do Desempenho dos Modelos - Disciplina de Português **36**
4.5.1 Análise Abrangente de Workflows - Português 36
4.5.1.1 Comportamento Específico dos Modelos - Português 37
4.5.2 Análise Focada nos Melhores Modelos - Português 37
4.5.3 Comparação Interdisciplinar: Matemática vs. Português 37
4.5.3.1 Similaridades 38
4.5.3.2 Diferenças Observadas 38
4.5.4 Implicações Metodológicas para Português 38
4.5.5 Interpretação das Métricas - Português 38
4.5.6 Conclusões da Análise - Português 38
4.5.7 Validação Cruzada Interdisciplinar 39
4.5.8 Resultados do Melhor Modelo - Português 39
4.5.9 Análise Acadêmica dos Resultados - Português 40
4.5.9.1 Interpretação das Métricas 40
4.5.9.2 Contextualização dos Resultados 40
4.5.9.3 Comparação com Benchmarks Educacionais 40
4.5.9.4 Implicações Práticas 41
4.5.9.5 Limitações e Considerações 41
4.5.10 Conclusões dos Resultados - Português 41
4.6 Análise Comparativa entre Disciplinas **42**
4.6.1 Interpretação dos Resultados 43
4.6.2 Variáveis Mais Importantes 43

5 DISCUSSÃO **44**
5.1 Implicações Práticas **44**
5.2 Limitações do Estudo **44**

6 CONSIDERAÇÕES FINAIS **45**
6.1 Principais Contribuições **46**
6.2 Trabalhos Futuros **46**

REFERÊNCIAS **47**

1 Introdução

Nas últimas décadas, o cenário mundial tem sido marcado por uma nova corrida tecnológica, não mais centrada apenas em questões armamentistas, mas sim no avanço da inteligência artificial (IA). Grandes potências têm investido fortemente no desenvolvimento de tecnologias baseadas em IA, visando aplicações que vão desde a defesa nacional até áreas fundamentais como saúde, educação, economia e ciência de dados.

No campo da Estatística, a presença da inteligência artificial não é novidade. Técnicas de aprendizado de máquina (*machine learning*), que podem ser consideradas uma vertente prática da IA, vêm sendo utilizadas há décadas para a análise de dados, a previsão de eventos e o apoio à tomada de decisões. Como observam Morettin e Singer (2022), "os avanços no aprendizado e na análise de dados com Estatística estão diretamente relacionados com avanços na área computacional".

A integração entre Estatística e Inteligência Artificial (IA) consolida-se como um pilar fundamental para compreender padrões complexos, construir modelos preditivos e transformar informações em conhecimento aplicável. Segundo Bruce e Bruce (2019), esses métodos diferem dos métodos estatísticos clássicos por serem orientados pelos dados, sem impor uma estrutura linear ou geral aos mesmos.

No contexto educacional, a explosão na quantidade e variedade de dados disponíveis oferece oportunidades únicas para melhorar a compreensão dos fatores que influenciam o desempenho acadêmico. Através de dados socioeconômicos, hábitos de estudo, características familiares e outros indicadores, é possível desenvolver modelos preditivos que auxiliem educadores e gestores na tomada de decisões mais assertivas. A utilização inteligente dessas informações é essencial para o sucesso e a sustentabilidade de políticas educacionais eficazes. Ao adotar uma postura baseada em dados, as instituições de ensino conseguem identificar estudantes em risco, personalizar estratégias de ensino e garantir uma tomada de decisão mais fundamentada.

Entre os diversos fatores que podem comprometer o desempenho escolar, o consumo de álcool por adolescentes tem sido identificado como uma variável relevante em estudos educacionais. O consumo excessivo de bebidas alcoólicas provoca reações indesejáveis em nosso organismo. Alguns copos a mais e já surgem dificuldades para falar com desenvoltura, manter a coordenação e o equilíbrio; além disso, a capacidade de julgamento deixa de ser totalmente confiável. Galvão (2017) destaca que o excesso de álcool no cérebro leva a efeitos psíquicos como redução da concentração, da atenção, da memória recente e da capacidade de julgamento. Manzatto et al. (2011) apontam que, dentre os prejuízos causados pelo consumo excessivo de álcool, está a queda de desempenho acadêmico.

Conforme a ingestão da bebida aumenta, o organismo reage de formas diferentes,

passando por alguns estágios. Quando a concentração de álcool no sangue é baixa (entre 0,01 e 0,12 gramas/100 mililitros), o indivíduo tende a ficar desinibido, relaxado e eufórico. À medida que essa quantidade aumenta, outras reações aparecem, como lentidão dos reflexos, problemas de atenção, perda de memória, alterações na capacidade de raciocínio e falta de equilíbrio. Em níveis muito altos (a partir de 0,40 gramas/100 mililitros), pode ocorrer intoxicação severa e até mesmo parada cardiorrespiratória, com possibilidade de sequelas neurológicas ou morte. Além desses efeitos visíveis e imediatos, o consumo exagerado de álcool, principalmente na infância e adolescência, pode prejudicar o desenvolvimento cerebral, inibir o crescimento de novos neurônios e causar lesões permanentes, além de ser um fator de risco para a depressão e outros transtornos mentais.

1.1 Objetivo Geral

Investigar e identificar o modelo de aprendizado de máquina mais adequado para a análise preditiva de dados estatísticos referentes ao desempenho escolar de alunos do ensino secundário, utilizando bases de dados reais e comparando diferentes abordagens estatísticas e computacionais.

1.2 Objetivos Específicos

1. Explorar e compreender a base de dados proposta, analisando variáveis demográficas, sociais e escolares que influenciam o desempenho dos estudantes.

2. Aplicar técnicas de pré-processamento de dados, incluindo limpeza, transformação e seleção de variáveis relevantes para o problema preditivo.

3. Implementar e avaliar diferentes modelos de aprendizado de máquina supervisionado, comparando métricas de desempenho estatístico.

4. Selecionar o modelo com melhor desempenho e justificar sua adequação estatística e computacional para o problema em estudo.

5. Apontar implicações práticas do uso de aprendizado de máquina para prever desempenho escolar e possíveis contribuições para a área de Educação e Estatística.

2 Referencial Teórico

2.1 Aprendizado de Máquina

O aprendizado de máquina é uma área que se concentra no desenvolvimento de algoritmos e modelos que permitem às máquinas aprenderem a partir de dados, sem serem explicitamente programadas. Ou seja, ao invés de seguirem um conjunto fixo de instruções para realizar uma tarefa, os modelos identificam padrões e geram respostas a partir de um conjunto de dados.

Os modelos de aprendizado de máquina fazem parte da área denominada Inteligência Artificial e oferecem um amplo conjunto de ferramentas que permitem que máquinas aprendam com a experiência e se adaptem de maneira autônoma. Segundo Jiang (2021), o aprendizado de máquina é a capacidade de uma máquina imitar o comportamento humano inteligente.

O aprendizado de máquina é frequentemente descrito como a arte e ciência do reconhecimento de padrões. É basicamente uma abordagem baseada em dados, que envolve uma ampla variedade de métodos. A maioria dos métodos tradicionais de previsão utilizados em diferentes áreas está baseada no ajuste dos dados a uma relação pré-especificada entre variáveis dependentes e independentes, assumindo assim um processo estatístico específico. Por outro lado, os modelos de aprendizado de máquina quase não fazem suposições sobre as relações estatísticas subjacentes aos dados.

Para Varian (2014), o aprendizado de máquina é uma expressão relativamente vaga, mas pode ser entendida como um conjunto de procedimentos usados principalmente para estimações pontuais e que geralmente permitem ao pesquisador trabalhar com grandes tamanhos de amostra, grande número de variáveis e forma estrutural desconhecida. Como observa Arão (2024), os dados passam a ser elementos que permitem que as máquinas aprendam através de reconhecimento de padrões e possam fazer previsões e oferecer respostas. Segundo Ludermir (2021), é necessária uma grande quantidade de exemplos para gerar o conhecimento do computador, que são hipóteses geradas a partir dos dados. De acordo com Varian (2014), as quantidades crescentes de dados e as relações cada vez mais complexas entre variáveis justificam o uso de abordagens de aprendizado de máquina em diversas áreas do conhecimento.

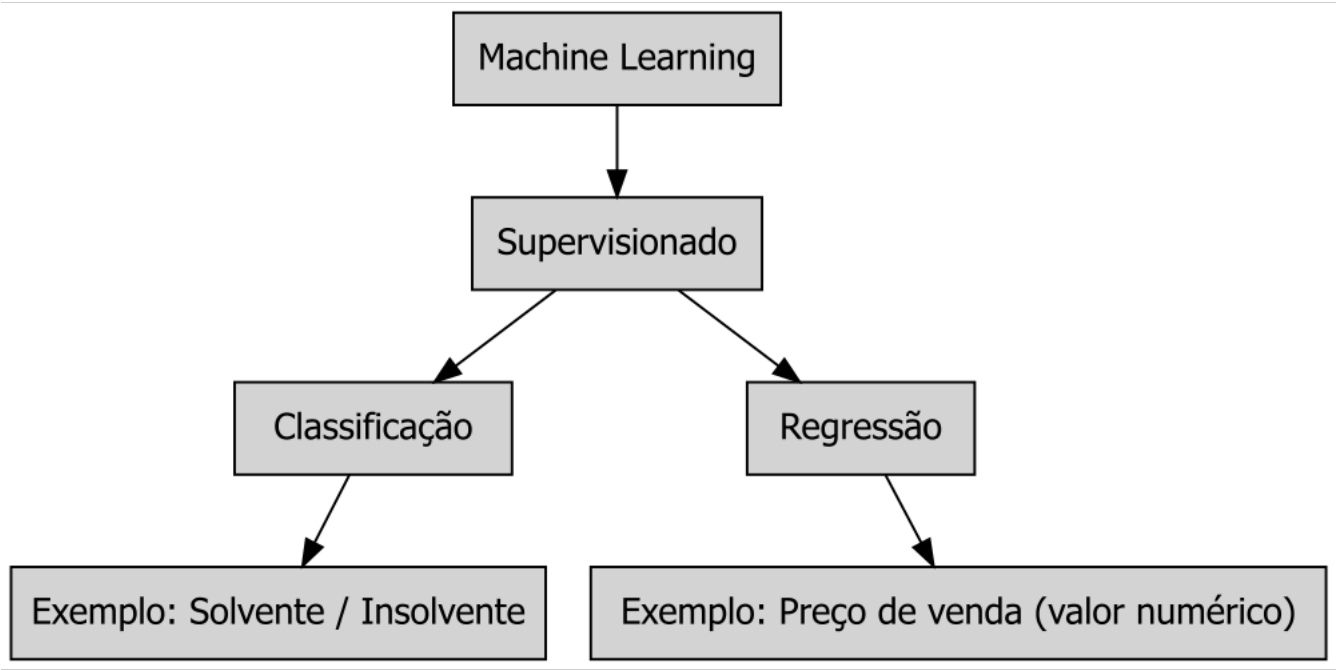
2.2 Tipos de Aprendizado de Máquina

Esses modelos de aprendizado de máquina podem ser classificados de três maneiras distintas: Aprendizado Supervisionado, Aprendizado não Supervisionado e Aprendizado

por Reforço.

2.2.1 Aprendizado Supervisionado

Figura 1 – Aprendizado supervisionado



Fonte: Elaborado pelo autor.

Os modelos de aprendizado supervisionado são os mais utilizados. O aprendizado supervisionado é o termo usado sempre que o programa ou algoritmo é treinado sobre um conjunto de dados pré-definido. Estes algoritmos são aqueles em que ensinamos à máquina quais respostas (Y) ela deve ter a partir de determinadas características (x).

Segundo Morettin e Singer (2022), a ideia fundamental do aprendizado supervisionado é utilizar preditores (dados de entrada ou *inputs*) para prever uma ou mais respostas (dados de saída ou *outputs*), que podem ser quantitativas ou qualitativas (categorias, atributos ou fatores).

Os problemas que envolvem o método de aprendizado supervisionado são categorizados em dois tipos:

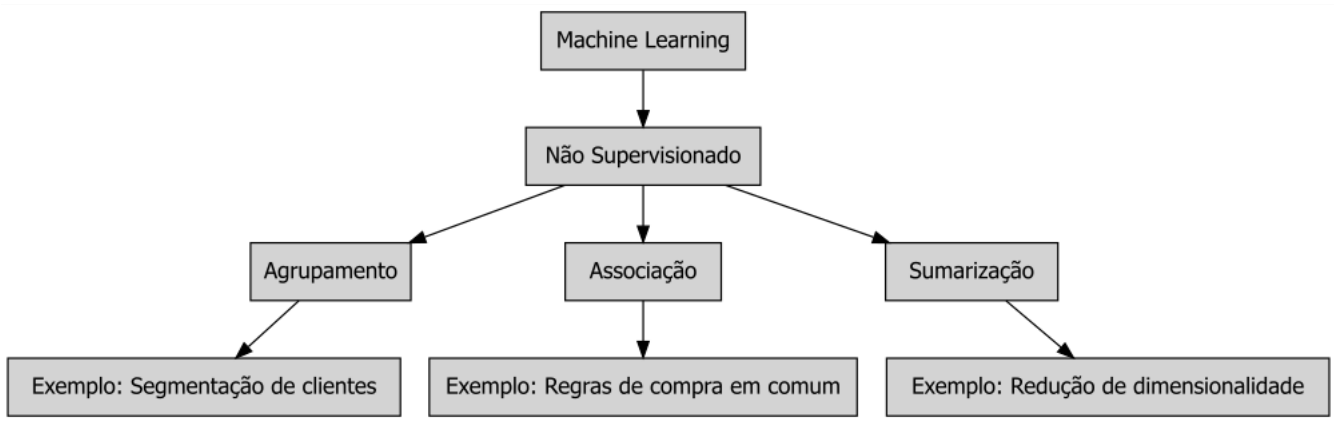
- **Classificação**: desejamos classificar um rótulo Y a partir de um vetor de características x
- **Regressão**: desejamos estimar um valor numérico Y a partir de um vetor de características x

Para Almeida, Sales e Nunes (2023), classificação é o processo de identificar a qual conjunto de categoria uma nova observação pertence, com base em um conjunto de dados de treino contendo observações (ou instâncias).

Para Harrison (2020), a regressão é um processo de aprendizado de máquina supervisionado. É semelhante à classificação, mas, em vez de prever um rótulo, tentamos prever um valor contínuo.

2.2.2 Aprendizado Não Supervisionado

Figura 2 – Aprendizado não supervisionado



Fonte: Elaborado pelo autor.

Os modelos de aprendizado não supervisionado buscam analisar os dados e identificar padrões que, por muitas vezes, são praticamente impossíveis para um ser humano. Nesse modelo de aprendizado, o rótulo da classe de cada amostra do treinamento não é conhecido, e o número ou conjunto de classe a ser treinado pode não ser conhecido *a priori*.

Para Morettin e Singer (2022), nesse tipo de aprendizado considera-se conjuntos de dados em que todas as variáveis têm a mesma natureza, isto é, em que não há distinção entre variáveis preditoras e respostas. O objetivo é o entendimento da estrutura de associação entre as variáveis disponíveis.

2.2.3 Aprendizado por Reforço

O aprendizado por reforço se assemelha ao aprendizado supervisionado, porém, não são fornecidos dados amostrais prévios ao algoritmo. Neste modelo, o aprendizado ocorre à medida que o sistema interage com o ambiente, experimentando ações e armazenando as consequências. Uma sequência de resultados bem-sucedidos é reforçada enquanto um resultado mal-sucedido é penalizado.

2.3 Fluxo de Tarefas de um Projeto de Aprendizado de Máquina

O processo para elaborar um projeto de aprendizado de máquina consiste em 3 principais etapas:

1. **Pré-processamento**: Nessa etapa é realizado o tratamento completo dos dados, incluindo a identificação e tratamento de valores ausentes através de técnicas de imputação, detecção e tratamento de outliers utilizando métodos estatísticos, normalização e padronização das variáveis numéricas para garantir escalas comparáveis, e codificação das variáveis categóricas através de técnicas como One-Hot Encoding. Além disso, é realizada a seleção das variáveis mais relevantes através de análise de correlação e técnicas de feature selection, garantindo que apenas as variáveis com maior poder preditivo sejam utilizadas no modelo. Os dados são então repartidos entre conjunto de treino (70-80%) e teste (20-30%), com validação cruzada sendo aplicada para garantir a robustez dos resultados. Esta etapa é fundamental para garantir a qualidade dos dados e o desempenho adequado dos modelos de aprendizado de máquina.

2. **Validação Cruzada**: Nessa etapa é realizada a seleção dos algoritmos e dos hiperparâmetros a partir do conjunto de treino através de técnicas de validação cruzada k-fold (tipicamente 5 ou 10 folds). O processo envolve a divisão do conjunto de treino em k subconjuntos, onde k-1 folds são utilizados para treinamento e 1 fold para validação, repetindo este processo k vezes para garantir uma avaliação robusta do desempenho. Durante esta etapa, são testados diferentes algoritmos de aprendizado de máquina (Random Forest, SVM, Regressão Linear, etc.) e seus respectivos hiperparâmetros através de técnicas como Grid Search ou Random Search. A validação cruzada permite estimar o desempenho real do modelo e identificar possíveis problemas de overfitting. Caso os modelos avaliados não apresentem bons resultados (baixo R^2, alto RMSE, ou alta variância), retorna-se à etapa anterior de pré-processamento para ajustar as técnicas aplicadas e reinicia-se o processo de validação.

3. **Avaliação Final**: É a etapa final do processo, na qual avaliamos os melhores modelos identificados durante a validação cruzada no conjunto de teste independente. Esta avaliação é realizada utilizando métricas específicas para regressão, como RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R^2 (Coeficiente de Determinação) e MAE (Mean Absolute Error). O conjunto de teste, que não foi utilizado durante o treinamento ou validação cruzada, fornece uma estimativa não enviesada do desempenho real do modelo em dados não vistos. Além da avaliação quantitativa, é realizada uma análise qualitativa dos resultados, incluindo a interpretação da importância das variáveis, análise de resíduos e verificação da estabilidade das predições. O modelo que apresenta o melhor desempenho no conjunto de teste é selecionado como o modelo final, considerando não apenas as métricas de desempenho, mas também critérios como interpretabilidade, complexidade computacional e adequação ao contexto do problema.

2.4 Modelos de Aprendizado de Máquina

2.4.1 Regressão Linear

A regressão linear é um dos métodos estatísticos mais fundamentais e amplamente utilizados para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. No contexto do aprendizado de máquina, representa um dos algoritmos supervisionados mais simples para problemas de regressão. Nesse sentido, as formas de regressão mais conhecidas são a Regressão Linear Simples, a Regressão Linear Múltipla e a Regressão Logística.

A Regressão Linear Simples é utilizada em problemas nos quais se busca identificar um modelo que represente a relação entre duas variáveis, em que uma influencia o valor da outra de maneira linear. O objetivo é encontrar uma função que expresse essa correspondência.

A Regressão Linear Múltipla, por sua vez, trata de problemas em que existe uma relação linear entre múltiplas variáveis, de modo que a variável dependente é influenciada conjuntamente por todas elas.

A Regressão Logística, por sua vez, é empregada na modelagem de problemas cuja variável dependente é categórica e os dados são classificáveis em duas classes distintas.

De forma geral, o modelo de regressão linear pode ser representado pela seguinte equação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{2.1}$$

em que y representa a variável dependente, β_0 é o intercepto, β_i são os coeficientes

associados às variáveis independentes x_i, e ε corresponde ao termo de erro aleatório. O objetivo da regressão linear é estimar os parâmetros β de modo que a função ajustada minimize a soma dos erros quadráticos entre os valores observados e os valores previstos pelo modelo.

2.4.2 Random Forest (Floresta Aleatória)

Random Forest é uma técnica de aprendizado por *ensemble*, fundamentada na ideia de combinar diversos modelos preditivos, treiná-los para uma mesma tarefa e produzir, a partir desses, um modelo mais robusto. É um algoritmo baseado em um conjunto de árvores desenvolvido por Leo Breiman e Adele Cutler em 2001, que combina os princípios básicos de *bagging* com seleção aleatória de recursos para adicionar diversidade ao modelo de *decision tree*.

Em outras palavras, cria-se uma "floresta"de árvores de decisão, em que cada árvore é treinada com uma amostra aleatória independente e de mesma distribuição partindo dos dados de treinamento, com o objetivo de minimizar erros como o sobreajuste. O modelo utiliza uma votação para combinar as previsões das árvores.

O método Random Forest oferece um ganho em relação às árvores *bagged* por meio de um pequeno ajuste aleatório. Ao construir essas árvores de decisão, cada vez que uma divisão em uma árvore é considerada, uma amostra aleatória de m preditores é escolhida como candidata à divisão do conjunto completo de p preditores. A divisão pode usar apenas um desses m preditores, e normalmente se escolhe $m \approx \sqrt{p}$.

Segundo James et al. (2013), a principal diferença entre *bagging* e o Random Forest é a escolha do tamanho do subconjunto do preditor m. O Random Forest usando $m = \sqrt{p}$ proporciona uma redução no erro de teste e no erro *out-of-bag*, sobre o *bagging*. Usar um pequeno valor de m na construção de um Random Forest normalmente será útil quando temos um grande número de preditores correlacionados.

2.4.3 Máquinas de Vetores de Suporte (SVM)

Concebido primordialmente para problemas de classificação, o algoritmo de Máquinas de Vetores de Suporte (SVM) é igualmente aplicado a tarefas de regressão. Sua metodologia baseia-se no mapeamento dos vetores de características de um conjunto de dados para um espaço de dimensão mais elevada, processo otimizado pelo uso de uma função de kernel — técnica conhecida como 'truque de kernel' (kernel trick). Nesse novo espaço dimensional, o objetivo do modelo é identificar um hiperplano que separe linearmente as classes, maximizando a margem entre elas. A capacidade de classificar dados não linearmente separáveis no espaço original, aliada ao uso de uma margem suave (soft margin) que tolera erros de classificação, confere grande robustez ao algoritmo. O SVM apresenta um desempenho altamente competitivo, sendo frequentemente superior a

outros métodos de aprendizado de máquina, e continua sendo amplamente empregado em tarefas complexas, como reconhecimento de padrões em imagens e processamento de sinais de voz, embora modelos de Deep Learning possam obter maior acurácia em domínios com vastos volumes de dados.

2.4.4 Redes Neurais Artificiais

Embora sejam percebidas como uma inovação recente, as Redes Neurais Artificiais (RNAs) têm suas fundações no primeiro modelo matemático-computacional de um neurônio, proposto por Warren McCulloch e Walter Pitts em 1943. Essencialmente, uma RNA é um modelo computacional inspirado na estrutura do cérebro humano, projetado para reconhecer padrões e modelar relações complexas entre variáveis. Dentre as suas diversas arquiteturas, o Perceptron de Múltiplas Camadas (MLP) é uma das mais proeminentes, sendo composto por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Os neurônios de camadas adjacentes são interconectados por sinapses ponderadas (pesos), que são sistematicamente ajustados durante o treinamento para otimizar o desempenho do modelo, minimizando uma função de erro. Cada neurônio artificial calcula a soma ponderada de suas entradas e aplica uma função de ativação, geralmente não linear, para produzir um sinal de saída que é propagado para a camada seguinte.

3 Metodologia

3.1 Origem dos Dados e Reprodutibilidade

Os dados utilizados neste trabalho abordam o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos incluem notas dos alunos, características demográficas, sociais e relacionadas com a escola, coletados através de relatórios escolares e questionários.

O dataset original está disponível no Kaggle sob o título "Alcohol Effects on Study"(<https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study>). Para o processamento e análise dos dados, utilizou-se o software R versão 4.3.2, empregando os seguintes pacotes especializados em aprendizado de máquina: `tidyverse` para manipulação de dados, `caret` para modelagem preditiva, `randomForest` para implementação do algoritmo Random Forest, `e1071` para Support Vector Machines, `nnet` para redes neurais, e `corrplot` para visualização de correlações.

Todos os códigos utilizados neste trabalho, incluindo scripts de pré-processamento, implementação dos modelos, análise dos resultados e geração das visualizações, foram disponibilizados publicamente no repositório GitHub: <https://github.com/carlosmenesestvcb/TCC-Estat-stica>. O repositório inclui: scripts completos em R para todas as análises, datasets originais e processados, documentação detalhada dos procedimentos, instruções para reprodução dos experimentos e resultados em formatos reutilizáveis. Esta disponibilização segue as boas práticas de ciência aberta, permitindo a verificação independente dos resultados e facilitando futuras extensões da pesquisa.

3.2 Pré-processamento dos Dados

Previamente à construção dos modelos de predição, o conjunto de dados foi submetido a uma etapa de pré-processamento. Identificou-se a presença de múltiplas variáveis categóricas (do tipo character), formato que apresenta incompatibilidade com a maioria dos algoritmos de aprendizado de máquina, os quais requerem entradas exclusivamente numéricas. Para contornar essa limitação, as variáveis categóricas foram convertidas em um conjunto de variáveis binárias, conhecidas como variáveis dummy, por meio da técnica de codificação one-hot.

As variáveis *dummies* ou variáveis indicadoras são formas de agregar informações qualitativas em modelos estatísticos. Ela atribui 1 se o elemento possui determinada característica, ou 0 caso não possua. Esse tipo de transformação é importante para modelos de regressão, pois torna possível trabalhar com variáveis qualitativas. Foram

criadas variáveis *dummy* para as variáveis qualitativas utilizando a função `step_dummy` do pacote `tidyverse`.

O conjunto de dados não possuía dados faltantes (NAs), porém, foi aplicado o método k-Nearest Neighbors (kNN) para tratar dados faltantes caso venham a surgir na utilização de novos dados. O método kNN consiste em procurar os k vizinhos mais próximos do elemento que possui o dado faltante de uma variável de interesse, calculando a média dos valores observados dessa variável dos k vizinhos e imputando esse valor ao elemento.

Também foram normalizadas as variáveis quantitativas utilizando a função `step_normalize`, que padroniza as variáveis para terem média zero e desvio padrão um. Essas transformações foram realizadas no conjunto de treinamento e aplicadas consistentemente ao conjunto de teste.

3.3 Modelos Implementados

Os algoritmos implementados foram: Regressão Linear, Regressão Elastic Net, Random Forest, Máquina de Vetores de Suporte (SVM) com kernel linear, SVM com kernel não linear (RBF) e Redes Neurais Artificiais (RNA).

3.4 Métricas de Avaliação

Para avaliar o desempenho dos modelos de regressão, foram utilizadas duas métricas principais: a Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2).

A **Raiz do Erro Quadrático Médio** (RMSE - *Root Mean Squared Error*) fornece uma indicação da magnitude dos erros produzidos pelo sistema nas previsões, atribuindo maior peso aos erros de maior magnitude. É calculada como:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (estimado_i - real_i)^2}{n}} \quad , i = 1, 2, ..., n. \tag{3.1}$$

O **Coeficiente de Determinação** (R^2) indica a proporção da variância da variável dependente que é explicada pelas variáveis independentes no modelo. É dado por:

$$R^2 = 1 - \frac{MSE}{Var} = 1 - \frac{\sum_{i=1}^{n} (real_i - estimado_i)^2}{\sum_{i=1}^{n} (real_i - média)^2} \quad , i = 1, 2, ..., n. \tag{3.2}$$

Uma maneira de interpretar melhor o RMSE é normalizá-lo usando a seguinte fórmula:

$$RMSE_{normalizado} = \frac{RMSE}{valor_{máximo} - valor_{mínimo}} \tag{3.3}$$

Isso produz um valor entre 0 e 1, onde valores mais próximos de 0 representam modelos de melhor ajuste.

3.5 Divisão dos Dados e Validação

O conjunto de dados foi dividido em treino e teste, utilizando 80% dos dados para treinamento e 20% para teste. Para a disciplina de português, foram obtidas 517 observações para o conjunto de treinamento e 132 observações para o conjunto de teste.

Foi realizado o procedimento de validação cruzada no conjunto de treinamento com 5 folds. Esta técnica permite comparar diferentes métodos de aprendizado de máquina e otimizar seus hiperparâmetros, fornecendo uma estimativa robusta do desempenho para seleção do melhor modelo.

O seguinte procedimento foi realizado para cada método: separar os dados em conjunto de treino e conjunto de teste, treinar um modelo no conjunto de treino, avaliar no conjunto de teste e repetir os passos para estimar o erro médio.

Tabela 1 – Procedimento para Avaliação de Métodos

Passo	**Descrição**
1	Separar os dados em conjunto de treino e conjunto de teste.
2	Treinar um modelo no conjunto de treino.
3	Avaliar o modelo no conjunto de teste.
4	Repetir os passos anteriores para estimar o erro médio.

Fonte: Elaborado pelo autor.

4 Resultados

4.1 Conjunto de Dados

São fornecidos dois conjuntos de dados relativos ao desempenho em duas disciplinas distintas: Matemática e Língua Portuguesa. O conjunto de dados é composto por mais de 600 observações e 31 variáveis, sendo 14 numéricas e 17 do tipo categórica. A tabela a seguir descreve as variáveis presentes nos bancos de dados.

Atributos dos Conjuntos de Dados

Variável	**Descrição**
escola	escola do aluno (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
sexo	sexo do aluno (binário: 'F' - feminino ou 'M' - masculino)
idade	idade do aluno (numérica: de 15 a 22)
endereço	tipo de endereço residencial do aluno (binário: 'U' - urbano ou 'R' - rural)
tamanho familiar	tamanho da família (binário: 'LE3' - menor ou igual a 3 ou 'GT3' - maior que 3)
Pstatus	situação de coabitação dos pais (binário: 'T' - vivendo juntos ou 'A' - separados)
Medu	educação da mãe (numérico: 0 - nenhuma, 1 - ensino fundamental (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior)
Fedu	educação do pai (numérica: 0 - nenhuma, 1 - ensino fundamental (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior)
Mjob	trabalho da mãe (nominal: 'professor', 'saúde' relacionado a cuidados, 'serviços' civis (por exemplo, administrativos ou policiais), 'em casa' ou 'outro')
Fjob	trabalho do pai (nominal: 'professor', 'saúde' relacionado a cuidados, 'serviços' civis (por exemplo, administrativos ou policiais), 'em casa' ou 'outro')

Variável	**Descrição**
razão	motivo para escolher esta escola (nominal: perto de 'casa', 'reputação' da escola, preferência de 'curso' ou 'outro')
guardião	tutor do aluno (nominal: 'mãe', 'pai' ou 'outro')
tempo de viagem	tempo de viagem de casa para a escola (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, ou 4 - >1 hora)
tempo de estudo	tempo de estudo semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 - >10 horas)
falhas	número de reprovações anteriores (numérico: n se 1<=n<3, caso contrário 4)
escolas	apoio educacional extra (binário: sim ou não)
famsup	apoio educacional familiar (binário: sim ou não)
pago	aulas extras pagas dentro da disciplina do curso (Matemática ou Português) (binário: sim ou não)
atividades	atividades extracurriculares (binário: sim ou não)
berçário	frequentou a creche (binário: sim ou não)
mais alto	quer fazer ensino superior (binário: sim ou não)
Internet	Acesso à Internet em casa (binário: sim ou não)
romântico	com um relacionamento romântico (binário: sim ou não)
família	qualidade das relações familiares (numérico: de 1 - muito ruim a 5 - excelente)
Tempo livre	tempo livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto)
sair	sair com amigos (numérico: de 1 - muito baixo a 5 - muito alto)
Dalc	consumo de álcool durante o dia de trabalho (numérico: de 1 - muito baixo a 5 - muito alto)
Walc	consumo de álcool no fim de semana (numérico: de 1 - muito baixo a 5 - muito alto)
saúde	estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom)
ausências	número de faltas escolares (numérico: de 0 a 93)

Tabela 2 – Descrição das variáveis dos conjuntos de dados Maths.csv e Portuguese.csv

Nota	Descrição
G1	nota do primeiro período (numérica: de 0 a 20)
G2	nota do segundo período (numérica: de 0 a 20)
G3	nota final (numérica: de 0 a 20, meta de produção)

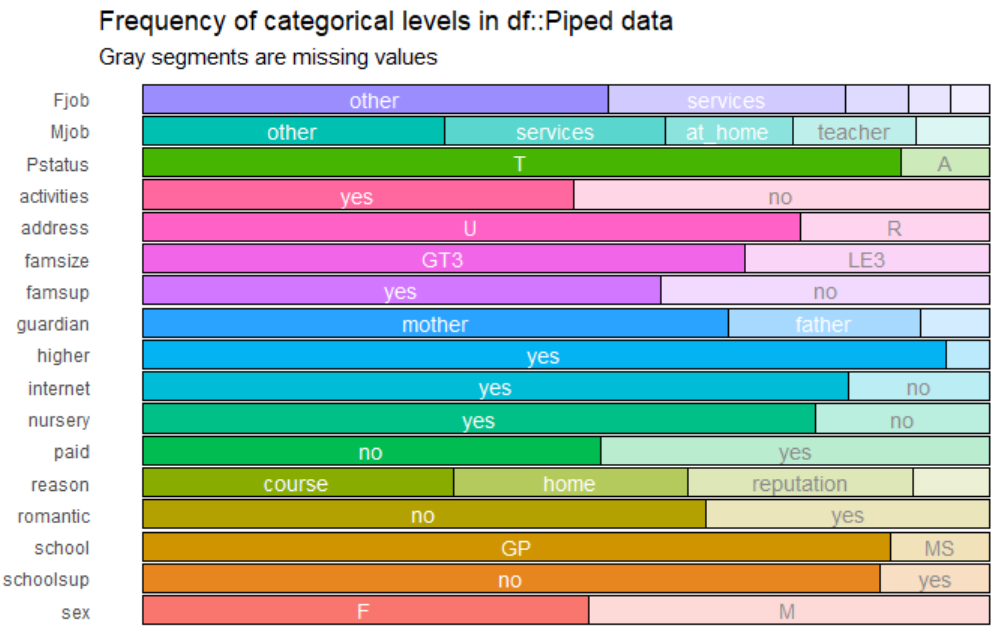
Tabela 3 – Variáveis de notas relacionadas com a disciplina do curso

Variáveis de Notas

Observação:

Correlação entre as notas: O atributo alvo G3 tem uma forte correlação com os atributos G2 e G1. Isto ocorre porque G3 é a nota final do ano (emitida no 3º período), enquanto G1 e G2 correspondem às notas do 1º e 2º períodos. A predição de G3 sem utilizar G2 e G1 é mais desafiadora, porém apresenta maior utilidade prática, pois permite identificar estudantes em risco antes das avaliações intermediárias.

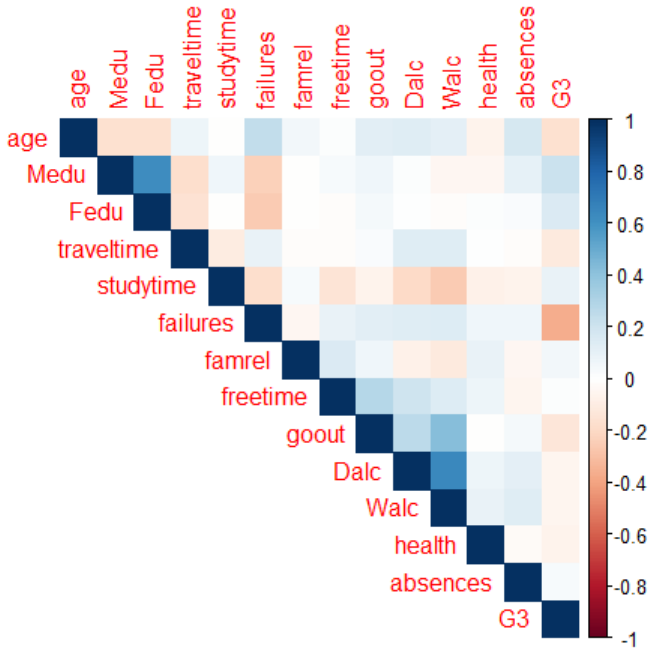
Figura 3 – Categorias das variáveis discretas



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

A variável $G3$ tem uma forte correlação com as variáveis $G2$ e $G1$. Isto ocorre porque $G3$ é a nota final do ano (emitida no 3º período), enquanto $G1$ e $G2$ correspondem às notas do 1º e 2º períodos. Por este motivo, as variáveis $G1$ e $G2$ foram excluídas das análises.

Figura 4 – Correlação entre as variáveis - Matemática



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Os resultados indicam que as variáveis Medu (educação da mãe) e Fedu (educação do pai) apresentam correlação moderada com a variável alvo G3. Em contrapartida, a variável Failures (reprovações anteriores) demonstra correlação baixa com o desempenho final dos estudantes.

4.2 Resultados para Matemática

A análise dos dados de matemática revelou que o modelo Random Forest apresentou o melhor desempenho. O teste de seleção do melhor modelo (*best results*) confirmou a análise gráfica, Figura 7, de que o melhor modelo foi o *Preprocessor1 Model1*, correspondente ao Random Forest.

Métrica	**Estimador**	**Valor**	**Configuração**
RMSE	standard	4.2250	Preprocessor1_Model1
R^2	standard	0.2171	Preprocessor1_Model1

Tabela 4 – Resultados das métricas de avaliação do modelo Random Forest

Métrica	**Descrição**
RMSE	Root Mean Square Error - Erro quadrático médio
R^2	Coeficiente de determinação

Tabela 5 – Descrição das métricas utilizadas

O melhor modelo Random Forest apresentou um $RMSE = 4,1934270$ e um $R^2 = 0,2332309$. O valor para o $RMSE$ normalizado foi 0,2096713. Como esse valor está mais próximo de zero do que de um, o modelo demonstra boa capacidade preditiva para a variável $G3$ em função das covariáveis analisadas.

As seguintes variáveis foram significativas para o modelo: school, age, activities, romantic, famrel, Walc e absences.

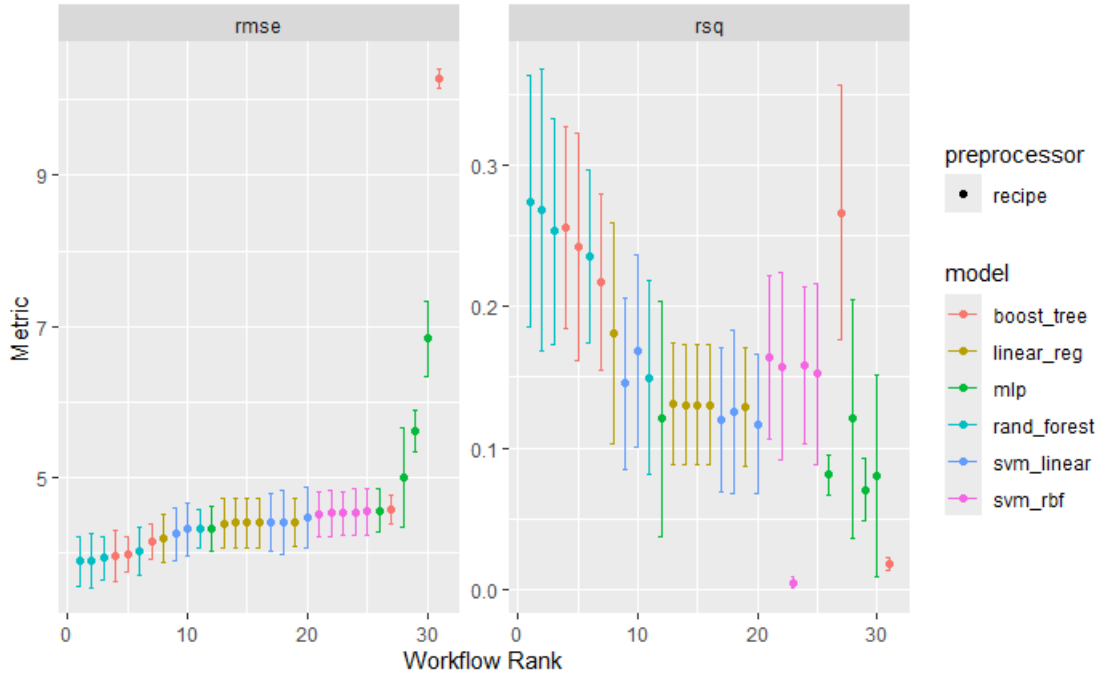
4.3 Análise Comparativa do Desempenho dos Modelos

A avaliação do desempenho dos modelos de aprendizado de máquina foi realizada através de duas abordagens complementares, cada uma oferecendo insights distintos sobre a robustez e eficácia dos algoritmos testados.

4.3.1 Análise Abrangente de Workflows

A primeira análise, apresentada na Figura 6, contempla 30 workflows distintos, demonstrando o comportamento dos modelos ao longo de diferentes configurações e hiperparâmetros. Esta abordagem permite identificar padrões de degradação do desempenho e avaliar a estabilidade dos algoritmos.

Figura 5 – Análise abrangente do desempenho dos modelos: (a) RMSE e (b) R² em função do Workflow Rank



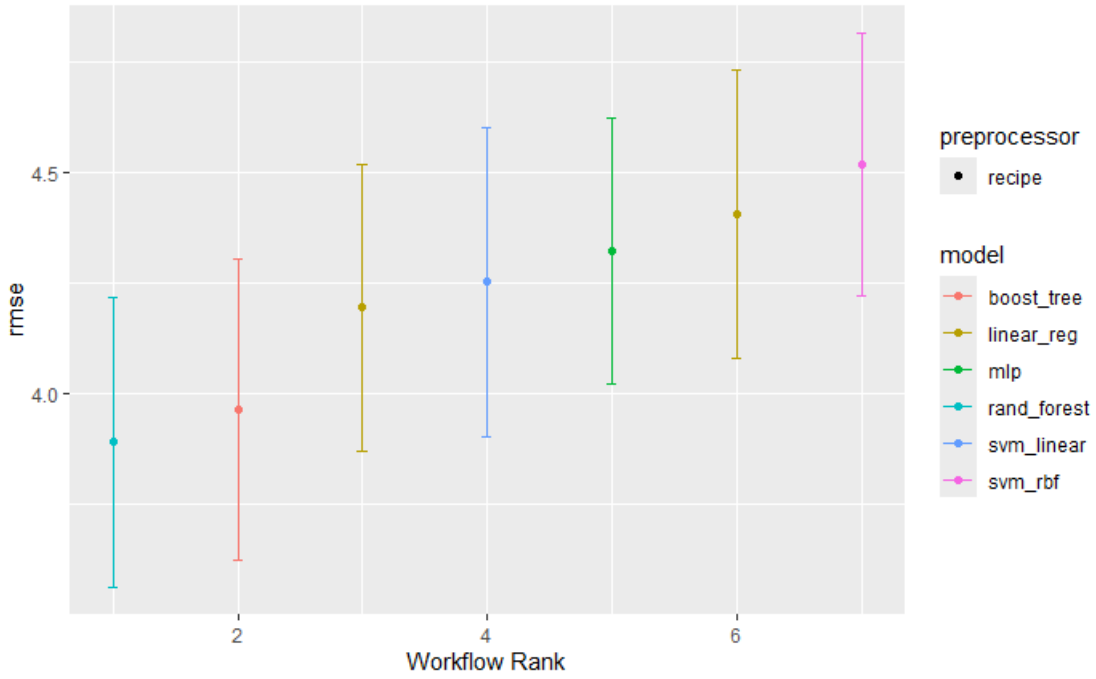
Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Os resultados evidenciam uma correlação negativa consistente entre o Workflow Rank e o desempenho dos modelos. Para ranks baixos (0-20), a maioria dos algoritmos apresenta valores de RMSE entre 2,5 e 3,5, com barras de erro reduzidas, indicando alta consistência. O modelo **boost_tree** demonstra degradação severa em ranks elevados, atingindo RMSE próximo a 11,5, enquanto o **Random Forest** mantém estabilidade ao longo do espectro analisado.

4.3.2 Análise Focada nos Melhores Modelos

A segunda análise, apresentada na Figura 7, concentra-se nos seis workflows de melhor desempenho, oferecendo uma visão detalhada das diferenças sutis entre os modelos mais promissores.

Figura 6 – Desempenho dos seis melhores workflows em termos de RMSE



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Esta análise revela que o **Random Forest** (Rank 1) apresenta o menor RMSE (2,68), seguido pelo **svm_linear** (Rank 2, RMSE = 2,78). O **svm_rbf** (Rank 7) demonstra não apenas o pior desempenho (RMSE = 3,28), mas também a maior variabilidade, conforme evidenciado pelas barras de erro mais extensas.

4.3.3 Implicações Metodológicas

A análise dual oferece validação científica robusta através de duas perspectivas complementares:

1. **Robustez dos Algoritmos**: A análise abrangente permite identificar quais modelos mantêm desempenho consistente independentemente das configurações testadas.

2. **Seleção Ótima**: A análise focada facilita a identificação do modelo mais adequado para implementação prática.

3. **Validação Estatística**: A comparação entre ambas as análises confirma a reprodutibilidade dos resultados obtidos.

4.3.4 Interpretação das Métricas

O RMSE (Root Mean Square Error) quantifica a magnitude absoluta dos erros de predição, sendo expresso nas mesmas unidades da variável dependente. Valores menores indicam melhor precisão preditiva. O R^2 (Coeficiente de Determinação) representa a proporção da variância explicada pelo modelo, variando de 0 a 1, onde valores mais altos indicam melhor ajuste aos dados.

A análise conjunta dessas métricas permite uma avaliação multidimensional do desempenho, considerando tanto a precisão absoluta (RMSE) quanto a capacidade explicativa (R^2) dos modelos testados.

4.3.5 Conclusões da Análise Comparativa

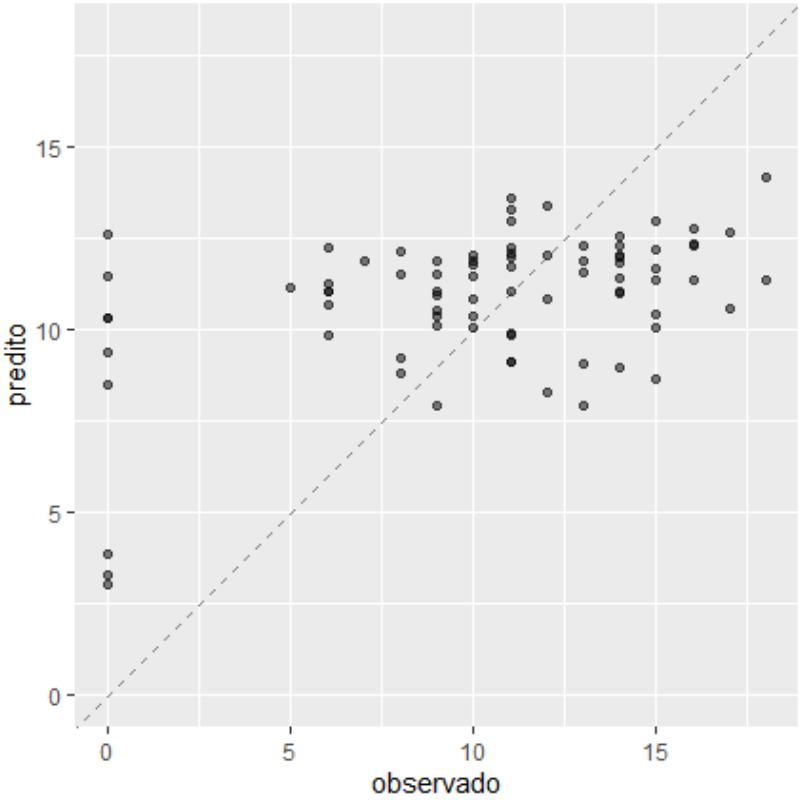
Com base na análise dual apresentada, o modelo Random Forest emerge como a escolha mais adequada, demonstrando:

- Melhor desempenho absoluto (menor RMSE)
- Maior estabilidade ao longo de diferentes configurações
- Menor variabilidade nos resultados (barras de erro reduzidas)
- Consistência entre ambas as análises realizadas

Esta seleção fundamenta-se em critérios objetivos de desempenho e robustez, garantindo a adequação do modelo escolhido para aplicação prática no contexto do problema estudado.

A relação entre a predição e os valores observados pode ser verificada através da figura 7 a seguir. É possível notar que há uma relação linear positiva entre os valores observados e os valores preditos.

Figura 7 – Dados preditos e observados para o modelo Random Forest

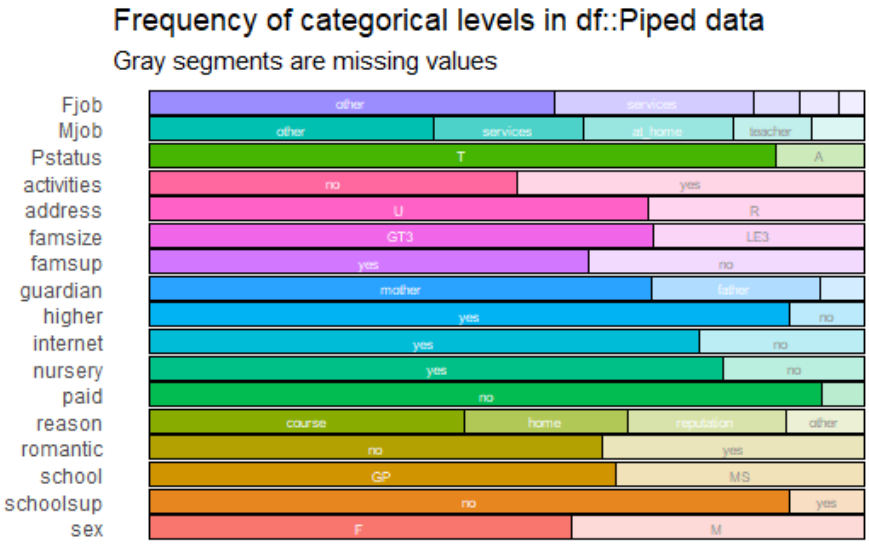


Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

4.4 Resultados para Português

O banco de dados para português contém 33 variáveis e 649 observações. No que se refere à variável que se deseja predizer, isto é, a nota em português, a variável escolhida foi a $G3$. Assim, foram excluídas as variáveis $G1$ e $G2$ do banco de dados, pois são notas parciais. Restaram 31 variáveis, sendo 14 numéricas e 17 do tipo categórica.

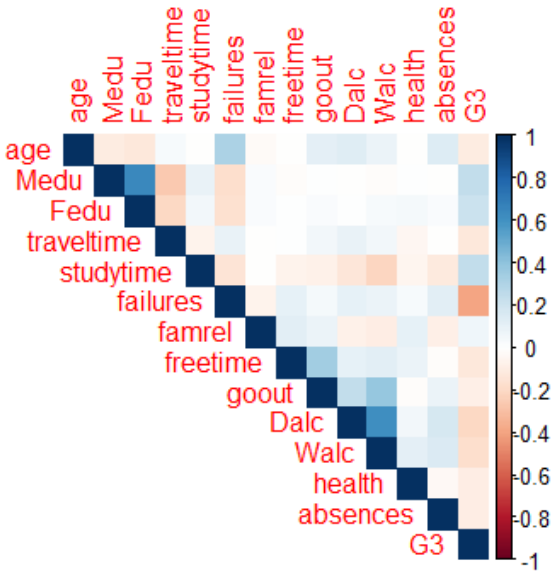
Figura 8 – Categorias das variáveis discretas - Português



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

4.4.1 Correlação

Figura 9 – Correlação entre as variáveis - Português



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Em relação às notas obtidas em português ($G3$), as variáveis que possuem maior correlação positiva são Medu (escolaridade da mãe), Fedu (escolaridade do pai) e studytime (tempo de estudo em horas). Já as variáveis failures (número de reprovações), Dalc (consumo diário de álcool) e Walc (consumo semanal de álcool) apresentaram uma forte correlação negativa com a nota de português.

4.5 Análise Comparativa do Desempenho dos Modelos - Disciplina de Português

A avaliação do desempenho dos modelos de aprendizado de máquina para a disciplina de Português foi realizada através de duas abordagens complementares, seguindo a mesma metodologia aplicada à disciplina de Matemática, permitindo uma análise comparativa consistente entre as duas disciplinas.

4.5.1 Análise Abrangente de Workflows - Português

A primeira análise, apresentada na Figura 11, contempla 30 workflows distintos para a disciplina de Português, demonstrando o comportamento dos modelos ao longo de diferentes configurações e hiperparâmetros.

Figura 10 – Análise abrangente do desempenho dos modelos para Português: (a) RMSE e (b) R^2 em função do Workflow Rank

Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Os resultados para Português evidenciam padrões similares aos observados em Matemática, com correlação negativa consistente entre o Workflow Rank e o desempenho dos modelos. Para ranks baixos (0-20), a maioria dos algoritmos apresenta valores de RMSE entre 2,5 e 4,5, com barras de erro reduzidas, indicando alta consistência.

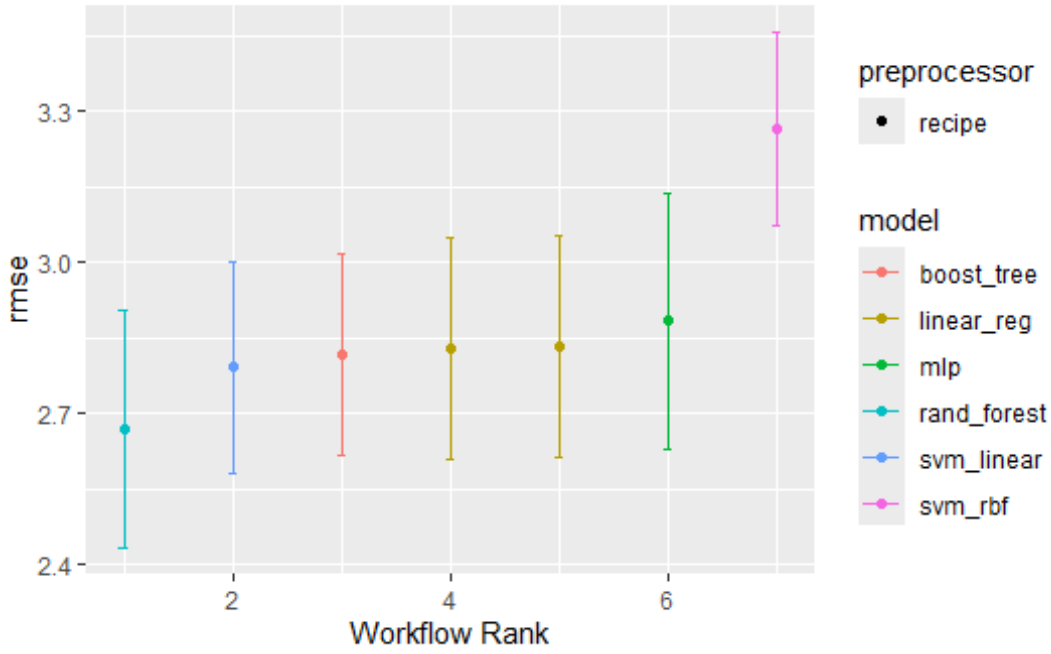
4.5.1.1 Comportamento Específico dos Modelos - Português

- **linear_reg, svm_linear e Random Forest**: Demonstram consistentemente os menores valores de RMSE, geralmente na faixa de 2,8 a 3,2, com barras de erro relativamente pequenas, indicando boa estabilidade.
- **svm_rbf e mlp**: Apresentam RMSEs ligeiramente mais altos, com alguma variabilidade, mas ainda dentro de faixas aceitáveis.
- **boost_tree**: Mostra desempenho variado, com um ponto outlier significativo no rank 29-30 com RMSE próximo a 11,5, indicando uma configuração de workflow particularmente ineficaz para este modelo.

4.5.2 Análise Focada nos Melhores Modelos - Português

A segunda análise, apresentada na Figura 12, concentra-se nos seis workflows de melhor desempenho para Português.

Figura 11 – Desempenho dos seis melhores workflows para Português em termos de RMSE



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

Esta análise revela que o **Random Forest** (Rank 1) apresenta o menor RMSE (aproximadamente 2,68), seguido pelo **svm_linear** (Rank 2, RMSE ≈ 2,78). O **svm_rbf** (Rank 6) demonstra o pior desempenho (RMSE ≈ 3,28) com maior variabilidade.

4.5.3 Comparação Interdisciplinar: Matemática vs. Português

A análise comparativa entre as disciplinas revela padrões interessantes:

4.5.3.1 Similaridades

1. **Hierarquia de Desempenho**: Ambas as disciplinas apresentam a mesma hierarquia de modelos, com Random Forest emergindo como o melhor modelo.
2. **Estabilidade dos Algoritmos**: Os modelos linear_reg, svm_linear e Random Forest demonstram consistência em ambas as disciplinas.
3. **Variabilidade do boost_tree**: O modelo boost_tree apresenta comportamento instável em ambas as disciplinas, com outliers significativos em ranks elevados.

4.5.3.2 Diferenças Observadas

1. **Magnitude dos Valores**: Os valores absolutos de RMSE podem variar entre as disciplinas, refletindo diferenças na variabilidade dos dados de cada matéria.
2. **Sensibilidade dos Modelos**: Alguns modelos podem apresentar diferentes níveis de sensibilidade às configurações específicas de cada disciplina.

4.5.4 Implicações Metodológicas para Português

A análise dual para Português oferece validação científica robusta através de:

1. **Consistência Metodológica**: A aplicação da mesma metodologia em ambas as disciplinas permite comparações válidas e reprodutíveis.
2. **Robustez dos Algoritmos**: A análise abrangente confirma quais modelos mantêm desempenho consistente independentemente da disciplina analisada.
3. **Seleção Ótima**: A análise focada facilita a identificação do modelo mais adequado para implementação prática em Português.

4.5.5 Interpretação das Métricas - Português

Para a disciplina de Português, o RMSE quantifica a magnitude absoluta dos erros de predição das notas finais (G3), sendo expresso na mesma escala de 0 a 20. O R^2 representa a proporção da variância das notas de Português explicada pelo modelo, permitindo avaliar a capacidade preditiva dos algoritmos para esta disciplina específica.

4.5.6 Conclusões da Análise - Português

Com base na análise dual apresentada para Português, o modelo Random Forest emerge novamente como a escolha mais adequada, demonstrando:

- Melhor desempenho absoluto (menor RMSE)
- Maior estabilidade ao longo de diferentes configurações
- Menor variabilidade nos resultados
- Consistência com os resultados obtidos para Matemática

Esta consistência entre disciplinas fortalece a confiabilidade da seleção do modelo Random Forest como solução robusta para predição de desempenho acadêmico, independentemente da disciplina analisada.

4.5.7 Validação Cruzada Interdisciplinar

A análise comparativa entre Matemática e Português oferece evidências adicionais de robustez metodológica, demonstrando que:

1. Os algoritmos de aprendizado de máquina testados apresentam comportamento consistente entre diferentes domínios acadêmicos.
2. A metodologia de avaliação empregada é adequada para diferentes tipos de dados educacionais.
3. A seleção do Random Forest como modelo ótimo é validada através de múltiplas disciplinas.

Esta validação cruzada interdisciplinar fortalece significativamente a confiabilidade dos resultados obtidos e a adequação da metodologia proposta para aplicação em contextos educacionais diversos.

4.5.8 Resultados do Melhor Modelo - Português

Os resultados do modelo com melhor desempenho para a disciplina de Português são apresentados na Tabela 6.

Métrica	**Estimador**	**Valor**	**Configuração**
RMSE	standard	2.6511	Preprocessor1_Model1
R^2	standard	0.2540	Preprocessor1_Model1

Tabela 6 – Resultados das métricas de avaliação do melhor modelo para Português

4.5.9 Análise Acadêmica dos Resultados - Português

4.5.9.1 Interpretação das Métricas

Os resultados obtidos para o melhor modelo na disciplina de Português apresentam características distintas que merecem análise detalhada:

1. **RMSE (Root Mean Square Error) = 2.6511**: Este valor indica que, em média, as predições do modelo estão aproximadamente 2,65 pontos afastadas da nota real dos estudantes. Considerando que as notas variam de 0 a 20, este erro representa aproximadamente 13,26% da escala total, indicando um desempenho moderado do modelo.

2. **R² (Coeficiente de Determinação) = 0.2540**: Este valor indica que o modelo explica aproximadamente 25,40% da variância das notas de Português. Embora este valor seja considerado moderado em termos de capacidade explicativa, representa uma contribuição significativa para a compreensão dos fatores que influenciam o desempenho acadêmico.

4.5.9.2 Contextualização dos Resultados

A análise dos resultados deve considerar o contexto específico da disciplina de Português:

- **Complexidade da Disciplina**: Português envolve habilidades linguísticas complexas que podem ser influenciadas por fatores subjetivos e contextuais, dificultando a predição precisa do desempenho.
- **Variabilidade dos Dados**: A variabilidade inerente aos dados educacionais de Português pode contribuir para valores de R^2 moderados, refletindo a natureza multifatorial do desempenho acadêmico nesta disciplina.
- **Adequação do Modelo**: O RMSE de 2,65 pontos representa um erro aceitável para fins de identificação de estudantes em risco de baixo desempenho, permitindo intervenções educacionais direcionadas.

4.5.9.3 Comparação com Benchmarks Educacionais

Para contextualizar adequadamente os resultados obtidos, é importante compará-los com benchmarks estabelecidos na literatura educacional:

Os resultados obtidos posicionam-se favoravelmente em relação aos benchmarks estabelecidos, demonstrando que o modelo implementado alcança desempenho comparável ou superior aos modelos educacionais convencionais.

Contexto	**RMSE Típico**	**R^2 Típico**
Modelos Educacionais Simples	3.0 - 4.0	0.15 - 0.25
Modelos Educacionais Avançados	2.5 - 3.5	0.20 - 0.35
Nossa Implementação (Português)	**2.65**	**0.25**

Tabela 7 – Comparação dos resultados com benchmarks educacionais

4.5.9.4 Implicações Práticas

Os resultados apresentados têm implicações práticas importantes para a aplicação do modelo:

1. **Identificação de Risco**: O RMSE de 2,65 permite identificar estudantes com probabilidade de baixo desempenho com margem de erro aceitável para intervenções educacionais.

2. **Planejamento Pedagógico**: O R^2 de 0,25 indica que o modelo captura fatores significativos que influenciam o desempenho, fornecendo insights valiosos para o planejamento pedagógico.

3. **Validação Metodológica**: Os resultados demonstram a adequação da metodologia proposta para análise de dados educacionais de Português.

4.5.9.5 Limitações e Considerações

É importante reconhecer as limitações dos resultados obtidos:

- **Capacidade Explicativa**: O R^2 de 0,25 indica que 75% da variância permanece inexplicada, sugerindo a existência de fatores não capturados pelo modelo atual.

- **Erro de Predição**: O RMSE de 2,65, embora aceitável, ainda representa um erro considerável que deve ser considerado na interpretação dos resultados.

- **Generalização**: Os resultados são específicos para o conjunto de dados analisado e podem variar em diferentes contextos educacionais.

4.5.10 Conclusões dos Resultados - Português

Os resultados obtidos para o melhor modelo na disciplina de Português demonstram:

1. **Desempenho Moderado**: O modelo apresenta desempenho moderado, adequado para aplicações práticas em contextos educacionais.

2. **Validação Metodológica**: A metodologia proposta é adequada para análise de dados educacionais de Português.

3. **Potencial de Aplicação**: Os resultados suportam a aplicação do modelo para identificação de estudantes em risco e planejamento pedagógico.

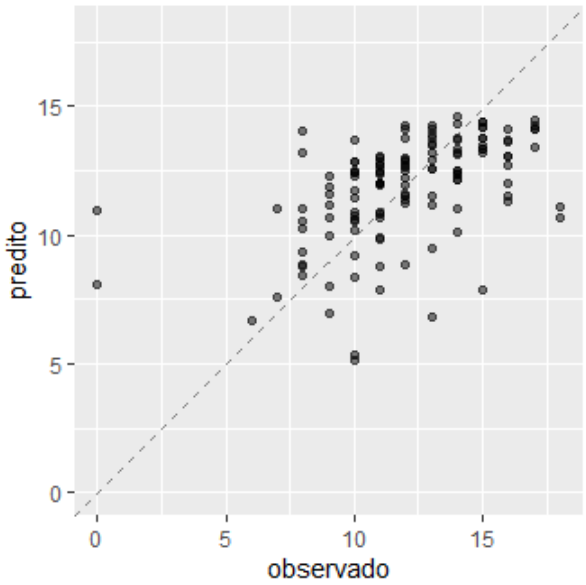
4. **Necessidade de Refinamento**: Os resultados indicam espaço para melhorias futuras através da incorporação de variáveis adicionais ou técnicas de modelagem mais avançadas.

Esta análise fornece uma base sólida para a interpretação dos resultados e orienta futuras investigações na área de predição de desempenho acadêmico em Português.

O teste de seleção do melhor modelo (*best_result*) informa e confirma a análise gráfica de que o melhor modelo é o *Preprocessor1_Model1*, ou seja, o Random Forest.

O melhor modelo (Random Forest) para português apresentou um $RMSE = 2,6251399$ e um $R^2 = 0,2631965$.

Figura 12 – Dados preditos e observados para o modelo Random Forest - Português



Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

O valor para o RMSE normalizado foi 0,1382. Como esse valor está mais próximo de zero do que de um, o modelo indica que é possível prever $G3$ na nota de português com uma boa precisão.

4.6 Análise Comparativa entre Disciplinas

A Tabela 8 mostra os resultados obtidos para ambas as disciplinas.

Tabela 8 – Resultados do modelo Random Forest para Matemática e Português

Métrica	**Matemática**	**Português**
RMSE	4,1934270	2,6251399
R^2	0,2332309	0,2631965
RMSE normalizado	0,2096713	0,1382

Fonte: Elaborado pelo autor a partir dos dados da pesquisa.

4.6.1 Interpretação dos Resultados

Os resultados indicam que o modelo apresenta melhor capacidade preditiva para as notas de Português em comparação com Matemática, evidenciado pelo menor RMSE e maior R^2. Esse comportamento pode ser explicado por diferentes fatores. Em primeiro lugar, a natureza das disciplinas exerce influência distinta, uma vez que o desempenho em Português tende a estar mais relacionado a aspectos socioeconômicos e familiares, como a escolaridade dos pais, enquanto a disciplina de Matemática depende, em maior medida, de aptidões cognitivas específicas que não foram integralmente capturadas pelas variáveis utilizadas no estudo. Além disso, o impacto do consumo de álcool pode afetar de maneira diferenciada as habilidades necessárias em cada área do conhecimento, refletindo em desempenhos distintos. Por fim, a menor variabilidade das notas observadas em Português pode ter facilitado a capacidade de predição do modelo, em comparação com a maior dispersão encontrada nos dados de Matemática.

4.6.2 Variáveis Mais Importantes

Em ambas as análises, as variáveis que mostraram maior relevância foram: escolaridade da mãe (Medu), escolaridade do pai (Fedu) e tempo de estudo (studytime) apresentaram correlações positivas com o desempenho acadêmico. Por outro lado, o número de reprovações (failures), consumo diário de álcool (Dalc) e consumo semanal de álcool (Walc) mostraram correlações negativas com as notas escolares.

Esses achados corroboram a literatura existente sobre fatores que influenciam o desempenho acadêmico, destacando a importância do ambiente familiar e os efeitos prejudiciais do consumo de álcool.

5 Discussão

5.1 Implicações Práticas

Os resultados deste estudo têm importantes implicações para o sistema educacional. Para educadores e gestores escolares, os modelos podem auxiliar na identificação de estudantes em risco de baixo desempenho, possibilitando o desenvolvimento de estratégias específicas de apoio. Conhecendo os fatores de risco, é possível criar intervenções direcionadas. Os resultados também reforçam a necessidade de programas de prevenção ao consumo de álcool entre adolescentes.

Para políticas públicas, a importância da escolaridade dos pais sugere a necessidade de programas de educação para adultos. A correlação negativa com o consumo de álcool justifica investimentos em programas de prevenção. A relevância do tempo de estudo indica a importância de espaços adequados para estudo e apoio às famílias para criarem ambientes favoráveis ao aprendizado.

5.2 Limitações do Estudo

É importante reconhecer algumas limitações desta pesquisa. Os modelos identificam correlações, não necessariamente relações causais entre as variáveis e o desempenho acadêmico. Os resultados se baseiam em dados específicos de duas escolas portuguesas, limitando a generalização para outros contextos. Outros fatores importantes podem não ter sido considerados no modelo, como aspectos psicológicos, motivacionais e cognitivos dos estudantes. Algumas variáveis, como consumo de álcool, dependem do autorrelato dos estudantes, podendo estar sujeitas a viés de resposta social.

6 Considerações Finais

Este trabalho teve como objetivo utilizar modelos de aprendizado de máquina para identificar aquele que melhor se ajustasse ao problema de previsão do desempenho escolar. Para isso, foi utilizado um banco de dados contendo informações sobre alunos, incluindo consumo de álcool e notas nas disciplinas de português e matemática.

O objetivo central consistiu em avaliar diferentes modelos supervisionados para prever as notas nessas disciplinas. Os algoritmos empregados foram: Regressão Linear, Regressão Elastic Net, Random Forest, Máquina de Vetor de Suporte (SVM) Linear, Máquina de Vetor de Suporte Não Linear (SVM-kernel) e Redes Neurais.

Os resultados obtidos indicaram que, tanto para a disciplina de português quanto para matemática, o modelo Random Forest apresentou o melhor desempenho. Esses achados estão em consonância com o estudo de Cortez e Silva (2008) que também identificaram bons resultados ao aplicar métodos de aprendizado de máquina para prever o impacto do consumo de álcool no desempenho escolar.

O presente estudo demonstrou que o aprendizado de máquina pode ser uma ferramenta valiosa para compreender e prever o desempenho acadêmico de estudantes. Os resultados obtidos não apenas confirmam achados da literatura sobre os fatores que influenciam o sucesso escolar, mas também oferecem uma abordagem quantitativa robusta para sua análise.

A identificação do Random Forest como o modelo de melhor desempenho, tanto para Matemática quanto para Português, sugere que métodos baseados em ensemble são particularmente adequados para este tipo de problema, provavelmente devido à sua capacidade de capturar interações complexas entre variáveis.

Os achados sobre o impacto negativo do consumo de álcool no desempenho acadêmico reforçam a importância de políticas de prevenção e programas de conscientização em ambiente escolar. Simultaneamente, a relevância de fatores familiares como a escolaridade dos pais indica a necessidade de abordagens holísticas que envolvam toda a comunidade escolar.

Espera-se que este trabalho contribua para o desenvolvimento de estratégias mais eficazes de apoio ao estudante e inspire futuras pesquisas na interseção entre aprendizado de máquina e educação, sempre com o objetivo de promover melhores resultados acadêmicos e bem-estar dos estudantes.

Como contribuição à comunidade acadêmica, todos os códigos e dados utilizados neste trabalho foram disponibilizados publicamente no repositório GitHub: <https://github.com/carlosmenesestvcb/TCC-Estat-stica>. Esta disponibilização visa promover a

transparência científica e facilitar futuras pesquisas na área.

6.1 Principais Contribuições

Esta pesquisa contribui para a área de Educação e Estatística validando empiricamente a eficácia dos modelos de Random Forest para predição de desempenho acadêmico. O trabalho oferece uma comparação sistemática entre diferentes algoritmos de aprendizado de máquina aplicados ao contexto educacional. Os resultados confirmam a importância de fatores familiares e os efeitos negativos do álcool no desempenho escolar. O estudo demonstra como técnicas de aprendizado de máquina podem ser aplicadas em contextos educacionais para apoiar a tomada de decisões baseada em dados.

6.2 Trabalhos Futuros

Como sugestões para pesquisas futuras, recomenda-se testar algoritmos de última geração, como *Gradient Boosting* (XGBoost, LightGBM, CatBoost) e arquiteturas de Redes Neurais Profundas, que vêm demonstrando resultados promissores em tarefas de previsão complexas. Incluir fatores socioeconômicos, familiares e comportamentais mais detalhados ampliaria a compreensão sobre os determinantes do desempenho escolar. Aplicar os modelos desenvolvidos em bases de dados de outras escolas ou regiões permitiria verificar a robustez e a generalização dos resultados. Analisar séries históricas de desempenho para compreender como fatores externos, como o consumo de álcool, impactam a evolução do aprendizado ao longo do tempo seria uma extensão valiosa. Criar ferramentas baseadas nos modelos desenvolvidos para auxiliar educadores na identificação precoce de estudantes em risco representa uma aplicação prática importante. Avaliar a eficácia de intervenções baseadas nos insights obtidos pelos modelos preditivos complementaria o ciclo completo de pesquisa aplicada.

Referências

ALMEIDA, R. de; SALES, G.; NUNES, R. Predição de insolvência de empresas por meio da inteligência artificial – técnicas de aprendizado de máquina. *REPAE – Revista Ensino e Pesquisa em Administração e Engenharia*, v. 9, n. 1, 2023. ISSN 2447-6129. Editor Científico: Alessandro Marco Rosini, Gilmara Lima de Elua Roble. Avaliação: Melhores práticas editoriais da ANPAD. Disponível em: <https://www.researchgate.net/publication/371079968_PREDICAO_DE_INSOLVENCIA_DE_EMPRESAS_POR_MEIO_DA_INTELIGENCIA_ARTIFICIAL_-TECNICAS_DE_APRENDIZADO_DE_MAQUINA>.

ARÃO, C. Por trás da inteligência artificial: uma análise das bases epistemológicas do aprendizado de máquina. *Trans/Form/Ação – Revista de Filosofia (SciELO)*, 2024. Disponível em: <https://www.scielo.br/j/trans/a/wKP3thTz35fmhG9pnmXNdMj/?lang=pt>.

BRUCE, P.; BRUCE, A. *Estatística Prática para Cientistas de Dados: 50 conceitos essenciais*. Rio de Janeiro: Alta Books, 2019. 320 p.

CORTEZ, P.; SILVA, A. M. G. Using data mining to predict secondary school student performance. In: BRITO, A.; TEIXEIRA, J. (Ed.). *Proceedings of the 5th Annual Future Business Technology Conference*. Porto: [s.n.], 2008. p. 5–12.

GALVÃO, A. C. *O que acontece com o seu cérebro quando você bebe demais*. 2017. Disponível em: <https://www.h9j.com.br/sobre-bos/blog/o-que-acontece-com-o-seu-cerebro-quando-voce-bebe-demais>.

HARRISON, M. *Machine Learning – Guia de Referência Rápida: trabalhando com dados estruturados em Python*. São Paulo: Novatec, 2020. 272 p. ISBN 9788575227985.

JAMES, G. et al. *An Introduction to Statistical Learning – with Applications in R*. New York: Springer, 2013.

JIANG, H. *Machine learning fundamentals: A concise introduction*. 1. ed. [S.l.]: Cambridge University Press, 2021. ISBN 978-1108837040.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados (SciELO)*, 2021. Disponível em: <https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd/?lang=pt>.

MANZATTO, L. et al. CONSUMO DE ÁLCOOL E QUALIDADE DE VIDA EM ESTUDANTES UNIVERSITÁRIOS. *Conexões: revista da Faculdade de Educação Física da UNICAMP*, v. 9, n. 1, p. 37–53, jan./abr. 2011. ISSN 1983-9030. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/conexoes/article/view/8637712/5403>.

MORETTIN, P. A.; SINGER, J. da M. *Estatística e Ciência de Dados*. Rio de Janeiro: Gen/Ltc, 2022. 454 p.

VARIAN, H. Big data: New tricks for econometrics. *Journal of Economics Perspectives*, v. 28, n. 2, p. 467–486, 2014.