# Predição de Desempenho Estudantil: Uma Abordagem com Modelos de Aprendizado de Máquina Supervisionados

**Jamile Guarda[1], Lucas L. Fernandes[2]**

[1]Departamento de Metalurgia – Universidade Federal do Rio Grande do Sul (UFRGS) – Porto Alegre – RS – Brasil

[2]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

jamileguarda@gmail.com, lima.fernandes@inf.ufrgs.br

***Abstract.*** *This paper analyzes the prediction of students' academic performance using supervised machine learning models. After exploratory analysis and data preprocessing, six classifiers were trained and validated: Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, AdaBoost e Decision Tree. Performance was compared using accuracy, precision, recall, and F1-score. The classifier SVC achieved the best results on the test set and was selected as the final model. The results highlight its predictive potential and applicability in the educational field.*

*Keywords: machine learning, student performance, supervised models, Support Vector Classifier (SVC).*

***Resumo.*** *Este artigo analisa a predição do desempenho acadêmico de estudantes com modelos supervisionados de aprendizado de máquina. Após análise exploratória e pré-processamento dos dados, seis classificadores foram treinados e validados: Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, AdaBoost e Decision Tree. O desempenho foi comparado por meio de acurácia, precisão, revocação e F1-score. O classificador SVC apresentou os melhores resultados no conjunto de testes, sendo selecionado como modelo final. Os resultados evidenciam seu potencial preditivo e aplicabilidade na área educacional.*

*Palavras-chave: aprendizado de máquina, desempenho estudantil, modelos supervisionados, Support Vector Classifier (SVC).*

## 1. Introdução

A análise preditiva do desempenho acadêmico de estudantes é um tópico de interesse crescente na interseção entre Educação e Aprendizado de Máquina, pois pode auxiliar na identificação precoce de alunos em risco e orientar intervenções pedagógicas personalizadas (Romero & Ventura, 2020; Kumar & Pal, 2020). Modelos de classificação supervisionada têm sido amplamente aplicados para esse fim, aprendendo automaticamente padrões a partir de conjuntos de dados contendo informações socioeconômicas, comportamentais e notas (Cortez and Silva 2008).

Nesse contexto, algoritmos como Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, AdaBoost e Decision Tree destacam-se como abordagens consolidadas, reconhecidas por sua simplicidade, interpretabilidade e bom desempenho em uma ampla variedade de domínios e tipos de dados.

Neste trabalho, essas técnicas são aplicadas à tarefa de classificação do desempenho estudantil utilizando um conjunto de dados real. O principal objetivo é comparar a eficácia dos modelos em termos de acurácia, precisão, revocação e F1-score, a fim de identificar a abordagem mais adequada para esse contexto.

## 2. Metodologia

### 4.1. Análise exploratória

O presente estudo utilizou o conjunto de dados Student Performance Dataset, disponibilizado publicamente na plataforma Kaggle por Rabie El Kharoua (2024) e originalmente proveniente do repositório UCI Machine Learning. O dataset contém informações de 2.392 estudantes do ensino médio, abrangendo variáveis relacionadas a características demográficas, hábitos de estudo, envolvimento dos pais, atividades extracurriculares e desempenho acadêmico. A variável alvo original, GradeClass, classifica as notas dos estudantes em cinco categorias (de 0 a 4), as quais foram posteriormente mapeadas para os conceitos A a F com base no GPA (nota média) dos alunos.

### 4.2. Pré-processamento dos dados

Inicialmente, os dados foram importados e inspecionados quanto à presença de valores nulos e inconsistências textuais, como strings vazias e espaços em branco em colunas do tipo object. A variável alvo foi reclassificada em um novo atributo binário denominado GradeBinary, no qual os conceitos A, B e C foram considerados como desempenho “bom” (valor 1), enquanto D e F foram classificados como “ruim” (valor 0).

Para mitigar o desbalanceamento entre classes, foi aplicada uma técnica combinada de Synthetic Minority Oversampling Technique (SMOTE) com Tomek Links, restrita ao conjunto de treino. Essa abordagem visa aumentar a representatividade da classe minoritária e remover amostras ambíguas entre as classes.

### 4.3. Treinamento e validação dos modelos

A base de dados foi dividida em três subconjuntos: 60% para treino, 20% para validação e 20% para teste, mantendo a proporção original das classes por meio de estratificação. Seis modelos de classificação supervisionada foram avaliados: Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, AdaBoost e Decision Tree.

Cada modelo foi encapsulado em um pipeline com normalização dos dados por meio do MinMaxScaler e submetido à otimização de hiperparâmetros com uso de GridSearchCV. O processo de validação foi realizado por meio de validação cruzada

estratificada (StratifiedKFold) com cinco subdivisões. A métrica adotada para seleção dos melhores parâmetros foi o F1-score macro.

Os modelos treinados foram avaliados com base nas métricas de acurácia, precisão, revocação (recall) e F1-score, considerando os três conjuntos de dados (treino, validação e teste).

### 4.4. Interpretação e análise crítica

O modelo com melhor desempenho no conjunto de teste foi submetido à análise interpretativa utilizando o método SHapley Additive exPlanations (SHAP), com o objetivo de identificar as variáveis mais relevantes para a predição do desempenho acadêmico.

O modelo SVC, por apresentar os melhores resultados entre os avaliados, foi retreinado com seus hiperparâmetros ótimos e utilizado na avaliação final. Essa etapa incluiu a geração da matriz de confusão e a produção de gráficos de importância dos atributos, contribuindo para a interpretação crítica dos fatores determinantes no desempenho estudantil.

## 3. Resultados e discussão

### 3.1. Avaliação quantitativa dos modelos

A Tabela 1 apresenta os resultados obtidos pelos seis classificadores testados no conjunto de teste, após o ajuste de hiperparâmetros por meio do GridSearchCV com validação cruzada estratificada. Os modelos foram avaliados segundo as métricas de acurácia, precisão, revocação e F1-score.

**Tabela 1. Desempenho dos modelos no conjunto de teste.**

| Modelo | Acurácia | Precisão | Revocação | F1-score |
|---|---|---|---|---|
| SVC | 0,864 | 0,763 | 0,838 | 0,799 |
| Random Forest | 0,854 | 0,766 | 0,786 | 0,776 |
| Logistic Regression | 0,841 | 0,712 | 0,851 | 0,775 |
| AdaBoost | 0,848 | 0,740 | 0,812 | 0,774 |
| Decision Tree | 0,841 | 0,724 | 0,818 | 0,768 |
| KNN | 0,770 | 0,620 | 0,740 | 0,675 |

Observa-se que o modelo SVC apresentou o melhor desempenho em termos de F1-score no conjunto de teste (0,799), superando os demais classificadores. Esse resultado demonstra sua capacidade de equilibrar sensibilidade e precisão, mesmo diante de dados originalmente desbalanceados.

Embora o modelo Random Forest tenha apresentado a maior acurácia no conjunto de treino (0,990) e excelente desempenho na validação, sua performance no teste foi inferior à do SVC, indicando possível overfitting. Já o KNN apresentou o pior

desempenho entre os modelos avaliados, com F1-score de 0,675, revelando limitações na generalização para dados não vistos.

### 3.2 Interpretação com SHAP

A interpretabilidade do modelo SVC foi analisada utilizando o método SHAP, que atribui importância às variáveis com base em sua contribuição para a predição individual. A Figura 1 apresenta o summary plot, destacando os atributos com maior impacto na classificação da classe positiva (desempenho bom).

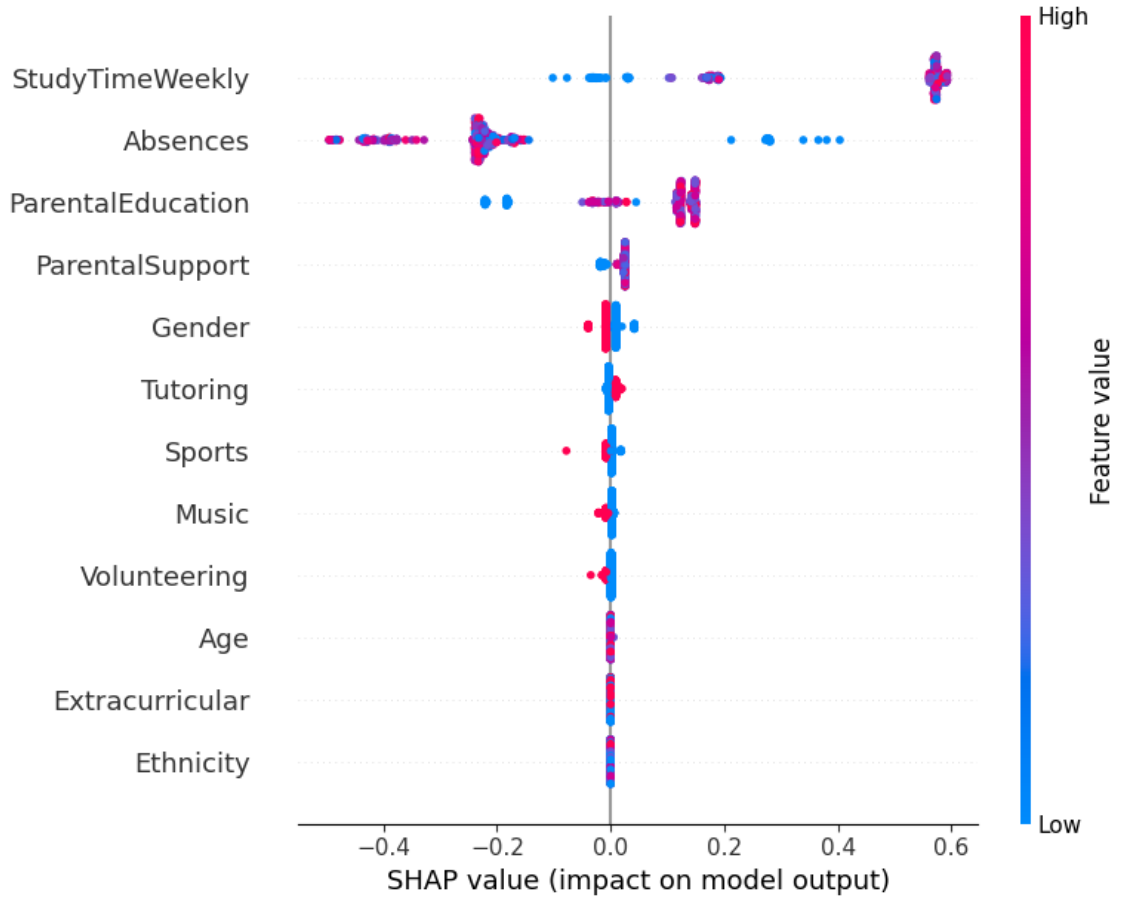


**Figura 1. Valores SHAP para as principais variáveis preditoras.**

As variáveis mais relevantes para o modelo foram: escolaridade dos pais (ParentalEducation), tempo de estudo semanal (StudyTimeWeekly), número de faltas (Absences) e envolvimento em atividades extracurriculares. Esses fatores têm sido amplamente reconhecidos como influências significativas no desempenho acadêmico de estudantes, conforme discutido por Epstein (2001) e pela revisão sistemática apresentada por Castro et al. (2015), o que reforça a confiabilidade dos resultados obtidos.

### 3.3 Avaliação final e matriz de confusão

O modelo SVC foi retreinado com seus hiperparâmetros ótimos — C = 10, kernel = 'rbf' e gamma = 'auto' — e testado novamente para avaliação final. A Figura 2 exibe a matriz

de confusão, permitindo verificar visualmente a performance do modelo na separação entre alunos de bom e mau desempenho.

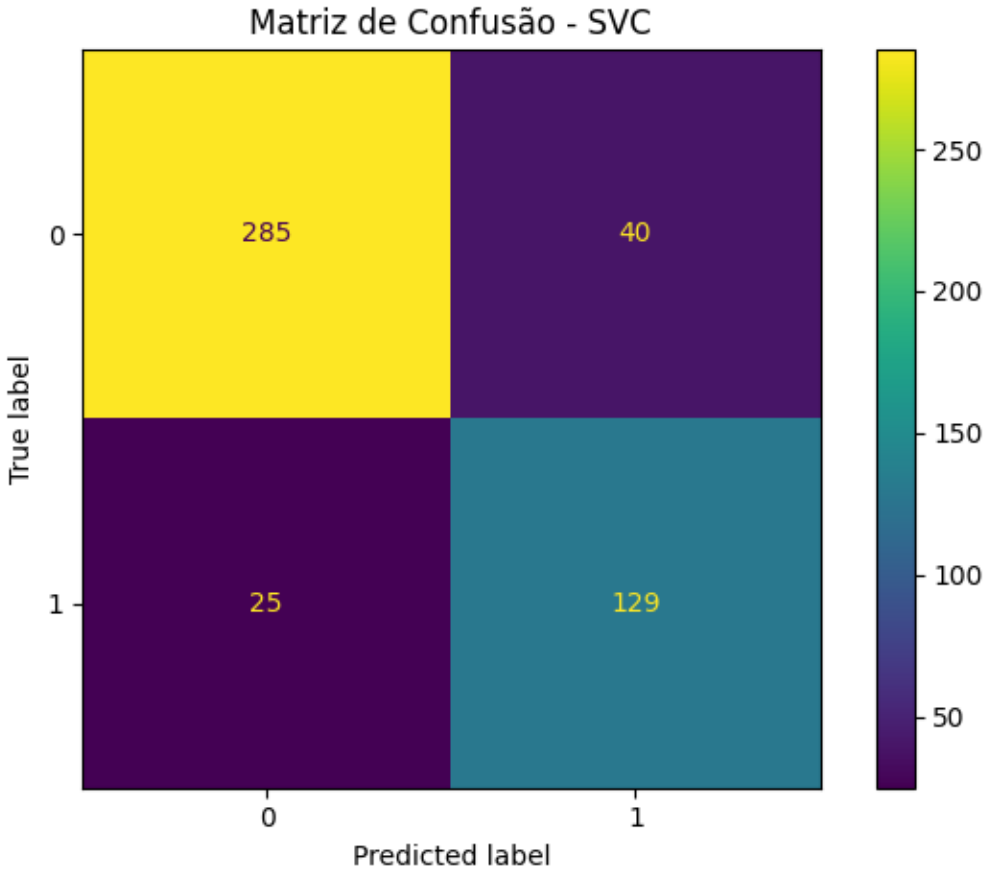


**Figura 2. Matriz de confusão para o modelo SVC.**

A matriz revela bom equilíbrio entre os acertos das duas classes, com um número relativamente baixo de falsos positivos e negativos, o que evidencia a robustez do modelo SVC para a tarefa de predição binária de desempenho.

### 3.4 Discussão

Os resultados demonstram que modelos supervisionados são eficazes para prever o desempenho acadêmico com base em variáveis comportamentais e demográficas. O modelo SVC destacou-se tanto em desempenho quanto em estabilidade entre os conjuntos de validação e teste, sendo considerado o mais adequado neste contexto.

A análise interpretativa com SHAP fornece insights valiosos para aplicações educacionais, evidenciando fatores que podem ser trabalhados por gestores escolares para melhorar os resultados dos alunos. No entanto, destaca-se como limitação a ausência de dados longitudinais (dados coletados ao longo do tempo para acompanhar a evolução dos mesmos alunos) e o uso de informações autodeclaradas (respostas fornecidas pelos próprios alunos ou responsáveis) o que pode introduzir viés. Estudos futuros podem incorporar dados históricos e institucionais para aprimorar a acurácia e generalização dos modelos.

## 4. Referências


Romero, C. and Ventura, S. (2010) “Educational Data Mining: A Review of the State of the Art”, *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, Vol. 40, No. 6, pp. 601–618. DOI: 10.1109/TSMCC.2010.2053532.

Kumar, M. and Pal, S. (2020) “Predicting students’ academic performance using data mining techniques”, *International Journal of Computer Science and Information Technologies*, Vol. 2, No. 2, pp. 686–690.

Cortez, P. and Silva, A. M. G. (2008) “Using Data Mining to Predict Secondary School Student Performance”, In: *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*, Porto, Portugal.

Rabie El Kharoua (2024) “Students Performance Dataset”, Kaggle. Disponível em: https://www.kaggle.com/ds/5195702. DOI: 10.34740/KAGGLE/DS/5195702.

Epstein, J. L. (2001) *School, Family, and Community Partnerships: Preparing Educators and Improving Schools*, Westview Press, USA.

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E. and Gaviria, J. L. (2015) “Parental involvement on student academic achievement: A meta-analysis”, *Educational Research Review*, Vol. 14, pp. 33–46. DOI: 10.1016/j.edurev.2015.01.002.