Fatores que influenciam no aprendizado dos estudantes da Olimpíada Pernambucana de Matemática

Flávio L. M. Melo, Mirele M. Lima, Filipe T. Soares, Roberta A. A. Fagundes, Wylliams B. Santos

Universidade de Pernambuco (UPE) - PE - Brasil

{flavio.leandromorais,mirele.moutinho, roberta.fagundes,wbs}@upe.br, fts@ecomp.poli.br

Abstract. *Data from the latest International Student Assessment Program show 68,1% of students have the worst level of proficiency in mathematics. The article investigates the importance of Pernambuco Olympics in Mathematics (OPEMAT) and its impact on teaching and learning mathematics in the state of Pernambuco, identifying the factors that influence students in school and social life. A case study was conducted using interviews for data collection and educational data mining for analysis and evaluation. The results show that factors, such as study time for OPEMAT and family incentives, positively influence performance.*

Resumo. *Dados do último Programa Internacional de Avaliação de Estudantes apresentam que 68,1% possuem o pior nível de proficiência em matemática. Este artigo investiga a importância da Olimpíada Pernambucana de Matemática (OPEMAT) e o seu reflexo no ensino e aprendizado da matemática no estado de Pernambuco, identificando os fatores que influenciam na vida escolar e social dos estudantes. Foi conduzido um estudo de caso com a utilização de entrevistas para coleta dos dados e mineração de dados educacionais para análise e avaliação. Os resultados mostram que fatores como o tempo de estudo para a OPEMAT e o incentivo familiar, influenciam positivamente no desempenho.*

1. Introdução

As Olimpíadas de Matemática têm se estabelecido como políticas públicas que visam elevar o interesse e a compreensão do ensino-aprendizado da matemática no Brasil. Segundo Dos Santos et al. (2018), "Olimpíadas de matemática estimulam e divulgam o conhecimento e o interesse pela matemática e pelas ciências, desperta nos alunos o interesse pelos desafios matemáticos e pelas resoluções de situações problemas, desenvolvem o raciocínio e melhoram a qualidade do ensino."

O Brasil possui o maior evento olímpico da área no mundo, a Olimpíada Brasileira de Matemática das Escolas Públicas e Privadas (OBMEP), que firmou a parceria entre escolas, universidades e instituições de pesquisa. Seu principal objetivo é promover o desenvolvimento de habilidades matemáticas e a inclusão social por meio da difusão de conhecimento OBMEP (2019). Ela inspirou a criação da Olimpíada

Pernambucana de Matemática (OPEMAT), um projeto de extensão desenvolvido pelo Departamento de Matemática da Universidade Federal Rural de Pernambuco em conjunto com outras IES de Pernambuco. A OPEMAT esteve em 2019 em sua quinta edição, com o desafio de contribuir para a melhoria da qualidade da educação básica.

A OPEMAT é uma competição para alunos dos ensinos fundamental e médio das escolas públicas e particulares, através de prova realizada em 10 (dez) polos espalhados pelo estado, numa única fase, e da realização de capacitação para professores do ensino fundamental e médio. A prova é fatorada em três níveis e com públicos diferentes, distribuídos da seguinte forma: o 6º e 7º anos do ensino fundamental correspondem aos competidores do nível 1; 8º e 9º do ensino fundamental, os do nível 2, e os três anos do ensino médio, são os do nível 3 OPEMAT (2019).

Nesta pesquisa procurou-se evidências do impacto da OPEMAT na vida dos seus participantes, tornando-se necessário coletas de dados através de entrevistas com professores, coordenadores e estudantes competidores. Para a análise do repositório dos dados colhidos, foi utilizada a mineração de dados educacionais (MDE). Os métodos de Regressão Linear (RL), Regressão Robusta (RR) e Multi-Layer Perceptron (MLP), buscaram investigar e explicar as variáveis educacionais que relacionam a OPEMAT e o desempenho da nota de matemática do estudante participante.

2. Revisão da literatura

O estudo de Vitorino et al. (2018), objetivou identificar o perfil do bom desempenho dos medalhistas da Olimpíada Paraibana de Informática, para que tais característica servisse como base para preparação e motivação para novos participantes. Para isso, foram realizadas entrevistas semiestruturadas e estruturadas com os medalhistas dos anos de 2016 e 2017, para a construção de um banco de dados, que posteriormente foi processado por meio das linguagens Python e R.

O trabalho de Nascimento e Júnior (2018) aplicou técnicas da MDE, com a finalidade de estimar o indicador de docentes com cursos superior no cenário da educação básica e também comparar o desempenho dos modelos de RL e RR pelo índice de erro de previsão. Portanto, apresentou a RR como o modelo que mais se adequa ao estudo, já que ele obteve o menor erro de previsão da variável elencada.

Já o estudo de Silva et al. (2019) utilizou os modelos paramétricos e não paramétricos com o objetivo de identificar padrões que influenciam no desempenho dos estudantes do 5° ano do ensino fundamental do estado de Pernambuco em Língua Portuguesa e Matemática. Propondo-se a investigar essa aplicabilidade da MDE para a previsão de desempenho de alunos e auxiliar professores e gestores na tomada de decisões estratégicas no cenário amplo da educação.

O artigo de Silva Pinto et al. (2019) conduziu um estudo de caso que identificou fatores que influenciam no resultado do IDEB das escolas municipais de Maceió-AL. Para isso, foram utilizadas técnicas de seleção de atributos, no contexto da MDE, para

avaliar e contribuir com informações relevantes para o apoio de educadores e gestores na tomada de decisão.

O presente estudo se diferencia dos demais por conduzir um estudo de caso baseado em entrevista através da utilização da MDE, no tocante a contribuição da OPEMAT no desenvolvimento e obtenção do conhecimento. Assim, através da aplicação de modelos de previsão (RL, RR e MLP) identifica-se informações precisas dos resultados que colaborem com a sociedade no que diz respeito à execução da OPEMAT, como também, quanto tem contribuído para o desenvolvimento cognitivo, a inclusão social e a melhoria na qualidade do ensino e aprendizado da matemática nas escolas pernambucanas, avaliando o desempenho da nota de matemática do estudante.

3. Estudo de Caso

O presente estudo utilizou as práticas do *guideline* para condução do estudo de caso proposto por Runeson e Host (2009), conforme etapas apresentadas na Figura 1.

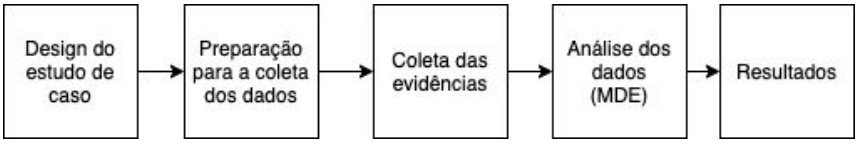

Figura 1. Etapas do estudo de caso inspirado em Runeson e Host (2009).

3.1 Design do estudo de caso

De modo preliminar ao planejamento do estudo de caso, foi realizado um estudo teórico sobre as práticas, aplicações, evoluções, funcionalidades e regulamentação da OPEMAT nos anos da sua execução. Também se averiguou sobre competições similares como a Olimpíada Brasileira de Matemática (OBM), OBMEP e a Olimpíada Internacional de Matemática (IMO), que são reconhecidas nacional e internacionalmente como políticas públicas de incentivo e melhoramento do ensino-aprendizagem da matemática.

A etapa seguinte consistiu no planejamento do estudo de caso, averiguando-se a influência da OPEMAT no ensino-aprendizado da matemática, através da rotina dos alunos participantes. Para isso, foram definidas coletas de dados com diferentes atores representativos (diretores, coordenadores, professores e alunos) envolvidos nessa olimpíada, por meio de entrevistas estruturadas e semiestruturadas.

3.2 Preparação para a coleta de dados

Para a coleta de evidências, foram aplicadas as técnicas de entrevistas baseadas nas diretrizes definidas por Strandberg (2019), considerando os seguintes aspectos e instrumentos: consentimento do entrevistado, benefício à sociedade, confidencialidade, valor científico, habilidade do pesquisador, justiça, respeito à lei e ética na revisão. A primeira etapa das entrevistas foi direcionada aos profissionais participantes da

OPEMAT e olimpíadas do gênero, tais como coordenadores acadêmicos e professores escolares responsáveis pela disciplina de matemática, da cidade de Caruaru.

A etapa seguinte, voltada ao estudante competidor, estabeleceu uma comissão composta por professores doutores em matemática, psicologia e estatística e uma doutoranda em educação, com o propósito de conceber um questionário que avalie o reflexo da competição nas vidas dos estudantes, adotando como base modelos aplicados pelo Índice de Desenvolvimento da Educação Básica (IDEB)[1], Instituto Nacional de Estudos e Pesquisa Educacionais Anísio Teixeira (INEP)[2].

O questionário elaborado é um instrumento de natureza quantitativa e segue um modelo pirâmide com a maioria das questões fechadas, contendo 24 perguntas, utilizando a escala Likert[3], que abordam a vida escolar e pessoal do participante. Salienta-se que foram desenvolvidos dois tipos de questionários, sendo um para o ensino fundamental[4] e o outro para o ensino médio[5]. Para sua de validação, os mesmos foram simulados como teste piloto, em um formulário google, a estudantes competidores de uma escola do município de Caruaru.

3.3 Coleta de evidências

No primeiro momento desta etapa foram realizadas entrevistas, caracterizadas por uma pesquisa de caráter qualitativa, com a abordagem de um questionário[6] aberto com 15 perguntas para profissionais, baseando-se em olimpíadas nessa temática. Dando continuidade, foram realizadas as transcrições dessas entrevistas para um documento, onde foram catalogadas todas as conversas ocorridas na pesquisa de campo. Em seguida foi feita análise do documento, permanecendo na descrição tudo que realmente tinha coerência com o caso estudado. Salienta-se que no decorrer de todo o processo, o anonimato dos entrevistados foi mantido. Por fim, ocorreu a geração de um documento com todas as análises.

Com a etapa de validação do questionário concluída, a coordenação geral da OPEMAT os imprimiu e anexou ao material destinado aos polos Recife e Caruaru. Foi acordada a aplicação dos questionários no mesmo dia da realização do concurso, após o ingresso do competidor à sala, ou imediatamente após sua conclusão da prova, harmonizando com a não obrigatoriedade do aluno a respondê-lo. Ao final o material foi coletado e as respostas foram transcritas, formando a base de dados deste estudo.

3.4 Análise dos Dados

Nesta seção serão apresentadas as etapas da análise dos dados através da utilização da metodologia de MDE, *Cross-Industry Standard for Data Mining* (CRISP-DM)

---

[1] INEB http://portal.inep.gov.br/ideb
[2] INEP http://www.inep.gov.br/
[3] Escala Likert http://encurtador.com.br/iBL07
[4] Questionário ensino fundamental http://encurtador.com.br/ckAO1
[5] Questionário ensino médio http://encurtador.com.br/mpC07
[6] https://drive.google.com/drive/folders/1Y3QwpXpi075JNfK0ZyuJHs2wlplNCqUK?usp=sharing

Chapman et al. (2000), a qual é composta de seis fases. O uso desta metodologia torna o processo de mineração de dados mais rígido, além de ser uma das metodologias mais populares no contexto da MDE.

3.4.1 Compreensão do Negócio

Nesta fase são definidos os objetivos do projeto. Nesse sentido, busca-se com este projeto identificar os fatores que influenciam o desempenho dos estudantes na OPEMAT, utilizando modelos de previsão (regressão) aplicados à base de dados coletada na subseção 3.3.

3.4.2 Compreensão dos Dados

Nessa pesquisa foram coletados dados por meio de entrevistas, cujas informações contidas apresentam o relacionamento do ensino-aprendizado da matemática refletidos nos estudantes participantes. O devido contexto buscou abordar dados do ensino fundamental, onde inicia as atividades no campo das olimpíadas, e depois fazendo uma junção com informações adquiridas de alunos do ensino médio, que já possuem habilidades específicas em competições. Todos os dados são compostos por diversos atributos de respostas de alunos, as quais incluem notas de proficiência na disciplina de matemática e o desempenho nas olimpíadas de matemática.

3.4.3 Preparação dos Dados

A base de dados é composta por 24 atributos, mas identificou-se atributos nominais, das quais foram desconsideradas para o estudo, pois os modelos de previsão (regressão) utilizam apenas variáveis quantitativas, assim apenas 19 variáveis numéricas foram consideradas para predição do desempenho do estudante na OPEMAT. Estas 19 variáveis foram classificadas em dois tipos: variável resposta ou dependente (y) e variáveis explicativas ou independentes (x) para construção do modelo de regressão.

A variável resposta utilizada foi Nota em Matemática (y) e as 18 variáveis explicativas (x) e sua respectiva correlação pearson com variável (y) foram: Série (0.197), Idade (0,248), Sexo (0,061), Tipo da escola (0,239) Acesso a computador em casa (0,008), Acesso a computador na escola (0,053), Quantidade de olimpíadas (0,092), Estuda matemática por conta própria (0,101), Quantidade de horas que estuda(0,063), Reprovado (0,383), Quantidade de reprovação (0,376), Notas em Português (0,930), Reforço de matemática por semana (0,206), Treinamento para olimpíadas (0,667), Estímulo familiar (0,848), Estímulo de amigos (0,781), Estímulo por influência da OPEMAT (0,874) e Estímulo em outras áreas devido à OPEMAT (0,683). Portanto, as variáveis independentes que têm maior influência na nota de matemática do estudante foram aspectos relacionados ao dia a dia do aluno. Família e Amigos, como também participação dos alunos em Olimpíadas, servem de grande estímulo.

Além disso, os dados ainda foram padronizados em 0.15 e 0.85 para realizar o processo de modelagem conforme apresentado em Nascimento et al. (2018). Isso se deu

pois muitos dados eram categorizados em 0 ou 1, e nesse sentido existiam muitos valores 0. Nessa padronização não existiria um sobre-ajuste nos modelos de regressão.

3.4.4 Modelagem

Os modelos de regressão propostos por Montgomery et al. (2012) e aplicadas neste trabalho são dados pela seguinte definição: seja $x = x_0, ..., x_n$ um vetor de variáveis independentes, $\beta = \beta_0, ..., \beta_n$ um vetor de parâmetros e seja $\varepsilon = \varepsilon_0, ..., \varepsilon_n$ um vetor de erro aleatório. A equação do modelo linear à coleção dos dados i=1,…, n é dada por:

$$yi = \beta_0 + \beta_{xi} + \varepsilon_i. \quad (1)$$

Conforme a Equação 1, o modelo de regressão utilizado neste trabalho: a RL, que possui por determinação o vetor estimado pelo método dos mínimos quadrados, minimizando uma função baseada na soma dos resíduos quadrados (ε_i), concedida por:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \quad (2)$$

onde y_i é a variável resposta e $\widehat{y_i}$ a variável resposta conjecturada pelo modelo. E o segundo modelo definido é a RR, no qual o vetor é estimado, minimizando uma função critério baseada na função p, que é dada por:

$$\sum_{i=1}^{n} p\left(\frac{y_i - \beta_{xi}}{\sigma}\right), \quad (3)$$

no qual σ é um estimador robusto e p uma função particular.

Adicionalmente, será feito o uso da rede *Multi-Layer Perceptron* (MLP), que é a generalização da rede Perceptron e possui como sua principal diferença a existência de uma terceira camada, denominada como camada escondida. De acordo com Ferreira et al. (2016) uma MLP possui na sua camada de entrada um conjunto de nós ou neurônios, uma ou mais camadas escondidas, e uma camada de saída. A rede MLP normalmente utiliza como suas funções de ativação a tangente hiperbólica e a sigmóide logística, e possui um treinamento supervisionado, o que implica que ao utilizá-la como modelo para este artigo, a rede possuirá uma nota desejada para os parâmetros de entrada, e buscará ver o quão próximo a sua predição chega da nota desejada.

Pode-se utilizar o exemplo dado por Matt (2015) em referência a como a rede neural trabalha na atualização de pesos, cálculo de erros e aprendizado.

3.4.5 Avaliação

Nesta fase foi realizada a avaliação dos modelos de previsão (RL, RR e MLP) através de uma métrica. Utilizou-se a métrica do erro médio absoluto, do inglês *Mean Absolute Error* (MAE). A Equação 4 apresenta esse erro de previsão e é representada por:

$$\frac{1}{n} \sum_{i=1}^{n} \left| y_i - \widehat{y_i} \right|, \quad (4)$$

onde n é o tamanho do conjunto de dados, y_i é a valor real da variável e $\widehat{y}_i$ é o valor estimado/predito pelo modelo de regressão (RL, RR e MLP). Dessa forma, possibilitou a aplicabilidade de teste estatístico para os modelos propostos.

Conforme as análises realizadas, foi proposto como modelo de estudo os fatores (variáveis/perguntas) que influenciam o desempenho da matemática na vida escolar dos alunos e desempenho na OPEMAT. Para o estudo, é importante considerar a nota de matemática dos alunos, como variável resposta (y), e as demais como variáveis explicativas (x). Portanto, o experimento se realizou pelas execuções do Monte Carlo (MC) descrito no Algoritmo 1, onde todos os experimentos e análises foram realizados no ambiente de desenvolvimento da linguagem Python.

Algoritmo 1. Simulação Monte Carlo

1: **Determinar** MC = 1000
2: **Para todo** i ← 1 até MC **faça:**
3: **Particionar** aleatoriamente a coleção de dados em dois conjuntos, sendo o primeiro em treino com (75% da base) e o outro em teste com (25% restante).
4: **Aplicação** dos modelos de regressão a partir da base de treino (Equações 1, 2, 3).
5: **Aplicação** da regra de predição para a base de teste.
6: **Calcular** o erro de prognóstico (MAE) para cada modelo aplicado (Equação 4).
7: **Guarda** os resultados de erro de cada modelo.
8: **Fim para**
9: **Apurar** os resultados das análises da média e desvio padrão dos erros {MAE} e teste estatístico de todos os modelos.

4. Discussão dos Resultados

Com o propósito de verificar a efetividade da rede MLP na predição das notas dos alunos e de comparar esta rede com os valores obtidos por regressões lineares e robustas, foram efetuadas 20 iterações de treinamentos e predições, e em cada iteração 1/3 da base fora sorteada randomicamente e utilizada como teste e 2/3 dela como treino.

Tabela 1. Erros dos modelos.

Modelos	RL	RR	MLP
Média MAE	0.0582	0.0535	0.0678
Desvio Padrão MAE	0.0073	0.0079	0.0153
Média MSE	0.0082	0.0075	0.0142
Desvio Padrão MSE	0.0045	0.0048	0.0097
Média R^2	0.8638	0.8732	0.7703
Desvio Padrão R^2	0.0882	0.0929	0.1421

A partir disto, tornou-se possível observar o resultado do R quadrado dos 3 modelos e fazer uma comparação dele por iterações, como pode-se observar no gráfico:

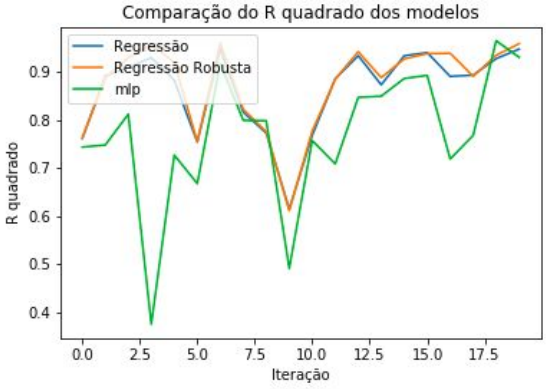

Figura 2. Comparação entre modelos.

Torna-se evidente que apesar da MLP possuir um resultado bastante satisfatório na predição das notas dos alunos, este resultado ainda deixa a desejar quando comparado com as técnicas de RL e RR.

Com o propósito de assegurar a maior eficácia das regressões comparadas com a MLP, foi efetuado o teste de hipótese de *wilcoxon*. O resultado obtido ao comparar o R quadrado de uma RR com uma MLP utilizando *wilcoxon* foi abaixo (*p-value* = 0.0003) de 0.05, o que confirma que os valores podem ser considerados distintos, assim como a comparação da RR com a RL (*p-value* = 0.0089) que também se mostrou abaixo de 0.05, o que nos permite afirmar que existe uma diferença entre os valores obtidos pelos modelos das regressões.

Por fim, levanta-se a hipótese de que um dos possíveis motivos dos resultados da MLP não serem tão satisfatórios como o das regressões, seja pela quantidade ínfima de dados de treinamento. Em algumas iterações onde houve uma discrepância entre os valores dos modelos, pode-se observar que a MLP torna-se enviesada pelos dados de treino, não conseguindo igualá-los aos dados de teste. Uma das iterações onde isto fica evidente é a iteração 5, e podemos observar este fato no gráfico de comparação do MAE entre os dados de teste e de treino.

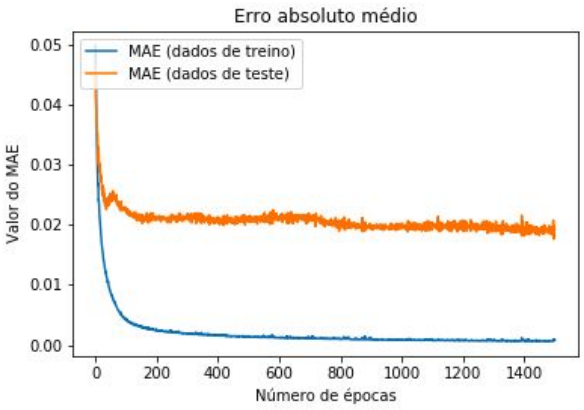

Figura 3. Comparação de erro absoluto médio entre treino e teste.

5. Conclusões

O devido trabalho buscou estudar e identificar os fatores que influenciam na vida estudantil dos competidores da OPEMAT. Para isso, construiu-se uma base de dados com questionário desenvolvido e entrevistas aplicadas aos participantes dos polos Recife e Caruaru no ano de 2019. Os modelos aplicados mostraram resultados significativos, os valores de RR se mostraram melhores em relação a RL e MLP , como também seu desempenho é ratificado com teste de hipótese.

As principais contribuições da pesquisa foram: a identificação de fatores baseados no estudo de caso e a aplicação de três modelos de previsão para análise do desempenho da nota de matemática na OPEMAT. Desse modo, o uso de MDE possibilitou identificar fatores como o estímulo da família, amigos e participação em outras olimpíadas como importantes para maximizar o desempenho do aluno na nota de matemática da OPEMAT. Além disso, sugere-se a utilização de políticas e investimentos para estímulos dos estudantes na participação de olimpíadas.

De acordo com Chapman et al. (2000), a última etapa da metodologia CRISP-DM é dedicada a disponibilização das informações e conhecimentos na indústria. Deste modo, identificamos essa etapa como uma limitação do estudo a ser endereçada como trabalho futuro.

6. Referências

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., (2000). Crisp-dm 1.0: Step-by-step data mining guide. SPSSinc, 9:13

da Silva Pinto, G., Júnior, O. F., Costa, E., Barbirato, J. C. C., and Rodrigues, W. R. M.(2019). Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. *In Brazilian Symposium On Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*,volume 30, page 1828.

Dos Santos, A. J., Batista, A. V., and Melo, F. S. (2018). A importância da olimpíada municipal de matemática nas escolas públicas de itabaianinha-se. *Encontro Internacional de Formação de Professores e Fórum Permanente de Inovação Educacional*, 11(1).

Ferreira, A., Ferreira, R. P., da Silva, A. M., Ferreira, A., and Sassi, R. J. (2016). Um estudo sobre previsão da demanda de encomendas utilizando uma rede neural artificial. *Blucher Marine Engineering Proceedings*, 2(1):353-364.

Matt, M. (2015). A Step by Step Backpropagation Example. Disponível em: <https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/> Acesso em: 16 jul 2020.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Introduction to linear regression analysis, volume 821. John Wiley & Sons.

Nascimento, R. and Júnior, G. C. (2018). Estudo sobre docentes do ensino básico através de indicadores educacionais e modelos de regressão. *In Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 7, page 379.

Nascimento, R. L. S., das Neves Junior, R. B., de Almeida Neto, M. A., and de Araújo Fagundes, R. A., (2018). Educational data mining:An application of regressors in predicting school dropout. *In International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 246–257.* Springer.

OBMEP. Disponível em: <http://www.obmep.org.br/> Acesso em: 09/09/2019.

OPEMAT. Disponível em <http://www.opemat.com.br/> Acesso em : 09/09/2019.

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2), 131–164.

Silva, P., do Nascimento, R. L. S., Lima, M., Fagundes, R., and de Souza, F. d. F. (2019). Modelos de regressão aplicados a predição do desempenho escolar de estudantes do ensino fundamental. *In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1621.

Strandberg, P. E. (2019). Ethical interviews in software engineering. *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Porto de Galinhas, Recife, Brazil, 2019, pp. 1-11.*

Vitorino, M., Silva, H., Sampaio, L., and Gheyi, R. (2018). Perfil dos premiados em olimpíadas de informática e sua influência sobre a educação em computação. *In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 228.