Aplicação de algoritmos de classificação para prever a evasão escolar

Caio E. S. Lopes, PPGEC - Universidade de Pernambuco,
cesl@ecomp.poli.br, https://orcid.org/0000-0001-7969-5754
João Antonio da Silva Lima, PPGEC - Universidade de Pernambuco,
jasl@ecomp.poli.br, https://orcid.org/0000-0001-5010-0024
Roberta A. A. Fagundes, Universidade de Pernambuco, roberta.fagundes@upe.br
https://orcid.org/0000-0002-7172-4183

Resumo: A evasão escolar é uma questão de grande impacto que afeta sistemas educacionais globalmente. A definição de uma análise mais rigorosa que identifique fatores que possam melhorar o desempenho e previsão dos algoritmos de classificação. Portanto, a preparação dos dados é indispensável para aplicação desses algoritmos, tanto para aumentar a qualidade dos resultados, como para um melhor entendimento da natureza dos dados. O objetivo é aplicar algoritmos de classificação para prever a evasão escolar, incorporando uma análise mais crítica na preparação dos dados através da redução de dimensionalidade e tempo de execução dos algoritmos. A metodologia definida para atingir esses objetivos possui 4 etapas (entendimento dos dados, preparação dos dados, modelagem e avaliação) e foram aplicadas de maneira cíclica para identificar melhor os processos. Nos resultados ao aplicar transformações nos dados e redução de características, observa-se uma melhor performance dos algoritmos aplicados quando comparados com os da literatura. Este estudo ressalta a importância na adequação de dados para um padrão que possa ser absorvido durante os diferentes algoritmos de classificação utilizados.

Palavras-chave: Algoritmos de Classificação, Evasão Escolar e Pré Processamento

Application of classification algorithms to predict school dropout.

Abstract: School dropout is a highly impactful issue that affects educational systems globally. The definition of a more rigorous analysis to identify factors that can enhance the performance and prediction of classification algorithms is essential. Therefore, data preparation is indispensable for the application of these algorithms, both to improve the quality of results and to gain a better understanding of the nature of the data. The goal is to apply classification algorithms to predict school dropout, incorporating a more critical analysis in data preparation through dimensionality reduction and algorithm execution time. The methodology defined to achieve these objectives consists of 4 steps (data understanding, data preparation, modeling, and evaluation) and was applied cyclically to better identify the processes. In the results, when applying data transformations and feature reduction, a better performance of the algorithms is observed when compared to those in the literature. This study highlights the importance of data adaptation to a standard that can be absorbed by the various classification algorithms used.

Keywords: Classification Models, School Dropout, Machine Learning, Academic Risk, Educational System Enhancement.

1. Introdução

A evasão escolar, como apontado por Torres Marques et al. (2022), representa um desafio significativo que exerce um impacto adverso sobre as instituições de ensino, acarretando consequências nas esferas social, acadêmica, econômica e ambiental. Esses impactos negativos, por sua vez, repercutem nas políticas de investimento e desenvolvimento educacional.

A melhoria da qualidade da educação com o objetivo de obter melhores resultados em avaliações educacionais, conforme discutido por Batista e Fagundes (2023), não é uma tarefa simples. No entanto, em um contexto orientado por dados, a utilização de ferramentas como mineração de dados e técnicas de aprendizado de máquina possibilita a identificação de fatores significativos que podem influenciar esses resultados.

A pesquisa de Hasan, Rabby, Islam e Hossain (2019) destaca a importância dos algoritmos de Machine Learning na esfera educacional, visando a previsão e otimização do desempenho dos estudantes. Nesse contexto, os registros acadêmicos dos alunos desempenham um papel essencial, fornecendo informações valiosas sobre seus hábitos de estudo, preferências em relação às disciplinas e, possivelmente, seus níveis de aptidão intelectual. A aplicação de algoritmos de Machine Learning, conforme destacado pelos autores, emerge como um meio eficaz para identificar tendências e padrões nesses dados, possibilitando a antecipação do desempenho acadêmico de um aluno. Isso evidencia a relevância da aplicação de algoritmos de Machine Learning na previsão da evasão escolar.

O presente trabalho tem como objetivo implementar melhorias e generalizações nos dados, visando uma classificação mais precisa da evasão de alunos com base no conjunto de dados: *"Predict students dropout and academic success"* (2011).

2. Trabalhos Relacionados

Edson e Solange (2021) em seu estudo sobre evasão escolar com base na análise de dados do curso de Sistemas de Informação da Universidade Federal de Santa Maria - Campus Frederico Westphalen, os autores propuseram um estudo sobre os padrões de evasão escolar no ensino superior, O estudo passou por um rigoroso processo de pré-processamento e limpeza de dados, resultando em 409 registros de alunos. Foi realizada a padronização da situação do aluno em "regular", "formado" ou "evadido". Além disso, os dados foram discretizados em conceitos, classificados como "MUITO BAIXO", "BAIXO", "MÉDIO", "ALTO" e "MUITO ALTO". O estudo investigou a possível relação entre a evasão escolar no curso de Sistemas de Informação e a distância entre a instituição e a moradia do estudante, utilizando a API do Google Maps para calcular a distância aproximada em quilômetros. Os resultados foram gerados por meio de árvores de decisão, destacando dados relacionados ao aluno e seu desempenho acadêmico como fatores importantes para a identificação da evasão escolar no ensino superior. A pesquisa aborda a relevância da investigação da evasão escolar no ensino superior, considerando seus impactos econômicos e na educação em geral.

De acordo com o estudo realizado por Ioanna Lykourentzou et al. (2009), o

artigo propõe um método de previsão de evasão em cursos de e-learning, baseado em três técnicas populares de aprendizado de máquina e dados detalhados dos alunos. As técnicas de aprendizado de máquina utilizadas incluem redes neurais feedforward, máquinas de vetores de suporte e ensemble probabilístico simplificado fuzzy ARTMAP. Dado que uma única técnica pode não ser capaz de classificar com precisão todos os estudantes de e-learning, enquanto outra pode obter sucesso, foram testados três esquemas de decisão que combinam de diferentes maneiras os resultados das três técnicas de aprendizado de máquina. O método foi avaliado com base em sua precisão geral, sensibilidade e precisão, e seus resultados demonstraram ser significativamente melhores do que os apresentados na literatura relevante. Segundo os autores, o método, apresentado no estudo, visa a previsão antecipada e precisa da evasão de estudantes em cursos de e-learning. Ele se fundamenta em registros detalhados dos alunos, obtidos a partir do Sistema de Gestão de Aprendizado que hospeda os cursos de e-learning, para realizar estimativas dinâmicas e adaptá-las ao progresso do aluno durante o curso. O uso das três técnicas de aprendizado de máquina mencionadas, aliado à combinação de seus resultados por meio de diferentes esquemas de decisão, tem como objetivo superar as limitações individuais dessas técnicas na identificação de estudantes em risco de evasão. Os resultados experimentais evidenciam a precisão deste método em comparação com outros estudos na área. Este método promete ser uma ferramenta valiosa para instrutores, permitindo-lhes identificar prontamente estudantes em risco e direcionar esforços para atender às necessidades específicas desses alunos, potencialmente aumentando as taxas de retenção nos cursos de e-learning. O futuro reserva possibilidades para a aplicação deste método em outros tipos de cursos, incluindo blended learning, educação a distância e ensino tradicional. Além disso, é necessário explorar como o desempenho do método pode ser aprimorado ao considerar diferentes atributos dos alunos. Outras técnicas e abordagens também podem ser investigadas, tanto em termos de desempenho individual quanto em combinação, para a previsão de evasão. Por fim, a incorporação deste método nas estratégias de retenção de instituições educacionais é uma perspectiva que pode contribuir para o aumento das taxas de retenção de alunos.

O estudo conduzido por Júlia e Hazra (2015) propõe um conjunto de sete variáveis destinadas a distinguir entre alunos concluintes e não concluintes em cursos que fazem uso de Ambientes Virtuais de Ensino e Aprendizagem (AVEA). Para verificar a eficácia dessas variáveis, os autores realizaram uma análise estatística comparativa dos dados de 1168 alunos, abrangendo 89 disciplinas em 5 cursos de uma instituição de ensino. Os resultados desta análise evidenciaram diferenças significativas no desempenho de alunos concluintes e não concluintes, sugerindo que as variáveis propostas podem ser empregadas com êxito na construção de modelos de previsão de alunos não concluintes. A metodologia adotada no estudo fez uso de técnicas estatísticas, incluindo o teste *t de Student,* para avaliar as disparidades entre os grupos de alunos concluintes e não concluintes. As variáveis de interesse foram selecionadas com base na literatura existente e nas características mais comuns dos AVEAs. Os dados foram coletados a partir das interações dos alunos nas plataformas de AVEA e dos registros acadêmicos da instituição. Para facilitar a análise, os valores das variáveis foram normalizados em uma escala de 0 a 1, garantindo uma comparação coerente do desempenho dos alunos. Uma limitação do estudo reside no fato de que apenas um número restrito de registros continha dados completos,

destacando a necessidade de pesquisas futuras que explorem técnicas de mineração de dados para lidar com a ausência de dados em contextos semelhantes. As conclusões desta pesquisa têm o potencial de identificar fatores críticos que podem auxiliar na previsão do desempenho dos alunos em cursos que fazem uso de AVEAs, sem depender unicamente de variáveis demográficas. Essa contribuição oferece insights valiosos para educadores e instituições de ensino, enriquecendo a compreensão dos fatores que impactam a conclusão de cursos nesse ambiente virtual.

O trabalho realizado por Martins, Tolledo, Machado, Baptista e Realinho (2021) tem como objetivo principal a redução do fracasso acadêmico no ensino superior por meio da aplicação de técnicas de aprendizado de máquina. O foco da pesquisa é identificar alunos em risco de insucesso em estágios iniciais de sua jornada acadêmica, possibilitando a implementação de estratégias de apoio adequadas.

Os pesquisadores utilizam um conjunto de dados proveniente de uma instituição de ensino superior para desenvolver modelos de classificação capazes de prever o desempenho acadêmico dos estudantes. Esse conjunto de dados inclui informações disponíveis no momento da matrícula dos alunos, tais como histórico acadêmico, dados demográficos e fatores socioeconômicos.

A abordagem do problema envolve uma tarefa de classificação de três categorias, caracterizada por um desequilíbrio significativo em relação a uma das classes. Para lidar com essa questão, os pesquisadores testam algoritmos de balanceamento de classes, como SMOTE e ADASYN, a fim de melhorar a representatividade das classes minoritárias.

No que diz respeito à modelagem, os autores exploram tanto métodos de aprendizado de máquina convencionais, como regressão logística, árvores de decisão e SVM, quanto algoritmos de boosting mais avançados, incluindo gradient boosting, extreme gradient boosting, logit boosting e catboosting.

Os resultados obtidos demonstram que os algoritmos de boosting respondem de maneira mais eficaz à tarefa de classificação específica em comparação aos métodos convencionais. No entanto, mesmo com o uso de algoritmos de ponta, ainda há desafios na correta identificação da maioria dos casos nas classes minoritárias. Como próximo passo, os pesquisadores planejam incluir informações sobre o desempenho dos alunos no primeiro ano, como notas acadêmicas dos primeiros semestres, para aprimorar ainda mais o desempenho dos modelos.

Dessa forma, o presente estudo visa aprimorar técnicas e modelos de aprendizado de máquina relacionados à evasão escolar, com base no trabalho prévio realizado por Martins, Tolledo, Machado, Baptista e Realinho (2021), intitulado "Early Prediction of student’s Performance in Higher Education: A Case Study".

3. Metodologia

Este projeto segue a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) e inicialmente replicou os passos da autora. Posteriormente, novos modelos de machine learning e técnicas de tratamento de dados foram aplicados para generalizar e aprimorar a precisão dos modelos.

Conforme definido por Chapman et al. (2000), o CRISP-DM é uma metodologia amplamente reconhecida para projetos de mineração de dados, composta

por seis fases: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação. Essas fases são aplicadas de maneira cíclica para identificar as melhores métricas e processos, fornecendo um guia passo a passo para profissionais conduzirem projetos de análise de dados de forma estruturada e eficiente.

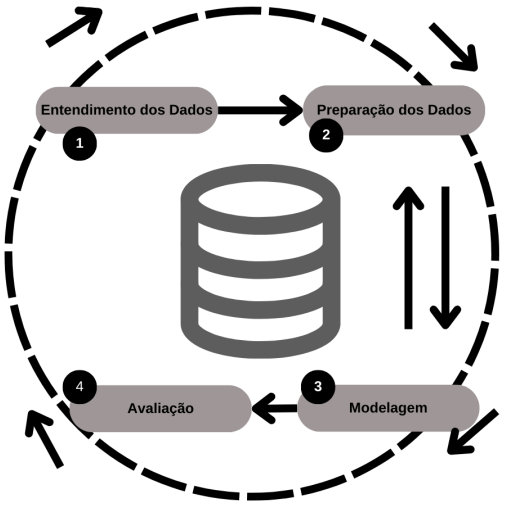

Figura 1. Processo CRISP-DM (O autor)

Este estudo segue uma abordagem simplificada do CRISP-DM, abrangendo compreensão de dados, preparação do modelo, modelagem de dados e avaliação dos resultados, com foco na importância da generalização de dados e modelos para adaptá-los a diversos cenários de evasão escolar.

Como destacado por Bousquet e Elisseeff (2002), a generalização é crucial em modelos de aprendizado de máquina para avaliar a capacidade de prever o desempenho em dados não observados. Portanto, a prioridade deste estudo foi promover a generalização do conjunto de dados, removendo variáveis regionais, visando melhorar a precisão na classificação.

3.1. Primeiro ciclo - Refazendo passos da autora

Conforme destacado por Young, Rose, Karnowski, Lim e Patton (2015), os hiperparâmetros em modelos de aprendizado de máquina exigem ajustes para adaptar o modelo a diferentes contextos. No presente estudo, a autora adotou técnicas de otimização de hiperparâmetros, incluindo o Grid Search, como detalhado por Petro e Pavlo Liashchynskyi (2019). O Grid Search é uma abordagem tradicional que explora minuciosamente um conjunto predefinido de hiperparâmetros para calibrar modelos como regressão logística, suport vector machine e árvore de decisão. Por outro lado, no caso da random forest, a autora optou por utilizar o Randomized Search CV, que, como explicado por Petro e Pavlo Liashchynskyi, realiza seleções aleatórias de combinações de parâmetros. Essa abordagem se mostrou particularmente vantajosa quando apenas alguns hiperparâmetros impactam significativamente o desempenho do modelo.

Após a otimização dos parâmetros, o modelo foi aplicado ao conjunto de dados e treinado, resultando na obtenção dos resultados do primeiro ciclo de avaliação.

3.2. Segundo ciclo - Alterações realizadas pelo autor

Com o objetivo de ampliar a capacidade de generalização do modelo, foi adotada a estratégia de eliminar colunas relacionadas à regionalidade, como a taxa de inflação do país, a nacionalidade dos alunos e seu status de aluno internacional, entre outras. Essa abordagem conferiu ao modelo uma perspectiva mais abrangente e flexível, reduzindo sua dependência de características geográficas específicas e expandindo o alcance da base de dados.

Posteriormente, os mesmos modelos previamente utilizados pela pesquisadora foram empregados com o intuito de aprimorar a eficácia e precisão na classificação dos perfis dos alunos. Uma decisão adicional foi tomada para simplificar o problema, que consistiu na remoção da classe minoritária da variável alvo que representava os alunos matriculados. Isso transformou o conjunto de dados em um problema de classificação binária em vez de multiclasse, facilitando a análise e interpretação dos resultados.

Além disso, houve a aplicação da normalização dos dados, conforme discutido por Dalwinder Singh e Birmohan Singh (2019). A normalização envolveu a escalonagem e transformação dos dados para igualar a contribuição de cada característica. A qualidade dos dados desempenha um papel crucial no sucesso dos algoritmos de aprendizado de máquina na construção de modelos preditivos para problemas de classificação. A importância da normalização de dados na melhoria da qualidade e, consequentemente, no aprimoramento do desempenho dos algoritmos de aprendizado de máquina tem sido objeto de análise em diversos estudos.

Optou-se também por não realizar o aumento dos dados minoritários, pois, após o tratamento de dados e a exclusão da classe minoritária, as classes remanescentes apresentaram uma variação adequada para a aplicação dos modelos de aprendizado de máquina.

Após a conclusão desses processos, os modelos de aprendizado de máquina foram aplicados aos dados tratados, resultando em melhorias nas métricas de desempenho, cujos resultados serão discutidos na seção 4.

4. Resultados e Análises

Após a realização de ajustes, os resultados obtidos nesta pesquisa alinharam-se de forma consistente com os da autora, particularmente no contexto de modelos tradicionais e técnicas de boosting. Isso resultou em métricas robustas e um desempenho significativamente melhorado. No entanto, ao aplicar transformações nos dados e efetuar a seleção de características, observou-se um desempenho excepcional dos modelos tradicionais, superando as abordagens da autora e obtendo uma classificação mais elevada na análise comparativa.

O estudo realizado destaca a importância crucial da generalização dos dados e da seleção de variáveis para aprimorar as métricas de desempenho dos modelos. As tabelas comparativas a seguir ilustram o desempenho dos modelos, replicando os processos delineados pela autora e apresentando os resultados após as modificações implementadas.

Tabela 1. Comparativo das métricas dos modelos analisados antes e pós tratamento e transformação de dados. - Clássicos

Comparativo de métricas dos modelos sem tratamento de dados e tratados - Clássicos				
Métricas	**Regressão Logística**	**SVM**	**Árvore de Decisão**	**Florestas Aleatórias**
F1-Score - Natural	62%	69%	67%	75%
F1-Score - Tratado	91%	91%	89%	90%
Diferença	29%	21%	22%	15%
Acurácia - Natural	67%	72%	66%	76%
Acurácia - Tratado	91%	91%	89%	90%
Diferença	24%	19%	23%	14%
Precisão - Natural	64%	71%	67%	75%
Precisão - Tratado	91%	91%	89%	91%
Diferença	27%	20%	22%	15%
Revocação - Natural	67%	72%	66%	76%
Revocação - Tratado	91%	91%	89%	90%
Diferença	24%	19%	23%	14%

Tabela 2. Comparativo das métricas dos modelos analisados antes e pós tratamento e transformação de dados. - Boosting

Comparativo de métricas dos modelos sem tratamento de dados e tratados - Boosting				
Métricas	**Gradient Boosting**	**Extreme Gradient Boosting**	**CatBoosting**	**LogitBoost**
F1-Score - Natural	75%	76%	75%	71%
F1-Score - Tratado	90%	91%	91%	91%
Diferença	15%	16%	16%	19%
Acurácia - Natural	76%	77%	76%	74%
Acurácia - Tratado	90%	91%	91%	91%
Diferença	14%	15%	15%	16%
Precisão - Natural	75%	76%	75%	71%

Precisão - Tratado	90%	92%	91%	91%
Diferença	15%	16%	16%	19%
Revocação - Natural	76%	77%	76%	74%
Revocação - Tratado	90%	91%	91%	91%
Diferença	14%	15%	15%	16%

5. Considerações Finais

A evasão escolar é um desafio premente que impacta de maneira profunda as instituições de ensino, repercutindo em esferas sociais, acadêmicas, econômicas e ambientais. Esses efeitos adversos permeiam as políticas de investimento e desenvolvimento educacional, destacando a melhoria da qualidade da educação como uma prioridade inquestionável.

O presente estudo foi empenhado em aprimorar e generalizar os dados, visando obter uma classificação mais precisa da evasão de alunos, baseando-se no conjunto de dados "Predict students' dropout and academic success" (2011). O uso estratégico da generalização dos modelos e da normalização dos dados, juntamente com a binarização do problema, resultou em melhorias notáveis, com algumas métricas apresentando ganhos expressivos, como no caso da Regressão Logística, que alcançou uma melhoria de até 30%.

Este estudo sublinha, portanto, a relevância da aplicação de técnicas de aprendizado de máquina na identificação e prevenção da evasão escolar. As melhorias alcançadas por meio da generalização dos modelos e do tratamento adequado dos dados indicam que essas abordagens têm um potencial substancial para contribuir não apenas para a retenção de alunos, mas também para aprimorar a qualidade geral da educação. Nesse sentido, a combinação de abordagens baseadas em dados e aprendizado de máquina abre perspectivas promissoras para abordar desafios complexos, como a evasão escolar, contribuindo para um futuro mais auspicioso no campo da educação.

A análise das métricas corrobora a importância da generalização dos modelos e a necessidade de um tratamento aprimorado dos dados para mitigar a alta dimensionalidade. A Tabela 3 realça os modelos que apresentaram as métricas mais favoráveis, evidenciando um notável aprimoramento no desempenho de nossos modelos.

Tabela 3. Comparativo das métricas dos modelos analisados antes e pós tratamento e transformação de dados. - Boosting

Melhores modelos e métricas		
Métricas	**Regressão Logística**	**Extreme Gradient Boosting**
F1-Score - Natural	62%	76%
F1-Score - Tratado	91%	91%
Diferença	29%	16%
Acurácia - Natural	67%	77%
Acurácia - Tratado	91%	91%
Diferença	24%	15%
Precisão - Natural	64%	76%
Precisão - Tratado	91%	92%
Diferença	27%	16%
Revocação - Natural	67%	77%
Revocação - Tratado	91%	91%
Diferença	24%	15%

Referências

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. **CRISP-DM 1.0: Step-by-step data mining guide.** SPSS Inc., 2000.

Realinho, Valentim, Vieira Martins, Mónica, Machado, Jorge, & Baptista, Luís. (2021). **Predict students' dropout and academic success.** UCI Machine Learning Repository. 2021

Bousquet, O., e Elisseeff, A. S**tability and Generalization.** Journal of Machine Learning Research, 2, 499-526. Submitted 7/01.. 2002

Hasan, H. M. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. **Machine Learning Algorithm for Student's Performance Prediction.** 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019

Singh, D., & Singh, B. **Investigating the impact of data normalization on classification performance.** Applied Soft Computing, 2019

TORRES MARQUES, L. .; TORRES MARQUES, B. .; MORAIS SILVA, C. A. . **Uma Abordagem de Descoberta de Conhecimento para Desvendar Causas da Evasão Escolar**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 20, n. 1, p. 284–294, 2022.

MARTINS, M. V. et al. **Early Prediction of student's Performance in Higher Education: A Case Study.** In: ROCHA, Á.; ADELI, H.; DZEMYDA, G.; MOREIRA,

F.; RAMALHO CORREIA, A. M. (Eds.). Trends and Applications in Information Systems and Technologies. In: WorldCIST 2021. Advances in Intelligent Systems and Computing, vol. 1365. Springer, Cham, 2021.

YOUNG, S. R. et al. **Optimizing deep learning hyperparameters through an evolutionary algorithm.** In: Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments - MLHPC '15, 2015.

NOETZOLD, E.; DE L. PERTILE, S. **Análise e predição de evasão dos alunos de um curso de graduação em sistemas de Informação por meio da mineração de dados educacionais.** Revista Novas Tecnologias na Educação, Porto Alegre, v. 19, n. 1, p. 351–360, 2021.

MARQUES CARVALHO DA SILVA, J.; IMRAN, H. **Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por Ambientes Virtuais de Ensino e Aprendizagem**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 13, n. 2, 2015

YKOUERNTZOU, Ioanna et al. **Dropout prediction in e-learning courses through the combination of machine learning techniques.** Computers & Education, [S.l.] Athens, Greece. 2009

BATISTA, M. R.; FAGUNDES, R. A. de A. **Mineração de dados educacionais aplicada a performance de estudantes: uma revisão sistemática da literatura**. Revista Novas Tecnologias na Educação, Porto Alegre, v. 21, n. 1, p. 271–280, 2023.

BANERJEE, Arindam. **Hyperparameter Tuning Using Randomized Search**. Disponível em: https://www.analyticsvidhya.com/blog/2022/11/hyperparameter-tuning-using-randomized-search/. Acesso em: 14/10/2023.

LIASHCHYNSKYI, Petro; LIASHCHYNSKYI, Pavlo. **Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS.** arXiv:1912.06059 (cs), 2019.