unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

GABRIEL MARTIN PHILOT

Data Science aplicado para obtenção de indicadores de desempenho em ambiente escolar remoto

Guaratinguetá - SP
2021

Gabriel Martin Philot

Data Science aplicado para obtenção de indicadores de desempenho em ambiente escolar remoto

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Física da Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Física.

Orientador: Prof. Dr. Marco Aurélio Alvarenga Monteiro

Guaratinguetá - SP
2021

Philot, Gabriel Martin
P568d Data Science aplicado para obtenção de indicadores de desempenho em ambiente escolar remoto / Gabriel Martin Philot – Guaratinguetá, 2021.
72 f.: il.
Bibliografia: f. 61-67

Trabalho de Graduação – Bacharelado em Física – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2021.
Orientador: Prof. Dr. Marco Aurélio Alvarenga Monteiro

1. Inteligência artificial. 2. Educação - Processamento de Dados.
3. Física - Estudo e ensino. 4. Gestão do conhecimento. I. Título.

CDU 371.39

Luciana Máximo

Bibliotecária CRB-8/3595

unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

GABRIEL MARTIN PHILOT

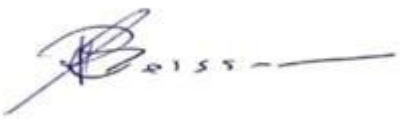
ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE "**GRADUADO EM FÍSICA**"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM FÍSICA

Prof. Dr. Julio M. Hoff da Silva
Coordenador

BANCA EXAMINADORA:

Prof. Dr. Marco Aurélio Alvarenga Monteiro
Orientador/UNESP-FEG



Prof. Dr. Rodolfo Meissner Rolando
UNESP-FEG

Prof. Dr. José Silvério Edmundo Germano
Membro Externo

Março/ 2021

DADOS CURRICULARES

GABRIEL MARTIN PHILOT

NASCIMENTO	25.05.1992 – São Paulo/ SP
FILIAÇÃO	George Guarany Philot Gisleine Martin Philot
2021	Graduação em Física – Bacharelado UNESP - Guaratinguetá

dedico este trabalho
de modo especial, à minha família

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, fonte da vida e da graça. Agradeço pela minha vida, minha inteligência, minha família e meus amigos;

ao meu orientador, *Prof. Dr. Marco Aurélio Alvarenga Monteiro* que jamais deixou de me incentivar. Sem a sua orientação, dedicação e auxílio, o estudo aqui apresentado seria praticamente impossível;

aos meus pais *George e Gisleine*, que apesar das dificuldades enfrentadas, sempre incentivaram meus estudos;

às funcionárias da Biblioteca do Campus de Guaratinguetá pela dedicação, presteza e principalmente pela vontade de ajudar;

aos funcionários da Faculdade de Engenharia do Campos de Guaratinguetá pela dedicação e alegria no atendimento.

"The measure of greatness in a scientific idea is the extent to which it stimulates thought and opens up new lines of research."

Paul A.M. Dirac

RESUMO

Existe um interesse crescente em utilizar *Data Science* (ciência de dados) em ambientes escolares obtendo inteligência a partir do seu grande volume de dados gerados. Em ambientes de educação a distância é essencial aos professores definirem quais alunos estão com baixo desempenho escolar. Neste trabalho, utilizando a metodologia de *Data Science* foram avaliados dados de atividades desempenhadas por alunos de uma escola integrante da plataforma de aprendizado contínuo da empresa Eduqo. Por regressão multilinear dos dados associada ao algoritmo de *machine learning "Train and Split*" foi obtido uma equação de modelo (R^2=0,69) e dois indicadores para previsão do desempenho dos alunos: média de nota de tarefa e média de nota de listas. A clusterização dos dados (*K-means*) a partir dos indicadores permitiu obter 4 grupos de desempenho dos alunos, dessa forma os professores conseguem verificar rapidamente quais alunos requerem acompanhamento para um melhor desempenho do aprendizado escolar em ambiente remoto.

PALAVRAS-CHAVE: Ciência de dados. Educação. EaD. *Machine Learning*. Regressão multilinear.

ABSTRACT

There has been growing interest in using Data Science in school environments obtaining intelligence from their large data volume generated. In e-learning environment it is essential for professors define which students have poor school performance. In this work, was used Data Science methodology to evaluate data from activities carried out by students from a school which is a member of a continuous learning platform of Eduqo's company. By multilinear regression of the data associated with a machine learning algorithm, Train and Split, were obtained a model equation (R^2=0,69) and two indicators for prevision of student's performance: average grade of task and average grade of lists. Data clustering (K-means) by indicators allowed obtain 4 groups of student performance, thus the professors can promptly verify which students require monitoring for better learning performance in a remote school environment.

KEYWORDS: Data science. Education. E-Learning. Machine learning. Multilinear regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AA	*Academic Analytic*
ABED	Associação Brasileira de Ensino à distância
CRISP	*Cross Industry Standard Process For Data Mining*
EaD	Educação à distância
EDM	*Educational Data Mining*
EM	Algoritmo de expectativa - maximização
EQM	Erro Quadrático Médio
GD	Gradiente Descendente
GPU	*Graphic processing units*
IA	Inteligência Artificial
IBM	*International Business Machines Corporation*
IE	Instituição de Ensino
IES	Instituição de Ensino Superior
ITS	*Intelligent Tutorial Systems*
LMS	*Learning Management Systems*
MAE	Erro Absoluto Médio
MSE	Média dos erros quadráticos
NLP	Processamento de linguagem natural
SQL	*Standard Query Language*
SQR	Soma dos quadrados dos resíduos
SSE	Soma Residual dos Quadrados
TI	Tecnologia da Informação
TICs	Tecnologia da Informação e da Comunicação
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura

LISTA DE SÍMBOLOS

α	Valor esperado de Y
β_1	Variação esperada de Y
β_k	Variação esperada de Y em um incremento unitário em X_k
e_i	Erro não explicado pelo modelo
H	Classe de hipótese
H_0	Hipótese nula
R^2	Coeficiente de determinação
ω_0	Ponto inicial da reta
ω_1	Inclinação da reta
x_1	Atributo de entrada
Y	Resíduo

SUMÁRIO

1 INTRODUÇÃO

A pandemia gerada pelo COVID-19 mudou abruptamente a rotina das escolas, segundo a UNESCO mais de 90% dos estudantes do mundo foi afetada com o fechamento de escolas e universidades (UNESCO, 2020). Para garantir a continuidade do ensino foi observado um grande número de escolas públicas e privadas mudarem para o modelo de educação à distância (EaD). Grande parte dessas escolas não estava preparada e nunca imaginaram que teriam que migrar para um modelo EaD, onde suas práticas escolares e pedagógicas teriam que ser totalmente revistas. Além disso, os professores não estavam tecnologicamente educados para adaptarem a maneira a qual transmitiam os seus conhecimentos e conteúdos aos alunos.

Sabe-se que a aprendizagem não ocorre da mesma forma para cada indivíduo, sendo necessária a intervenção do professor para entender as dificuldades de seus alunos e promover o aprendizado homogêneo. Dessa forma uma mudança repentina interação entre aluno e professor pode afetar o processo aprendizagem.

Vygotsky (1998) sugere que existem dois níveis de desenvolvimento de aprendizado: desenvolvimento real (ou atual) e desenvolvimento potencial. O desenvolvimento real se refere ao que o aluno já possui consolidado e sabe desenvolver sem o auxílio externo, sendo ligado a aprendizados prévios fora do ambiente escolar. Enquanto que o desenvolvimento potencial é referente ao que o aluno não sabe fazer ainda sem o auxílio de um indivíduo. Neste caso o professor é o indivíduo mediador que conduzirá o aluno por meio de conceitos, experiências e atividades ao domínio de conteúdo, resoluções de problemas e consequentemente à consolidação do aprendizado contínuo (GONÇALVES; JESUS, 2010).

O desafio da escola e principalmente do professor é justamente compreender as dificuldades de cada um de seus alunos e quais temas não ficaram claros, para que possam atuar na capacitação do aluno. Esse desafio se torna ainda maior em um ambiente de educação à distância, somado a uma falta de estrutura ou preparo dos educadores e alunos a esta dinâmica de ensino. A falta de preparo dos alunos a forma de educação à distância bem como fatores externos a escola como contexto de pandemia, contribuem para a queda do seu aprendizado (DIAS; PINTO, 2020).

Esse contexto traz a necessidade de indicadores efetivos aos professores para que possam acompanhar mediante as atividades feitas pelos alunos os pontos de maior dificuldade não só de uma turma inteira como também relacionado a cada aluno.

As Tecnologias da Informação e da Comunicação (TICs) são um conjunto de recursos tecnológicos utilizados para reunir e compartilhar informações (LOBO; MAIA, 2015). O modelo de educação à distância (EaD) utiliza esses recursos cada vez mais para o aprendizado on-line eficiente dos alunos (PAESE, 2012). A facilidade de acesso e a diversidade de maneiras de se aprender um mesmo conteúdo, são grandes vantagens das TICs e propiciam novas formas para o aprendizado do aluno (KENSKI, 2003). Há um crescente interesse em aliar a ciência de dados ao campo da educação devido a grande disponibilidade de dados gerados por escolas e docentes (LIU; HUANG, 2017).

O uso da ciência de dados permite o desenvolvimento de metodologias para coleta e análise de dados educacionais possibilitando tomadas de decisões por parte da escola e professores, para melhoria não só do ensino como também do engajamento dos alunos com o ambiente escolar ou virtual (GHAZARIAN; KWON, 2015)(DEJESUS, 2019)(MORRIS; FINNEGAN; WU, 2005).

Neste sentido alguns trabalhos são encontrados na literatura. Como o caso de Vettori e Zaro (2016), que estudaram o aplicativo *Socrative App* como uma ferramenta de ensino aliada a metodologia "*Peer instruction*". O aplicativo *Socrative App* a partir de questionários ou atividades dissertativas, desenvolvidas por professores, coleta dados de respostas dos alunos e permite verificar o desempenho instantâneo deles.

Os autores concluíram que o aplicativo contribuiu para o engajamento da turma analisada, bem como melhor entendimento do professor da situação da turma. Entretanto, o aplicativo por si só não consegue ser efetivo sem uma metodologia de ensino. Morris e Wu (2005) fizeram um estudo do engajamento de alunos de graduação de um curso online desenvolvido pelo sistema de Universidades da Geórgia (USG). Os autores analisaram o comportamento dos alunos frente a sua persistência, analisando alunos que concluíram ou não o curso, e participação, avaliando o tempo gasto e o número de materiais visualizados para o estudo de conteúdo e desenvolvimento de discussões.

Por meio de testes estatísticos foi obtida diferença significativa de participação entre alunos concluintes e não concluintes, ou seja, alunos que concluíram o curso realizavam mais discussões e gastavam mais tempo estudando os conteúdos. Para avaliar como a participação poderia prever a conquista de melhor desempenho, os autores fizeram uma análise de regressão multivariada obtendo que aproximadamente 31% (R^2=0,31) da variabilidade no desempenho se deve pelas variáveis de participação do aluno. As variáveis que mais conseguiram prever o desempenho do aluno neste estudo foram número de páginas de

conteúdo visualizadas, número de publicações de discussão vistas e o tempo gasto vendo essas publicações (MORRIS & WU, 2005).

Em grande maioria as escolas que iniciaram o modelo EaD tinham pouco ou nenhum preparo para um modelo integralmente digital, o que dificulta o acompanhamento do aprendizado de seus alunos. Independente das demais dificuldades encontradas pelas escolas, ainda não se tem certeza sobre a evolução e/ou sucesso desses modelos de ensino criados de modo forçado e repentino. Mesmo assim, essa manobra vai gerar mais dados para diversos estudos, bem como metodologias baseadas em ciência de dados para analisar o que acontece em modelos EaD para ensino médio.

Esse estudo tem como objetivo mostrar uma aplicação da metodologia de ciência de dados na área da educação. Para isso foi realizado um estudo com base nos dados, presentes na plataforma da empresa Eduqo, de atividades alunos do ensino fundamental ao médio de uma escola privada, aplicando a metodologia de data science para encontrar possíveis indicadores preditivos do desempenho dos alunos, antes que propriamente esses façam as provas. Dessa forma professores poderiam acompanhar o desempenho dos alunos e promover engajamento de todos de maneira eficiente.

2 EDUCAÇÃO À DISTÂNCIA

O ensino a distância (EaD), também chamado de educação a distância e *e-learning* (MOORE & KEARLSEY,2008) é uma variante de ensino que nos últimos anos vem alcançando maior espaço e visibilidade nas instituições de ensino básico e superior (IES). Essa modalidade é conceituada como um processo de ensino-aprendizagem na qual sua principal característica está na separação física e espacial entre alunos e professores e pela presença de tecnologias, que possibilitem a interação entre eles (TESTA E FREITAS, 2002).

No mundo educacional, a EaD já não é novidade, porém, na atualidade, a palavra que identifica cada vez mais esse processo é a "interação". As tecnologias de comunicação mais eficazes e fluentes permitem esse novo modo de educar. Já é possível falar até em interação em tempo real, como as videoconferências, por exemplo, em que pessoas espacialmente separadas assistem aulas de forma síncrona (TESTA E FREITAS, 2002).

Os primeiros registros de utilização da EaD foram identificados em 1728,em Boston, nos Estados Unidos, através de um curso por correspondência Os processos de formação à distância existem desde a Idade Antiga, já que Alexandre, o Grande, foi aluno de Aristóteles por correspondência (MATTA,2003). Porém, pode-se afirmar mesmo que em 2000, mais de 80 países, representantes de todos os continentes territoriais, já utilizam esse recurso em todos os níveis de ensino (LITTO E FORMIGA, 2009).

A história da educação a distância no Brasil começou em 1904, de acordo com a Associação Brasileira de Ensino a Distância (ABED), com uma matéria publicada no Jornal do Brasil, onde foi encontrado um anúncio nos classificados oferecendo curso de datilografia por correspondência (ABED, 2011). Desde então, muito se evoluiu no EaD. Entretanto, oficialmente, a educação a distância surgiu pelo Decreto nº 5.622 de 19 de dezembro de 2005, que posteriormente foi revogado. A sua atualização ocorreu pelo Decreto nº 9.057, de 25 de maio de 2017, vigente até a atualidade, que define, no seu primeiro artigo:

> Art. 1º Para os fins deste Decreto,considera-se educação a distância a modalidade educacional na qual a mediação didático-pedagógica nos processos de ensino e aprendizagem ocorra com a utilização de meios e tecnologias de informação e comunicação, com pessoal qualificado, com políticas de acesso, com acompanhamento e avaliação compatíveis, entre outros, e desenvolva atividades educativas por estudantes e profissionais da educação que estejam em lugares e tempos diversos. (BRASIL, 2017)

A EaD vem assumindo uma proporção considerável no mercado educacional, mas não se pode ignorar as peculiaridades que essa modalidade de educação apresenta como a implementação e gerenciamento em relação ao ensino tradicional. Uma das problemáticas se refere ao fato de ser um sistema suscetível à grande influência do meio, por ser aberto, especialmente quando é necessária a utilização da internet no processo de ensino-aprendizagem (FRANTZ E KING, 2000).

São considerados como fatores críticos de sucesso na EaD: a capacitação de pessoal, o envolvimento dos participantes do processo de aprendizagem (equipe, professores etc.), os estudantes, o plano pedagógico, a tecnologia escolhida e as parcerias estratégicas, como a terceirização (TESTA,2002).Acredita-se que se deve considerar ainda: a acessibilidade cultural (diferenças culturais dos estudantes), a flexibilidade estrutural do programa (horário e lugar de estudo flexíveis) e o suporte metacognitivo (monitoramento do progresso dos alunos) (JOIA E LIMA, 2007).

Os professores assumem papel e responsabilidades fundamentais para que o EaD funcione efetivamente, exercendo três funções básicas que são : organizacional (organizador do programa do curso), social (responsável por criar um ambiente agradável e amigável para a aprendizagem) e intelectual (como facilitador da educação, por meio de discussões instigantes) (BAPTISTA,2005).Os fatores que influenciam realmente na satisfação do aluno na EaD são o controle do aluno por parte do professor/tutor, as relações interpessoais, o entusiasmo e a interação do grupo (AH-FAHAD,2010).

Foram levantados seis fatores críticos, mais recentemente: estratégia pedagógica utilizada, aspecto administrativo das instituições de ensino (treinamento para alunos, funcionários e professores, entre outros), tecnologia empregada, avaliação contínua do desempenho dos estudantes e da plataforma de aprendizagem, suporte dado aos alunos (tanto tecnológico quanto pedagógico e administrativo) e *design*, ou aparência, dos programas virtuais utilizados (PURI, 2012).

3 DATA SCIENCE

3.1 DATA SCIENCE PARA EDUCAÇÃO

O crescimento exponencial da geração de dados na área educacional, por usuários, dispositivos, sistemas e tecnologias está cada dia mais associado à *Data Science*, ou ciência de dados, e apresentam-se como ricas oportunidades para análise, entendimento, modelagem e predição do grande volume de dados gerados na educação.

Assim como nas demais áreas afetadas pela *Data Science*, o campo educacional vem incorporando cenários dessas tecnologias, em virtude das diversas abordagens educacionais gerarem cada vez mais dados, demandando análises detalhadas e voltadas para um melhor planejamento e execução de ações na área da educação. A análise de dados educacionais, de uma maneira geral, representa uma área de pesquisa emergente em Informática em Educação para o desenvolvimento de métodos que exploram dados oriundos de ambientes educacionais e também administrativos com a finalidade de entender melhor os estudantes e os cenários em que eles aprendem (DANIEL, 2016). A análise de dados é realidade em muitas áreas do conhecimento, com motivações claras e reais de utilização e de estratégia de uso com a finalidade de descobrir relações não óbvias ou ainda não experimentadas, mas que residem de forma implícita nos volumosos bancos de dados (SILVA, PERES, BOSCARIOLI, 2016).

No campo educacional, o interesse em usar e analisar dados já foi despertado, porém de maneira segregada em três linhas de pesquisa: *Educational Data Mining*, *Learning Analytics* e *Academic Analytics*. Conceituando cada uma destas linhas, a Mineração de Dados Educacionais (do inglês *Educational Data Mining*, EDM) é uma área de pesquisa que utiliza as tarefas da Mineração de Dados como Análise Preditiva, Agrupamento e Associação de Dados aplicados a problemas de contexto educacional (ROMERO E VENTURA, 2007; SIEMENS E BAKER, 2012; COSTA et al.,2013).

A *Educational Data Mining* tem como objetivos fazer descobertas sobre o comportamento dos estudantes e o ambiente no qual a aprendizagem ocorre, fornecendo insumos para o professor ou aluno investigar eventuais padrões descobertos (ROMERO E VENTURA, 2007; ROMERO et al., 2016). Os trabalhos na área de *Educational Data Mining* incluem, basicamente, dados oriundos de Sistemas de Gerenciamento de Aprendizagem (do inglês *Learning Management Systems*, LMS), Sistemas Tutoriais Inteligentes (do inglês *Intelligent Tutorial Systems,* ITS), *e-learning*, repositórios de objetos de aprendizagem e aplicações web utilizadas na educação.

Learning Analytics foi definido como um "processo para a medição, coleta, análise e comunicação de dados sobre os alunos e os seus contextos, para fins de compreensão e otimização da aprendizagem nos ambientes em que esse processo ocorre" (SOUZA et al.,2016;SIEMENS et. al., 2012). É uma área de pesquisa que envolve o uso de ferramentas de análise de dados para avaliar processos de aprendizagem estabelecidos por educadores aos seus educandos.

Nessa mesma linha, qualquer tipo de estratégia de aprendizado, seja com uso de recursos de aprendizagem (objetos de aprendizagem) elaborados pelos educadores ou por uma equipe de TI, ao final espera-se que os resultados sejam analisados e, portanto, é nesse momento em que se inserem os estudos de *Learning Analytics* (PAPAMITSIOU e ECONOMIDES, 2014; MARTINEZ-MALDONADO et al.,2016; KNIGHT e LITTLETON,2016;QUIGLEY et al. 2017).

As ferramentas de análise de dados usadas no *Learning Analytics* são as mais amplas possíveis, incluindo as tarefas da Mineração de Dados. É nesse ponto que se inicia a geração de conflito conceitual entre a *Educational Data Mining* e *Learning Analytics*. Enquanto *Educational Data Mining* e *Learning Analytics* focam na aprendizagem do aluno, há também uma série de outras informações dos mesmos que saem do escopo do aprendizado, mas que podem estar intrínsecas ao aprendizado do aluno ou mesmo ser incorporadas de alguma maneira nas análises. Portanto, trata-se então dos dados acadêmicos (PAPAMITSIOU e ECONOMIDES, 2014; MARTINEZ-MALDONADO et al.,2016; KNIGHT e LITTLETON,2016;QUIGLEY et al. 2017).

A Análise de Dados Acadêmicos, tradução nossa à *Academic Analytics* ou AA, tem como foco o uso dos dados oriundos dos sistemas de informação da Instituição de Ensino (IE) para tentar entender os dados cadastrais dos alunos e outros que se relacionam com a vivência acadêmica do aluno na instituição (CAMPBELL e OBLINGER, 2007; BAEPLER e MURDOCH, 2010). Exemplo é o uso de dados demográficos, desempenho acadêmico, histórico escolar, censo da instituição, uso dos recursos computacionais, financeiro (para instituições particulares) e uma série de outros dados que podem implicar de alguma maneira no desempenho do aluno (CAMPELL, 2007).

No entanto, a análise deve ter um olhar mais amplo aos instrutores e alunos, mas também aos gestores. E nesse aspecto, pode-se considerar a *Academic Analytics* como apoio aos gestores, a partir do momento em que se usam as análises para avaliar, por exemplo, projetos pedagógicos, processos administrativos, uso dos recursos da biblioteca, entre outros (BAEPLER E MURDOCH, 2010).

Para aplicação da *Data Science*, deve-se tentar inicialmente responder algumas perguntas sobre o problema em questão, no caso deste trabalho quer se entender como está o engajamento dos alunos dentro da plataforma e verificar se é possível obter um indicador para ter o desempenho o aluno de forma indireta às provas da plataforma. Para guiar a escola e professores sobre como ajudar os alunos com dificuldades antes mesmo de estas serem visualizadas como resultados nas provas.

3.2 INTRODUÇÃO À DATA SCIENCE

Dhar (2013) define *Data science* (ciência de dados) como o estudo que envolve dados e o seu estudo sistemático baseado fundamentos da estatística, de forma a organizar e analisar esses dados. A ciência de dados se difere da estatística principalmente por envolver dados heterogêneos interdisciplinares como textos, imagens e vídeos que se interagem entre si por relações complexas (DHAR,2013).

O interesse em *Data science* se tornou popular quando a tecnologia conseguiu dar maior suporte a este campo dentro ciência da computação. Além disso, a ciência de dados é atrativa pela análise dados de forma flexível e prática por meio de linguagens como *Python* e Programação R que possuem fácil desenvolvimento e portabilidade (MILLER, 2015). Existe demanda crescente da ciência de dados, segundo Miller (2015) é estimado um aumento de 364.000 para 2.729.000 de posições para cientistas de dados até 2020 (MILLER; HUGHES, 2017). Para organizações, a *Data science* é empregada para transformar dados em valor seja na forma de faturamento, redução de custos, agilidade na tomada de decisões e melhora do serviço ao cliente (OLAVSRUD, 2019).

Os conceitos *Data mining* (mineração de dados) e *Machine learning* (aprendizagem de máquina) são essenciais e acompanham a *Data Science*. O *Data mining* é um passo da descoberta de conhecimento que compreende a análise de dados, geralmente em grande volume, para obter padrões consistentes, relações e agrupamentos podendo entender melhor os dados (DHAR, 2013).

Um algoritmo de aprendizagem é capaz de avaliar um conjunto de dados e extrair padrões para obtenção de inteligência. Esses algoritmos são moldados em ferramentas que compõem o *data mining* como agrupamentos, árvores de decisão e hipóteses (MIKUT; REISCHL, 2011). *Machine learning* por sua vez são ferramentas usadas para fazer predições e testes de hipóteses sobre os dados. Segundo Gutierrez (2015) o aprendizado de máquina é

uma combinação da ciência da computação, teoria da probabilidade, estatística e visualização de dados. O poder preditivo do *Machine learning* está ligado com a quantidade de dados, assim com o passar do tempo e coleta de mais dados mais acurado o algoritmo (LOHR, 2012).

Em relação à metodologia de *Data Science*, não existe um único caminho para o seu desenvolvimento, todavia os caminhos têm um comum objetivo de extrair informação sobre um conjunto de dados. Como representação abrangente, a metodologia da *Data Science* geralmente é dividida em cinco etapas (ROLLINS, 2015):

1º Etapa: Nesta etapa o pesquisador busca o embasamento para entender o problema em questão, visando reconhecer as variáveis necessárias para montagem do estudo.

2º Etapa: Após entendimento do problema, ocorre a coleta de dados das variáveis relevantes ao estudo. Um fator importante nesta etapa é a qualidade e/ou quantidade dos dados coletados, já que os modelos que avaliam esses dados são fundamentados com ferramentas estatísticas que requerem não somente dados representativos como também um volume de dados suficientes de amostra para estimativas de uma população.

3º Etapa: Nesta etapa é feita a preparação dos dados coletados. São feitos procedimentos como retirada de dados duplicados e verificação se há dados ausentes, para garantir a qualidade das análises.

4º Etapa: Com os dados prontos, estes são submetidos a adequação em diversos modelos, sejam eles estatísticos ou modelos obtidos por ferramentas de *Machine learning* para realização de análises.

5º Etapa: Para avaliar a qualidade do modelo para os dados, são feitos diversas análises e testes. Nesta etapa, caso seja necessário, são realizados ajustes nos modelos para se adequarem melhor aos dados. Após essa etapa feita com sucesso é possível obter inteligência ou conhecimentos sobre os dados coletados.

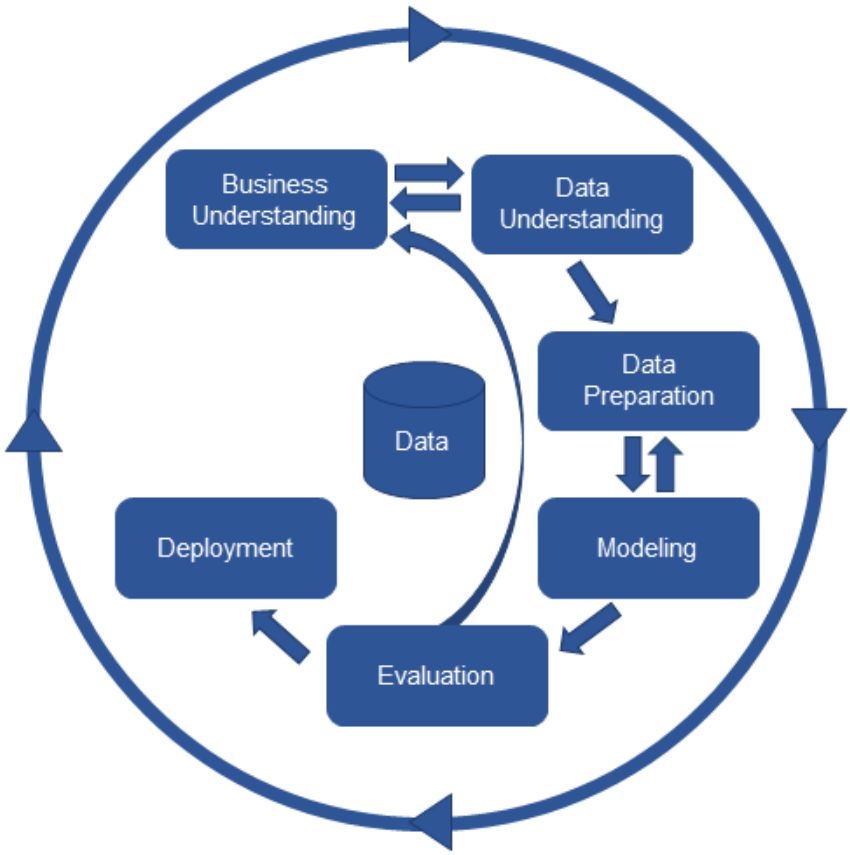
As etapas descritas acima são baseadas na famosa Metodologia CRISP DM. Como dito, o *Data Mining* faz parte da *Data Science*, que utiliza estatística e matemática como base para cruzamento de dados, por meio de técnicas de indução para propor hipóteses e solucionar questões. Nesse sentido, a Metodologia CRISP DM reúne as melhores práticas para que o *Data Mining* seja o mais produtivo e eficiente possível, analisando dados financeiros, de recursos humanos, produção, hábitos dos clientes e outros, para propor modelos de melhoria ou solução de problemas.

CRISP DM é a abreviação de *Cross Industry Standard Process for Data Mining* e surgiu justamente para atender aos projetos que estão diretamente envolvidos com o processamento e a análise de um grande volume de dados.

As vantagens em utilizar a metodologia CRISP DM são a melhoria no relacionamento com o cliente, entendendo seus hábitos de consumo e necessidades, atraindo e fidelizando a clientela; orientação na tomada de decisões, já que a mineração de dados alinhada à análise preditiva é uma grande ajuda no gerenciamento e controle de riscos, dessa forma decisões mais inteligentes podem ser tomadas; implementação de novos modelos de resolução de problemas; e o fornecimento de análises em tempo real, depois de implantado o ciclo do CRISP DM, é possível ter análises em tempo real conforme a situação e o cenário vão mudando, possibilitando mudanças imediatas e personalizadas para cada momento. A agilidade na tomada de decisões e a resolução de problemas, com certeza, são vantagens competitivas importantes.

É válido ressaltar que reajustes devem e são feitos ao modelo conforme mais dados são coletados, para se obter um modelo mais aprimorado que permita fazer análises representativas. Este caráter de flexibilidade da metodologia de *Data Science* a torna muito atraente principalmente em campos com um volume de dados variáveis como os da educação e finanças, que sofrem constantes alterações de objetivos de análise.

Figura 1- Fases do modelo CRISP-DM



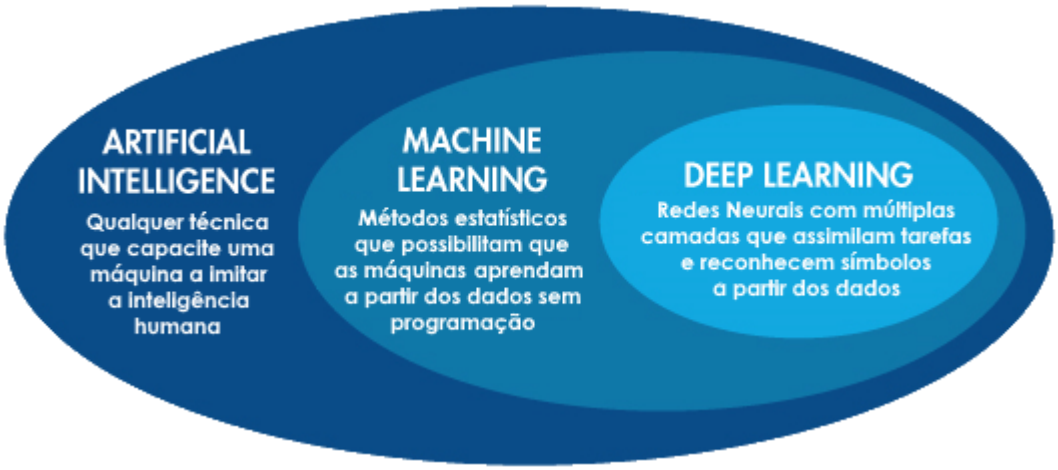
Fonte: Gonzalez (2019).

3.3 MACHINE LEARNING

3.3.1 O que é o Machine Learning?

Nos últimos anos a popularidade da Inteligência Artificial explodiu, especialmente desde 2015. Muito disso tem a ver com a disponibilidade dos GPUs (*graphic processing units*) que fazem com que processamento paralelo seja mais rápido, mais barato e mais poderoso. O *Big Data* também influencia diretamente neste processamento com a maciça quantidade de dados virtuais como imagens, transações, dados de mapas, textos e outros (SAS,2020).

Figura 2: Inteligência Artificial e seus subgrupos



Fonte: Lavagnoli (2019).

Inteligência artificial é o termo amplamente usado para abranger todas as máquinas, ferramentas, métodos e dispositivos que realizam atividades inteligentes, como o próprio nome já indica. Isso significa que tudo aquilo que é dotado de IA tem a capacidade de raciocinar, solucionar problemas, planejar, manipular objetos ou reconhecer vozes, rostos e imagens (NEUROTECH, 2020).

A inteligência artificial é uma subárea da ciência da computação direcionada ao desenvolvimento de computadores que são capazes de atuar em situações que costumeiramente são realizadas por pessoas. Ou, de maneira mais geral, capazes de resolver problemas complexos, nos quais hoje a melhor referência de solução é a humana. As máquinas executam funções cognitivas — associadas ao aprendizado humano e ao raciocínio para decidir — por meio de algoritmos inovadores, simulando a forma humana de assimilar informações, generalizar sobre casos passados e tomar decisões em situações futuras (NEUROTECH, 2020).

Com o advento da Inteligência artificial, que é a ciência de reproduzir capacidades humanas, o *Machine Learning* vem como um subconjunto específico da Inteligência Artificial que treina uma máquina para aprender com dados. O M.L é capaz de correlacionar um volume imenso de informações para extrair conhecimento e tomar decisões. Não é necessário programar o computador para procurar por informações específicas, ele aprende de maneira "autodidata". Isso significa que em vez de programar normas do tipo "se-então", os desenvolvedores apresentam exemplos de situações passadas e informam o que aquela

informação representa. Pode ser a ocorrência de uma fraude, uma tentativa de invasão a um sistema, uma compra realizada ou uma imagem. A partir da alimentação de um número de casos, o computador "descobre" as características dos casos e generaliza para fornecer respostas corretas em situações futuras (NEUROTECH, 2020).

O *Deep learning* é uma subárea do *Machine Learning*. São métodos evoluídos das redes neurais. As redes neurais se inspiram no cérebro humano e suas propriedades, como a capacidade de aprender e generalizar. No *Deep Learning*, as redes neurais são mais sofisticadas, de tamanho muito maior e com habilidade superior de extrair e representar informações. Antes, a análise de imagens, de vídeos, o reconhecimento de voz e o processamento de textos eram muito inferiores à capacidade humana. Hoje, os resultados são surpreendentes (NEUROTECH, 2020).

O *Machine Learning* é um método de análise de dados da área da Inteligência Artificial que automatiza a criação de modelos analíticos**.** Por meio de algoritmos que aprendem a partir de diversas bases de dados e de experiências acumuladas, ele possibilita a predição e o aprendizado de certos padrões e comportamentos automaticamente, sem a intervenção humana (OLIVEIRA, 2020).

A ideia básica do *Machine Learning* é permitir que máquinas desenvolvam modelos e façam predições sem a necessidade de que sejam reprogramadas para isso. E conforme vão sendo expostas a novos dados, elas vão aprendendo mais e se adaptando de forma independente, para fornecerem resultados mais precisos; gerarem resultados mais confiáveis e reproduzíveis, contribuírem para melhores tomadas de decisão e a realização de ações inteligentes em tempo real e ganho em velocidade de produção (OLIVEIRA, 2020).

O *Machine Learning* permite desenvolver algoritmos capazes de realizar automaticamente tarefas que poderiam consumir muitas horas de trabalho (e raciocínio lógico) de seres humanos.

E como as máquinas são ensinadas? O ensino se dá através de exemplos, usados como referências, que podem vir de bases de dados de fraudes, compras, interações com *call centers*, seguros contratados ou de imagens, vídeos, registros de voz, dados de dispositivos, textos em documentos e que representam situações do dia-a-dia.

Basicamente o ensino das máquinas se constitui como um processo de inserção de dados. O programador insere dados com atributos ou recursos diferentes que os algoritmos irão compreender e fornecer um limite de decisão com base nos dados fornecidos. Quando o algoritmo entender e interpretar os dados, ou seja, ele estará treinado por si mesmo.

Posteriormente, o programador poderá utilizar o algoritmo em uma fase de teste e sem uma programação explícita, irá inserir um ponto de dados de teste e esperar os resultados. Por fim, com a máquina ensinada por meio de exemplos, inserção de dados e repetições, ela estará ajustada para compreender os dados e absorver a aprendizagem dos mesmos, predizendo soluções futuras.

O enorme volume e variedade de dados disponíveis em um mundo hiperconectado, o atual poder de processamento computacional e as novas e mais acessíveis soluções de armazenamento de informações abrir todo um novo leque de possibilidades para as empresas. Tornou-se viável analisar dados complexos em grande escala e, principal mente, aprender com eles. É possível criar algoritmos capazes de realizar previsões de cenários muito mais assertivas do que as feitas por seres humanos, e em uma velocidade que nunca conseguiria alcançar. Além de uma forma de aumentar a produtividade da equipe, o uso de aprendizagem de máquina também permite substituir processos por soluções mais eficazes e obter uma redução de custos utilizando o *Machine Learning* (JUST DIGITAL,2017).

Existem dois tipos de aprendizado**:** o supervisionado e o não supervisionado**.** O supervisionado é responsável pela maior parte do *Machine Learning*. Neste tipo de aprendizado, por meio de um conjunto de exemplos, os algoritmos aprendem um modelo para poderem prover a variável de interesse, baseando-se em variáveis dependentes. Esse tipo de aprendizado envolve a participação de um ‘agente externo’, e é geralmente utilizado em aplicações nas quais os dados históricos preveem possíveis acontecimentos futuros (OLIVEIRA, 2020).

Atualmente, o *Machine Learning* é uma das áreas de aplicação mais promissoras no campo da Tecnologia da Informação onde seu escopo de aplicação é quase ilimitado. A aplicação do *Machine learning* em uma área educacional é muito interessante para pesquisadores e cientistas (SAS, 2020).

O objetivo do aprendizado de máquina é programar computadores para usar dados de exemplo ou experiências anteriores para resolver um determinado problema. Reconhecimento de padrão, educação, visão computacional, bioinformática, processamento de linguagem natural, etc. são apenas alguns dos campos onde o *machine learning* pode ser aplicado (SAS, 2020).

Existem muitas outras implementações empresariais de aprendizado de máquina na área de educação como: prever o desempenho do aluno e aprender sobre cada aluno, já que com o modelo de aprendizado de máquina pode-se descobrir pontos fracos e sugerir maneiras de

melhorar, como aulas adicionais ou estudar literatura adicional; *Test Students & Grade Students Fairly* (o aprendizado de máquina pode ajudar a criar avaliações adaptativas computadorizadas (SAS, 2020).

A avaliação baseada em aprendizado de máquina fornece *feedback* constante para professores e alunos sobre como o aluno aprende, o suporte de que precisa e o progresso que estão fazendo em relação aos seus objetivos de aprendizagem);melhorar a retenção (aprendizado de máquina, como análise de aprendizagem, também ajudará a melhorar as taxas de retenção com a identificação de alunos "em risco", as escolas podem chegar até esses alunos e obter a ajuda de que precisam diretamente; apoiar professores e a instituição (algoritmos baseados em aprendizado de máquina podem ajudar na classificação de papéis de avaliação manuscritos dos alunos) (SAS,2020).

Com o auxílio das análises de dados através do aprendizado de máquina é possível aumentar a eficiência, já que o machine learning na forma de inteligência artificial tem o potencial de promover uma melhor eficiência aos educados, concluindo tarefas como gerenciamento de sala de aula, programação,etc. Os educadores podem se concentrar em tarefas que não podem ser realizados pela inteligência artificial e que exigem ação humana (WU, HSIAO & NIAN,2018). Por sua vez, os educadores são livres para se concentrar em tarefas que não podem ser realizadas pela IA e que exigem um ser humano tocar. Já o aprendizado de máquina como análise de aprendizagem poder ajudar os professores a obterem informações sobre dados que não podem ser coletados humanamente.

Dessa forma, os computadores podem trabalhar com profundidade nos dados, filtrando peças de conteúdo, fazendo conexões e conclusões que impactam positivamente o processo de ensino – aprendizagem. É possível também uma análise preditiva que gera conclusões sobre acontecimentos futuros. Por exemplo, usando um conjunto de dados de registros cumulativos de alunos do ensino médio, a análise preditiva pode nos dizer quais são os mais propensos a desistir devido ao fracasso acadêmico ou até mesmo sua pontuação prevista em um exame padronizado (ĐAMBIC, KRAJCAR, & BELE, 2016; DELEN, 2010; LYKOYRENTZOU et al.,2009; RAM, WANG, CURRIM,F. & CURRIM, S.,2015; CHAI & GIBSON,2015; JIA, & MAREBOYANA,2014).

O aprendizado adaptativo pode ser usado para remediar alunos com dificuldade ou desafiar alunos superdotados. A aprendizagem adaptativa é um sistema educacional online ou baseado em tecnologia que analisa o desempenho do aluno em tempo real e modifica os métodos de ensino e o currículo com base nesses dados. O aprendizado de máquina na forma

de aprendizado personalizado pode ser usado para dar a cada aluno uma experiência educacional individualizada (ANAND et al.,2018;ALAM et al.,2018; CIOLACU et al.,2017).

A aprendizagem personalizada é um modelo educacional em que os alunos orientam sua própria aprendizagem, seguindo seu próprio ritmo e, em alguns casos, tomando suas próprias decisões sobre o que aprender. O ideal é que, em uma sala de aula com aprendizagem personalizada, os alunos escolham o que lhes interessa e os professores ajustam o currículo e os padrões aos interesses dos alunos. Além disso, o aprendizado de máquina na forma de inteligência artificial pode ser usado para avaliar as tarefas e exames dos alunos com mais precisão do que um ser humano. Pode requerer alguma contribuição de um ser humano, mas os resultados terão maior validade e confiabilidade (ANAND et al.,2018;ALAM et al.,2018; CIOLACU et al.,2017).

Quando se desenvolve um sistema de aprendizado de máquina, a estrutura utilizada na programação é diferente da programação de *software* tradicional. No método tradicional se cria um conjunto de regras para gerar uma resposta a partir do processamento dos dados introduzidos. Já os algoritmos de *Machine Learning* são criados a partir dos dados que serão analisados e as respostas/resultados que se esperam dessa análise. No final do processo o sistema cria as próprias regras ou perguntas (MILLER,2015).

A tecnologia *Machine Learning* permite que os modelos sejam treinados em conjuntos de dados antes de serem implantados. Um aplicativo ou *software* com *Machine Learning* é um tipo de programa que melhora automaticamente e gradualmente com o número de experiências em que ele é colocado para treinar. Nessa primeira etapa o treinamento é assistido. O processo iterativo leva a uma melhoria nos tipos de associações feitas entre elementos e dados, os quais são apresentados em uma grande quantidade. Devido a essa grande quantidade de dados que serão analisados, os padrões e associações feitas somente por observação humana poderiam ser ineficientes, caso sejam feitas sem um suporte das tecnologias *Machine Learning* (SIMON& GIROLAMI, 2013).

O aprendizado de máquina usa uma variedade de algoritmos que iterativamente aprenda com os dados para melhorar, descrever dados e prever resultados. Conforme os algoritmos ingerem dados de treinamento, é possível produzir modelos mais precisos com base nesses dados (SAS, 2020).

Após o treinamento inicial de um aplicativo ou *software* de *Machine Learning* ele poderá ser usado em tempo real para aprender sozinho com os dados apresentando maior precisão nos resultados com o passar do tempo (SIMON& GIROLAMI, 2013).

O objetivo dos processamentos de aprendizado de máquina é como a derivação de modelos preditivos a partir de dados atuais e históricos. Dessa forma, a tendência é que o algoritmo aumente a precisão e acurácia à medida que ocorram mais iterações (SIMON& GIROLAMI, 2013).

Os algoritmos de aprendizado de máquina precisam considerar todos os recursos em um campo de jogo uniforme. Isso significa que os valores de todos os recursos devem ser transformados na mesma escala. O processo de transformar recursos numéricos para usar a mesma escala é conhecido como dimensionamento de recursos. É uma etapa importante de pré-processamento de dados para a maioria dos algoritmos de aprendizado de máquina baseados em distância, porque pode ter um impacto significativo no desempenho do algoritmo (ARVAI, 2020).

Com o uso de algoritmos de aprendizagem de máquina pode-se alcançar resultados extremamente eficientes para domínios muito restritos usando modelos treinados a partir de grandes conjuntos de dados. Com o uso de ferramentas de *machine learning* podem ser detectados certos padrões. Estes padrões permitem mapear o processo e realizar predições de forma que se assume que o futuro não será muito diferente do passado onde o dado dessa amostra foi coletado, e assim é possível esperar que as predições futuras estejam corretas (ALPAYDIN, 2014).

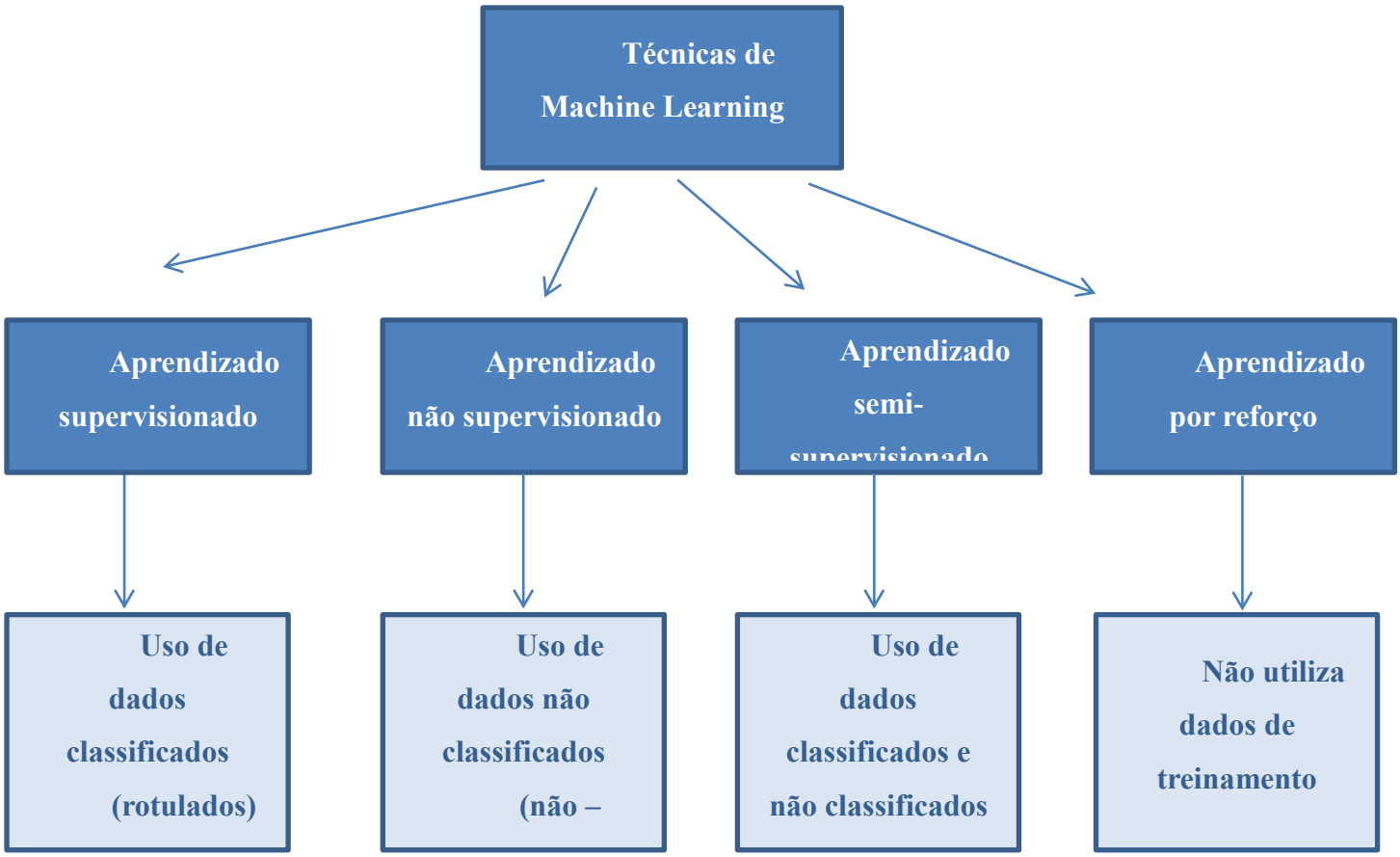
3.3.2 Técnicas de Machine Learning

Para que haja um processo de aprendizagem da máquina é necessário utilizar um conjunto de dados. O *Big Data* permite que os dados sejam virtualizados e então armazenados de maneira mais eficiente e econômica, seja *on premises* ou na *cloud*. Além da eficiência o *Big Data* é ferramenta auxiliar na melhoria da velocidade e confiabilidade da rede, removendo outras limitações físicas associadas ao gerenciamento de dados em grande quantidade (SHALEV-SCHWARTZ e BEN-DAVID, 2014).

No entanto, deve-se notar que, em contraste com a Inteligência Artificial tradicional, o aprendizado de máquina não tem por objetivo único automatizar uma tarefa que o ser humano faria de forma repetitiva ou demorada, mas usar os pontos fortes e habilidades especiais dos computadores para complementar a inteligência humana. Por exemplo, a capacidade de digitalizar e processar enormes bancos de dados permite que programas de aprendizado de máquina detectem padrões que estão fora do escopo da percepção humana (SHALEV-SCHWARTZ e BEN-DAVID, 2014).

De forma geral, há quatro principais técnicas de modelagem de machine learning: Aprendizado supervisionado, Aprendizado não supervisionado, Aprendizado semisupervisionado e Aprendizado por reforço. Salientando que no trabalho aqui descrito, a modelagem utilizada foi o aprendizado supervisionado.

Organograma 1 - Técnicas de Machine Learning por categorias



Fonte: Adaptado de Mohammed,Khan e Bashier, (2017).

A seleção do algoritmo de *machine learning* que melhor se adequa a aplicação pretendida de aprendizado de máquina para alcançar melhores resultados, às vezes, pode ser a parte mais difícil. A decisão de qual algoritmo usar pode ser orientada respondendo a perguntas importantes, como: Qual é o tamanho e a natureza dos dados? O que você deseja atingir com o modelo? Quão preciso o modelo precisa ser? Quanto tempo há disponível para treinar o modelo? Quão interpretável e compressível o modelo precisar ser? (WUJEK,HALLe GUNES,2016).

A maior diferença entre o aprendizado de máquina supervisionado e não supervisionado é o fato dos algoritmos de aprendizado de máquina supervisionados serem treinados em

conjuntos de dados rotulados que orientam o algoritmo a entender quais recursos são importantes para o problema em questão Por outro lado, os não supervisionados, são treinados em dados não rotulados e devem determinar a importância do recurso por conta própria, com base nos padrões inerentes à amostra (CASTLE,2018).

O Aprendizado supervisionado é utilizado, normalmente, para a predição de eventos. Na aprendizagem supervisionada, o objetivo é inferir uma função ou mapeamento a partir de dados de treinamento. Os dados de treinamento consistem no vetor de entrada X e no vetor de saída Y de rótulos. Um rótulo do vetor Y é a explicação de seus respectivos dados de entrada. Em outras palavras, ao utilizar o aprendizado supervisionado, obtém-se conhecimento prévio de quais devem ser os valores de saída para nossas amostras, podendo dados incorretos interferir na eficácia do modelo (MOHAMMED, KHAN e BASHIER, 2017).

O aprendizado supervisionado geralmente é realizado no contexto de classificação, quando se quer mapear a entrada para os rótulos de saída. Tanto na regressão quanto na classificação, o objetivo é encontrar relacionamentos ou estruturas específicas nos dados de entrada que nos permitam produzir efetivamente dados de saída corretos. O objetivo do aprendizado supervisionado é aprender uma função que, dada uma amostra de dados e resultados desejados, se aproxima melhor da relação entre entrada e saída observável nos dados. Ao conduzir o aprendizado supervisionado, as principais considerações a serem feitas são em relação à complexidade do modelo e o *tradeoff* de viés e variância (SONI, 2018).

Como já dito o *machine learning* supervisionado baseia-se na ideia de aprender com o exemplo. O algoritmo é alimentado com dados que se relacionam com o domínio do problema e metadados que atribuem rótulos (ou *labels*) aos dados. O processo de atribuição de rótulos aos dados é chamado de "*labeling*", e desempenha um papel crucial na obtenção de bons resultados no *machine learning* supervisionado (MANDIC, 2019).

No aprendizado de máquina, a rotulagem de dados é o processo de identificar dados brutos (imagens, arquivos de texto, vídeos etc.) e adicionar um ou mais rótulos informativos e significativos para fornecer contexto para que um modelo de aprendizado de máquina possa aprender com eles. A rotulagem de dados é necessária para uma variedade de casos de uso, incluindo visão computacional, processamento de linguagem natural e reconhecimento de fala (MANDIC, 2019).

Hoje, a maioria dos modelos práticos de aprendizado de máquina utiliza aprendizado supervisionado, que aplica um algoritmo para mapear uma entrada para uma saída. Para que o aprendizado supervisionado funcione, é necessário um conjunto rotulado de dados com os quais o modelo pode aprender para tomar decisões corretas. A rotulagem de dados

normalmente começa pedindo aos humanos que façam julgamentos sobre um dado não rotulado (MANDIC,2019).

O modelo de aprendizado de máquina usa rótulos fornecidos por humanos para aprender os padrões subjacentes em um processo chamado "treinamento de modelo". O resultado é um modelo treinado que pode ser usado para fazer previsões sobre novos dados.

No aprendizado de máquina, um conjunto de dados devidamente rotulado que usa como o padrão como objetivo para treinar e avaliar um determinado modelo costuma ser chamado de "verdade básica". A precisão do modelo treinado dependerá da precisão de sua verdade básica, portanto, é essencial gastar tempo e recursos para garantir uma rotulagem de dados altamente precisa (MANDIC, 2019).

Para este trabalho utilizou-se algoritmos como a regressão linear múltipla e o *K-means cluster* no aprendizado da máquina.

3.4 FUNDAMENTOS DOS MODELOS UTILIZADOS

3.4.1 Regressão linear

Quando é falado em *Machine Learning*, está se falando também de estatística. Isso porque o Aprendizado de Máquina só pôde ser criado graças à ampla variedade de técnicas estatísticas desenvolvidas nos últimos tempos. E uma das técnicas estatísticas mais utilizadas no aprendizado de máquina é a regressão linear (OLIVEIRA, 2020).

Existem dois tipos de aprendizado**:** o supervisionado e o não supervisionado. O supervisionado é responsável pela maior parte do *Machine Learning*. Neste tipo de aprendizado, por meio de um conjunto de exemplos, os algoritmos aprendem um modelo para poderem prover a variável de interesse, baseando-se em variáveis dependentes. Esse tipo de aprendizado envolve a participação de um ‘agente externo’, e é geralmente utilizado em aplicações nas quais os dados históricos preveem possíveis acontecimentos futuros (OLIVEIRA,2020).

A regressão linear é uma abordagem que está entre as mais simples. Foi desenvolvida no campo da estatística e é tratada como um modelo útil para entender o relacionamento entre valores numéricos de entrada e saída. O *Machine Learning* a "tomou emprestado" e, assim, a regressão linear se tornou uma das ferramentas mais úteis neste campo. Torna-se uma abordagem útil sempre que se deseja estimar uma resposta quantitativa. Trata-se de um método largamente utilizado e é importante que se tenha um bom entendimento em torno dos princípios da regressão linear, antes de partir para metodologias de aprendizado mais

complexas, as quais costumam ser vistas como generalizações ou extensões da mesma (OLIVEIRA,2020).

Algumas diferenças entre o *machine learning* e a regressão linear devem ser destacadas já que seu uso deve ser feito após uma cautelosa análise. O método estatístico de regressão apenas considera um relacionamento linear, é sensível a valores discrepantes, toma como base a média da variável dependente e assume que os dados são independentes. Ao contrário do *machine learning*, que pode modificar seu comportamento autonomamente tendo como base a sua própria experiência e tem uma interferência humana mínima, a regressão linear deve ser manipulada humanamente. Esse viés impõe a possibilidade de erros no processamento da técnica de regressão, enquanto no *machine learning* a eficiência da máquina se destaca em minimizar erros e diminuir o risco de indisponibilidade da aplicação por falha humana.

A regressão linear é feita manualmente, enquanto o *machine learning* usa a inteligência artificial como aplicação do método. Além disso, a aplicação dos métodos matemáticos da regressão linear é algo complexo e demorado e com a ascensão dos modelos de aprendizagem de máquina o tempo destes processamentos algorítmicos diminuiu, já que o crescente volume de dados e variedade de dados disponíveis impossibilita o uso apenas dos cálculos matemáticos por meio de regressão linear.

Diferente da técnica de regressão linear que é melhor trabalhada com dados mais simples e não escalados, o *machine learning* tem uma boa capacidade de *data preparation*, utiliza processos automatizados e iterativos, trabalha com escalabilidade e a modelagem é conjunta.

Existem dois tipos de regressão linear: simples e a múltipla: a regressão linear simples refere-se quando temos somente uma variável independente (X) para fazer a predição. Já a regressão linear múltipla refere-se a várias variáveis independentes (X) usadas para fazer a predição (DAMASCENO, 2020).

O algoritmo de regressão linear simples pode ser usado em variados problemas desde que as variáveis de entrada e saída sejam contínuas. Algumas das suas aplicações são em previsões de venda de produtos, valores de imóveis no setor imobiliário, cálculo de expectativa de vida em um país, entre outros (DAMASCENO,2020).

A regressão linear simples deve ser aplicada quando há uma boa correlação linear (positiva ou negativa) entre os dados, ou seja, quando o relacionamento ou associação entre os dados pode ser definido com uma reta. Quando se trata de correlação linear definimos como uma medida estatística usada no cálculo da associação entre os pontos. Neste caso, a Correlação de Pearson é indicada (DAMASCENO,2020).

O objetivo da regressão linear é encontrar uma reta que consiga definir bem os dados e minimizar a diferença entre o valor real e a saída calculada pelo modelo. A função que representa a regressão linear é dada a seguir (DAMASCENO,2020):

$$f(x)= w_0 + w_1 * x_1 \quad \text{(I)}$$

Onde w_0 (representa o ponto inicial da reta) e w_1 (representa a inclinação da reta, ou seja, o quanto que essa variável cresce conforme o tempo passa) são variáveis que o algoritmo calcula para poder definir a reta, e x_1 seria o atributo de entrada que foi dada ao modelo. E com esses valores é possível fazer as previsões (DAMASCENO,2020).

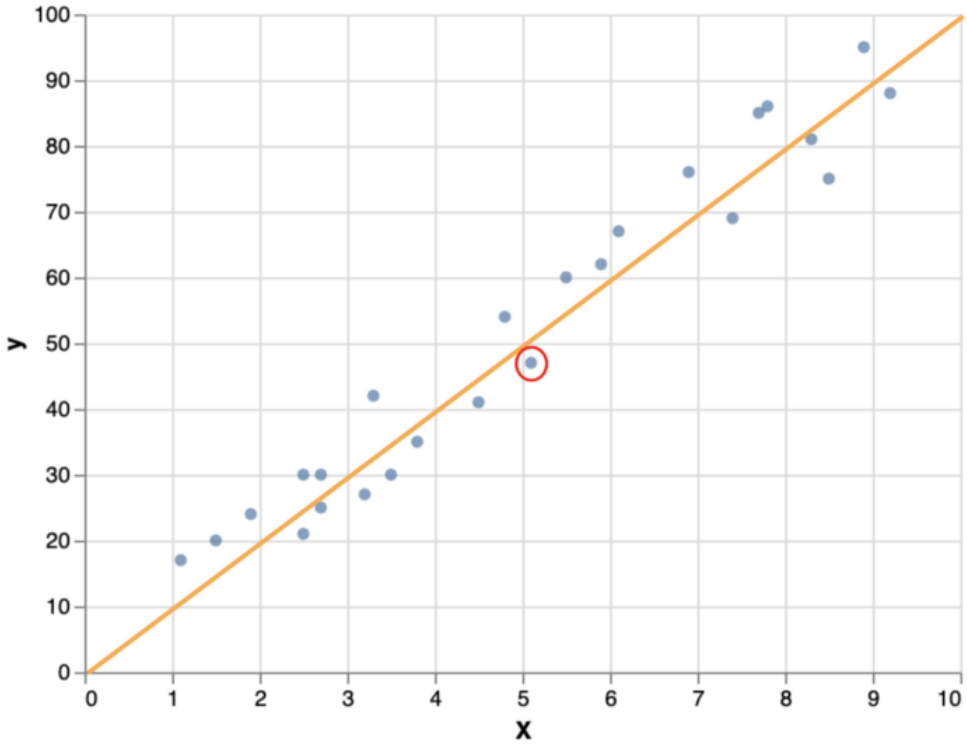
Por exemplo, é suposto que o algoritmo calculou o valor de w_0 e w_1 e definiu que seria respectivamente 0 e 10. O valor de x_1 será 5.1, portanto o cálculo realizado será:

$$\hat{Y} = f(x) = 0 + 10 * 5.1 \quad \text{(II)}$$

$$\hat{Y} = 51 \quad \text{(III)}$$

Esse cálculo será realizado para todas as previsões que serão realizadas, portanto se essas previsões forem plotadas, chegará ao seguinte gráfico:

Figura 3-Previsões da regressão linear simples plotadas



Fonte: Damasceno (2020).

Os pontos azuis representam os valores reais, já a reta representa a reta estimada pelo modelo. O ponto circulado em vermelho é o valor real, da variável que calculada anteriormente, e chegou-se a um valor de 51, já o valor real é de mais ou menos 49, portanto há um erro no valor estimado do modelo (DAMASCENO, 2020).

Se for calculado o valor real menos o valor previsto, se obterá o erro/ resíduo. Portanto a equação ficaria:

Resíduo = Y - Yˆ (IV)

Neste caso o resíduo é -2 para esse ponto em específico. O resíduo representa a quantidade da variabilidade de Y que o modelo ajustado não consegue explicar. Os resíduos contém informação sobre o motivo do modelo não ter se ajustado bem aos dados (DAMASCENO, 2020).No entanto, dentre as regressões lineares é importante destacar a regressão linear múltipla tomando como base de que no descrito trabalho ela foi utilizada. A regressão múltipla é uma junção de técnicas estatísticas para construir modelos que descrevem de maneira satisfatória relações entre várias variáveis explicativas de um determinado processo.

A diferença entre a regressão linear simples e a múltipla é que na múltipla são tratadas duas ou mais variáveis explicativas. Enquanto a regressão clássica envolve uma única variável independente e uma variável dependente, a regressão linear múltipla envolve duas ou mais variáveis independentes que contribuem para uma única variável dependente (FARIA, 2006). Entretanto os fundamentos teóricos de ambas são similares.

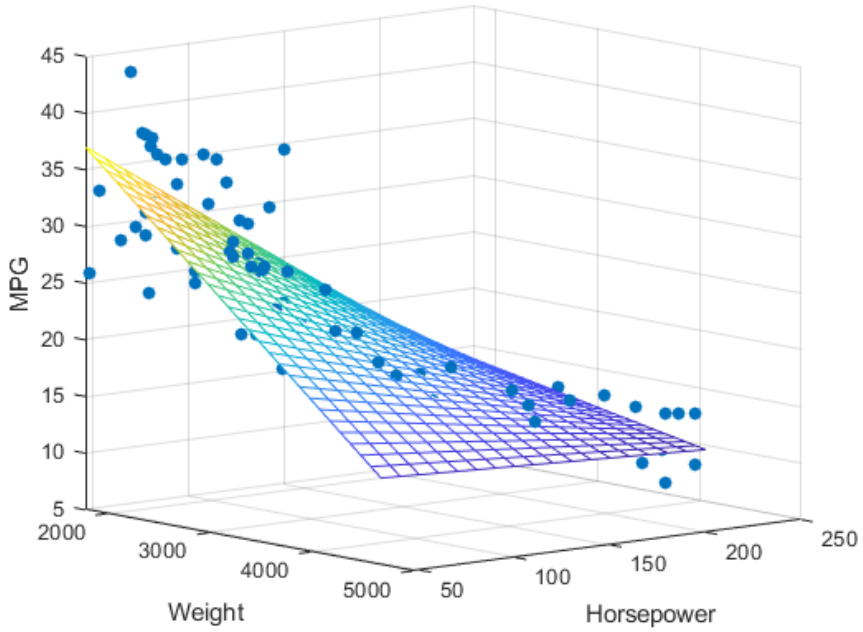
A regressão linear múltipla ou multilinear é usada quando há *n* variáveis de entrada e uma variável de resposta. É definida da seguinte maneira:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki} + e_i \quad \text{(V)}$$

No qual, α é o valor esperado de Y quanto todas as variáveis independentes forem nulas.β_1 é a variação esperada em Y dado um incremento unitário em X_1, mantendo-se constantes todas as demais variáveis independentes. β_k é a variação esperada em Y dado um incremento unitário em X_k , mantendo-se constantes todas as demais variáveis independentes, e_i é o erro não explicado pelo modelo (FARIA,2006).

Plotando em gráfico as *n* variáveis de entrada de uma regressão linear múltipla obtêm-se:

Figura 4 - Gráfico de regressão linear múltipla plotado com n variáveis



Fonte: Ge & Wu (2020).

Devem ser citadas as métricas de validação usadas no algoritmo de regressão linear, seja simples ou múltipla,com ou sem uso de *machine learning*. O primeiro consiste no SQR (Soma dos Quadrados dos Resíduos) e é definido pela Soma dos quadrados dos resíduos, mostra a variação de Y que não é explicada pelo modelo. É a medida da variação que não pode ser explicada (DAMASCENO,2020).

SQR = $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (yi - \check{y}i)^2$ (VI)

O R^2 é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. Ele também é conhecido como o coeficiente de determinação ou o coeficiente de determinação múltipla para a regressão múltipla (DAMASCENO, 2020).O R^2 está sempre entre 0 e 1. O 0 indica que o modelo não explica nada da variabilidade dos dados de resposta ao redor de sua média e o 1 indica que o modelo explica toda a variabilidade dos dados de resposta ao redor de sua média.

Há também o erro médio absoluto (MAE) que consiste na métrica de erro de regressão mais simples. Ele calcula o valor dos resíduos para cada um dos pontos e depois é calculada a média de todos esses resíduos (DAMASCENO,2020).

MAE = $\frac{1}{2}\sum \left| y - \check{y} \right|$ (VII)

Por fim, a Média dos erros ao quadrado (MSE) é apenas o cálculo do erro, mas elevado ao quadrado (DAMASCENO,2020).

MSE = $\frac{1}{2}\sum \left| y - \check{y} \right|^2$ (VIII)

O modelo de regressão multilinear para os dados utilizado no estudo descrito obteve um R^2 de 0,69 implicando que 69% da variabilidade é explicada pelo modelo de regressão através da equação:

$$Y = 0{,}01980 * (\text{n\_tarefa}) + 0{,}12561 * (\text{media\_nota\_tarefa}) + 0{,}001802 * (\text{n\_listas}) + 0{,}038463 * (\text{n\_listas}) + 0{,}75121 * (\text{media\_nota\_listas}) + 0{,}01790 * (\text{n\_pdfs}) + 0{,}01193 * (\text{n\_vídeo}) + 0{,}11942 \quad \text{(IX)}$$

Logo, identificou-se que o modelo obteve um erro absoluto médio (MAE) de 0,05 e erro quadrático médio (EQM) de 0,01 implicando uma baixa diferença nos valores do que é predito pelo modelo e o que foi observado (SAMMUT; WEBB, 2010). Para estudos relativos a ciências sociais e educacionais devido à interferência de variáveis imprevisíveis no comportamento humano, geralmente são aceitáveis R^2 mais baixos (ONDITI, 2013).

3.4.2 K- Means Cluster

A análise de agrupamento, também conhecida como *clustering*, é um conjunto de técnicas computacionais que consiste em separar objetos em grupos (*clusters*) baseados nas suas características (LINDEN, 2005). Esta técnica tem o objetivo de separar os grupos por função de dissimilaridade, com o intuito de encontrar características parecidas dentro de seus grupos e características distintas entre eles, ao mesmo tempo. O algoritmo *K-means* categoriza os dados coletados buscando uma similaridade entre eles, sendo cada similaridade apresentada como um agrupamento dos dados (LINDEN, 2005).

Os algoritmos de agrupamento podem ser baseados em métodos hierárquicos, capazes de realizar análise de *clusters*. Eles têm como principal característica a possibilidade de, em determinado passo do algoritmo, mesclar um cluster com o outro, fazendo assim vários agrupamentos. Eles organizam os dados baseados em uma estrutura hierárquica de acordo com a proximidade entre os indivíduos, o que resulta em uma árvore binária (dendograma), onde a raiz da mesma representa o conjunto de dados inteiros e as folhas representam os indivíduos finais (DONI, 2004).

Os métodos hierárquicos podem ser subdivididos em métodos aglomerativos e divisivos. Na forma aglomerativa, cada elemento se inicia representando um grupo e com o passar do tempo o elemento ou grupo se liga a outros por forma de similaridade, até o último passo, onde se forma um grupo único com todos os elementos. A forma divisiva acontece de forma oposta a aglomerativa, onde o grupo inicial contem todos os elementos do grupo de dados e com o decorrer do tempo, é dividido em subgrupos, de forma que os elementos de um subgrupo possuam certa distância dos elementos de outro (DONI, 2004).

Métodos de agrupamento permitem aprender a mistura parâmetros de dados. Além da modelagem probabilística, discute-se quantização vetorial e agrupamento hierárquico.Na prática, o *clustering* auxilia a identificação de duas qualidades de dados: clusters significativos e clusters úteis (ALPAYDIN, 2014).

Um dos métodos não hierárquicos mais conhecidos e utilizados atualmente é o *k-means. K-Means* é um algoritmo de agrupamento de dados não hierárquico que utiliza uma técnica iterativa para particionar um conjunto de dados. Ele foi proposto num trabalho pioneiro de S. Lloyd em 1957, contudo, só foi publicado no ano de 1982. Esse algoritmo busca minimizar essa função e é um problema muito difícil. É possível mostrar que o algoritmo *k-means* converge para um mínimo local (JAIN, 2010).

O *k-means* é o mais popular e mais simples algoritmo particional. *K-means* foi descoberto independentemente em diferentes campos científicos, primeiramente por Steinhaus e mesmo tendo sido proposto há mais de 50 anos, ainda é um dos algoritmos mais utilizados para clusterização devido à facilidade de implementação, simplicidade, eficiência e sucesso empírico e possui várias extensões desenvolvidas em várias formas (JAIN, 2010).

O método de clusterização *K-means* tem como característica de formação o modelo particional, no qual há a realocação iterativa baseada em erro quadrático. E seu objetivo é particionar *n* observações dentre *k* grupos onde cada observação pertence ao grupo mais próximo da média (JAIN, 2010).

O algoritmo *K-means* tem como finalidade particionar os dados em cluster, de forma que não exista área conjunta entre os *cluster* (conjuntos), erradicando a possibilidade de um cluster permanecer dentro de outro. Logo, amostras dentro de um cluster são parecidas, enquanto amostras de diferentes *clusters* seguem diferentes. O elemento principal do algoritmo funciona por um processo de duas etapas denominado maximização da expectativa. A etapa de expectativa atribui cada ponto de dados ao seu centróide mais próximo. Em seguida, a etapa de maximização calcula a média de todos os pontos para cada cluster e define o novo centróide (ANASTACIO, 2020).

A função *k-means* particiona os dados em k *clusters* mutuamente exclusivos e retorna o índice do cluster ao qual atribui cada observação. O *K-means* trata cada observação em seus dados como um objeto que tem uma localização no espaço. A função encontra uma partição na qual os objetos em cada *cluster* estão os mais próximos possíveis uns dos outros e os mais distantes possíveis dos objetos em outros *clusters*. Pode-se escolher uma métrica de distância para usar o *k-means* com base nos atributos de seus dados. Como muitos métodos de

agrupamento, k-*means clustering* requer que se especifique o número de *clusters* k antes do *clustering* (ANASTACIO, 2020).

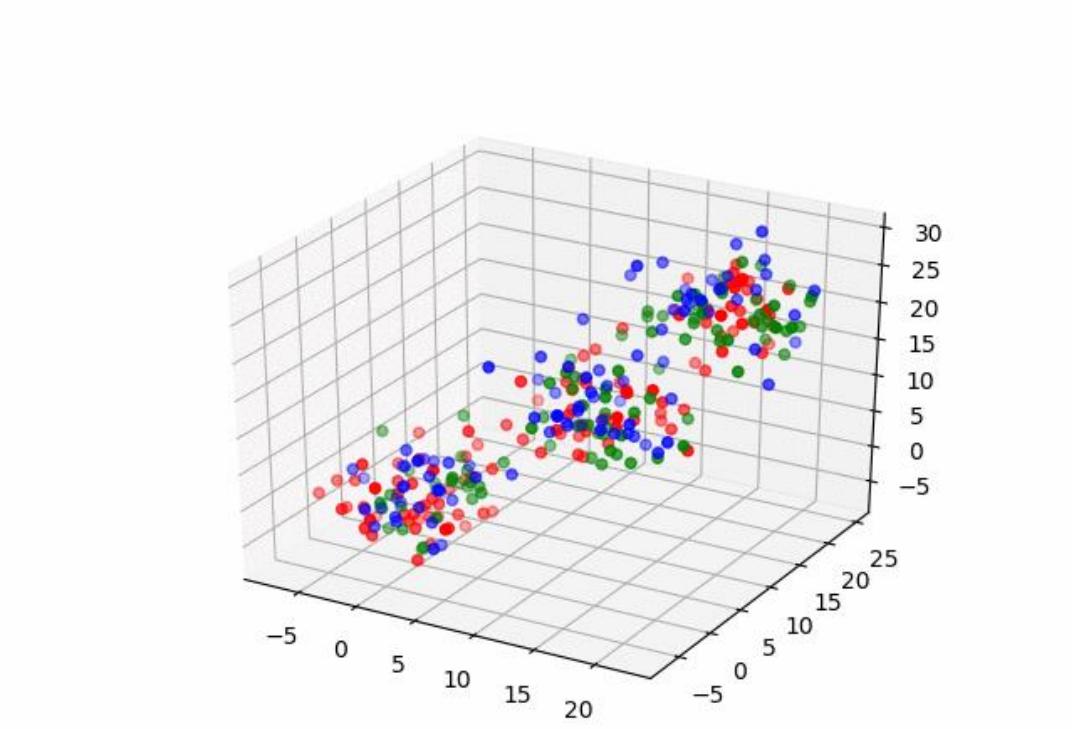
A ideia básica é que elementos que componham um mesmo *cluster* devem apresentar alta similaridade, mas devem ser muito dissimilares de objetos de outros *clusters.* Em outras palavras, toda clusterização é feita com objetivo de maximizar a homogeneidade dentro de cada cluster e maximizar a heterogeneidade entre clusters (JAIN, 2010).

A grande vantagem do uso das técnicas de Clusterização é que, ao agrupar dados similares, pode-se descrever de forma mais eficiente e eficaz as características peculiares de cada um dos grupos identificados. Isso fornece um maior entendimento do conjunto de dados original, além de possibilitar o desenvolvimento de esquemas de classificação para novos dados e descobrir correlações interessantes entre os atributos dos dados que não seriam facilmente visualizadas sem o emprego de tais técnicas (JAIN, 2010).

Para uma clusterização com o algoritmo *K-means* é necessário tratar os dados para o formato de entrada do algoritmo. Dessa maneira, o algoritmo é iniciado com o valor de K arbitrário e há o cálculo da distância (raio) para cada ponto de um centroide, este processo será feito por uma matriz de distância, na qual cada linha representa a distância de cada dado para cada centroide. Então, cada dado é designado para o centroide mais próximo. O novo ponto médio de centroide será medido pelo ponto médio de seus pontos ao redor e a etapa será repetida até não ocorrer mais mudanças (ANASTACIO, 2020).

No exemplo a seguir será mostrado como o *K-means* se comporta graficamente. Serão usados três *clusters* e consequentemente, o K será igual a três. Inicialmente, o algoritmo insere o K pontos (centroide).

Figura 5- Início da clusterização com o algoritmo *K-means*

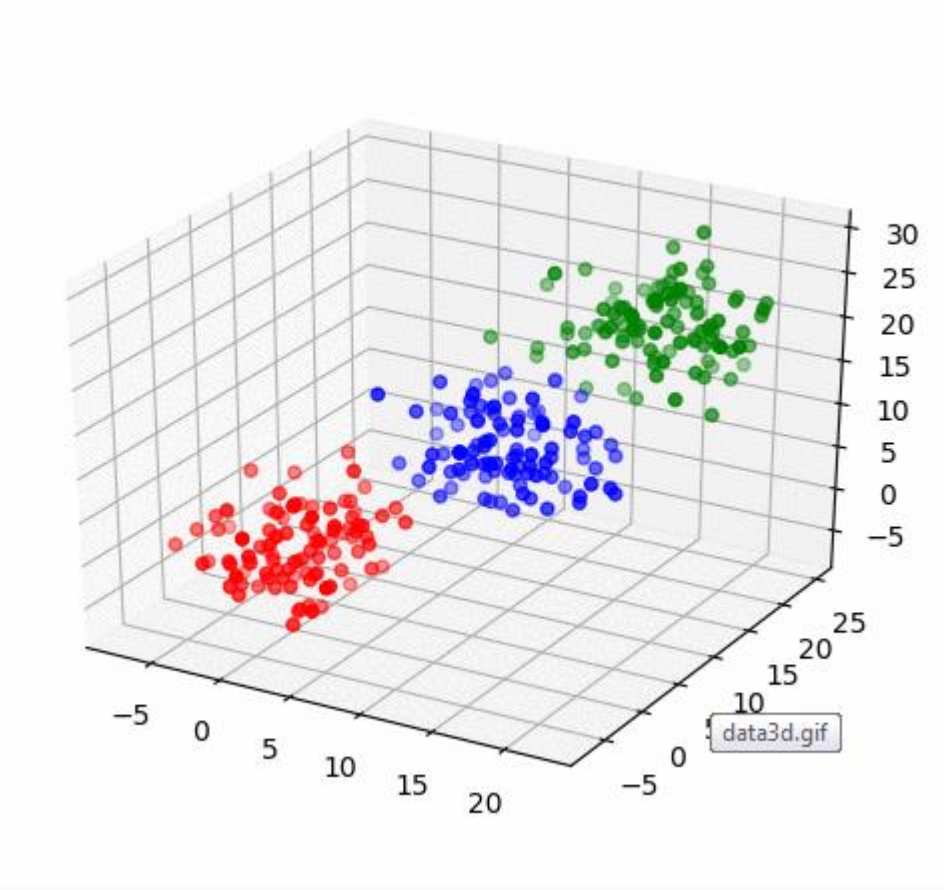


Fonte: Aydoğdu (2018).

A primeira iteração do algoritmo é calcular a distância média de todos os pontos que estão atrelados ao centróide, e então mudar a posição do centróide para o novo ponto que foi calculado, que é a distância média de todos os pontos que se ligaram à aquele centróide. Essa mudança de posição do centróide pode alterar os itens que fazem parte daquele grupo.

Após a iteração do cálculo da média, alguns pontos mudaram de centróide, os pontos que estão marcados em verde passaram do centróide azul para o vermelho, e o que está marcado em azul passou do centróide vermelho para o azul. Essa iteração de cálculo da média da distância dos pontos até o centróide ocorre em loop até que nenhum ponto mude de centróide, isso acontece quando os centróides param de se mover porque já estão na posição central da distância entre os pontos.

Figura 6- Finalização da clusterização e grupos separados pelo algoritmo *K-means*



Fonte: Aydoğdu (2018).

Entre a penúltima iteração e esta não ouve mais mudança de pontos entre o gráfico e o centróide, fazendo com que o algoritmo *K-Means* pare sua execução chegando ao resultado esperado e criando três grupos. Assim, quando um novo item for incluído no gráfico, ele já terá um grupo que atende aquela região e o computador já saberá do que se trata o dado novo.

4 APLICAÇÃO DO MÉTODO DE DATA SCIENCE

A metodologia *Data Science* traz ferramentas para lidar com o grande volume de dados desde a extração ao desenvolvimento com o *Machine learning* de um modelo preditivo de qualidade, que possa fornecer inteligência para tomada de ações ágeis (MITCHELL, 1999).

Os dados da escola que foram utilizados neste estudo foram extraídos diretamente do banco de dados da empresa Eduqo, que fornece uma plataforma de personalização da aprendizagem. Por meio da plataforma as escolas podem adicionar e armazenar atividades escolares de diversos tipos, como por exemplo, provas e tarefas. Com a pandemia e o fechamento das escolas, o número de clientes desta empresa aumentou tendo diversas escolas que se adaptaram repentinamente ao EaD com uso da plataforma. O estudo dos dados gerados pela plataforma por metodologia *Data Science*, pode fornecer informações de como professores podem atuar para aprimorar o aprendizado dos seus alunos.

Na plataforma cada professor adiciona os dados de forma exclusivamente determinada por eles mesmos e pela coordenação da escola. Assim os professores podem realizar todas as suas rotinas pedagógicas e disponibilizar seu material de forma personalizada. Existe a opção de o professor adicionar diversos tipos de conteúdo tais como vídeos, materiais em PDF: artigos e textos; também tarefas e listas de exercício. Em relação às provas os professores podem adicionar sua própria prova montar prova diretamente na plataforma e ainda podem obter questões do banco de dados públicos disponibilizados pela Eduqo. Uma biblioteca com todos os materiais adicionados na plataforma pelo professor fica salva para este aproveitar o que foi adicionado em atividades anteriores.

Os dados coletados não necessariamente seguem um padrão de nomenclatura. Os dados relacionados ao comportamento do aluno frente aos tipos de materiais são coletados e modelados com objetivo de obter uma inteligência sobre cada modelo, identificando padrões e grupos de alunos com dificuldades e quais conteúdos precisam de reforço. Esse estudo busca assim obter análises personalizadas e indicadores do desempenho dos alunos da escola. A tabela 1 resume as categorias de dados coletados bem como apresenta o que foi coletado de cada categoria.

Tabela 1 - Categoria dos dados coletados

Categoria	O que foi coletado
Tarefas	Quantidade de tarefas, média de nota das tarefas
Listas de exercício	Quantidade de listas, Quantidade de exercícios por lista e média da nota das listas
Textos com conteúdo em PDF	Quantidade de textos em PDF
Vídeo e vídeo aulas	Quantidade de vídeos
Atividades avaliativas - Prova online	Média de notas das provas
Apresentação em Power Points com conteúdo	Quantidade de apresentações em PPT
Tempo total	Tempo total de uso da plataforma nestas categorias

Fonte: O autor.

Os registros de tudo que o aluno fez na plataforma podem ser extraídos via linguagem de Consulta Estruturada (*SQL, Standard Query Language*) um tipo de linguagem considerada padrão para banco de dados (MELTON; SIMON, 1993). Os dados coletados são de uma escola privada de São Paulo, durante o período de 15 de março de 2020 até 27 de agosto de 2020 de todos seus alunos do primeiro ano do fundamental ao terceiro ano do ensino médio ativos na plataforma da empresa Eduqo.

4.1 ANÁLISE PRÉ-EXPLORATÓRIA

Neste estudo os dados extraídos da escola em questão foram avaliados de forma a encontrar algum tipo de inteligência a partir deles. Na parte de extração desses dados, todos os dados das atividades documentadas na própria plataforma desde resultados em provas, frequência do aluno e se ele viu o vídeo disponibilizado por completo foram extraídos.

No contexto deste estudo apesar de a escola tentar controlar tudo o que seus alunos assistem, existem muitos alunos que por diversos motivos, sejam estes sociais e/ou tecnológicos, não possuem todos os dados dentro da plataforma, ou seja, são perfis de alunos que nunca avançaram com o programa escolar durante a pandemia. Assim, na preparação dos dados, foi necessário limpar os dados com ferramentas de programação que auxiliam a visualização dos dados não completos, para que a análise destes seja representativa com alunos que realmente participaram da maioria das atividades disponibilizadas. No que diz

respeito à normalização, os dados foram normalizados a partir das turmas de cada aluno para garantir que permaneçam na escala estabelecida.

Após os dados coletados, foi feita uma pré-análise estatística deles para entender sobre a correlação entre as variáveis, desvio padrão e médias destes dados. Basicamente são propriedade estatística puderam ser obtidas de maneira bem simples por códigos prontos na biblioteca Pandas, na linguagem *Python*. Objetivo dessa etapa é conhecer como está a disposição dos dados, para ter uma pré-visualização de qual seriam as variáveis mais relevantes e assim obter uma visão sobre qual modelo pode ser utilizado.

Os dois testes estatísticos mais comumente usados para estabelecer a relação entre as variáveis são a correlação de Pearson e o *p-value*. A correlação é uma forma de testar se duas variáveis têm algum tipo de relacionamento, enquanto o *p-value* diz se o resultado de um experimento é estaticamente significativo (MIOT, 2018). No desenvolvimento do trabalho o teste estatístico utilizado foi o Coeficiente de Correlação de Pearson.

O Coeficiente de Correlação de Pearson também é chamado de "coeficiente de correlação produto-momento" ou simplesmente de "ρ de Pearson". É um teste estatístico que explora a intensidade e o sentido do comportamento mútuo entre variáveis. Este coeficiente pode assumir apenas valores entre -1 e 1. A correlação indica a interdependência entre duas variáveis. O cálculo do Coeficiente de Correlação de Pearson serve para detectar o grau de correlação entre as variáveis quando não se é facilmente compreendida sua interdependência (MIOT,2018). Existem várias possibilidades de interpretação da correlação. Pode-se considerar a inclinação da reta que representa a relação entre as variáveis, podem-se considerar as séries de valores como vetores, e o ρ, em uma interpretação geométrica, representaria o cosseno do ângulo formado entre os vetores, etc (SALLES, 2018).

Tabela 2 - Graus de intensidade das correlações entre variáveis através do Coeficiente de Pearson.

Graus de intensidade	Correlação
0.9 a 1	positivo ou negativo indica uma correlação muito forte.
0.7 a 0.9	positivo ou negativo indica uma correlação forte.
0.5 a 0.7	positivo ou negativo indica uma correlação moderada.
0.3 a 0.5	positivo ou negativo indica uma correlação fraca.
0 a 0.3	positivo ou negativo indica uma correlação desprezível.

Fonte: Salles (2018).

Já o *P-value* depende diretamente de uma dada amostra, e fornece uma medida da força dos resultados de um teste, em contraste a uma rejeição ou não rejeição. Se a hipótese nula for verdadeira e a chance da variação aleatória for a única razão para as diferenças amostrais, então o *P-value* é uma medida quantitativa para alimentar o processo de tomada de decisão como evidência (NUZZO,2014).

Tabela 3- Interpretação do *P-value*

P-value	Interpretação
P< 0,01	evidência muito forte contra H0
0,01< = P < 0,05	evidência moderada contra H0
0,05< = P < 0,10	evidência sugestiva contra H0
0,10< = P	pouca ou nenhuma evidência real contra H0

Fonte : Nuzzo (2014).

Para uma amostra de tamanho fixo, quando o número de realizações é decidido antecipadamente, a distribuição de p é uniforme (assumindo a hipótese nula). Isto é expresso como P(p < x) = x. E significa que o critério de p <0,05 atinge um de 0,05, ou seja, o nível alfa (NUZZO, 2014).

O valor de P é uma medida que demonstra quanto de evidência se tem contra uma hipótese nula. Quanto menor o valor de P, mais evidência se tem. Deve-se combinar o *P-value* com o nível de significância para tomar decisões sobre um dado teste de hipótese. Em tal caso, se o *P-value* for menor que algum corte (usualmente 0,05, algumas vezes um pouco mais como 0,1 ou um pouco menos como 0,01) então a hipótese nula é rejeitada (NUZZO, 2014).

Entende-se que a distribuição dos *P-values* sob-hipótese nula H_0 é uniforme, e não depende de uma forma particular do teste estatístico. Em um teste de hipótese estatístico, o *P-value* é a probabilidade de observar um teste estatístico com valor observado, assumindo que a hipótese nula seja verdadeira (NUZZO, 2014).

O valor de p é definido com respeito a uma distribuição. Portanto, pode-se chamar de "hipótese de modelo-distribucional " ao invés de " hipótese nula".Neste caso há a significância de que se a hipótese nula fosse verdadeira, o *P-value* é a probabilidade contra a hipótese nula deste caso (NUZZO,2014).

4.2 MODELAGEM PARA REGRESSÃO MULTILINEAR

Como explicado anteriormente, à regressão multilinear refere-se a diferentes variáveis independentes (X) usadas para fazer a predição (DAMASCENO, 2020).

O objetivo na regressão linear é ajustar os valores reais o mais próximo possível. Quantificar a diferença entre os valores previstos e reais usando uma função de perda (ou custo). Essa função de perda pode ser baseada nos algoritmos de erro quadrático médio ou erro quadrático médio. O exercício é repetido até que a função perda convirja e alcance o valor mínimo. Quando um algoritmo atinge seu alvo, ele é conhecido como convergente. Isso é obtido usando o algoritmo de gradiente descendente, que depende internamente das técnicas de cálculo parcial multivariado (MALIK, 2019).

O gradiente descendente é usado em vários algoritmos, incluindo regressão para redes neurais. Ele depende muito de cálculos multivariados para encontrar os pontos mínimos.O algoritmo, conhecido como gradiente descendente (GD), é usado para encontrar mínimos e / ou máximos de uma função. Essa função pode ser a função de custo de um algoritmo de aprendizado de máquina (MALIK, 2019).

O gradiente descendente encontra a taxa de mudança das variáveis e se ajusta para se mover em direção ao ponto mínimo. O ponto mínimo, neste caso, são os valores das variáveis de entrada que nos darão o valor mínimo para a função de custo(MALIK,2019).

Cada vez que o algoritmo decide se precisa aumentar ou diminuir os valores das variáveis de entrada. Ele faz isso calculando a taxa de variação das variáveis. A taxa de variação das variáveis é calculada usando a técnica de derivada parcial de cálculo multivariado (MALIK,2019).

Quando atinge o ponto mínimo, a derivada da função será 0. Isso implica que a descida do gradiente atingiu o ponto mínimo. O ponto mínimo é quando a derivada é "0".Embora haja uma solução algébrica numericamente para encontrar o mínimo local, a natureza iterativa derivada da descida de gradiente escala melhor quando temos um conjunto de dados maior (MALIK, 2019).

Para encontrar a descida do gradiente para várias variáveis é necessário tratar cada variável separadamente, tornando todas as outras variáveis constantes e, em seguida, encontrar a derivada parcial da função. Este é um caso de uso crucial de cálculo parcial multivariado em algoritmos de aprendizado de máquina. Esta técnica é usada em algoritmos de otimização, regressão e redes neurais (MALIK, 2019).

O modelo é inicialmente ajustado em um conjunto de dados de treinamento (GARETH, 2013), que é um conjunto de exemplos usados para ajustar os parâmetros (por exemplo, pesos de conexões entre neurônios em redes neurais artificiais) do modelo. O modelo é treinado no conjunto de dados de treinamento usando um método de aprendizado supervisionado, por exemplo, usando métodos de otimização como gradiente descendente ou gradiente descendente estocástico (GARETH, 2013).

Na prática, o conjunto de dados de treinamento geralmente consiste em pares de um vetor de entrada (ou escalar) e o vetor de saída correspondente (ou escalar), onde a chave de resposta é comumente indicada como o destino (ou rótulo). O modelo atual é executado com o conjunto de dados de treinamento e produz um resultado, que é então comparado com o destino, para cada vetor de entrada no conjunto de dados de treinamento. Com base no resultado da comparação e no algoritmo de aprendizagem específico que estão sendo usados, os parâmetros do modelo são ajustados. O ajuste do modelo pode incluir seleção de variável e estimativa de parâmetro (GARETH, 2013).

Na execução desse procedimento foi utilizada a linguagem *Python* e uma biblioteca específica, a *Scikit Learn*. Essa biblioteca possui o algoritmo de *machine learning* para

regressão multilinear já desenvolvida, no qual o algoritmo é comum e amplamente utilizado no mundo dos dados (ARVAI, 2020).

O *train test split*, após aprender com os dados realiza a regressão. Neste algoritmo é necessário dados de entrada e algumas pré-configurações. Sendo assim, os dados de entrada foram escolhidos apenas dentre os dados que foram limpos, ajustados durante a pré-análise e demonstraram boas correlações entre as variáveis e significância estatística, obtendo um resultado assertivo e limpo.

As configurações foram testadas de maneira a obter os melhores resultados. Há algumas técnicas para encontrar a configuração mais adequada, entretanto especificamente nesse modelo, 20% dos dados foram utilizados para teste e 80% para treino.

4.3 VALUATION

Sucessivamente, o modelo ajustado é usado para prever as respostas para as observações em um segundo conjunto de dados denominado conjunto de dados de validação (GARETH, 2013). O conjunto de dados de validação fornece uma avaliação imparcial de um modelo ajustado no conjunto de dados de treinamento enquanto ajusta os hiperparâmetros do modelo (BRIAN, 1996). Os conjuntos de dados de validação podem ser usados para regularização por parada antecipada (parando o treinamento quando o erro no conjunto de dados de validação aumenta, pois isso é um sinal de sobreajuste do conjunto de dados de treinamento (PRECHELT e ORR, 2012).

Este procedimento simples é complicado na prática pelo fato de que o erro do conjunto de dados de validação pode flutuar durante o treinamento, produzindo vários mínimos locais. Essa complicação levou à criação de muitas regras ad-hoc para decidir quando o sobreajuste realmente começou (PRECHELT e ORR,2012).

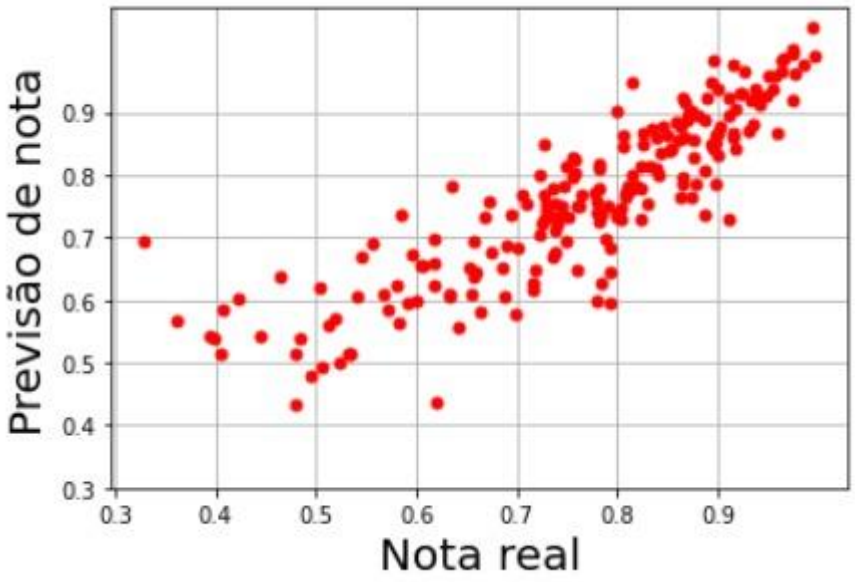
Finalmente, o conjunto de dados de teste é um conjunto de dados usado para fornecer uma avaliação imparcial de um ajuste do modelo final no conjunto de dados de treinamento. Se os dados no conjunto de dados de teste nunca foram usados no treinamento (por exemplo, na validação cruzada), o conjunto de dados de teste também é chamado de conjunto de dados de validação. O termo "conjunto de validação" às vezes é usado em vez de "conjunto de teste" em alguma literatura (por exemplo, se o conjunto de dados original foi particionado em apenas dois subconjuntos, o conjunto de teste pode ser referido como o conjunto de validação) (PRECHELT e ORR,2012).

Um conjunto de teste é, portanto, um conjunto de exemplos usados apenas para avaliar o desempenho (ou seja, generalização) de um classificador totalmente especificado (RIPLEY, 1996). Para fazer isso, o modelo final é usado para prever classificações de exemplos no conjunto de teste. Essas previsões são comparadas às classificações verdadeiras dos exemplos para avaliar a precisão do modelo (LAROSE e LAROSE, 2014).

Em um cenário em que os conjuntos de dados de validação e teste são usados, o conjunto de dados de teste é normalmente usado para avaliar o modelo final selecionado durante o processo de validação. No caso em que o conjunto de dados original é particionado em dois subconjuntos (conjuntos de dados de treinamento e teste), o conjunto de dados de teste pode avaliar o modelo apenas uma vez (por exemplo, no método de validação) (KOHAVI, 1998).

No estudo deste trabalho através do teste de validação, foram encontradas previsões com o conjunto de dados testes que possibilitaram uma avaliação da precisão do modelo adotado. No gráfico abaixo, os dados de teste foram correlacionados entre a previsão de nota e a nota real dos estudantes selecionados.

Figura 5 - Validação da previsão de nota versus nota real dos estudantes



Fonte: O autor.

O modelo adotado obteve um R^2 de 0,69, predizendo que 69% da variabilidade é explicada pelo modelo de regressão. No contexto de estudos envolvendo o desempenho escolar foram reportados alguns R^2 de regressões multilineares como Lacruz; Américo (2017)

R^2 =0,229; Alves; Soares (2013) R^2 =0,39; Morris; Finnegan; Wu (2005) R^2 =0,31 e Mazulo (2015) R^2 =0,219. Portanto, pode-se considerar que o modelo de regressão multilinear obtido apresenta R^2 adequado ao estudo.

4.4 MODELAGEM K-MEANS

Na etapa de Modelagem, o algoritmo *K-means* foi escolhido pelo fato de que embora a quantificação que a Regressão Multilinear demonstrou no modelo aplicado tenha sido válida, ela não responderia de forma prática qual dos alunos apresentavam dificuldade de uma maneira otimizada.

Então, o *K-means* foi ajustado para obter os valores da regressão com maior peso, ou seja, os valores que realmente importam e são os indicadores do modelo. Dessa forma, os dados foram selecionados para separar os alunos em grupos possibilitando a discretização de uma maneira rápida e eficiente para os professores e grupo pedagógico.

Na execução desse procedimento foi utilizada a linguagem *Python* e a biblioteca própria *Scikit Learn*, que conta com o algoritmo de *machine learning* desenvolvido para realizar o *K-means* (ARVAI, 2020). O *K-means* foi executado com os indicadores relevantes para separar os alunos. Portanto, os dados de entrada foram exclusivamente os dados dos indicadores.

Na escolha do número ideal de *cluster* são utilizados o método do cotovelo ou o coeficiente de silhueta. Quando o SSE é plotado como função do número de clusters, à medida que mais centroides são adicionados, a distância de cada ponto até seu centroide mais próximo diminui. Então, o ponto ideal onde a curva SSE se dobra é definido como ponto de cotovelo, no qual o valor x deste ponto é considerado uma compensação razoável entre o erro e o número de clusters (ARVAI, 2020).

Existem diferentes maneiras para determinar o número de *cluster* que seja o mais próximo da realidade. Neste projeto, o número de K (*clusters*) para a separação dos grupos escolhido foi o K=4. Essa escolha foi tomada levando em conta um valor otimizado e a otimização do trabalho da escolha, já que se o valor de k fosse muito alto, isso dificultaria o trabalho da escola em separar todos os níveis de alunos com dificuldades. Portanto, foi pensado que quatro *clusters* seriam o ideal neste processo. É importante lembrar que o projeto trabalha com a *Data Science* e esse valor poderia ser mudado após uma experimentação da escola ou uma abrupta mudança nos dados.

5 RESULTADOS E DISCUSSÕES

Na parte de extração desses dados, os dados das atividades documentadas na própria plataforma foram obtidos para cada categoria apresentada na seção Dados deste artigo. Na Tabela 4 é apresentada a nomenclatura adotada para cada categoria para melhor compreensão dos resultados.

Tabela 4 - Nomenclatura adotada para cada categoria dos dados

Categoria	O que foi coletado	Nomenclatura
Tarefas	Quantidade de tarefas, média de nota das tarefas	n_tarefa média_nota_tarefa
Listas de exercício	Quantidade de listas, Quantidade de exercícios por lista e média da nota das listas	n_listas n_exercícios média_nota_listas
Textos com conteúdo em PDF	Quantidade de textos em PDF	n_pdfs
Vídeo e vídeo aulas	Quantidade de vídeos	n_vídeo
Atividades avaliativas - Prova online	Média de notas das provas	média_nota_provas
Apresentação em Power Points com conteúdo	Quantidade de apresentações em PPT	n_ppt
Tempo total	Tempo total de uso da plataforma	tempo_total

Fonte: O autor.

Para obter indicadores de desempenho do aluno, foram coletados dados dos alunos do ensino fundamental ao médio de uma escola privada usando a plataforma da Eduqo. Inicialmente a população coletada continha 1424 alunos e foi reduzida a 1009 alunos (71% do total) após a retirada de alunos desistentes e com dados contendo faltas de informações e/ou campos nulos. Essa etapa de preparação dos dados é importante para otimizar as análises com uma amostragem mais representativa. Em 1009 alunos, foram avaliadas 57 turmas com média 17 de alunos por turma com desvio padrão de 7,5.

Avaliando os dados foi visto que a variável de número de apresentações (n_ppt) continha muitos dados nulos comparados com as demais variáveis, portanto, para não

interferirem estas foram retiradas para as análises subsequentes. Os dados de número de apresentação não foram retirados no tratamento dos dados, pois ocasionaria redução de 40% dos dados de análise, podendo afetar na obtenção de um modelo representativo. Análises estatísticas simples dos dados dos 1009 alunos como contagem, média e desvio padrão foram realizadas em relação a cada variável e estão resumidas na Tabela 5.

Tabela 5 - Análises estatísticas dos dados

	n_tarefas	média_nota_tarefas	n_listas	n_exercícios	média_nota_listas	n_pdfs	n_vídeo	média_nota_provas	tempo_total
Média	31,305	0,608	1 24,904	479,861	0,726	6 58,568	1 42,069	7,709	16 81,701
Desvio Padrão	29,312	0,310	6 8,058	304,231	0,128	1 55,695	6 9,197	1,478	12 31,343

Fonte: O autor.

Devido à alta variabilidade da variável de tempo total (Desvio padrão=1231,343) para realização de todas as atividades (tempo_total), concluiu-se que esta variável poderia estar contabilizando o tempo que o aluno deixa a plataforma aberta, mas não necessariamente estaria utilizando-a para acessar os conteúdos e realizar tarefas. Portanto para não se obter conclusões incorretas da atividade do aluno, optou-se pela retirada dessa variável.

Os dados foram normalizados a partir da turma do aluno, já que os alunos avaliados possuem idades e níveis diferentes de aprendizado. Assim, a normalização por turma auxilia a ter um melhor modelo para analisar a escola inteira com todos os seus anos escolares (fundamental ao ensino médio).

Na etapa de exploração e análise de dados, foram avaliadas as correlações entre as variáveis (Tabela 6), visando identificar quais destas teriam maior relevância para análise. Pela Tabela 6 é possível observar que média de nota de tarefas (0,553) e média de nota de listas (0,789) possui uma maior correlação com a média de nota de provas. Foi obtido que todas as variáveis eram estatisticamente significativas a 95% de confiança (p<000,1).

Tabela 6 - Correlações entre as variáveis

	n_tarefa	média_nota_tarefas	n_listas	n_exercícios	Média_nota_listas	n_pdfs	n_vídeo	média_nota_provas
média_nota_prova	0,449	0,553	0,378	0,447	0,789	0,329	0,328	1,000

Fonte: O autor.

O teste estatístico de Correlação de Pearson foi realizado na análise estatística entre as variáveis do presente trabalho. Essa análise demonstrou que todos os dados apresentam boas correlações entre si como apresentado na tabela 5. Então, foram realizados testes para se conhecer a significância estatística entre todas as variáveis correlacionadas através do *P-value* e foi validado que todas as variáveis são estatisticamente significantes.

Com o conhecimento da linearidade dos dados e significância das variáveis foi montado um modelo preditivo de regressão multilinear, utilizando o treinamento e teste de dados do algoritmo de *machine learning "Train and Split*". Neste algoritmo parte dos dados são separados para treino do modelo e parte dos dados na base de teste são usados para teste da performance do modelo. Após o teste do modelo para a amostra de dados, o modelo é testado novamente para toda a população de dados de forma a avaliar a previsibilidade do modelo (DATA FLAIR, 2018c).

O modelo de regressão multilinear para os dados (equação IX) obteve um R^2 de 0,69 implicando que 69% da variabilidade é explicada pelo modelo de regressão. Além disso, o modelo obteve um erro absoluto médio (MAE) de 0,05 e erro quadrático médio (EQM) de 0,01 implicando uma baixa diferença nos valores do que é predito pelo modelo e o que foi observado (SAMMUT; WEBB, 2010).

Para estudos relativos a ciências sociais e educacionais devido a interferência de variáveis imprevisíveis no comportamento humano, geralmente são aceitáveis R^2 mais baixos (ONDITI, 2013). No contexto de estudos envolvendo o desempenho escolar foram reportados alguns R^2 de regressões multilineares como Lacruz; Américo (2017) R^2 =0,229; Alves; Soares (2013) R^2 =0,39; Morris; Finnegan; Wu (2005) R^2 =0,31 e Mazulo (2015) R^2 =0,219.

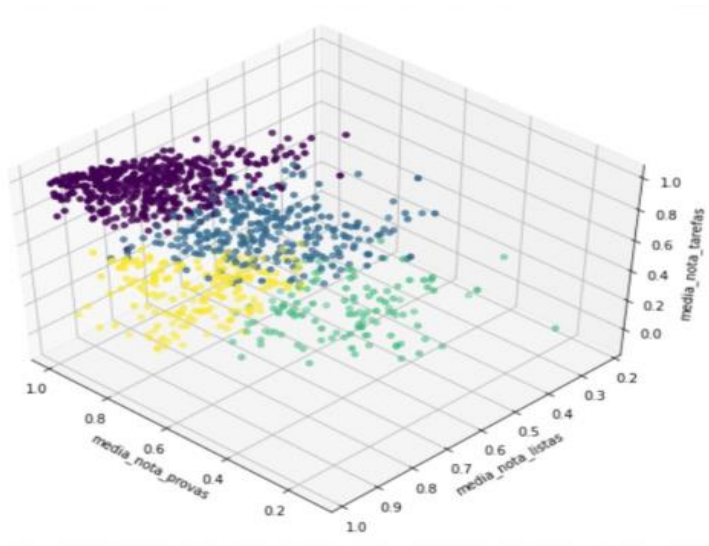
Portanto, pode-se considerar que o modelo de regressão multilinear obtido apresenta R^2 adequado ao estudo.

$$Y = 0{,}01980 * (n\_tarefa) + 0{,}12561 * (media\_nota\_tarefa) + 0{,}001802 * (n\_listas) + 0{,}038463 * (n\_listas) + 0{,}75121 * (media\_nota\_listas) + 0{,}01790 * (n\_pdfs) + 0{,}01193 * (n\_vídeo) + 0{,}11942 \quad \text{(IX)}$$

O modelo de regressão multilinear envolvendo essas variáveis fornecem indicadores que intuitivamente tem impacto conhecido na média de provas, ou seja, é esperado que alunos que tiram notas boas nas tarefas tenham bom desempenho em provas. Entretanto, o grande diferencial deste tipo de análise é poder quantificar a magnitude de influência dessas variáveis no aspecto de desempenho aprendizagem do aluno (HAIDER, 2015).

Partindo da regressão multilinear os coeficientes com maior peso, sendo eles média de nota de lista e de tarefa (com índices na regressão, respectivamente, de 0,751 e 0,126) foram utilizados como indicadores. Esses dois indicadores, foram utilizados para prever a média de nota das provas. Além disso, foi utilizado um algoritmo de clusterização *K-means* para criar 4 clusters (ou agrupamentos), separando os alunos conforme seu desempenho (Figura 8) com base nos dados de média dos indicadores (Tabela 5).

Figura 8 - Agrupamentos obtidos pelo algoritmo de clusterização *K-means*



Fonte: O autor.

O agrupamento 1, em roxo, representa os alunos com melhor desempenho em ambos indicadores, com médias acima de 0,79 (Tabela 5), correspondendo a 421 alunos dos 1009. O agrupamento 2, em amarelo, correspondem a 192 alunos que possuíram um bom desempenho, porém inferior ao agrupamento 1. Além disso, foi possível verificar que esse agrupamento tem uma média cerca de 3 vezes inferior do indicador nota de tarefas em relação ao indicador de nota de listas.

O agrupamento 3, em azul, são alunos que podem ser considerados com desempenho intermediário, correspondendo neste estudo a 279. Esses alunos apesar de terem um desempenho menor possuem ambos indicadores com médias semelhantes. O agrupamento 4 com 177 alunos, é o agrupamento de alunos que requer mais atenção do professor já que contemplam os alunos com o pior desempenho dentre os demais. Neste agrupamento, o indicador de nota de tarefas tem uma média cerca de 5 vezes menor do que a do indicador nota de listas.

Tabela 7 - Médias e desvio padrão para cada indicador e agrupamento

		média_nota_tarefas	**média_nota_listas**	**média_notas_provas**
Agrupamento 1 - 421 Alunos	**Média**	0,896	0,796	0,882
	Desvio Padrão	0,097	0,097	0,0765
Agrupamento 2 - 192 Alunos	**Média**	0,276	0,775	0,781
	Desvio Padrão	0,144	0,089	0,079
Agrupamento 3 - 279 Alunos	**Média**	0,607	0,655	0,693
	Desvio Padrão	0,112	0,096	0,111
Agrupamento 4 - 117 Alunos	**Média**	0,121	0,564	0,544
	Desvio Padrão	0,129	0,096	0,122

Fonte: O autor.

Desta forma a clusterização com base nos indicadores pode ser utilizada por professores para avaliarem rapidamente quais alunos necessitam de intervenção, no caso deste trabalho os alunos do agrupamento 4, podendo atuar para entender quais são as dificuldades destes alunos sejam elas no conteúdo ou no acesso da plataforma e assim poderem fornecer reforços e treinamentos.

Especificamente para os indicadores, os professores conseguem avaliar, de acordo com as suas diferentes médias, quais deles requerem uma maior preocupação. Podendo assim, levar mais alunos para o agrupamento 1 e ter turmas mais engajadas com bons desempenhos. Além disso, os professores conseguem ter um acompanhamento da sua turma, entendendo como o EaD e a digitalização da escola está impactando a turma por meio das análises dos indicadores em cenários pré-pandemia e com pandemia.

É importante ressaltar que o modelo de regressão multilinear obtido neste artigo pode ser melhorado com o passar do tempo a partir da obtenção de mais dados, traduzindo em um modelo com maior previsibilidade.

6 CONSIDERAÇÕES FINAIS

Ambientes escolares são responsáveis por gerar uma grande quantidade de dados que são de interesse ao estudo por campos de *Data Science* (ciência de dados). Existe uma dificuldade dos professores em definir em um meio de educação à distância (EaD) quais alunos estão tendo dificuldades seja na matéria ou de engajamento com as ferramentas digitais, requerendo assim um acompanhamento.

Este trabalho obteve pela coleta de dados de uma escola integrada a plataforma de aprendizado Eduqo, indicadores para acompanhamento do desenvolvimento do aluno. Esses indicadores foram obtidos seguindo a metodologia *data science* com posterior análise de regressão multilinear das variáveis com algoritmo de *machine learning "Train and Split",* obtendo um modelo com R^2=0,69, sendo obtidos os indicadores de média de nota de tarefa e média de nota de listas para clusterização pelo algoritmo *K-means*. Com a clusterização dos dados por meio dos indicadores obtidos, professores conseguem definir quais grupos de alunos em suas turmas estão tendo dificuldades, podendo agir para melhoria do aprendizado.

Este trabalho contribuiu para um melhor entendimento dos dados que são gerados em ambientes escolares e mostrou como a metodologia de ciência de dados pode ser aplicada a eles. Para os trabalhos futuros seria interessante avaliar quantitativamente a melhora do desempenho de alunos de uma turma por meio do monitoramento dos indicadores obtidos por esse modelo.

No contexto de uso de metodologias *data science* em ambiente escolar para obtenção de indicadores, seria válido realizar um estudo de processamento de linguagem natural (NLP), ou seja, um estudo correlacionando a interação dos alunos no ambiente de aprendizado remoto com a forma de expressão dos alunos no que diz respeito a sentimentos e emoções. Dessa forma seria possível compreender mais o aluno no que diz respeito à motivação em relação à rotina pedagógica.

REFERÊNCIAS

ABED. Associação brasileira de educação a distância: conceitos e história no Brasil e no mundo. **Associação Brasileira de Educação a Distância**, São Paulo, [2011]. Disponível em: http://www.abed.org.br/revistacientifica/Revista_PDF_Doc/2011/Artigo_07.pdf. Acesso em: 17 jan. 2021.

ALPAYDIN, E. **Introduction to machine learning.** 3rd ed. Massachusetts: MIT Press, 2014.

ALVES, M. T. G.; SOARES, J. F. Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. **Educação e Pesquisa**, São Paulo, v. 39, n.1, p.177 – 194, 2013.

ANAND, V. K.; RAHIMAN, S. A.; GEORGE, E. B.; HUDA, A. S. Recursive clustering technique for students' performance evaluation in programming courses. *In:* MAJAN INTERNATIONAL CONFERENCE, 1., 2018, Mascate. **Proceedings** [...]. Mascate: MIC, 2018. Disponível em: https://ieeexplore.ieee.org/abstract/document/8363153/. Acesso em: 16 jan. 2021.

ANASTACIO, B. ***K-means* :** o que é, como funciona, aplicações e exemplo em *python*. São Paulo, [2020]. Disponível em**:** https://medium.com/programadores-ajudando-programadores/k-means-o-que-%C3%A9-como-funciona-aplica%C3%A7%C3%B5es-e-exemplo-em-python-6021df6e2572. Acesso em: 17 jan. 2021.

ALAM, M. M.; MOHIUDDIN, K.; DAS, A. K.; ISLAM, M. K.; KAONAIN, M. S.; ALI, M. H. A reduced feature based neural network approach to classify the category of students. *In*: 2ND INTERNATIONAL CONFERENCE ON INNOVATION IN ARTIFICIAL INTELLIGENCE. **Proceedings** […].Shanghai: ACM, 2018. p. 28-32.

ARVAI, K. **K-means clustering in python:** a practical guide. Alexandria, **[**2020]. Disponível em: https://realpython.com/k-means-clustering-python/#evaluating-clustering-performance-using-advanced-techniques. Acesso em: 18 jan. 2021.

AYDOĞDU, Ç. **3D visualization of k-means clustering**. Ankara, [2018]. Disponível em: https://medium.com/analytics-vidhya/3d-visualization-of-k-means-clustering-47d3d3e82117. Acesso em: 10 fev. 2021.

BAEPLER, P.; MURDOCH, C. J. Academic analytics and data mining in higher education. **International Journal for the Scholarship of Teaching and Learning**, Minnesota, v.4, n. 2, p. 11- 18, 2010.

BAPTISTA, A. C. P. **Avaliação do mestrado multimédia em educação da Universidade de Aveiro**. Dissertação (Mestrado em Educação) - Universidade de Aveiro, Aveiro, 2005.

BRASIL. Presidência da República. Lei nº 9.394, de 20 de dezembro de 1996. **Lei das Diretrizes e Bases da Educação.** Estabelece as diretrizes e bases da educação nacional, Brasília, DF: Câmara dos Deputados, 1996.

CAMPBELL, J. P.; OBLINGER, D. G. Academic analytics: a new tool for a new era. **EDUCAUSE review**, London, v.42, n.4, p. 40-57, 2007. Disponível em:

https://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era. Acesso em: 10 jan. 2021.

CASTLE, N. **What is semi-supervised learning?** São Paulo: Oracle, 2018. Disponível em: https://blogs.oracle.com/datascience/what-is-semi-supervised-learning. Acesso em: 02 jan. 2020.

CHAI, K. E.; GIBSON, D. Predicting the risk of attrition for undergraduate students with time based modelling**.** *In*: INTERNATIONAL ASSOCIATION FOR DEVELOPMENT OF THE INFORMATION SOCIETY, 8., 2015, Greater Dublin. **Proceedings** […]. Greater Dublin, Ireland: IADIS, 2015. Disponível em: https://eric.ed.gov/?id=ED562154. Acesso em: 25 fev. 2021.

CIOLACU, M.; TEHRANI, A. F.; BEER, R.; POPP, H. Education 4.0: fostering student's performance with machine learning methods. *In:* IEEE 24TH INTERNATIONAL SYMPOSIUM FOR DESIGN AND TECHNOLOGY IN ELECTRONIC PACKAGING (SIITME)*,* 1., 2018, Iasi. **Proceedings** […]. Iasi, Romênia: SIITME, 2018, p. 438-443.

COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *In*: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO, 1., Rio de Janeiro. **Anais** [...]. Rio de Janeiro: Sociedade Brasileira de Computação, 2013, p.1-29.

DAMASCENO, L. **Regressão linear?.** Rio de Janeiro, [2020]. Disponível em: https://medium.com/@lauradamaceno/regress%C3%A3o-linear-a7f247c3e29. Acesso em: 18 fev. 2021.

ĐAMBIC, G.; KRAJCAR, M; BELE, D. Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. **International Journal of Digital Technology & Economy**, Zagreb, v.1, n.1, p. 1-11, 2016.

DANIEL, B. K. **Big data and learning analytics in higher education:** current theory and practice. Zurique: Springer, 2016. *E-book*.

DATA FLAIR. **Train and test set in python machine learning :** how to split. Nova Delhi, [2018]. Disponível em: https://data-flair.training/blogs/train-test-set-in-python-ml/. Acesso em: 20 set. 2020.

DEJESUS, J. **Data science applications for schools**. Alabama, [2017]. Disponível em: https://towardsdatascience.com/data-science-applications-for-schools-d8913d21d363. Acesso em: 10 set. 2020.

DELEN, D. A comparative analysis of machine learning techniques for student retention management. **Decision Support Systems**, Storrs, v.49, n.4, p. 498-506, 2010.

DHAR, V. Data science and prediction. **Communications of the ACM**, New York, v.56, n.12, p. 64 – 73, 2013.

DIAS, E.; PINTO, F. A. Educação e a covid-19. **Ensaio**: avaliação e políticas públicas em educação, Rio de Janeiro, v.28, n.108, p.545 -554, 2020.

DONI, M. V. **Análise de cluster:** métodos hierárquicos e de particionamento. 2004. Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie, São Paulo, 2004. Disponível em: http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF. Acesso em: 03 mar. 2021.

FARIA, J. C. **Notas de aulas expandidas**. Ilhéus: UESC, 2006.

FRANTZ, G. L.; KING, J. The distance education learning model (DEL). **Educational Technology**, Lincoln, v.1, n.1, p. 33-39, 2000.

GARETH, J. **Uma introdução à estatística:** aprendizagem com aplicações em r. Nova York : Springer, 2013. *E-book.*

GE, Y.; WU, H. Prediction of corn price fluctuation based on multiple linear regression analysis model under big data**. Neural Computing and Applications**, Seattle, v. 32, n.13, p. 16843–16855, 2020. Disponível em: https://www.researchgate.net/publication/330644308_Prediction_of_corn_price_fluctuation_based_on_multiple_linear_regression_analysis_model_under_big_data/citation/download. Acesso em: 19 fev. 2021.

GHAZARIAN, P.; KWON, S. The future of american education : trends, strategies, & realities. **Philosophy of Education Archive**, Oxford, v. 56, n. 18, p. 147-177, 2015.

GONÇALVES, E.; JESUS, A. **Ensino e aprendizagem na perspectiva histórico-crítica:** algumas reflexões. Curitiba, [2010]. Disponível em: http://www.diaadiaeducacao.pr.gov.br/portals/cadernospde/pdebusca/producoes_pde/2010/2010_uel_ped_artigo_edilene_brancalhao.pdf. Acesso em: 8 set. 2020.

GONZALEZ, M. **Crisp-dm na prática**. São Paulo, 2019. Disponível em: https://medium.com/matgonz/crisp-dm-na-pr%C3%A1tica-65be0ee92ada. Acesso em: 22 fev. 2021.

GUTIERREZ, D. **Machine learning and data science**: an introduction to statistical learning methods with r. London: Technics Publications, 2015. *E-book.*

HAIDER, M. **Getting started with data science**. Armonk: IBM press, 2015. *E-book.*

JAIN, K. **Data clustering:** 50 years beyond k-means. Michigan: Pattern Recognition Letters, 2010. *E-book*.

JIA, J. W.; MAREBOYANA, M. Predictive models for undergraduate student retention using machine learning algorithms. *In:* AO, S. I.; KIM, H. K.; CASTILLO, O.; CHAN**,** A. H. S.; KATAGIRI**,** H. **Transactions on engineering technologies**, Dordrecht: Springer, 2014. v. 1, cap. 10, p. 315-329.

JOIA, L. A.; LIMA, N. C. C. Fatores críticos de sucesso em treinamentos corporativos a distância via web: evidências empírico-exploratórias a partir de um estudo de caso. *In*: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 31., 2007, Rio de Janeiro. **Anais** [...]. Rio de Janeiro, RJ: ANPAD, 2007. p. 1-16.

JUST DIGITAL. **Três vantagens do machine learning e o deep learning para empresas**. São Paulo, 2017. Disponível em: https://blog.justdigital.com.br/3-vantagens-do-machine-learning-e-o-deep-learning-para-empresas/. Acesso em: 07 jan. 2021.

KENSKI, V.M. Aprendizagem mediada pela tecnologia. **Revista diálogo educacional**, Curitiba, v. 4, n.10, p.47-56, 2003.

KNIGHT, S.; LITTLETON, K. Dialogue as data in learning analytics for productive educational dialogue. **Journal of learning analytics**, Beaumont, v. 2, n.3, p.111-143, 2016.

KOHAVI, R.; PROVOST, F. Glossary of terms. **Journal of machine learning**, London, v. 30, n. 2, p. 271 - 274, 1998.

LACRUZ, A. J.; AMÉRICO, L. B. The school environment and its performance: analysis of the scores obtained by schools from the state of Espírito Santo in the nationwide exam prova Brasil, using multiple linear regression. **Revista de Administração Pública**, Rio de Janeiro, v. 51, n. 5, p. 854-878, 2017.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data**: an introduction to data mining. 2nd. ed. Hoboken: John Wiley & Sons, 2014.

LAVAGNOLI**,** S. **Machine learning ou deep learning?.** São Paulo, [2019]. Disponível em: https://opencadd.com.br/machine-learning-ou-deep-learning/. Acesso em: 20 fev.2020.

LINDEN, R. **Um algoritmo híbrido para extração de conhecimento em bioinformática**, Tese (Doutorado em Engenharia Elétrica) – COPPE, UFRJ, Rio de Janeiro, 2005.

LIU, M.; HUANG, Y. The use of data science for education: the case of social-emotional learning. **Smart Learning Environments**, Dordrecht, v. 4, n. 1, p. 2 – 13, 2017.

LOBO, A. S. M.; MAIA, L. C. G. O uso das TICs como ferramenta de ensino-aprendizagem no ensino superior. **Caderno de Geografia**, Ponte Nova, v.25, n.44, p. 16 – 26, 2015.

LOHR, S. **The age of big data**. New York, 2012. Disponível em: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html. Acesso em: 14 set. 2020.

LYKOURENTZOU, I.; GIANNOUKOS, I.; NIKOLOPOULOS, V.; MPARDIS, G.; LOUMOS, V. Dropout prediction in elearning courses through the combination of machine learning techniques. **Journal Computers & Education**, New York, v. 53, n. 3, p.950-965, 2009.

MALIK, F. **Calculus multivariate calculus and machine learning**: a must-know concept for every professional. California, 2019. Disponível em: https://medium.com/fintechexplained/calculus-multivariate-calculus-and-machine-learning-242b9efcb41c**.** Acesso em: 21 dez.2020.

MANDIC, C. L. **Uma abordagem de machine learning para o log analytics**. São Paulo, 2020. Disponível em: https://blog.mandic.com.br/artigos/uma-abordagem-de-machine-learning-para-o-log-analytics/. Acesso em: 17 fev.2021

MARTINEZ-MALDONADO, R.; SCHNEIDER, B.; CHARLEER, S.; SHUM, S. B.; KLERKX, J.; DUVAL, E. Interactive surfaces and learning analytics: data, orchestration aspects, pedagogical uses and challenges. *In*: SIXTH INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS & KNOWLEDGE, 9., 2016, Edinburgh. **Proceedings** […].Edinburgh: ACM, 2016. p. 124-133.

MATTA, A. E. R. Comunidades em rede de computadores: abordagem para a educação a distância: EAD acessível a todos**. Revista Brasileira de Aprendizagem Aberta e a Distância**. São Paulo, v.1, n. 1, p. 1 – 9, 2003. Disponível em: http://www.abed.org.br/revistacientifica/Revista_PDF_Doc/2003_Comunidades_Rede_Computadores_Alfredo_Matta.pdf. Acesso em: 21 nov. 2020.

MAZULO, E. S. Análise da proficiência em matemática por meio de regressão linear múltipla. **Revista Intersaberes**, São Paulo, v. 10, n. 21, p. 613 – 625, 2015.

MELTON, J.; SIMON, A. R. M. **Understanding the new SQL**: a complete guide. California*:* Morgan Kaufmann, 1993.

MIKUT, R.; REISCHL, M. Data mining tools. **WIREs Data Mining and Knowledge Discovery**, New York, v.1, n.5, p.431 – 443, 2011.

MILLER, T. W. **Modeling techniques in predictive analytics with python and r**: a guide to data science. Upper Saddle River: Pearson FT Press, 2015.

MILLER, S.; HUGHES, D. **The quant crunch**: how the demand for data science skills is disrupting the job market. Armonk, 2017. Disponível em: https://www.ibm.com/downloads/cas/3RL3VXGA. Acesso em: 12 set. 2020.

MITCHELL, T. M. Machine learning and data mining. **Communications of the ACM**, New York, v. 42, n. 11, p. 42 – 48, 1999.

MOORE, M. G; KEARSLEY, G. **Educação a distância:** uma visão integrada. São Paulo: Cengage Learning, 2008.

MORRIS, L.; FINNEGAN, C.; WU, S. Tracking student behavior, persistence and achievement in online courses. **The Internet and Higher Education**, London, v.8 ,n.3, p. 221-231, 2005.

MOHAMMED, M.; KHAN, M. B.; BASHIER, E.B.M. **Machine learning**: algorithms and applications. Boca Raton: CRC Press, 2017.

NEUROTECH. **Quais as diferenças entre inteligência artificial, machine learning e deep learning?.** São Paulo, 2020. Disponível em: https://www.neurotech.com.br/quais-as-diferencas-entre-inteligencia-artificial-machine-learning-e-deep-learning/. Acesso em: 18 fev. 2021.

NUZZO, R. **P-values, the "gold standard" of statistical validity, are not as reliable as many scientists assume.** Nature, 2014. Disponível em: https://institutoisaia.com.br/unidade-pesquisa-clinica/pdf/2014/06-08-2014-Discussao-do-artigo-Statistical-errors-P-values-the-gold-standard-of-statistical-validity-are-not-as-reliable-as-many-scientists-assume-Nuzzo-R.pdf. Acesso em: 18 fev.2021.

OLAVSRUD, T. **What is data science? transforming data into value**. Florida, 2019. Disponível em: https://www.cio.com/article/3285108/what-is-data-science-a-method-for-turning-data-into-value.html#. Acesso em: 12 set. 2020

OLIVEIRA, B. **Qual a relação entre machine learning e a estatística?.** Belo Horizonte, 2020. Disponível em: https://operdata.com.br/blog/a-relacao-entre-machine-learning-e-a-estatistica/#:~:text=O%20Machine%20Learning%20%C3%A9%20um,a%20cria%C3%A7%C3%A3o%20de%20modelos%20anal%C3%ADticos.&text=A%20ideia%20b%C3%A1sica%20do%20Machine,que%20sejam%20reprogramadas%20para%20isso. Acesso em: 19 fev. 2021.

ONDITI, A. A. Relationship between customer personality, service features and customer loyalty in the banking sector: a survey of banks in Homabay County, Kenya. **International Journal of Business and Social Science**, New York, v.4, n.15, p.132 – 150, 2013.

PAESE, C. R. Educação a distância (ead) e o uso das tecnologias de informação e comunicação (tics), baseada em ambientes virtuais de aprendizagem (ava) : Algumas reflexões sobre a importância da tutoria on-line. **Itinerarius Reflectionis**, Jataí , v. 8, n. 1, p. 1- 21, 2012. Disponível em: https://www.revistas.ufg.br/rir/article/view/20377. Acesso em: 3 fev. 2021.

PAPAMITSIOU, Z. K.; ECONOMIDES, A. A. Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. **Educational Technology & Society**, New York, v.17, n. 4, p. 49-64, 2004.

PRECHELT, L.; ORR, G. B. Parada antecipada - mas quando?. *In* : MONTAVON, G.; MULLER, K. R.(ed.). **Redes neurais:** truques do comércio. Heidelberg: Springer, 2012. p. 53-67. Disponível em*:* https://link.springer.com/chapter/10.1007%2F978-3-642-35289-8_5. Acesso em: 14 mar. 2021.

PURI, G. Critical success factors in e-learning: an empirical study. **International Journal of Multidisciplinary Research**, Haryana, v.*2*, n.1, p. 149 – 161, 2012.

QUIGLEY, D.; OSTWALD, J.; SUMNER, T. Scientific modeling: using learning analytics to examine student practices and classroom variation. *In*: INTERNATIONAL LEARNING ANALYTICS & KNOWLEDGE CONFERENCE, 11., 2017, Vancouver. **Proceedings** […]. Vancouver: ACM, 2017. p. 329 -338.

RAM, S.; WANG, Y.; CURRIM, F.; CURRIM, S. **Using big data for predicting freshmen retention**, Alabama: FD Press, 2015.

RIPLEY, B. D. **Pattern recognition and neural networks.** Cambridge: Cambridge University Press, 1996.

ROLLINS, J. **Foundational methodology for data science.** Armonk: IBM Analytics, 2015.

ROMERO, C.; VENTURA, S. Educational data mining: a survey from 1995 to 2005. **Expert systems with applications**, New York*,* v. 33, n. 1, p.135-146, 2007.

SALLES, R. **Correlação**: direto ao ponto. São Paulo, 2018. Disponível em: https://medium.com/brdata/correla%C3%A7%C3%A3o-direto-ao-ponto-9ec1d48735fb. Acesso em: 04 jan. 2021.

SAMMUT, C.; WEBB, G. I. **Mean absolute error**. Encyclopedia of Machine Learning. Boston: Springer, 2010.

SAS. **Inteligência artificial:** O que é e qual sua importância?. São Paulo, 2020. Disponível em: https://www.sas.com/pt_br/insights/analytics/inteligencia-artificial.html. Acesso em: 12 fev. 2021.

SHALEV-SCHWARTZ, S.; BEN-DAVID, S. **Understanding machine learning:** from theory to algorithms. Cambridge: Cambridge University Press, 2013.

SIEMENS, G.; BAKER, R. S. Learning analytics and educational data mining: towards communication and collaboration. *In*: INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS AND KNOWLEDGE, 2., 2012, New York. **Proceedings** […]. New York: ACM, 2012. p. 252-254.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados:** com aplicações em r. Rio de Janeiro: Elsevier, 2016.

SIMON, T. M. R.; GIROLAMI, M. A first course in machine learning. **Stat Papers**, Glasglow, v. 56, n. 1, p. 271 – 271, 2013.

SONI, D. **Supervised vs. unsupervised learning**. Kansas, 2018. Disponível em: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d. Acesso em: 02 out. 2020.

SOUZA, R.; NETO, F. M.; SANTOS, A.; FONTES, L.; NAASSON, E.; VALENTIM, R. Um ambiente inteligente de avaliação de comportamentos de tutores e turmas no ambiente virtual de aprendizagem moodle. *In :* WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 11., 2016, Uberlândia. **Anais** [...]. Uberlândia, MG : WCBIE, 2016. p.764 – 773.

TESTA, M. G.; FREITAS, H. M. R. Fatores importantes na gestão de programas de educação a distância via internet: a visão dos especialistas. *In*: ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO, 1,. 2002, Salvador. **Anais** [...]. Salvador: ANPAD, 2002. Disponível em: https://www.scielo.br/scielo.php?script=sci_nlinks&ref=000127&pid=S1679-39512007000500010000022&lng=es. Acesso em 12 dez. 2020.

UNESCO. **Global education monitoring report**. Toronto: GEM Report, 2020. *E-book.*

VETTORI, M.; ZARO, M. A. Avaliação do socrative app como ferramenta auxiliar de ensino para a construção de aprendizagens significativas em uma disciplina de física geral a partir do peer instruction. *In* : SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO,9., 2016. **Anais** [...]. Porto Alegre: SBIE, 2016. Disponível em: https://www.br-ie.org/pub/index.php/sbie/article/view/6699/4587. Acesso em 17 fev. 2021.

VIGOTSKY, L. **Pensamento e linguagem**. São Paulo: Martins Fontes, 1998.

WU, J. Y.; HSIAO, Y. C.; NIAN, M. W. Using supervised machine learning on large-scale online forums to classify course-related facebook messages in predicting learning achievement within the personal learning environment**. Interactive Learning Environments**, London, v. 28, n. 1, p. 65 -80, 2018.

WUJEK, B.; HALL, P.; GUNES, F. **Best practices for machine learning applications.** Cary: SAS Institute Inc., 2016.

ANEXO A – Características dos principais algoritmos de *Machine Learning*

Tipo de Algoritmo	Aplicações comuns	Aplicações sugeridas	Escala dos dados	Capacidade de interpretação
Regressão penalizada	− Regressão supervisionada − Classificação supervisionada	− Especificação manual não-linear e termos de interação explícitos.	Conjunto de dados pequeno à grande	Alta
Naive Bayes	− Classificação supervisionada	− Modelo linear ou fenômeno linearmente separável. − Adequado para conjuntos de dados extremamente grandes, nos quais métodos complexos são intratáveis.	Conjunto de dados pequeno à extremamente grande	Moderada
Árvore de decisão	− Regressão supervisionada − Classificação supervisionada	− Modelo de fenômenos não linearmente separáveis em grande quantidade de dados não trabalhados. − Interações automáticas implícitas. − Valores faltantes e *outliers* nas variáveis de entrada tratados automaticamente. − Conjuntos de árvore de decisão (exemplo: florestas aleatórias, aumento de gradiente) podem aumentar a acurácia da predição e diminuir o sobre ajuste, no entanto, também diminui a escalabilidade e capacidade de interpretação.	Conjunto de dados médio à grande	Moderada
K-ésimo Vizinhos mais próximos (kNN)	− Regressão supervisionada − Classificação supervisionada	− Modelo de fenômenos não linearmente separáveis. − Pode ser usado para combinar à precisão de técnicas mais sofisticadas, mas com	Conjunto de dados pequeno à médio	Baixa

		menos parâmetros de ajuste.		
Máquina de Vetores de Suporte	− Regressão supervisionada − Classificação supervisionada − Detecção de anomalias	− Modelo linear ou fenômeno linearmente separável usando núcleos lineares. − Modelo de fenômenos não linearmente separáveis usando núcleos não-lineares. − Detecção de anomalias utilizando uma classe de máquina de vetores de suporte.	Conjunto de dados pequeno à grande usando núcleos lineares. Conjunto de dados de médio à grande usando núcleos nãolineares.	Baixa
Rede neural artificial	− Regressão supervisionada − Classificação supervisionada − Clustering não supervisionado − Extração de característica não supervisionado − Detecção de anomalias	− Modelo de fenômenos não linearmente separáveis. − Rede neurais profundas (exemplo: deep learning) para reconhecimento de padrão de imagens, vídeos e sons. − Todas as interações consideradas em topologias multicamadas totalmente conectadas. − Extração de elementos não lineares com e redes de máquinas Boltzmann restritas. − Armazenamento em cluster e visualização com mapas auto-organizados − Detecção de anomalias com redes.	Conjunto de dados pequeno à médio	Baixa
Regras de associação	− Construção de regra supervisionada. − Construção de regra não supervisionada.	− Construir um conjunto de regras complexas usando a simultaneidade de itens ou eventos em conjuntos de dados transacionais.	Conjunto de dados transacionais médio à grande	Moderada

K-Significados	− Agrupamento não supervisionado	− Criação de um número conhecido a priori de clusters esféricos, disjuntos e de tamanho igual. − O método k-mode pode ser usado para dados categóricos. − O método	Conjunto de dados transacionais médio à grande	Moderada

		k-prototypes pode ser usado para dados mistos.		
Máquinas de fatoração	− Regressão e classificação supervisionada − Extração de características não supervisionada	− Extração de um número conhecido a priori de características não interpretáveis, oblíquas de conjuntos de dados sparse e transacionais. − Pode explicar automaticamente interações variáveis. − Criar modelos a partir de um grande número de características sparse. podendo superar as Máquinas de Vetores de Suporte.	Conjunto de dados sparse ou transacionais médio à extremamente grande	Moderada
Cluster hierárquico	− Agrupamento não supervisionado	− Criação de um número conhecido a priori de clusters não esféricos, disjuntos ou sobre ajustes de clusters de tamanhos diferentes.	Conjunto de dados pequeno	Moderada
Cluster Espectral	− Agrupamento não supervisionado	− Criação de um número dependente de dados de clusters arbitrariamente modelados, disjuntos ou sobrepostos de tamanhos diferentes.	Conjunto de dados pequeno	Moderada
Análise de componentes principais	− Extração de características não supervisionada	− Extração de um número dependente de dados de recursos lineares ortogonais. − A Decomposição de valor singular é frequentemente usada em vez de Análise de componentes principais em dados amplos. − A Análise de componentes principais Sparse pode ser usada para criar características mais interpretáveis, mas a ortogonalidade é perdida. − A Análise de componentes principais Kernel pode ser usado para extrair	Conjunto de dados pequeno à grande para Análise de componentes principais tradicional e Decomposição de valor singular. Conjunto de dados pequeno à médio para Análise de componentes principais Sparse e Kernel.	Normalmente baixa

		características não lineares.		
Fatoração de matriz não negativa	− Extração de características não supervisionada	− Extração de um número conhecido a priori de características interpretáveis, lineares, oblíquas e não negativas.	Conjunto de dados pequeno à grande	Alta
Projeções aleatórias	− Extração de características não supervisionada	− Extração de um número dependente de características lineares, não interpretáveis, orientados arbitrariamente de igual importância	Conjunto de dados médio à extremamente grandes	Baixa

Fonte: Características dos principais algoritmos de Machine Learning

Fonte: Adaptado de Wujet, Hall e Gunes (2016).