# Previsão da colocação final e dos times vencedores de rodadas do Campeonato Brasileiro de Futebol Masculino da série A com aprendizado de máquina

Marcos de Pinho Tavares Proença

***Resumo*—Este artigo apresenta um estudo a respeito da utilização de aprendizado de máquina e de outras técnicas computacionais de análise de dados no contexto do futebol. O objetivo é utilizar os conhecimentos adquiridos na disciplina Introdução ao aprendizado de máquina e à mineração de dados (PO450) oferecida pela universidade Unicamp-sp para construir modelos de machine learning que tentem prever com considerável acurácia os times vencedores de rodadas do Campeonato Brasileiro de Futebol Masculino, assim como a colocação final deste torneio de futebol.**

***Palavras-Chave*—*Aprendizado de máquina, futebol, previsão, análise de dados, Campeonato Brasileiro de Futebol Masculino.***

***Abstract*—This article shows the usage of machine learning and other computational data analysis techniques in the context of soccer. The objective is to utilize the knowledge acquired in the subject Introdução ao aprendizado de máquina e à mineração de dados (PO450) offered by the college Unicamp-sp to build models of machine learning that try to predict with a good accuracy the winning teams of each round of the Campeonato Brasileiro de Futebol Masculino, as the final placement of this soccer tournament.**

***Keywords*—*Machine learning, soccer, predict, data analysis, Campeonato Brasileiro de Futebol Masculino.***

## I. Introdução

O futebol é um esporte secular mundialmente estabelecido e adorado por milhares de pessoas. No Brasil, o país do futebol, como é rotulado por ter muitas conquistas e por revelar muitos jogadores talentosos, isso não é diferente, sendo o futebol um importante fator que movimenta a economia brasileira. Dessa forma, de acordo com a Confederação Brasileira de Futebol (CBF), entidade responsável por organizar e regulamentar campeonatos nacionais, o futebol em 2019 representou 0,79% do PIB nacional, o que corresponde a um valor de cerca de R$53 bilhões. Portanto, esse dado mostra o grande mercado que é o futebol no Brasil, assim como o valor deste trabalho para, por exemplo, sites de apostas de futebol e amantes do esporte.

No Brasil, os campeonatos de futebol são muito disputados e ao contrário de alguns países no mundo, onde há uma grande hegemonia de poucos times com mais recursos financeiros, como na Alemanha atualmente onde o Bayern München é campeão há nove anos consecutivos, o Campeonato Brasileiro de Futebol se mostra um torneio com muitos candidatos ao título, com sete campeões diferentes nos últimos dezoito anos, desde o estabelecimento da modalidade dos pontos corridos.

Para contextualizar melhor, o Campeonato Brasileiro de Futebol Masculino (Brasileirão) é um torneio anual onde vinte times de futebol de todo o país disputam o posto de campeão por meio de 38 rodadas, onde cada time joga duas vezes com o seu adversário, uma vez em casa e outra vez como visitante na casa do adversário. Logo, cada partida vale três pontos para o time vitorioso, um ponto em caso de empate e zero pontos para o time perdedor, sendo campeão o time que obtiver a maior pontuação. Em caso de empate na pontuação final alguns critérios de desempate são adotados como, maior número de vitórias e maior saldo de gols.

Além disso, é importante explicar, pois foi a principal abordagem computacional utilizada para realização deste projeto, que aprendizado de máquina, do inglês machine learning, é uma técnica que visa treinar a máquina (computador) para reconhecer padrões de forma inteligente e automática. Para realizar esse reconhecimento de padrões é preciso fornecer ao algoritmo uma base de dados com atributos que não sejam, preferencialmente, enviesados e que serão usados pelo modelo para chegar em um resultado. Dessa forma, os principais métodos para se ensinar a máquina são: o aprendizado supervisionado que ocorre quando o modelo aprender a partir de resultados conhecidos pré-estabelecidos, o aprendizado não supervisionado, quando não são fornecidas respostas ao algoritmo o obrigando a aprender sozinho e o aprendizado por reforço quando o programa deve alcançar determinado objetivo interagindo com o ambiente. Logo, com o objetivo de explorar essa técnica foram construídos os códigos deste trabalho.

Diante dos aspectos apresentados, neste trabalho acadêmico foram utilizadas ferramentas de aprendizado de máquina supervisionado para tentar realizar a difícil tarefa de prever resultados de rodadas do Campeonato Brasileiro de Futebol Masculino, por meio de uma classificação de um rótulo discreto no caso vitória, derrota e empate, assim como a colocação final através de uma regressão que tenta prever uma quantidade contínua, no caso a pontuação dos times.

## II. Metodologia

Para realizar este trabalho foram construídos dois códigos na linguagem de programação python, com o intuito de comparar os resultados, por meio de duas abordagens de aprendizado de máquina diferentes, regressão e classificação.

Vale ressaltar que os dados foram obtidos por meio da técnica do *web scraping* que coleta dados de sites da internet e os transforma em informação estruturada com a ajuda da biblioteca do python pandas para uma posterior manipulação. No caso, os dados do futebol brasileiro de 2013-2021 foram extraídos do site worldfootball que contém informações das principais ligas de futebol do mundo.

No código das rodadas, como foi denominado o código que tenta prever os resultados dos jogos dos times mandantes,

utilizou-se o aprendizado de máquina para classificar as partidas de futebol que foram divididas em variáveis categóricas discretas: 2 correspondente à vitória, 0 correspondente à derrota e 1 correspondente à empate. Neste caso, após tratar os dados coletados, que continham os horários dos jogos, os times mandantes e visitantes e o placar do jogo, foram construídos os atributos preditores (variáveis independentes ou explicativas) já que eles não eram encontrados prontos no *dataset*. Dessa forma, os atributos foram definidos para o time mandante e para o time visitante, sendo eles: aproveitamento, a pontuação atual do time dividida pela pontuação máxima possível; taxa de gols, a divisão de gols marcados por gols sofridos; saldo de gols, a quantidade de gols marcados menos a quantidade de gols sofridos; derrotas, o número de derrotas acumulado do time até a rodada e na temporada do jogo, vitórias e empates seguem a mesma lógica do atributo derrotas; ranking corresponde à classificação do time de acordo com a CBF na respectiva temporada, sendo essa classificação construída com base em alguns critérios, como colocação no campeonato brasileiro, fase alcançada na Copa do Brasil, dentre outros com o objetivo de definir os participantes da série D do Brasileirão e da Copa do Brasil; valor do time corresponde ao somatório do valor de mercado dos jogadores que compõem o elenco do time em cada temporada de acordo com o site de análises Transfermarkt.

Os atributos que mais se destacaram neste código foram o aproveitamento do mandante e o aproveitamento do visitante, com respectivamente, 266.69 e 200.29 de F Score. Essa medida, F Score, mostra, de forma resumida, o quão informativo cada atributo é em relação à classe (variável dependente ou resposta), ou seja, a resposta dos dados (vitória, derrota ou empate) usada como base para treinar e tentar prever os resultados. Logo, quanto maior o F Score do atributo, mais impacto ele terá no treinamento do modelo.

Diante dos aspectos apresentados, o próximo procedimento feito para construção do modelo foi normalizar os dados para deixar os atributos com grandezas similares e não atrapalhar o treinamento, posteriormente realizou-se uma técnica conhecida como PCA que selecionou os atributos normalizados que explicam 99% da variância dos dados. Essa técnica busca reduzir as dimensões de um conjunto de dados, diminuindo a quantidade de atributos e agrupando os atributos que explicam pouco o modelo para tentar preservar as propriedades do conjunto e facilitar o treinamento dos algoritmos. Logo em seguida, dividiu-se os anos de 2013 até 2019 para treinamento e o ano de 2020 para teste, ou seja, um total de 3040 amostras, já que em cada ano são realizados 380 jogos e foram usados 8 anos. É importante destinar a maior parte dos dados para treinamento para contribuir com um modelo mais otimizado, assim como é importante uma boa quantidade de dados para avaliar a acurácia do modelo. Então, essa divisão dos dados foi passada para os modelos que no caso do código das rodadas foram a Regressão logística, o Random Forest com trezentas árvores e a Árvore de decisão que foram treinados para prever o conjunto dos testes.

Para contextualizar a respeito dos modelos, a Regressão logística é uma técnica utilizada para tentar prever variáveis categóricas que consiste em, por meio da avaliação da probabilidade de se encontrar determinada categoria da classe, obter a probabilidade de ocorrência de um evento e a influência dos atributos no evento de estudo. Já a técnica da Árvore de decisão ou Decision tree consiste em criar uma estrutura parecida com um fluxograma com regras para tomada de decisão que percorre os nós dos ramos da árvore para encontrar o melhor resultado. Por fim, a técnica do Random Forest, em português floresta aleatória, cria várias Árvores de decisão que escolhem as variáveis de forma aleatória para tentar melhorar o resultado, já que quanto mais árvores forem criadas, melhor serão os resultados encontrados até determinado limite onde as melhorias não serão tão significativas, porém maior será o tempo de criação do modelo. Vale salientar que, a divisão dos dados, assim como esses modelos, foram feitos com a ajuda da biblioteca scikit-learn.

O último procedimento feito no código das rodadas foi montar a tabela do ano de 2020 para os modelos de acordo com a previsão dos modelos. Como a previsão é dada em rótulos discretos, foi preciso convertê-los para a pontuação de cada time e para isso bastou multiplicar as vitórias por 3 e os empates por 1, já as derrotas não contam como pontuação.

No código das tabelas, o outro código criado, para tentar prever a pontuação final e consequentemente a colocação dos times na tabela, utilizou-se o *machine learning* para efetuar uma regressão da pontuação. O procedimento realizado nesta parte foi bem semelhante ao feito no código das rodadas, contudo o *dataset* extraído do site continha o time, a posição do time, o número de partidas, de vitórias, derrotas e empates, os gols feitos e sofridos e a pontuação dos anos de 2013 até 2020 com um total de 6080 amostras (20 times em uma tabela vezes 38 rodadas vezes 8 anos). A partir do *dataset* construiu-se os atributos, sendo eles: gols marcados e gols sofridos pelo time até a rodada em questão; saldo de gols, gols marcados menos gols sofridos; aproveitamento, a pontuação atual do time dividida pela pontuação máxima possível até a rodada em questão; taxa de gols, a quantidade de gols marcados dividida pela quantidade de gols sofridos; o valor do time em milhões e o ranking da CBF, feitos da mesma forma que no código das rodadas. Logo, os atributos que obtiveram maiores F Scores foram, os gols marcados (733.62) e gols sofridos com (80.30).

O próximo passo foi começar a parte do *machine learning*. Para isso, normalizou-se os atributos e a pontuação dos times foi definida como classe. Posteriormente os dados foram divididos de 2013 até 2018 para treinamento e 2019 para teste dos modelos. É importante salientar que para tentar evitar e verificar se houve *overfitting*, quando o modelo se ajusta muito bem ao *dataset*, os dados do ano de 2020 não foram utilizados para teste ou treinamento para servirem como dados de validação, ou seja, dados que o modelo nunca viu. Por fim, esses dados foram atribuídos aos seguintes modelos, Regressão linear, Árvore de decisão, Random Forest com cem árvores e XGBoost para se tentar obter o modelo com menor taxa de erro.

Para entender melhor os modelos, a Regressão linear é uma técnica que busca estabelecer uma relação entre os dados a partir de uma reta, que considera as variáveis dependentes e independentes do modelo e cria uma reta a partir delas. Já a Árvore de decisão, assim como o algoritmo Random Forest foram explicados anteriormente e são técnicas poderosas que podem ser aplicadas tanto para problemas de classificação quanto para problemas de regressão, em que geralmente o Random Forest apresenta uma maior acurácia, pois ele utiliza várias árvores de decisão. O último algoritmo utilizado, o XGBoost, assim como o Random Forest é um algoritmo de *ensemble*, logo utiliza modelos mais simples para formar um algoritmo robusto e com melhores resultados. Dessa forma, o XGBoost (*Extreme Gradient Boosting*) se baseia na técnica

*Gradient Boosting,* um algoritmo que utiliza um agrupamento de modelos de predição mais fracos, para conseguir resultados superiores, ajustando o modelo a partir da minimização do erro de ajuste por meio da técnica de gradiente descendente.

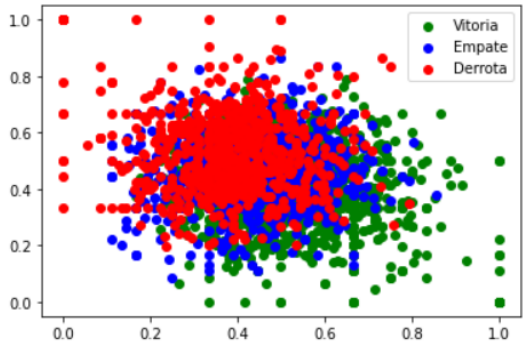
Para finalizar o código das tabelas, foi feita a previsão da tabela final do Campeonato Brasileiro de futebol de 2020 e para isso foi dada como entrada para os modelos uma tabela de determinada rodada desta edição do torneio e a partir dela os resultados foram obtidos de forma satisfatória, levando em consideração a dificuldade de se prever resultados relacionados à futebol pela imprevisibilidade do esporte. Vale observar que este trabalho foi realizado após o ano de 2020, no primeiro semestre de 2021, logo foi possível verificar a acurácia dos modelos de forma melhor. Além disso, tentou-se prever quem seria o campeão do campeonato de 2021, utilizando os dados referentes da oitava rodada, com a ressalva de que alguns times apresentavam jogos a menos e isso pode alterar o resultado da previsão.

Portanto, a construção dos modelos, tanto no código das rodadas, assim como no código das tabelas se mostrou bem sucedido, sendo possível aplicar grande parte dos conhecimentos adquiridos na disciplina.

## III. Resultados

Os modelos de aprendizado de máquina do código das rodadas obtiveram acurácia de teste relativamente boa, levando em consideração que se trata de um problema multiclasse (três rótulos) e que o futebol é um esporte difícil de se prever os resultados, como é possível perceber pelo gráfico de dispersão:

Figura 1 - gráfico de dispersão com eixo x correspondente ao atributo aproveitamento do mandante e eixo y aproveitamento do visitante normalizados.




Fonte: autoria própria.

Pelo gráfico da Figura 1 é possível perceber o quanto os dados se misturam e a complexidade do problema, já que nem sempre, para este caso, o time com melhor aproveitamento no campeonato consegue vencer o time com menor aproveitamento. Para visualizar melhor essa problemática, a seguir se encontra uma tabela com os modelos usados e as respectivas acurácias deles:

Tabela 1 - acurácia dos modelos do código das rodadas.

| Modelo | Acurácia (teste) |
|---|---|
| Regressão logística | 57,6% |
| Random Forest | 57,9% |
| Árvore de decisão | 51,3% |

Fonte: autoria própria.

A partir da tabela, percebe-se que o modelo que obteve a melhor acurácia foi o Random Forest com 57,9% e, como esperado, com uma acurácia maior que a técnica da Árvore de decisão. Além disso, o índice de acerto dos algoritmos, não foi tão alto como já era esperado, sendo que a Regressão logística e o Random Forest tiveram valores de acurácia próximos, mas a técnica da Árvore de decisão ficou com uma taxa de acerto um pouco distante em relação aos outros modelos.

No código das tabelas foram encontrados valores para os testes relativamente altos de $R^2$, ou seja, o coeficiente de determinação, que representa a proporção da variância explicada pelos atributos do modelo (variáveis independentes) e ademais baixos RMSEs (*root mean squared error*) que são o erro quadrático médio do modelo e medem a diferença entre os valores observados e os valores obtidos pelo modelo, sendo RMSEs baixos considerados bons. Esses resultados para os modelos utilizados se encontram na Tabela 4:

Tabela 2 - métricas $R^2$ e RMSE dos modelos do código das tabelas.

| Modelo | $R^2$ (teste) | RMSE (teste) |
|---|---|---|
| Regressão linear | 91,4% | 4,95 |
| Árvore de decisão | 91,3% | 5,29 |
| Random Forest | 94,3% | 4,27 |
| XGBoost | 93,7% | 4,49 |

Fonte: autoria própria.

Como é possível perceber pela tabela, o algoritmo que obteve os melhores resultados, tanto por ter o maior coeficiente de determinação, quanto por ter o menor erro quadrático médio foi, assim como no código das rodadas, o Random Forest, o que demonstra o poder dessa técnica. Além disso, observa-se que os quatro modelos utilizados obtiveram as métricas relativamente próximas, com poucas diferenças.

A seguir se encontra uma comparação entre a tabela final do campeonato de 2020 prevista pelo código das tabelas à esquerda, utilizando como entrada a rodada trinta e dois e à direito a previsão do código das rodadas.

Figura 2 - times e respectivas posições e pontuações de acordo com as previsões dos códigos do trabalho (ano 2020).

**Random Forest**

| Posicao | Time | Pontos |
|---|---|---|
| 1 | Flamengo RJ | 82 |
| 2 | Palmeiras | 73 |
| 3 | Atlético Mineiro | 67 |
| 4 | São Paulo FC | 66 |
| 5 | Santos FC | 56 |
| 6 | Internacional | 53 |
| 7 | Grêmio Porto Alegre | 53 |
| 8 | Fluminense RJ | 53 |
| 9 | Ceará - CE | 49 |
| 10 | Corinthians SP | 49 |
| 11 | Fortaleza | 44 |
| 12 | Sport - PE | 43 |
| 13 | Athletico Paranaense | 38 |
| 14 | Botafogo - RJ | 35 |
| 15 | Atlético Goianiense | 30 |
| 16 | Bahia - BA | 27 |
| 17 | Red Bull Bragantino | 23 |
| 18 | Coritiba FC | 23 |
| 19 | Vasco da Gama | 22 |
| 20 | Goiás | 22 |

**Random Forest**

| Posicao | Time | Pontos |
|---|---|---|
| 1 | Internacional | 88 |
| 2 | Atlético Mineiro | 88 |
| 3 | Flamengo RJ | 78 |
| 4 | São Paulo FC | 78 |
| 5 | Santos FC | 64 |
| 6 | Palmeiras | 62 |
| 7 | Fluminense RJ | 62 |
| 8 | Grêmio Porto Alegre | 57 |
| 9 | Corinthians SP | 46 |
| 10 | Ceará - CE | 45 |
| 11 | Red Bull Bragantino | 44 |
| 12 | Atlético Goianiense | 43 |
| 13 | Vasco da Gama | 43 |
| 14 | Athletico Paranaense | 43 |
| 15 | Fortaleza | 42 |
| 16 | Sport - PE | 39 |
| 17 | Bahia - BA | 38 |
| 18 | Coritiba FC | 29 |
| 19 | Goiás | 24 |
| 20 | Botafogo - RJ | 22 |

Fonte: autoria própria.

Diante das tabelas da Figura 2, que foram feitas com base nos modelos com melhores resultados em ambos os códigos, o Random Forest, é possível perceber que, no caso do código das tabelas, foi preciso passar como entrada uma rodada relativamente perto do fim do campeonato para que ele conseguisse prever o campeão do torneio, o Flamengo, e adicionou-se à pontuação um desvio padrão para deixá-la mais perto da realidade. Assim, no geral ele se mostrou satisfatório em prever os melhores colocados e os piores colocados, acertando a posição de alguns times. Já na tabela final do código das rodadas percebe-se algo semelhante ao encontrado no código das tabelas, conseguindo prever relativamente bem os melhores e piores times da edição de 2020 do campeonato brasileiro, mas não acertando o campeão.

Além disso, foi feita a previsão para o brasileirão de 2021, utilizando como entrada a rodada oito (rodada mais atual até a realização deste trabalho) com o código das tabelas. O resultado encontrado se encontra a seguir para os modelos com melhores métricas de avaliação de desempenho:

Figura 3 - times e respectivas posições e pontuações de acordo com as previsões dos códigos do trabalho (ano 2021).

**Random Forest**

| Posicao | Time | Pontos |
|---|---|---|
| 1 | Palmeiras | 78 |
| 2 | Red Bull Bragantino | 76 |
| 3 | Bahia - BA | 68 |
| 4 | Fortaleza | 61 |
| 5 | Athletico Paranaense | 60 |
| 6 | Atlético Mineiro | 48 |
| 7 | Santos FC | 46 |
| 8 | Flamengo RJ | 45 |
| 9 | Internacional | 44 |
| 10 | Ceará - CE | 44 |
| 11 | Juventude - RS | 41 |
| 12 | Fluminense RJ | 38 |
| 13 | Chapecoense | 38 |
| 14 | América - MG | 34 |
| 15 | Corinthians SP | 31 |
| 16 | Atlético Goianiense | 31 |
| 17 | São Paulo FC | 25 |
| 18 | Cuiabá - MT | 24 |
| 19 | Sport - PE | 23 |
| 20 | Grêmio Porto Alegre | 22 |

**XGBoost**

| Posicao | Time | Pontos |
|---|---|---|
| 1 | Palmeiras | 82 |
| 2 | Red Bull Bragantino | 79 |
| 3 | Bahia - BA | 71 |
| 4 | Fortaleza | 64 |
| 5 | Athletico Paranaense | 64 |
| 6 | Atlético Mineiro | 47 |
| 7 | Santos FC | 47 |
| 8 | Internacional | 46 |
| 9 | Ceará - CE | 44 |
| 10 | Flamengo RJ | 43 |
| 11 | Juventude - RS | 41 |
| 12 | Chapecoense | 38 |
| 13 | Fluminense RJ | 36 |
| 14 | Corinthians SP | 32 |
| 15 | América - MG | 32 |
| 16 | Atlético Goianiense | 30 |
| 17 | Cuiabá - MT | 27 |
| 18 | São Paulo FC | 25 |
| 19 | Sport - PE | 23 |
| 20 | Grêmio Porto Alegre | 23 |

Fonte: autoria própria.

Por meio dessa previsão do ano de 2021, é possível perceber que os modelos seguem claramente o aproveitamento do time, assim como a quantidade de gols na rodada de entrada, ou seja, os atributos que mais explicam o modelo. Por isso, o líder atual, Red Bull Bragantino se encontra em uma posição elevada, assim como o Palmeiras, o campeão previsto por ser um time muito regular, conseguindo boas colocações e títulos nas últimas edições do campeonato.

## IV. Conclusões

A partir dos resultados encontrados em ambos os códigos programados, o das tabelas e os das rodadas, é possível concluir que os modelos de *machine learning* construídos, assim como as técnicas de mineração de dados utilizadas se mostraram satisfatórios e condizentes com o que se esperava, demonstrando assim o poder dessas técnicas computacionais para resolver problemas reais do dia a dia, como é o caso do futebol, um esporte com regras simples, mas com grande imprevisibilidade.

Ademais, vale ressaltar que vários outros atributos poderiam ser considerados para contribuir com a criação dos modelos de aprendizado de máquina, como o número de cartões amarelos e cartões vermelhos do time, o retrospecto de partidas entre os times, já que alguns times apresentam bons números relacionados a vitórias contra outros times específicos, o número de jogos sem perder, dentre outros. Contudo, esses atributos não foram colocados, uma vez que são difíceis de serem encontrados, além de que alguns sites não permitem a realização do *web scraping,* o que não deixa os códigos programados ruins, mas é algo importante a ser mencionado para futuros projetos pela possível melhoria que eles podem trazer para os resultados.

Logo, conclui-se que, o conteúdo passado na disciplina PO450, Introdução ao aprendizado de máquinas e à mineração de dados, foi aplicado corretamente e de forma satisfatória.

## Agradecimentos

Agradecimentos ao professor Leonardo Tomazeli Duarte, um excelente docente, à Unicamp que proporciona um ambiente propício para o desenvolvimento pessoal, acadêmico e profissional e aos meus pais que me deram estrutura e oportunidades.

## Referências

[1] Assessoria CBF. CBF apresenta relatório sobre papel do futebol na economia do Brasil. 14 de dezembro de 2019. Disponível em:<https://www.cbf.com.br/a-cbf/informes/index/cbf-apresenta-relatorio-sobre-papel-do-futebol-na-economia-do-brasil>. Acesso em: 03 de julho de 2021.

[2] Confederação Brasileira de Futebol. In Wikipédia: a enciclopédia livre. Disponível em:<https://pt.wikipedia.org/wiki/Confedera%C3%A7%C3%A3o_Brasileira_de_Futebol>. Acesso em: 03 de julho de 2021.

[3] WATT, J.; BORHANI, R.; KATSAGGELOS, A.; Machine Learning Refined: Foundations, Algorithms, and Applications. Cambridge: Cambridge University Press, 2016.

[4] Campeonato Brasileiro de Futebol. In Wikipédia: a enciclopédia livre. Disponível em:<https://pt.wikipedia.org/wiki/Campeonato_Brasileiro_de_Futebol>. Acesso em: 03 de julho de 2021.

[5] GONZALEZ, L. Regressão Logística e suas Aplicações. Tese (bacharel em Ciências da Computação -Centro de Ciências Exatas e Tecnologias, Universidade Federal do Maranhão. São Luís, p. 46. 2018.