

INSTITUTO FEDERAL MINAS GERAIS

PRÓ-REITORIA DE PESQUISA, INOVAÇÃO E PÓS-GRADUAÇÃO

PROJETO DE PESQUISA

EDITAL 47/2022 - EDITAL DE FLUXO CONTÍNUO PARA O REGISTRO DE PROJETOS DE PESQUISA VOLUNTÁRIOS

UNIDADE PROPONENTE

Campus:
FORMIGA

IDENTIFICAÇÃO DO PROJETO

Título do Projeto:

Modelos de Aprendizagem Estatística em Finanças: Aplicações e Desenvolvimento de Soluções Reproduzíveis e Auditáveis

Grande Área de Conhecimento:
CIÊNCIAS SOCIAIS APLICADAS

Área de Conhecimento:
ADMINISTRAÇÃO

Período de Execução:

Início: 01/08/2023 | Término: 01/08/2025

Nome do Responsável

(Coordenador): Washington Santos da Silva
Titulação: DOUTORADO
Matrícula: 1553273
Vínculo: Voluntário

Departamento de Lotação: CFO-IFMG
Telefone: (00037) 3322-8434

E-mail: washington.silva@ifmg.edu.br

EQUIPE PARTICIPANTE

Professores e/ou Técnicos Administrativos do IFMG

Membro	Contatos	Vínculo	Titulação
Nome: Washington Santos da Silva Matrícula: 1553273	Tel.: (00037) 3322-8434 E-mail: washington.silva@ifmg.edu.br	Voluntário	DOUTORADO

DISCRIMINAÇÃO DO PROJETO

Resumo

Este projeto objetiva identificar problemas e aplicar modelos de aprendizagem estatística, incluindo aprendizagem profunda e reforço de aprendizagem, a problemas da área e de subáreas das Finanças e de áreas correlatas, com a finalidade de produção de artigos científicos. Além disso, outro objetivo pretende desenvolver materiais de aprendizagem (tutoriais) para capacitar os mestrandos e estudantes do bacharelado em Administração em metodologias e recursos computacionais necessários para a produção de pesquisas, relatórios e outros produtos baseados em dados auditáveis e reproduzíveis, envolvendo as linguagens R, Python, o sistema de controle de versão Git e o GitHub, o sistema de publicação Quarto, entre outros.

Introdução

Neste século XXI, os computadores estão envolvidos em muitas transações econômicas e podem capturar dados associados a essas transações, que podem então ser manipulados e analisados, incluindo grandes bancos de dados de texto. As técnicas estatísticas e econométricas convencionais, como a regressão, apresentam desempenho relativamente bom para o estudo de alguns fenômenos, mas existem questões específicas dos grandes conjuntos de dados que podem demandar ferramentas diferentes (EINAV; LEVIN (2014)).

Primeiro, o tamanho dos bancos de dados atualmente disponíveis pode exigir ferramentas de manipulação de dados mais poderosas. Em segundo lugar, podemos ter mais preditores potenciais do que os apropriados para estimação, por isso, em geral é necessário implementar alguma técnica de seleção de variáveis (VARIAN (2014)).

Terceiro, os grandes conjuntos de dados disponíveis podem permitir o ajuste de relações mais flexíveis do que as implicadas pelos modelos lineares. Técnicas de aprendizagem estatística, tais como árvores de decisão, máquinas de vetores de suporte, redes neurais, aprendizado profundo entre outras, podem permitir formas mais eficazes de modelar relações complexas (MULLAINATHAN; SPIESS (2017)).

Isto posto, a utilização de modelos de aprendizagem estatística (ou de máquina) pode ajudar a melhorar a tomada de decisões em subáreas das Finanças, tais como: análise de crédito, gestão de risco, otimização de portfólios, algorithmic trading e previsão de variáveis econômico-financeiras. Destaque-se que serão também exploradas aplicações em áreas correlatas (ATHEY (2018)). Importante destacar ainda que, dado o substancial volumes de dados textuais disponíveis, avalia-se no âmbito deste projeto, que a mineração de textos pode ser aplicada para extrair informações valiosas dos dados textuais disponíveis, uma área relativamente ainda pouco explorada na área de Finanças e nas áreas correlatas.

Neste contexto, este projeto de pesquisa tem como um dos objetivos explorar e aplicar modelos de aprendizagem estatística, incluindo aprendizagem profunda e reforço de aprendizagem, para abordar uma ampla gama de problemas na área de Finanças e áreas afins.

No entanto, muitos profissionais da área de Finanças ainda não estão preparados para aplicar esses modelos e técnicas. Isto posto, outro objetivo do projeto é desenvolver diferentes tipos de materiais de aprendizagem para capacitar os mestrandos e graduandos em Administração e outros profissionais da área de negócios nas metodologias e recursos computacionais necessárias para a criação de relatórios e de outros produtos baseados em dados auditáveis e reproduzíveis.

Justificativa

A justificativa para este projeto de pesquisa é baseada na crescente importância da aplicação da aprendizagem estatística na tomada de decisões em outros aspectos das Finanças, áreas correlatas e nas ciências sociais aplicadas em geral. Os modelos de aprendizagem estatística tem demonstrado sua eficácia na previsão, otimização e gestão de riscos, bem como na identificação de oportunidades de investimento em mercados altamente dinâmicos (GOGAS; PAPADIMITRIOU (2021)).

A estatística e a econometria modernas e os modelos de aprendizagem estatística, são fundamentalmente disciplinas computacionais, mas é fácil constatar que este fato não se reflete na formação dos profissionais das ciências sociais aplicadas. Com a ascensão do *big data* e da ciência de dados, tornou-se cada vez mais claro que para formarmos pesquisadores e profissionais preparados para os desafios atuais, é necessária um formação explícita em técnicas e ferramentas computacionais. Além disso, as diretrizes curriculares recentes afirmam claramente que trabalhar com dados requer extensas habilidades de computação e que os profissionais de administração, e certamente de todas as demais áreas, devem ser fluentes no acesso, manipulação, análise e modelagem com softwares profissionais de análise estatística.

Enfatizando, a aplicação de modelos de aprendizagem estatística exige não apenas expertise no entendimento dos modelos, mas também na implementação de fluxos de trabalho robustos que garantam a reproducibilidade e auditabilidade das análises, entretanto, é fácil constatar que muitos profissionais da área de finanças ainda não são formados nestas técnicas e métodos (ÇETINKAYA-RUNDEL; RUNDEL (2018)).

Isto posto, este projeto é relevante não apenas para a evolução dos procedimentos de modelagem de dados em Finanças e áreas correlatas, mas também para a formação de pessoal qualificado para a implementação de abordagens inovadoras para resolver os desafios apresentados neste início de Séc. XXI. A ênfase na aplicação de procedimentos e métodos computacionais compatíveis com a reproducibilidade e auditabilidade dos produtos resultantes de projetos profissionais e de pesquisa pretende criar capacidades para que as descobertas e soluções propostas sejam confiáveis e transparentes, promovendo a confiança nas decisões baseadas na modelagem avançada de dados.

Fundamentação Teórica

Este referencial teórico é basendo em HASTIE; TIBSHIRANI; FRIEDMAN (2009) e JAMES (2021)

4.1 O que é Aprendizagem Estatística?

De forma geral, suponha que observamos uma resposta quantitativa Y e p preditores diferentes, X_1, X_2, \dots, X_p . Assumimos que existe alguma relação entre Y e $X = X_1, X_2, \dots, X_p$, que pode ser escrita de forma muito geral como:

$$Y = f(X) + \epsilon$$

Aqui f é alguma função fixa, mas desconhecida, de X_1, \dots, X_p , e ϵ é um termo de erro aleatório, que é independente de X e tem média zero. Nesta formulação f representa a informação sistemática que X fornece sobre Y . Em geral, a função f pode envolver mais de uma variável de entrada.

Em essência, a aprendizagem estatística refere-se a um conjunto de abordagens para estimar f .

Nesta seção, descrevemos alguns dos principais conceitos teóricos que surgem na estimativa de f , bem como ferramentas para avaliar as estimativas obtidas.

4.1.1 Por que estimar f ?

Existem duas razões principais pelas quais podemos querer estimar f : *previsão e inferência*.

Previsão

Em muitas situações, um conjunto de entradas X^X está disponível, mas a saída Y^Y não pode ser obtida facilmente. Nesta configuração, como a média do termo de erro é zero, podemos prever Y usando:

$$\hat{Y} = \hat{f}(X),$$

sendo que \hat{f} representa a estimativa para f e \hat{Y} representa a previsão resultante para Y . Neste cenário, \hat{f} é frequentemente tratado como uma caixa preta, no sentido de que normalmente não se preocupa com a forma exata de f , desde que produza previsões precisas para Y .

A acurácia de \hat{Y} como previsão para Y depende de duas quantidades, que chamaremos de erro redutível e erro irredutível. Em geral, f não será uma estimativa perfeita para f , e esta imprecisão introduzirá algum erro. Este erro é redutível porque podemos potencialmente melhorar a precisão de f usando a técnica de aprendizagem estatística mais apropriada para estimar f . No entanto, mesmo que fosse possível formar uma estimativa perfeita para f , de modo que a resposta estimada assumisse a forma $\hat{Y} = f(X)$, a previsão ainda teria algum erro! Isso ocorre porque Y também é uma função de ε , que, por definição, não pode ser previsto usando X . Portanto, a variabilidade associada a ε também afeta a precisão das previsões. Isso é conhecido como *erro irredutível*, porque não importa quão bem estimamos f , não podemos reduzir o erro introduzido por ε .

Considere uma determinada estimativa \hat{f} e um conjunto de preditores X , que produz a previsão $\hat{Y} = \hat{f}(X)$. Suponha por um momento que \hat{f} e X sejam fixos, de modo que a única variabilidade venha de ε . Então é fácil mostrar que:

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2, \\ &= [f(X) - \hat{f}(X)]^2 + V(\varepsilon) \end{aligned}$$

sendo que $E(Y - \hat{Y})^2$ representa o valor esperado da diferença quadrática entre o valor previsto e o atual de Y , e $V(\varepsilon)$ representa a variância do termo do erro ε . O foco deste projeto está na aplicação

de técnicas de aprendizagem estatística em problemas de finanças para estimar f com o objetivo de minimizar o erro redutível.

Inferência

Muitas vezes estamos interessados em entender a associação entre Y e X_1, \dots, X_p . Nesta situação desejamos estimar f , mas nosso objetivo não é necessariamente fazer previsões para Y . Agora f não pode ser tratado como uma caixa preta, porque precisamos saber sua forma exata. Neste cenário, pode-se estar interessado em responder às seguintes perguntas:

- Quais preditores estão associados à resposta? Muitas vezes acontece que apenas uma pequena fração dos preditores disponíveis está substancialmente associada a Y . Identificar os poucos preditores importantes entre um grande conjunto de variáveis possíveis pode ser extremamente útil, dependendo da aplicação.
- Qual é a relação entre a resposta e cada preditor? Alguns preditores podem ter um relacionamento positivo com Y , no sentido de que valores maiores do preditor estão associados a valores maiores de Y . Outros preditores podem ter a relação oposta. Dependendo da complexidade de f , a relação entre a resposta e um determinado preditor também pode depender dos valores dos outros preditores.
- A relação entre Y e cada preditor pode ser resumida adequadamente usando uma equação linear ou a relação é mais complicada? Historicamente, a maioria dos métodos para estimar f assumiu uma forma linear. Em algumas situações, tal suposição é razoável ou mesmo desejável. Mas muitas vezes a verdadeira relação é mais complicada e, nesse caso, um modelo linear pode não fornecer uma representação precisa da relação entre as variáveis de entrada e de saída.

Neste projeto, objetivamos tratar problemas que se enquadram na configuração de previsão, na configuração de inferência ou em uma combinação de ambas. Dependendo se o objetivo final é a previsão, a inferência ou uma combinação dos dois, diferentes métodos para estimar f podem ser apropriados.

Por exemplo, os modelos lineares permitem inferências previsíveis de modelos relativamente simples e interlineares, mas podem não produzir previsões tão precisas quanto algumas outras abordagens. Em contraste, algumas das abordagens altamente não lineares que pretendemos aplicar a problemas de Finanças podem potencialmente fornecer previsões bastante precisas para Y , mas isto ocorre às custas de um modelo menos interpretável, para o qual a inferência é mais desafiadora.

4.1.2 Como estimar f ?

Ao longo deste projeto, exploramos muitas abordagens lineares e não lineares para estimar f . No entanto, estes métodos geralmente partilham certas características. Fornecemos uma visão geral dessas características compartilhadas nesta seção. Sempre assumiremos que observamos um conjunto de n pontos de dados diferentes, sendo que selecionamos uma fração dessas observações como sendo dados de treinamento, porque usaremos essas observações para treinar, ou ensinar, nosso método como estimar f .

Considere que x_{ij} representa o valor do j-ésimo preditor para observação i , sendo $i=1,2,...,n$ e $j=1,2,...,p$. Da mesma forma, seja y_i a variável resposta para a i -ésima observação. Assim, os dados de treinamento consistem em $x_1, y_1, x_2, y_2, \dots, x_n, y_n$, sendo $x_i = x_{i1}, x_{i2}, \dots, x_{ip}$.

Conforme exposto anteriormente, um dos objetivos deste projeto é aplicar diversos métodos de aprendizagem estatística aos dados de treinamento para estimar a função desconhecida f . Em outras palavras, queremos encontrar uma função f tal que $Y \approx f$ para qualquer observação X,Y.

Em termos gerais, a maioria dos métodos de aprendizagem estatística para esta tarefa podem ser caracterizados como paramétricos ou não paramétricos. Existem vantagens e desvantagens em ambos os métodos de aprendizagem estatística. Exploraremos ambos os tipos de métodos no desenvolvimento deste projeto. Discutiremos agora brevemente esses dois tipos de abordagens.

Métodos Paramétricos

Os métodos paramétricos envolvem uma abordagem baseada em duas etapas:

1. Primeiro, fazemos uma suposição sobre a forma funcional, ou formato, de f . Por exemplo, uma suposição muito simples é que f é linear em $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$.

Este é um modelo linear. Uma vez que assumimos que f é linear, o problema de estimar f é bastante simplificado. Em vez de tendo que estimar uma função p -dimensional totalmente arbitrária $f(X)$, basta estimar os coeficientes $\beta_0, \beta_1, \dots, \beta_p$.

2. Após a seleção de um modelo, precisamos de um procedimento que use os dados de treinamento para ajustar ou treinar o modelo. No caso do modelo linear, precisamos estimar os parâmetros $\beta_0, \beta_1, \dots, \beta_p$. Ou seja, queremos encontrar valores desses parâmetros tais que:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

A abordagem mais comum para ajustar este modelo é referida como mínimos quadrados (comuns). No entanto, os mínimos quadrados são uma das muitas maneiras possíveis de ajustar o modelo linear.

A abordagem baseada em modelo que acabamos de descrever é chamada de paramétrica; reduz o problema de estimar f a estimar um conjunto de parâmetros. Assumir uma forma paramétrica para f simplifica o problema de estimar f porque geralmente é muito mais fácil estimar um conjunto de parâmetros, como $\beta_0, \beta_1, \dots, \beta_p$ no modelo linear, do que isto é ajustar uma função totalmente arbitrária f .

A desvantagem potencial de uma abordagem paramétrica é que o modelo que escolhemos geralmente não corresponderá à verdadeira forma desconhecida de f . Se o modelo escolhido estiver muito longe do verdadeiro f , então nossa estimativa será ruim. Podemos tentar resolver esse

problema escolhendo modelos flexíveis que possam se ajustar a muitas formas funcionais possíveis e flexíveis para f .

Mas, em geral, ajustar um modelo mais flexível requer estimar um maior número de parâmetros. Esses modelos mais complexos podem levar a um fenômeno conhecido como overfitting dos dados, o que essencialmente significa que o overfitting segue os erros ou ruídos muito de perto.

Métodos Não-Paramétricos

Os métodos não paramétricos não fazem suposições explícitas sobre a forma funcional de f . Em vez disso, eles buscam uma estimativa de f que chegue o mais próximo possível dos pontos de dados, sem ser muito grosseira ou distorcida.

Tais abordagens podem ter uma grande vantagem sobre as abordagens paramétricas: ao evitar a suposição de uma forma funcional específica para f , elas têm o potencial de ajustar com precisão uma gama mais ampla de formas possíveis para f . Qualquer abordagem paramétrica traz consigo a possibilidade de que a forma funcional usada para estimar f seja muito diferente da verdadeira f , caso em que o modelo resultante não se ajustará bem aos dados.

Em contraste, as abordagens não paramétricas evitam completamente este perigo, uma vez que essencialmente nenhuma suposição sobre a forma de f é feita. Mas as abordagens não paramétricas sofrem de uma grande desvantagem: uma vez que não reduzem o problema de estimar f a um pequeno número de parâmetros, um número muito grande de observações (muito mais do que normalmente é necessário para uma abordagem paramétrica) é necessário. necessário para obter uma estimativa precisa de f .

4.2 Aprendizagem Supervisionada versus Aprendizagem Não Supervisionada

A maioria dos problemas de aprendizagem estatística se enquadra em uma de duas categorias: supervisionado ou não supervisionado. Sem dúvida, a maioria das aplicações a serem exploradas neste projeto se enquadram no domínio da aprendizagem supervisionada. Para cada observação

da(s) medida(s) preditora(s) x_i , $i=1,\dots,n$ há uma medida da variável resposta associada y_i .

Desejamos ajustar um modelo que relate a resposta aos preditores, com o objetivo de prever com precisão a resposta para futuras observações ou compreender melhor a relação entre a resposta e os preditores (inferência). Muitos modelos de aprendizagem estatísticos clássicos, tais como regressão linear e regressão logística, bem como métodos mais modernos como Modelos Aditivos Generalizados (GAMM), Árvores de Decisão, Boosting e Suporte Vector Machines (SVM), operam no domínio da aprendizagem supervisionada.

Por outro lado, a aprendizagem não supervisionada descreve o desafio um pouco mais desafiador em que para cada observação $i = 1,\dots,n$, observamos um vetor de medições x_i mas nenhuma resposta associada y_i . Assim, não é possível ajustar um modelo de regressão, pois não há variável resposta a ser prevista. Neste cenário, estamos, de certa forma, trabalhando às cegas; a situação é referida como não supervisionada porque não temos variável resposta que pode supervisionar

nossa análise. Que tipo de análise estatística é possível nestes casos? Podemos buscar entender as relações entre as variáveis ou entre as observações.

Uma ferramenta de aprendizagem estatística que podemos usar neste cenário é a análise de agrupamentos. O objetivo da análise de agrupamentos é verificar, com base em x_1, \dots, x_n , se as observações podem ser alocadas em grupos relativamente distintos. Por exemplo, num estudo de segmentação de mercado, podemos observar múltiplas características (variáveis) para clientes potenciais, como CEP, renda familiar e hábitos de compra. Podemos acreditar que os clientes se enquadram em grupos diferentes, como grandes compradores versus pequenos compradores, em termos do volume dispendido. Se as informações sobre os padrões de gastos de cada cliente estivessem disponíveis, então uma análise supervisionada seria possível.

Contudo, esta informação não está disponível, ou seja, não sabemos se cada cliente potencial gasta muito ou não. Nesta configuração, podemos tentar agrupar os clientes com base nas variáveis mensuradas, a fim de identificar grupos distintos de clientes potenciais. Identificar esses grupos pode ser interessante porque pode ser que os grupos diferem em relação a alguma propriedade de interesse, como hábitos de consumo, por exemplo.

4.2.1 Avaliando a Acurácia de Modelos

Nesta seção, discutimos alguns dos conceitos mais importantes que surgem na seleção de um modelo de aprendizagem estatística para um conjunto de dados específico.

4.2.2 Medindo a Qualidade do Ajuste

Para avaliar o desempenho de um método de aprendizagem estatística em um determinado conjunto de dados, precisamos de alguma forma de medir até que ponto suas previsões realmente correspondem aos dados observados. Ou seja, precisamos quantificar até que ponto o valor da resposta prevista para uma determinada observação está próximo do valor verdadeiro da resposta para essa observação. No cenário de regressão, a medida mais comumente usada é o *Mean Squared Error* (MSE), dado por:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

sendo $\hat{f}(x_i)$ a previsão que f fornece para a i -ésima observação. O MSE será pequeno se as respostas previstas estiverem muito próximas das respostas verdadeiras, e será grande se, para algumas das observações, as respostas previstas e verdadeiras diferirem substancialmente.

O MSE é calculado usando os dados de treinamento que foram usados para se ajustar ao modelo e, portanto, deve ser chamado com mais precisão de treinamento MSE. Mas, em geral, não nos importamos realmente quanto bem o método funciona treinando nos dados de treinamento. Em vez disso, estamos interessados na precisão das previsões que obtemos quando aplicamos o nosso método a dados de teste nunca antes vistos. Por que é com isso que nos importamos? Suponha que estejamos interessados em desenvolver um algoritmo para prever o preço de uma ação com base nos retornos anteriores de ações. Podemos treinar o método usando retornos de ações dos últimos 6 meses. Mas realmente não nos importamos com o quanto bem nosso método prevê o preço das ações da semana passada. Em vez disso, nos preocupamos com o quanto bem ele irá prever o preço de amanhã ou o preço do próximo mês.

Para afirmar isso de forma mais matemática, suponha que ajustamos nosso método de aprendizagem estatística em nossas observações de treinamento $x_1, y_1, x_2, y_2, \dots, x_n, y_n$, e obtemos a

estimativa \hat{f} . Podemos então calcular $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$.

Se estes forem aproximadamente iguais a y_1, y_2, \dots, y_n , então o MSE de treinamento é pequeno. No entanto, não estamos realmente interessados em saber se $\hat{f}(x_i) \approx y_i$; em vez disso, queremos saber se

$f(x_0)$ é aproximadamente igual a y_0 , onde x_0, y_0 é uma observação de teste nunca antes vista e não usada para treinar o método de aprendizagem estatística. Queremos escolher o método que fornece o menor MSE de teste, em oposição para o MSE de treinamento mais baixo. Em outras palavras, se tivéssemos um grande número de observações de teste, poderíamos calcular:

$$\text{Média}(y_0 - f(x_0))^2$$

o erro quadrático médio de previsão para essas observações de teste x_0, y_0 . Gostaríamos de selecionar o modelo para o qual essa quantidade seja a menor possível.

Como podemos tentar selecionar um método que minimize o MSE de teste? Em algumas situações, podemos ter um conjunto de dados de teste disponível – ou seja, podemos ter acesso a um conjunto de observações que não foram usadas para treinar o método de aprendizagem estatística. Podemos então simplesmente avaliar as observações do teste e selecionar o método de aprendizagem para o qual o MSE do teste é menor.

Na prática, geralmente é possível calcular o MSE de treinamento com relativa facilidade, mas estimar o MSE de teste é consideravelmente mais difícil porque normalmente não há dados de teste disponíveis. O nível de flexibilidade correspondente ao modelo com o MSE de teste mínimo pode variar consideravelmente entre os conjuntos de dados. Ao longo deste projeto, discutimos uma variedade de abordagens que podem ser usadas na prática para estimar esse ponto mínimo. Um método importante é a validação-cruzada, que é um método para estimar o MSE do teste usando os dados de treinamento.

Objetivo Geral

Este projeto envolve dois objetivos. Primeiramente, visa identificar problemas atuais na área de Finanças e áreas correlatas que possam ser tratados pela aplicação de modelos de aprendizagem estatística, abrangendo áreas como análise de crédito, gestão de risco, otimização de portfólios, algorithmic trading e previsão de variáveis econômico-financeiras, além de envolver a exploração de aplicações em grandes áreas correlatas, tais como auditoria, contabilidade.

Em segundo lugar, o projeto se propõe a desenvolver materiais de aprendizagem para alunos do Mestrado Profissional em Administração e do bacharelado para capacitar-los para a adoção de métodos e recursos computacionais que possibilitem o desenvolvimento de produtos e soluções reproduzíveis e auditáveis (VILHUBER et al. (2023), HOYNES (2023)), o que envolverá a capacitação em linguagens de programação (WICKHAM; ÇETINKAYA-RUNDEL; GROLEMUND (2023), JAMES et al. (2023)), sistemas de controle de versão (BRYAN (2018)) e recursos para a manipulação de grandes arquivos de dados (WICKHAM; ÇETINKAYA-RUNDEL; GROLEMUND (2023)).

Metas

1 - Elaborar a produção científica, técnica e tecnológica pertinente

Metodologia da Execução do Projeto

Para alcançar os objetivos propostos, os seguintes procedimentos serão executados:

1. Revisão Sistemática da literatura - Parte 1: Realizar uma revisão sistemática da literatura sobre a aplicação de modelos de aprendizagem estatística na área de finanças e em áreas correlatas para identificar oportunidades de pesquisas inovadoras.

2. Revisão Sistemática da literatura - Parte 2: será realizada uma revisão sistemática da literatura referente à aplicação de metodologias adequadas para a elaboração de relatórios, artigos e produtos baseados em dados auditáveis e reproduzíveis na área de Finanças e áreas correlatas.
3. Identificação de problemas: A partir das revisões sistemáticas, serão identificados problemas e oportunidades de pesquisa e de desenvolvimento de soluções técnicas em Finanças e em áreas correlatas que possam ser tratadas pela aplicação de modelos de aprendizagem estatística e/ou pela adoção de metodologias computacionais para a elaboração de relatórios, artigos e produtos baseados em dados auditáveis e reproduzíveis.
4. Elaboração e submissão de pelo menos dois artigos científicos a revistas indexadas.
5. Desenvolvimento de materiais de aprendizagem: serão desenvolvidos materiais de aprendizagem, tais como tutoriais para capacitar os estudantes do Mestrado profissional em Administração e, se houver demanda, estudantes do Bacharelado em Administração:
 1. na aplicação de modelos de aprendizagem estatística em problemas da área de Finanças e de áreas correlatas;
 2. para a adoção de métodos e técnicas para a elaboração de relatórios, artigos e produtos baseados em dados auditáveis e reproduzíveis;
 3. os materiais versarão sobre linguagens de programação (R, Python), controle de versão com git e GitHub, o sistema de publicação Quarto, entre outros recursos e procedimentos computacionais.
6. Todos os procedimentos anteriores serão realizados, preferencialmente, pela orientação e co-orientação de orientados do mestrado profissional em Administração com a participação de estudantes do bacharelado em Administração.

Acompanhamento e Avaliação do Projeto

- Serão elaborados e apresentados relatórios semestrais sobre os resultados parciais obtidos.
- A eficácia dos materiais de aprendizagem desenvolvidos será avaliada pela realização de pesquisas com os usuários atuais e potenciais.

Disseminação dos Resultados

1. Submissão de pelo menos dois artigo científico para publicação em revistas científicas indexadas
2. Elaboração de pelo menos três tutoriais sobre a utilização das linguagens, R, python e sobre o sistema de controle de versão git para o desenvolvimento de pesquisas e produtos reproduzíveis e auditáveis. Esses tutoriais serão contabilizados como produção técnica para o Mestrado Profissional em Administração.

Por fim, é importante destacar que uma co-orientação em andamento no âmbito do Mestrado Profissional em Administração já obteve resultados promissores para publicação.

Referências Bibliográficas

ATHEY, S. The impact of machine learning on economics. Em: **The economics of artificial intelligence: An agenda.** [s.l.] University of Chicago Press, 2018. p. 507–547.

BRYAN, J. Excuse me, do you have a moment to talk about version control? **The American Statistician**, v. 72, n. 1, p. 20–27, 2018.

ÇETINKAYA-RUNDEL, M.; RUNDEL, C. Infrastructure and tools for teaching computing throughout the statistical curriculum. **The American Statistician**, v. 72, n. 1, p. 58–65, 2018.

EINAV, L.; LEVIN, J. Economics in the age of big data. **Science**, v. 346, n. 6210, p. 1243089, 2014.

GOGAS, P.; PAPADIMITRIOU, T. Machine learning in Economics and Finance. **Computational Economics**, v. 57, p. 1–4, 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. [s.l.] Springer-Verlag, 2009.

HOYNES, H. Reproducibility in Economics: Status and Update. **Harvard Data Science Review**, v. 5, n. 3, 2023.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., TAYLOR, J. An Introduction to Statistical Learning: With Applications in Python. 2023, Springer, 2023.

JAMES. G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: With Applications in R. 2. ed. [s.l.] Springer, 2021.

MULLAINATHAN, S.; SPIESS, J. Machine learning: An Applied Econometric Approach. **Journal of Economic Perspectives**, v. 31, n. 2, p. 87–106, 2017.

VARIAN, H. R. Big data: New tricks for econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3–28, 2014.

VILHUBER, L. et al. Reinforcing reproducibility and replicability: An introduction. **Harvard Data Science Review**, v. 5, n. 3, 2023.

WICKHAM, H.; ÇETINKAYA-RUNDEL, M.; GROLEMUND, G. R for data science. 2. ed. [s.l.] "O'Reilly Media, Inc.", 2023.

CRONOGRAMA DE EXECUÇÃO

Meta Atividade	Especificação	Indicador(es) Qualitativo(s)	Indicador Físico	Período de Execução		
				Unid.de Medida	Qtd.	Início
1 1	Orientação Coorientação pelo menos Profissional Administração.	a) Elaboração do artigo e científico resultante da revisão sistemática sobre aplicação de um aprendizagem estatística em mestrando Mestrado Finanças e áreas correlatas b) Desenvolvimento dos tutoriais que constituirão a produção técnica e tecnológica.	b) Desenvolvimento dos tutoriais que constituirão a produção técnica e tecnológica.			01/08/2023 01/08/2025

PLANO DE APLICAÇÃO

	Classificação da Despesa	Especificação PROPI (R\$)	DIGAE (R\$)	Campus Proponente (R\$)	Total (R\$)
TOTAIS	0	0	0	0	

Anexo A

MEMÓRIA DE CÁLCULO

CLASSIFICAÇÃO DE DESPESA	ESPECIFICAÇÃO	UNIDADE DE MEDIDA	QUANT.	VALOR UNITÁRIO	VALOR TOTAL
TOTAL GERAL					-