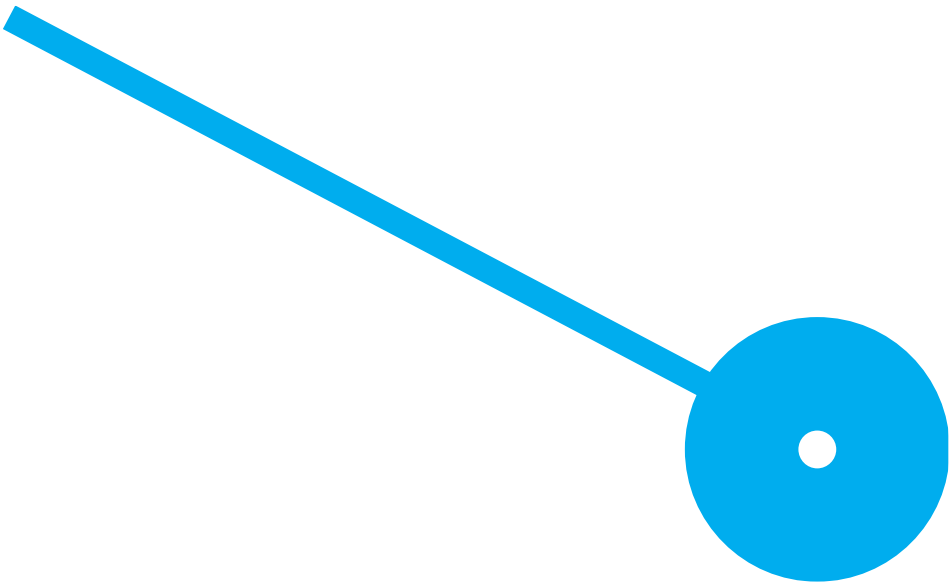
ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

MESTRADO
Engenharia Informática

Aplicação de Modelos de *Machine Learning* para Previsão de Eventos de Stress Financeiro

Ana Beatriz Esteves Fernandes

10/2024



ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

MESTRADO
Engenharia Informática

Aplicação de Modelos de *Machine Learning* para Previsão de Eventos de Stress Financeiro

Ana Beatriz Esteves Fernandes
8180039

Orientadores:

Professora Doutora Mariana Valério de Carvalho
Professora Doutora Ana Isabel Borges

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática pela Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto.

10/2024

Declaração de Integridade

Eu, Ana Beatriz Esteves Fernandes, estudante nº 8180039, do Mestrado de Engenharia Informática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, declaro que não fiz plágio nem auto-plágio, pelo que o trabalho intitulado "Aplicação de Modelos de *Machine Learning* para Previsão de Eventos de Stress Financeiro" é original e da minha autoria, não tendo sido usado previamente para qualquer outro fim. Mais declaro que todas as fontes usadas estão citadas, no texto e na bibliografia final, segundo as regras de referenciação adotadas na instituição.

Agradecimentos

Após dois anos repletos de desafios e momentos altos e baixos, quero agradecer a todos aqueles que estiveram constantemente ao meu lado e contribuíram para tornar esta jornada mais fácil. Compartilharam comigo experiências, histórias, risos e, acima de tudo, possibilitaram o meu crescimento tanto pessoal quanto profissional. Portanto, este espaço é reservado a todos vocês.

Em primeiro lugar começo por agradecer às minhas orientadoras, Professora Doutora Mariana Carvalho e Professora Doutora Ana Isabel Borges, por me terem dado a oportunidade de realizar este projeto sob a vossa orientação, por todo o apoio prestado durante todo o processo de pesquisa e desenvolvimento deste projeto e por todo o conhecimento e ensinamentos partilhados ao longo deste ano. À Professora Mariana tenho de agradecer não só por este ano e por este projeto, mas também pelos 3 anos de Licenciatura que passaram, pelos muitos projetos que desenvolvemos juntas nas várias UC's que lecionou e, principalmente, pelo projeto final da Licenciatura, o estágio que também o desenvolvi com a sua orientação. À Professora Ana Isabel, que embora só tenha conhecido este ano, obrigada pelos ensinamentos durante este projeto e por todo o apoio prestado. Este projeto deve-se, também, a vocês que foram excecionais ao orientar-me para o sucesso deste projeto.

Para o meu pai, obrigada por me ajudares a tornar na pessoa que sou hoje. Obrigada por me apoiares, por dares os melhores conselhos, por cuidares de mim e por me ajudares a crescer. A minha vida sem ti teria sido muito mais difícil, por isso, obrigada por não desistires de mim e me teres dado a melhor vida possível. Estarei para sempre grata!

Aos meus irmãos, Hugo e Renato, obrigada por estarem presentes e por serem os melhores irmãos que eu poderia ter pedido! Partilhar a vida com vocês é das melhores coisas do mundo, por isso, obrigada por serem os meus melhores amigos e por estarem sempre lá quando eu preciso. Gosto imensamente de vocês, maninhos!

Sobrinha Leonor, um dia quando cresceres vou-te mostrar isto para te inspirares e superares-me. Ser tia foi das melhores prendas que tive e quero muito ser para ti um

suporte, uma amiga e uma inspiração. Por isso, obrigada por me fazeres querer ser melhor! Amo-te infinitamente!

Inês e Leonor, as melhores primas do mundo inteiro que são muito mais do que isso. Vocês são como irmãs! Obrigada por cada palavra, cada gesto, por me porem um sorriso na cara com as vossas parvoíces, por cada jantar no sushi, e por serem quem são. Orgulha-me ver-vos crescer e espero continuar a acompanhar-vos pela vida fora. Amo-vos daqui até à lua!

À minha madrinha/tia Rosária, obrigada pelos ensinamentos de vida, pelas tuas palavras certas no momento certo, por seres uma inspiração e por me teres dado as duas melhores pessoas que tenho na vida. És como uma mãe para mim e obrigada por isso!

Ao tio Sérgio, obrigada pelos conselhos, pelos incentivos e pelas brincadeiras. Os teus incentivos de ingressar na universidade quando eu não o queria deram frutos. Obrigada por isso, pela tua companhia nos treinos, pela tua música/guitarra e pelos teus poemas. Gosto muito de ti!

Aos avós, Bernardino e Olinda, obrigada pelas vossas histórias, ensinamentos, partilhas e amor! Estarei para sempre grata por tudo o que fizeram por mim!

Obrigada aos meus sogros que foram (e são) como pais! Obrigada por me receberem tão bem, por me ajudarem quando eu precisei, pelas gargalhadas e, acima de tudo, por serem família. Quero poder retribuir em dobro todo o bem que me proporcionam!

Gui, conheci-te tão pequeno e estás a tornar-te um homem! Vou estar sempre aqui para te ajudar nos exercícios de matemática :) e naquilo que precisares. És como um irmão!

À Lili e ao Guga, obrigada por me "distraírem" e tornarem os dias de trabalho neste projeto menos difíceis. Os dias das "ganâncias" são sempre os melhores!

Agora às minhas manas que são as melhores amigas do mundo!

Débora, és a amiga mais antiga que tenho, e passem os anos que passarem a nossa amizade continua igual. Por isso obrigada por ainda estares comigo e por saber que posso contar contigo em qualquer circunstância!

Bia6, "trocaste-me" por Lisboa e, embora passemos menos tempo juntas (menos do que aquilo que gostaria) sei que basta uma mensagem para me dares os teus melhores conselhos. Obrigada por esse riso inconfundível, por colocares um sorriso em qualquer pessoa, mesmo às vezes sendo tu a precisar mais. Mesmo longe, estás sempre perto!

Xica, sabes que os nossos melhores momentos foram os do complexo. Essas memórias ficarão para sempre guardadas na minha memória e no meu coração. Vais ser sempre a minha vitoriana preferida!

Guida, além de amiga agora és a minha farmacêutica pessoal :) as tuas histórias são sempre as melhores, mais icônicas e divertidas. Obrigada por seres essa amiga, a vida é mais divertida contigo (ou às vezes, a rir da tua desgraça).

DD, a primeira amiga com quem saí à noite e bebi o meu 1º shot, até podia dizer que és má influência só por causa disso, mas estaria a mentir! És sim uma amiga para lá de excelente, a minha mana zebra, parceira das conversas de tecnologia que quero levar para o resto da vida!

Bia2, a amiga mais parecida comigo (para não dizer igual), até no nome. Não que seja necessário dizer, mas és a minha melhor amiga, a minha confidente, amiga de todas as horas, parceira de jogos do Porto, de noitadas, festas, concertos, jantares, compras, cinema, tudo e mais alguma coisa. Não existem palavras suficientes que descrevam o tanto que eu gosto de ti, és a irmã que nunca tive e estou muito agradecida por naquele dia do 6º ano, nos terem juntado. Nunca mais te larguei nem vou largar, aconteça o que acontecer. Juntas para sempre!

A todas vocês, obrigada! Cada uma à vossa maneira, foram muito importantes para terminar este capítulo. Obrigada por estarem comigo, quero levar-vos para a vida toda! Amo-vos!

Por último, mas não menos importante, um especial agradecimento ao meu namorado Pedro que, principalmente neste último ano, foi (e é) o meu pilar! Obrigada pela tua ajuda, pelo teu apoio incondicional e por me incentivares a não desistir. Obrigada por me dares a mão quando eu mais precisava, por não me deixares cair e, acima de tudo, por estares ao meu lado sempre. Não existem palavras suficientes que descrevam o quão grata estou por te ter comigo. O fim desta etapa deve-se a ti!

P.S: Tio Ângelo, o meu anjo da guarda, estejas onde estiveres, obrigada! As memórias que tenho tuas, as brincadeiras, piadas, conversas, sorrisos, caminhadas, férias e muitas outras, foram muitas vezes o meu combustível para terminar esta jornada. O meu maior sonho era que estivesses cá a celebrar este capítulo comigo, mas sei o quão feliz e orgulhoso de mim estás.

No final do dia, os "parabéns" serão para mim, mas o brinde será sempre a ti! Por teres sido muito mais do que um tio. Amo-te para sempre!

Resumo

O stress financeiro nas organizações pode manifestar-se através de eventos críticos, como falência, e a capacidade de prever esses eventos é crucial para a gestão de riscos e a tomada de decisões estratégicas. O presente estudo envolveu a aplicação e comparação de cinco modelos distintos de sobrevivência para prever eventos de stress financeiro: Regressão de *Cox, Random Survival Forest* (RSF), Kernel *SVM, Multi-Task Logistic Regression* (MTLR) e *DeepSurv*. Cada modelo foi selecionado com base nas suas características específicas e o seu potencial para lidar com dados de sobrevivência, oferecendo uma abordagem abrangente para a análise preditiva.

Este trabalho detalha também o processo de seleção e preparação dos dados, abordando todo o processo seguido desde a recolha dos dados até à análise de correlações entre variáveis. A identificação e remoção de variáveis altamente correlacionadas ajudaram a otimizar o desempenho dos modelos e a simplificar a interpretação dos resultados.

Os resultados obtidos indicam que todos os modelos aplicados foram eficazes na previsão de eventos de stress financeiro, com o RSF destacando-se pela sua *performance* superior. O estudo demonstra a aplicabilidade e a eficácia dos modelos de sobrevivência baseados em *Machine Learning* (ML) na identificação de riscos financeiros, oferecendo informações valiosas para a gestão financeira e a tomada de decisões estratégicas.

Em conclusão, este trabalho contribui para a literatura existente ao aplicar e comparar uma vasta gama de técnicas de ML de sobrevivência na previsão de eventos de stress financeiro. As descobertas oferecem uma base sólida para futuras pesquisas e práticas na área, enfatizando a importância da escolha adequada do modelo para a previsão e a gestão eficaz dos riscos financeiros.

Palavras-chave: Stress Financeiro; Análise de Sobrevivência; Modelos de *Machine Learning* de Sobrevivência

Abstract

Financial stress within organizations can manifest through critical events such as bankruptcy, and the ability to predict these events is crucial for risk management and strategic decision-making. This study involved the application and comparison of five distinct survival models to predict financial stress events: Cox Regression, Random Survival Forest, Kernel SVM, Multi-Task Logistic Regression (MTLR), and DeepSurv. Each model was selected based on its specific characteristics and potential to handle survival data, providing a comprehensive approach to predictive analysis.

This project also details the process of data selection and preparation, covering the entire process from data collection to correlation analysis between variables. The identification and removal of highly correlated variables helped to optimize model performance and simplify result interpretation.

The results obtained indicate that all applied models were effective in predicting financial stress events, with Random Survival Forest standing out for its superior performance. The study demonstrates the applicability and effectiveness of machine learning-based survival models in identifying financial risks, providing valuable insights for financial management and strategic decision-making.

In conclusion, this work contributes to the existing literature by applying and comparing a wide range of survival machine learning techniques for predicting financial stress events. The findings provide a solid foundation for future research and practice in the field, emphasizing the importance of selecting the appropriate model for effective prediction and management of financial risks.

Keywords: Financial Distress; Survival Analysis, Survival Machine Learning Models

Índice

Lista de figuras

Lista de quadros

Lista de abreviaturas

AUC – *Area Under the Curve*

CAE – Classificação Portuguesa das Atividades Económicas

C-Index – Índice de Concordância

CRISP-DM – *Cross Industry Standard Process for Data Mining*

DL – *Deep Learning*

IA – Inteligência Artificial

Kernel SVM – *Kernel Support Vector Machine*

KNN – *K-Nearest Neighbors*

ML – *Machine Learning*

MTLR – *Multi-Task Logistic Regression*

NaN – *Not a Number*

RSF – *Random Survival Forest*

SABI – *Iberian Balance Sheet Analysis System*

SVM – *Support Vector Machines*

1. Introdução

O rápido desenvolvimento tecnológico dos últimos anos tem revolucionado inúmeras áreas do conhecimento, e o setor financeiro não é exceção. Estes avanços tecnológicos têm transformado profundamente a área financeira, redefinindo a forma como as organizações lidam com riscos, tomam decisões e enfrentam desafios emergentes.

O stress financeiro pode-se manifestar de diversas formas, desde crises de liquidez e solvência até choques macroeconómicos e eventos sistémicos. As suas causas são multifacetadas, podendo ser desencadeadas por fatores económicos, políticos, regulatórios ou comportamentais. No entanto, independentemente da origem, a capacidade de prever e mitigar eventos de stress financeiro é fundamental para garantir estabilidade e resiliência de qualquer organização.

Neste contexto, a aplicação de modelos de *Machine Learning* (ML) surge como uma abordagem promissora para lidar com a complexidade e a incerteza associadas ao stress financeiro. Estes modelos têm a capacidade única de analisar grandes volumes de dados em tempo real, identificar padrões ocultos e antecipar mudanças no mercado financeiro antes que elas se tornem crises em larga escala.

1.1. Justificação do tema

No cenário económico atual, as empresas enfrentam uma série de desafios financeiros que podem comprometer a sua estabilidade e crescimento. Flutuações no mercado, mudanças nas políticas económicas, instabilidade geopolítica e outros fatores podem desencadear eventos de stress financeiro que impactam diretamente a saúde financeira das organizações. Diante desse contexto, é fundamental compreender e antecipar esses eventos para garantir a sustentabilidade e a resiliência das empresas.

Sun *et al.* (2014) referem que a previsão do stress financeiro é de grande importância para as empresas, pois permite uma gestão mais eficaz dos riscos financeiros e a adoção de medidas preventivas para evitar crises. Antecipar eventos de stress financeiro possibilita às empresas prepararem-se adequadamente, implementando estratégias de mitigação de

riscos, como a revisão de políticas de crédito, o ajuste de fluxo de caixa, a procura por fontes alternativas de financiamento e a diversificação de investimentos.

Além disso, os mesmos autores indicam que a previsão do stress financeiro é crucial para a manutenção da confiança dos investidores, credores e todos os *stakeholders*, demonstrando uma gestão prudente e responsável por parte da empresa. A capacidade de identificar e gerir proativamente os riscos financeiros contribui para a sustentabilidade do negócio e para a sua capacidade de enfrentar adversidades de forma resiliente.

Neste contexto, os modelos de ML surgem como uma ferramenta poderosa para prever eventos de stress financeiro nas empresas. Ao analisar grandes volumes de dados financeiros e económicos, estes modelos são capazes de identificar padrões e tendências ocultas que escapam à análise humana tradicional. Com algoritmos sofisticados, os modelos de ML podem antecipar eventos de stress com maior precisão e rapidez, permitindo que as empresas ajam proativamente para mitigar os seus impactos.

1.2. Objetivos

O objetivo geral desta dissertação é desenvolver e comparar modelos de ML para previsão de eventos de stress financeiro nas empresas. No entanto, e para atingir este objetivo, surgem uma série de objetivos específicos:

1. **Revisão da Literatura:** Realizar uma revisão abrangente da literatura relacionada com o stress financeiro, modelos de ML e as suas aplicações na previsão de eventos financeiros adversos.

2. **Recolha e Preparação de Dados:** Identificar e recolher dados relevantes para a construção e treino dos modelos de previsão de stress financeiro. Isto inclui dados financeiros históricos, indicadores económicos, informações setoriais e outras variáveis relevantes.

3. **Desenvolvimento de Modelos de ML:** Desenvolver e implementar diferentes modelos de ML para prever o stress financeiro nas empresas.

4. **Avaliação de Desempenho:** Avaliar o desempenho dos modelos desenvolvidos utilizando métricas apropriadas, como o índice de concordância.

5. **Interpretação dos Resultados:** Interpretar os resultados obtidos a partir dos modelos de ML, destacando informações relevantes para a compreensão dos fatores que contribuem para o stress financeiro nas empresas.

6. **Contribuição Académica e Prática:** Contribuir para o avanço do conhecimento académico no campo da previsão de stress financeiro, bem como fornecer orientações práticas para empresas, investidores e reguladores interessados em melhorar a gestão de riscos financeiros.

Os objetivos estabelecidos nesta dissertação são fundamentais para alcançar uma compreensão abrangente e robusta da previsão do stress financeiro nas empresas. A revisão da literatura proporciona uma base sólida de conhecimento teórico, enquanto a recolha e preparação de dados garantem a qualidade e a relevância das informações utilizadas nos modelos de ML. O desenvolvimento e a avaliação desses modelos permitem identificar abordagens eficazes para prever o stress financeiro, enquanto a interpretação dos resultados fornece informações valiosas sobre os fatores subjacentes ao fenómeno do stress financeiro.

1.3. Metodologia de trabalho

A metodologia adotada neste projeto foi o *Cross Industry Standard Process for Data Mining* (CRISP-DM). O CRISP-DM é um modelo padrão amplamente utilizado para a mineração e análise de dados e fornece uma estrutura sistemática e organizada para conduzir projetos de mineração de dados, garantindo uma abordagem coerente e eficiente desde a compreensão inicial do problema até a implementação das soluções encontradas.

Esta metodologia baseia-se na descrição de abordagens recorrentes e amplamente aplicadas por especialistas no campo da mineração de dados. O CRISP-DM é uma estrutura iterativa que divide o processo de mineração de dados em etapas distintas e bem

definidas. Esta segmentação permite que as equipas de projeto organizem de forma estruturada e sistemática todas as fases do trabalho, desde a compreensão inicial do problema até à implementação das soluções. A abordagem iterativa do CRISP-DM facilita a revisão contínua e a adaptação das estratégias, assegurando que os esforços sejam direcionados de forma eficiente e eficaz para atingir os objetivos do projeto.

Segundo Chapman (2000), as principais etapas do CRISP-DM são as seguintes:

1. **Compreensão de Negócio*:*** Nesta fase inicial, são estabelecidos os objetivos e requisitos do projeto. A equipa de projeto deve obter uma compreensão clara do contexto empresarial e das metas que a análise de dados pretende atingir. Isso envolve interações com os *stakeholders* para definir de forma precisa o problema a ser resolvido.

2. **Compreensão dos Dados*:*** Nesta fase, a equipa recolhe, analisa e avalia os dados disponíveis para o projeto. Isto engloba a identificação das fontes de dados, a obtenção dos dados necessários e a análise da qualidade e relevância desses dados.

3. **Preparação dos Dados*:*** Nesta fase, os dados recolhidos são tratados e organizados para serem analisados. Isto pode envolver a limpeza dos dados, a sua transformação e a seleção de características relevantes. O objetivo é criar um conjunto de dados que esteja pronto para a análise.

4. **Modelação*:*** Nesta etapa, a equipa desenvolve modelos de mineração de dados utilizando técnicas adequadas, como árvores de decisão, redes neuronais, regressão, entre outras. Estes modelos são ajustados e avaliados para alcançar o melhor desempenho possível.

5. **Avaliação*:*** Os modelos desenvolvidos são avaliados segundo métricas relevantes, como C-Index, *Brier Score* e AUC. Posteriormente, os modelos são avaliados verificando se os mesmos cumprem os objetivos do negócio.

6. **Implementação*:*** Os modelos validados são postos em funcionamento num ambiente de produção. Isto pode incluir o desenvolvimento de sistemas ou

aplicações que utilizam os resultados da mineração de dados para apoiar a tomada de decisões.

7. **Monitorização*:*** Após a implementação, é fundamental vigiar o desempenho dos modelos de forma contínua e realizar ajustes conforme os dados mudam ou evoluem.

O CRISP-DM é um modelo altamente flexível que reconhece e integra a natureza iterativa do processo de mineração de dados. Este modelo permite que as etapas sejam repetidas conforme necessário para aprimorar os modelos existentes ou incorporar novos conjuntos de dados. Tal abordagem adaptativa oferece uma estrutura robusta e dinâmica para o desenvolvimento de projetos de mineração de dados, auxiliando as organizações na obtenção de informações valiosas a partir dos seus dados e facilitando a tomada de decisões informadas e estratégicas.

Cada fase do CRISP-DM é descrita de forma detalhada e meticulosa, elucidando as atividades específicas que devem ser realizadas, os métodos e técnicas aplicadas, e os resultados esperados e obtidos. Esta documentação minuciosa fornece uma visão abrangente e clara do processo de análise de dados ao longo do desenvolvimento do projeto. Com isso, é possível entender profundamente como cada etapa contribuiu para o alcance dos objetivos definidos, e como as informações geradas podem ser utilizados para agregar valor e apoiar a tomada de decisões. O CRISP-DM, portanto, não só organiza e estrutura o processo de mineração de dados, mas também facilita a avaliação e a melhoria contínua dos modelos e das estratégias adotadas.

A Figura 1 representa o diagrama de Gantt para este projeto. Esta técnica funciona como uma ferramenta eficaz, que permite monitorizar o progresso da atividade em relação ao tempo. Dá também uma visão mais clara das tarefas, durações e dependências, o que facilita o planeamento e a organização das etapas do processo de desenvolvimento da investigação. Cada barra é uma representação de uma tarefa e a sua colocação no eixo do tempo permite ver os prazos de início e de fim, a sobreposição e os pontos críticos.

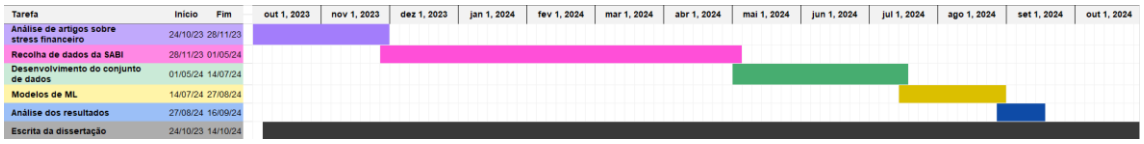


Figura 1 - Diagrama de Gantt

1.4. Resultados

Os resultados deste estudo mostraram que todos os modelos de sobrevivência aplicados foram eficazes na previsão de eventos de stress financeiro nas organizações. O modelo RSF destacou-se como o mais eficaz, alcançando o valor mais elevado de C-Index (0.89), indicando uma capacidade superior de discriminação entre diferentes tempos de ocorrência de eventos de stress financeiro.

O *Kernel* SVM e o MTLR apresentaram desempenhos semelhantes, ambos com um C-Index próximo de 0.87, demonstrando também uma elevada capacidade de previsão.

O modelo de *Cox* mostrou um bom desempenho com um C-Index de 0.86, reafirmando a sua relevância como um método clássico na análise de sobrevivência, apesar de não alcançar os resultados dos modelos mais avançados. Por outro lado, o *DeepSurv* foi o modelo com o desempenho mais modesto, atingindo um C-Index de 0.79.

Estes resultados ressaltam a eficácia dos modelos de ML de sobrevivência na previsão de eventos de stress financeiro, com destaque para o RSF como a melhor escolha para este tipo de análise, sendo que, com a utilização da técnica "*hyperparameter tuning"* foram descobertos os melhores parâmetros para este modelo e o seu valor de C-Index subiu ligeiramente para 0.8932.

1.5. Estrutura da dissertação

Este documento está estruturado em seis capítulos, cada um abordando aspetos específicos do projeto:

1. **Introdução:** Fornece uma introdução ao tema abordado, explicando o seu conceito e alguns dos objetivos que se pretende atingir.

2. **Revisão da literatura:** Oferece uma revisão abrangente da literatura relevante, estabelecendo as bases conceituais do tema abordado.

3. **Métodos e conceitos importantes:** Oferece uma revisão abrangente de conceitos teóricos importantes para o tema abordado nesta dissertação.

4. **Processo de Previsão de Stress Financeiro:** Descreve a metodologia adotada para a recolha, preparação e análise dos dados, bem como os procedimentos utilizados para o treino e validação dos modelos de ML.

5. **Análise dos Resultados:** Apresenta os resultados obtidos na aplicação dos modelos de ML para previsão do stress financeiro, destacando métricas de desempenho e informações relevantes.

6. **Conclusão:** Sintetiza as principais descobertas da pesquisa e destaca as suas contribuições, oferecendo recomendações para profissionais e pesquisadores interessados em explorar ainda mais o tema.

2. Revisão da literatura

O capítulo de revisão da literatura tem como objetivo oferecer uma visão completa dos estudos e teorias já existentes sobre o tema em discussão. A revisão é organizada para explorar e examinar as principais abordagens, métodos e descobertas anteriores que são relevantes para o campo de estudo, fornecendo uma análise crítica e contextualizada das contribuições de outros autores. Por meio desta análise, pretende-se identificar lacunas no conhecimento, entender novas tendências e avaliar de que forma as pesquisas anteriores influenciam e direcionam o estudo.

2.1. Definição de Stress Financeiro

O autor Beaver (1996) refere que as dificuldades financeiras incluem a incapacidade de pagar dívidas ou dividendos preferenciais e as consequências correspondentes, como a liquidação por interesse dos credores até a entrada em processo de falência. Como Beaver (1996) mencionou, uma empresa é como um reservatório formado pelo fluxo de caixa, composto por entradas e saídas de dinheiro. Uma empresa em dificuldades financeiras é como um reservatório cuja água é drenada.

Carminchael (1972) acredita que a dificuldade financeira é uma situação em que uma empresa encontra frustração no cumprimento das suas obrigações. Estas frustrações incluem: insuficiência de liquidez, insuficiência de capital próprio, incumprimento de dívidas e insuficiência de capital líquido. Foster (1986) definiu as dificuldades financeiras como um problema grave de liquidez que não pode ser resolvido sem uma restruturação em grande escala do funcionamento ou da estrutura das entidades económicas. Para Doumpos (1999) e Zopounidis (1999), as dificuldades financeiras não só incluem a incapacidade de reembolsar pagamentos obrigatórios importantes e as consequências acima mencionadas, como também incluem a situação de valor líquido negativo dos ativos, o que significa que o passivo total de uma empresa excede o seu ativo total do ponto de vista contabilístico.

Ross *et al.* (1999) resumiram estudos anteriores e concluíram que as dificuldades financeiras consistem nas quatro seguintes condições: (1) falência da empresa, ou seja,

uma empresa não pode pagar a dívida pendente após a liquidação; (2) falência legal, ou seja, uma empresa ou os seus credores solicitam ao tribunal uma declaração de falência; (3) falência técnica, ou seja, uma empresa não pode cumprir o contrato dentro do prazo para reembolsar o capital e os juros; (4) falência contabilística, ou seja, os ativos líquidos contabilísticos de uma empresa são negativos.

Bose (2006) definiu dificuldades financeiras como a condição em que o preço das ações de uma empresa é inferior a 10 cêntimos, o que é seguido por Ravisankar *et al.* (2010). Ao estudar a previsão do stress financeiro das empresas de Taiwan, Lin (2009) definiu dificuldades financeiras como a incapacidade de uma empresa pagar as suas obrigações financeiras à medida que estas se vencem. Em termos operacionais, diz-se que uma empresa está em situação de falência quando ocorre qualquer um dos seguintes eventos: falência, incumprimento de obrigações, eventos que significam uma incapacidade de pagar as dívidas à medida que estas se vencem, entrada num processo de falência, um acordo explícito com os credores para reduzir as dívidas, ou ser classificada como "ação de entrega total" pela Bolsa de Valores de Taiwan ou pelo *Gre Tai Securities Market*.

Os autores Ding (2008) e Sun *et al*. (2014) referem que, enquanto nos países em desenvolvimento, como a China e o Irão, as dificuldades financeiras são geralmente definidas como um certo grau de deterioração financeira determinado pela instituição de gestão da segurança nacional. Por exemplo, as dificuldades financeiras das empresas chinesas cotadas em bolsa são definidas como tratamento especial ("*Special Treatment"* - ST) pela Bolsa de Valores Chinesa, pelo facto dos seus lucros continuarem a ser negativos durante dois anos consecutivos ou de os seus ativos líquidos por ação serem inferiores ao valor nominal das ações. Os autores Rafiei *et al.* (2011), referem que as empresas iranianas cujas perdas acumuladas são superiores a 50% do seu capital são classificadas como empresas em dificuldades financeiras de acordo com a lei comercial 141 da Bolsa de Valores de Teerão.

Sun *el al.* (2014) referem que, para a definição de stress financeiro, existem muitos pontos de vista distintos. Os diferentes artigos e autores podem dar explicações diferentes de acordo com o seu próprio objetivo de estudo.

2.2. Condicionantes do Stress Financeiro

Habib *et al.* (2020) organizam os fatores que contribuem para o stress financeiro em três categorias: (1) condicionantes fundamentais ao nível da empresa; (2) condicionantes macroeconómicos; e (3) condicionantes da direção empresarial.

Habib *et al.* (2020) afirmam que ao nível dos fatores fundamentais da empresa diversas varáveis foram documentadas como afetando o stress financeiro como por exemplo: políticas de cobertura, relações com os funcionários, divulgação de narrativas de gestão, atividades de responsabilidade social corporativa, opiniões de auditoria qualificadas, entre outras. Magee (2013) conclui que a cobertura da moeda estrangeira reduz o risco de stress financeiro, uma vez que a cobertura pode minimizar a volatilidade do valor da empresa, sendo esta uma fonte proeminente de risco de stress financeiro. Além disso, a cobertura reduz a probabilidade de dificuldades financeiras, reduzindo os pagamentos de impostos e aumentando a capacidade de endividamento, como referido por Smith *et al.* (1985). Kane *et al.* (2005) concluem que as empresas com boas relações com os trabalhadores registam um menor risco de stress financeiro. Em caso de adversidade, as empresas podem renegociar com os trabalhadores na tentativa de obter concessões salariais temporárias. Também é possível que as empresas financeiramente saudáveis sejam capazes de fazer investimentos adequados nos funcionários, o que reforçaria as boas relações com estes. Tennyson *et al.* (1990) concluem que as divulgações de narrativas têm um poder explicativo incremental sobre a informação quantitativa baseada no estado financeiro na previsão de falências. As empresas que enfrentam potenciais falências são mais propensas a incluir narrativas que se concentram nas suas estratégias de sobrevivência e/ou recuperação.

Para Habib *et al.* (2020) é intuitivo prever que o risco de dificuldades financeiras aumenta durante períodos de recessão económica devido ao declínio das vendas, fluxos de caixa e da rentabilidade do negócio. Para Liou *et al.* (2007) a condição económica de um país também tem impacto no ambiente de negócios através das alterações nas taxas de inflação, taxas de juro, taxas de emprego, disponibilidade de crédito e política monetária. A investigação de Chordia *et al.* (2005) e Bonsall *et al.* (2013) sugere que a incorporação de fatores macroeconómicos nos modelos de previsão de dificuldades ao nível da empresa

melhora o seu poder explicativo, o que é consistente com a evidência de que as variáveis macroeconómicas representam quase metade da variação dos resultados das empresas.

Shleifer *et al.* (1997) referem que a governança empresarial é um conjunto de mecanismos através dos quais os investidores externos se protegem contra a desapropriação por parte dos investidores internos. Uma direção empresarial fraca cria uma oportunidade para os acionistas que detêm o controlo e os gestores beneficiarem de uma empresa às custas dos acionistas que não detêm o controlo, tal como mencionado por Habib *et al.* (2020).

Johnson *et al.* (2000) sugerem que a direção das empresas explica melhor as crises financeiras do que as variáveis macroeconómicas, pelo que a direção empresarial tem sido um dos temas mais discutidos pelos investigadores.

2.3. Consequências do Stress Financeiro

Habib *et al.* (2020) dividem as consequências do stress financeiro em quatro principais categorias: (1) consequências da informação financeira da auditoria; (2) consequências operacionais ao nível da empresa; (3) consequências para o mercado de capitais; (4) consequências para a administração empresarial.

As conclusões gerais deste tema sugerem que as empresas em dificuldades financeiras envolvem-se em escolhas contabilísticas que aumentam e diminuem o rendimento em diferentes fases de dificuldades.

De acordo com Habib *et al.* (2020) a emissão de um parecer de auditoria com reservas e a cobrança de honorários de auditoria mais elevados parecem ser as principais reações dos auditores externos às empresas em dificuldades. Do ponto de vista da auditoria externa, conclui-se que os auditores avaliam riscos empresariais mais elevados para as empresas com dificuldades financeiras e, consequentemente, emitem pareceres de auditoria mais elaborados.

No que diz respeito às consequências operacionais ao nível da empresa, os autores Habib *et al.* (2020) concluem que as empresas em dificuldades financeiras evitam impostos,

ajustam o montante das contas a receber e/ou a pagar e impõem custos de financiamento indiretos aos concorrentes sem dificuldades. Uma empresa em dificuldades financeiras pode escolher o planeamento fiscal como uma ferramenta adicional para superar a dificuldade, além disso, empresas em dificuldades financeiras com problemas de fluxo de caixa podem reduzir as condições de crédito comercial.

Habib *et al.* (2020) referem as consequências das dificuldades financeiras para o mercado de capitais concluem, em geral, que o risco de dificuldades representa um certo número de anomalias do mercado. Dentro desta principal categoria, os autores dividem em três categorias secundárias: (1) ajustamento de dividendos; (2) estratégias de reorganização; (3) anomalias de mercado. Relativamente, à primeira categoria de ajustamento de dividendos, Giroux (1984) e Wiggins (1984) concluem que as empresas em dificuldades financeiras prolongadas são suscetíveis de se confrontarem com acordos de dívida vinculativos e o ajustamento dos pagamentos pode ser um instrumento para atenuar as violações dos acordos. Sudarsanam (2001) e Lai (2001) demonstram que as empresas em dificuldades financeiras que não conseguiram passar o processo de recuperação são mais suscetíveis de se concentrarem na restruturação financeira, como a redução/omissão de dividendos e a restruturação financeira. Em relação à categoria de estratégias de reorganização, Jostarndt *et al*. (2008), referem que uma delas é a restruturação da dívida, através da qual as empresas em dificuldades financeiras podem aliviar o controlo dos credores, através de uma injeção de dinheiro, da redução ou adiamento de pagamentos contratuais ou de uma conversão da dívida em capital. Theodossiou *et al.* (1996) examinam as características específicas das empresas em dificuldades financeiras que atraem mais interesse na aquisição. De facto, concluem que as empresas em dificuldades financeiras com baixa eficiência de gestão são mais suscetíveis de serem adquiridas.

Por último, as consequências das dificuldades financeiras para a administração empresarial incluem más práticas de governação por parte das organizações em dificuldades, referido por Habib *et al*. (2020).

Gilson (1993) e Vetsuypens (1993) concluem que os *CEO's* que conseguem manter a sua posição numa empresa em dificuldades sofrem cortes salariais e de bónus. Guo *et al.* (2017) concluem que as empresas em dificuldades financeiras reduzem a remuneração do

CEO, intensificam a monitorização do conselho de administração e ajustam as nomeações dos gestores.

3. Métodos e Conceitos

Neste capítulo, será apresentado o enquadramento teórico dos dados utilizados para prever o stress financeiro nas empresas, fundamentando as bases conceptuais e metodológicas que sustentam esta investigação. Inicialmente, serão discutidas as principais teorias e modelos financeiros que explicam os fatores determinantes do stress financeiro, destacando as suas implicações práticas e académicas. Em seguida, será analisada a literatura existente sobre previsão financeira, abordando os métodos e técnicas mais utilizados para a identificação de sinais de stress financeiro. Este capítulo também explorará os diferentes tipos de dados financeiros e não financeiros relevantes para a análise, bem como as fontes de onde esses dados podem ser obtidos. Por fim, será feita uma revisão de casos estudos e pesquisas anteriores que aplicaram modelos preditivos em contextos semelhantes, fornecendo uma visão abrangente dos desafios e oportunidades associados à previsão do stress financeiro nas empresas. Este enquadramento teórico é essencial para contextualizar a abordagem metodológica e justificar as escolhas feitas ao longo do estudo.

3.1. *Machine Learning*

Para clarificar a estrutura da Inteligência Artificial (IA), a Figura 2 mostra a interconexão dos componentes principais a partir de uma perspetiva geral. Para Ribeiro (2022), a IA constitui o âmbito mais abrangente, englobando o ML e o *Deep Learning* (DL). O ML é uma subdivisão da IA, e o DL, por sua vez, é uma subdivisão do ML.

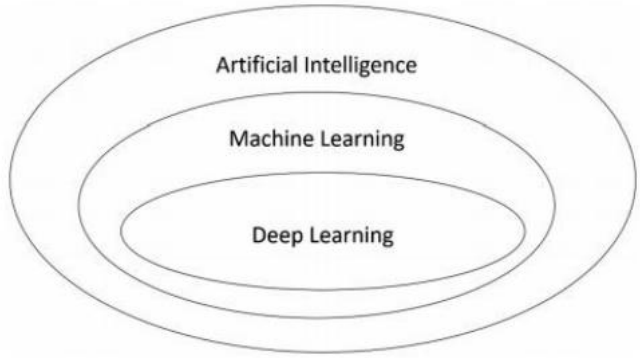


Figura 2 - Estrutura da IA
Fonte: *Ribeiro, P.M.F. (2022). Universidade do Minho School of Engineering Machine Learning Applied to Companies Management. Disponível em: https://hdl.handle.net/1822/84499*

O conceito de IA surgiu com a intenção de capacitar os computadores a realizarem tarefas semelhantes às do cérebro humano. A obtenção de um conjunto de dados de entrada é fundamental para criar parâmetros ou valores numéricos que permitem gerar dados de saída, normalmente expressos em probabilidades.

Monteiro *et al.* (2022) referem que a IA visa permitir que as máquinas realizem tarefas humanas. Isso implica dotar as máquinas com a capacidade de aprender, reconhecer, perceber e tomar decisões. Para tal, é essencial dispor de uma vasta quantidade de dados de alta qualidade, garantindo consistência, integridade, precisão e conformidade.

A IA é um conceito vasto que engloba diversas subáreas, incluindo o ML, que faz parte da IA, e o DL, que é uma ramificação do ML. O DL permite que os computadores reconheçam padrões, aprendam de forma contínua e façam previsões baseadas em dados, sem a necessidade de programação específica para cada tarefa. Esse processo automatiza a criação de modelos analíticos e possibilita que as máquinas se adaptem autonomamente a novos cenários, como explicam Monteiro *et al.* (2022).

Diversos especialistas e desenvolvedores acreditam que, num futuro próximo, a IA poderá superar a capacidade humana de aprender e raciocinar sobre qualquer assunto. No entanto, há quem permaneça cético, argumentando que toda atividade cognitiva está fundamentada em julgamentos de valor que são influenciados pela experiência humana. Atualmente, a IA possui um leque vasto de aplicações, abrangendo desde a saúde e as finanças até os jogos e outras áreas.

As técnicas de ML baseiam-se em reconhecimento de padrões, ciência da computação e inferência estatística. Estas abordagens são especialmente eficazes em domínios que trabalham com grandes volumes de dados, como finanças, economia e medicina. O principal objetivo dessas técnicas é gerar informações significativas e realizar previsões precisas a partir de dados históricos e eventos passados.

Nas abordagens tradicionais de programação, os desenvolvedores criam algoritmos que instruem o computador de forma precisa sobre como realizar uma tarefa. Por exemplo, para um programa reconhecer imagens de gatos, seria necessário elaborar um código detalhado que definisse as características específicas de um gato e orientasse o

computador a identificar essas características nas imagens. Contudo, essa abordagem é limitada, pois exige que os programadores antecipem e codifiquem todas as possíveis variações e situações que o programa pode encontrar.

Por outro lado, no ML, em vez de programar regras detalhadas, os algoritmos são concebidos para aprender a partir dos dados. Eles são treinados com um conjunto de dados de treino, que inclui exemplos com as suas respetivas entradas e respostas. O algoritmo analisa esses exemplos para identificar padrões e relações entre as entradas e as respostas, permitindo-lhe fazer previsões ou tomar decisões baseadas nas informações aprendidas.

Depois de treinado, o algoritmo pode fazer previsões ou tomar decisões com base em novos dados que não foram utilizados durante o processo de formação. Em essência, o algoritmo aperfeiçoa o seu desempenho numa tarefa específica à medida que é exposto a mais dados e adquire mais experiência.

Atualmente, é possível encontrar várias aplicações de ML à nossa volta. Por exemplo, *websites* que sugerem produtos, filmes e músicas com base nas nossas compras, visualizações ou audições anteriores. Os filtros de *spam* impedem que mensagens indesejadas cheguem à nossa caixa de correio. Os sistemas de análise de imagens médicas ajudam os profissionais de saúde a identificar tumores que poderiam passar despercebidos. E, claro, já existem carros autónomos equipados com tecnologia que permite conduzir sem intervenção humana.

3.2. *K-Fold Cross Validation*

Esta técnica é amplamente utilizada como um método eficaz para avaliar o desempenho de um modelo de forma mais precisa e confiável. Os autores Gorriz *et al.* (2024) explicam que esta abordagem consiste em dividir o conjunto de dados em K subconjuntos (ou *"folds"*) de tamanho aproximadamente igual, com o objetivo de garantir que todas as partes do conjunto de dados sejam utilizadas tanto para o treino quanto para a validação. O modelo é treinado K vezes, sendo que em cada uma dessas iterações, K-1 *folds* são usados para treino e o *fold* restante é destinado à validação. Este processo repete-se até que todos os *folds* tenham sido usados como conjunto de validação. No final, o

desempenho global do modelo é obtido calculando-se a média dos resultados de todas as iterações, o que permite uma avaliação mais robusta e confiável.

Gorriz *et al.* (2024) indicam que esta técnica tem várias vantagens significativas. Primeiramente, ao treinar e validar o modelo em diferentes subconjuntos, a técnica ajuda a reduzir a variabilidade que pode surgir em função de uma única divisão de treino e teste, oferecendo uma visão mais precisa da capacidade de generalização do modelo. Além disso, auxilia na prevenção do *overfitting*, pois o modelo é testado em diferentes porções do conjunto de dados.

3.3. *Overfitting*

O *overfitting*, segundo Sliusarenko *et al.* (2024), é um dos principais desafios ao treinar modelos de ML e refere-se à situação em que o modelo se ajusta excessivamente aos dados de treino, capturando não apenas os padrões relevantes, mas também o ruído e as peculiaridades específicas desse conjunto de dados. Quando isto acontece, o modelo torna-se altamente especializado nos dados que aprende durante o treino, o que resulta num desempenho excecional nesse conjunto, porém, ao ser testado em dados novos (como um conjunto de validação ou teste), ele falha em generalizar bem, apresentando um desempenho significativamente inferior.
Sliusarenko *et al.* (2024), explicam que o *overfitting* geralmente acontece quando o modelo é muito complexo, com um número elevado de parâmetros ou quando se utiliza uma quantidade insuficiente de dados de treino. Modelos mais complexos, como redes neuronais profundas, estão especialmente sujeitos a esse problema, pois possuem uma grande capacidade de ajuste e podem acabar por modelar até os detalhes irrelevantes dos dados. Além disso, o *overfitting* pode ser agravado por dados com ruído ou por *features* que não têm relação direta com a variável que se deseja prever.

Existem diversas técnicas que podem ser aplicadas para mitigar o *overfitting*, explicadas por Sliusarenko *et al.* (2024). A regularização, que adiciona penalidades a modelos excessivamente complexos, ajuda a reduzir o ajuste excessivo ao impedir que os parâmetros se tornem muito grandes. Outra estratégia é a utilização de mais dados de treino, o que facilita a identificação de padrões mais gerais em vez de detalhes específicos.

O *cross validation*, como mencionado anteriormente, também é uma ferramenta crucial para avaliar a capacidade de generalização do modelo e detetar *overfitting*, permitindo que o desempenho seja testado em diferentes subconjuntos dos dados.

3.4. Análise de sobrevivência

Wang *et al.* (2019) explicam que a análise de sobrevivência é uma ferramenta estatística que analisa e modela dados em que o resultado é o tempo até a ocorrência de um evento de interesse. Souza *et al.* (2022) sintetiza que, a análise de sobrevivência é essencialmente estudar a probabilidade de um determinado evento ocorrer num dado período de tempo.
Para Perrigot *et al.* (2004) a análise de sobrevivência dedica-se a estudar o risco de ocorrência de um evento específico ao longo do tempo. Esta abordagem estatística tem dois objetivos fundamentais. O primeiro é estimar a duração do período durante o qual o evento em questão pode acontecer, permitindo prever quando é mais provável que ocorra. O segundo objetivo é examinar e descrever a distribuição temporal desse evento, ou seja, analisar como o tempo afeta a sua probabilidade de ocorrência, e quantificar o impacto de diferentes fatores independentes, conhecidos como covariáveis, sobre essa distribuição. A recolha de dados, essencial para esta análise, é feita através de um registo longitudinal, que consiste no acompanhamento contínuo dos eventos que ocorrem em indivíduos, organizações, entre outros. Este registo detalhado permite captar informações precisas e ao longo do tempo, necessárias para realizar uma análise robusta e informada.

3.4.1. Conceitos principais da Análise de Sobrevivência

Neste subcapítulo, serão apresentados os principais conceitos que fundamentam esta técnica, fornecendo uma base sólida para a sua compreensão e aplicação em diferentes contextos.

De acordo com Souza *et al.* (2022) existem cinco conceitos importantes na análise de sobrevivência: evento/acontecimento, tempo, escala, censura e função de sobrevivência.

Para se compreender de forma mais aprofundada a análise de sobrevivência, é essencial começar por explicar o conceito de acontecimento. Formalmente, um acontecimento é

definido como "uma alteração de estado, determinada por uma ou mais variáveis qualitativas, que ocorre dentro de um período de observação e no espaço de estado relevante". Em termos simples, refere-se a qualquer mudança significativa de estado que possa ser monitorizada ao longo do tempo. De acordo com Perrigot *et al.* (2004) estas mudanças qualitativas são consideradas acontecimentos quando existe uma "rutura relativamente acentuada entre o que acontece antes e o que acontece depois" dessa mudança, num dado intervalo temporal.

Um exemplo típico de acontecimento pode ser a abertura de uma nova loja, o encerramento de uma loja existente, a conclusão de um contrato de trabalho, entre outros. Estes eventos são cruciais para a análise de sobrevivência, pois constituem os pontos de mudança que a análise procura estudar e prever.

Neste contexto, surgem três questões fundamentais: 1. O evento ocorreu? 2. Em que momento específico do tempo aconteceu? e 3. Como é que diferentes fatores, ou covariáveis, influenciam tanto a ocorrência como o momento exato em que o acontecimento se dá? Estas perguntas são centrais para a análise, permitindo não só descrever e compreender a natureza dos acontecimentos, mas também prever o impacto de várias condições sobre quando e como estes ocorrem.

Para Souza *et al.* (2022) o evento ou acontecimento é o objeto de análise em termos de se ou quando vai ocorrer.

A janela de medição define o intervalo de tempo durante o qual o investigador conduz as suas observações e recolhe dados. Colosimo (2006) indica que a escolha da duração dessa janela é uma decisão que cabe ao próprio investigador, sendo, portanto, subjetiva e muitas vezes arbitrária. Não existe uma base sólida de evidências teóricas ou empíricas que possam orientar claramente essa decisão, de acordo com Perrigot *et al*. (2004). Souza *et al.* (2022) explicam que o tempo é o período de tempo, contado a partir de uma determinada origem, até que o evento ocorra, sendo que a origem varia de acordo com o problema.

Em diferentes estudos que analisam organizações, pacientes, trabalhadores, entre outros, observa-se que a duração das janelas de medição pode variar amplamente. Por exemplo,

alguns estudos são realizados ao longo de apenas três ou quatro meses, enquanto outros se estendem por períodos muito mais longos, chegando a durar vários anos. Perrigot *et al*. (2004) mencionam que esta variação na duração da janela de medição não é trivial, pois pode ter um impacto significativo nos resultados obtidos. Devido à natureza arbitrária da escolha da duração, é crucial ter em mente que os resultados podem flutuar de acordo com o período de observação escolhido.

Esta variabilidade nos resultados, influenciada pela duração da janela de medição, é algo que se verifica em vários estudos, particularmente naqueles que investigam a rotatividade de pessoal e outros fenómenos dinâmicos. A consequência desta diversidade, de acordo com Perrigot *et al*. (2004), não é apenas a diferença nos resultados em si, mas também a dificuldade acrescida em comparar estudos entre si, uma vez que os períodos de observação distintos podem levar a conclusões diferentes, mesmo quando investigam questões semelhantes.

Em vários campos de estudo, especialmente nas ciências empresariais e gestão, é comum encontrarem-se dados que medem a duração entre dois eventos. Para os autores Perrigot *et al*. (2004), por exemplo, podemos estar a analisar o tempo decorrido entre duas compras realizadas por um cliente, a duração do emprego de um funcionário numa empresa, ou o tempo de operação de uma empresa até ao seu encerramento.

A escala é um dos conceitos importantes na análise de sobrevivência. Souza *et al.* (2022) referem que este conceito é o tamanho do intervalo de tempo que será utilizado na análise, que pode ser segundos, minutos, horas, dias, semanas, meses, anos, ou qualquer intervalo entre essas medidas.

Perrigot *et al.* (2004) referem que a censura é um conceito fundamental na análise de sobrevivência e refere-se a situações em que o tempo de sobrevivência é apenas parcialmente conhecido. Este fenómeno pode ocorrer por várias razões. Por exemplo, pode haver censura à esquerda, onde não conseguimos determinar a data de início do evento que estamos a estudar, como no caso de não saber quando exatamente uma pessoa fez uma compra. Alternativamente, pode haver censura à direita, onde não conseguimos identificar a data final do evento, como quando uma empresa ainda está em operação ao

final do período de observação. Também pode ocorrer censura à direita se perdermos a pista de um cliente, que desaparece da amostra antes do fim da janela de medição.

O termo "censurado" significa que se ignora a duração exata do evento, porque falta a data inicial e/ou a data final desse evento. Deste modo, considera-se que os dados censurados são dados incompletos. Para que os dados fossem completos, a informação teria de satisfazer três condições principais: 1. É necessário saber o tempo durante o qual o sujeito está exposto a um risco específico; 2. Devemos ser capazes de identificar o fim desse período de exposição; e 3. O fim do período deve ser devido a um evento que está a ser investigado. No entanto, muitas vezes essas condições não podem ser plenamente cumpridas, uma vez que o processo pode cessar por razões que não estão relacionadas com o evento de interesse, tal como indicam os autores Perrigot *et al*. (2004).

Além disso, os autores Perrigot *et al.* (2004) indicam que a censura pode ocorrer de duas formas principais: à direita e à esquerda. A censura à direita refere-se à situação em que não conseguimos determinar o ponto final do período entre os dois eventos em estudo. Isto significa que, apesar de se saber quando o evento começou, não se sabe quando ou se ele vai terminar, porque o processo ainda está em curso. Por exemplo, um dado censurado à direita pode surgir quando uma empresa ainda está ativa no mercado no fim do período de observação. Em termos mais gerais, a censura à direita ocorre quando o evento de interesse não aconteceu até ao final da janela de medição, ou ainda não aconteceu até ao presente momento.

Por outro lado, a censura à esquerda ocorre quando não se consegue determinar a data inicial do evento. Por exemplo, se estivermos a investigar a data em que um cliente adquiriu um produto, mas apenas sabemos que a compra ocorreu em algum momento antes do início da nossa janela de medição, isso é censura à esquerda. Neste caso, os autores Perrigot *et al.* (2004), referem que temos um ponto final conhecido, mas a data inicial do evento é desconhecida ou ocorreu antes do início do período que estamos a estudar.

Para os autores Perrigot *et al.* (2004), a censura representa um desafio significativo na análise de dados, particularmente quando se utilizam técnicas estatísticas tradicionais, devido à sua natureza incompleta. No entanto, a análise de sobrevivência oferece métodos

robustos para lidar com dados censurados, permitindo uma interpretação mais precisa e útil dos resultados, mesmo na presença de censura.

Os autores Souza *et al.* (2022) explicam que, em alguns casos, o período do estudo encerra-se e existem alguns indivíduos para os quais o evento nunca ocorreu, como por exemplo, o paciente não morreu durante o tempo de realização do estudo. Nestes casos costuma-se censurar os dados, ou seja, ao invés de retirar este indivíduo do estudo, regista-se o indivíduo com o tempo máximo de duração do estudo.

Por fim, o último conceito importante referente à análise de sobrevivência é a função de sobrevivência. Esta função, como indicam os autores Souza *et al.* (2022), é a probabilidade do evento objeto de estudo não ter ocorrido no momento *t*. A função é definida como, $S(t) = P(T > t)$, onde T é o tempo de ocorrência do evento e *t* é o tempo máximo que se deseja observar.

3.4.2. Metodologia da Análise de Sobrevivência

Neste subcapítulo serão explorados os princípios fundamentais da metodologia de análise de sobrevivência, com a introdução de algumas definições básicas essenciais para compreender esta abordagem. A análise de sobrevivência centra-se principalmente em dois conceitos-chave: a função de sobrevivência e a função de risco.

A função de sobrevivência refere-se à probabilidade de um indivíduo ou unidade de análise sobreviver além de um determinado tempo. Por outras palavras, ela indica a probabilidade de o evento de interesse ainda não ter ocorrido até um ponto específico no tempo. Por outro lado, função de risco (ou função de *hazard*) descreve a taxa instantânea de ocorrência do evento num determinado momento, dado que o indivíduo já sobreviveu até esse ponto.

Colosimo (2006) explica que os métodos utilizados na análise de sobrevivência podem ser classificados em dois tipos principais: não-paramétricos e semi-paramétricos. Os métodos não-paramétricos, como o método de *Kaplan-Meier*, são utilizados quando não há covariáveis envolvidas na análise. Este método permite estimar a função de sobrevivência sem assumir uma forma específica para a distribuição dos tempos de

sobrevivência. É particularmente útil para analisar dados onde o objetivo é simplesmente estimar a probabilidade de sobrevivência ao longo do tempo.

Por outro lado, Colosimo (2006) explica que, quando há covariáveis que podem influenciar a sobrevivência, recorre-se a métodos semi-paramétricos. O modelo de *Cox* é um exemplo clássico de abordagem semi-paramétrica. Este modelo permite a inclusão de múltiplas covariáveis e examina como essas variáveis independentes afetam a taxa de risco de ocorrência do evento. O modelo de *Cox* é flexível e não requer a especificação detalhada da forma da função de risco, sendo, portanto, adequado para situações em que se deseja avaliar o impacto de diferentes fatores sobre a sobrevivência.

Em resumo, a escolha do método adequado depende da presença ou ausência de covariáveis nos dados. Se os dados não contêm covariáveis, o método de *Kaplan-Meier* é geralmente suficiente e eficaz. No entanto, se existirem covariáveis que precisam de ser consideradas, o modelo de *Cox* oferece uma abordagem mais robusta para a análise.

Colosimo (2006) refere que quando se investiga eventos específicos, a variável dependente frequentemente representa o período até que o evento ocorra. A função de sobrevivência é, portanto, a probabilidade incondicional de que o evento ainda não tenha ocorrido até um determinado ponto no tempo *t*. Se denotarmos a função de sobrevivência por *S(t)*, podemos expressá-la da seguinte forma:

$$S(t) = Prob(T > t) = 1 - F(t) \qquad (1)$$

Aqui, *T* é a variável aleatória que representa o tempo até o evento, e *F(t)* é a função de distribuição acumulada da variável *T*. A função de distribuição acumulada, *F(t)*, dá a probabilidade de o evento ter ocorrido até o tempo *t*, ou seja, a probabilidade de o tempo até o evento ser menor ou igual a *t*. Assim, a função de sobrevivência é calculada subtraindo-se *F(t)* de 1:

$$S(t) = 1 - F(t) \qquad (2)$$

Além disso, a função de distribuição acumulada pode ser obtida a partir da função de densidade de probabilidade *f(t)*, que descreve a distribuição do tempo até o evento. A relação entre a função de sobrevivência e a função de densidade é dada por:

$$S(t) = 1 - \int_t^0 f(u)\, du \qquad (3)$$

Neste contexto, $\int_t^0 f(u)\, du$representa a probabilidade acumulada de o evento ter ocorrido até o tempo *t*. Assim, *S(t)* representa a probabilidade complementar, ou seja, a chance de o evento ainda não ter ocorrido até o tempo *t*.

Portanto, a função de sobrevivência fornece uma visão clara e quantitativa sobre a durabilidade ou persistência das unidades no estudo, oferecendo uma perspetiva essencial sobre a dinâmica do evento de interesse ao longo do tempo. Este conceito é amplamente utilizado em vários campos, incluindo a medicina, a economia, e as ciências sociais, para analisar o tempo até a ocorrência de eventos significativos.

Para compreender adequadamente os modelos de risco proporcional, é fundamental introduzir dois conceitos essenciais, explicados por Colosimo (2006). O primeiro conceito é o de conjunto em risco, que se refere ao grupo de unidades, sejam elas indivíduos, organizações, empregados, etc., que estão expostas à possibilidade de ocorrer o evento de interesse num determinado momento. Por exemplo, durante o primeiro período de tempo analisado (como um dia, semana, mês ou ano), todas as unidades da amostra são consideradas em risco.

O segundo conceito-chave é a taxa de risco, que representa a probabilidade condicional de um evento ocorrer para uma unidade da amostra num momento específico, dado que essa unidade está em risco nesse momento. Se designarmos essa taxa de risco por *h(t)*, ela pode ser expressa da seguinte forma:

$$h(t) = \lim_{dt \to 0} \frac{Prob(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)} \qquad (4)$$

Aqui, *f(t)* é a função de densidade de probabilidade do tempo até o evento e *S(t)* é a função de sobrevivência no tempo *t*. A fórmula demonstra como a taxa de risco é calculada como a razão entre a função de densidade de probabilidade e a função de sobrevivência. Esta taxa de risco é uma medida crucial para analisar como o risco de ocorrência do evento varia ao longo do tempo, fornecendo informações detalhadas sobre a dinâmica do evento de interesse dentro da amostra estudada.

Os métodos não paramétricos são particularmente eficazes para realizar uma análise exploratória inicial de dados de sobrevivência, pois não pressupõem um modelo específico para a taxa de risco, como explica Colosimo (2006). Isso significa que, ao utilizar estes métodos, não é necessário assumir uma forma predeterminada para como o risco varia ao longo do tempo, o que proporciona uma maior flexibilidade na análise.

Colosimo (2006) indica que estes métodos permitem que se estimem funções de sobrevivência separadas para diferentes grupos dentro de uma amostra. Por exemplo, se estamos a analisar dados de pacientes, empresas ou lojas, podemos calcular a probabilidade de cada um desses grupos ainda estar ativo ou "sobrevivente" após um determinado intervalo de tempo, que pode variar de dias a semanas, meses ou anos.

Através destes métodos, Colosimo (2006) refere que podemos comparar a função de sobrevivência entre os grupos e identificar se existem diferenças significativas. Por exemplo, podemos verificar se pacientes de diferentes categorias de tratamento têm taxas de sobrevivência distintas, se empresas de diferentes setores têm probabilidades variadas de continuidade no mercado, ou se lojas localizadas em diferentes regiões têm padrões diferentes de permanência abertas.

Essencialmente, os métodos não paramétricos permitem uma análise detalhada da probabilidade de um evento de interesse ainda não ter ocorrido após um determinado período. Esta abordagem é particularmente útil em situações onde não se conhece ou não se deseja assumir uma distribuição específica para o tempo até o evento, proporcionando uma visão inicial valiosa sobre o comportamento dos dados sem imposições excessivas sobre a estrutura do modelo.

Estes métodos permitem calcular a probabilidade de um paciente ainda estar vivo, de uma empresa continuar a operar no mercado ou de uma loja permanecer aberta após um determinado período de tempo, como dias, semanas, meses ou anos.

Para um intervalo específico *j*, onde *Qj* é a probabilidade condicional de sobrevivência após o tempo *j*, dado que o indivíduo, empresa ou loja sobreviveu até o tempo *j−1*, e *qj* é o estimador dessa probabilidade, a função de sobrevivência é calculada como:

$$S(t) = Qt * Qt - 1 * ... * Q1 \qquad (5)$$

E o estimador da função de sobrevivência é:

$$S^t = qt * qt - 1 * ... * q1 \quad (6)$$

Para determinar *qj,* o estimador da taxa de sobrevivência em cada intervalo *j*, utiliza-se as seguintes definições:

- *nj*: número de indivíduos, empresas, ou lojas que ainda estão em risco imediatamente antes do intervalo *j*.
- *mj*: número de eventos observados durante o intervalo *j* (como mortes para indivíduos, falências para empresas, ou encerramentos para lojas).

A taxa de sobrevivência estimada para o intervalo *j*, *qj*, é calculada pela fórmula:

$$Qj = \frac{nj - mj}{nj} \qquad (7)$$

Esta fórmula representa a proporção de unidades que continuam a "sobreviver" no intervalo *j*, após a ocorrência dos eventos observados.

3.4.3. Modelo *Random Survival Forest*

Ishwaran (2019) refere que o RSF é uma extensão do algoritmo *Random Forest* (RF), especificamente adaptado para lidar com dados de sobrevivência, que são caracterizados por observações associadas a um tempo até a ocorrência de um evento de interesse, como morte, recaída de uma doença ou falência de uma empresa. Esses dados podem incluir

censura, o que ocorre quando o evento de interesse não foi observado para todas as unidades de análise até o final do estudo. Vallarino (2024) explica que os RSF foram desenvolvidos para superar limitações dos métodos tradicionais de análise de sobrevivência, como o modelo de *Cox*, combinando a robustez e a capacidade de utilizar variáveis preditivas diversas dos RF com as necessidades específicas da análise de sobrevivência.

Os principais componentes dos RSF incluem uma floresta de árvores de decisão, onde as árvores são construídas para prever a função de sobrevivência em vez de uma variável de resposta categórica ou contínua. Ishwaran (2019) explica que durante a construção das árvores, as divisões dos dados são realizadas com base no critério do *log-rank*, que maximiza a diferença na função de sobrevivência entre os grupos criados por cada divisão. Este método permite que o RSF capture de forma eficaz a influência de variáveis preditivas sobre o tempo de sobrevivência. Além disso, o modelo incorpora o conceito de censura, incluindo as observações para as quais o evento de interesse não ocorreu até o término do estudo, de modo que essas observações não enviesem os resultados. A predição da função de sobrevivência para uma nova observação é obtida ao agregar os resultados de todas as árvores da floresta, cada uma gerando uma estimativa da função de sobrevivência, que são combinadas para formar uma predição final mais robusta.

Ishwaran (2019) indica que os RSF oferecem várias vantagens importantes em relação aos métodos tradicionais de análise de sobrevivência, como a capacidade de lidar com conjuntos de dados de alta dimensionalidade, onde o número de variáveis pode ser maior que o número de observações. Eles são também capazes de captar interações complexas entre variáveis sem a necessidade de especificar essas interações previamente, e são robustos a dados ruidosos, pois o impacto desses dados tende a ser minimizado na predição final. Além disso, os RSF fornecem predições dinâmicas da função de sobrevivência ao longo do tempo, permitindo uma visão detalhada de como o risco de ocorrência do evento varia com o tempo e com diferentes valores das covariáveis.

Como referido por Ishwaran (2019), este modelo é amplamente utilizado em áreas como medicina, biologia, economia e engenharia, onde a análise de dados de sobrevivência é crítica. Exemplos de aplicações incluem a previsão do tempo até a recaída em pacientes com cancro, a análise de durabilidade de produtos e o estudo do tempo até a falência de

empresas. Contudo, apesar das suas vantagens, o uso de RSF apresenta alguns desafios, como a complexidade na interpretação dos resultados e o impacto na compreensão do impacto de variáveis individuais, que pode ser mais difícil do que em modelos paramétricos tradicionais. Além disso, a criação e o treino de uma grande floresta de árvores de decisão podem ser computacionalmente exigentes, especialmente em grandes volumes de dados.

O modelo de RSF pode ser representado pela seguinte expressão:

$$h(t|x) = \left(\frac{1}{B}\right)\sum_{b=1}^{B} hb(t|x) \qquad (8)$$

Em que $hb(t|x)$ é a taxa de risco para um indivíduo com covariáveis x na árvore de decisão bth e B é o número de árvores na floresta aleatória, como explicado por Vallarino (2024).

3.4.4. Modelo *Kernel Support Vector Machine*

Hofmann *et al.* (2008) explicam que os *Kernel* SVM são uma extensão do método básico de *Support Vector Machine* (SVM), que é amplamente utilizado para classificação e regressão. O SVM clássico funciona bem para dados que são linearmente separáveis, ou seja, quando é possível traçar uma linha (ou um hiperplano em dimensões superiores) que separa perfeitamente as classes de dados. No entanto, muitos problemas do mundo real envolvem dados que não são linearmente separáveis, o que limita a aplicabilidade direta do SVM linear. Para superar essa limitação, os métodos de *Kernel* são introduzidos, permitindo que o SVM lide eficientemente com problemas de classificação não linear.
A ideia fundamental por trás dos *Kernel* SVMs, explicam Hofmann *et al.* (2008), é a transformação dos dados originais para um espaço de características de dimensão superior, onde se espera que as classes sejam linearmente separáveis. Essa transformação é realizada por meio de uma função de *Kernel*, que permite calcular produtos internos entre os dados nesse espaço transformado sem precisar de calcular explicitamente as coordenadas nesse novo espaço. Este processo é conhecido como o "truque do *Kernel*", e é central para a eficiência computacional do método. Por outras palavras, o *Kernel* permite que o SVM encontre uma superfície de separação não linear nos dados originais, mas de forma indireta e computacionalmente eficiente.

Hofmann *et al.* (2008) referem que a capacidade dos SVM's de lidar com margens de erro também é uma característica importante. Quando os dados não são perfeitamente separáveis, é permitido que alguns pontos estejam do lado "errado" do hiperplano separador, mas com uma penalização associada. Essa abordagem, conhecida como margem suave ou *soft margin*, permite que o modelo tenha uma melhor capacidade de generalização, evitando a sensibilidade excessiva a ruído ou a *outliers*.

Vallarino (2024) explica que os *Kernel* SVM's são amplamente aplicados em várias áreas, incluindo reconhecimento de padrões, análise de imagem, bioinformática e processamento de texto. Estes modelos são especialmente úteis em situações onde as relações entre as classes são complexas e não lineares, oferecendo uma combinação poderosa de flexibilidade e precisão. Contudo, é importante destacar que a interpretação dos resultados de um SVM com *Kernel* pode ser menos intuitiva do que a de métodos lineares, uma vez que as transformações para o espaço de características tornam o modelo menos transparente.

O modelo de *Kernel* SVM pode ser representado por:

$$f(x) = sign(\sum_{i=1}^{n} aiyiK(xi, x) + b) \qquad (9)$$

Em que $K(xi, x)$ é uma função de *kernel* que mede a semelhança entre os vetores de caraterísticas xi e x, yi é o rótulo da classe da i-ésima instância, αi são os pesos dos vetores de apoio e b é o enviesamento, como explicado por Vallarino (2024).

3.4.5. Modelo *Multi-task Logistic Regression*

Bisaso *et al.* (2018) referem que os modelos de sobrevivência MTLR surgem como uma extensão tanto da regressão logística quanto dos modelos de sobrevivência, sendo adaptados para lidar com múltiplas tarefas ou eventos simultaneamente. Estes modelos são particularmente úteis quando há interesse em modelar diferentes tipos de eventos ou aspetos do tempo até a ocorrência de eventos, proporcionando uma abordagem mais abrangente e flexível.

A essência dos modelos MTLR reside na aplicação de uma série de regressões logísticas ao longo de diferentes intervalos de tempo, modelando a probabilidade de sobrevivência em cada um desses intervalos. Esta abordagem difere dos modelos de *Cox*, que se baseiam nas razões de risco (*hazard ratios*). No MTLR, o tempo é dividido em intervalos discretos, e para cada intervalo uma regressão logística é aplicada. A função de sobrevivência é então obtida a partir das probabilidades de sobrevivência em cada um desses intervalos.

Os modelos MTLR têm ampla aplicação em diversas áreas, incluindo medicina, onde podem ser utilizados para modelar o tempo até diferentes eventos clínicos, como morte, recaída de doença ou hospitalização. Na engenharia, são utilizados em análises de confiabilidade, onde é possível modelar simultaneamente múltiplas falhas ou tipos de falhas.

3.4.6. Modelo *DeepSurv*

Wang *et al.* (2024) referem que os modelos *DeepSurv* representam uma aplicação inovadora das técnicas de DL na análise de sobrevivência, uma área tradicionalmente dominada por métodos estatísticos como o modelo de *Cox*. *DeepSurv* é essencialmente uma extensão do modelo de riscos proporcionais de *Cox*, mas utiliza redes neuronais profundas para modelar a relação não linear entre as covariáveis e o risco de ocorrência de um evento.
Vallarino (2024) explica que o modelo *DeepSurv* é projetado para superar as limitações dos modelos de sobrevivência tradicionais, que frequentemente assumem relações lineares entre as covariáveis e o tempo até o evento. Essa suposição de linearidade pode ser restritiva em muitos cenários reais, onde as relações entre as variáveis preditivas e o risco de evento são complexas e não lineares. As redes neuronais profundas, com a sua capacidade de capturar essas relações complexas, oferecem uma solução robusta para essa limitação.

Wang *et al.* (2024) explicam que a arquitetura básica do *DeepSurv* envolve o uso de redes neuronais profundas para aprender uma função de risco a partir das covariáveis. Essa função de risco é uma generalização do que é utilizado no modelo de *Cox*, mas em vez de ser linear, é modelada através das camadas da rede neuronal. Essas camadas permitem

que o modelo capture interações não lineares e complexas entre as covariáveis, resultando numa previsão de risco mais precisa e adaptada às particularidades dos dados.

Uma das principais vantagens do *DeepSurv* é sua capacidade de lidar com dados de alta dimensionalidade e complexidade, algo que é comum em áreas como a biomedicina, onde o número de covariáveis pode ser muito grande e as interações entre elas são complexas. Por exemplo, em estudos genéticos, onde milhares de genes podem ser considerados como covariáveis, os métodos tradicionais de sobrevivência muitas vezes falham em capturar toda a complexidade dos dados. O *DeepSurv*, por outro lado, pode integrar essas informações de forma mais eficaz, oferecendo previsões de risco mais precisas.

O treino do modelo *DeepSurv* envolve a minimização de uma função de perda que é baseada na função de verossimilhança parcial do modelo de *Cox*, tal como é explicado por Wang *et al.* (2024). Essa função de perda é adaptada para o contexto das redes neuronais, permitindo a otimização dos pesos da rede durante o processo de treino.

Além da sua aplicação em biomedicina, o *DeepSurv* tem sido utilizado em diversas outras áreas, como a análise de confiabilidade de sistemas, onde a previsão do tempo até a falha de componentes pode ser crítica. Em cada um desses contextos, o *DeepSurv* demonstra a sua flexibilidade e poder ao lidar com diferentes tipos de dados e estruturas de covariáveis.

3.4.7. As condições de utilização e os limites da Análise de Sobrevivência

Neste subcapítulo, serão discutidas as condições necessárias para a aplicação adequada da análise de sobrevivência, bem como os seus principais limites. Compreender esses fatores é fundamental para garantir a precisão dos resultados e evitar interpretações incorretas das análises.

A análise de sobrevivência é adequada para problemas que envolvem eventos específicos. Este método é particularmente eficaz para quatro tipos de dados:

1. **Variabilidade Temporal:** As variáveis dependentes mudam ao longo do tempo.

2. **Distribuição das Variáveis:** Não se exige que as variáveis dependentes sejam distribuídas normalmente, nem que sejam distribuídas de forma independente ou idêntica.

3. **Censura:** A presença de censura, que ocorre quando os dados não fornecem uma observação completa do evento, não representa um problema significativo para a análise.

4. **Dados Longitudinais:** Os dados são recolhidos ao longo de vários períodos de tempo, permitindo observar mudanças e eventos ao longo do tempo.

Estas características fazem da análise de sobrevivência uma ferramenta valiosa para compreender e modelar o tempo até a ocorrência de eventos, independentemente das condições de distribuição e da presença de dados censurados.

Alguns desafios enfrentados pelos investigadores nas ciências sociais surgem devido à aplicação de modelos de duração. Um dos principais problemas é que essas técnicas são frequentemente apresentadas na literatura estatística, que abrange várias tradições e áreas de estudo. Como resultado, há uma grande variedade de terminologias e métodos, o que pode causar confusão e dificultar a aplicação consistente dessas técnicas em contextos de negócios.

O problema associado à janela de observação refere-se à forma como os resultados do estudo são interpretados, o que pode mudar dependendo da duração da janela escolhida para a análise. Se a janela de observação for demasiado curta, pode não capturar informações suficientes para uma análise completa. Por outro lado, se for excessivamente longa, pode incluir dados irrelevantes ou desatualizados, o que também pode afetar a interpretação dos resultados. Assim, a escolha da duração da janela de observação é crucial para garantir uma interpretação precisa e significativa dos dados.

Os problemas associados à censura à direita afetam a validade interna dos estudos, pois não é possível garantir que os resultados seriam os mesmos se a variável dependente, que representa o estado estudado, tivesse sido avaliada num momento diferente. O mesmo se aplica à censura à esquerda. Apesar de a análise de sobrevivência ter sido desenvolvida

para lidar com casos censurados, há uma carência de literatura sobre este fenómeno. É difícil entender como cada método lida com os casos censurados; frequentemente, são apresentadas fórmulas complexas sem explicações adequadas ou comparações com outros métodos. Além disso, não há estudos teóricos sobre a sensibilidade dos resultados à taxa de censura, nem limites definidos que indiquem quando a censura pode comprometer a fiabilidade dos resultados.

3.4.8. Métricas de desempenho dos modelos de sobrevivência

As métricas de desempenho possuem um papel crucial na avaliação de modelos de sobrevivência, permitindo quantificar a precisão e a capacidade preditiva dessas ferramentas estatísticas. No presente trabalho, foram utilizadas três métricas principais para avaliar o desempenho dos modelos: o Índice de Concordância (C-Index), que mede a capacidade de discriminação do modelo em ordenar corretamente os tempos de eventos; o *Brier Score*, que avalia a calibração e a precisão das probabilidades preditas ao longo do tempo; e a área sob a curva ROC (AUC), que fornece uma visão adicional sobre o desempenho discriminativo em diferentes pontos de tempo.

Este subcapítulo tem como objetivo descrever e explicar em detalhe cada uma destas métricas de desempenho utilizadas neste projeto, com o objetivo de avaliar a qualidade dos modelos de sobrevivência desenvolvidos.

3.4.8.1. Índice de Concordância

O C-Index é uma métrica amplamente utilizada para avaliar o desempenho de modelos de sobrevivência, cujo objetivo é prever o tempo até a ocorrência de um evento de interesse, como por exemplo, a morte, a falha de um equipamento ou a recorrência de uma doença. Segundo Longato *et al.* (2020), o C-Index mede a capacidade do modelo em realizar previsões corretas de ordem relativa para os tempos de sobrevivência, ou seja, avalia quão bem o modelo classifica corretamente os indivíduos em termos de "quem sobreviverá por mais tempo".

Longato *et al.* (2020) referem que o funcionamento do C-Index se baseia no conceito de concordância entre as previsões do modelo e os dados observados. Para isso, considera-

se a comparação entre pares de indivíduos. Caso o modelo preveja que um indivíduo possui maior risco do que o outro (isto é, menor tempo de sobrevivência), e essa previsão corresponder à realidade observada (isto é, o indivíduo com maior risco realmente experimenta o evento primeiro), a predição é considerada "concordante".

Matematicamente, o C-Index é calculado como a razão entre o número de pares de indivíduos corretamente classificados (concordantes) e o número total de pares comparáveis. Um par de indivíduos é considerado comparável quando, no mínimo, um dos dois experimenta o evento de interesse, ou seja, não está censurado. Já um par é considerado concordante quando o indivíduo que teve o evento mais cedo também foi aquele que o modelo previu como tendo o maior risco.

A interpretação do C-Index varia entre 0 e 1. Longato *et al.* (2020) explicam que um valor de C-Index igual a 1 indica que o modelo é perfeitamente capaz de ordenar corretamente os tempos de sobrevivência para todos os pares de indivíduos comparáveis. Um valor de C-Index igual a 0,5 indica que o modelo tem desempenho equivalente ao acaso, ou seja, não discrimina os riscos entre os indivíduos de forma útil. Já um C-Index abaixo de 0,5 sugere que o modelo frequentemente faz previsões inversas à realidade, sendo menos eficaz que um modelo aleatório.

No contexto dos modelos de sobrevivência, como o Modelo de *Cox*, o C-Index é particularmente relevante, pois avalia a habilidade do modelo em ordenar corretamente os riscos relativos entre diferentes indivíduos. Esse índice também apresenta a vantagem de ser robusto frente à censura, uma característica comum em dados de sobrevivência, onde o evento de interesse não é observado em todos os indivíduos até o final do estudo.

Contudo, apesar da sua ampla utilização, o C-Index possui algumas limitações. Esta métrica não tem em consideração o tempo absoluto de sobrevivência, focando exclusivamente na ordenação dos eventos. Além disso, o índice pode ser insensível a diferenças sutis nos tempos de sobrevivência previstos, concentrando-se apenas na classificação dos indivíduos. Diante disso, o C-Index é uma métrica importante e poderosa para medir o desempenho relativo dos modelos de sobrevivência, mas pode ser complementado com outras métricas que considerem o tempo absoluto e outras características dos dados.

3.4.8.2. *Brier Score*

O *Brier Score* é uma métrica amplamente utilizada para avaliar o desempenho de modelos preditivos, incluindo os modelos de sobrevivência. Segundo Assel *et al.* (2017), o *Brier Score* mede a precisão das probabilidades previstas para um determinado evento em comparação com os resultados observados, funcionando como uma medida de erro quadrático médio entre as probabilidades previstas pelo modelo e os resultados reais.
No contexto de modelos de sobrevivência, o objetivo é prever a probabilidade de um indivíduo sobreviver até um tempo específico *t*. Assel *et al.* (2017) explicam que o cálculo do *Brier Score* envolve a média do erro quadrático entre a probabilidade de sobrevivência prevista pelo modelo e o valor observado, considerando tanto os indivíduos que experimentaram o evento quanto aqueles que foram censurados, ou seja, para os quais o evento não foi observado até o fim do estudo.

Assel *et al.* (2017) referem que o *Brier Score* varia de 0 a 1, sendo que um valor de 0 indica previsões perfeitas, onde o modelo previu exatamente o que aconteceu para todos os indivíduos, enquanto um valor de 1 indica um desempenho muito mau, com previsões completamente incorretas. Valores intermédios refletem diferentes graus de precisão, com valores menores indicando melhor desempenho do modelo.

Uma das vantagens do *Brier Score* é que oferece uma medida direta da precisão das previsões probabilísticas feitas pelo modelo. Enquanto métricas como o C-Index avaliam a capacidade do modelo de ordenar corretamente os tempos de sobrevivência de diferentes indivíduos, o *Brier Score* foca na exatidão das probabilidades absolutas de sobrevivência em momentos específicos. Além disso, o *Brier Score* é ajustado para lidar com dados censurados, utilizando técnicas como o método de *Kaplan-Meier*, que ajusta a métrica para compensar a falta de informações completas sobre os indivíduos censurados. Desta forma, o *Brier Score* é uma métrica robusta e adequada para avaliar modelos em cenários onde a censura é frequente, como nos estudos de sobrevivência.

Apesar da sua utilidade, o *Brier Score* apresenta algumas limitações. Uma delas é a sua dependência de um ponto específico no tempo, o que pode tornar a métrica menos informativa se não for calculada em vários momentos relevantes ao longo do período de

estudo. Além disso, Assel *et al.* (2017) explicam que, por ser uma média de erros quadráticos, o *Brier Score* é mais sensível a previsões extremas ou altamente imprecisas, ampliando o impacto dessas previsões na métrica final.

Resumindo, o *Brier Score* é uma métrica poderosa para avaliar o desempenho de modelos de sobrevivência, fornecendo uma análise detalhada da calibração e da precisão das previsões probabilísticas. A sua robustez em lidar com censura e a sua capacidade de avaliar diretamente a precisão das previsões tornam-no uma ferramenta valiosa na análise de modelos de sobrevivência, especialmente quando utilizado em conjunto com outras métricas, como o C-Index, para fornecer uma visão mais abrangente da *performance* do modelo.

3.4.8.3. AUC

O AUC é uma métrica amplamente utilizada para avaliar o desempenho de modelos preditivos, incluindo os modelos de sobrevivência. No contexto dos modelos de sobrevivência, a AUC é adaptada para lidar com a censura dos dados e mede a capacidade do modelo em discriminar entre indivíduos que irão experimentar o evento de interesse e aqueles que não irão.

Para calcular a AUC em modelos de sobrevivência, é comum utilizar a Curva ROC (*Receiver Operating Characteristic*), que segundo Huang *et al.* (2005) é uma representação gráfica do desempenho do modelo ao variar o ponto de corte para classificar eventos como positivos ou negativos. No contexto de sobrevivência, o ponto de corte pode variar com o tempo, refletindo a capacidade do modelo de classificar corretamente o risco relativo de eventos futuros. A Curva ROC é gerada ao representar a Taxa de Verdadeiros Positivos (Sensibilidade) versus a Taxa de Falsos Positivos (1 - Especificidade) em diferentes pontos de corte.

A AUC é calculada como a área sob a Curva ROC e varia entre 0 e 1. Huang *et al.* (2005) explicam que um valor de 1 indica que o modelo é perfeito na discriminação entre eventos e não eventos, ou seja, pode classificar todos os casos corretamente. Um valor de 0,5 sugere que o modelo não é melhor do que um classificador aleatório, enquanto uma AUC

abaixo de 0,5 indica um desempenho inferior ao acaso, onde o modelo frequentemente faz previsões incorretas.

Em modelos de sobrevivência, a adaptação da AUC para lidar com censura é crucial. A censura ocorre quando o evento de interesse não foi observado para alguns indivíduos até o fim do estudo, o que pode complicar o cálculo da AUC. Para lidar com isso, é comum utilizar o C-Index, que é uma adaptação da AUC para dados de sobrevivência.

Huang *et al.* (2005) explicam que a interpretação da AUC em modelos de sobrevivência segue princípios semelhantes aos da AUC em contextos de classificação binária. Uma AUC alta indica que o modelo é eficaz na discriminação entre aqueles que irão experimentar o evento e aqueles que não irão, mesmo com a presença de censura. Por outro lado, uma AUC baixa sugere que o modelo tem dificuldade em discriminar corretamente entre eventos e não eventos.

Embora a AUC seja uma métrica valiosa, esta possui algumas limitações. A principal limitação é que a AUC não considera a magnitude do risco, apenas a capacidade do modelo em discriminar entre eventos e não eventos. Além disso, a AUC pode ser afetada pela presença de censura e a interpretação pode ser mais complexa em dados com alta taxa de censura.

4. Processo de Previsão de Stress Financeiro

Neste capítulo são abordadas as fases da metodologia CRISP-DM e as tarefas incluídas em cada uma das fases. A qualidade e integridade dos dados são fundamentais para a precisão das previsões e, por isso, é necessário realizar recolha, limpeza, transformação e integração dos dados provenientes de diversas fontes. A preparação dos dados inclui a identificação e tratamento de valores em falta, a normalização e padronização das variáveis, bem como a criação de novas variáveis, como indicadores financeiros que sejam relevantes para a análise. Além disso, são exploradas técnicas de manipulação de dados que permitam uma análise mais profunda e a extração de características essenciais que alimentam os algoritmos de ML. Neste capítulo são então estabelecidas as bases para a construção de um modelo robusto e confiável, essencial para a antecipação de situações de stress financeiro e para a implementação de estratégias de mitigação de risco nas empresas.

4.1. Compreensão de Negócio

Este projeto tem como foco enfrentar o desafio do stress financeiro nas organizações que se pode manifestar de várias formas. O impacto desse stress é significativo, pois pode resultar em falências ou na redução das operações, salientando a importância de prever e mitigar esses riscos.

O projeto visa desenvolver modelos de sobrevivência capazes de prever o nível de stress financeiro das empresas ao longo do tempo. Isso permitirá identificar as empresas em risco e aplicar estratégias eficazes para melhorar a sua saúde financeira.

O objetivo do projeto inclui a análise de dados financeiros históricos para criar modelos preditivos e gerar informação de valor. Entre os possíveis desafios estão a qualidade dos dados disponíveis e a resistência das empresas em adotar as soluções propostas, os quais serão enfrentados com o uso de técnicas robustas de modelação.

4.2. Compreensão dos Dados

A primeira tarefa, e uma das mais importantes, a ser realizada no âmbito deste projeto foi a recolha de dados. Esta etapa torna-se de grande importância, uma vez que os dados recolhidos, bem como a sua qualidade e quantidade, podem determinar se o projeto terá sucesso ou não. A forma como esta tarefa inicial é conduzida pode ter um impacto significativo em todas as fases seguintes do projeto, influenciando diretamente a sua viabilidade e os resultados a serem alcançados.

Conforme mencionado anteriormente, os dados necessários para o projeto em questão dizem respeito a organizações reais e foram extraídos de uma base de dados autêntica, a SABI (*Iberian Balance Sheet Analysis System*; Informa D&B, n.d.), que é disponibilizada pelo ISCAP (Instituto Superior de Contabilidade e Administração do Porto). A SABI é uma base de dados abrangente que contem uma vasta gama de informações sobre empresas situadas em Portugal e Espanha. Entre os dados disponíveis, encontram-se detalhes identificativos da organização, como o número de contribuinte, o número de colaboradores, a morada e outros dados similares que permitem uma identificação clara e precisa das entidades empresariais. Adicionalmente, a SABI oferece um conjunto de dados financeiros essenciais para a análise de desempenho das empresas. Estes dados incluem o resultado líquido do período, que indica o lucro ou prejuízo obtido pela empresa, bem como o valor do ativo, que representa os recursos controlados pela empresa. Também estão disponíveis informações sobre o passivo, que são as obrigações financeiras da empresa, e o capital próprio. Estes dados financeiros, entre outros tantos disponíveis na SABI, são cruciais para a avaliação da saúde financeira das empresas e para a realização de análises mais aprofundadas no âmbito do projeto. Portanto, a utilização desta base de dados robusta e detalhada permite que o projeto se baseie em informações concretas e atualizadas, proporcionando uma base sólida para o desenvolvimento de análises e conclusões precisas e relevantes.

Esta tarefa de recolha de dados revelou-se bastante extensa e demorada, devido a vários fatores que complicaram o processo. A base de dados SABI apresentava uma limitação técnica que só permitia a extração de um número reduzido de dados de cada vez. Esta restrição implicou a necessidade de realização de múltiplas operações de extração. O

resultado foi a acumulação de um grande número de ficheiros Excel, cada um contendo uma parte dos dados extraídos.

Após a conclusão do processo de recolha de dados, verificou-se que muitos dos dados recolhidos possuíam valores nulos. Esta elevada quantidade de valores nulos impediu o avanço imediato para a próxima etapa do projeto, pois a ausência de dados completos comprometeria a precisão das análises subsequentes.

O conjunto de dados obtido a partir da base de dados SABI é composto com informação de empresas portuguesas do setor da indústria transformadora, abrangendo o período de 2011 a 2022. Para assegurar que as empresas selecionadas pertenciam realmente ao setor da indústria transformadora, foram aplicados critérios baseados na Classificação Portuguesa das Atividades Económicas (CAE, revisão 3). Segundo esta classificação, as atividades relacionadas com a indústria transformadora estão incluídas nos códigos CAE 11 a CAE 33. Com a aplicação rigorosa destes critérios, foi possível estabelecer um conjunto de dados composto por 65 195 empresas.

Assim, o conjunto de dados desenvolvido é composto por 65195 linhas e 615 colunas, sendo elas:

Tabela 1 - Colunas do conjunto de dados

Nome da coluna	Descrição
Nº Contribuinte	Número de identificação fiscal da empresa
Total_Divida	Soma total das dívidas da empresa
Total_do_ativo	Valor total dos ativos da empresa
Crescimento_do_ativo	Taxa de crescimento do valor total dos ativos ao longo do tempo

Total_do_passivo_corrente	Soma dos passivos com vencimento em curto prazo
Fluxos_de_Caixa_das_Atividades_Operacionais	alor líquido gerado ou consumido pelas atividades operacionais
Caixa_e_seus_equivalentes	Recursos em caixa e ativos altamente líquidos
Depositos_bancarios_e_Caixa	Soma dos depósitos bancários e caixa disponível
Rácio_de_Tesouraria	Indicador de liquidez baseado em caixa, equivalentes e passivos a curto prazo
Ativos_Correntes	Total de ativos com liquidez elevada ou realizáveis em curto prazo
Passivos_curto_prazo	Total de obrigações financeiras a serem liquidadas em curto prazo
Rácio_Corrente	Indicador de liquidez que compara ativos correntes com passivos de curto prazo

Total_do_passivo	oma total das obrigações financeiras da empresa
Rácio_da_Divida_em_relação_ao_Capital_Próprio	Proporção entre dívida total e capital próprio
EBITDA	Lucros antes de juros, impostos, depreciação e amortização
Dividas_a_Terceiros_MLP	Dívidas a terceiros com vencimento em longo prazo
Divida_a_Curto_Prazo	Obrigações financeiras com vencimento em curto prazo
EBITDA_para_Divida_a_Curto_Prazo_e_Juros	Indicador de capacidade de pagamento de dívidas e juros
Crescimento_do_Capital_Próprio	Taxa de aumento do capital próprio da empresa
Rendimento_Bruto_do_Ativo	Retorno bruto obtido pelos ativos
Resultado_Liquido_do_Exercicio	Lucro líquido apurado no exercício financeiro
Taxa_de_Crescimento_do_Rendimento	Taxa de crescimento do rendimento ao longo do tempo

Fluxos_de_caixa	Total dos fluxos de caixa gerados ou consumidos
Taxa_de_Crescimento_do_Fluxo_de_Caixa_Líquido	Aumento percentual do fluxo de caixa líquido
Custo_Mercadorias_Vendidas_e_Materias_Consumidas	Custo direto de mercadorias vendidas e matérias-primas
Inventarios	Valor dos *stocks* da empresa
Período_de_Rotação_do_Inventário	Tempo médio para converter estoques em vendas
Capital_Proprio	Recursos próprios disponíveis da empresa
Dívidas_a_Longo_Prazo_em_relação_ao_Capital_Próprio	Proporção de dívidas de longo prazo sobre o capital próprio
Nome	Nome da empresa
Localidade	Local onde a empresa está situada
País	País de registo da empresa
Região	Região geográfica onde a empresa opera
Código CAE	Código de Atividade Económica da empresa.

Data_de_constituição	Data de fundação da empresa
Estado	Estado atual da empresa (ativa, inativa, etc.)
Número_Empregados	Quantidade de empregados
Número_de_Diretores_e_Gestores_Atuais,	Total de diretores e gestores da empresa
Dívidas_a_Longo_Prazo_em_relação_ao_Total_do_Ativo	Proporção de dívidas de longo prazo sobre o total de ativos
Rácio_entre_o_Fluxo_de_Tesouraria_e_Total_da_Dívida	Indicador de solvência com base no fluxo de tesouraria
Proveitos_Operacionais	Receitas geradas pelas operações principais
Crescimento_do_Rendimento_Operacional	Taxa de crescimento das receitas operacionais
Rácio_Rápido	Indicador de liquidez que exclui *stocks* dos ativos correntes
Total_do_Capital_Próprio_e_do_Passivo	Soma do capital próprio e passivo
Ativos_Fixos	Valor dos bens permanentes da empresa

Rácio_entre_os_Ativos_Fixos_e_o_Capital_a_Longo_Prazo	Relação entre ativos fixos e capital de longo prazo
Dividas_financeiras_CP	Dívidas financeiras de curto prazo
Rácio_Fluxo_Tesouraria_Operacional_e_Dívida_Longo_Prazo	Indicador de solvência de longo prazo
Resultados_transitados	Lucros acumulados não distribuídos
Lucro_Retido_sobre_Total_do_Ativo	Proporção de lucros retidos em relação aos ativos totais
Rácio_de_Retenção	Percentagem do lucro retido pela empresa
ROA	Retorno sobre os ativos
Dívida_Curto_Prazo_em_relação_ao_Capital_Próprio	Proporção de dívidas de curto prazo sobre o capital próprio
Divida_a_Curto_Prazo_em_relação_ao_Total_do_Ativo	Proporção de dívidas de curto prazo sobre os ativos totais
Total_do_Ativo_em_relação_ao_Total_do_Passivo	Relação entre ativos totais e passivos totais
Resultado_Liquido_Exercicio_dividido_total_ativo	Indicador de rentabilidade do ativo

FINL	Mede o grau de endividamento da empresa em relação ao capital próprio
LIQ	Avalia a capacidade da empresa de cumprir obrigações de curto prazo com os ativos disponíveis
Stress_Financeiro	Indicador de risco financeiro da empresa

Não esquecendo que cada variável financeira está representada para cada ano entre 2011 e 2022.

O conjunto de dados recolhido ficou com um total de 615 colunas e 65195 registos de informação sobre empresas desde 2011 até 2022. Neste momento estamos preparados para seguir para a fase de preparação dos dados. Nesta fase iremos aplicar várias técnicas de pré-processamento de dados de forma a preparar o conjunto de dados para a modelação.

4.3. Preparação dos Dados

A fase de preparação dos dados envolve processos de limpeza, transformação e integração dos dados, para que sejam consistentes, relevantes e adequados às necessidades do projeto, permitindo assim a construção de modelos analíticos precisos e confiáveis.

4.3.1. Remoção de colunas

Conforme detalhado anteriormente, o conjunto de dados inclui colunas que inicialmente representam dados diretamente extraídos da base de dados. Estes dados serviram como base primária para o desenvolvimento das fórmulas analíticas discutidas no capítulo

anterior, as quais são fundamentais para a avaliação financeira detalhada das empresas em estudo.

A decisão de remover as colunas originais que contêm esses dados brutos foi tomada visando otimizar a estrutura e a utilidade do conjunto de dados final. A remoção dessas colunas não apenas elimina duplicações desnecessárias, mas também simplifica a interpretação dos dados, concentrando-se nos resultados derivados das análises específicas realizadas. Isso garante que o conjunto de dados final seja mais direto e focado nas métricas financeiras elaboradas, que são cruciais para avaliar a saúde financeira e o desempenho das empresas ao longo do tempo.

Ao remover estas colunas, evita-se a redundância de informações e reduz-se o risco de confusão na análise dos dados. Esta abordagem promove uma maior eficiência na utilização do conjunto de dados, facilitando a aplicação de modelos analíticos e a interpretação dos resultados obtidos. Além disso, ao manter apenas as colunas pertinentes às métricas derivadas, assegura-se que os dados sejam mais acessíveis e úteis para os analistas e gestores que dependem de informações precisas para tomadas de decisão estratégicas.

Portanto, a exclusão das colunas de dados brutos, uma vez integrados às fórmulas desenvolvidas, não apenas simplifica a estrutura do conjunto de dados, mas também fortalece a sua utilidade como uma ferramenta eficaz para análises financeiras profundas e informadas.

Assim, foram removidas as seguintes colunas: Total_Divida, Total_do_activo, Total_do_passivo_corrente, Fluxos_de_Caixa_das_Atividades_Operacionais, Caixa_e_seus_equivalentes, Depositos_bancarios_e_Caixa, Ativos_Correntes, Passivos_curto_prazo, Total_do_passivo, EBITDA, Dividas_a_Terceiros_MLP, Divida_a_Curto_Prazo, Fluxos_de_caixa, Custo_Mercadorias_Vendidas_e_Materias_Consumidas Inventarios, Capital_Proprio, Proveitos_Operacionais, Total_do_Capital_Próprio_e_do_Passivo, Ativos_Fixos, Dividas_financeiras_CP e Resultados_transitados, sendo que o conjunto de dados passou a ser constituído por 375 colunas.

Além disso, durante uma análise inicial do conjunto de dados e ao desenvolver as fórmulas mencionadas anteriormente, identificou-se que algumas delas requeriam dados históricos de anos anteriores. Por exemplo, para calcular o valor da métrica de stress financeiro para o ano de 2011 (stress_financeiro_2011), seria necessário ter acesso aos resultados líquidos dos anos de 2011 e 2010. No entanto, dado que não estavam disponíveis dados para o ano de 2010, foi decidido remover todas as colunas referentes ao ano de 2011.

Esta decisão foi tomada para garantir a consistência e a integridade das análises realizadas com base nas fórmulas estabelecidas. Ao eliminar as colunas que não poderiam ser adequadamente calculadas devido à falta de dados históricos completos, assegurou-se que o conjunto de dados fosse utilizado de forma precisa e que as métricas derivadas fossem calculadas de forma válida e informativa. Isto é fundamental para garantir que as conclusões tiradas das análises sejam robustas e confiáveis, proporcionando informações claras sobre a saúde financeira das empresas analisadas.

Assim, procedeu-se à redução do número de colunas no conjunto de dados, resultando num total de 350 colunas no conjunto de dados.

4.3.2. Substituição de valores não divisíveis por 0

Conforme explicado anteriormente, este conjunto de dados foi construído a partir de dados adquiridos de uma base de dados existente, complementados com fórmulas desenvolvidas especificamente para utilizar esses dados. No decorrer deste processo, foi inevitável encontrar situações onde os resultados das fórmulas incluíam valores que não podiam ser divididos por 0. Isso ocorreu com frequência porque muitas vezes um dos valores utilizados na fórmula era 0, levando a tentativas de divisão por 0, uma operação que é matematicamente indefinida e, portanto, impossível de executar.

Devido a estas circunstâncias, o conjunto de dados final continha um número significativo de valores não divisíveis por 0. Este problema apresentou um desafio significativo, pois a presença destes valores indefinidos dificultava a continuidade e a precisão das análises futuras. Foi, então, necessário tomar uma decisão sobre como lidar com estes valores para não comprometer a integridade dos resultados das análises subsequentes.

Em todas as novas colunas criadas a partir dos dados recolhidos na base de dados tinham este problema de existirem valores não divisíveis por 0, desta forma, depois de considerar várias abordagens, a solução adotada foi substituir todos os valores não divisíveis por 0 por zeros. Esta decisão foi tomada apesar da consciência de que esta não era a abordagem ideal. É importante reconhecer que valores que resultam de uma divisão por 0 não são verdadeiramente zeros, e esta substituição pode introduzir uma certa distorção nos dados. No entanto, entre as opções disponíveis, esta foi considerada a mais prática e viável para resolver o problema. A substituição permitiu que as análises prosseguissem sem a interrupção causada pelos valores indefinidos, ainda que com a compreensão de que essa solução não era perfeita. Esta abordagem visou equilibrar a necessidade de manter a continuidade das análises com a imperfeição inerente à substituição dos valores não divisíveis por 0.

4.3.3. Substituição de valores em falta e NaN

A verificação de valores em falta ou *Not a Number* (NaN) é um dos passos cruciais na análise de dados. Este processo envolve identificar e quantificar a presença de valores em falta ou indefinidos no conjunto de dados. Entender a percentagem de tais valores é fundamental, pois a sua presença pode afetar a qualidade e a integridade das análises subsequentes. Após identificar a extensão dos valores em falta ou NaN, é necessário tomar decisões informadas sobre como lidar com eles. Estas decisões podem incluir a remoção dos registos que contêm valores nulos, a substituição dos valores em falta por estimativas ou médias, ou a utilização de métodos avançados de imputação para preencher os dados em falta. Este passo é essencial para assegurar que as análises realizadas sejam precisas e que os resultados obtidos sejam confiáveis.

Assim sendo, o primeiro passo consistiu em identificar a percentagem de valores em falta e NaN presentes em cada coluna do conjunto de dados. Após essa análise detalhada, constatou-se que havia colunas específicas onde esses tipos de dados em falta eram predominantes. A tabela seguinte mostra quais as colunas com maior percentagem de valores em falta e a ação realizada em cada uma delas para resolver este problema.

Tabela 2 - Percentagem de valores em falta em cada coluna

Coluna	**Percentagem de valores em falta**	**Ação realizada**
"Localidade"	47%	Remoção da coluna
"Estado"	90%	Remoção da coluna
"Número de Empregados"	50%	Remoção da coluna

Todas as colunas mencionadas, exceto a coluna "Número de Empregados", incluem dados referentes às empresas e são compostas por informações não numéricas. Esta característica torna a substituição dos valores em falta por médias, zeros ou qualquer outro valor de preenchimento uma tarefa bastante complexa e, muitas vezes, impraticável. Dados não numéricos, como os presentes nas colunas "Localidade" e "Estado", não podem ser simplesmente substituídos por médias, pois esses conceitos não se aplicam a informações categóricas ou textuais. Devido à dificuldade de realizar uma substituição adequada e ao impacto que isso teria na integridade e precisão da análise dos dados, foi decidido que a melhor abordagem seria eliminar essas colunas do conjunto de dados. Assim, ao remover essas colunas, assegura-se que as análises subsequentes não sejam prejudicadas por tentativas inadequadas de imputação de dados em falta.

Por outro lado, a coluna "Número de Empregados" já contém dados numéricos e, teoricamente, poderia ser substituída por médias ou outras métricas similares com relativa facilidade. Inicialmente, a estratégia era preencher os valores em falta com a média do número de empregados de cada empresa. Dessa forma, esperava-se utilizar esta coluna nos dados sem introduzir distorções significativas ou desvios muito grandes em relação à realidade. No entanto, ao tentar implementar este procedimento através do desenvolvimento do código correspondente, foi detetado um problema nos dados que impediu a substituição dos valores em falta pelas médias adequadas. Após investir tempo significativo neste problema e enfrentar várias tentativas fracassadas de resolução, a decisão tomada foi de remover esta coluna do conjunto de dados. Esta medida foi adotada para garantir a qualidade e a integridade dos dados utilizados nas análises subsequentes, evitando assim qualquer distorção ou erro que poderia resultar de uma imputação inadequada dos valores em falta.

4.3.4. Criação de Novas Variáveis

As colunas presentes neste conjunto de dados estão classificadas em duas categorias distintas, cada uma desempenhando um papel específico na análise em questão.

Na primeira categoria existem colunas que representam os dados propriamente ditos, os quais foram extraídos diretamente da base de dados. Estes dados são essenciais, pois fornecem a matéria-prima necessária para qualquer tipo de análise posterior. São os valores concretos, numéricos ou categóricos, que refletem informações específicas sobre as entidades estudadas.

Na segunda categoria existem colunas que contêm fórmulas que utilizam os dados obtidos da SABI. Estas fórmulas não foram criadas aleatoriamente, mas sim baseadas numa revisão extensa e detalhada da literatura existente. Esta revisão da literatura foi realizada previamente, com o objetivo de identificar as metodologias e fórmulas mais relevantes e reconhecidas pela comunidade científica para a análise de situações de stress financeiro. Durante esta revisão, foram analisados diversos artigos científicos, teses, dissertações e outros documentos técnicos que tratam do tema em estudo. Foi verificado que muitos desses documentos utilizavam fórmulas similares para prever e analisar situações de stress financeiro, indicando uma aceitação e validação dessas fórmulas na prática.

Com base nesta revisão, recolheram-se as fórmulas mais adequadas, com a finalidade de serem aplicadas aos dados recolhidos. A ideia é que, após a recolha dos dados iniciais através da SABI, essas fórmulas possam ser desenvolvidas e aplicadas com o auxílio dos dados obtidos, permitindo uma análise mais robusta e fundamentada. Este processo de integração das fórmulas com os dados reais visa proporcionar uma melhor compreensão das dinâmicas financeiras e ajudar a prever possíveis situações de stress financeiro com maior precisão.

Para proporcionar uma compreensão mais clara e abrangente dos dados utilizados, apresenta-se a seguir uma explicação detalhada das colunas que contêm fórmulas. Esta explicação tem como objetivo esclarecer a origem e a utilidade dessas fórmulas, bem como a maneira como foram aplicadas aos dados extraídos da SABI:

- Crescimento_do_ativo: esta métrica mede a variação dos ativos totais de uma empresa de um ano para o outro. Calcula-se dividindo os ativos no ano *t* pelos ativos no ano *t – 1*.

$$\text{Crescimento\_do\_ativo} = \frac{Ativos\ Totais\ no\ ano\ t}{Ativos\ Totais\ no\ ano\ t-1}$$

- Fluxos_de_Caixa_das_Atividades_Operacionais: Fluxo de Caixa das Operações sobre Passivos Correntes mede a capacidade de uma empresa de cobrir os seus passivos correntes (dívidas de curto prazo) com o fluxo de caixa gerado pelas operações. A fórmula é a seguinte:

$$\text{Fluxos\_de\_Caixa\_das\_Atividades\_Operacionais} = \frac{Fluxo\ de\ Caixa\ das\ Operações}{Passivos\ Correntes}$$

- Rácio_de_Tesouraria: Índice de Liquidez Imediata mede a capacidade de uma empresa de pagar as suas dívidas de curto prazo apenas com as suas reservas de caixa e equivalentes de caixa. A fórmula é a seguinte:

$$\text{Rácio\_de\_Tesouraria} = \frac{Caixa + Equivalentes\ de\ Caixa}{Passivos\ Correntes}$$

- Rácio_Corrente: Índice de Liquidez Corrente mede a capacidade de uma empresa pagar as suas dívidas de curto prazo com os seus ativos de curto prazo. A fórmula é a seguinte:

$$\text{Rácio\_Corrente} = \frac{Ativos\ Correntes}{Passivos\ a\ Curto\ Prazo}$$

- EBITDA_para_Divida_a_Curto_Prazo_e_Juros: EBITDA sobre Dívida de Curto Prazo e Juros avalia a capacidade de uma empresa pagar a sua dívida de curto prazo e os juros associados utilizando o seu EBITDA. A fórmula é a seguinte:

$$\text{EBITDA\_para\_Divida\_a\_Curto\_Prazo\_e\_Juros} = \frac{Dívida\ (Longo\ Prazo\ e\ Curto\ Prazo)}{EBITDA}$$

- Crescimento_do_Capital_Próprio: Crescimento do Património Líquido avalia a variação percentual do patrimônio líquido de uma empresa de um ano para o outro. A fórmula é calculada dividindo o património líquido no ano *t* pelo patrimônio líquido no ano *t−1*:

$$\text{Crescimento\_do\_Capital\_Próprio} = \frac{Património\ Líquido\ no\ ano\ t}{Património\ Líquido\ no\ ano\ t-1}$$

- Rendimento_Bruto_do_Ativo: Retorno Bruto sobre Ativos mede a eficiência com que uma empresa utiliza os seus ativos para gerar lucros antes de juros, impostos, depreciação e amortização (EBITDA). A fórmula é a seguinte:

$$\text{Rendimento\_Bruto\_do\_Ativo} = \frac{EBITDA}{Ativos\ Totais\ Médios}$$

- Taxa_de_Crescimento_do_Rendimento: Taxa de Crescimento do Lucro Líquido mede a variação percentual do lucro líquido de uma empresa de um ano para o outro. A fórmula é a seguinte:

$$\text{Taxa\_de\_Crescimento\_do\_Rendimento} = \frac{Lucro\ Líquido\ no\ ano\ t}{Lucro\ Líquido\ no\ ano\ t-1}$$

- Taxa_de_Crescimento_do_Fluxo_de_Caixa_Líquido: Taxa de Crescimento do Fluxo de Caixa Líquido mede a variação percentual do fluxo de caixa líquido de uma empresa de um ano para o outro. A fórmula é a seguinte:

$$\text{Taxa\_de\_Crescimento\_do\_Fluxo\_de\_Caixa\_Líquido} = \frac{Fluxo\ de\ Caixa\ Líquido\ no\ ano\ t}{Fluxo\ de\ Caixa\ Líquido\ no\ ano\ t-1}$$

- Período_de_Rotação_do_Inventário: Período de Rotatividade do Inventário mede o tempo médio que uma empresa leva para vender o seu inventário completo em dias. A fórmula é a seguinte:

$$\text{Período\_de\_Rotação\_do\_Inventário} = \frac{Inventário * 365}{Custo\ das\ Mercadorias\ Vendidas}$$

- Dívidas_a_Longo_Prazo_em_relação_ao_Capital_Próprio: Dívidas de Longo Prazo sobre Capital Próprio mede a proporção das dívidas de longo prazo de uma empresa em relação ao seu capital próprio. A fórmula é a seguinte:

$$\text{Dívidas\_a\_Longo\_Prazo\_em\_relação\_ao\_Capital\_Próprio} = \frac{Dívida\ de\ Longo\ Prazo}{Capital\ Próprio}$$

- Dívidas_a_Longo_Prazo_em_relação_ao_Total_do_Ativo: Dívidas de Longo Prazo sobre Ativos Totais mede a proporção das dívidas de longo prazo de uma empresa em relação ao total dos seus ativos. A fórmula é a seguinte:

$$\text{Dívidas\_a\_Longo\_Prazo\_em\_relação\_ao\_Total\_do\_Ativo} = \frac{Dívida\ de\ Longo\ Prazo}{Ativos\ Totais}$$

- Rácio_entre_o_Fluxo_de_Tesouraria_e_Total_da_Dívida: Rácio do Fluxo de Caixa Operacional sobre o Total de Dívidas mede a capacidade de uma empresa de gerar caixa a partir das suas operações para cobrir as suas obrigações totais (dívidas). A fórmula é a seguinte:

$$\text{Rácio\_entre\_o\_Fluxo\_de\_Tesouraria\_e\_Total\_da\_Dívida} = \frac{Fluxo\ de\ Caixa\ Operacional}{Passivos\ Totais}$$

- Crescimento_do_Rendimento_Operacional: Crescimento do Rendimento Operacional mede a variação percentual do rendimento operacional de uma empresa de um ano para o outro. A fórmula é a seguinte:

$$\text{Crescimento\_do\_Rendimento\_Operacional} = \frac{Rendimento\ Operacional\ no\ ano\ t}{Rendimento\ Operacional\ no\ ano\ t-1}$$

- Rácio_Rápido: Rácio Rápido mede a capacidade de uma empresa de cumprir as suas obrigações de curto prazo sem depender da venda de inventários. A fórmula é a seguinte:

$$\text{Rácio\_Rápido} = \frac{Ativos\ Correntes - Inventários}{Passivos\ Correntes}$$

- Rácio_entre_os_Ativos_Fixos_e_o_Capital_a_Longo_Prazo: Rácio de Ativos Fixos sobre Capital de Longo Prazo avalia a proporção dos ativos fixos de uma empresa em relação ao seu capital de longo prazo, descontando os passivos correntes. A fórmula é a seguinte:

$$\text{Rácio\_entre\_os\_Ativos\_Fixos\_e\_o\_Capital\_a\_Longo\_Prazo} = \frac{Ativos\ Fixos}{Capital\ Próprio - Passivos\ Correntes}$$

- Rácio_entre_o_Fluxo_de_Tesouraria_Operacional_e_a_Dívida_a_Longo_Prazo : Rácio do Fluxo de Caixa Operacional sobre a Dívida de Curto Prazo avalia a capacidade de uma empresa de gerar fluxo de caixa operacional suficiente para cobrir a sua dívida de curto prazo. A fórmula é a seguinte:

$$\text{Rácio\_entre\_o\_Fluxo\_de\_Tesouraria\_Operacional\_e\_a\_Dívida\_a\_Longo\_Prazo} = \frac{Fluxo\ de\ Caixa\ Operacional}{Dívida\ de\ Curto\ Prazo}$$

- Lucro_Retido_sobre_Total_do_Ativo: Lucros Retidos sobre Ativos Totais avalia a eficiência com que uma empresa utiliza os seus lucros retidos em relação ao total dos seus ativos. A fórmula é a seguinte:

$$\text{Lucro\_Retido\_sobre\_Total\_do\_Ativo} = \frac{Lucros\ Retidos}{Ativos\ Totais}$$

- Rácio_de_Retenção: Taxa de Retenção avalia a proporção dos lucros líquidos de uma empresa que são retidos e reinvestidos na empresa, em vez de serem distribuídos aos acionistas como dividendos. A fórmula é a seguinte:

$$\text{Rácio\_de\_Retenção} = \frac{Lucros\ Retidos}{Lucro\ Líquido}$$

- ROA (*Return on Assets*): Retorno sobre Ativos é uma medida da eficiência com que uma empresa utiliza os seus ativos para gerar lucro. A fórmula para calcular o ROA é a seguinte:

$$ROA = \frac{Lucro\ Líquido}{Média\ dos\ Ativos\ Totais}$$

- Dívida_Curto_Prazo_em_relação_ao_Capital_Próprio: Dívida de Curto Prazo sobre Património Líquido avalia a proporção da dívida de curto prazo em relação ao património líquido de uma empresa. A fórmula para calcular esta métrica é:

$$\text{Dívida\_Curto\_Prazo\_em\_relação\_ao\_Capital\_Próprio} = \frac{Dívida\ de\ Curto\ Prazo}{Património\ Líquido}$$

- Divida_a_Curto_Prazo_em_relação_ao_Total_do_Ativo: Dívida de Curto Prazo sobre Ativos Totais avalia a proporção da dívida de curto prazo em relação ao total de ativos de uma empresa. A fórmula para calcular esta métrica é:

$$\text{Divida\_a\_Curto\_Prazo\_em\_relação\_ao\_Total\_do\_Ativo} = \frac{Dívida\ de\ Curto\ Prazo}{Ativos\ Totais}$$

- Total_do_Ativo_em_relação_ao_Total_do_Passivo: Ativos Totais sobre Passivos Totais avalia a relação entre os ativos totais e os passivos totais de uma empresa. A fórmula para calcular esta métrica é:

$$\text{Total\_do\_Ativo\_em\_relação\_ao\_Total\_do\_Passivo} = \frac{Ativos\ Totais}{Passivos\ Totais}$$

- FINL (*Financial Leverage*): Alavancagem Financeira avalia o grau de endividamento de uma empresa em relação aos seus ativos totais. A fórmula para calcular esta métrica é:

$$FINL = \frac{Total\ Dívida}{Total\ Ativo}$$

- LIQ (*Liquidity Ratio*): Rácio de Liquidez calcula-se dividindo os ativos correntes pelos passivos correntes de uma empresa. A fórmula é:

$$LIQ = \frac{Total\ Ativos\ Correntes}{Total\ Passivos\ Correntes}$$

Além das colunas que foram anteriormente identificadas e explicadas, existe ainda outra coluna fundamental para avaliar o estado financeiro de uma empresa, estando preenchida com valores 1, que indica se a empresa está em situação de stress financeiro, ou valores 0, que indica que não está. Esta coluna foi adicionada com base em critérios específicos e análises detalhadas, conforme descrito na revisão da literatura. Ela desempenha um papel crucial na análise financeira, ajudando a identificar empresas que podem estar a enfrentar desafios financeiros significativos ao longo do tempo. A utilização desta coluna referente ao stress financeiro consiste em utilizar uma definição encontrada na literatura para definir o stress financeiro para utilizar em modelos de ML que possam prever situações de stress financeiro nas organizações, sendo que esta coluna será a variável dependente que será utilizada posteriormente para treinar os modelos de ML:

- Stress_Financeiro: Durante a revisão da literatura sobre o tema, descobriu-se que esta nova coluna visa determinar se uma empresa está em stress financeiro (indicado pelo valor 1) ou não (indicado pelo valor 0). A definição e o método de cálculo dessa métrica foram identificados e detalhados na literatura especializada. A definição baseia-se no cálculo do índice B de Zmijewski, que é calculado da seguinte forma:

$$B = -4{,}3 - 4{,}5ROA + 5{,}7FINL + 0{,}004LIQ$$

Segundo essa definição, as empresas são classificadas como estando em stress financeiro se o valor de B for superior a 0. Este índice leva em consideração diversos fatores financeiros cruciais, como o retorno sobre ativos (ROA), os resultados líquidos (FINL) e a liquidez (LIQ), fornecendo uma medida compreensiva da saúde financeira e da capacidade de uma empresa para enfrentar desafios económicos. Esta abordagem analítica é fundamental para identificar e monitorizar empresas que podem estar a enfrentar dificuldades financeiras significativas, permitindo uma intervenção proativa e estratégias de gestão adequadas.

Inicialmente, o conjunto de dados estava estruturado de forma que cada linha representava uma empresa específica, e cada coluna continha uma variável financeira com informações anuais dessa empresa, como por exemplo, EBITDA_2022, EBITDA_2021, e assim sucessivamente. No entanto, para adequar o conjunto de dados às exigências dos modelos de sobrevivência, foi necessário transformar esta estrutura.

As alterações realizadas consistiram em reconfigurar o conjunto de dados para que cada linha representasse uma empresa num determinado ano (ano esse que seria respetivo ao ano em que ocorreu o evento de stress financeiro ou, no caso de nunca ter experienciado eventos de stress financeiro, seria considerado o ano mais recente), sendo que essa linha seria composta por colunas financeiras relativas a esse ano e uma coluna denominada "stress_financeiro" que indica se a empresa enfrentou (1) ou não (0) um evento de stress financeiro naquele ano específico.

Para além disso, cada linha será ainda composta por variáveis financeiras respetivas aos 3 anos anteriores em que se observa o evento. Ou seja, imaginando que o ano da linha onde ocorre o evento de stress financeiro é 2020. Essa linha será constituída, não só pelas variáveis financeiras correspondentes a esse mesmo ano, mas também pelas variáveis financeiras dos 3 anos anteriores (2019, 2018 e 2017).

Esta reformulação do conjunto de dados permite uma análise temporal detalhada de cada empresa, possibilitando que os modelos de sobrevivência identifiquem padrões e correlacionem as variáveis financeiras com a ocorrência de stress financeiro ao longo do

tempo. Este formato é fundamental para capturar a dinâmica temporal das variáveis e fornecer uma base sólida para previsões mais precisas sobre a estabilidade financeira das empresas ao longo dos anos.

Nova variável "*years_to_event*":

Para o treino de modelos de sobrevivência, é crucial a presença de uma variável de tempo, que desempenha um papel fundamental na análise. Esta variável específica, denominada "*years_to_event*" é responsável por registar o período de tempo que decorre desde o momento da criação da empresa até à ocorrência de um evento de stress financeiro. A sua importância reside na capacidade de capturar e quantificar o intervalo temporal, o que permite uma compreensão mais profunda de como o fator tempo influencia o surgimento de dificuldades financeiras.

Ao analisar esta variável, é possível avaliar de forma detalhada o impacto que a duração da atividade empresarial tem na sua estabilidade financeira ao longo do tempo. Ou seja, a variável " *years_to_event*" permite correlacionar o tempo de existência da empresa com a probabilidade de enfrentar crises financeiras, fornecendo assim uma perspetiva valiosa sobre como o tempo pode ser um determinante crítico para a saúde financeira da organização. Esta análise torna-se essencial para identificar padrões de risco ao longo do ciclo de vida da empresa, permitindo, desta forma, que sejam tomadas medidas preventivas ou corretivas em momentos oportunos, aumentando a resiliência e sustentabilidade das organizações no mercado.

Conforme mencionado anteriormente, a estrutura do conjunto de dados foi reorganizada de forma que cada empresa seja representada por apenas 1 linha, cada uma correspondendo a um ano específico. Nessas linhas, será indicado se a empresa experimentou um evento de stress financeiro naquele ano ou não. Dada essa reformulação, tornou-se essencial definir com precisão a variável "*years_to_event*", que irá refletir o tempo decorrido em anos desde a criação da empresa até o momento de cada evento ou ausência dele. A definição da variável "*years_to_event*" foi estabelecida da seguinte forma:

1. **Empresas sem eventos de stress financeiro**: Se uma empresa não enfrentou nenhum evento de stress financeiro ao longo dos anos, ou seja, todas as linhas na coluna "stress_financeiro" estão preenchidas com 0, a variável "*years_to_event*" será calculada como o número de anos decorridos desde o ano de criação da empresa até o ano em que os dados foram recolhidos. Sendo que apenas a linha correspondente ao ano mais recente será mantida no conjunto de dados.

2. **Empresas com apenas eventos de stress financeiro**: Se uma empresa passou por eventos de stress financeiro em todos os anos considerados, com todas as linhas da coluna "stress_financeiro" preenchidas com 1, a variável "*years_to_event*" será definida como a diferença entre o ano de criação da empresa e o ano em que ocorreu o primeiro evento de stress financeiro. Apenas a linha correspondente ao primeiro 1 na coluna "stress_financeiro" será mantida, pois esse é o ponto crítico de análise.

3. **Empresas com eventos mistos (com e sem stress financeiro)**: Se uma empresa experimentou tanto anos com stress financeiro quanto anos sem, ou seja, a coluna "stress_financeiro" contém uma combinação de 0 e 1, a variável "*years_to_event*" será definida como o número de anos desde a criação da empresa até ao ano específico em que ocorreu o primeiro evento de stress financeiro. Apenas a linha onde aparece o primeiro 1 na coluna "stress_financeiro" será mantida.

Estas definições são cruciais para garantir que a análise de sobrevivência capture de forma precisa a dinâmica temporal de cada empresa, permitindo uma avaliação robusta dos fatores que influenciam a estabilidade financeira ao longo do tempo.

4.3.5. Correlação entre variáveis

A análise de correlação entre as variáveis é um passo crucial antes de iniciar o treino dos modelos, pois permite identificar quais variáveis apresentam uma relação forte entre si. Quando duas ou mais variáveis têm uma correlação elevada, positiva ou negativa, isso significa que estão a representar informações muito semelhantes. Manter as variáveis no conjunto de dados pode levar a redundâncias e comprometer a eficácia do modelo, uma

vez que essas variáveis repetem praticamente o mesmo padrão de informação. Assim, é importante remover essas variáveis redundantes para simplificar o modelo e melhorar o seu desempenho.

Para proceder à eliminação das variáveis com correlação elevada, foi realizada uma análise detalhada da correlação entre todas as variáveis presentes no conjunto de dados. Nessa análise, foram consideradas apenas as variáveis cuja correlação absoluta (ou seja, a correlação tanto positiva quanto negativa) fosse superior a 0.7. Isto significa que, quando duas variáveis apresentavam uma correlação maior do que 0.7 ou menor do que -0.7, uma delas foi removida do conjunto de dados, por se considerar que ambas estavam a transmitir essencialmente a mesma informação. Este processo é fundamental para garantir que o modelo não fique sobrecarregado com informações redundantes e pode assim focar-se nas variáveis que realmente contribuem de forma única para a previsão.

As seguintes colunas foram removidas por apresentarem elevada correlação com outras:

Tabela 3 - Colunas removidas

Ativos_Correntes
Rácio_entre_o_Fluxo_de_Tesouraria_Operacional_e_a_Dívida_a_Longo_Prazo
Rácio_Corrente
Custo_Mercadorias_Vendidas_e_Materias_Consumidas
Dividas_a_Terceiros_MLP
EBITDA_para_Divida_a_Curto_Prazo_e_Juros
Fluxos_de_Caixa_das_Atividades_Operacionais
LIQ
Dívidas_a_Longo_Prazo_em_relação_ao_Capital_Próprio
Lucro_Retido_sobre_Total_do_Ativo
Divida_a_Curto_Prazo_em_relação_ao_Total_do_Ativo
Total_Divida

4.3.6. Normalização das variáveis

Um dos passos efetuados antes de realizar o treino dos modelos foi a normalização das variáveis. A normalização das colunas antes de treinar modelos de ML é uma etapa crucial para garantir que as variáveis sejam comparáveis e que o modelo funcione de forma eficiente e precisa. A normalização desempenha diversos papéis importantes no processo de modelação. Primeiro, estabiliza o processo de treino, especialmente em algoritmos que utilizam gradientes, como regressão logística, redes neuronais e SVM. Sem normalização, variáveis de entrada com diferentes escalas podem levar o algoritmo a dar mais peso a certas características em detrimento de outras, resultando em modelos enviesados ou com dificuldades de convergência.

Além disso, a normalização acelera a convergência do algoritmo, facilitando a navegação pelo espaço de parâmetros de forma mais uniforme, o que é particularmente benéfico em métodos baseados em gradiente descendente (Li *et al.,* 2024; Ding *et al.,* 2023). Outro benefício importante é a prevenção do problema de dominação das variáveis. Sem normalização, variáveis com valores numéricos maiores podem dominar as variáveis com valores menores, não necessariamente por serem mais importantes, mas simplesmente devido à sua escala maior (Chaopeng *et al.,* 2024). Isso pode levar a interpretações incorretas dos pesos das variáveis no modelo.

A normalização também melhora a interpretação dos coeficientes em modelos lineares, permitindo que os coeficientes atribuídos às variáveis possam ser comparados diretamente. Caso as variáveis não sejam normalizadas, os coeficientes podem não refletir a importância relativa das variáveis, sendo afetados pelas diferenças de escala. Por fim, a normalização ajuda a prevenir problemas numéricos, especialmente em redes neuronais profundas, onde variáveis com escalas muito diferentes podem causar resultados imprecisos ou aumentar significativamente o tempo de computação.

A normalização de variáveis refere-se ao processo de transformação de dados para ajustar a escala das variáveis, de modo que elas apresentem valores dentro de um intervalo comum, geralmente entre 0 e 1. A normalização garante que nenhuma variável com maior amplitude domine as restantes durante o processo de otimização, facilitando a convergência e melhorando o desempenho dos modelos. Normalmente, a transformação é realizada através da fórmula:

$$x' = \frac{x - xmin}{xmax - xmin} \qquad (10)$$

onde *x* é o valor original, *xmin* e *xmax* são, respetivamente, o menor e o maior valor da variável.

4.3.7. Conjunto de Dados Final

Após todo o pré-processamento de dados realizado, o novo conjunto de dados passa a ser constituído por 65110 linhas e 148 colunas, sendo elas:

Tabela 4 - Colunas que compõem o conjunto de dados final

Coluna	Mínimo	Máximo	Média	Desvio Padrão
Código da CAE	0	1	0,49216	0,30967
year	2012	2022	2019,6	3.9
Crescimento_do_ativo	0	1	2,3e-05	4,0e-03
Caixa_e_seus_equivalentes	0	1	0,50530	0,00567
Capital_Proprio	0	1	0,08758	0,00513
Cash_Flow_from_Operations	0	1	0,67384	0,00299
Rácio_de_Tesouraria	0	1	0,10711	0,00535
Debt-to-Equitity_Ratio	0	1	0,07783	0,0037
Depositos_bancarios_e_Caixa	0	1	0,5	0.00386
Dividas_financeiras_CP	0	1	1,5e-05	3,9e-03
EBITDA	0	1	0,47	0.00366
Crescimento_do_Capital_Próprio	0	1	1,5e-05	3,9e-03
FINL	0	1	0,0039	0,01
Ativos_Fixos	0	1	1,2e-04	7,7e-03
Fluxos_de_caixa	0	1	0,58	0,0035
Rendimento_Bruto_do_Ativo	0	1	0,93244	0,00408
Growth_Rate_of_Income	0	1	0,68932	0,00418
Taxa_de_Crescimento_do_Fluxo_de_Caixa_Líquido	0	1	0,98769	0,00388

Inventarios	0	1	1,2e-04	7,3e-03
Período_de_Rotação_do_Inventário	0	1	0,00023	0,00816
Dívidas_a_Longo_Prazo_em_relação_ao_Total_do_Ativo	0	1	2,5e-05	4,4e-03
Crescimento_do_Rendimento_Operacional	0	1	2,4e-05	4,2e-03
Operating_Cash_Flow_to_Total_Debt_Ratio	0	1	0.27764	0.00315
Passivos_curto_prazo	0	1	4,7e-05	5,1e-03
Proveitos_Operacionais	0	1	5,5e-05	5,5e-03
Rácio_Rápido	0	1	2,0e-05	4,0e-03
ROA	0	1	0.99994	0.00671
Rácio_entre_os_Ativos_Fixos_e_o_Capital_a_Longo_Prazo	0	1	0.28150	0.00336
Resultado_Liquido_do_Exercicio	0	1	0.99983	0.00745
Resultados_transitados	0	1	0.74959	0.00439
Rácio_de_Retenção	0	1	0.8792	0.00402
Dívida_Curto_Prazo_em_relação_ao_Capital_Próprio	0	1	0.81399	0.00347
Stress_Financeiro	0	1	0.32207	0.46727
Total_do_Ativo_em_relação_ao_Total_do_Passivo	0	1	0.00043	0.00718
Total_do_ativo	0	1	1,4e-04	9,1e-03
Total_do_Capital_Próprio_e_do_Passivo	0	1	8,5e-05	7,1e-03
Total_do_passivo	0	1	0.00021	0.00444
Total_do_passivo_corrente	0	1	0.00006	0.00438
years_to_event.	0	347	21,005	16,167

Não esquecendo que para cada linha, as variáveis financeiras repetem-se 4 vezes devido ao facto de existirem os valores financeiros relativos ao ano da linha e aos 3 anos anteriores.

5. Análise dos Resultados

Neste capítulo, são apresentados e discutidos os resultados obtidos com o treino dos diferentes modelos de sobrevivência aplicados ao estudo. A análise comparativa entre os modelos permitiu avaliar a sua eficácia na predição de eventos de stress financeiro, utilizando métricas como o C-Index para medir o desempenho de cada abordagem. Com base nos resultados obtidos, identifica-se o modelo que demonstrou o melhor desempenho em termos de precisão preditiva, justificando a sua superioridade relativamente aos restantes modelos analisados. Este capítulo visa, assim, proporcionar uma compreensão clara e fundamentada sobre a escolha do modelo mais adequado para o contexto específico do estudo.

Em todos os modelos treinados, foi utilizada a técnica *K-Fold Cross Validation*. Todos os modelos foram treinados com 5 *folds*.

5.1. Modelo *CoxPHFitter*

Os resultados obtidos com o modelo *CoxPHFitter* indicam um desempenho robusto na predição de eventos de sobrevivência. A concordância (C-Index) de 0.86 sugere que o modelo tem uma capacidade preditiva elevada, uma vez que valores próximos de 1 indicam um bom ajuste entre as previsões do modelo e os eventos reais observados.

A análise de regressão utilizando o modelo *CoxPH* revelou importantes informações sobre os fatores associados à probabilidade de ocorrência de eventos de stress financeiro em empresas.

A variável "*year*" apresentou um coeficiente negativo de -0.95, indicando que anos mais recentes estão associados a uma menor probabilidade de ocorrência de stress financeiro, com o IC 95% entre -0.97 e -0.93. O valor p também foi inferior a 0.005, confirmando a significância estatística desta variável.

Algumas variáveis, como "Crescimento_do_ativo" e "Caixa_e_seus_equivalentes", apresentaram coeficientes relevantes, mas não estatisticamente significativos ($p > 0.05$),

sugerindo que o seu impacto sobre o risco de eventos de stress financeiro não é claro, possivelmente devido à variabilidade nos dados.

O "Capital_Proprio" teve um coeficiente de 0.79, com um IC 95% entre -0.51 e 2.10, o que sugere uma possível relação positiva com a probabilidade de stress financeiro, embora o valor p não tenha sido significativo. Variáveis relacionadas à liquidez, como "Rácio_de_Tesouraria" também tiveram coeficientes positivos, sugerindo que maior liquidez pode estar associada a um maior risco de stress financeiro, mas, mais uma vez, sem significância estatística robusta.

Por outro lado, variáveis como "Resultado_Liquido_do_Exercicio" apresentaram coeficientes negativos (-1.69), sugerindo que melhores resultados financeiros estão associados a uma menor probabilidade de ocorrência de stress financeiro. Estes resultados foram estatisticamente significativos, com valores p inferiores a 0.005.

Além disso, a inclusão de variáveis com defasagem de um ano (ano-1) permitiu observar o impacto dos dados históricos sobre o risco atual. Por exemplo, "FINL_ano-1" apresentou um coeficiente de -1.11, sugerindo que o aumento no valor desta variável no ano anterior está associado a uma menor probabilidade de stress financeiro.

Além da avaliação realizada com o valor do C-Index, outras métricas de desempenho foram aplicadas para garantir uma análise mais completa da capacidade preditiva dos modelos. Entre essas métricas adicionais, destacam-se o *Brier Score* e o AUC. O *Brier Score*, que mede a precisão das probabilidades previstas, resultou num valor de 0.26, o que indica que as previsões feitas pelo modelo estão razoavelmente próximas dos resultados reais, embora haja margem para melhorias.

Já o AUC, que avalia a capacidade do modelo de distinguir entre eventos ocorridos e não ocorridos ao longo do tempo, apresentou um valor de 0.78. Este valor demonstra que o modelo tem uma boa habilidade de separação entre indivíduos que experienciarão o evento e aqueles que não o farão, sendo que quanto mais próximo de 1, melhor a discriminação.

5.2. Modelo *Random Survival Forest*

O valor do C-Index neste modelo obtido foi 0.8853, o que indica um desempenho preditivo elevado do modelo. O C-Index varia entre 0.5 e 1.0, onde um valor de 0.5 significa que o modelo não é melhor do que uma escolha aleatória, enquanto um valor de 1.0 representa uma previsão perfeita da ordem dos eventos. No caso específico deste modelo, o C-Index de 0.8853 sugere que o modelo consegue ordenar corretamente os tempos de sobrevivência em aproximadamente 88.5% das vezes, demonstrando uma forte capacidade de discriminação entre os tempos dos eventos e não eventos.

O modelo apresentou um *Brier Score* de 0.19, o que indica uma alta precisão nas previsões probabilísticas realizadas. Este valor, que está abaixo de 0.20, sugere que o modelo consegue prever com bastante exatidão a probabilidade de sobrevivência ao longo do tempo, com pouca diferença entre as previsões e os resultados reais.

Além disso, a métrica AUC foi calculada em 0.83, o que reforça a eficácia do modelo na discriminação entre os indivíduos que terão ou não o evento de interesse. Um AUC próximo de 1 indica que o modelo tem uma forte capacidade de distinguir corretamente os casos positivos dos negativos, o que é um sinal de um desempenho robusto e confiável. Estes resultados, em conjunto, indicam que o modelo tem um bom equilíbrio entre a precisão das previsões e a capacidade de identificar corretamente os eventos.

Em resumo, o modelo de RSF treinado apresentou uma elevada capacidade preditiva, como indicado pelos valores obtidos em cada uma das métricas utilizadas, sugerindo que o modelo é eficaz na discriminação entre os tempos de sobrevivência dos indivíduos, sendo que o treino deste modelo foi efetuado consoante os parâmetros definidos como *default.*

5.2.1. Definição dos Hiperparâmetros do modelo RSF

A escolha de ajustar os hiperparâmetros foi essencial para otimizar o desempenho dos modelos preditivos utilizados neste projeto. Hiperparâmetros, ao contrário dos parâmetros aprendidos diretamente pelo modelo, são configurações definidas antes do processo de treino e influenciam diretamente a capacidade de generalização e a precisão

das previsões. Ajustá-los permite melhorar o equilíbrio entre viés e variância, evitando tanto o sobre ajuste quanto o subajuste aos dados. Ao recorrer à otimização dos hiperparâmetros, foi possível adaptar o modelo às especificidades dos dados financeiros analisados, maximizando a sua capacidade de prever eventos de stress financeiro de forma mais robusta e precisa.

Os hiperparâmetros em ML referem-se a configurações pré-definidas que não são aprendidas diretamente a partir dos dados durante o processo de treino. Em vez disso, são definidos pelo utilizador antes de iniciar o treino e desempenham um papel crucial na determinação do desempenho do modelo. Ajustar os hiperparâmetros de forma eficaz significa explorar e identificar as configurações mais adequadas para um modelo específico, que maximizem o seu desempenho e levem à obtenção dos melhores resultados possíveis. Este processo envolve a escolha criteriosa de valores para várias opções, como o número de camadas numa rede neuronal, o tamanho do lote, entre outros. A combinação correta desses hiperparâmetros pode fazer uma diferença significativa na capacidade do modelo de generalizar bem para novos dados, evitando problemas como *overfitting* ou *underfitting*. Portanto, a tarefa de ajuste de hiperparâmetros é uma fase crítica no desenvolvimento de modelos de ML, pois visa encontrar as configurações ideais que permitam ao modelo realizar previsões com a maior precisão e eficiência possível, dentro do contexto específico em que está a ser aplicado.

Dado que o modelo que apresentou os melhores resultados utilizando os parâmetros definidos como *default* foi o RSF, optou-se por aplicar uma técnica de ajuste de hiperparâmetros, conhecida como "*hyperparameter tuning*". Essa decisão foi motivada pela procura contínua de otimização do desempenho do modelo. Embora os resultados iniciais tenham sido promissores com as configurações padrão, existe a possibilidade de melhorar ainda mais a eficácia do RSF ajustando cuidadosamente esses hiperparâmetros. O objetivo central desta abordagem é identificar as configurações de parâmetros que maximizem o desempenho do modelo, especificamente visando elevar o valor do C-Index, que é uma métrica crucial para avaliar a qualidade das previsões no contexto de sobrevivência. Por outras palavras, a intenção é refinar o modelo RSF para que ele se torne ainda mais preciso e robusto, proporcionando previsões que estejam mais alinhadas com a realidade observada nos dados. Este processo de "*hyperparameter tuning*" envolve explorar diferentes combinações de parâmetros e avaliar sistematicamente o impacto de

cada uma no desempenho do modelo, garantindo assim que a versão final do RSF seja a mais eficiente possível para a tarefa em questão.

Para alcançar uma compreensão mais profunda e otimizar o desempenho do modelo RSF, foi conduzido um estudo detalhado que envolveu o treino do modelo em cinco ocasiões distintas. Cada uma dessas sessões de treino utilizou diferentes combinações de parâmetros, permitindo uma análise comparativa dos resultados. A escolha por múltiplas execuções do treino com variações nos hiperparâmetros foi feita com o intuito de identificar como diferentes configurações afetam o desempenho geral do modelo. Durante cada sessão de treino, o modelo foi avaliado com base no C-Index, uma métrica estatística que quantifica a capacidade do modelo de fazer previsões corretas em análises de sobrevivência.

Ao longo do estudo, cada iteração do treino foi meticulosamente monitorizada para capturar as nuances de desempenho resultantes de cada conjunto específico de parâmetros. Ao final desse experimento, os resultados foram comparados entre si para determinar qual das cinco execuções produziu o modelo mais eficaz, ou seja, aquele com o valor de C-Index mais elevado. Essa comparação permitiu não apenas identificar o melhor modelo, mas também mapear os parâmetros específicos que levaram a esse desempenho superior.

Portanto, o estudo não forneceu apenas um modelo otimizado, mas também informações valiosas sobre como diferentes configurações de parâmetros podem influenciar a eficácia do RSF. Esta abordagem sistemática e cuidadosa garantiu que o modelo final fosse o mais robusto e preciso possível, equipado com os melhores parâmetros identificados durante o processo de experimentação.

Os resultados obtidos ao longo deste estudo e os parâmetros utilizados em cada treino estão expressos na tabela 3:

Tabela 5 - Parâmetros obtidos em cada treino e respetivos resultados

	n estimators	***max depth***	***min samples split***	***min samples leaf***	**C-Index**	***Brier Score***	**AUC**
1º	798	11	19	9	0.8898	0.188	0.837
2º	298	14	11	1	0.8929	0.189	0.843
3º	874	15	20	2	0.8932	0.192	0.845
4º	810	18	11	10	0.8901	0.191	0.839
5º	499	9	4	7	0.8900	0.187	0.835

Ao examinar detalhadamente os resultados apresentados na tabela, é possível concluir que o modelo mais eficiente, dentre todos os avaliados, é o terceiro. Este modelo destacou-se por utilizar uma configuração específica de hiperparâmetros, que inclui 874 estimadores (*n_estimators*), uma profundidade máxima das árvores (*max_depth*) fixada em 15, um valor mínimo de 20 para a divisão dos nós (*min_samples_split*), e um mínimo de 2 amostras por folha (*min_samples_leaf*).
Estes parâmetros foram ajustados com o objetivo de maximizar o desempenho do modelo, e a combinação específica adotada pelo terceiro modelo provou ser a mais eficaz. O critério utilizado para determinar a superioridade deste modelo em relação aos outros foi os valores do C-Index, *Brier Score* e AUC, sendo métricas amplamente utilizadas para avaliar a precisão das previsões em modelos de sobrevivência.

O modelo 3 obteve um C-Index de 0.8932, que é o maior valor registado entre todos os modelos testados, indicando que ele possui a melhor capacidade de discriminar entre os diferentes tempos de eventos nos dados de sobrevivência analisados. Além disso, o modelo 3 obteve valores superiores de *Brier Score* (0.192) e AUC (0.845), embora seja um aumento significativo em relação aos outros modelos. Por outras palavras, este modelo demonstrou ser mais preciso e confiável na previsão dos resultados, tornando-o o melhor para aplicações práticas. A escolha deste conjunto específico de hiperparâmetros resultou num modelo otimizado que combina tanto profundidade na análise quanto robustez nas previsões, destacando-se como a melhor opção dentro do estudo realizado.

5.3. Modelo *Kernel Support Vector Machine*

Com um C-Index de 0.8671, o modelo demonstra uma elevada capacidade de discriminação, ou seja, consegue ordenar corretamente os tempos de sobrevivência em aproximadamente 86.7% das comparações realizadas. Este resultado sugere que o modelo *Kernel* SVM é bastante eficaz na predição de tempos de sobrevivência, sendo capaz de diferenciar de forma robusta entre indivíduos com diferentes riscos de ocorrência do evento.

Este valor do C-Index reflete um bom equilíbrio entre sensibilidade e especificidade, indicando que o modelo está bem ajustado para capturar as complexidades dos dados de sobrevivência, enquanto minimiza os erros na ordenação dos tempos de eventos. Em contextos práticos, um C-Index superior a 0.85 é geralmente considerado um sinal de que o modelo tem uma performance forte, tornando-o adequado para aplicações preditivas e análises de risco em situações reais.

O valor calculado para o *Brier Score* foi de 0.23, o que sugere que, embora o modelo apresente alguma margem de erro, a discrepância média entre as probabilidades previstas e os resultados reais observados é relativamente modesta. Este valor indica que há espaço para melhorias na precisão das previsões, mas ainda assim reflete uma performance razoável.

Por outro lado, o valor da AUC foi de 0.80, o que demonstra que o modelo possui uma capacidade considerável de distinguir entre eventos positivos e negativos. Uma AUC de 0,80 é um indicativo de que o modelo é bastante eficaz em classificar corretamente os diferentes estados dos eventos, mostrando uma performance sólida na discriminação.

Portanto, estes resultados são indicadores positivos da qualidade do modelo *Kernel* SVM treinado, sugerindo que ele poderá ser útil para prever a ocorrência de eventos em dados de sobrevivência, auxiliando na tomada de decisões em cenários onde o tempo até o evento é uma variável crítica.

5.4. Modelo *Multi-Task Logistic Regression*

Com um C-Index de 0.8729, este modelo demonstra uma excelente habilidade de discriminação, conseguindo ordenar corretamente os tempos de sobrevivência em cerca de 87.29% das comparações realizadas. Esse resultado indica que o modelo MTLR é altamente eficiente na previsão dos tempos até a ocorrência de eventos de stress financeiro, mostrando-se robusto na diferenciação entre entidades com diferentes níveis de risco.
Este valor do C-Index reflete um desempenho excecional, sugerindo que o modelo está bem ajustado para capturar as complexidades dos dados financeiros analisados, ao mesmo tempo em que minimiza erros na ordenação dos tempos de eventos. Em cenários práticos, um C-Index superior a 0.85 é geralmente visto como um sinal de que o modelo tem uma performance sólida, tornando-o adequado para previsões de risco e apoio à tomada de decisões em contextos onde o tempo até o evento é uma variável crucial.

Para o modelo MTLR, o valor do *Brier Score* obtido foi de 0.24. Este resultado indica uma margem de erro moderada nas previsões feitas pelo modelo, refletindo uma discrepância média de 0.24 entre as probabilidades previstas e os resultados reais observados. Embora este valor sugira que o modelo ainda apresenta algum nível de imprecisão, ele também aponta para uma performance razoável em termos de previsões.

Em contraste, a AUC foi de 0.80, o que evidencia uma boa capacidade do modelo em distinguir entre eventos positivos e negativos. Com uma AUC de 0.80, o modelo MTLR demonstra uma performance sólida, sendo eficaz na separação correta dos diferentes estados dos eventos.

Assim, estes resultados são um forte indicativo da eficácia do modelo MTLR treinado, sugerindo que ele pode ser uma ferramenta útil para antecipar a ocorrência de eventos de stress financeiro, oferecendo suporte eficaz na identificação e gestão de riscos em situações reais.

5.5. Modelo *DeepSurv*

Ao treinar um modelo *DeepSurv* para prever eventos de stress financeiro, o C-Index foi utilizado como uma métrica chave para avaliar o desempenho do modelo. No caso específico deste treinamento, o C-Index obtido foi de 0.7931, com o modelo sendo treinado utilizando os parâmetros definidos como *default*.

Com um C-Index de 0.7931, o modelo demonstra uma boa capacidade de discriminação, conseguindo ordenar corretamente os tempos de sobrevivência em aproximadamente 79.31% das comparações realizadas. Esse resultado indica que o modelo *DeepSurv* é eficaz na previsão dos tempos até a ocorrência de eventos de stress financeiro, capturando as relações não lineares entre as variáveis preditivas e o risco associado.

Embora o C-Index de 0.7931 seja ligeiramente inferior ao de outros modelos, ele ainda reflete uma performance respeitável, especialmente considerando a complexidade dos dados e a natureza não linear das interações capturadas pelo *DeepSurv*. Em cenários práticos, um C-Index próximo de 0.8 ainda é considerado bom, sugerindo que o modelo pode ser útil para aplicações preditivas e análises de risco, embora possa haver espaço para melhorias adicionais.

No caso do modelo *DeepSurv*, o *Brier Score* registado foi de 0.31, indicando uma margem de erro relativamente ampla nas previsões realizadas. Este valor reflete uma diferença média de 0.31 entre as probabilidades estimadas pelo modelo e os resultados reais observados, sugerindo que há uma necessidade significativa de melhorar a precisão das previsões.

Relativamente ao valor da AUC, este de 0.69, o que demonstra uma capacidade moderada do modelo em diferenciar entre eventos positivos e negativos. Com uma AUC de 0.69, o modelo *DeepSurv* apresenta uma habilidade de discriminação que é aceitável, mas que ainda pode ser aperfeiçoada para oferecer uma distinção mais clara entre os estados dos eventos. Assim, enquanto o *Brier Score* sugere que o modelo necessita de ajustes para refinar a precisão das previsões, a AUC indica que, embora o modelo tenha um desempenho razoável na classificação, há espaço para melhorias na sua capacidade de discriminar eficazmente entre diferentes tipos de eventos.
Portanto, estes resultados sugerem que o modelo *DeepSurv* treinado é uma ferramenta válida para prever a ocorrência de eventos de stress financeiro, contribuindo para a

identificação e gestão de riscos em cenários onde o tempo até o evento é um fator determinante.

5.6. Comparação dos Resultados dos Modelos

Como mencionado anteriormente, na tentativa de prever eventos de stress financeiro nas organizações, foram avaliados cinco diferentes modelos de sobrevivência: *Cox*, RSF, *Kernel* SVM, MTLR e *DeepSurv*. Cada um desses modelos foi treinado e testado para determinar a sua eficácia na previsão de eventos de stress financeiro. Esta subsecção tem como objetivo comparar os resultados obtidos em cada uma das métricas utilizadas (C-Index, *Brier Score* e AUC) para avaliar o melhor que obteve melhores resultados e concluir qual deles é o melhor modelo para prever eventos de stress financeiro.

5.6.1. C-Index

No geral, todos os modelos demonstraram um desempenho bastante positivo. Especificamente, todos eles conseguiram alcançar um C-Index superior a 0.8, o que indica uma excelente capacidade de discriminação e uma boa habilidade para ordenar corretamente os tempos até a ocorrência dos eventos. Este valor do C-Index sugere que os modelos foram capazes de identificar e classificar de forma eficaz os diferentes níveis de risco de stress financeiro nas organizações, fornecendo previsões robustas e confiáveis.

Contudo, é importante notar que o modelo *DeepSurv* apresentou um C-Index ligeiramente inferior, com um valor de 0.79. Embora este valor seja um pouco menor comparado com os C-Index dos outros modelos, ainda reflete uma boa performance na previsão dos eventos de stress financeiro. O *DeepSurv* continua a ser uma ferramenta útil para a análise de sobrevivência, capturando as relações complexas entre as variáveis financeiras e o risco associado, mas com uma precisão ligeiramente menor em comparação com os outros modelos avaliados.

A Figura 3, que se encontra apresentada abaixo, exibe um gráfico de barras que ilustra os valores de C-Index obtidos para os diversos modelos de sobrevivência treinados. Este

gráfico fornece uma visão clara e comparativa do desempenho de cada modelo na previsão de eventos de stress financeiro.

A análise detalhada deste gráfico revela que o modelo *DeepSurv* apresentou o menor valor de C-Index, com um resultado de 0.79. Este valor indica que, embora o modelo seja útil, ele teve um desempenho relativamente inferior em comparação com os outros modelos testados. Em seguida, o modelo *Cox* obteve um valor de C-Index de 0.86, demonstrando uma capacidade de discriminação superior ao *DeepSurv*, mas ainda abaixo de alguns outros modelos avaliados.

Os modelos *Kernel* SVM e MTLR mostraram um desempenho idêntico, ambos apresentando um C-Index aproximado de 0.87. Isso indica que estes modelos tiveram uma capacidade comparável de ordenar corretamente os tempos até a ocorrência dos eventos, colocando-os entre os melhores desempenhos entre os modelos analisados.

No entanto, o modelo que se destacou em termos de eficácia foi o RSF, que alcançou o valor mais alto de C-Index, com um resultado de 0.89. Este valor superior sugere que o RSF foi o modelo mais eficaz na previsão de eventos de stress financeiro, demonstrando a melhor capacidade de discriminação e ordenação dos tempos de sobrevivência entre todos os modelos testados.

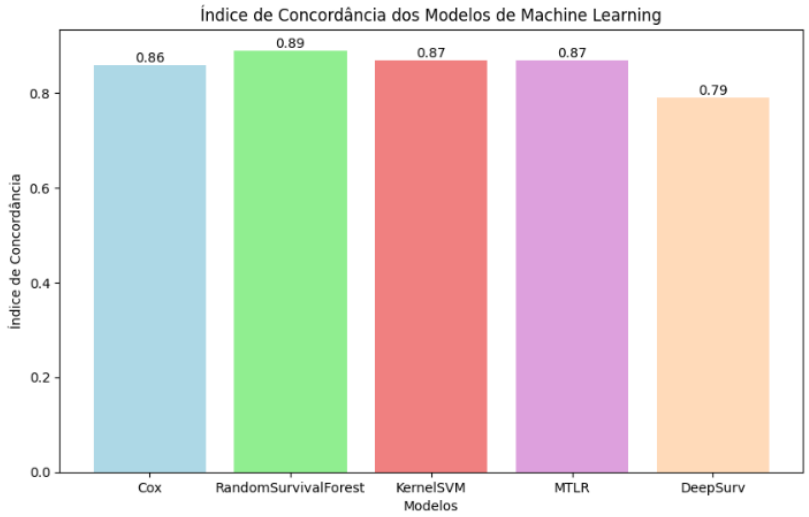


Figura 3 - Índices de Concordância

5.6.2. *Brier Score*

A análise dos valores do *Brier Score* revela diferenças marcantes na precisão das previsões fornecidas por cada modelo. O modelo RSF destaca-se como o mais preciso, apresentando um *Brier Score* de 0.19. Este valor é o mais baixo entre os modelos avaliados e indica que o RSF tem a menor margem de erro nas suas previsões, o que reflete uma excelente capacidade de prever eventos com alta precisão. Ou seja, as previsões feitas pelo RSF estão, em média, mais próximas dos resultados reais, sugerindo que este modelo é o mais eficaz na tarefa de previsão.

O modelo *Kernel* SVM, com um *Brier Score* de 0.23, também demonstra uma performance sólida, embora não alcance o nível de precisão do RSF. Com uma margem de erro relativamente baixa, o *Kernel* SVM é eficaz na previsão dos eventos, mas há uma ligeira discrepância em comparação com o RSF. Isso ainda indica uma boa capacidade preditiva, mas com um pequeno espaço para melhorias na precisão.

O MTLR apresenta um *Brier Score* de 0.24, situando-se entre o *Kernel* SVM e o *Cox*. Este valor indica uma precisão razoável, sendo mais preciso que o modelo *Cox*, mas com uma margem de erro ligeiramente maior em comparação com o *Kernel* SVM. Assim, o MTLR oferece um desempenho aceitável, mas há uma oportunidade para otimizar o modelo e reduzir ainda mais a discrepância nas previsões.

O modelo *Cox*, com um *Brier Score* de 0.26, revela um desempenho intermediário. Este valor sugere que o modelo *Cox* tem uma margem de erro moderada nas suas previsões, o que indica que, embora seja funcional, há necessidade de melhorias para alcançar uma precisão superior. A discrepância média entre as previsões e os resultados reais é maior em comparação com os modelos RSF, *Kernel* SVM e MTLR.

Por fim, o modelo *DeepSurv* apresenta o *Brier Score* mais alto, de 0.31. Este valor indica a maior margem de erro entre todos os modelos avaliados, o que reflete uma precisão relativamente baixa. O *DeepSurv* tem a menor capacidade preditiva dos modelos analisados, sugerindo que é necessário realizar ajustes substanciais para melhorar a precisão das suas previsões. Em resumo, o RSF destaca-se como o modelo mais preciso, seguido pelo *Kernel* SVM e MTLR, com o *Cox* oferecendo um desempenho intermediário e o *DeepSurv* apresentando a maior imprecisão. As diferenças nos *Brier Scores* destacam

a necessidade de otimizações específicas para os modelos com maior margem de erro, a fim de aprimorar sua capacidade de previsão.

A Figura 4, apresentada a seguir, demonstra um gráfico de barras que ilustra os valores de *Brier Score* obtidos para os diversos modelos de sobrevivência treinados:

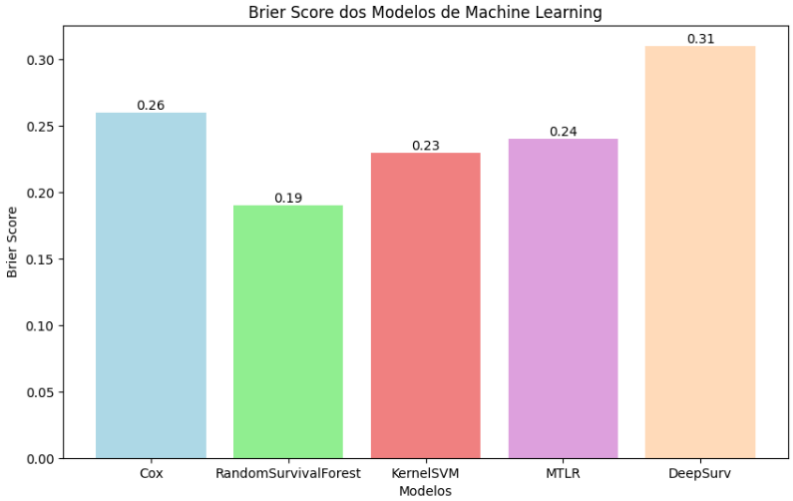


Figura 4 - Brier Score

5.6.3. AUC

A análise dos valores da AUC para os diferentes modelos, conforme ilustrado no gráfico de barras, representado na Figura 5, revela diferenças significativas na capacidade discriminatória de cada abordagem.

O modelo RSF destaca-se com a AUC mais alta, que é 0.83. Este valor sugere que o RSF possui uma excelente capacidade de distinguir entre eventos positivos e negativos, com uma performance superior na classificação dos casos. A AUC de 0.83 indica que o modelo tem uma alta taxa de verdadeiros positivos em relação aos falsos positivos, proporcionando uma boa separação entre os diferentes estados dos eventos.

O *Kernel* SVM e o MTLR apresentam ambos uma AUC de 0.80, indicando uma capacidade de discriminação bastante robusta. Esses valores mostram que tanto o *Kernel* SVM quanto o MTLR são eficazes em classificar eventos, com um desempenho similar e consistente. A AUC de 0.80 sugere que estes modelos têm uma boa capacidade de

separar os eventos positivos dos negativos, embora não alcancem o nível superior de precisão mostrado pelo RSF.

O modelo *Cox*, com uma AUC de 0.78, demonstra uma capacidade de discriminação ligeiramente inferior em comparação com os modelos RSF, *Kernel* SVM e MTLR. Embora ainda apresente uma AUC robusta, indicando uma boa habilidade em separar os diferentes estados dos eventos, o *Cox* não é tão eficaz quanto os modelos com AUC mais alta. A diferença, embora não seja extremamente grande, sugere que o *Cox* tem um desempenho um pouco mais limitado na identificação correta dos eventos.

Por fim, o modelo *DeepSurv* possui a AUC mais baixa, de 0.69. Este valor indica que o *DeepSurv* tem a menor capacidade discriminatória entre os modelos analisados. A AUC de 0.69 sugere que o modelo tem uma menor taxa de verdadeiros positivos em relação aos falsos positivos, o que reflete uma capacidade reduzida de diferenciar entre eventos positivos e negativos. Este resultado sugere que o *DeepSurv* tem uma performance menos eficaz na classificação, o que pode ser uma área significativa para melhorias.

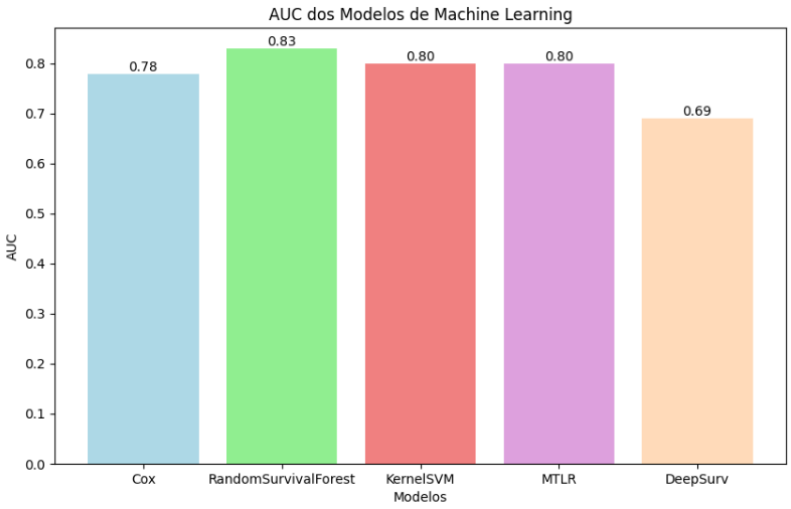


Figura 5 - AUC

Em resumo, o RSF é o modelo que oferece a melhor capacidade discriminatória, seguido pelos modelos *Kernel* SVM e MTLR, que têm desempenhos semelhantes e bastante fortes. O *Cox* apresenta uma capacidade de discriminação um pouco menor, e o *DeepSurv*, com a AUC mais baixa, mostra a menor eficiência na separação entre eventos positivos e negativos. As diferenças nas AUC's destacam a eficácia variável dos modelos

em termos de discriminação e sugerem que o *DeepSurv* pode necessitar de ajustes significativos para melhorar sua capacidade preditiva.

5.6.4. Comparação com os resultados obtidos na revisão de literatura

Este capítulo analisa e compara os resultados de dois estudos anteriores com os obtidos no presente trabalho, que tem como objetivo prever o stress financeiro nas organizações. Os artigos em questão pertencem aos autores Ha *et al.* (2023) e Tran *et al.* (2022), ambos explorando diferentes modelos preditivos para a previsão de stress financeiro.

Ha *et al.* (2023), utilizaram seis métodos de classificação tradicionais (Regressão Logística, SVM, Árvores de Decisão, RF, KNN e Redes Neuronais) com base em três modelos financeiros: Altman, Fich e Slezak, e Zmijewski. A métrica de avaliação foi a *accuracy*, e os melhores resultados foram observados para modelos baseados em árvores, com a RF alcançando uma *accuracy* de 0.98 para Altman e Zmijewski. Modelos como KNN e Redes Neuronais tiveram desempenhos mais fracos, com *accuracy* de 0.81 a 0.87 nos diferentes modelos financeiros.

O estudo de Tran *et al.* (2022) investigou o desempenho de vários modelos de classificação, incluindo o Extreme Gradient Boosting (XGBoost), RF, Regressão Logística, Redes Neuronais Artificiais, Árvores de Decisão e SVM. Aqui, a principal métrica foi a AUC, com o RF atingindo o valor mais alto (0.9788), seguido de perto pelo XGBoost (0.9702). A *accuracy* geral foi mais alta para o XGBoost (95.66%). No entanto, o SVM apresentou a menor AUC (0.7889), com *recall* e *F1-score* relativamente baixos, o que reflete a sua dificuldade em identificar instâncias positivas de insolvência.
Em termos de *accuracy* e AUC, observa-se que os modelos baseados em árvores (RF e Árvores de Decisão) destacaram-se tanto nos artigos analisados quanto no presente estudo. No artigo referente aos autores Ha *et al.* (2023), a RF alcançou uma *accuracy* de 0.98, e no artigo referente aos autores Tran *et al.* (2022), o RF teve uma AUC de 0.9788. No presente estudo, o RSF manteve essa superioridade, com os melhores resultados em todas as métricas.

Em contrapartida, o desempenho do SVM apresentou variabilidade. No artigo dos autores Ha *et al.* (2023), o SVM alcançou *accuracy* entre 0.89 e 0.97, enquanto no artigo dos autores Tran *et al.* (2022), a AUC foi a mais baixa entre todos os modelos (0.7889). No presente estudo, o Kernel SVM mostrou-se competitivo, com C-Index de 0.87 e AUC de 0.80, mas não superou o desempenho de modelos baseados em árvores.

Os modelos de Regressão Logística também apresentaram um comportamento interessante. No artigo dos autores Ha *et al.* (2023), a Regressão Logística apresentou *accuracy* de até 0.97 (para o modelo Zmijewski), e no artigo dos autores Tran *et al.* (2022), o valor mais alto da AUC foi 0.9303. Já no presente estudo, o modelo Cox atingiu um C-Index de 0.86 e AUC de 0.78, que, embora sejam números sólidos, não se aproximam dos valores obtidos pelas árvores de decisão.

Os resultados dos dois artigos analisados e do presente estudo apontam para um padrão claro: modelos baseados em árvores, especialmente o RF e RSF, consistentemente demonstram forte desempenho na previsão de stress financeiro. Essa robustez reflete-se em métricas como *accuracy*, AUC e C-Index. Embora outros métodos, como Regressão Logística e SVM, tenham mostrado resultados promissores em alguns cenários, estes não foram capazes de igualar a eficiência dos modelos de árvores. No contexto de modelos de sobrevivência, o RSF destaca-se como o método mais confiável, oferecendo um equilíbrio eficaz entre diferentes métricas.

6. Conclusões

Neste capítulo, são apresentadas as principais conclusões do projeto, que visou a previsão de eventos de stress financeiro em organizações utilizando modelos de sobrevivência baseados em técnicas de ML. Através da análise e comparação de cinco modelos distintos — Regressão de *Cox*, RSF, *Kernel* SVM, MTLR e *DeepSurv* — foi possível avaliar a eficácia de cada um na identificação e antecipação de eventos de stress financeiro.

Os resultados obtidos revelaram que todos os modelos testados foram capazes de prever eventos de stress financeiro com um nível elevado de precisão, como evidenciado pelos valores do C-Index, do *Brier Score* e da AUC. O modelo RSF destacou-se dos demais, apresentando o melhor desempenho global, com um C-Index de 0.89, um *Brier Score* de 0.19 e uma AUC de 0.83. Este resultado sublinha a capacidade superior deste modelo em discriminar corretamente entre diferentes tempos de ocorrência de eventos financeiros críticos, tornando-o uma ferramenta poderosa para a previsão e gestão de risco.

O *Kernel* SVM e o MTLR também se mostraram bastante eficazes, ambos com C-Index próximo de 0.87, o que indica que esses modelos são particularmente robustos na análise de sobrevivência em contextos financeiros. Os valores de AUC obtidos de ambos os modelos foram também semelhantes, registando valores de 0.80. No entanto, os valores obtidos de *Brier Score* foram ligeiramente superiores no modelo MTLR (0.24) do que no modelo *Kernel* SVM (0.23). A Regressão de *Cox*, embora tenha sido superada pelos modelos mais avançados, manteve-se como uma opção confiável, com um C-Index de 0.86, um *Brier Score* de 0.26 e uma AUC de 0.78 provando que ainda é uma metodologia relevante e útil em análises de sobrevivência.

Por outro lado, o modelo *DeepSurv*, baseado em redes neuronais, apresentou o menor desempenho entre os modelos analisados, com um C-Index de 0.79, um *Brier Score* de 0.31 e uma AUC de 0.69. Apesar de ser o modelo com a menor capacidade preditiva, o *DeepSurv* mostrou-se valioso na captura de relações não lineares complexas, o que sugere que, com aprimoramentos adicionais ou em cenários específicos, este modelo pode oferecer contribuições importantes.

No final, com a utilização da técnica "*hyperparameter tunning*" foi possível descobrir quais os melhores parâmetros para o modelo com maiores valores obtidos em cada

métrica, o modelo RSF, sendo eles: 874 para o *n_estimators*, 15 para o *max_depth*, 20 para o *min_samples_split* e 2 para o *min_samples_leaf*.
Com a utilização destes parâmetros os valores de cada métrica não se alteraram muito, sendo que inicialmente o valor de C-Index era de 0.89, passando para 0.8932. No caso do *Brier Score*, inicialmente o seu valor era de 0.19, passando para 0.192 e, relativamente à métrica AUC, o seu valor era de 0.83 e passou para 0.845. No entanto, esta técnica permitiu otimizar ainda mais o modelo, aumentando ligeiramente o seu valor de cada uma das métricas utilizadas para medir o desempenho dos modelos.

Em termos práticos, os resultados deste estudo demonstram que a aplicação de modelos de sobrevivência baseados em ML é uma abordagem viável e eficaz para a previsão de eventos de stress financeiro. A capacidade de prever com precisão quando estes eventos podem ocorrer oferece às organizações uma vantagem estratégica significativa, permitindo-lhes implementar medidas preventivas e de mitigação antes que os eventos ocorram.

Além disso, o processo de seleção e preparação dos dados, incluindo a identificação e remoção de variáveis altamente correlacionadas, foi crucial para a otimização do desempenho dos modelos. Este cuidado na preparação dos dados reflete-se na qualidade dos resultados obtidos, reforçando a importância de uma abordagem meticulosa na análise de dados.

No entanto, este estudo também aponta para algumas limitações. A variabilidade nos desempenhos dos modelos sugere que a escolha do modelo deve ser feita com base nas características específicas do conjunto de dados e no contexto da aplicação. Além disso, embora os resultados tenham sido promissores, futuras pesquisas poderiam explorar a integração de outros modelos ou técnicas, assim como a aplicação de métodos de seleção de características mais avançados para potencialmente melhorar ainda mais a precisão preditiva.

Em conclusão, este trabalho contribui significativamente para o campo da análise de sobrevivência e da previsão de stress financeiro, ao demonstrar a aplicabilidade e eficácia de modelos de ML neste domínio. As descobertas fornecem uma base sólida para futuras pesquisas e práticas, reforçando a importância da escolha adequada do modelo e da

preparação rigorosa dos dados para a previsão precisa e gestão eficaz de eventos financeiros críticos.

6.1. Trabalho futuro

Como trabalho futuro, a previsão de eventos de stress financeiro utilizando modelos de sobrevivência com ML pode ser aperfeiçoada em várias direções. Uma possibilidade é a exploração de novos algoritmos, como redes neuronais profundas de sobrevivência, que podem capturar relações não lineares complexas nos dados. Além disso, o uso de técnicas de *explainability* (métodos que tornam os modelos de ML mais interpretáveis, ou seja, explicam como e porquê de um modelo tomar determinadas decisões), como SHAP ou LIME, pode fornecer maior transparência aos modelos, ajudando a interpretar melhor as principais variáveis que contribuem para o stress financeiro.

Além disso, existem várias melhorias que podem ser seguidas para enriquecer este projeto no futuro, como por exemplo, a utilização de dados não estruturados, como notícias económicas ou relatórios de auditoria, através do processamento de linguagem natural. A obtenção de dados financeiros mais detalhados incluindo variáveis externas como fatores macroeconómicos, mudanças regulatórias ou crises setoriais, pode melhorar a capacidade preditiva dos modelos. A criação de novas variáveis que capturem interações ou comportamentos complexos das empresas ao longo do tempo pode enriquecer o conjunto de dados e tornar as previsões mais precisas. Isso inclui a identificação de padrões sazonais, tendências ou mudanças abruptas em indicadores financeiros que possam antecipar o stress financeiro.

Referências bibliográficas

Altman E.I. (1968). Financial ratios discriminant analysis and the prediction of corporate bankruptcy, J. Finance 23, 589–609.

Assel, M.; Sjoberg, D. D.; Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and prognostic research, v. 1, n. 1, 2017. https://doi.org/10.1186/s41512-017-0020-3

Beaver, W. (1966). Financial ratios as predictors of failure, J. Account. Res. 4, 71–111.

Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., & Castelnuovo, B. (2018). A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. Em BMC Medical Informatics and Decision Making (Vol. 18, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s12911-018-0659-x

Bose, I. (2006). Deciding the financial health of dot-coms using rough sets, Inform. Manage. 43, 835–846.

Bonsall, S.B., Bozanic, Z., & Fischer, P.E. (2013). What do management earnings forecasts convey about the macroeconomy?, Journal of Accounting Research 51, 225–266.

Carminchael, D.R. (1972). The auditor's reporting obligation. Auditing Res. Monogr. (1) (New York: AICPA) 94–94.

Chaopeng, L., Huang, X., & Huang, W. (2024). Optimizing recurrent neural networks: A study on gradient normalization of weights for enhanced training efficiency. *Applied Sciences*, 14(15), 6578. https://doi.org/10.3390/app14156578

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Chordia, T., & Shivakumar, L. (2005). Inflation illusion and post-earnings-announcement drift, Journal of Accounting Research 43, 521–556.

Colosimo, E. A., & Giolo, S. Ruiz. (2006). *Análise de sobrevivência aplicada*. Edgard Blücher.

Deakin, E.B. (1972). A discriminant analysis of prediction of business failure, J. Account. Res. 3 (spring), 167–169.

Ding, F., Yang, H. Z., & Liu, F. (2023). Performance analysis of stochastic gradient algorithms under weak conditions. *Science China Information Sciences*, 51(8), 1269-1280. https://doi.org/10.1007/s11432-023-16109-x

Ding, Y., Song, X., & Zeng, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine, Expert Syst. Appl. 34, 3081–3089.

Doumpos, M., & Zopounidis, C. (1999). A multinational discrimination method for the prediction of financial distress: the case of Greece, Multinatl. Finan. J. 3, 71–101.

Foster, G. (1986). Financial Statement Analysis, 2nd ed., Prentice Hall, NJ.

Gilson, S.C., & Vetsuypens, M.R. (1993). CEO compensation in financially distressed firms: an empirical analysis, The Journal of Finance 48, 425–458.

Giroux, G.A., & Wiggins, C.E. (1984). An events approach to corporate bankruptcy, Journal of Bank Research 15, 179–187.

Gorriz, J. M., Segovia F., Ramirez J. (2024). Is K-fold cross validation the best model selection method for Machine Learning? https://doi.org/10.48550/arXiv.2401.16407

Gu, M. (2017). Distress risk, investor sophistication, and accrual anomaly, Journal of Accounting, Auditing and Finance. Forthcoming. https://doi.org/10.1177/0148558X17696762

Ha, H. H.; Dang, N. H.; Tran, M. D. (2023) Financial distress forecasting with a machine learning approach. Corporate governance and organizational behavior review, v. 7, n. 3, p. 90–104. https://doi.org/10.22495/cgobrv7i3p8

Habib, A., Costa, M.D., Huang, H.J., Bhuiyan, M.B.U., & Sun, L. (2020). Determinants and consequences of financial distress: review of the empirical literature. Account Finance, 60, 1023-1075. https://doi.org/10.1111/acfi.12400

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. Em The Annals of Statistics (Vol. 36, Issue 3). Institute of Mathematical Statistics. https://doi.org/10.1214/009053607000000677

Huang, J.; Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE transactions on knowledge and data engineering, v. 17, n. 3, p. 299–310, 2005.

Informa D&B. (n.d.). SABI: Iberian Balance Sheet Analysis System. https://sabi.informa.es/ip

Ishwaran, H., & Lu, M. (2019). Random Survival Forests. Em Wiley StatsRef: Statistics Reference Online (pp. 1–13). Wiley. https://doi.org/10.1002/9781118445112.stat08188

Johnson, S., Boone, P., Breach, A., & Friedman, E. (2000). Corporate governance in the Asian financial crisis, Journal of Financial Economics 58, 141–186.

Jostarndt, P., & Sautner, Z. (2008). Financial distress, corporate control, and management turnover, Journal of Banking and Finance 32, 2188–2204.

Kane, G.D., Velury, U., & Ruf, B.M. (2005). Employee relations and the likelihood of occurrence of corporate financial distress, Journal of Business Finance and Accounting 32, 1083–1105.

Li, C., Zhang, J., & He, T. (2024). Why gradient clipping accelerates training: A theoretical justification. Proceedings of International Conference on Learning Representations, Addis Ababa.

Lin, T.-H. (2009). A cross model study of corporate financial distress prediction in Taiwan: multiple discriminant analysis, logit, probit and neural networks models, Neurocomputing 72, 3507–3516.

Liou, D.K., & Smith, M. (2007). Macroeconomic variables and financial distress, Journal of Accounting, Business and Management 14, 17–31.

Longato, E.; Vettoretti, M.; Di Camillo, B. (2020) A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. Journal of biomedical informatics, v. 108, n. 103496, p. 103496, 2020. https://doi.org/10.1016/j.jbi.2020.103496

Magee, S. (2013). The effect of foreign currency hedging on the probability of financial distress, Accounting and Finance 53, 1107–1127.

Monteiro, R., de Castro Machado Rabello, G., Vidal de Arruda Junior, F., & Biscegli Jatene, F. (2022). Inteligência Artifical, Deep Learning, Machine Learning, Redes Neuronais na Medicina e Biomarcadores Vocais: Conceitos, Onde Estamos e para Onde Vamos. Revista Da Sociedade de Cardiologia Do Estado de São Paulo, 32(1), 11–17. https://doi.org/10.29381/0103-8559/2022320111-7

Perrigot, R., Cliquet, G., & Mesbah, M. (2004). Possible applications of survival analysis in franchising research. The International Review of Retail, Distribution and Consumer Research, 14(1), 129-143.

Ravisankar, P., Ravi, V., & Bose, I. (2010). Failure prediction of dotcom companies using neural network-genetic programming hybrids, Inf. Sci. 180, 1257–1267.

Rafiei, F.M., Manzari, S.M., & Bostanian, S. (2011). Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence, Expert Syst. Appl. 38, 10210–10217.

Ribeiro, P.M.F. (2022). Universidade do Minho School of Engineering Machine Learning Applied to Companies Management. Disponível em: https://hdl.handle.net/1822/84499

Ross, S.A., Westerfield, R.W., & Jaffe, J.F. (1999). Corporate finance, second ed., Homewood IL.

Shleifer, A., & Vishny, R.W. (1997). A survey of corporate governance, The Journal of Finance 52, 737–783.

Sliusarenko, T., Pohurska, M. (2024). STOP OVERFITTING WITH PROVEN TECHNIQUES. Grail of Science, n. 40, p. 373–375. https://doi.org/10.36074/grail-of-science.07.06.2024.057

Smith, C.W., & Stulz, R.M. (1985). The determinants of firms' hedging policies, Journal of Financial and Quantitative Analysis 20, 391–405.

Souza, J. A. De; Komati, K. S.; Andrade, J. O. (2022) Análise de Sobrevivência: um estudo de caso em um Curso de Sistemas de Informação. Anais do XXX Workshop sobre Educação em Computação. https://doi.org/10.5753/wei.2022.223357

Sudarsanam, S., & Lai, J. (2001). Corporate financial distress and turnaround strategies: an empirical analysis, British Journal of Management 12, 183–199.

Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. Knowledge-Based Systems, 57, 41–56. https://doi.org/10.1016/j.knosys.2013.12.006

Sun, J., Jia, M., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies, Expert Syst. Appl. 38, 9305–9312.

Sun, J., & Hui, X.-F. (2006). Financial distress prediction based on similarity weighted voting CBR, Lect. Notes Artif. Int. 4093, 947–958.

Theodossiou, P., Kahya, E., Saidi, R., & Philippatos, G. (1996). Financial distress and corporate acquisitions: further empirical evidence, Journal of Business Finance and Accounting 23, 699–719.

Tennyson, B.M., Ingram, R.W., & Dugan, M.T. (1990). Assessing the information content of narrative disclosures in explaining bankruptcy, Journal of Business Finance and Accounting 17, 391–410.

Tran, K. L. *et al.* (2022) Explainable machine learning for financial distress prediction: Evidence from Vietnam. Data, v. 7, n. 11, p. 160. https://doi.org/10.3390/data7110160

Vallarino, D. (2024). A Comparative Machine Learning Survival Models Analysis for Predicting Time to Bank Failure in the US (2001-2023). *Journal of Economic Analysis*, *3*(1), 50. https://doi.org/10.58567/jea03010007

WANG, P.; LI, Y.; REDDY, C. K. (2019) Machine learning for survival analysis: A survey. ACM computing surveys, v. 51, n. 6, p. 1–36.

Wang, Z., Lee, J. W., Chakraborty, T., Ning, Y., Liu, M., Xie, F., Ong, M. E. H., & Liu, N. (2024). Survival modeling using deep learning, machine learning and statistical methods: A comparative analysis for predicting mortality after hospital admission (Versão 1). arXiv. https://doi.org/10.48550/ARXIV.2403.06999

ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

M MESTRADO
ENGENHARIA INFORMÁTICA

Aplicação de Modelos de *Machine Learning* para Previsão de Eventos de Stress Financeiro

Beatriz Fernandes

Orientadores:

Professora Doutora Mariana Valério de Carvalho

Professora Doutora Ana Isabel Borges