Correlação entre Características de Riscos Cibernéticos e Possibilidade de Ocorrência em Empresas - Uma Abordagem com Regressão Logística e *Machine Learning*

José Henrique Viana Santos, Universidade Federal do Ceará
Germano Fenner, Universidade Federal do Ceará
Fábio Lotti Oliva, Universidade de São Paulo
Elisângela da Silva Rodrigues, Universidade Federal do Ceará
Celso Cláudio de Hildebrand e Grisi, Universidade de São Paulo

Resumo

Empresas ao identificarem seus riscos, podem atribuir diferentes características tais como impacto, probabilidade, resposta a ser adotada, área de maior impacto na organização. As características variam de acordo com tamanho, maturidade e cultura de riscos do negócio. Este estudo propõe uma análise de risco cibernético aplicada a três empresas distintas, utilizando modelos de regressão logística ordinal e de *machine learning*. A pesquisa identifica as características de um risco e como elas influenciam na sua possibilidade de ocorrência em cada uma das organizações onde o modelo foi aplicado. Os resultados sugerem que características como eliminar riscos e o tempo percepção são os fatores mais críticos para a perspectiva de ocorrência. As conclusões apontam para a necessidade de uma abordagem flexível, capaz de lidar com a complexidade dos riscos cibernéticos de diferentes naturezas e intensidades.

Palavras-chave: Riscos cibernéticos; características; regressão logística ordinal; *machine learning*.

Abstract

When companies identify their risks, they can assign different characteristics such as impact, probability, response to be adopted, and area of greatest impact in the organization. The characteristics vary according to the size, maturity, and risk culture of the business. This study proposes a cyber risk analysis applied to three different companies, using ordinal logistic regression and machine learning models. The research identifies the characteristics of a risk and how they influence its possibility of occurrence in each of the organizations where the model was applied. The results suggest that characteristics such as eliminating risks and time perception are the most critical factors for the probability of occurrence. The conclusions point

to the need for a flexible approach, capable of dealing with the complexity of cyber risks of different natures and intensities.

***Keywords*:** Cyber risks; characteristics; ordinal logistic regression; machine learning.

1. Introdução

No contexto da Gestão de Riscos (GR) e da Segurança da Informação (SI), empresas buscam continuamente estratégias eficazes para lidar com incertezas e minimizar as probabilidades de materialização dos riscos identificados. Uma das abordagens amplamente utilizadas envolve atribuir características ou atributos específicos aos riscos identificados, permitindo, assim, diferentes categorizações e perspectivas de controle, alinhadas às orientações da ABNT ISO/IEC 27002 (2022).

Entretanto, apesar do crescente reconhecimento da importância da Segurança Cibernética (SC), especialmente em virtude do aumento significativo de ataques cibernéticos (AC) (Pandey, Singh, Gunasekaran e Kaushik, 2020), observa-se que a literatura existente ainda apresenta lacunas relevantes. Particularmente, há uma insuficiência teórica relacionada à identificação clara das características específicas dos riscos cibernéticos (RC) que mais influenciam na sua materialização, bem como a falta de modelos estruturados que relacionem essas características à probabilidade de ocorrência efetiva dos ataques nas organizações (Demirkan e Mckee, 2020; Mouton e Coning, 2020).

Considerando essa lacuna, este estudo pretende contribuir diretamente para a literatura, propondo um modelo capaz de estabelecer correlações entre as características dos riscos cibernéticos e sua probabilidade de ocorrência em empresas. Desta forma, o estudo é orientado pela seguinte questão de pesquisa: Quais características específicas dos riscos cibernéticos exercem maior influência sobre a probabilidade de ocorrência desses riscos nas organizações?

Para responder a essa questão, este trabalho utiliza métodos quantitativos, incluindo técnicas de regressão logística e *machine learning*, visando identificar e analisar precisamente os atributos que impactam significativamente a materialização dos riscos cibernéticos nas organizações estudadas.

2. Referencial Teórico

a) Modelos Estatísticos

Para apoiar decisões relacionadas ao gerenciamento de riscos cibernéticos (GRCb), é essencial empregar abordagens analíticas capazes de identificar fatores críticos e prever com precisão a ocorrência de eventos adversos. Dentre essas abordagens, os modelos estatísticos desempenham papel fundamental, especialmente a Regressão Logística Ordinal (RLO). A RLO foi escolhida por ser particularmente adequada para variáveis categóricas ordenadas, que frequentemente surgem em contextos de segurança cibernética, onde níveis de severidade dos riscos apresentam uma ordem natural, mas não necessariamente intervalos constantes entre eles (Agresti, 2018). Hosmer et al. (2013) destacam que a RLO oferece coeficientes interpretáveis na forma de *odds ratios*, permitindo uma análise clara e direta sobre como determinadas características aumentam ou diminuem a probabilidade da ocorrência de riscos em níveis mais críticos. Kleinbaum e Klein (2010) reforçam a robustez dessa técnica na análise da influência relativa de múltiplas variáveis, contribuindo diretamente para o entendimento dos fatores de risco mais relevantes. Estudos recentes, como os de Long e Freese (2014) e Williams (2020), consolidam ainda mais a eficácia da RLO na análise categórica ordenada, enfatizando sua aplicabilidade em diferentes contextos organizacionais, incluindo gestão de riscos cibernéticos.

b) Modelos de Aprendizado de Máquina

Embora modelos estatísticos como a RLO sejam eficazes na análise linear e na interpretação direta dos resultados, o gerenciamento de riscos cibernéticos frequentemente requer a identificação de padrões complexos e interações não lineares entre múltiplas variáveis. Neste contexto, os modelos de aprendizado de máquina, especificamente Random Forest (RF) e *Gradient Boosting* (GB), são abordagens complementares ideais, capazes de capturar relações complexas que modelos estatísticos tradicionais podem não evidenciar claramente. O RF foi escolhido pela sua capacidade comprovada de lidar com grandes volumes de dados, combinando múltiplas árvores de decisão para alcançar previsões robustas e estáveis (Breiman, 2001). Essa técnica facilita a identificação das variáveis mais relevantes, ajudando as organizações a priorizar esforços e recursos na mitigação de riscos cibernéticos críticos (Zhang e Ma, 2012). Por outro lado, o GB oferece uma abordagem iterativa e incremental, construindo árvores de decisão sequenciais, cada uma corrigindo erros das anteriores. De acordo com Friedman (2001), essa metodologia é particularmente eficaz em destacar variáveis importantes,

mesmo quando estas não aparecem inicialmente com grande relevância nos modelos tradicionais, proporcionando insights valiosos sobre padrões e tendências ocultas nos dados de riscos cibernéticos (Gareth et al., 2013).

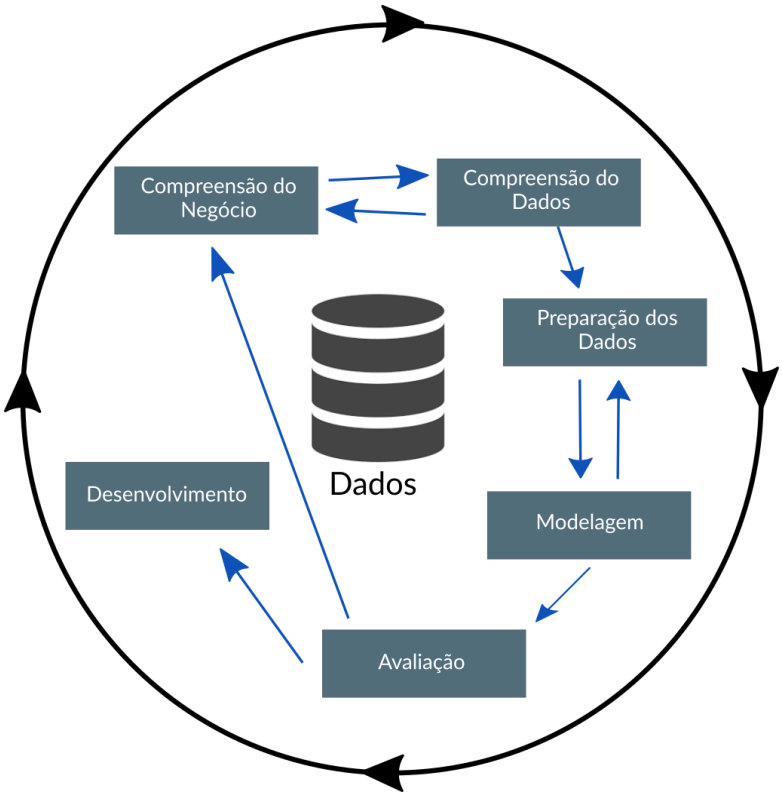
c) Gerenciamento de Riscos Cibernéticos

O gerenciamento de riscos, especialmente em cibersegurança, é um tema amplamente debatido tanto no meio acadêmico quanto no setor corporativo. Essa atenção reflete-se na variedade de metodologias e frameworks já consolidados, como os modelos tradicionais de gestão de projetos (HERMES, NBC, PRINCE2, PMBOK), padrões específicos como o Management of Risk (M_o_R), o Standard for Risk (PMI, 2019), o The Orange Book (2003), e os controles internos organizacionais (COSO, 2007). Além disso, as normas ISO/IEC 27001, 27002, 27032 e ISO 31000 são amplamente utilizadas para guiar práticas efetivas de segurança da informação. No contexto específico da segurança cibernética, modelos como o Cyber Security Body of Knowledge (CyBOK) e os Critical Security Controls (CIS Controls) fornecem orientações práticas para enfrentar desafios de segurança emergentes. O BSI-Standard 100-1 complementa essas orientações detalhando procedimentos claros para a implementação de Sistemas de Gestão da Segurança da Informação (SGSI), alinhados às normas ISO mencionadas. Finalmente, frameworks como o NIST *Privacy Framework* e o NIST *Cybersecurity Framework* se destacam ao fornecer uma estrutura clara e sistemática para identificação, prevenção e resposta a riscos cibernéticos, contribuindo significativamente para melhorar as comunicações entre stakeholders e fortalecer as práticas gerais de segurança nas organizações.

3. Metodologia

Esta pesquisa utilizou a metodologia *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*), abordagem amplamente utilizada para projetos com foco em análise de dados. A *CRISP-DM* consiste em seis etapas principais: (1) entendimento do problema, (2) entendimento dos dados, (3) preparação dos dados, (4) modelagem, (5) avaliação e o (6) desenvolvimento. A Figura 1 apresenta a integração das seis etapas do modelo *CRISP-DM*.

Figura 1 - Ilustração gráfica da metodologia *CRISP-DM*



Fonte: Estatidados, 2024

As etapas são iterativas permitindo que o processo volte a etapas anteriores sempre que necessário, garantindo flexibilidade e adaptabilidade ao longo do projeto. No contexto deste estudo, a metodologia foi adaptada para organizar a coleta, o pré-processamento, a modelagem e a análise dos dados de risco das empresas avaliadas, visando gerar *insights* para a GRCb. Estudos anteriores demonstram a eficácia do *CRISP-DM* em projetos similares de análise de dados que envolvem riscos justificando sua aplicação neste contexto (Rawat, 2023).

3.1 Etapas da Metodologia

3.1.1 Compreensão do negócio

O primeiro passo deste estudo foi definir como diferentes características de risco afetam a probabilidade de ocorrência de eventos em diferentes níveis de gravidade. Buscou-se entender os objetivos do negócio e transformá-los em problemas específicos de análise de dados. Tendo como referência para a pesquisa o modelo do COSO (2007) e a proposta de Oliva (2016), o objetivo principal foi entender a influência dessas características nos riscos e gerar *insights* estratégicos que possam apoiar a GR nas empresas em que o modelo de pesquisa foi aplicado.

3.1.2 Compreensão dos Dados

Nesta etapa, foram coletados dados no período de março até julho de 2024 em três empresas distintas, cada uma fornecendo informações detalhadas sobre os agentes de risco, a relação deles com o negócio, os riscos cibernéticos (RC) a eles associados, probabilidade de ocorrência, impacto ao negócio, características e respostas a serem adotadas.

Tabela 1 - Exemplo de dados coletados das empresas

Agente	**Risco Associado ao Agente**	**Tempo de Percepção**	**Pode ser Eliminado**	**Amplitude do Impacto**	**Possibilidade de Ocorrer**
Fornecedor de Software	Sistemas Vulneráveis	Longo Prazo	Não	Nacional	Baixo
Fornecedor de Hardware	Defeitos no Dispositivo	Instantâneo	Não	Regional	Médio
Grupos de Sabotagem	Vazamento de dados	Instantâneo	Não	Internacional	Baixo
Engenharia Social	Golpes	Instantâneo	Sim	Nacional	Muito Alto
Hacker	Ataque Cibernético	Uso no dia a dia	Não	Local	Muito Alto

Fonte: Autor

As avaliações em relação os riscos identificados consideraram aspectos como duração, possibilidade de ser eliminado, geração de resíduos pela solução adotada, necessidade de apoio para resposta, impacto, área mais afetada, possibilidade de ocorrência, tempo de percepção caso o risco ocorra, resposta clássica (escalar, evitar, aceitar, transferir e mitigar) adotada, existência de *SLA* acima de 99% e o tempo de percepção da empresa (semanas, dias, minutos, segundos).

Durante o entendimento dos dados, foram realizadas análises iniciais para identificar inconsistências, avaliar a qualidade dos dados e entender a representatividade das informações disponíveis. Esta etapa ajudou a garantir que os dados coletados fossem apropriados e suficientemente completos para alcançar os objetivos da pesquisa.

3.1.3 Preparação dos Dados

Para garantir a qualidade da análise, os dados coletados foram submetidos a um pré-processamento abrangente. Esta fase considera:

a. **Remoção de Colunas Irrelevantes** - Variáveis como identificadores internos e outras informações específicas de cada empresa foram removidas, pois não contribuem diretamente para a análise de risco;
b. **Codificação de Variáveis Categóricas** - Algumas variáveis eram categóricas (por exemplo, "Pode Ser Eliminado" com valores "Sim" ou "Não"). Para permitir a aplicação de técnicas estatísticas e modelos de aprendizado de máquina, essas variáveis foram transformadas em valores numéricos utilizando o método de *Label Encoding*, o que assegura que os modelos possam lidar adequadamente com essas informações. O *Label Encoding* transforma os valores categóricos em números inteiros, facilitando a aplicação dos modelos, mas sem atribuir significados hierárquicos aos valores;
c. **Normalização dos Nomes das Colunas** - As variáveis foram padronizadas em português, garantindo consistência na interpretação dos resultados e facilitando o uso dos dados em diferentes etapas da análise.

3.1.4 Modelagem

A etapa de modelagem incluiu a aplicação de duas abordagens principais para análise dos dados que são:

- **Regressão Logística Ordinal** - Utilizada para avaliar como cada característica contribui para a probabilidade de um risco ocorrer em diferentes níveis de gravidade. Esse modelo é adequado para variáveis dependentes ordinais e forneceu uma análise detalhada sobre como cada característica influencia os riscos. A RLO ajusta uma função logística cumulativa para prever a probabilidade de ocorrência de cada nível de risco, sendo especialmente útil para variáveis categóricas ordenadas, como "Muito Baixo", "Baixo", "Médio", "Alto" e "Muito Alto" conforme a fórmula apresenta a seguir.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Onde $(\sigma(z))$ é a função que transforma o valor de (z) em uma probabilidade entre 0 e 1, e, após essa transformação, (σ) passa a representar uma razão de chances (*odds*) de um evento

ocorrer em comparação com a sua não ocorrência, (z) é a combinação linear das variáveis preditora, dada por:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

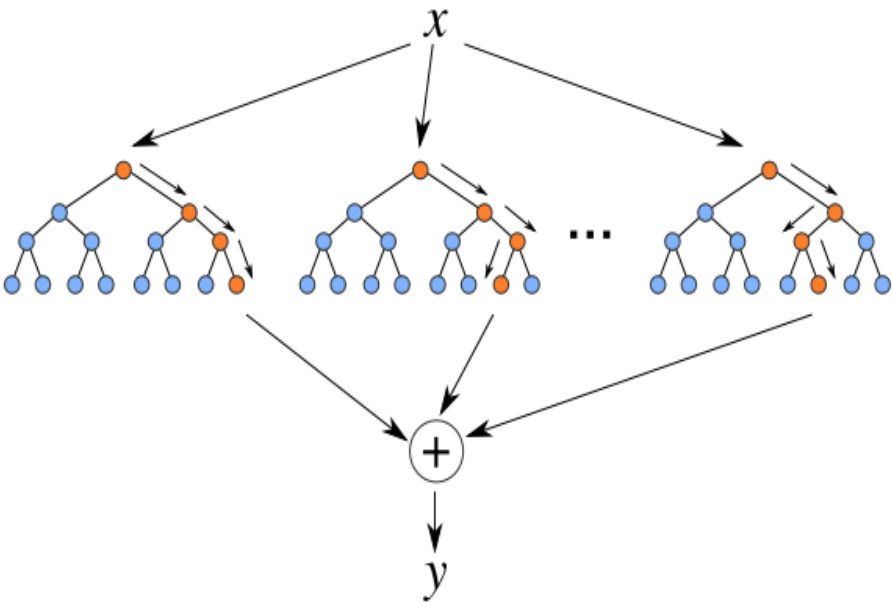
Nesta fórmula:

- $(\beta_0, \beta_1, \ldots, \beta_n)$ são os coeficientes (influência das variáveis);
- $(x_1, x_2, \ldots, x_n)$ representam as características de risco;
- (e) é a base do logaritmo natural, usada para transformar o escore linear (z) em uma probabilidade compreensível entre 0 e 1, facilitando a interpretação dos resultados.

Em resumo, a função logística transforma o escore linear (z) em uma probabilidade de ocorrência de um risco em determinado nível, facilitando a interpretação dos resultados da análise. No contexto deste estudo, a função logística calcula a probabilidade cumulativa de que um risco esteja em um nível específico. Cada coeficiente presente na fórmula reflete a influência de uma característica de risco, como "Impacto Financeiro" ou "Tempo de Percepção", sobre a gravidade do risco. Valores positivos dos coeficientes indicam que o aumento dessa característica está associado a uma maior probabilidade do risco se manifestar em níveis mais graves, enquanto valores negativos indicam o contrário, diminuindo essa probabilidade. Assim, ao interpretar os coeficientes da regressão, podemos compreender diretamente como cada característica afeta a gravidade dos riscos analisados nas empresas pesquisadas.

- **Random *Forest*** - O *Random Forest* é um modelo de aprendizado de máquina que combina várias árvores de decisão para realizar previsões mais robustas e precisas. Cada árvore é treinada com uma amostra aleatória dos dados e suas previsões são combinadas para gerar um resultado. Essa abordagem permite capturar a importância relativa de cada variável no modelo, destacando quais fatores têm maior influência na previsão dos riscos. A Figura 2 apresenta uma árvore de decisão para AM.

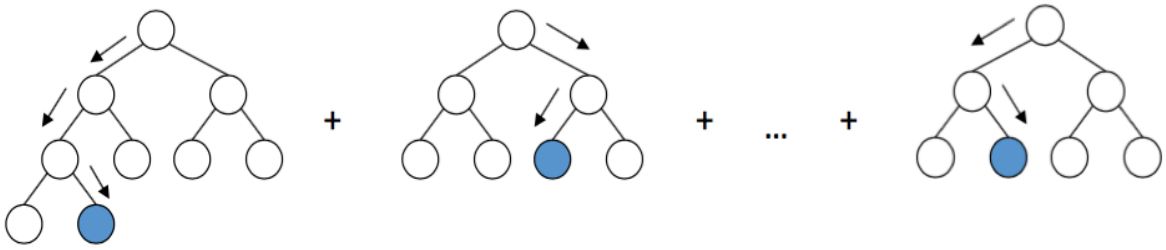
Figura 2 - Modelo de aprendizado de máquina *Random Forest*



Fonte: *Decision Tree e Random Forest*, 2024.

- ***Gradient Boosting*** - Assim como o *Random Forest*, o *Gradient Boosting* também utiliza árvores de decisão, mas adota uma abordagem diferente para construir suas previsões. Enquanto o *Random Forest* treina múltiplas árvores de forma independente e combina seus resultados, o *Gradient Boosting* constrói as árvores de maneira sequencial. A cada etapa, uma nova árvore é adicionada para corrigir os erros das previsões anteriores, permitindo que o modelo capture padrões mais profundos e complexos. Essa técnica de aprendizado iterativo é ideal para identificar interações sutis entre as variáveis. A Figura 3 apresenta um exemplo de árvore de decisão do tipo *Gradient Boosting*.

Figura 3 - Modelo de aprendizado de máquina *Gradient Boosting*



Fonte: *Gradient Boosting explained*, 2024.

3.1.5 Avaliação

A avaliação dos modelos foi realizada utilizando métricas e ferramentas específicas para cada abordagem, garantindo uma análise detalhada dos resultados. Para a RLO, utilizamos gráficos

de coeficiente (*Coefficient plot*) para ilustrar a influência de cada característica de risco nos diferentes níveis de ocorrência. Para os modelos de aprendizado de máquina, *Random Forest* e Gradient Boosting, a avaliação se baseou na importância das variáveis, destacada através de gráficos de barras que mostram a contribuição de cada característica para a previsão dos riscos.

Tabela 2 - Método de avaliação dos resultados para cada modelo

Modelo	**Método utilizado para avaliação**
Regressão Logística Ordinal	Gráfico de Coeficiente (*Coefficient Plot*)
Random Forest	Gráfico de Importância das Variáveis (*Feature Importance Plot*)
Gradient Boosting	Gráfico de Importância das Variáveis (*Feature Importance Plot*)

Fonte: Autor

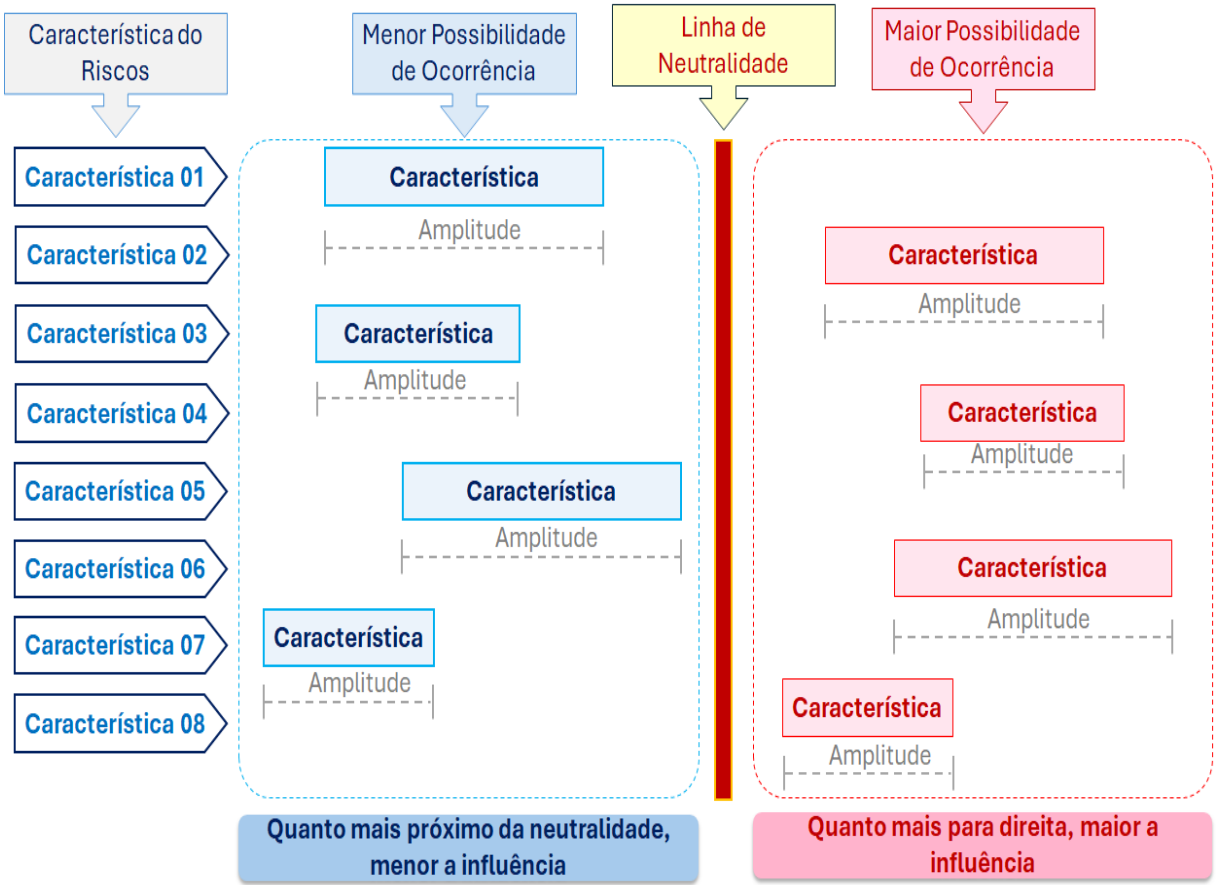
Cada modelo foi analisado individualmente em termos de sua capacidade de oferecer *insights* práticos e úteis para a gestão de riscos, destacando as variáveis com maior impacto.

3.1.6 Comunicação dos Resultados

Os resultados foram organizados e comunicados de maneira clara e visualmente acessível, utilizando uma variedade de gráficos para facilitar a compreensão dos achados do estudo. Como foi dito anteriormente, para a RLO, foram apresentados gráficos de Coeficiente (*Coefficient plot*), que mostram a influência positiva ou negativa de cada característica no risco.

Conforme dito anteriormente, os riscos identificados podem receber diferentes características que a GR do negócio pode atribuir. Nosso modelo considerou atributos como duração, possibilidade de ser eliminado, geração de resíduos pela solução adotada, necessidade de apoio para resposta, impacto, área mais afetada, possibilidade de ocorrência, exposição do negócio, tempo de percepção, impacto financeiro, resposta clássica (escalar, evitar, aceitar, transferir e mitigar), existência de *SLA* acima de 99% e o tempo de percepção (semanas, dias, minutos, segundos). Na Figura 6 temos, ao lado esquerdo, as características dos riscos.

Figura 4 - Gráfico de Coeficiente

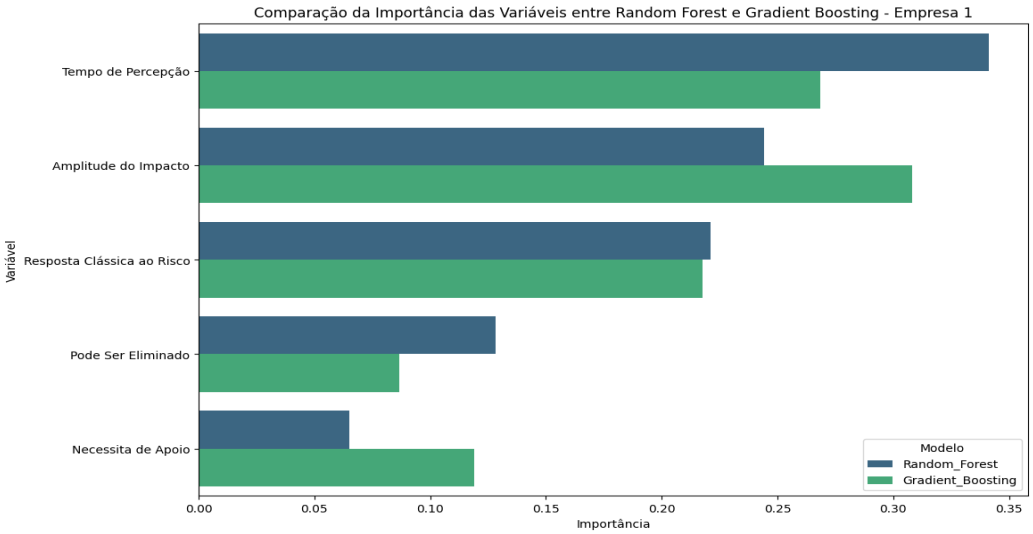


Fonte: Autor.

Quanto mais para a esquerda estiver uma característica, menor será a sua influência sobre a possibilidade de o risco ocorrer. Quanto mais para a direita, maior será essa possibilidade de o risco vir a se materializar. Ao centro, temos a linha de neutralidade.

Para os modelos *Random Forest* e *Gradient Boosting*, foram apresentados gráficos (Figura 5) de importância das variáveis (*Feature Importance Plot)* que mostram a importância das variáveis utilizadas pelo modelo. Além disso, gráficos comparativos foram utilizados para mostrar as diferenças e semelhanças entre os modelos, permitindo uma visão mais integrada dos fatores de risco que influenciam de maneira consistente o risco de ocorrência.

Figura 5 - Exemplo Gráfico de Importância das Variáveis (*Feature Importance* Plot)



Fonte: Autor

A metodologia *CRISP-DM* por ser interativa, permitindo revisões constantes em cada uma das etapas conforme novas informações são obtidas. Essa característica garantiu que o processo de análise de risco fosse contínuo e flexível, atendendo aos requisitos do projeto e produzindo resultados significativos e práticos para as empresas participantes.

4. Resultados

Nesta seção, apresentamos os resultados obtidos a partir da análise dos dados coletados. Utilizamos a RLO e os modelos de aprendizado de máquina (*Random Forest* e *Gradient Boosting*) para explorar a influência das características de risco na "Possibilidade de Ocorrência" dos eventos.

4.1 Análise dos Resultados com Regressão Logística Ordinal

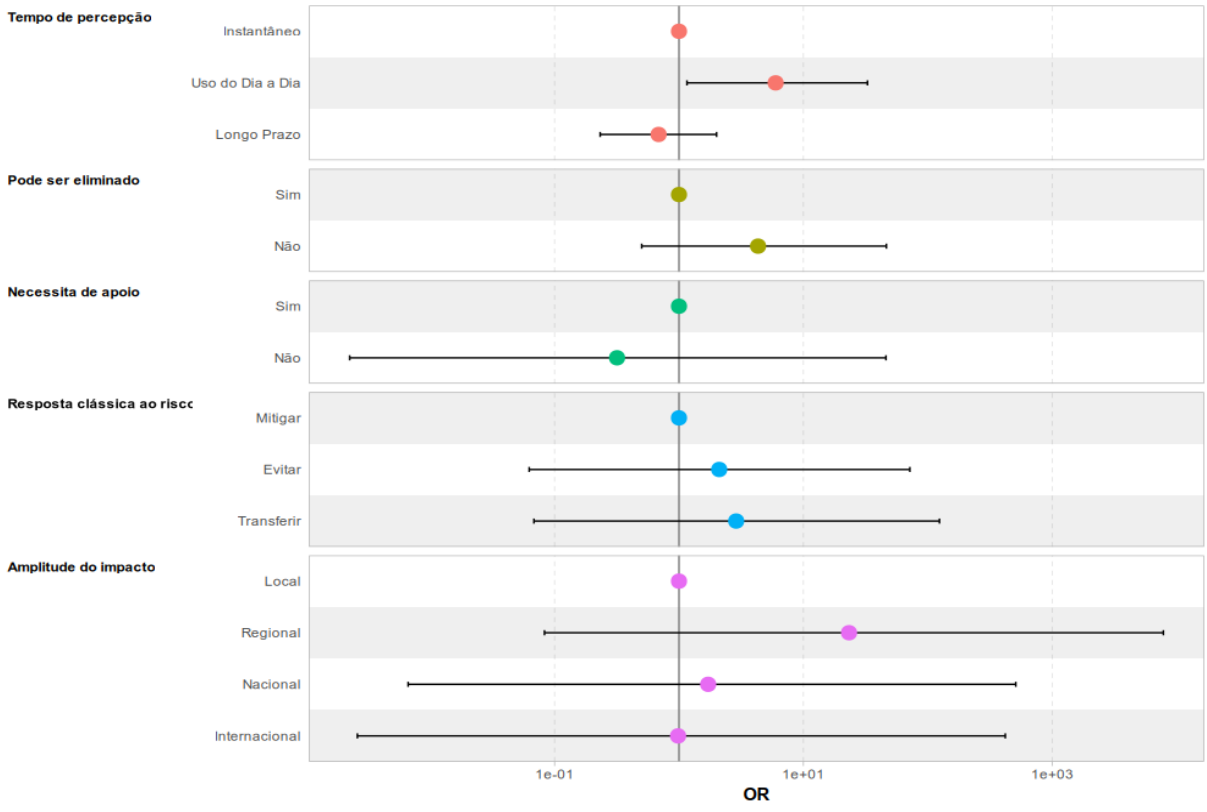
As simulações desta pesquisa utilizaram o software R que é uma linguagem de programação multi-paradigma orientada a objetos. R é voltada à manipulação, análise, modelagem e visualização de dados. Os modelos de aprendizado de máquina *Random Forest* e *Gradient Boosting* foram implementados usando a linguagem de programação *Python* que oferece recursos de desenvolvimento de alto nível, interpretada por script e imperativa. A análise foi conduzida usando a RLO, o que nos permitiu identificar como cada variável preditora impacta a gravidade dos riscos. Os resultados são apresentados através do Gráfico de Coeficiente que mostra os *odds ratios*, indicando se uma característica está associada a um aumento ou diminuição na probabilidade de ocorrer um risco mais severo. O modelo foi aplicado a três

casos reais de avaliação tendo os entrevistados disponibilizado informações contidas em planilhas anonimizadas compostas por abas temáticas e de variáveis quantitativas e qualitativas relacionadas à identificação, avaliação e classificação de riscos. Os contextos organizacionais representados nesses conjuntos de dados serão apresentados a seguir.

4.2. Empresa do Segmento de Mineração

A empresa estudada atua na mineração e metalurgia de estanho e minerais industriais. Com atuação global, comercializa seus produtos para diversos países como China, França, Noruega entre outros. Sua produção de metais gira em torno de 2000 toneladas ao mês. Sua operação é bastante complexa envolvendo e envolve todo um controle relacionado a segurança das pessoas e respeito ao meio ambiente. Neste contexto, participam do processo mais de 1400 colaboradores da empresa e 1200 terceirizados.

Figura 6 - Gráfico de Coeficiente gerado pela RLO – Caso 1



Fonte: Autor

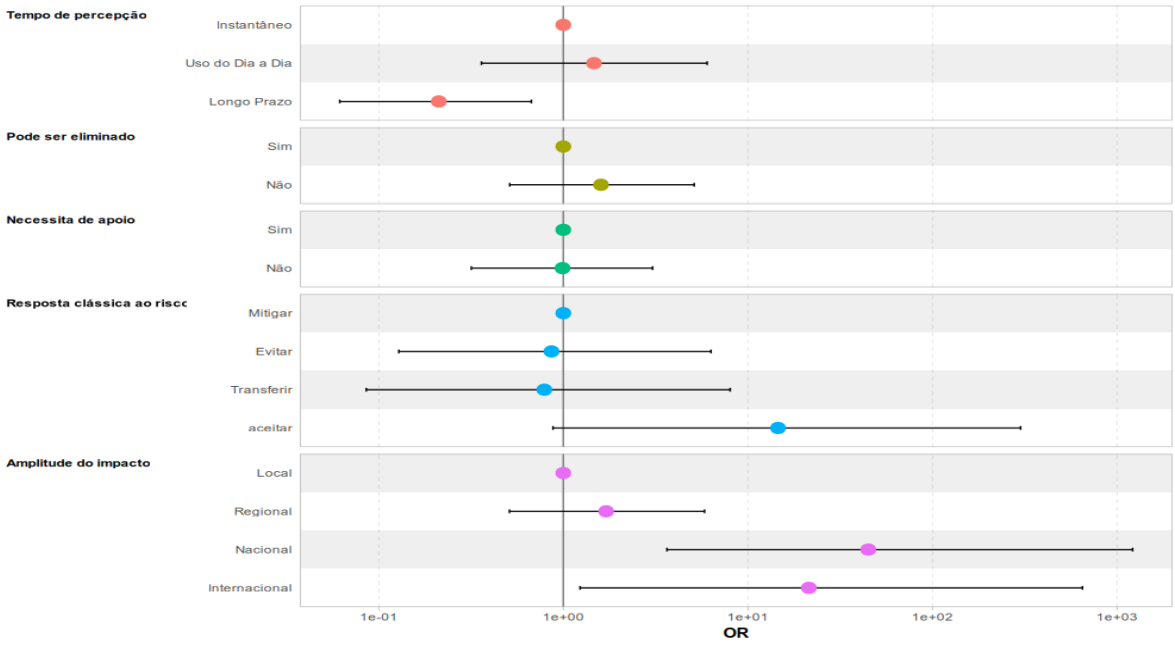
O gráfico gerado para a empresa 1 mostra tanto semelhanças quanto diferenças na priorização das variáveis. Em ambos os modelos, “Tempo de Percepção”, “Amplitude do Impacto” e

"Resposta Clássica ao Risco" aparecem como as três variáveis mais importantes, embora em ordens diferentes, indicando consenso sobre sua relevância geral. Por outro lado, "Necessita de Apoio" e "Pode Ser Eliminado" ocupam as duas últimas posições em ambos os modelos, reforçando que essas variáveis têm menor impacto nas decisões. Apesar das diferenças na ordem de prioridade, é possível perceber que os modelos possuem uma constância sobre quais variáveis são mais influentes e quais são menos significativas na análise.

4.3 Empresa do Segmento Público

A segunda empresa em que o modelo foi aplicado é uma empresa de economia mista com capital aberto. Ela está presente em mais de 150 municípios do estado do Ceará, beneficiando cerca de 5 milhões de cearenses. Por ser uma empresa pública e prestando um benefício exclusivo a população, ela não possui concorrentes no seu segmento, porém, para atender aos usuários e manter a performance dos seus serviços, ela conta com uma complexa organização dos recursos de TI (software, hardware, processo, pessoas, regulamentos) que garantem o desempenho dos seus serviços.

Figura 7 - Gráfico de Coeficiente pela RLO – Caso 2



Fonte: Autor

No gráfico de coeficiente da empresa 2 podemos observar que a possibilidade de ocorrência do risco é menor entre os agentes com tempo de percepção a longo prazo, indicando uma redução

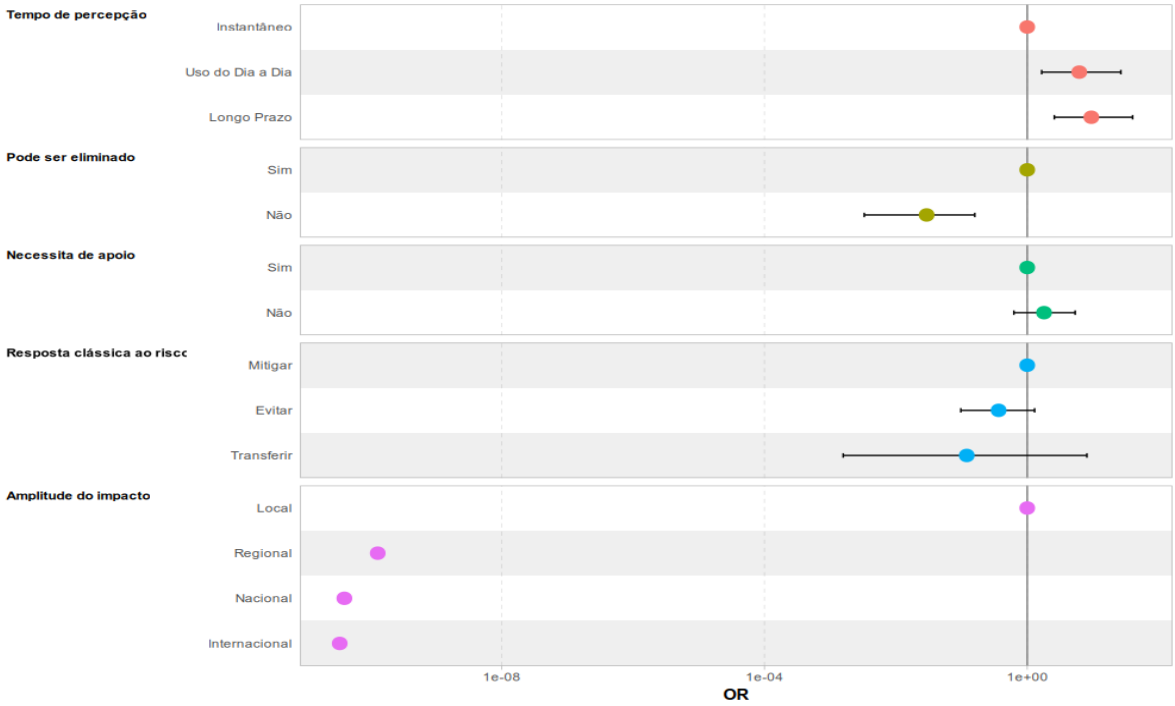
na chance de ocorrência desses riscos. Por outro lado, a amplitude do impacto, quando classificada como "Nacional" ou "Internacional", demonstra uma maior probabilidade de ocorrência do risco, sugerindo que esses contextos ampliados possuem influência significativa.

Para essa empresa, riscos de longo prazo tendem a ser menos frequentes, enquanto aqueles com amplitude de impacto em nível nacional ou internacional apresentam maior chance de ocorrer.

4.4 Empresa do Segmento de Segurança da Informação

A terceira empresa em que o modelo foi aplicado é especializada na oferta de serviços em segurança cibernética para organizações. Seu modelo de negócio considera um portfólio completo de serviços que vão desde consultoria em SI, treinamento de colaboradores, diagnósticos e auditorias. Atuando desde o ano de 2010, ela atende clientes locais e na região metropolitana da sua cidade. A empresa comercializa soluções de hardware e software incluindo serviços de instalação, configuração e manutenção.

Figura 8 - Gráfico de Coeficiente gerado pela RLO – Caso 3



Fonte: Autor

No gráfico gerado para a empresa 3, observa-se que a possibilidade de ocorrência do risco é maior tanto para os agentes com tempo de percepção no uso do dia a dia quanto para aqueles

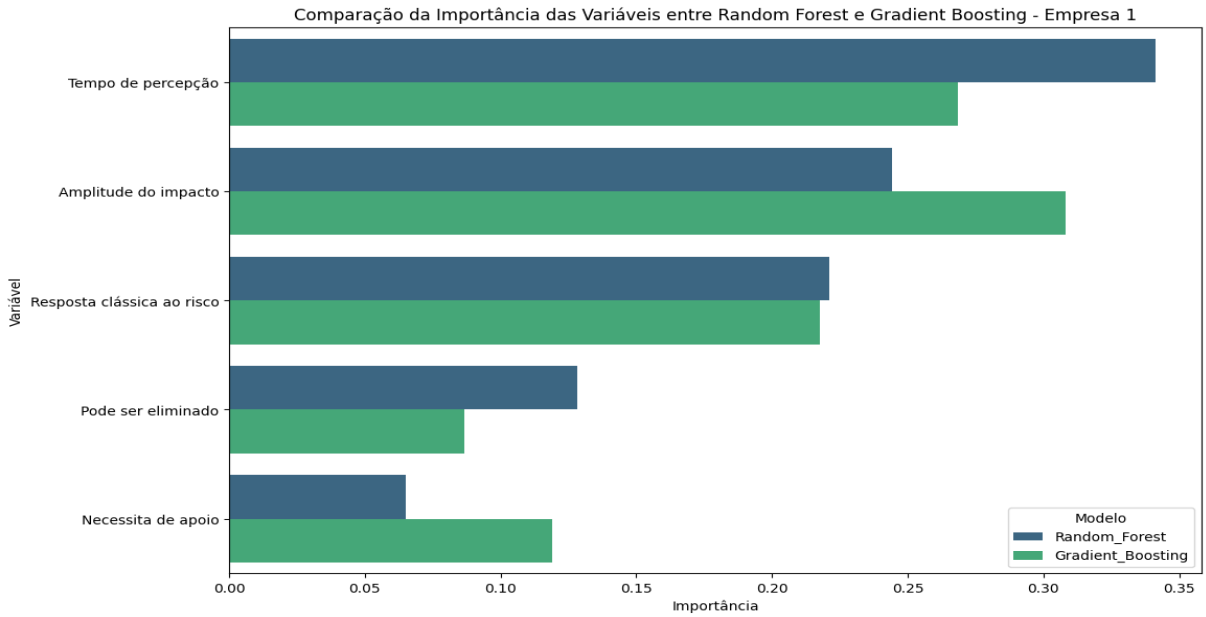
com percepção a longo prazo, indicando que atividades cotidianas e percepções mais prolongadas estão associadas a um aumento na probabilidade de ocorrência do risco. Por outro lado, o fato de o risco não poder ser eliminado reduz significativamente a chance de sua ocorrência. Em resumo, para essa empresa, riscos percebidos no dia a dia ou a longo prazo apresentam maior chance de ocorrer, enquanto a impossibilidade de eliminar o risco atua como um fator redutor de sua probabilidade.

4.5 Resultados dos Modelos de Aprendizado de Máquina

Para uma análise mais detalhada, aplicamos modelos de aprendizado de máquina como *Random Forest* e *Gradient Boosting*, com o objetivo de identificar padrões mais complexos e observar a importância relativa de cada característica na previsão dos riscos.

Os gráficos de importância das variáveis gerados pelos dois modelos foram sobrepostos para fins de comparação. Essa comparação nos permite observar semelhanças e divergências entre as abordagens, destacando as características que influenciam de forma consistente a probabilidade de ocorrência dos riscos.

Figura 9 - Gráfico de Importância de Variáveis para a – Caso 1

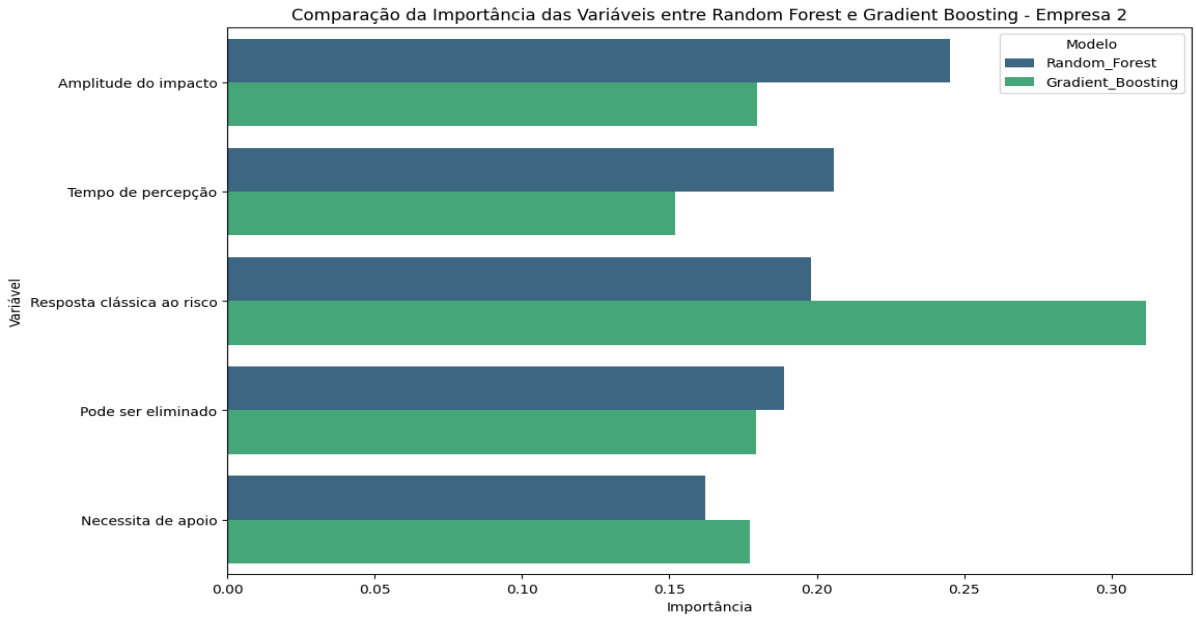


Fonte: Autor

O gráfico gerado para a empresa 1 mostra tanto semelhanças quanto diferenças na priorização das variáveis. Em ambos os modelos, “Tempo de Percepção”, “Amplitude do Impacto” e

"Resposta Clássica ao Risco" aparecem como as três variáveis mais importantes, embora em ordens diferentes, indicando consenso sobre sua relevância geral. Por outro lado, "Necessita de Apoio" e "Pode Ser Eliminado" ocupam as duas últimas posições em ambos os modelos, reforçando que essas variáveis têm menor impacto nas decisões. Apesar das diferenças na ordem de prioridade, é possível perceber que os modelos possuem uma constância sobre quais variáveis são mais influentes e quais são menos significativas na análise.

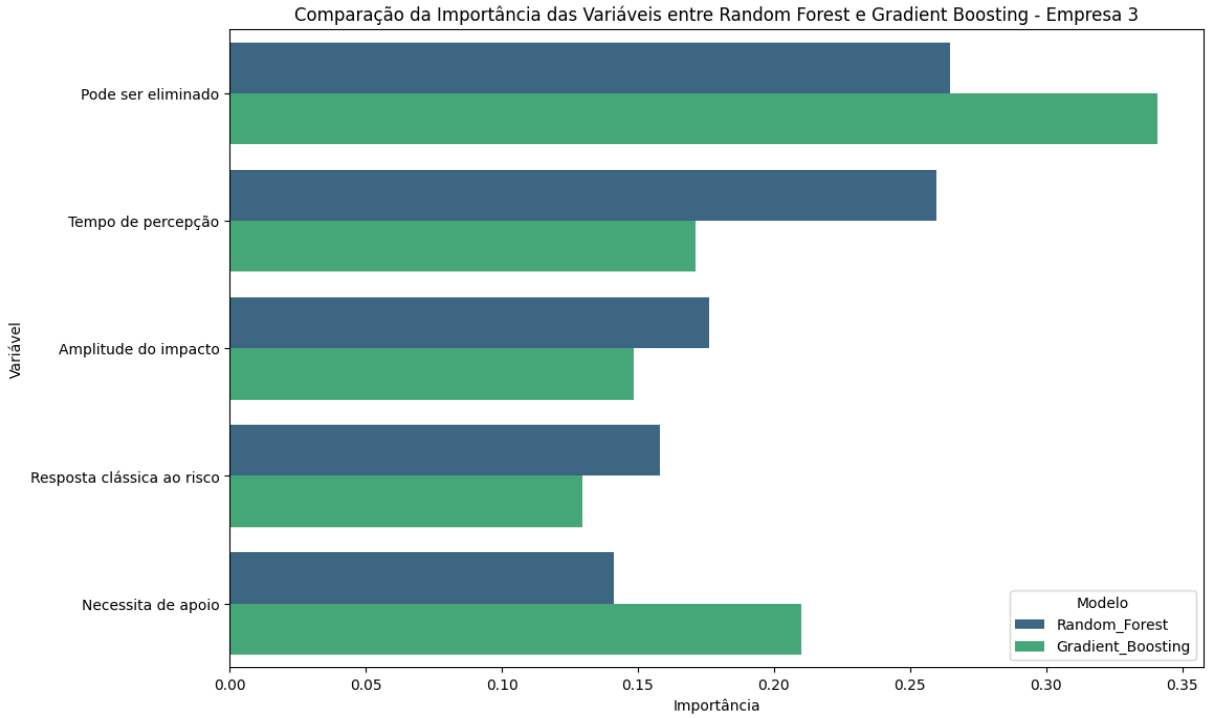
Figura 10 - Gráfico de Importância de Variáveis para a – Caso 2



Fonte: Autor

Para a empresa 2, o gráfico destaca diferenças significativas na priorização. No RF, "Amplitude do Impacto" é a variável mais importante, seguida pelo "Tempo de Percepção" e pela "Resposta Clássica ao Risco", enquanto "Pode Ser Eliminado" e "Necessita de Apoio" têm menor relevância. Já no GB, a "Resposta Clássica ao Risco" assume a liderança com a "Amplitude do Impacto" em segundo lugar. Além disso, "Pode Ser Eliminado" e "Necessita de Apoio" aparecem como intermediárias, enquanto o "Tempo de Percepção" tem a menor relevância. Apesar das diferenças, ambos os modelos destacam a relevância da "Amplitude do Impacto" e da "Resposta Clássica ao Risco", ainda que em ordens distintas, enquanto concordam que "Necessita de Apoio" e "Pode Ser Eliminado" possuem papéis menos determinantes.

Figura 11 - Gráfico de Importância de Variáveis para a – Caso 3



Fonte: Autor

Na Empresa 3, “Pode Ser Eliminado” é a variável mais importante tanto para o RF quanto para o GB. No entanto, os modelos divergem na ordem das demais variáveis. No Random Forest, o “Tempo de Percepção” ocupa o segundo lugar, seguido por “Amplitude do Impacto”, enquanto “Resposta Clássica ao Risco” e “Necessita de Apoio” têm menor relevância. Já no Gradient Boosting, “Necessita de Apoio” aparece em segundo, com “Tempo de Percepção” e “Amplitude do Impacto” em posições intermediárias, e “Resposta Clássica ao Risco” sendo a menos importante. Apesar do consenso em relação à liderança de “Pode Ser Eliminado”, os modelos apresentam diferenças significativas na classificação das demais variáveis, refletindo abordagens distintas na análise dos dados.

5. Conclusões e Limitações

As considerações finais iniciam-se com a apresentação do (1) atendimento ao objetivo da pesquisa e aos aspectos metodológicos, seguidas das (2) contribuições teóricas, (3) implicações gerenciais e, por fim, das (4) limitações da pesquisa e sugestões para futuros estudos.

(1) Atendimento ao Objetivo da Pesquisa e aos Aspectos Metodológicos

O objetivo principal desta pesquisa foi analisar a relação entre diversas características de risco

em empresas e a probabilidade de ocorrência desses riscos, utilizando diferentes abordagens analíticas (RLO, RF e GB). O estudo seguiu o modelo CRISP-DM, estruturado em etapas definidas.

Inicialmente, foram identificadas variáveis relevantes ao contexto dos riscos cibernéticos, tais como "Tempo de Percepção", "Pode Ser Eliminado", "Amplitude do Impacto" e "Resposta Clássica ao Risco". Essas variáveis foram escolhidas com base em sua relevância teórica e prática. A preparação dos dados envolveu tratamentos específicos para garantir a qualidade e a consistência necessária à análise. A aplicação inicial da RLO revelou a influência direta das variáveis, destacando especialmente a relevância do "Tempo de Percepção", "Amplitude do Impacto" e da possibilidade de eliminação do risco.

Na etapa seguinte, técnicas de ML (RF e GB) forneceram análises complementares, capturando relações mais complexas e permitindo uma visão mais profunda sobre os fatores determinantes na ocorrência dos riscos. Assim, a pesquisa respondeu à questão formulada inicialmente: "Quais características específicas dos riscos cibernéticos exercem maior influência sobre sua probabilidade de ocorrência nas organizações?", indicando quais fatores devem receber maior atenção.

(2) Contribuições Teóricas Este estudo avança significativamente o debate teórico sobre gerenciamento de riscos cibernéticos ao integrar métodos estatísticos tradicionais (RLO) com técnicas modernas de aprendizado de máquina (RF e GB). Essa combinação proporciona uma perspectiva que complementa abordagens existentes, demonstrando a relevância da análise híbrida para a compreensão profunda dos fatores críticos. O estudo reforça e amplia a discussão sobre a importância relativa do tempo de percepção, amplitude do impacto e possibilidade de eliminação de riscos, indicando como esses fatores devem ser considerados juntos e não isoladamente.

(3) Implicações Gerenciais

Os resultados desta pesquisa possuem implicações diretas e relevantes para a gestão dos riscos cibernéticos. Gestores podem utilizar os *insights* obtidos para priorizar recursos e esforços preventivos, especialmente ao reconhecer a importância crucial do tempo de percepção e da antecipação aos riscos. Além disso, a compreensão detalhada da amplitude do impacto

potencial e das possibilidades práticas de eliminação dos riscos cibernéticos oferece subsídios ao negócio para uma tomada de decisão mais assertiva e eficaz, resultando em uma gestão de riscos mais estratégica e menos reativa para a organização. Esses resultados podem apoiar diretamente as empresas na elaboração de políticas internas mais robustas e na implementação de mecanismos específicos para aumentar sua resiliência cibernética.

(4) Limitações e Sugestões para Futuros Estudos

Esta pesquisa possui limitações relacionadas principalmente ao contexto específico de aplicação dos modelos, restrita a três empresas distintas, o que limita a generalização dos achados. Outro aspecto limitante foi a seleção das variáveis, condicionada pela disponibilidade dos dados, potencialmente deixando de fora outros fatores relevantes.

Para futuras pesquisas, recomenda-se ampliar significativamente a amostra, incluindo empresas de diferentes regiões, setores e perfis organizacionais, visando maior representatividade. Sugere-se também o uso de técnicas analíticas adicionais, como redes neurais, capazes de identificar interações ainda mais complexas e dinâmicas entre variáveis. Além disso, seria valiosa a realização de estudos longitudinais para avaliar como essas relações evoluem ao longo do tempo, permitindo uma compreensão mais robusta e completa do comportamento dos riscos cibernéticos nas organizações.

Referência

ABNT NBR ISO/IEC 27002. **Norma Brasileira ABNT NBR ISO/IEC 27002 – Segurança da Informação, Segurança Cibernética e Proteção à Privacidade – Controles de Segurança da Informação**. Associação Brasileira de Normas Técnicas. Rio de Janeiro: ABNT, 2022.

AGRESTI, A. **An Introduction to Categorical Data Analysis**. Wiley, 2018.

AXELOS. **Management of Risk: Guidance for Practitioners 3rd Edition**. (ISBN 9780113312740). Published by TSO (The Stationery Office), 2010.

BREIMAN, L. **Random Forests**. *Machine Learning*, *45*(1), 2001, 5-32.

COSO - **Gerenciamento de Riscos Corporativos - Estrutura Integrada**. *Price Waterhouse Coopers - PwC*, 2007.

DECISION TREE E RANDOM FOREST. Disponível em: http://carlosbaia.com/2016/12/24/decision-tree-e-random-forest/. Acessado em: 18 nov 2024.

DEMIRKAN, Sebahattin; DEMIRKAN, Irem; MCKEE, Andrew. **Blockchain technology in the future of business cyber security and accounting**. Journal of Management Analytics, v. 7, n. 2, p. 189-208, 2020.

ESTATIDADOS. **CRISP-DM - Processo Padrão Inter-Indústrias para Mineração de Dados**. Disponível em: https://estatidados.com.br/crisp-dm-processo-padrao-inter-industrias-para-mineracao-de-dados/. Acessado em: 18 nov 2024.

FRIEDMAN, J. H. **Greedy Function Approximation: A Gradient Boosting Machine**. Annals of Statistics, 29(5), 2001, 1189-1232.

FRIEDMAN, J., Hastie, T., & Tibshirani, R. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2017.

GARETH, J., Witten, D., Hastie, T., & Tibshirani, R. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013.

GRADIENT BOOSTING EXPLAINED. Disponível em: https://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html. Acessado em: 18 nov 2024.

HOSMER, D. W., Lemeshow, S., & Sturdivant, R. X. ***Applied Logistic Regression***. John Wiley & Sons, 2013.

KLEINBAUM, D. G., & Klein, M. **Logistic Regression: A Self-Learning Text**. Springer Science & Business Media, 2010.

LONG, J. S., FREESE, J.***Regression Models for Categorical Dependent Variables Using Stata***. Stata Press, 2014.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. MIT Press, 2012.

NBC**. IPMA - Referencial Brasileiro de Competências**. Tradução Raphael de Oliveira Albergarias Lopes. 1.ed. - Escola Politécnica/UFRJ, Rio de Janeiro: 2012.

OLIVA, F.L. **Knowledge management barriers, practices and maturity model**. Journal of Knowledge Management, v. 18, n. 6, p. 1053–1074, 2014.

OLIVA, F. L.; SOBRAL, M. C.; DAMASCENO, F.; TEIXEIRA, H. J.; GRISI, C. C. H.; FISCHMANN, A. A.; SANTOS, S. A. dos. **Risks and strategies in a Brazilian innovation – flexfuel technology**. Journal of Manufacturing Technology Management, v. 25, n. 6, p. 916–930, 2014.

OLIVA, F. L. ***A Maturity Model for Enterprise Risk Management***. International Journal of Production Economics, v. 173, p. 66-79, 2016.

OLIVA, F.L.; KOTABE, M. **Barriers, practices, methods and knowledge management tools in startups.** Journal of Knowledge Management, v. 23, n. 9, p. 1838–1856, 2019.

OLIVA, F. L.; SEMENSATO, B. I.; PRIOSTE, D. B.; WINANDY, E. J. L.; BUTION, J. L.; COUTO, M. H. G.; BOTTACIN, M. A.; MAC LENNAN, M. L. F.; TEBERGA, P. M. F.; SANTOS, R. F.; da SILVA, S. F.; MASSAINI; SINGH, S. K. **Innovation in the main Brazilian business sectors: characteristics, types and comparison of innovation.** Journal of Knowledge Management, v. 23, n.1, p. 135–175, 2019.

OLIVA, F.L.; COUTO, M.H.G.; SANTOS, R.F.; BRESCIANI, S. **The integration between knowledge management and dynamic capabilities in agile organizations**. Management Decision, v. 57, n. 8, p. 1960-1979, 2019.

OLIVA, F. L.; TEBERGA, P. M. F.; TESTI, L. I. O.; KOTABE M.; GIUDICE, M. D.; KELLE, P.; CUNHA, M. P. **Risks critical success factors in the internationalization of born global startups of industry 4.0: A social, environmental, economic and institutional analysis**. Technological Forecasting and Social Change, v. 175, p. 121346, 2021.

OLIVA, F.L.; PAZA, A.C.T.; BUTION, J.L.; KOTABE, M.; KELLE, P.; VASCONCELLOS, E.P.G; de, GRISI, C.C. de H. e, *et al.* **A model to analyze the knowledge management risks in open innovation: proposition and application with the case of GOL Airlines**. Journal of Knowledge Management, v. 26, n. 3, p. 681–721, 2022.

OLIVA, F.L.; BUTION, J.L.; MOTTA, F.G.; FENNER, G.; RANDOLPH-SENG, B.; PAPA, M.; NAQSHBANDI, M.M. **Appetite for risk: theoretical framework and practical application in a technology-based environment.** Journal of Intellectual Capital, v. 26, n. 1, p. 71-103, 2025.

PANDEY, Shipra; SINGH, Rajesh Kumar; GUNASEKARAN, Angappa; KAUSHIK, Anjali**. Cyber security risks in globalized supply chains: conceptual framework**. Journal of Global Operations and Strategic Sourcing, 2020.

PMI. ***A Guide to the Project Management Body of Knowledge and the Standard for Project Management Seventh Edition***. PMI: Filadélfia, Pensilvânia - USA, 2021.

OGC - *Office of Government Commerce*. ***Managing Successful Projects with* PRINCE2®.** London: TSO (The Stationery Office), 2009.

OFCL - **HERMES - Conduite et Déroulement de Projets Dans le Domaine des Technologies de l'information et de la Communication (TIC)**. Unité de stratégie informatique de la Confédération USIC, CH-3003 Berne. OFCL, CH-3003 Berne, 2005.

RAWAT, K. **Applying CRISP-DM Methodology in Developing Machine Learning Model for Credit Risk Prediction. In: Arai, K. (eds) Intelligent Computing**. SAI 2023. Lecture Notes in Networks and Systems, vol 739. Springer, Cham. https://doi.org/10.1007/978-3-031-37963-5_37.

***The Cyber Security Body of Knowledge* - CyBOC**. University of Bristol, 2019. Disponível em: https://www.cybok.org/. Acessado em: 29 set. 2023.

The Orange Book Management of Risk. ***The Orange Book Management of Risk - Principles and Concepts***. Disponível em: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1154709/HMT_Orange_Book_May_2023.pdf. Acessado em: 24 set. 2023.

WILLIAMS, R. **Ordinal Regression Models: Problems, Solutions, and Application**. *Sociological Methods & Research*, 2020.

ZHANG, C., & MA, Y. **Ensemble Machine Learning: Methods and Applications**. Springer, 2012.

Agradecimentos:

O artigo é baseado em estudo apoiado por:

Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES

Faculdade de Economia, Administração, Contabilidade e Atuária - FEAUSP.

Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP.

Fundação Instituto de Administração - FIA.

Universidade de São Paulo – USP.

Universidade Federal do Ceará - UFC.