Instituto Superior
de Contabilidade
e Administração

Politécnico de Coimbra

COIMBRA BUSINESS SCHOOL
ISCAC.pt

Bárbara Sobral – 2024104488
Maria Ribeiro – 2024104121
Mariana Almeida 2024102807

Ataques Cibernéticos
Projeto Big Data 24/25

Coimbra, março de 2025

Instituto Superior
de Contabilidade
e Administração

Politécnico de Coimbra

iscac Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL ISCAC.pt

Ataques Cibernéticos – Projeto Big Data 24/25

Bárbara Xavier Sobral - 2024104488
Maria Margarido Ribeiro – 2024104121
Mariana Sofia Mendes Prata de Almeida - 2024102807

ISCAC | 2025

Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Bárbara Sobral – 2024104488
Maria Ribeiro – 2024104121
Mariana Almeida 2024102807

Ataques Cibernéticos
Projeto Big Data 24/25

Relatório de projeto submetido ao Instituto Superior de Contabilidade e Administração de Coimbra no âmbito da unidade curricular Big Data do Mestrado em Análise de Dados e Sistemas de Apoio à Decisão.

Coimbra, 20 março de 2025

TERMO DE RESPONSABILIDADE

Declaramos ser os autores deste trabalho, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atestamos ainda que todas as citações estão devidamente identificadas e que temos consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação do presente trabalho.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

RESUMO

Este projeto foca-se na análise e predição do volume de dados transmitidos durante ataques cibernéticos, utilizando tecnologias de *Big Data* e aprendizagem automática. Através do *Apache Spark* e da biblioteca *PySpark*, foi realizada a exploração, transformação e modelagem dos dados do conjunto UNSW-NB15. Diferentes modelos de regressão foram implementados e comparados para identificar a abordagem mais eficaz na estimativa da variável ***dbytes***. Os resultados obtidos fornecem *insights* valiosos para a cibersegurança, contribuindo para o desenvolvimento de sistemas mais eficientes na deteção e mitigação de tráfego malicioso.

Palavras-chave: Cibersegurança; *Big Data*; *Machine Learning*; Ataques Cibernéticos

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ABSTRACT

This project focuses on analysing and predicting the volume of data transmitted during cyberattacks using Big Data technologies and Machine Learning. Through Apache Spark and the PySpark library, data from the UNSW-NB15 dataset was explored, transformed, and modelled. Various regression models were implemented and compared to identify the most effective approach for estimating the ***dbytes*** *variable. The results provide valuable insights for cybersecurity, contributing to the development of more efficient systems for detecting and mitigating malicious traffic.*

Keywords: *Cybersecurity; Big Data; Machine Learning; Cyber Attacks*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ÍNDICE GERAL

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ÍNDICE DE TABELAS E FIGURAS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Lista de abreviaturas, acrónimos e siglas

GBT	*Gradient Boosting Trees*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
RFR	*Random Forest Regressor*
RL	Regressão Linear
RMSE	*Root Mean Squared Error*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

INTRODUÇÃO

A segurança cibernética é um dos desafios mais prementes da era digital, com o crescimento exponencial de ataques a redes e infraestruturas críticas. A capacidade de identificar, caracterizar e prever esses ataques é essencial para o desenvolvimento de estratégias de defesa eficazes. Neste contexto, a análise de grandes volumes de dados de tráfego de rede desempenha um papel fundamental na deteção de padrões maliciosos e na antecipação de comportamentos anómalos.

Este projeto utiliza o conjunto de dados UNSW-NB15, que contém registos de tráfego de rede, incluindo conexões normais e diferentes tipos de ataques cibernéticos. O objetivo principal é desenvolver modelos preditivos para estimar o volume de dados transmitidos durante ataques, representado pela variável ***dbytes***. Para isso, são aplicadas técnicas de *Big Data* e *Machine Learning* (ML), explorando e comparando diferentes algoritmos de regressão para identificar a abordagem mais eficaz.

Ao longo do projeto, os dados foram explorados e transformados para garantir a sua qualidade e adequação à modelagem preditiva. Foram implementados diferentes modelos de ML, incluindo Regressão Linear (RL), *Random Forest Regressor* (RFR) e *Gradient Boosting Trees* (GBT), com o objetivo de prever o volume de *bytes* transmitidos durante os ataques. A performance dos modelos foi avaliada com métricas padrão de regressão, permitindo identificar a abordagem mais eficaz para este problema específico. Com os *insights* obtidos, pretende-se contribuir para a melhoria de sistemas de deteção e mitigação de ataques cibernéticos, tornando-os mais eficientes na identificação de tráfego malicioso.

Este trabalho faz uso de diversas ferramentas, incluindo o *framework PySpark* para processamento distribuído de *Big Data*, a biblioteca *Spark MLlib* para a implementação dos modelos de ML, e Pandas para manipulação e análise dos dados. Além disso, foram utilizadas as bibliotecas *Matplotlib* e *Seaborn* para a visualização dos dados e dos resultados obtidos.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1 Escolha do Dataset e Definição do Problema

1.1 Definição do Problema

A crescente sofisticação e frequência dos ataques cibernéticos representam um grande desafio para a segurança de redes e sistemas informáticos. Entre as diversas abordagens para mitigar essas ameaças, a previsão do volume de tráfego malicioso surge como uma estratégia fundamental para a deteção precoce e resposta a incidentes. O objetivo deste projeto é prever a taxa de transferência de dados durante ataques cibernéticos, especificamente a quantidade de *bytes* recebidos pelo atacante (variável ***dbytes***), utilizando modelos de regressão.

Dado que os ataques podem apresentar variações significativas em termos de intensidade e impacto, a capacidade de prever o volume de tráfego associado pode permitir a identificação de atividades suspeitas em tempo real. Dessa forma, ao compreender os padrões de transferência de dados durante ataques, é possível adotar medidas preventivas para minimizar os danos e reforçar a segurança dos sistemas. Para isso, foram aplicadas técnicas de ML e processamento distribuído de *Big Data*, utilizando o *Apache Spark* para processar os dados de forma eficiente.

Neste projeto, foi utilizado o conjunto de dados UNSW-NB15, um dos mais reconhecidos na área de cibersegurança. Antes da modelagem preditiva, foram removidos os registos de tráfego normal (sem ataque), de modo a concentrar a análise apenas nos casos maliciosos. Em seguida, diferentes algoritmos de regressão foram implementados para prever a quantidade de *bytes* transmitidos ao atacante durante os incidentes.

1.2 Descrição do Dataset

O *dataset* UNSW-NB15 é amplamente utilizado para pesquisas na área de segurança cibernética e análise de tráfego de rede. Por sua vez, foi criado através da captura de pacotes de rede em ambiente realista, contendo conexões normais e tráfego malicioso de diferentes tipos de ataques. O conjunto de dados está dividido em quatro arquivos CSV, contendo um vasto conjunto de atributos relacionados às conexões de rede.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Cada registo no *dataset* representa uma interação entre dois dispositivos conectados, armazenando informações relevantes para a análise de tráfego. As principais variáveis presentes no UNSW-NB15 incluem:

- Informações de origem e destino: Endereço IP e porta utilizada na conexão.
- Protocolos de comunicação: Tipo de protocolo utilizado, como TCP, UDP ou ICMP.
- Duração da conexão: Tempo total de cada interação entre dispositivos.
- Volume de dados transmitidos: Quantidade de *bytes* enviados e recebidos durante a conexão.
- Características técnicas dos pacotes: Informações sobre as propriedades dos pacotes de rede envolvidos.
- Classificação do tráfego: Indicação se a conexão foi normal ou um ataque, além da especificação do tipo de ataque identificado.

Este conjunto de dados possibilita a análise detalhada das características do tráfego malicioso, permitindo a construção de modelos preditivos capazes de antecipar o comportamento de ataques cibernéticos. A partir dessas informações, este projeto visa identificar padrões de tráfego associados a ataques e prever a quantidade de dados comprometidos durante essas ocorrências.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2 Exploração e Transformação dos Dados

2.1 Importação e Preparação dos Dados

O primeiro passo na análise dos dados consistiu na importação dos ficheiros do *dataset* UNSW-NB15 e na sua conversão para um *DataFrame Spark*, permitindo um processamento eficiente e escalável. O *dataset* original estava dividido em quatro ficheiros CSV distintos (UNSW-NB15_1, UNSW-NB15_2, UNSW-NB15_3 e UNSW-NB15_4), os quais não continham os nomes das variáveis. Para garantir a correta identificação das colunas, os nomes das variáveis foram extraídos do ficheiro "NUSW-NB15_features.csv" e aplicados a todos os dados. A unificação dos ficheiros permitiu criar um único *dataset* estruturado, facilitando as análises subsequentes.

Uma vez que o objetivo do projeto é prever a quantidade de dados transferidos durante ataques cibernéticos, foi necessário remover os registos onde não ocorreram ataques. Assim, os dados foram filtrados para manter apenas os registos classificados como ataques, ou seja, aqueles onde a coluna ***label*** tem o valor 1. Essa filtragem reduziu significativamente o volume de dados, passando de 2.540.047 registos para 321.283, otimizando o processamento e garantindo que o modelo se concentrasse exclusivamente no tráfego malicioso.

2.2 Verificação e Tratamento de Valores Anómalos

Durante a exploração inicial dos dados, foram identificados valores ausentes e símbolos inconsistentes. Em particular, verificou-se que o símbolo "-" estava presente na variável categórica ***service***, representando serviços menos comuns ou não identificados. Para garantir a consistência dos dados, esses valores foram substituídos por "*another*", categorizando-os como serviços desconhecidos.

Além disso, foi realizada uma análise das colunas numéricas para identificar valores em falta. Nesses casos, a estratégia adotada foi a imputação pela média da respetiva variável, garantindo que não houvesse perda de informação e que a distribuição dos dados fosse preservada.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

```
Colunas com valores '-':
service: 79877

Colunas com valores ausentes (null):
ct_flw_http_mthd: 282939
is_ftp_login: 297183
```

Figura 1 - Valores Anómalos

2.3 Normalização das Variáveis Numéricas

Para evitar que variáveis com escalas muito diferentes influenciassem o desempenho do modelo, foi realizada a normalização das características numéricas utilizando a técnica *StandardScaler* do *Apache Spark*. Esse método ajusta os valores das variáveis para que tenham média zero e desvio padrão igual a um, garantindo uma contribuição equilibrada no modelo de ML. A normalização foi aplicada às seguintes variáveis:

- ***dur*** (duração da conexão)
- ***Spkts*** (número de pacotes enviados)
- ***Dpkts*** (número de pacotes recebidos)
- ***sloss*** (número de pacotes perdidos na origem)
- ***dloss*** (número de pacotes perdidos no destino)

Essa transformação permitiu que todas as variáveis numéricas fossem analisadas na mesma escala, evitando que valores muito elevados dominassem o modelo preditivo.

2.4 Transformação de Variáveis Categóricas

Como os algoritmos de ML não conseguem processar diretamente variáveis categóricas no formato de texto, foi necessário convertê-las para representações numéricas. Para isso, foi utilizado um *pipeline* composto por *StringIndexer* e *OneHotEncoder*, onde:

1. *StringIndexer* converteu as categorias textuais em índices numéricos.
2. *OneHotEncoder* transformou esses índices em vetores binários, garantindo que o modelo pudesse interpretar corretamente as relações entre as categorias.

Essa abordagem foi aplicada às principais variáveis categóricas, garantindo que a informação semântica original fosse preservada e devidamente representada no modelo.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.5 Seleção de Variáveis mais relevantes

Para melhorar o desempenho do modelo e reduzir a dimensionalidade dos dados, foram aplicadas duas técnicas complementares para a seleção das variáveis mais relevantes na previsão da variável-alvo ***dbytes***:

1. Correlação de *Pearson*: Avalia a relação entre ***dbytes*** e todas as variáveis numéricas.
2. Teste Qui-Quadrado: Identifica as variáveis categóricas mais significativas na previsão da variável-alvo.

Os resultados dessa análise permitiram selecionar as 15 variáveis mais relevantes para o modelo:

```
Top 10 características numéricas mais correlacionadas com dbytes:
dloss: 0.9992
Dpkts: 0.9761
res_bdy_len: 0.4224
Dload: 0.4121
dmeansz: 0.3027
dur: 0.1570
sttl: 0.1299
Spkts: 0.1172
dwin: 0.0825
swin: 0.0824

Top 5 características categóricas mais significativas para prever dbytes:
srcip: 1.0000
sport: 1.0000
dstip: 1.0000
dsport: 1.0000
proto: 1.0000
```

Figura 2 - Variáveis mais relevantes

A seleção dessas variáveis permitiu otimizar a eficiência do modelo, reduzindo a quantidade de atributos utilizados no treino e garantindo que apenas as características mais informativas fossem incluídas na análise preditiva.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.6 Análise Estatística

Após a seleção das variáveis, foi realizada uma análise estatística detalhada para compreender a distribuição dos dados e identificar possíveis anomalias. Foram calculadas métricas como média, mediana e desvio padrão para avaliar a dispersão dos valores. A análise revelou que a variável-alvo ***dbytes*** apresenta uma distribuição assimétrica, com a maioria dos valores concentrados em níveis baixos, mas com alguns casos extremos de transferências de dados muito elevadas.

```
Estatísticas descritivas da variável alvo (dbytes):
Média: 4446.658612500506
Desvio Padrão: 114309.2436858934
Mínimo: 0.0
Máximo: 14657531.0
```

Figura 3 - Estatísticas descritivas da variável alvo

A grande diferença entre a média e o valor máximo sugere a presença de *outliers*, indicando que alguns ataques cibernéticos envolvem volumes de transferência de dados extremamente elevados.

2.7 Visualização dos Dados

Para complementar a análise estatística, foram criadas diversas visualizações que ajudaram a identificar padrões nos dados, incluindo:

- Histograma da variável ***dbytes***: Evidenciou que a maioria das conexões tem um volume de tráfego reduzido, mas há alguns valores extremamente altos.

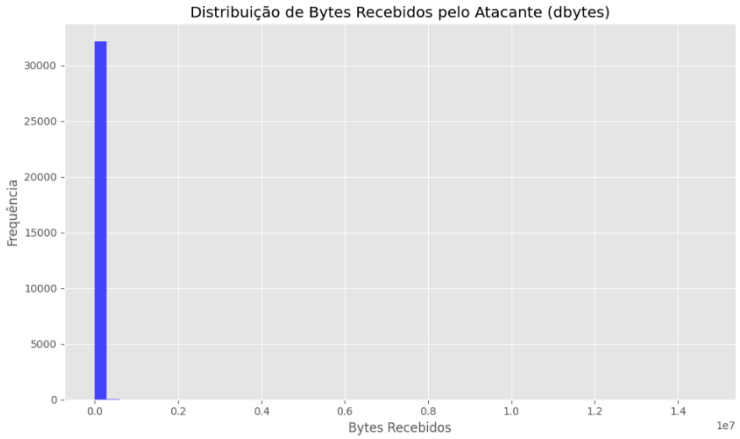

Figura 4 – Gráfico de Distribuição de Bytes Recebidos pelo Atacante (dbytes)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- *Boxplot* da variável-alvo por tipo de ataque: Permitiu identificar que determinados ataques apresentam volumes de transferência de dados muito superiores à média, no entanto, a maioria dos ataques não transfere muitos dados para o atacante. Os ataques DoS (*Denial of Service*) são conhecidos por consumir muitos recursos, o que pode explicar os valores elevados de ***dbytes***. Por outro lado, os ataques *Exploits* também apresentam grande variação, podendo estar associados a ataques que exploram vulnerabilidades para extrair ou transferir grandes volumes de dados.

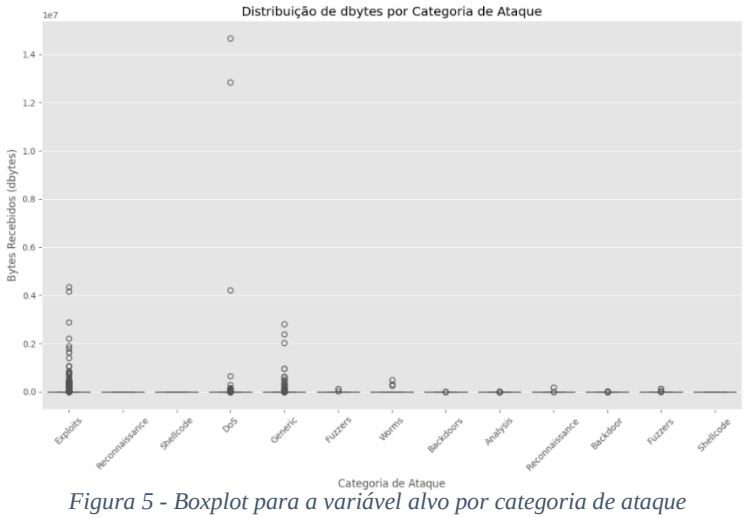

Figura 5 - Boxplot para a variável alvo por categoria de ataque

- *Heatmap* de correlação: Demonstrou que ***dbytes*** tem uma forte relação com ***Dpkts*** e ***dloss***, indicando que o número de pacotes recebidos e a perda de pacotes são fatores críticos na previsão do volume de dados transferidos.

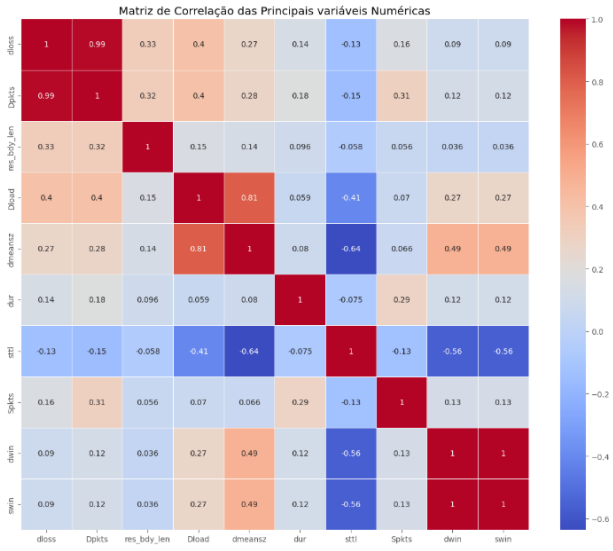

Figura 6 - Matriz de correlação para as variáveis numéricas

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Gráficos de dispersão: Ilustraram a relação entre as principais variáveis numéricas e ***dbytes***, evidenciando padrões no tráfego malicioso.

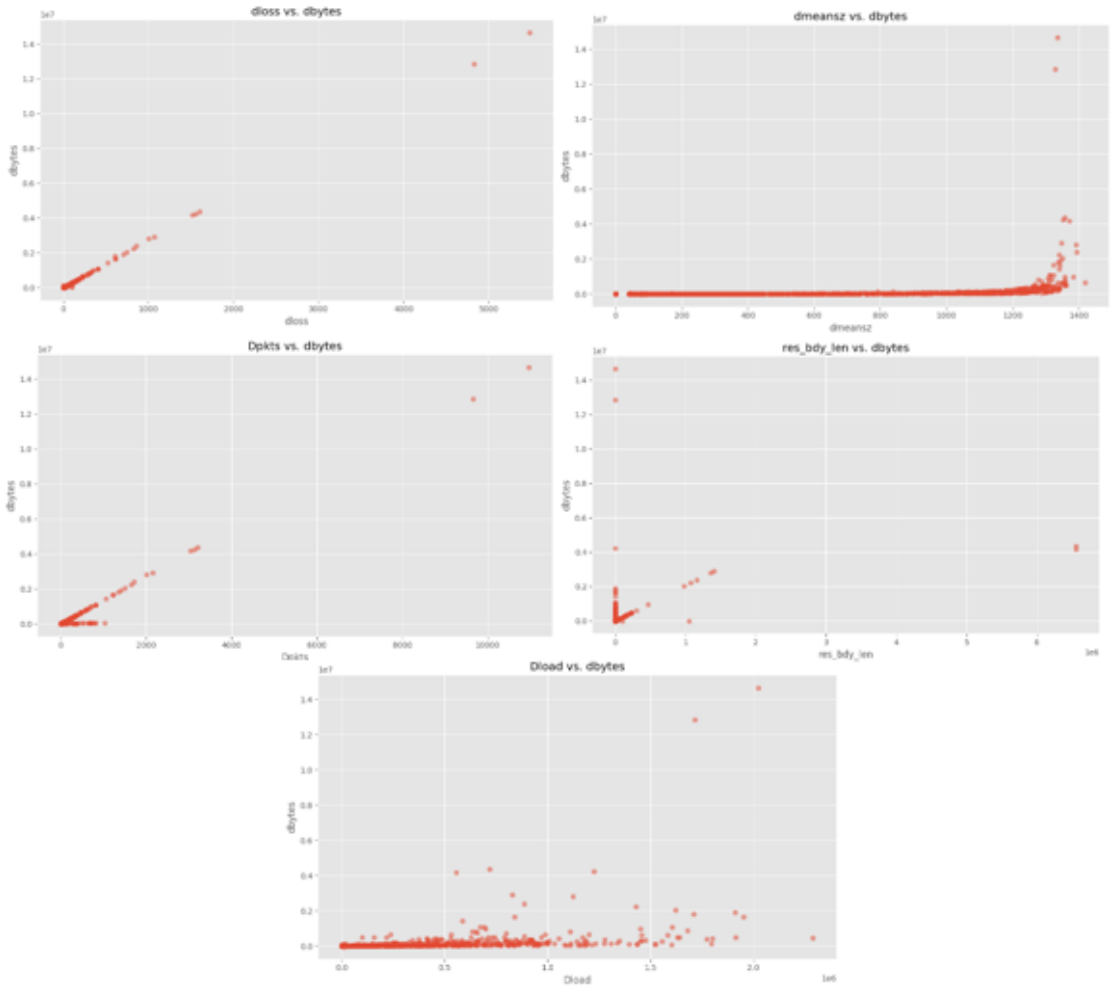

Figura 7 - Gráficos de dispersão das principais variáveis numéricas vs. variável alvo

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3 Treino e Avaliação do Modelo

3.1 Dividir os dados em conjuntos de treino e teste

Para garantir um processo de modelagem robusto e evitar problemas como *overfitting*, os dados foram divididos em dois subconjuntos: 80% para treino e 20% para teste. Esta abordagem proporciona um equilíbrio adequado entre a quantidade de dados disponível para aprendizagem e a capacidade de avaliar o desempenho dos modelos em dados não vistos.

A divisão foi realizada utilizando o método *randomSplit* do *PySpark*, que distribui os dados de forma aleatória de acordo com as proporções definidas. Para assegurar a reprodutibilidade dos resultados, foi utilizada uma semente aleatória (*seed* = 42), garantindo que a mesma divisão possa ser reproduzida em execuções futuras.

Após esta etapa, os conjuntos de treino e teste ficaram com os seguintes tamanhos:

- Conjunto de treino: 257.111 registos
- Conjunto de teste: 64.172 registos

Para avaliar a qualidade das previsões dos modelos, foi criado um avaliador de regressão utilizando a métrica *Root Mean Squared Error* (RMSE) como principal indicador. Esta métrica mede o erro médio quadrático entre os valores reais da variável-alvo ***dbytes*** e as previsões geradas pelos modelos. Além disso, foram consideradas outras métricas complementares, como o *Mean Absolute Error* (MAE) e o Coeficiente de Determinação (R^2), permitindo uma avaliação abrangente do desempenho preditivo de cada modelo.

3.2 Implementação e Treino dos modelos

Para a previsão da quantidade de *bytes* transferidos durante ataques cibernéticos (***dbytes***), foram testados três algoritmos de regressão com diferentes abordagens:

- RL: Um modelo paramétrico simples que assume uma relação linear entre as variáveis, sendo eficiente e rápido na obtenção de previsões.
- RFR: Um modelo baseado num conjunto de árvores de decisão, robusto e capaz de capturar relações não lineares nos dados.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- GBT**:** Um modelo que treina árvores de forma sequencial, corrigindo erros das anteriores, resultando, geralmente, em maior precisão, mas com um tempo de treino mais elevado.

Cada modelo foi configurado com parâmetros específicos para equilibrar desempenho e eficiência computacional. A RL foi ajustada com um número máximo de 10 iterações (*maxIter = 10*), garantindo uma execução rápida. Para evitar *overfitting*, foi aplicada uma regularização moderada (*regParam = 0.1*). Além disso, foi utilizada uma combinação equilibrada de regularização L1 (Lasso) e L2 (Ridge), com um *elasticNetParam* de 0.5. No caso do RFR, optou-se por um número reduzido de 20 árvores (*numTrees = 20*), o que aumenta a velocidade do modelo, embora possa comprometer sua capacidade preditiva. A profundidade máxima das árvores foi limitada a 5 (*maxDepth = 5*), equilibrando precisão e risco de *overfitting*. Para garantir a reprodutibilidade dos resultados, foi utilizada uma semente fixa (*seed = 42*). Já o modelo de GBT foi configurado com 10 iterações (*maxIter = 10*), mantendo um número reduzido de árvores para otimizar o desempenho computacional. A profundidade máxima das árvores foi limitada a 5 (*maxDepth = 5*), assim como no RFR, para controlar a complexidade do modelo. A taxa de aprendizagem foi definida como 0.1 (*stepSize = 0.1*), adotando assim uma abordagem conservadora para reduzir o risco de *overfitting*. Por fim, foi utilizada a mesma semente fixa (*seed = 42*) para assegurar a consistência nos resultados.

Para padronizar o processo de treino e avaliação, foi implementada uma função que:

1. Regista o tempo de início para medir a duração do treino.
2. Treina o modelo utilizando o conjunto de treino.
3. Calcula o tempo total de treino.
4. Gera previsões no conjunto de teste.
5. Avalia o modelo com métricas como RMSE, MAE e R^2.
6. Retorna um dicionário contendo todas as informações relevantes para comparação posterior.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3.3 Comparação dos modelos

Para avaliar o desempenho dos modelos na previsão da quantidade de *bytes* transferidos durante ataques cibernéticos (***dbytes***), foram utilizadas três métricas principais:

- RMSE: Mede o erro médio quadrático entre os valores reais e previstos. Quanto menor, melhor o desempenho do modelo.
- MAE: Mede a diferença absoluta média entre os valores reais e previstos, sendo útil para interpretar o erro médio das previsões.
- R^2: Indica a proporção da variabilidade da variável-alvo explicada pelo modelo. Quanto mais próximo de 1, melhor a precisão.

A tabela abaixo apresenta os resultados obtidos:

Tabela comparativa de métricas:

	Modelo	RMSE	MAE	R^2	Tempo de Treino (s)
0	Regressão Linear	5404.401779	1065.563480	0.997345	106.846408
1	Random Forest	88500.209152	2438.816345	0.287949	3631.903082
2	Gradient-Boosted Trees	88018.524067	2758.999182	0.295679	11332.102098

Figura 8 - Tabela comparativa de métricas dos modelos

Podemos assim observar que o modelo de RL demonstrou um desempenho excecionalmente superior, apresentando o menor RMSE (5404.40) e um R^2 próximo de 1 (0.9973), indicando que quase toda a variação da variável ***dbytes*** foi explicada pelo modelo. Já os modelos RFR e GBT tiveram um desempenho significativamente inferior, com RMSE 16 vezes maior que o da RL e R^2 abaixo de 0.30, sugerindo baixa capacidade preditiva.

É possível ainda aferir que RL treinou em apenas 106.85 segundos, sendo 34 vezes mais rápida que o RFR e 106 vezes mais rápida que o GBT. O modelo GBT apresentou um tempo de treino extremamente alto (11.332 segundos, ou mais de 3 horas), tornando-se impraticável para aplicações com restrições de tempo.

O desempenho surpreendentemente superior da RL em comparação com modelos mais complexos, RFR e GBT merece uma análise mais aprofundada. Este resultado pode ser explicado pela natureza intrínseca dos dados analisados, onde a relação entre as variáveis preditoras e a variável alvo apresenta uma forte componente linear. Em particular, a alta correlação entre ***dbytes*** e variáveis como *Dpkts* e *dloss* sugere uma relação proporcional

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

direta entre o número de pacotes recebidos e o volume de dados transferidos durante ataques.

Os modelos baseados em árvores, embora teoricamente mais capazes de capturar relações não lineares, podem ter sofrido com a limitação de hiperparâmetros implementada para reduzir o tempo computacional. A restrição na profundidade das árvores (*maxDepth* = 5) e no número de árvores (*numTrees* = 20 para RFR e *maxIter* = 10 para GBT) pode ter impedido estes modelos de capturar adequadamente a complexidade dos dados. Além disso, a grande dispersão dos valores de ***dbytes***, com alguns registos apresentando valores extremamente altos, pode ter afetado negativamente o desempenho destes algoritmos.

3.4 Visualização dos resultados

Para uma melhor interpretação do desempenho dos modelos, foram geradas diversas visualizações que permitiram comparar as métricas de erro, a capacidade explicativa e o tempo de treino de cada abordagem.

- Comparação de RMSE e MAE entre Modelos

Foi criado um gráfico de barras para comparar os valores de RMSE e MAE entre os três modelos testados. Esta visualização evidenciou a grande disparidade no desempenho dos modelos, destacando a superioridade da RL em relação ao RFR e ao GBT.

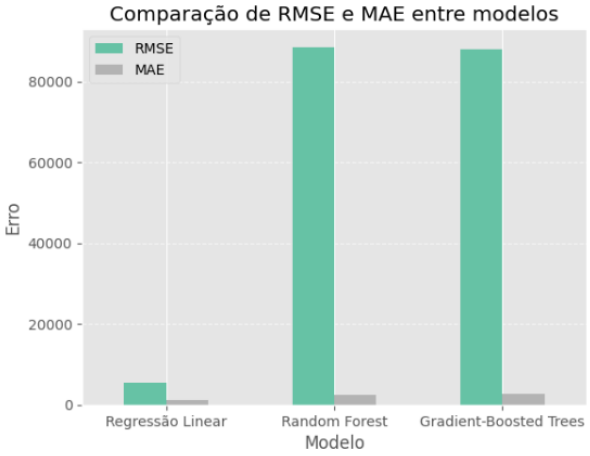

Figura 9 - Comparação de RMSE e MAE entre modelos

Podemos observar que o gráfico demonstra claramente que a RL obteve erros significativamente menores do que os outros modelos. A escala vertical foi dominada pelos valores extremamente elevados de RMSE do RFR e do GBT, fazendo com que os

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

erros da RL parecessem relativamente pequenos. Esta visualização reforçou a escolha da RL como o modelo mais adequado para a previsão da variável ***dbytes***.

- Coeficiente de Determinação (R^2) por Modelo

Foi gerado um gráfico de barras para comparar os valores de R^2 entre os modelos. Esta métrica indica a proporção da variabilidade da variável-alvo explicada por cada modelo.

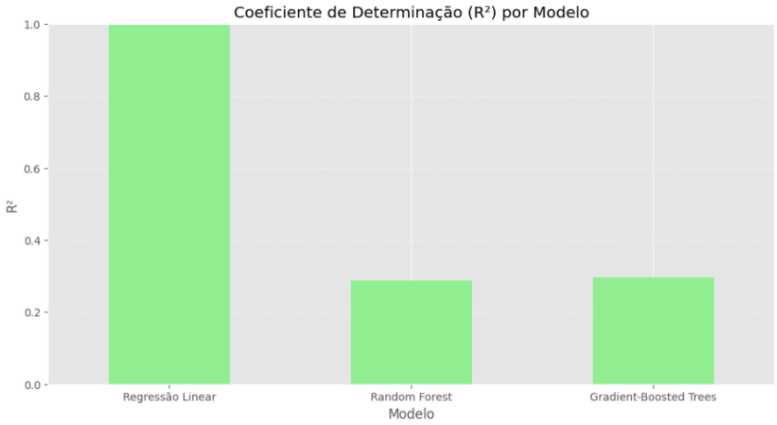

Figura 10 - Coeficiente de Determinação (R^2) por Modelo

Enquanto a RL apresentou um R^2 próximo de 1 (0.9973), demonstrando uma excelente capacidade explicativa, os modelos RFR e GBT tiveram valores de R^2 muito baixos (~0.29), indicando uma baixa capacidade preditiva.

- Tempo de Treino por Modelo

Foi criado um gráfico de barras para visualizar o tempo de treino de cada modelo. Este gráfico destacou as diferenças de eficiência computacional entre as abordagens.

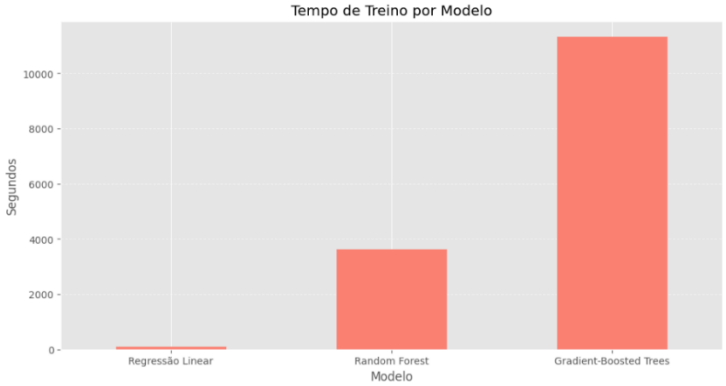

Figura 11 - Tempo de Treino por Modelo

O modelo GBT apresentou um tempo de treino extremamente alto, atingindo 11.332 segundos, o que o torna impraticável para aplicações com restrições de tempo. O RFR teve um tempo de treino significativamente menor, aproximadamente três vezes inferior

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

ao do GBT. No entanto, ainda assim, seu tempo de processamento foi 34 vezes maior que o da RL. Por outro lado, a RL demonstrou ser a abordagem mais eficiente, concluindo o treino em apenas 106.85 segundos, reforçando sua viabilidade para implementação em tempo real. A discrepância entre os tempos de treinamento foi tão acentuada que, no gráfico gerado, o tempo da RL pareceu insignificante devido à escala utilizada.

A análise conjunta dessas visualizações forneceu uma compreensão abrangente do desempenho dos modelos. Os gráficos confirmaram a superioridade da RL, tanto em termos de precisão preditiva quanto em eficiência computacional, consolidando-a como a melhor escolha para a previsão do volume de dados transferidos durante ataques cibernéticos.

3.5 Análise das previsões do melhor modelo

Após identificar a RL como o melhor modelo, foram realizadas análises mais detalhadas para avaliar a qualidade das previsões. Para garantir que a análise fosse computacionalmente eficiente e estatisticamente representativa, foi extraída uma amostra aleatória de 1% das previsões.

- Gráfico de Dispersão: Valores Reais vs. Previstos

Foi gerado um gráfico de dispersão para comparar os valores reais de ***dbytes*** com os valores previstos pelo modelo.

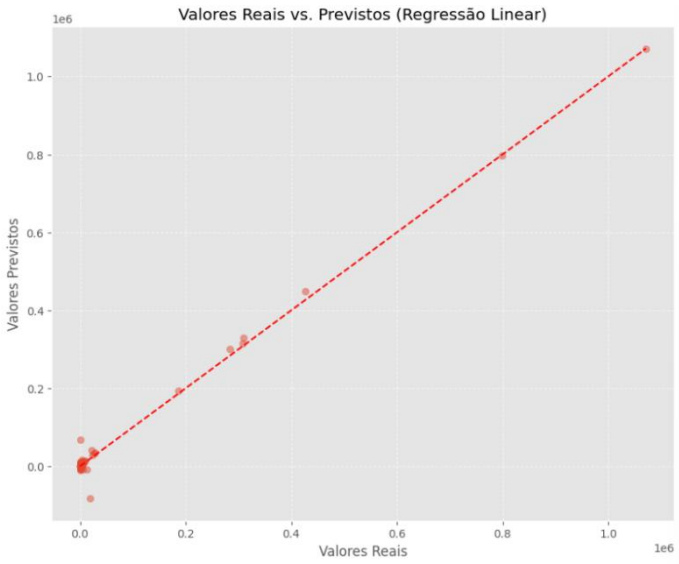

Figura 12 - Valores Reais vs. Previstos (Regressão Linear)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

A linha diagonal vermelha tracejada representa a previsão perfeita (y = x).

A maioria dos pontos está concentrada ao longo dessa linha, confirmando a forte correlação entre os valores previstos e os valores reais, consistente com o alto R^2 (0.9973) observado anteriormente. Há uma ligeira dispersão para valores extremos, indicando que o modelo pode ser menos preciso na previsão de transferências de dados muito elevadas.

O gráfico reforça que o modelo capturou eficazmente a relação entre as variáveis preditoras e a variável-alvo.

- Histograma dos Erros de Previsão

Para compreender a distribuição dos erros do modelo, foi gerado um histograma dos erros, definido como a diferença entre os valores reais e previstos (***dbytes*** - *prediction*).

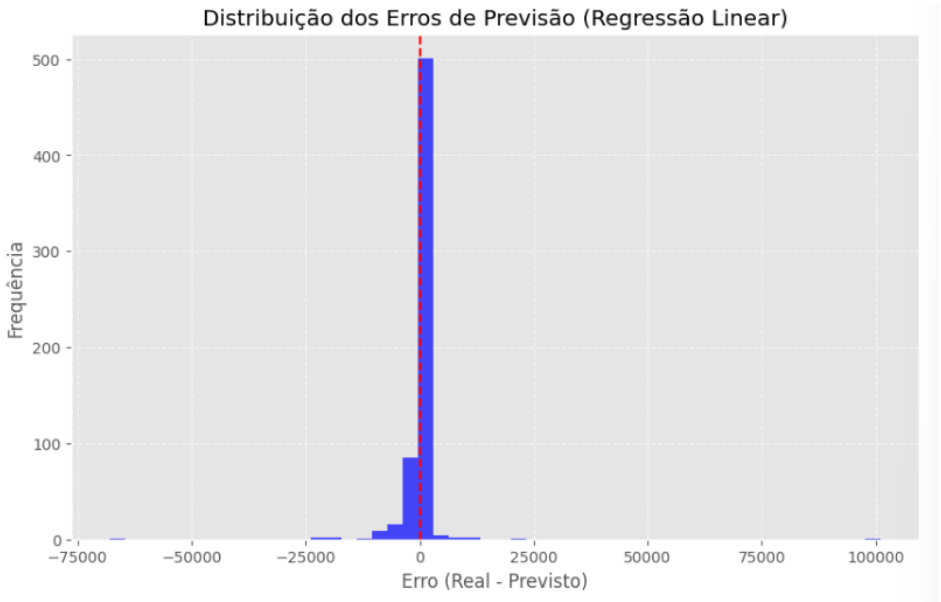

Figura 13 - Distribuição dos Erros de Previsão (Regressão Linear)

A distribuição dos erros é centrada em zero, com um pico pronunciado nessa região, indicando que a maioria das previsões tem um erro pequeno. A linha vermelha tracejada em zero destaca a ausência de viés sistemático, ou seja, o modelo não tende a subestimar ou superestimar os valores de forma consistente. A distribuição é aproximadamente simétrica, reforçando a confiabilidade do modelo. Existem alguns erros extremos nas caudas da distribuição, mas esses casos são relativamente raros, o que é coerente com o baixo RMSE obtido na avaliação dos modelos.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Esta análise detalhada confirmou a precisão e confiabilidade da RL na previsão do volume de dados transferidos durante ataques cibernéticos. Apesar de uma leve dispersão em valores extremos, o modelo demonstrou uma capacidade robusta de previsão, reforçando sua adequação para aplicações práticas em segurança cibernética.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

CONCLUSÃO

Este projeto demonstrou a aplicação de técnicas de *Big Data* e ML na análise e previsão do volume de dados transmitidos durante ataques cibernéticos, utilizando o conjunto de dados UNSW-NB15. A abordagem metodológica adotada permitiu processar eficientemente grandes volumes de dados, transformá-los adequadamente e implementar modelos preditivos com resultados significativos.

Contrariando as expectativas iniciais, a RL destacou-se como o modelo mais eficaz para a previsão da variável ***dbytes***, superando consideravelmente os modelos mais complexos como RFR e GBT. Com um RMSE de apenas 5404.40 e um coeficiente de determinação (R^2) de 0.9973, o modelo linear demonstrou uma capacidade explicativa excecional, capturando quase toda a variabilidade presente nos dados. Adicionalmente, a eficiência computacional, com tempo de treino de apenas 106.85 segundos, reforça a viabilidade para implementação em sistemas de deteção de ataques em tempo real.

Os resultados obtidos sugerem que a relação entre as características selecionadas do tráfego malicioso e o volume de dados comprometidos segue um padrão predominantemente linear. Esta descoberta é particularmente relevante para o desenvolvimento de sistemas de segurança cibernética, pois indica que modelos simples e computacionalmente eficientes podem ser tão ou mais eficazes do que abordagens mais complexas quando aplicados a problemas específicos de previsão em cibersegurança.

A análise detalhada das previsões confirmou a precisão e confiabilidade do modelo de RL, com a distribuição dos erros centrada em zero e sem evidência de viés sistemático. Embora o modelo apresente uma ligeira dispersão na previsão de valores extremos, sua capacidade geral de prever o volume de dados transferidos durante diferentes tipos de ataques demonstra o potencial para aplicações práticas na identificação e mitigação de ameaças cibernéticas.

Este trabalho contribui para o campo da cibersegurança ao demonstrar que a análise preditiva do volume de tráfego malicioso pode ser realizada com alta precisão utilizando técnicas de *Big Data* e modelos estatísticos relativamente simples. A abordagem adotada pode ser expandida para outros contextos de segurança da informação, auxiliando no desenvolvimento de sistemas mais eficientes para deteção e resposta a incidentes.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Como propostas para trabalhos futuros, sugere-se a exploração de modelos híbridos que combinem a eficiência da RL com a capacidade de capturar relações não lineares de outros algoritmos. Além disso, seria relevante investigar a aplicabilidade do modelo em dados de tráfego de rede em tempo real, permitindo uma abordagem mais dinâmica para a deteção de ataques cibernéticos. Outra possibilidade seria expandir a análise para incluir a previsão de outras variáveis relevantes para a cibersegurança, proporcionando uma visão mais abrangente do comportamento do tráfego malicioso. Por fim, recomenda-se aprimorar os modelos baseados em árvores através de uma otimização mais abrangente dos hiperparâmetros, de forma a melhorar o desempenho sem comprometer significativamente a eficiência computacional.

Em síntese, este projeto demonstrou que a combinação de tecnologias de *Big Data* com modelos estatísticos apropriados pode fornecer *insights* valiosos para a compreensão e previsão do comportamento de ataques cibernéticos, contribuindo para o desenvolvimento de estratégias de defesa mais eficazes na proteção de sistemas e infraestruturas críticas.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt