Detecção e Prevenção de Colisões Espaciais

Nico I. G. Ramos GRR20210574

***Resumo*—Este relatório apresenta o desenvolvimento de um modelo para detecção de risco de colisões espaciais, proposto no contexto de um desafio da Agência Espacial Europeia (ESA). O trabalho aborda a análise de séries temporais de mensagens de satélites, a complexidade da análise exploratória de dados (EDA) em um dataset de alta dimensionalidade e a evolução da metodologia: partindo da tentativa de previsão de valores contínuos com modelos ARIMA para uma abordagem de classificação baseada em clusterização com KMeans. O objetivo principal é minimizar os falsos negativos em eventos de alto risco.**

I. Introdução e Definição do Problema

A órbita terrestre encontra-se cada vez mais congestionada, aumentando drasticamente o risco de colisão entre satélites ativos e outros objetos espaciais, sejam eles detritos ou outros satélites. O problema central abordado neste trabalho é prever o risco final de colisão entre um satélite e outro objeto espacial.

O dataset utilizado, proveniente do *Collision Avoidance Challenge* [1] da Agência Espacial Europeia (ESA), consiste em alertas reais enviados periodicamente pelos satélites à base. A base de operações recebe uma grande quantidade de avisos, mas apenas uma fração muito pequena representa um risco real e elevado. As manobras de evasão precisam ser planejadas com pelo menos 2 dias de antecedência em relação ao TCA, sendo a decisão final tomada 1 dia antes.

Cada satélite envia mensagens contendo alertas de aproximação, formando uma série temporal para cada evento de possível colisão. Cada mensagem contém informações, tais como:

- Data estimada da colisão (TCA - *Time of Closest Approach*);
- Risco estimado;
- Incertezas associadas à medição e à órbita.

Os scripts desenvolvidos podem ser encontrados no repositório do autor [2].

A. Desafio Proposto

O desafio, proposto pela Agência Espacial Europeia (ESA), consiste em treinar um modelo capaz de prever o risco final estimado pelo satélite. O principal objetivo é minimizar os **Falsos Negativos**, ou seja, evitar que eventos de alto risco sejam incorretamente classificados como de baixo risco pelo modelo. Mais informações sobre o desafio podem ser encontradas na página oficial da competição [1].

II. Dados e Séries Temporais

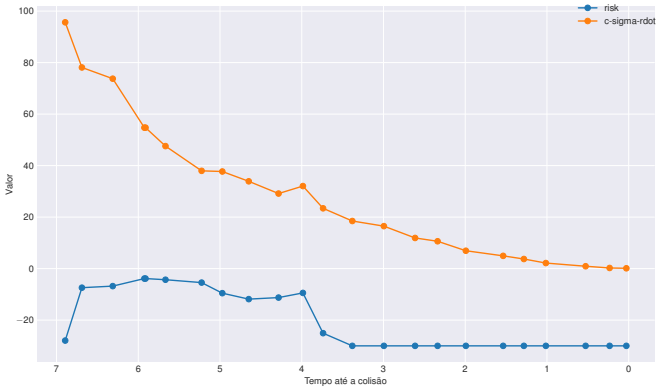
As séries temporais no dataset são sequências de observações de variáveis coletadas em intervalos regulares e ordenadas cronologicamente, permitindo acompanhar a evolução dos atributos (velocidade, incerteza, risco) ao longo do tempo até o TCA (Time of Closest Approach).

Os eventos possuem a seguinte estrutura:

- Cada linha do dataset representa uma observação (mensagem);
- Todas as linhas referentes a um mesmo *event_id* formam a série temporal daquele evento;

A Figura 1 exemplifica parte de uma série temporal de um evento.

Figura 1. Exemplo de série temporal de um evento.



A. Objetivo do Projeto

Inicialmente, o objetivo era projetar e treinar um modelo capaz de prever o valor numérico do risco final estimado, comparando o desempenho de dois otimizadores (*descida de encosta* e *Adam*) aplicados a modelos ARIMA (*Auto Regressive Integrated Moving Average*).

Esses modelos tentam prever valores futuros de uma série com base em seus próprios valores passados (lags) e nos erros de previsão anteriores. Trata-se de uma abordagem clássica para dados estacionários ou que podem ser tornados estacionários; uma explicação mais detalhada sobre esses modelos pode ser encontrada nos materiais didáticos do curso de Series Temporais do professor Lucambio Perez [3].

Conforme detalhado na seção de Metodologia, dificuldades com esses modelos levaram à redefinição do objetivo para a **classificação** dos eventos em alto ou baixo risco.

III. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados foi o principal desafio enfrentado neste projeto. A complexidade advém de ser um problema de domínio específico (dinâmica orbital) e da alta dimensionalidade inicial, com 103 colunas.

Em uma fase preliminar da análise (anterior aos resultados finais aqui apresentados), o processo foi iterativo. Tentou-se usar técnicas estatísticas para descobrir as colunas mais relevantes e filtrar as de menos relevância. Contudo, nenhum

conjunto obtido era considerado satisfatório e optou-se por descartar a seleção puramente empírica e buscar apoio na literatura:

1) *PACEcraft Collision Avoidance Challenge: Design and Results* [4]: Analisa os resultados da competição, fornecendo *insights*.
2) *Implementation and Comparison of Data-Based Methods* [5]: Compara métodos estatísticos e de ML para o problema.

Com base na interseção das features citadas nestes artigos, chegou-se ao conjunto inicial de variáveis, focado principalmente nas métricas de incerteza.

A Tabela 1 mostra o conjunto inicial e o significado de cada atributo.

Tabela 1
COLUNAS SELECIONADAS INICIALMENTE

Coluna	Significado
event_id	Identificador único do evento de conjunção (ID da CDM).
risk	Valor de risco autocomputado no instante de cada CDM ($\log_{10}$ da probabilidade).
time_to_tca	Intervalo de tempo entre a criação do CDM e o instante de aproximação máxima (TCA), em dias.
max_risk_scaling	Fator de escala usado para calcular a probabilidade máxima de colisão.
mahalanobis_distance	Distância de Mahalanobis baseada na covariância combinada.
c_sigma_t	Desvio-padrão da posição transversal (along-track) do chaser [m].
max_risk_estimate	Probabilidade máxima de colisão obtida pela covariância combinada escalada.
c_sigma_rdot	Desvio-padrão da velocidade radial do chaser [m/s].
miss_distance	Distância relativa entre chaser e target no TCA [m].
c_position_covariance_det	Determinante da matriz de covariância de posição do chaser.
c_sigma_n	Desvio-padrão da posição normal (cross-track) do chaser [m].
c_sigma_r	Desvio-padrão da posição radial do chaser [m].
c_obs_used	Número de observações usadas na determinação de órbita.
c_sigma_ndot	Desvio-padrão da velocidade normal (cross-track) do chaser [m/s].
relative_position_n	Posição relativa entre os objetos no eixo normal/cross-track [m].
c_recommended_od_span	Intervalo recomendado para determinação orbital (dias).
relative_position_r	Posição relativa entre os objetos no eixo radial [m].
c_sedr	Taxa de dissipação de energia do chaser (SED rate) [W/kg].
SSN	Número de manchas solares (Sunspot Number).
c_crdot_t	Correlação entre velocidade radial e posição transversal do chaser.
relative_speed	Velocidade relativa entre chaser e target no TCA [m/s].
c_time_lastob_end	Fim do intervalo da última observação usada na determinação orbital [dias].
c_time_lastob_start	Início do intervalo da última observação usada na determinação orbital [dias].
c_cr_area_over_mass	Coeficiente relacionado à pressão de radiação solar (área/massa).
c_cd_area_over_mass	Coeficiente balístico do chaser (área/massa).

A. Dimensões do Dataset

O dataset apresenta as dimensões detalhadas na Tabela 2. A quantidade de observações por evento varia, sendo que 50% dos eventos possuem 13 ou menos observações, com um máximo de 23 e mínimo de 1.

Tabela 2
DIMENSÕES DOS CONJUNTOS DE DADOS

Conjunto	Linhas	Eventos Únicos	Média Obs./Evento
Treino	162.634	13.154	12
Teste	24.484	2.167	11

B. Características e Limpeza dos Dados

Os dados são contínuos e as séries possuem uma frequência aproximada de 8 horas. A detecção de anomalias e limpeza seguiu os seguintes passos:

- Identificação de Valores Nulos
- Identificação de Outliers no tempo
- Identificação de Valores Constantes no tempo

Durante a etapa de identificação de valores nulos, foi observada que as colunas *SSN*, *c_sigma_ndot*, *c_sigma_rdot* e *c_crdot_t* possuem valores nulos significativos e foram removidas. A Tabela 3 ilustra a quantidade de valores nulos nessas colunas.

Tabela 3
COLUNAS COM VALORES NULOS SIGNIFICATIVOS

Coluna	Null Count
SSN	6.822
c_sigma_ndot	9.241
c_sigma_rdot	9.241
c_crdot_t	9.241

Para a escolha dos outliers, optou-se por utilizar o método Hampel. Contudo, janelas temporais pequenas ($window = 1$) não detectaram nulos, enquanto janelas maiores detectaram muitos, não sendo encontrado um meio tempo em nenhuma combinação de *window* e tolênrancia. A Tabela 4 demonstra a quantidade de outliers detectados em algumas das principais variáveis com ($windowz > 1$).

Tabela 4
QUANTIDADE DE OUTLIERS DETECTADOS (SELEÇÃO)

Coluna	Qtd. Outliers
SSN	29
c_sedr	304
relative_speed	572
mahalanobis_distance	761
miss_distance	823
risk	1.075
max_risk_estimate	1.108

Para a detecção de valores constantes, foi utilizada uma janela de tamanho 3 e um threshold de 1% do IQR. Observou-se que todas as colunas possuem uma grande quantidade de valores constantes ao longo do tempo, indicando que as séries tem um período estacionário. A Tabela 5 mostra a quantidade de valores constantes por variável.

IV. CORRELAÇÕES

A. Correlação com o risco nas observações finais

As principais colunas correlacionadas com o risco nas observações finais são aquelas que medem a incerteza da observação. Entre elas, a incerteza na determinação da órbita do chaser apresenta a maior correlação, cerca do dobro da segunda maior, seguida pela incerteza na medida da distância,

Tabela 5
QUANTIDADE DE VALORES CONSTANTES POR VARIÁVEL

Feature	Qtd. Constantes
miss_distance	**6,964**
relative_position_n	**7,379**
mahalanobis_distance	**8,134**
relative_position_r	**8,778**
max_risk_estimate	13,532
relative_speed	17,718
c_sigma_rdot	31,706
c_sigma_t	32,086
max_risk_scaling	38,655
c_crdot_t	40,961
c_position_covariance_det	51,916
c_sigma_r	51,874
risk	57,037
c_time_lastob_end	67,493
c_sigma_ndot	70,013
c_time_lastob_start	70,350
c_cr_area_over_mass	72,551
c_obs_used	75,628
c_sigma_n	76,022
c_recommended_od_span	77,530
c_sedr	77,669
c_cd_area_over_mass	78,316
SSN	85,524

que também possui impacto significativo. Ao longo dos dias, a incerteza na determinação da órbita e na posição do chaser se tornam cada vez mais relevantes, enquanto a importância do tamanho do objeto e do arrasto aumentam até a metade do período, para depois diminuir. Já a estimativa de risco máximo e a quantidade de observações começam com grande relevância, mas tendem a perder importância com o tempo.

A Figura 2 mostra a evolução da correlação das variáveis com o risco ao longo dos dias e a Tabela 6 a correlação das variáveis com o risco nos dois dias que antecedem o TCA.

V. RELAÇÃO ENTRE COLUNAS

A. Multicolinearidade (VIF)

As colunas relacionadas ao volume do erro do chaser apresentam os maiores níveis de multicolinearidade. A Tabela 7 apresenta os valores de VIF correspondentes às variáveis.

B. Auto Correlação (ACF)

A análise da auto correlação (ACF) das observações anteriores é apresentada na Figura 3. Observa-se que as auto correlações decrescem monotonicamente, permanecendo significativas até o *lag 3*, a partir do qual deixam de ser relevantes.

Figura 2. Evolução da correlação com o risco ao longo dos dias.

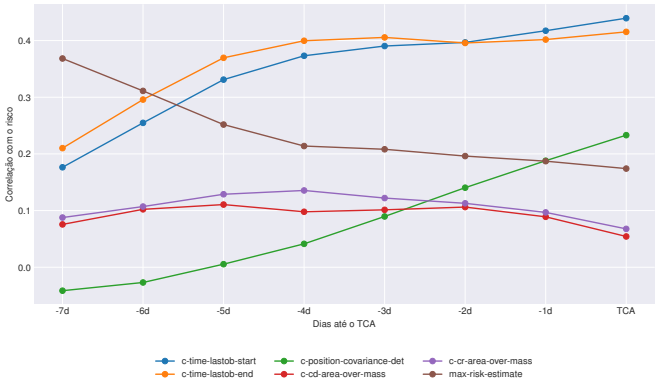


Tabela 6
CORRELAÇÃO COM O RISCO NOS DOIS DIAS ANTERIORES

Feature	-1 day	-2 days
c_time_lastob_start	0.4164	0.4036
c_time_lastob_end	0.4078	0.3923
mahalanobis_distance	-0.2593	-0.2825
c_obs_used	-0.1670	-0.1653
c_cr_area_over_mass	0.1557	0.1563
c_sedr	0.1425	0.1527
c_sigma_t	0.1410	0.1220
c_sigma_r	0.1331	0.1137
c_sigma_n	0.1330	0.1136
c_position_covariance_det	0.1330	0.1136
c_cd_area_over_mass	0.1239	0.1247
c_sigma_rdot	0.1180	0.1168
c_recommended_od_span	0.1171	0.0926
c_sigma_ndot	0.1095	0.1089
max_risk_estimate	0.0999	0.0891
max_risk_scaling	-0.0805	-0.0724
SSN	0.0498	0.0422
c_crdot_t	0.0188	0.0568
miss_distance	0.0037	-0.0324
relative_position_r	0.0008	0.0188
relative_speed	-0.0315	-0.0213
relative_position_n	-0.0150	-0.0084

Isso indica que informações de observações depois do *lag 3* têm menor impacto na estimativa do risco atual.

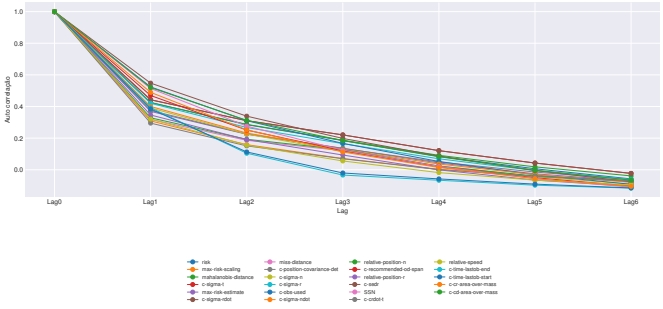
VI. EVOLUÇÃO DO RISCO

No que diz respeito à evolução do risco, a maior parte dos eventos classificados inicialmente como de alto risco é posteriormente atualizada para baixo risco antes da última observação. Poucos eventos apresentam oscilações entre estados, e as transições tendem a se tornar menos frequentes à medida que o TCA se aproxima. Além disso, é evidente que

Tabela 7
VARIANCE INFLATION FACTOR (VIF)

Feature	VIF
c_crdot_t	0.0161
max_risk_estimate	0.0254
relative_speed	0.1402
c_recommended_od_span	0.3911
miss_distance	0.4256
SSN	0.6051
c_time_lastob_end	0.6592
c_obs_used	0.7163
c_time_lastob_start	0.7828
mahalanobis_distance	0.8142
c_cr_area_over_mass	0.8983
c_cd_area_over_mass	0.9302
c_sedr	0.9587
max_risk_scaling	0.9864
relative_position_r	0.9961
relative_position_n	1.0020
c_sigma_t	**73.6501**
c_sigma_rdot	**101.1216**
c_sigma_r	**217,687.1774**
c_sigma_n	**73,747,676.0191**
c_sigma_ndot	**87,927,389.8727**
c_position_covariance_det	**311,512,889.3774**

Figura 3. Auto correlação das observações anteriores



há muito mais eventos que passam de alto para baixo risco do que o contrário.

A Figura 4 ilustra a evolução diária das transições de risco, mostrando como os estados variam ao longo do tempo. Já a Figura 5 apresenta o total das transições por dia, reforçando a tendência de redução das mudanças à medida que o TCA se aproxima.

VII. DIVISÃO DO CONJUNTO DE DADOS

O conjunto de treino foi dividido em 80% para treino e 20% para validação, mantendo a proporção de eventos de alto risco em cada subconjunto. Como esses eventos são raros, eles

Figura 4. Evolução das transições de risco por dia

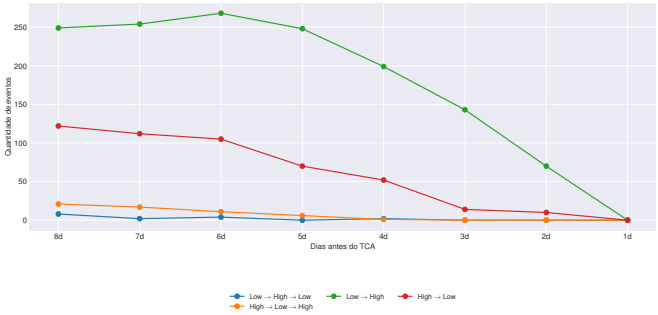
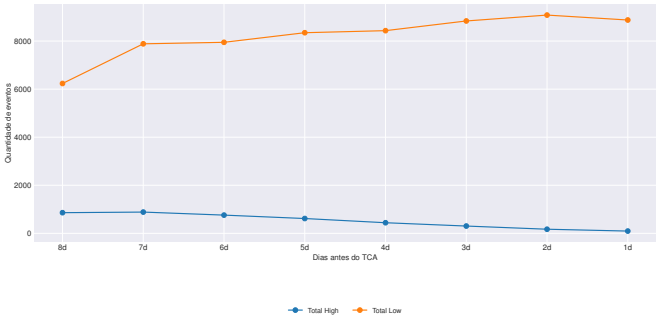


Figura 5. Total das transições de risco por dia



foram selecionados de forma a super-representar a quantidade real. Todos os eventos do conjunto de teste apresentam a última observação a menos de um dia do TCA, enquanto todas as outras observações estão a pelo menos 2 dias do TCA, garantindo tempo suficiente para planejar e executar a manobra. Por outro lado, os eventos do conjunto de teste não são filtrados.

A Tabela 8 ilustra os critérios de observação por conjunto, e a Tabela 9 apresenta a distribuição de eventos por conjunto e categoria de risco.

Tabela 8
CRITÉRIOS DE OBSERVAÇÃO POR CONJUNTO

TCA	Última observação	Todas antes da última
Teste	< 1 dia	> 2 dias
Treino	Qualquer data	Qualquer data

Tabela 9
DISTRIBUIÇÃO DE EVENTOS POR CONJUNTO E CATEGORIA DE RISCO

Conjunto	Alto Risco	Baixo/Médio Risco	Total de Eventos
Treino (Original)	2.77%	97.23%	13154
Validação	2.89%	97.11%	2630
Treino	2.75%	97.25%	10524
Teste	8.21%	91.79%	2167

VIII. DATASET FINAL

A Tabela 10 apresenta as variáveis selecionadas do dataset ao final do EDA, enquanto a Tabela 11 ilustra as estatísticas do dataset final.

Tabela 10
DICIONÁRIO DE VARIÁVEIS SELECIONADAS DO DATASET

Coluna	Significado
event_id	Identificador único do evento de conjunção (ID da CDM).
risk	Valor de risco autocomputado no instante de cada CDM ($\log_{10}$ da probabilidade).
time_to_tca	Intervalo de tempo entre a criação do CDM e o instante de aproximação máxima (TCA), em dias.
max_risk_scaling	Escala usada para calcular a probabilidade máxima de colisão a partir da covariância combinada.
mahalanobis_distance	Distância de Mahalanobis baseada na covariância combinada (métrica estatística não descrita oficialmente em detalhes).
c_position_covariance_det	Determinante da matriz de covariância de posição do *chaser* (volume da incerteza).
c_obs_used	Número de observações usadas na determinação de órbita do *chaser*.
c_recommended_od_span	Intervalo recomendado para atualização da determinação de órbita do *chaser* (dias).
c_sedr	Taxa de dissipação de energia do *chaser* (*Solar Energy Dissipation Rate*).
c_time_lastob_end	Fim do intervalo (em dias) da última observação aceita antes da criação do CDM.
c_time_lastob_start	Início do intervalo (em dias) da última observação aceita antes da criação do CDM.
c_cr_area_over_mass	Coeficiente balístico associado à pressão de radiação solar (área/massa).
c_cd_area_over_mass	Coeficiente balístico associado ao arrasto atmosférico (área/massa).

Tabela 11
ESTATÍSTICAS DO DATASET FINAL

Coluna	Mean	Std	Min	25%	50%	75%	Max
c_cd_area_over_mass	0.7843	2.3417	-128.1786	0.1774	0.4387	0.6928	147.9127
c_cr_area_over_mass	0.3465	0.9655	-0.7121	0.0518	0.1813	0.3052	59.1550
c_obs_used	59.1585	84.9738	3.0000	21.0000	30.0000	57.0000	2227.0000
c_position_covariance_det	1.10e+45	8.53e+45	-8.18e+18	2.36e+11	4.09e+13	6.06e+15	6.73e+58
c_recommended_od_span	12.6636	9.9371	0.0000	6.5900	11.5200	16.4700	234.4100
c_sedr	0.0030	0.0150	-0.1073	0.0003	0.0007	0.0015	0.8762
c_time_lastob_end	0.5897	0.8232	0.0000	0.0000	0.0000	1.0000	2.0000
c_time_lastob_start	40.1508	73.8095	1.0000	1.0000	1.0000	2.0000	180.0000
mahalanobis_distance	192.6028	433.6808	0.0000	22.4056	71.1696	198.4767	15427.1608
max_risk_scaling	5.37e+4	9.09e+5	0.0000	8.3239	31.7449	304.7437	4.98e+7
risk	-19.3406	10.0116	-30.0000	-30.0000	-17.8706	-9.1733	-1.4429

IX. CONCLUSÃO DO EDA

No EDA, houve uma redução de 50% das colunas exógenas originais, com 11 colunas removidas de um total de 21. O dataset final passou a conter 12 colunas, sendo 10 colunas exógenas, 1 identificador da série e 1 identificador do tempo. As colunas quantificam principalmente a incerteza e a confiança na órbita do chaser, bem como sua resistência ao movimento em alta e baixa órbita. Os eventos tendem a apresentar uma parte estável e são fortemente correlacionados com eventos próximos, embora as observações anteriores percam relevância rapidamente. Uma vez que um evento muda de estado, ele tende a permanecer nele, sem oscilar.

X. ESCOLHA DO MODELO

Inicialmente, foi escolhido o modelo ARIMA, entretanto, sua limitação de ser univariado permitiria analisar apenas a evolução do risco ao longo do tempo, sem considerar as interações com as variáveis exógenas, o que não é adequado ao problema. O ARIMAX apresenta a limitação de exigir os valores futuros das variáveis para prever o risco. O VARMAX, por sua vez, não conseguiu gerar modelos confiáveis para a maior parte das séries, as tabelas 12, 13 e 14 mostram os resultados para as ordens (1,0,1), (1,1,0) e (1,1,1), respectivamente. Os modelos foram implementados com a biblioteca statsmodels [6].

Foi feita uma tentativa de treinar um modelo global, capaz de prever padrões com base em todas as séries temporais

Tabela 12
RESULTADOS DA REGRESSÃO: SARIMAX(1, 0, 1)

Dep. Variable: risk — No. Obs: 22 Log Likelihood: -2.196 — AIC: 10.393 — BIC: 11.586						
Termo	Coef.	Std. Err.	z	P> $\|z\|$	[0.025	0.975]
ar.L1	0.7364	0.137	5.384	0.000	0.468	1.004
ma.L1	12.0574	6.143	1.963	0.050	0.017	24.098
sigma2	0.0003	0.000	0.907	0.364	-0.000	0.001
Diagnóstico dos Resíduos: Ljung-Box (Q): 0.05 (Prob: 0.82) — Jarque-Bera: 0.57 (Prob: 0.75) Heteroskedasticity (H): 0.91 (Prob: 0.93) — Skew: -0.03 — Kurtosis: 1.89						

Tabela 13
RESULTADOS DA REGRESSÃO: SARIMAX(1, 1, 0)

Dep. Variable: risk — No. Obs: 22 Log Likelihood: -2.775 — AIC: 9.550 — BIC: 10.346						
Termo	Coef.	Std. Err.	z	P> $\|z\|$	[0.025	0.975]
ar.L1	-0.1733	0.873	-0.199	0.843	-1.884	1.537
sigma2	0.0482	0.026	1.843	0.065	-0.003	0.099
Diagnóstico dos Resíduos: Ljung-Box (Q): 0.06 (Prob: 0.81) — Jarque-Bera: 0.30 (Prob: 0.86) Heteroskedasticity (H): 0.55 (Prob: 0.57) — Skew: 0.15 — Kurtosis: 2.24						

disponíveis. No entanto, algoritmos como o ARIMA capturam apenas o padrão de uma única série temporal. A primeira abordagem consistiu em buscar os melhores pesos para uma nova série temporal de risco, com base nos pesos obtidos no conjunto de treino via *grid search*. Porém, muitas séries eram muito pequenas, fazendo com que o modelo não convergisse.

Também foi tentado reduzir a frequência das séries (aumentar o período) para atingir um tamanho mínimo, mas isso suavizou os dados e impediu que os modelos capturassem padrões temporais relevantes. Em seguida, considerou-se gerar dados artificiais com base em séries similares através de clusterização.

Algumas tentativas com SARIMAX foram feitas, devido à limitação do ARIMAX de prever apenas uma variável, mas os resultados não foram satisfatórios. Os modelos que convergiram não conseguiram generalizar bem para novas séries, principalmente devido ao desalinhamento temporal entre sequências, dificultando encontrar pontos temporais coincidentes entre a série global e a nova série.

Posteriormente, explorou-se a possibilidade de utilizar modelos que considerassem múltiplas

Tabela 14
RESULTADOS DA REGRESSÃO: SARIMAX(1, 1, 1)

Dep. Variable: risk — No. Obs: 22 Log Likelihood: -1.790 — AIC: 9.579 — BIC: 10.487						
Termo	Coef.	Std. Err.	z	P> $\|z\|$	[0.025	0.975]
ar.L1	0.8251	0.177	4.666	0.000	0.479	1.172
ma.L1	-2.0437	NaN	NaN	NaN	NaN	NaN
sigma2	0.0113	NaN	NaN	NaN	NaN	NaN
Diagnóstico dos Resíduos: Ljung-Box (Q): 0.00 (Prob: 1.00) — Jarque-Bera: 2.18 (Prob: 0.34) Heteroskedasticity (H): 1.17 (Prob: 0.90) — Skew: -0.98 — Kurtosis: 4.19						

variáveis e múltiplas séries simultaneamente. Foram escolhidos dois algoritmos da biblioteca *skforecast* [7]: `ForecasterDirectMultiVariate` e `ForecasterRecursiveMultiSeries`. Entretanto, eles não funcionam em conjunto, e não foi encontrada forma de utilizar um regressor que fosse multivariado com o `ForecasterRecursiveMultiSeries`, ou multissérie com o `ForecasterDirectMultiVariate`. Uma alternativa considerada foi utilizar o `ForecasterRecursiveMultiSeries` com SARIMAX, mas isso não resolveu os problemas de convergência e desalinhamento temporal entre séries.

Como os modelos não conseguiram generalizar para novas séries ou não foram adequados para o problema, o objetivo de prever o valor numérico do risco foi abandonado e optou-se por uma abordagem baseada em clusterização seguida de classificação. A ideia é agrupar séries similares e utilizar as características dos clusters para classificar os eventos em alto ou baixo risco.

XI. Treino

No treino, as séries foram padronizadas por meio de ressample, ajustando o tamanho para o da maior série e uniformizando a frequência entre os eventos de cada série. Em seguida, aplicou-se uma normalização com variância média, ignorando diferenças de amplitude e escala entre as colunas, de modo a comparar a forma das séries em vez dos valores absolutos.

A clusterização foi realizada com o algoritmo KMeans, utilizando a métrica DTW para comparar séries desalinhadas, com máximo de 100 iterações e três clusters definidos.Por fim, os eventos foram classificados como alto risco quando 5% dos riscos na última observação excediam o threshold de -6 ($\log 10^{-6}$), conforme especificado no desafio [1]. Todos os demais eventos foram classificados como baixo risco, considerando que casos de alto risco são raros.

No conjunto de validação e teste, a previsão e classificação foram feitas utilizando a penúltima observação, e os resultados comparados com a última observação de cada evento. O conjunto de validação foi utilizado para ajustar o threshold de classificação, enquanto o conjunto de teste foi avaliado com o threshold definido na validação

XII. Resultados

O cluster 0 foi o único cluster a indentificar eventos de baixo risco, a Tabela 15 apresenta a média e o percentil 75% do risco para cada cluster identificado.

Tabela 15
Média e 75% de risco por cluster

Cluster	Mean Risk	75% Risk
0	-27.3440	-28.5429
1	-9.7131	-6.8949
2	-20.7393	-11.4110

As Tabelas 16 apresentam as matrizes de confusão obtidas nos conjuntos de validação e teste. Na validação, observa-se que a maior parte dos eventos de baixo/médio risco foi corretamente identificada, enquanto quatro casos de alto risco foram classificados incorretamente. No conjunto de teste, o comportamento da classificação se manteve similar, classificando quase todos os eventos como de baixo/médio risco, com poucos falsos negativos para alto risco.

Tabela 16
Matriz de confusão na validação e no teste

Conjunto	Real \Previsto	High	Low/Medium
Validação	High	71	4
	Low/Medium	2117	401
Teste	High	172	1
	Low/Medium	1663	256

XIII. Conclusões

Modelos de séries temporais como o Arima, Arimax e Varmax não se mostraram adequados para identificar e **generalizar padrões entre séries distintas**, além de dependerem da **estimativa futura das outras variáveis** para fazer previsões. Eles também não funcionam bem com séries curtas, como as presentes no dataset, embora sejam eficientes quando há observações suficientes.

Além disso, o *resample* das séries pode levar à perda de informações e características importantes, prejudicando a qualidade das previsões. Destaca-se que o modelo tenta prever o risco com base em quão **incertas** as medidas são.

A divisão em clusters priorizou o *recall*, evitando falsos negativos em eventos de alto risco. Na prática, porém, essa estratégia resultou na classificação quase universal de eventos como de alto risco, tornando-se ineficaz para diferenciar confiavelmente entre eventos de alto e baixo risco.

XIV. Trabalhos Futuros

Para melhorar os resultados, é possível explorar a possibilidade de treinar para outros tamanhos de clusters e procurar uma maneira de combinar a classificação do risco com a predição do valor final. A ideia é utilizar a identificação de padrões gerais similares (classificação) para auxiliar na predição do valor final dos eventos, generalizando informações de eventos semelhantes para as séries muito curtas. Também seria interessante explorar outras combinações de variáveis para classificar os eventos.

Referências

[1] AGENCY, E. S.; TEAM, A. C.; OFFICE, S. D. *Collision Avoidance Challenge*. 2019. Plataforma online Kelvins – ESA. Acesso em: 19 nov. 2025. Disponível em: ⟨https://kelvins.esa.int/collision-avoidance-challenge⟩.
[2] IG, N. *Collision-Detection*. 2025. Repositorio GitHub. Acesso em: 19 nov. 2025. Disponível em: ⟨https://github.com/nico-ig/Collision-Detection⟩.
[3] PEREZ, F. L. *Series Temporais III*. 2025. Material didatico disponivel na Internet. Acesso em: 19 nov. 2025. Disponível em: ⟨http://leg.ufpr.br/~lucambio/STemporais/STemporaisIII.html⟩.

[4] IZZO, D. et al. *PACEcraft Collision Avoidance Challenge: Design and Results of a Machine Learning Competition*. 2020. ArXiv preprint. Acesso em: 19 nov. 2025. Disponível em: ⟨https://arxiv.org/pdf/2008.03069⟩.
[5] European Space Agency. Implementation and comparison of data-based methods for collision avoidance in satellite operations. In: *8th European Conference on Space Debris*. [s.n.], 2021. Acesso em: 19 nov. 2025. Disponível em: ⟨https://conference.sdo.esoc.esa.int/proceedings/sdc8/paper/33/SDC8-paper33.pdf⟩.
[6] SEABOLD, S.; PERKTOLD, J.; AL. et. *statsmodels: Statistical Modeling in Python*. 2024. Acesso em 20 de novembro de 2025. Disponível em: ⟨https://www.statsmodels.org/⟩.
[7] RODRIGO, J. A.; ORTIZ, J. E. *skforecast*. 2025. Disponível em: ⟨https://skforecast.org/⟩.