Algoritmo de superamostragem de dados baseado em Curvas Principais

William Bento Pereira*, Fernando Elias de Melo Borges*, Bruno Henrique Groenner Barbosa*,
Marcin Kamiński†, Danton Diego Ferreira*,
*Departamento de Automática, Universidade Federal de Lavras, Lavras, Minas Gerais
Emails: william.pereira1@estudante.ufla.br, danton@ufla.br, fernando.borges4@estudante.ufla.br, brunohb@ufla.br
†Departamento de Máquinas Elétricas, Universidade de Ciências e Tecnologia de Breslávia, Breslávia, Polónia
Email: marcin.kaminski@pwr.edu.pl

***Resumo*—O desbalanceamento entre classes é um dos principais desafios em tarefas de classificação, afetando negativamente a capacidade dos modelos de identificar corretamente instâncias da classe minoritária. Este trabalho propõe e avalia uma técnica de superamostragem baseada em Curvas Principais extraídas por meio do algoritmo *k*-segmentos como alternativa aos métodos tradicionais, como SMOTE e a superamostragem aleatória. A abordagem explora a estrutura e relações não lineares dos dados para gerar amostras sintéticas mais representativas e equilibradas. Foram realizados experimentos em bases de dados sintéticas com diferentes graus de complexidade, e os resultados demonstram que o método baseado em Curvas Principais preserva melhor a distribuição original dos dados e melhora o desempenho de modelos preditivos, especialmente *Random Forest*, em comparação com as demais abordagens. Os resultados encontrados indicam que a técnica proposta se apresenta como uma alternativa eficiente para lidar com desbalanceamento em problemas de classificação complexos, principalmente quando as classes possuem distribuição alongada e em Anel no espaço de características.**

***Palavras-chave*—Desbalanceamento de classes; Superamostragem; Curvas Principais; *k*-segmentos; Dados sintéticos; Random Forest.**

I. Introdução

No cenário atual de aprendizagem de máquina, o desbalanceamento entre classes permanece como um dos grandes desafios para a eficácia de modelos de classificação. Situações em que uma classe está significativamente representada em menor número do que as demais, ocorrem em muitos cenários reais. Tal situação pode prejudicar diretamente a capacidade dos algoritmos de aprenderem de maneira equilibrada, favorecendo as classes majoritárias e marginalizando as minoritárias. Como consequência, o modelo tende a apresentar uma alta acurácia aparente, mas falha na tarefa de identificar corretamente os casos da classe menos representada, que, em muitas vezes, é a de maior interesse na aplicação prática. Esse problema é particularmente crítico em aplicações sensíveis, como a detecção de fraudes financeiras [1], identificação de doenças raras [2] ou monitoramento de falhas em sistemas industriais [3].

Para ilustrar esse desafio de maneira concreta, considere o caso do diagnóstico de câncer de mama em programas de triagem populacional. Nesses cenários, a maioria dos exames pertence a pacientes saudáveis, enquanto um pequeno percentual corresponde a casos positivos da doença. Essa desproporção faz com que, se não forem adotadas estratégias específicas para lidar com o desbalanceamento, os modelos de classificação acabem por negligenciar os casos de positivos, sendo estes os que demandam maior atenção. Estudos como o de Liu et al. [4] evidenciam que o desbalanceamento impacta severamente a eficácia dos diagnósticos automatizados, aumentando o risco de falsos negativos e comprometendo a utilidade clínica dos sistemas de suporte à decisão médica.

Para mitigar os efeitos do desbalanceamento, diferentes estratégias têm sido propostas na literatura. Uma das abordagens mais amplamente adotadas é a técnica de *oversampling*, que busca aumentar artificialmente a representatividade da classe minoritária por meio da criação de novas amostras. Desta forma, pode-se gerar um equilíbrio no conjunto de dados sem a necessidade de coletar mais exemplos reais. Dentre as técnicas mais utilizadas está o SMOTE (*Synthetic Minority Over-sampling Technique*), desenvolvido por Chawla et al. [5], que gera novas instâncias interpolando exemplos existentes da classe minoritária, contribuindo para uma melhor generalização dos modelos de aprendizado de máquina.

No entanto, apesar da popularidade do SMOTE, desafios importantes ainda persistem, como a geração de amostras redundantes ou pouco representativas da real distribuição da classe minoritária, especialmente em cenários com fronteiras de decisão complexas e não lineares [6]. Diante disso, torna-se relevante investigar alternativas que considerem melhor a geometria e a estrutura intrínseca dos dados. Neste contexto, é proposto uma nova abordagem de sobremostragem baseada em Curvas Principais (CP) [7], com o objetivo de explorar de forma mais eficaz a distribuição da classe minoritária.

A proposta central consiste em utilizar as Curvas Principais como guias para a geração de amostras sintéticas ao longo de trajetórias que refletem a topologia dos dados, em vez de simples interpolação linear entre vizinhos, como no SMOTE. Dessa forma, espera-se obter amostras mais informativas, diversas e condizentes com a estrutura real do espaço de características. As Curvas Principais são uma generalização não linear do conceito de Componentes Principais, introduzidas por Hastie e Stuetzle [8], e têm se mostrado eficazes em tarefas como agrupamento [9] e classificação [10], especialmente em domínios com dados de alta dimensão.

A principal contribuição deste trabalho está, portanto, na

proposta de um mecanismo de geração de amostras sintéticas que respeita a geometria dos dados da classe minoritária, explorando as Curvas Principais como ferramenta central para capturar sua estrutura não linear de forma mais fidedigna.

A estrutura do estudo está organizada nas seguintes seções. A Seção 2 aborda o referencial teórico, apresentando os principais conceitos relacionados ao desbalanceamento entre classes, as técnicas de tratamento para dados desbalanceados, a utilização de Curvas Principais na análise de dados e a relevância desses tópicos no contexto do estudo. A Seção 3 detalha os materiais e métodos, descrevendo os recursos utilizados e os procedimentos adotados, incluindo as abordagens para coleta e análise dos dados. Na Seção 4, são apresentados os resultados e a discussão, com ênfase nas bases de dados utilizadas, na aplicação de técnicas de *oversampling*, na análise qualitativa dos dados e nos resultados quantitativos obtidos. Por fim, a Seção 5 traz a conclusão, oferecendo uma síntese das principais descobertas, discutindo suas implicações e sugerindo direções para futuras investigações ou aplicações.

II. Referencial Teórico

A. Problemas de desbalanceamento entre classes

O desbalanceamento entre classes é uma problemática recorrente em tarefas de classificação supervisionada, caracterizada pela presença de uma disparidade significativa no número de amostras entre as classes. Em cenários extremos, a classe minoritária pode representar apenas uma fração ínfima do total de dados, levando os algoritmos de aprendizado a um viés em favor da classe majoritária [11]. Como consequência, o modelo tende a obter altas taxas de acurácia geral, mas apresenta desempenho insatisfatório na identificação de exemplos da classe minoritária.

Esse desequilíbrio dificulta a capacidade dos classificadores de aprender padrões relevantes da classe minoritária, uma vez que as métricas tradicionais, como a acurácia, tornam-se insuficientes para refletir o real desempenho do modelo. Em vez disso, torna-se necessário adotar métricas mais sensíveis, como a precisão, revocação (*recall*), *F1-score*, ou até mesmo a área sob a curva ROC (AUC-ROC), que consideram de forma mais equilibrada o desempenho nas diferentes classes.

Esse problema de avaliação em cenários com classes desbalanceadas é amplamente discutido na literatura. Powers [12] destaca que métricas como acurácia podem ser altamente enganosas quando a distribuição das classes é desigual, e reforça a necessidade de utilizar alternativas como precisão, revocação, *F1-score* e AUC-ROC para refletir de maneira mais fiel o desempenho do modelo.

B. Técnicas de tratamento para dados desbalanceados

Diversas abordagens têm sido propostas para mitigar os efeitos do desbalanceamento entre classes. Entre elas, destacam-se as técnicas de reamostragem, que buscam equilibrar a distribuição das classes por meio da superamostragem da classe minoritária ou da subamostragem da classe majoritária [13].

Dentre essas, a superamostragem aleatória é uma técnica amplamente utilizada para lidar com o desequilíbrio de classes em problemas de aprendizado de máquina. Ela consiste em replicar aleatoriamente exemplos da classe minoritária até que o número de instâncias dessa classe seja equilibrado em relação à classe majoritária. De acordo com o estudo de [5], essa abordagem é uma das mais simples e intuitivas, pois visa aumentar a representação da classe minoritária, tornando-a mais visível para os classificadores. No entanto, apesar de sua simplicidade, a sobreamostragem aleatória pode levar a problemas como o *overfitting*, uma vez que a duplicação das instâncias originais não adiciona novas informações ao modelo. Além disso, a técnica pode resultar em uma sobrecarga computacional, especialmente em conjuntos de dados grandes. Apesar disso, ela é frequentemente usada como uma linha de base em estudos comparativos e pode ser eficaz em cenários onde as classes são levemente desbalanceadas e não há risco significativo de *overfitting*.

O SMOTE (*Synthetic Minority Over-sampling Technique*) tem se mostrado particularmente eficaz na geração de novas amostras sintéticas a partir das existentes na classe minoritária, promovendo uma representação mais equilibrada do espaço de atributos. Estudos recentes, ampliam ainda mais as aplicações do SMOTE ao empregá-lo na análise de colaboração de dados interpretáveis, integrando este método em contextos mais complexos de análise e interpretação de padrões em conjuntos de dados desequilibrados [14].

C. Utilização de Curvas Principais na análise de dados

As Curvas Principais representam uma generalização não-linear do conceito de Componentes Principais. Elas foram originalmente definidas por Hastie e Stuetzle [8] como curvas unidimensionais parametrizadas, suaves (infinitamente diferenciáveis), que possuem a propriedade de auto-consistência e atravessam o conjunto de dados no espaço original.

Seja $g(t)$ uma curva suave no espaço $\mathbb{R}^d$, parametrizada por $t \in \mathbb{R}$, tal que:

$$g(t) = \begin{bmatrix} g_1(t) & g_2(t) & \cdots & g_d(t) \end{bmatrix}^T \tag{1}$$

para cada ponto $x \in \mathbb{R}^d$, define-se o índice de projeção $t_g(x)$ como o maior valor de t para o qual a distância entre x e $g(t)$ é minimizada:

$$t_g(x) = \sup \left\{ t : \|x - g(t)\| = \inf_{\mu} \|x - g(\mu)\| \right\} \tag{2}$$

onde $\|\cdot\|$ representa a norma euclidiana e $\mu \in \mathbb{R}$ é uma variável auxiliar.

Segundo Hastie e Stuetzle [8], a curva $g(t)$ é considerada uma curva principal se satisfizer as seguintes condições:

1) $g(t)$ não se auto-intersecta.
2) $g(t)$ possui comprimento finito em qualquer subconjunto limitado de $\mathbb{R}^d$.
3) A curva é auto-consistente, isto é:

$$g(t) = \mathbb{E}[X \mid t_g(X) = t], \quad \text{para todo } t \in \mathbb{R} \tag{3}$$

Intuitivamente, isso significa que a curva passa pelo meio da nuvem de dados, sendo $g(t)$ a média das observações cujas projeções estão próximas de t.

Neste trabalho, as curvas principais foram extraídas utilizando o algoritmo dos *k*-segmentos, proposto por Verbeek [15].

D. Algoritmo k-segmentos

O algoritmo dos *k*-segmentos, proposto por Verbeek [15], é uma técnica que aproxima uma curva não linear por meio da divisão dos dados em *k*-segmentos lineares. Cada segmento representa uma parte da estrutura dos dados, permitindo capturar variações locais em trajetórias complexas. A ideia central é ajustar sucessivamente retas que se conectam de forma a minimizar o erro de aproximação entre a curva original e os segmentos lineares, resultando em uma representação simplificada, porém fiel, da geometria subjacente dos dados. Essa abordagem é útil, por exemplo, na estimação de Curvas Principais, onde busca-se representar caminhos centrais dos dados com boa precisão e menor complexidade.

O funcionamento do algoritmo inicia-se pelo cálculo do centro dos dados, definido como a média de todas as amostras. A partir desse ponto central, insere-se o primeiro segmento na direção da primeira componente principal, com comprimento proporcional à variância nessa direção (tipicamente $\frac{3}{2}$ do desvio padrão).

Segmentos adicionais são inseridos iterativamente com base em agrupamentos de dados gerados por uma versão modificada do algoritmo *k*-médias, utilizando regiões de Voronoi. A cada nova iteração, os agrupamentos são atualizados e os segmentos recalculados, de modo que cada novo segmento afeta também a posição dos anteriores, garantindo uma construção progressivamente mais precisa da curva.

O processo é repetido até que se atenda a uma das seguintes condições de parada:

- o número máximo de segmentos pré-definido pelo usuário é atingido; ou
- o maior agrupamento contém menos de três amostras, indicando convergência.

Ao final da execução, os segmentos formam uma linha poligonal que serve como aproximação da curva principal. Essa linha pode, opcionalmente, ser suavizada para recuperar a continuidade e suavidade esperadas.

E. Relevância para o contexto do estudo

No presente trabalho, exploram-se a aplicação das Curvas Principais como uma ferramenta alternativa às estratégias de tratamento já existentes, investigando seu potencial para melhorar o desempenho preditivo dos modelos. A integração dessa abordagem visa não apenas aumentar a sensibilidade do classificador à classe minoritária, mas também oferecer uma representação mais fiel da complexidade dos dados analisados.

III. Material e Métodos

A. Método proposto

Como mostrado na Figura 2, as imagens representam as três bases de dados desbalanceadas. Essas bases apresentam distribuições assimétricas entre suas classes, o que as torna apropriadas para a análise de métodos voltados ao tratamento de desbalanceamento. Cada base possui características próprias que desafiam os algoritmos em diferentes contextos.

Os experimentos foram conduzidos em ambiente Python, com o uso de bibliotecas amplamente reconhecidas para aprendizado de máquina e manipulação de dados, garantindo a reprodutibilidade e a eficiência dos testes conduzidos.

A metodologia adotada consiste em uma análise comparativa entre diferentes técnicas de superamostragem aplicadas a conjuntos de dados com classes desbalanceadas. O foco está em investigar o desempenho do método proposto em comparação com o algoritmo SMOTE (do inglês *Synthetic Minority Over-sampling Technique*) e o método de superamostragem aleatória (ROS, do inglês, *Random Over-Sampling*).

Para a extração das Curvas Principais, foi utilizada a biblioteca `ocpc_py`, que possui como parte de seus módulos, um algoritmo de extração de CP via *k*-segmentos na linguagem Python. O processo de superamostragem ocorre da seguinte forma:

1) **Obtenção da Curva Principal (CP):** realizada conforme descrito anteriormente, com a construção dos segmentos ao longo das principais direções dos dados.
2) **Divisão do conjunto de dados:** os dados são divididos em agrupamentos definidos pelos segmentos, sendo cada ponto do conjunto projetado sobre o segmento mais próximo.
3) **Seleção dos pontos projetados:** dentro de cada agrupamento, são selecionados dados conforme o número de dados sintéticos a ser gerado e respeitando a densidade de dados por segmento.
4) **Movimentação dos pontos projetados:** os pontos selecionados são levemente deslocados ao longo do segmento, dentro de uma região limitada por um hiperparâmetro de ruído, com o objetivo de gerar variações realistas dos dados.
5) **Reconstrução da nova amostra:** a partir do ponto projetado modificado, gera-se uma nova amostra considerando a distância do ponto original ao segmento, garantindo a preservação da geometria dos dados.

O processo busca gerar novas amostras respeitando sua distribuição original, evitando, assim, uma interpolação de espaços vazios dos dados, sem replicação. O modelo possui dois hiperparâmetros: o primeiro, diz respeito ao tamanho do conjunto superamostrado e o segundo é um vetor de ruído r que controla o movimento do ponto projetado ao longo do segmento. Desta forma, podemos escrever o ponto projetado sintético conforme a equação 4:

$$\mathbf{p_{si}} = \mathbf{p_{re}} + \mathbf{d}(r) \tag{4}$$

Onde $\mathbf{p_{si}}$ corresponde ao ponto projetado sintético, $\mathbf{p_{re}}$ é o ponto projetado real e $\mathbf{d}(r)$ corresponde à uma distância em função de r.

Cabe ressaltar que a distância segue na direção do segmento, evitando que o ponto projetado fuja do segmento, mantendo

seu vetor canônico igual ao original. Além disto, o algoritmo satura movimentos que possam extrapolar o segmento, limitando a alocação do ponto projetado sintético aos vértices do segmento. De forma a ilustrar o procedimento de geração de dados sintéticos, um exemplo é apresentado na Figura 1.

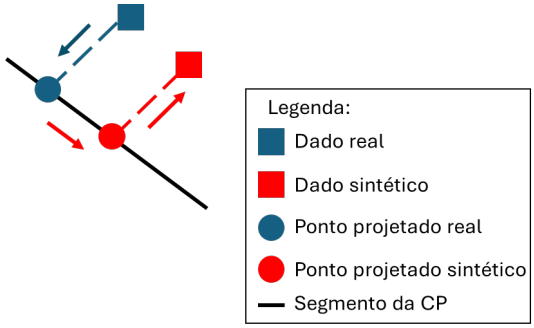


Figura 1: Exemplo de funcionamento do método proposto.

Como custo computacional, o modelo de superamostragem requer a construção da CP por meio do algoritmo k-segmentos, o qual possui complexidade computacional $O(kn^2)$, onde k é o número de segmentos e n o tamanho do conjunto de dados. Na fase de operacional da superamostragem, a função de maior complexidade é o cálculo das distâncias euclidianas de cada amostra à Curva Principal e a movimentação dos pontos projetados conforme descrito na equação 4.

B. Bases de dados e procedimento experimental

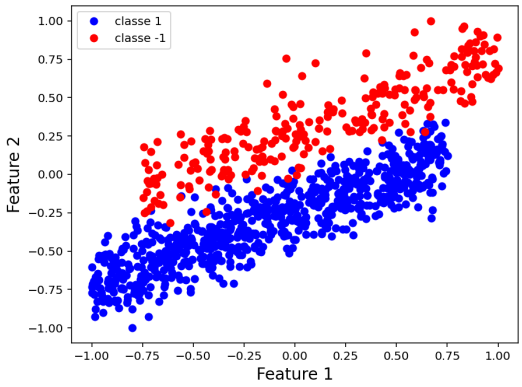
Os dados utilizados consistem em três conjuntos simulados, gerados especificamente para simular diferentes cenários de desbalanceamento (ambas a divisão de treino e teste foi 80/20):

- **Alongada**: proporção das classes majoritária de 75,0% e minoritária de 25,0%, totalizando 1.000.
- **Flame**: proporção das classes majoritária de 66,7% e minoritária de 33,3%, totalizando 1.500 amostras.
- **Anel**: proporção das classes majoritária de 66,7% e minoritária de 33,3%, totalizando 1.500 amostras.

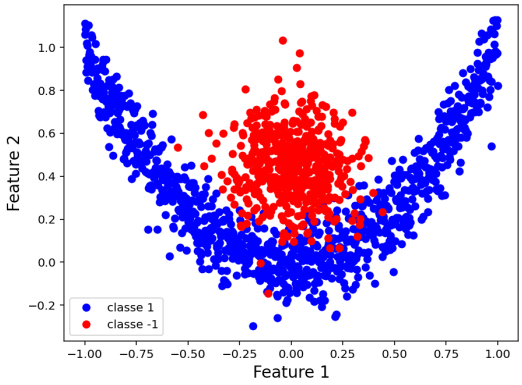
Com o objetivo de ilustrar as bases de dados utilizadas, Na Figura 2 estão dipostos os gráficos de dispersão de cada uma das bases de dados simuladas.

Antes da aplicação dos modelos de aprendizado de máquina, foi realizada uma análise da qualidade dos dados superamostrados, com o objetivo de verificar a adequação das amostras sintéticas geradas. Para isso, utilizaram-se duas abordagens complementares: o teste estatístico de *Kolmogorov-Smirnov* para duas amostras [16], implementado via biblioteca SciPy, e a distância de Wasserstein (*Earth Mover's Distance*). O primeiro avalia se as distribuições acumuladas das amostras originais e sintéticas são estatisticamente equivalentes; o segundo fornece uma medida quantitativa da divergência entre essas distribuições.

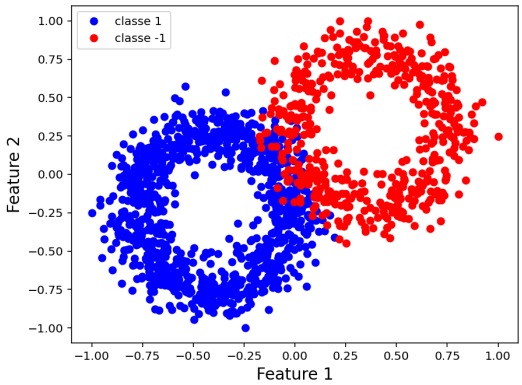
Após essa etapa de verificação, os conjuntos combinados (originais e sintéticos) foram utilizados no treinamento e avaliação do modelo de aprendizado de máquina com *Random Forest* que é um método baseado em múltiplas árvores de



(a) Alongado



(b) Flame



(c) Anel

Figura 2: Bases de dados desbalanceadas (-1) com diferentes características.

decisão, conhecido por sua robustez frente a dados desbalanceados [17].

O treinamento foi realizado mediante validação cruzada do tipo *k-fold* com 5 *folds*. Os conjuntos de dados reais (sem superamostragem) e superamostrados foram treiandos sob o mesmo conjunto de hiperparâmetros do classificador, com o objetivo de isolar o impacto do método de superamostragem nos resultados de classificação. Como resultados, foram dispostos os resultados de classificação relativos ao

conjunto de teste, aquele que não é visto pelo classificador na etapa de validação cruzada. Assim, além dos métodos de subamostragem, foi treinado o conjunto de dados real para ser tomado como *baseline*.

Para a avaliação do modelo, foram utilizadas as principais métricas de desempenho: acurácia, precisão, *recall*, *F1-score*, índice *Kappa* e AUC (Área sob a Curva ROC). A acurácia foi empregada como medida global de acertos, indicando a proporção de previsões corretas em relação ao total de amostras avaliadas.

A precisão e o *recall* foram analisados individualmente por classe. A precisão mede a proporção de verdadeiros positivos entre todas as previsões positivas realizadas pelo modelo, enquanto o *recall* avalia a capacidade do modelo em identificar corretamente todos os exemplos positivos existentes.

O *F1-score*, definido como a média harmônica entre precisão e *recall*, foi utilizado como uma métrica complementar, especialmente útil em cenários com desbalanceamento entre as classes. Já a AUC foi adotada para medir a capacidade do modelo em distinguir entre as classes ao longo de diferentes limiares de decisão, oferecendo uma visão mais ampla da performance discriminativa dos classificadores. Essas métricas, combinadas, possibilitaram uma análise equilibrada e detalhada da eficácia dos modelos aplicados.

Os experimentos foram realizados em um computador com processador Intel® Core™ i5 de 12ª geração, 8 GB de RAM e SSD de 256 GB. Mesmo com limitações de *hardware*, a técnica de superamostragem com Curvas Principais demonstrou-se viável e eficiente, sendo aplicável em contextos computacionalmente restritos

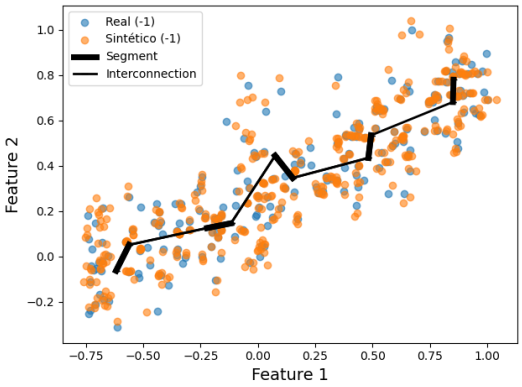
IV. Resultados e discussão

A. Análise visual

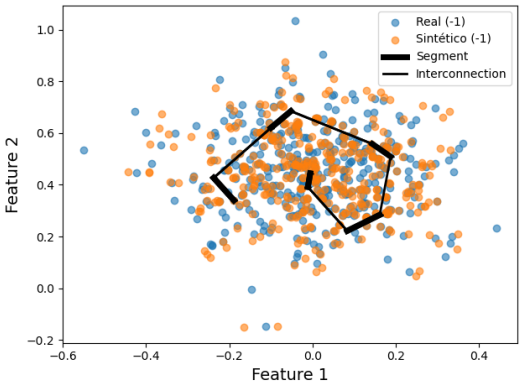
Com o objetivo de visualizar o comportamento de cada método de superamostragem, foram gerados os gráficos de dispersão para as bases de dados relativos à cada um dos modelos de superamostragem aplicados neste estudo. Os gráficos relativos à superamostragem via CP, ROS e SMOTE estão dispostos, respectivamente nas Figuras 3, 4 e 5. Para a superamostragem por CP, também foi disposta a Curva Principal obtida.

É possível observar pela Figura 3 que, o método de superamostragem utilizando CP apresenta uma preservação da estrutura original dos dados, respeitando a distribuição natural da classe minoritária.

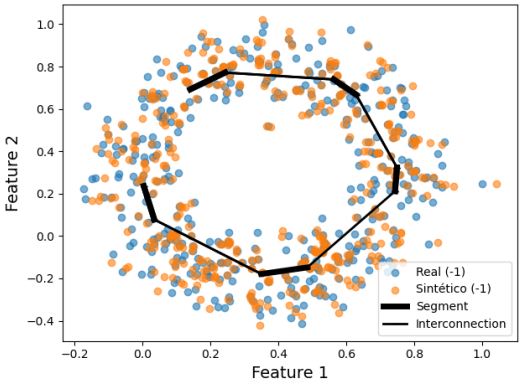
Por outro lado, a superamostragem aleatória (Figura 4) gera novas amostras de forma puramente aleatória, o que pode resultar em uma distribuição artificial dos dados. Já o método SMOTE (Figura 5), embora mais sofisticado, tende a produzir muitas amostras sobrepostas, o que pode levar à perda de variabilidade e ao aumento do risco de *overfitting*.



(a) Alongado



(b) Flame



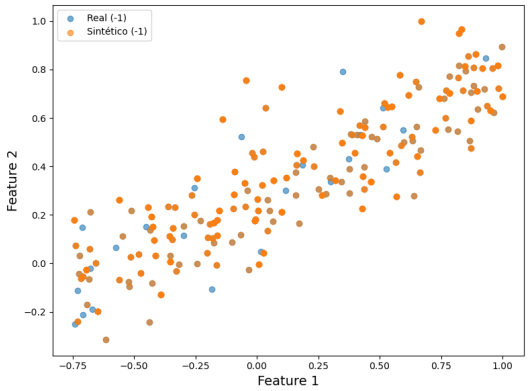
(c) Anel.

Figura 3: Superamostragem via CP.

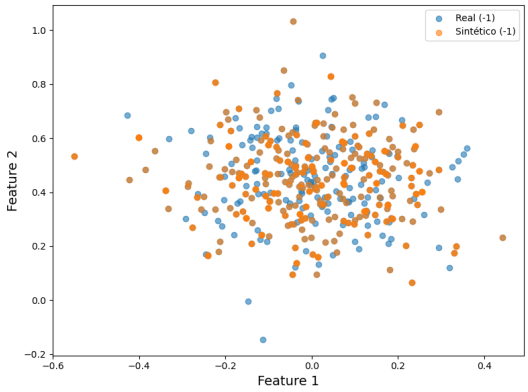
B. Análise da qualidade dos dados

Após a aplicação de técnicas de *oversampling*, foi realizada uma análise de qualidade dos dados utilizando duas métricas estatísticas: a distância de Wasserstein e o teste de Kolmogorov-Smirnov. Essa análise tem como objetivo avaliar o quão bem os métodos de balanceamento preservam a distribuição original dos dados, identificando possíveis distorções introduzidas durante o processo de geração de novas amostras.

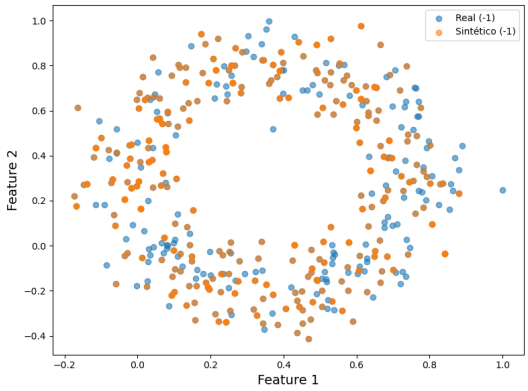
As Tabelas I e II apresentam os valores da distância de Wasserstein obtidos ao comparar os três métodos de balanceamento — Curvas Principais, SMOTE e superamostragem aleatória — aplicados sobre as bases de dados, considerando



(a) Alongado

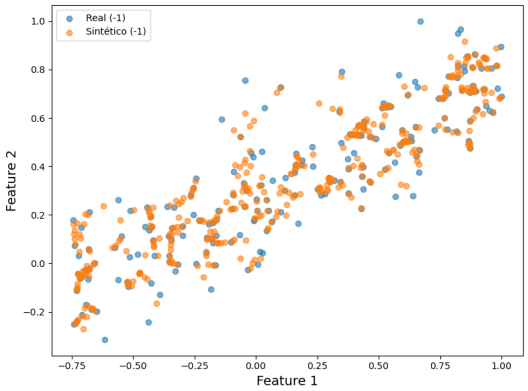


(b) Flame

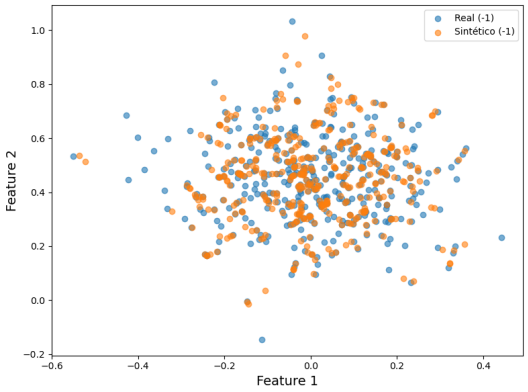


(c) Anel.

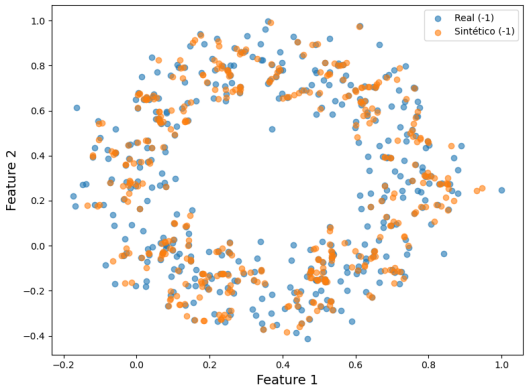
Figura 4: Superamostragem via ROS.



(a) Alongado



(b) Flame



(c) Anel.

Figura 5: Superamostragem via SMOTE.

as duas dimensões das bases de dados *Feature* 1 (F1) e *Feature* 2 (F2), respectivamente. Os valores mais baixos indicam menor diferença entre as distribuições das classes, sugerindo melhor desempenho do método em termos de manutenção da distribuição original dos dados.

Tabela I: Distância de Wasserstein - Feature 1

Base de Dados	Curvas (F1)	SMOTE (F1)	ROS (F1)
Flame	**0,0081**	0,0111	0,0091
Alongada	**0,0143**	0,0156	0,0295
Anel	**0,0135**	0,0140	0,0373

Tabela II: Distância de Wasserstein - Feature 2

Base de Dados	Curvas (F2)	SMOTE (F2)	ROS (F2)
Flame	0,0091	**0,0064**	0,0101
Aloganda	0,0143	**0,0109**	0,0239
Anel	**0,0128**	0,0147	0,0196

Analisando os resultados da distância de Wasserstein, podem-se destacar 2 cenários: o primeiro, com relação a *Feature* 1, o algoritmo proposto apresenta os melhores valores para todas as bases de dados. Em seguida, para a *Feature* 2, o SMOTE apresentou resultado superior, destacando-se nas bases de dados Flame e Alongada. Na base de dados em anel, o método proposto obteve o melhor valor. A superamostragem aleatória obteve resultados gerais inferiores aos dois outros. Tal cenário indica que as amostras geradas pelo método proposto possuem baixa divergência com as amostras reais, indicando boa qualidade da superamostragem. Adicionalmente, o método proposto se mostra equivalente ou, em alguns casos, superior ao SMOTE.

Dessa forma, os resultados numéricos se alinham ao visualizado pelos gráficos das Figuras 3, 4 e 5. Isto reforça a eficácia do método proposto, demonstrando sua capacidade de preservar as características das classes minoritárias, ao mesmo tempo em que minimiza a introdução de réplicas na distribuição dos dados.

As Tabelas III e IV apresentam os *p-values* do teste de Kolmogorov-Smirnov, aplicados às distribuições originais e balanceadas nas *Features* 1 e 2.

Tabela III: Kolmogorov-Smirnov - *pvalue* para *Feature* 1

Base de Dados	Curvas (F1)	SMOTE (F1)	**ROS (F1)**
Flame	**0,9824**	0,8069	0,9358
Alongada	**0,9996**	0,9956	0,9237
Anel	**0,9532**	0,8211	0,2792

Tabela IV: Kolmogorov-Smirnov - *pvalue* para *Feature* 2

Base de Dados	Curvas (F2)	SMOTE (F2)	ROS (F2)
Flame	0,7461	**0,9794**	0,9067
Alongada	0,9996	**0,9999**	0,8341
Anel	**0,9952**	0,9546	0,7963

Os resultados apresentados nas Tabelas III e IV, permitem avaliar o grau de similaridade entre as distribuições originais e aquelas obtidas após a aplicação das técnicas de superamostragem. Valores mais próximos de 1 indicam uma maior aderência à distribuição original, sugerindo menor impacto do processo de geração sintética sobre a estrutura estatística dos dados.

Em linhas gerais, os resultados do teste estatístico seguiu o mesmo padrão dos resultados de divergência pela distância de Wasserstein, com a subamostragem por CP se destacando na *Feature* 1 em todas as bases de dados e possuindo resultados inferiores ao SMOTE para as bases Flame e Alongada na *Feature* 2. Logo, o resultado estatístico corrobora com o resultado de divergência, indicando em quais bases de dados o método proposto possui maior capacidade de superamostragem. Assim, podendo identificar a eficácia do método de Curvas Principais na geração de amostras sintéticas que mantêm elevada fidelidade à distribuição original dos dados, o que é crucial para a construção de modelos preditivos robustos e generalizáveis.

C. Resultados em problemas de classificação

Os resultados de classificação referentes aos métodos de oversampling para as bases Alongada, Flame e Anel são apresentados, respectivamente, nas Tabelas V, VI e VII. Para a obtenção desses resultados, foi utilizado o classificador Random Forest, treinado com os dados originais juntamente com os exemplos sintéticos gerados por cada método de oversampling.

Tabela V: Resultados de classificação para a base de dados Alongada

Métrica	Baseline	CP	ROS	SMOTE
Precisão (-1)	**100.0**	87.5	92.2	94.0
Precisão (1)	80.3	**100.0**	98.7	98.7
Recall (-1)	24.5	**100.0**	95.9	95.9
Recall (1)	**100.0**	95.4	97.4	98.0
F1-Score	89.1	97.6	98.0	**98.3**
AUC	62.2	**97.7**	96.6	97.0
Kappa	32.9	91.0	92.0	**93.3**
ACC	81.5	96.5	97.0	**97.5**

Tabela VI: Resultados de classificação para a base de dados Flame

Métrica	Baseline	CP	ROS	SMOTE
Precisão (-1)	**94.7**	91.0	**94.7**	89.1
Precisão (1)	95.2	**96.0**	95.2	95.5
Recall (-1)	89.9	**91.9**	89.9	90.9
Recall (1)	**97.5**	95.5	**97.5**	94.6
F1-Score	**96.3**	95.8	**96.3**	95.0
AUC	**93.7**	**93.7**	**93.7**	92.7
Kappa	**88.6**	87.2	**88.6**	85.0
ACC	**95.0**	94.4	**95.0**	93.4

Tabela VII: Resultados de classificação para a base de dados Anel

Métrica	Baseline	CP	ROS	SMOTE
Precisão (-1)	**96.9**	91.2	**96.9**	94.8
Precisão (1)	**97.1**	97.0	**97.1**	96.6
Recall (-1)	**93.9**	**93.9**	**93.9**	92.9
Recall (1)	**98.5**	95.5	**98.5**	97.5
F1-Score	**97.8**	96.3	**97.8**	97.0
AUC	**96.2**	94.7	**96.2**	95.2
Kappa	**93.2**	88.8	**93.2**	90.9
ACC	**97.0**	95.0	**97.0**	96.0

Analisando os resultados de classificação para a classe Alongada, observa-se que o método proposto apresentou métricas com valores próximos ou superiores ao *baseline* e aos demais algoritmos de superamostragem. É importante ressaltar que as CP possuem boa capacidade de representação em bases alongadas, conforme reportado em [18]. Em relação à base de dados Flame, pode-se observar que o treinamento convergiu igualmente para o *baseline* e a superamostragem por ROS. Os resultados atingidos pelas CP atingiram valores próximos aos

melhores valores obtidos nas métricas. Para a base de dados em Anel, ocorreu o mesmo efeito da base de dados Flame.

A análise conjunta dos resultados das três bases de dados permite constatar que o método proposto possui potencial de atingir resultados equivalentes ou superiores aos métodos existentes na literatura. Todavia, faz-se importante analisar comparativamente o método proposto com os existentes na literatura, visto que há bases de dados onde as CP possuem melhor desempenho e outras que os demais métodos possuem melhor capacidade de superamostragem.

No caso da base de dados em Anel, destaca-se o uso da CP aberta, apontando uma lacuna dentro do espaço de *features*. Como alternativa de evitar tal lacuna, avaliar o uso da CP fechada faz-se importante, no qual pode-se melhorar o resultado a partir da representação dos dados em todo o espaço.

Além disto, a avaliação do método proposto em relação aos hiperparâmetros se faz necessária para trabalhos futuros, avaliando como o controle do ruído interfere na qualidade dos dados sintéticos gerados e em problemas de classificação. Adicionalmente, com base nos resultados obtidos, a superamostragem por CP se faz viável para ensaios em bases de dados reais, onde se faz possível a observação em dados multidimensionais com dados de diferentes formatos, mesclando *features* reais (valores decimais) e discretas.

Em suma, o método proposto apresenta boa capacidade de gerar dados representativos, o que aponta para aplicação em bases de dados reais, com diferentes tipos de dados e campos de aplicação. Além disso, a aplicação em outros modelos de classificação é importante para avaliar o impacto da subamostragem frente á diferentes classificadores e o comportamento destes nos conjuntos de dados superamostrados pelas CP. Assim, busca-se uma melhoria do processo de validação do modelo proposto neste trabalho.

V. Conclusão

Foi proposto um método alternativo de superamostragem de dados baseado em Curvas Principais. Como forma de avaliar o método, este foi aplicado em 3 diferentes bases de dados simuladas e comparado com 2 outros métodos de superamostragem existentes na literatura. Os resultados apresentados mostram a viabilidade do algoritmo proposto, atingindo resultados equivalentes ou superiores aos demais. Também foram observados os resultados em problemas de classificação, atingindo valores equivalentes aos demais e superando na base de dados Alongada.

Como perspectivas futuras, sugere-se a ampliação do escopo de avaliação do método, utilizando bases de dados reais de repositórios públicos. Outro ponto a ser avaliado é o uso da CP fechada em determinadas bases de dados, haja vista a presença de lacunas de representação da CP em bases de dados específicas, como a base em Anel. Adicionalmente, faz-se importante a avaliação do efeito sobre diferentes classificadores e diferentes tipos de dados. Assim, podendo oferecer um método eficaz de superamostragem que permite gerar novos dados, mantendo as características dos dados reais.

Agradecimentos

Os autores agradecem às agências de fomento CAPES, CNPq e FAPEMIG pelo aporte financeiro à esta pesquisa.

Referências

[1] P. C. Y. Cheah, Y. Yang, and B. G. Lee, "Enhancing financial fraud detection through addressing class imbalance using hybrid smote-gan techniques," *International Journal of Financial Studies*, vol. 11, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2227-7072/11/3/110

[2] P. Branco, L. Torgo, and R. Ribeiro, "A survey of predictive modelling under imbalanced distributions," 2015. [Online]. Available: https://arxiv.org/abs/1505.01658

[3] M. Jalayer, A. Kaboli, C. Orsenigo, and C. Vercellis, "Fault detection and diagnosis with imbalanced and noisy data: A hybrid framework for rotating machinery," *Machines*, vol. 10, no. 4, p. 237, Mar. 2022. [Online]. Available: http://dx.doi.org/10.3390/machines10040237

[4] R. Walsh and M. Tardy, "A comparison of techniques for class imbalance in deep learning classification of breast cancer," *Diagnostics*, vol. 13, no. 1, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/1/67

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] O. Belmokhtar, M. Mokhtari, and A. Benyettou, "Challenges and limitations of synthetic minority oversampling techniques in machine learning: An application to medical data," *PLOS ONE*, vol. 19, no. 1, p. e0287744, 2024.

[7] Y.-R. Yeh, Z.-Y. Lee, and Y.-J. Lee, "Anomaly detection via oversampling principal component analysis," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, 07 2013.

[8] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[9] E. C. C. Moraes, "Método não supervisionado baseado em curvas principais para reconhecimento de padrões," Dissertação de Mestrado, Universidade Federal de Lavras, 2015.

[10] F. E. de Melo Borges, O. F. Mota, D. D. Ferreira, and B. H. G. Barbosa, "One-class classifier based on principal curves," *Neural Computing and Applications*, vol. 35, no. 26, pp. 19 015–19 024, 2023.

[11] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[12] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[13] H. M. P. Marques, "Imbalanced learning: A comparative study of oversampling and undersampling techniques," 2024, acesso em: 22 abr. 2025. [Online]. Available: https://run.unl.pt/handle/10362/175263

[14] A. Imakura, M. Kihira, Y. Okada, and T. Sakurai, "Another use of smote for interpretable data collaboration analysis," *Expert Systems with Applications*, vol. 228, p. 120385, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423008874

[15] J. Verbeek, N. Vlassis, and B. Kröse, "A k-segments algorithm for finding principal curves," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 1009–1017, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865502000326

[16] F. J. Massey, "Kolmogorov–smirnov test: Overview," in *Wiley StatsRef: Statistics Reference Online*. Wiley, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06558

[17] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, J. Hemanth, X. Fernando, P. Lafata, and Z. Baig, Eds. Cham: Springer International Publishing, 2019, pp. 758–763.

[18] E. C. C. Moraes, D. D. Ferreira, G. B. Vitor, and B. H. G. Barbosa, "Data clustering based on principal curves," *Advances in Data Analysis and Classification*, vol. 14, no. 1, pp. 77–96, 2020.