Aplicação de Algoritmos de Aprendizado de Máquina na Análise da Vulnerabilidade Social e Insegurança Alimentar

Pedro L. S. Lobo[1], **Rogerio Salvini**[1], **Juliana Paula Felix**[1]

[1]Instituto de Informática - Universidade Federal de Goiás (UFG)

pedro_lemes@discente.ufg.br, rogeriosalvini@ufg.br, julianafelix@ufg.br

Abstract. *Hunger is a political and economic problem with profound social repercussions, with social vulnerability being an essential indicator for this understanding. This study applies machine learning methods to analyze food insecurity within the context of social vulnerability. Socioeconomic data from Fundação SEADE, RAIS, and CAISAN were used to create classification models that achieved F-score between 80% and 87% in categorizing the Paulista Index of Social Vulnerability (IPVS). Household income emerged as the most relevant factor. The results corroborate previous studies, indicating that socioeconomic data can be explored to identify indicators of vulnerability and food insecurity.*

Resumo. *A fome é um problema político e econômico com profundas repercussões sociais, sendo a vulnerabilidade social um indicador essencial para essa compreensão. Este estudo aplica métodos de aprendizado de máquina para analisar a insegurança alimentar no contexto da vulnerabilidade social. Dados socioeconômicos da Fundação SEADE, RAIS e CAISAN foram usados para gerar modelos de classificação que alcançaram F-score de 80% a 87% na categorização do Índice Paulista de Vulnerabilidade Social (IPVS). A renda domiciliar destacou-se como o fator mais relevante. Os resultados corroboram estudos anteriores, apontando que dados socioeconômicos podem ser explorados na identificação de indicadores de vulnerabilidade e insegurança alimentar.*

1. Introdução

A insegurança alimentar constitui um dos mais cruciais desafios globais contemporâneos, afetando milhões de pessoas, com repercussões profundas nas esferas social, econômica e política [Pérez-Escamilla and Segall-Corrêa 2008]. Em áreas urbanas, como a cidade de São Paulo, a insegurança alimentar se manifesta de maneira multifacetada, sendo influenciada por fatores como desigualdade de renda, acesso limitado a recursos alimentares e a ineficácia de políticas públicas. A compreensão dos determinantes desse problema é essencial para o desenvolvimento de intervenções eficazes que promovam a segurança alimentar de maneira sustentável.

Nos últimos anos, métodos de aprendizado de máquina têm emergido como ferramentas promissoras para a análise de grandes volumes de dados, possibilitando a identificação de padrões complexos e a geração de modelos preditivos aplicáveis a diversas áreas, incluindo a segurança alimentar [Kolisetty and Rajput 2019]. A aplicação dessas técnicas permite a classificação de dados e a identificação de possíveis indicadores

de fome e insegurança alimentar, oferecendo uma abordagem quantitativa e sistemática para a compreensão do fenômeno.

Este estudo tem como propósito aplicar o processo de mineração de dados [Fayyad et al. 1996] em dados das 96 regiões da cidade de São Paulo-SP que estão relacionados à fome e insegurança alimentar. O objetivo principal é a construção de modelos de classificação a partir de indicadores socioeconômicos das regiões como variáveis preditoras e verificar preditivas do Índice Paulista de Vulnerabilidade Social (IPVS). A vulnerabilidade social, medida pelo IPVS, é um indicador essencial para entender a relação com a fome pois reflete condições socioeconômicas desfavoráveis que dificultam o acesso adequado à alimentação. Para alcançar esse objetivo, foram conduzidos experimentos utilizando técnicas de aprendizado supervisionados de classificação [Singh et al. 2016]. Os resultados apresentados neste artigo refletem os principais achados dessas abordagens.

Além desta introdução, este trabalho possui mais três seções. Na seção 2 são apresentados os materiais e métodos, na qual é apresentada a base de dados utilizada, o processamento de dados, e processo de geração e avaliação dos modelos de classificação. Na sequência, a seção 3 com os resultados obtidos. E, por fim segue a seção 4, com a conclusão deste trabalho.

2. Materiais e Métodos

O método aplicado neste trabalho, apresentado na Figura 1, está disposto nas seções seguintes.

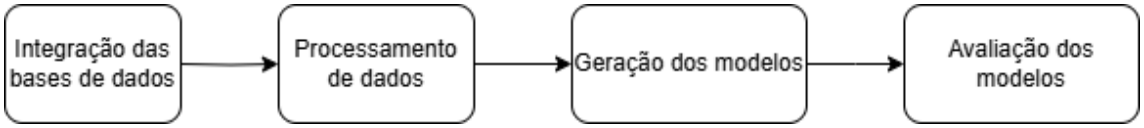

Figura 1. Fluxograma da metodologia aplicada.

2.1. Base de Dados

A base de dados utilizada neste estudo foi construída a partir de três fontes públicas distintas, cada uma contribuindo com informações relevantes para a análise proposta. Estas fontes são Fundação Sistema Estadual de Análise de Dados (SEADE)[1], Relação Anual de Informações Sociais (RAIS), Câmara Interministerial de Segurança Alimentar e Nutricional (CAISAN)[2].

A Fundação Sistema Estadual de Análise de Dados (Fundação SEADE) desenvolveu o Índice Paulista de Vulnerabilidade Social (IPVS), um indicador fundamental para avaliar a vulnerabilidade social no estado de São Paulo. O IPVS, com dados coletados em 2010, considera diversas dimensões, como renda, educação e acesso a serviços básicos, e oferece dados detalhados em nível municipal.

A Relação Anual de Informações Sociais (RAIS), organizada pelo antigo Ministério da Economia, constitui uma base de dados abrangente sobre o mercado de trabalho brasileiro. Essa fonte detalha os empregos em diversos setores, com ênfase em

[1] `http://ipvs.seade.gov.br/view/index.php`

[2] `https://aplicacoes.mds.gov.br/sagirmps/portal-san/artigo.php?link=23`

estabelecimentos que comercializam alimentos frescos. Neste trabalho foram utilizados dados do ano de 2016.

Por fim, foram utilizados dados de 2016 sobre feiras livres, coletados pela Câmara Interministerial de Segurança Alimentar e Nutricional (CAISAN), vinculada ao Ministério do Desenvolvimento Social. Essa fonte fornece informações sobre a quantidade, localização e frequência desses mercados, além de dados sobre os tipos de produtos comercializados, com ênfase especial em itens perecíveis como frutas, vegetais e outros alimentos.

A partir dessas três bases de dados construiu-se o conjunto final de dados utilizado nos experimentos deste trabalho. Essa construção foi realizada por meio da integração das informações de cada base para as regiões da cidade de São Paulo que estavam presentes em todas as três fontes. Com este processo, a base de dados final resultou em 96 observações, representando cada uma das 96 regiões da cidade de São Paulo. Isso garantiu uma representação completa das diferentes áreas da cidade. Além disso, com base em conversas com outros pesquisadores experientes no estudo da vulnerabilidade social, foram identificadas 10 variáveis de interesse, apresentadas na Tabela 1.

Tabela 1. Descrição das variáveis utilizadas.

Variável	**Tipo**	**Valores**
Renda média domiciliar em reais (R$)	Numérico (contínuo)	1.226,46 a 12.333,36
Densidade de estabelecimentos saudáveis (*in natura* + misto) por 10 mil habitantes	Numérico (contínuo)	8,31 a 192,41
Densidade de estabelecimentos não saudáveis (ultraprocessados) por 10 mil habitantes	Numérico (contínuo)	0,0 a 172,92
Número de feiras livres	Numérico (discreto)	0 a 22
Número de estabelecimentos *in natura* na base da RAIS	Numérico (discreto)	1 a 125
Índice Paulista de Vulnerabilidade Social	Numérico (discreto)	0 a 7
Rendimento médio domiciliar dos domicílios particulares (R$)	Numérico (contínuo)	1.211,52 a 10.807,24
Renda per capita dos domicílios particulares (R$)	Numérico (contínuo)	345,74 a 4.726,07
Proporção de pessoas responsáveis alfabetizadas	Numérico (contínuo)	89,9 a 100,0
Número de Estabelecimentos que vendem alimentos *in natura* no distrito	Numérico (discreto)	3 a 147

2.2. Processamento de Dados

A etapa de processamento de dados envolve a limpeza e transformação das informações. O objetivo da limpeza é melhorar a qualidade dos dados, lidando com valores ausentes, redundantes, inconsistentes, ruidosos ou atípicos (*outliers*). Na análise exploratória inicial,

utilizando estatísticas descritivas, não foram identificados problemas de qualidade nos dados integrados. No entanto, como mostrado na Tabela 1, todas as variáveis são numéricas, mas apresentam escalas muito diferentes — como, por exemplo, 'Número de feiras livres' e 'Renda média domiciliar'. Portanto, foi realizada uma transformação dos dados através da padronização pelo método Z-Score, a fim de tornar os valores comparáveis, otimizando o desempenho dos algoritmos na construção dos modelos de classificação [Imron and Prasetyo 2020]. A Equação 1 ilustra essa transformação, onde x representa o valor original da variável, μ a média dos valores, e σ o desvio padrão.

$$x_{new} = \frac{x - \mu}{\sigma} \tag{1}$$

O Índice Paulista de Vulnerabilidade Social (IPVS) foi utilizado como variável dependente (variável alvo) para a tarefa de classificação. Os valores do IPVS variam de zero a sete, indicando níveis crescentes de vulnerabilidade social. A Tabela 2 apresenta os níveis do IPVS com a respectiva distribuição de regiões em cada nível.

Tabela 2. Distribuição dos níveis de IPVS por região.

Nível	IPVS	# Regiões
0	Não classificado	20
1	Baixíssima vulnerabilidade	56
2	Vulnerabilidade muito baixa	7
3	Vulnerabilidade baixa	8
4	Vulnerabilidade média	1
5	Vulnerabilidade alta (urbanos)	0
6	Vulnerabilidade muito alta (aglomerados subnormais urbanos)	4
7	Vulnerabilidade alta (rurais)	0

Em 20 regiões, o valor do IPVS é igual a zero (não classificado), razão pela qual foram excluídas da análise. Dado o desequilíbrio na distribuição dos níveis de IPVS, as regiões com valores de dois a sete foram agrupadas. Com isso, a variável IPVS foi binarizada, com o valor 0 (56 observações) representando uma condição de "não vulnerável" e o valor 1 (20 observações) representando a condição de "vulnerável", conforme apresentado na Tabela 3.

Tabela 3. Distribuição de classes do IPVS binarizado por região.

Classe	IPVS	# Regiões
0	Não vulnerável	20
1	Vulnerável	56

2.3. Classificação

Foram aplicados algoritmos clássicos de aprendizado de máquina para realizar a classificação do IPVS entre regiões sem vulnerabilidade e com vulnerabilidade. Os algoritmos selecionados incluíram: *Support Vector Machine* (SVM), *K-Nearest Neighbors*

(KNN), *Linear Discriminant Analysis* (LDA), *Naïve Bayes* (NB), *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosting* (GB) e *XGBoost* (XGB). Abaixo, apresentamos alguns detalhes sobre cada um destes métodos.

Support Vector Machine (SVM) é um algoritmo de classificação que funciona buscando um hiperplano que melhor separa as diferentes classes. A principal ideia é maximizar a margem entre os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte.

K-Nearest Neighbors (KNN) classifica uma amostra com base nos k vizinhos mais próximos no espaço das características, utilizando uma métrica de distância (como a distância euclidiana) para determinar a proximidade. A suposição subjacente é que amostras com características semelhantes estão localizadas próximas umas das outras.

Linear Discriminant Analysis (LDA) busca encontrar combinações lineares das variáveis preditivas que melhor separam as classes. Esse método é baseado na suposição de que as variáveis seguem uma distribuição normal multivariada e que as classes têm a mesma covariância.

Naïve Bayes (NB) é baseado no Teorema de Bayes e na suposição de independência condicional entre as características, assumindo que o valor de uma característica é independente dos outros.

A *Decision Tree* (DT) atua dividindo os dados em subconjuntos com base em características, onde cada nó representa uma condição e os ramos indicam os possíveis resultados dessa condição.

Random Forest (RF) é um método de aprendizado de conjunto que consiste em várias árvores de decisão, no qual cada árvore é treinada em uma amostra aleatória dos dados. A previsão final é feita com base na média (para problemas de regressão) ou no voto majoritário (para problemas de classificação) das previsões das árvores individuais.

Gradient Boosting (GB) cria um modelo que combina vários modelos fracos, como árvores de decisão de baixa profundidade, em um modelo preditivo forte, adicionando cada novo modelo para corrigir os erros dos modelos anteriores.

XGBoost (XGB) é uma implementação otimizada do Gradient Boosting que se destaca por sua eficiência, escalabilidade e precisão. Ele incorpora técnicas como regularização para reduzir o risco de sobreajuste, além de ser altamente paralelizado, o que o torna mais rápido que outras implementações.

2.4. Otimização dos Modelos

Para otimizar o desempenho dos modelos de classificação foi utilizada a técnica de validação cruzada aninhada (*nested cross validation*).

A validação cruzada [Bro et al. 2008] é uma técnica utilizada para avaliar o desempenho de modelos preditivos, com o objetivo de garantir sua robustez e generalização. Ela consiste em dividir o conjunto de dados disponível em múltiplos subconjuntos. O modelo é treinado em uma combinação de subconjuntos e testado em um subconjunto diferente, repetindo esse processo para cada subconjunto. O desempenho do modelo é então avaliado com base na média dos resultados obtidos em cada iteração, o que ajuda a reduzir o risco de sobreajuste e fornece uma estimativa mais confiável da capacidade de

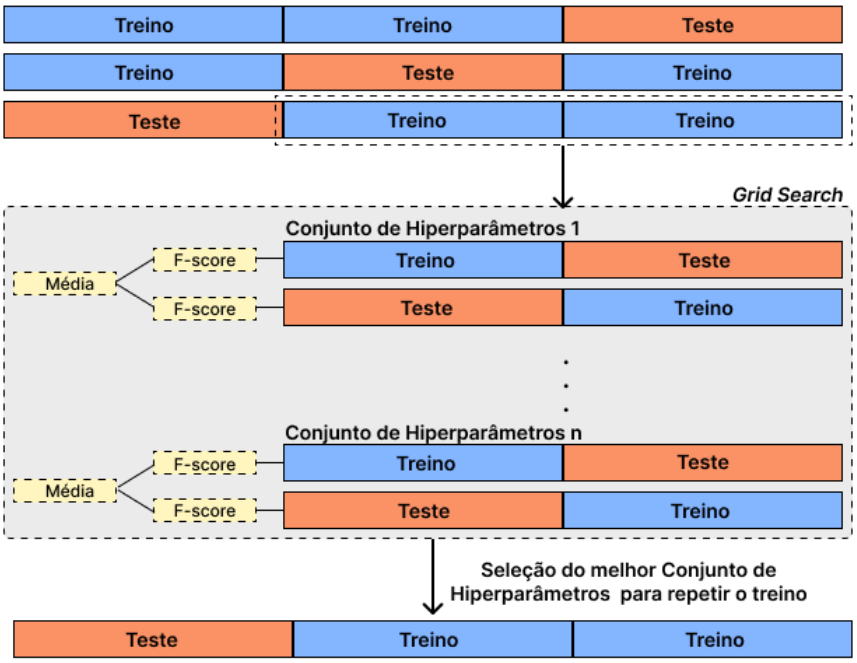

Figura 2. Ilustração do processo de validação *Nested Cross Validation*.

generalização do modelo para dados novos.

A validação cruzada aninhada é um procedimento composto por dois níveis de validação cruzada para a seleção de hiperparâmetros dos modelos. A Figura 2 ilustra este procedimento. O laço interno, utilizando 2 subconjuntos, tem como objetivo a busca dos melhores parâmetros do modelo. A técnica de busca em grade (*Grid Search*) [Yuanyuan et al. 2017] foi aplicada para testar sistematicamente combinações de valores de hiperparâmetros, com o objetivo de identificar a configuração que resulta no melhor desempenho do modelo em termos de uma métrica específica, neste trabalho foi utilizada a *F-score*. Foram avaliados parâmetros como distância, número de vizinhos, tipo de kernel, fator de regularização, critério de decisão, profundidade máxima, grau de suavização, taxa de aprendizado, número de estimadores, métodos de redução, tamanho mínimo de folha, esquema de ponderação, *solver*, poda por complexidade e número máximo de características. O laço externo, com 3 subconjuntos, utiliza um subconjunto como treino para ajustar os hiperparâmetros, usando o loop interno, e os demais de teste para avaliar o desempenho do modelo final. Esse processo garante que a avaliação de performance seja feita em dados independentes dos usados para ajuste dos hiperparâmetros, prevenindo o viés de sobreajuste. Dentre os três conjuntos de parâmetros obtidos, um para cada conjunto do laço externo, foi escolhido o que obteve melhor desempenho da *F-score*.

2.5. Avaliação

A avaliação final do desempenho dos modelos gerados foi feita por meio da validação cruzada *Leave One-Out* (LOO) [Wong 2015]. Diferente de outras formas de validação cruzada, o LOO testa o modelo iterativamente, deixando uma única instância fora do conjunto de treinamento em cada rodada, o que permite que o modelo seja treinado com o máximo de dados disponíveis em cada iteração. Essa abordagem ajuda a obter uma

estimativa mais precisa da performance do modelo em relação ao conjunto completo de dados, reduzindo o viés associado à seleção de amostras de validação. O uso do LOO foi justificado pela necessidade de maximizar a utilização do conjunto de dados, especialmente em situações onde o número de amostras é limitado.

Devido ao desbalanceamento dos dados do IPVS, a métrica de avaliação utilizada foi a *F-score*, apresentada na Equação 2. Em contextos de dados desbalanceados, a *F-score* é considerada uma das métricas mais adequadas, pois leva em consideração tanto a precisão quanto a sensibilidade, equilibrando a avaliação entre falsos positivos e falsos negativos [Riyanto et al. 2023]. Nesta métrica, VP são verdadeiros positivos, VN são verdadeiros negativos, FP são falsos positivos e FN são falsos negativos. Ou seja, VP representa os casos em que o modelo identificou corretamente uma ocorrência positiva; VN indica as previsões corretas de ocorrências negativas; FP corresponde aos casos em que o modelo previu uma ocorrência positiva, mas era negativa; FN refere-se aos erros em que o modelo previu uma ocorrência negativa, mas era positiva. O valor 1 da classe IPVS é considerado como positivo.

$$F\text{-}score = \frac{2 \times \mathrm{VP}}{2 \times \mathrm{VP} + \mathrm{FP} + \mathrm{FN}}. \tag{2}$$

No código desenvolvido para a classificação do IPVS e otimização dos modelos, foi utilizando o Google Colaboratory com a linguagem de programação Python, versão 3.10, e o uso da biblioteca Scikit-Learn.

3. Resultados

Os resultados do *F-score* médio obtidos pelos modelos gerados são apresentados na Tabela 4.

Tabela 4. F-score médio e desvio padrão dos modelos gerados.

Modelo	F-score	Desvio padrão
SVM	0.84	0.36
KNN	0.84	0.36
LDA	0.87	0.34
DT	0.83	0.38
RF	0.87	0.34
NB	0.83	0.38
GB	0.87	0.34
XGB	0.80	0.40

Os hiperparâmetros selecionados que foram utilizados para a geração de cada modelo de classificação são apresentados na Tabela 5.

Conforme apresentado na Tabela 4, o *F-score* variou entre 0.80 e 0.87 dentre os classificadores utilizados. A análise de importância das variáveis no modelo *Random Forest* [Fan et al. 2011] revelou que algumas características têm maior relevância na previsão IPVS. Na Figura 3 são apresentadas essas características. As variáveis relacionadas

Tabela 5. Hiperparâmetros utilizados nos modelos de classificação.

Modelo	Hiperparâmetros
SVM	C=1000, gamma=1, kernel='linear'
KNN	leaf_size=1, n_neighbors=11, p=2, weights='distance'
LDA	shrinkage=None, solver='svd'
DT	ccp_alpha=0.01, criterion='gini', max_depth=None, max_features='sqrt', random_state=0
RF	criterion='gini', max_depth=None, max_features='sqrt', n_estimators=200, random_state=0
NB	var_smoothing=0.12328467394420659
GB	learning_rate=0.01, max_depth=3, n_estimators=50, random_state=0
XGB	gamma=1.0, learning_rate=0.01, max_depth=3, min_child_weight=0, n_estimators=200, random_state=0

à renda, como renda per capita e rendimento médio domiciliar dos domicílios particulares, e renda média domiciliar, são as que possuem maior influência na classificação de vulnerabilidade.

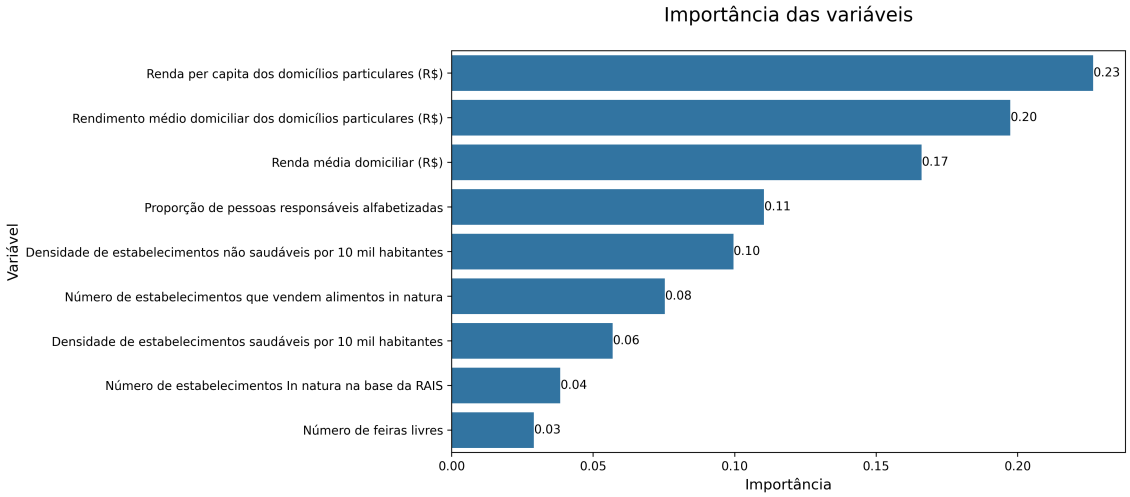

Figura 3. Importância das variáveis no modelo de *Random Forest*.

A vulnerabilidade social, mensurada por meio do IPVS, constitui um indicador fundamental para a compreensão da relação com a fome, uma vez que reflete as condições socioeconômicas adversas que dificultam o acesso adequado à alimentação [Silva et al. 2020]. O IPVS abrange fatores como renda, emprego e acesso a serviços públicos, os quais estão diretamente relacionados à segurança alimentar. Nas regiões com maior vulnerabilidade social, observa-se uma incidência mais elevada de fome, em razão da escassez de recursos e do suporte social, que comprometem a capacidade das famílias de assegurar uma alimentação suficiente e de qualidade.

Os resultados alcançados confirmam que a renda das famílias é o principal fator de vulnerabilidade e, consequentemente, à insegurança alimentar. Essa situação está ligada a um baixo status socioeconômico, refletindo-se em padrões de consumo alimentar inadequados e em um estado nutricional precário. Famílias com baixa renda mensal,

nível educacional reduzido e altos índices de desemprego enfrentam limitações significativas na aquisição de alimentos. Dessa forma, a pobreza emerge como a principal causa desse fenômeno, tornando a insegurança alimentar um problema recorrente entre lares vulneráveis, que correm o risco de enfrentar um suprimento alimentar doméstico insuficiente [Sharif and Ang 2001].

Estudos anteriores, como o de [Furness et al. 2004], também evidenciam uma forte correlação entre a insegurança alimentar e a renda, demonstrando que famílias com rendimentos mais baixos apresentam uma probabilidade significativamente maior de enfrentar dificuldades no acesso a alimentos. Além da baixa renda, famílias que incluem crianças ou que apresentam histórico de desabrigamento tendem a enfrentar maior vulnerabilidade. Esses fatores reforçam a hipótese de que restrições financeiras constituem um impedimento ao acesso a uma alimentação adequada, contribuindo, assim, para a perpetuação da insegurança nutricional.

4. Conclusão

Neste trabalho, foram aplicados algoritmos de aprendizado de máquina em dados socioeconômicos das regiões da cidade de São Paulo para analisar a vulnerabilidade social no contexto de estudos sobre fome e insegurança alimentar. Os modelos gerados atingiram um F-score médio de 0,84, indicando um bom nível de predição. Além disso, os fatores relacionados à renda domiciliar demonstraram ter um papel fundamental nessa questão.

Trabalhos futuros podem realizar a ampliação e diversificação das bases de dados utilizadas, visando a criação de um conjunto de dados mais robusto e equilibrado. Deste modo, os modelos podem apresentar melhores desempenhos, o que permite aprimorar a compreensão do problema abordado e contribuir para o desenvolvimento de indicadores mais eficazes para o combate da insegurança alimentar.

Referências

Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. (2008). Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5):1241–1251.

Fan, Y., Xuan, L., Qifeng, Z., and Linkai, L. (2011). Margin based variable importance for random forest. In *2011 6th International Conference on Computer Science Education (ICCSE)*, pages 1361–1366.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

Furness, B. W., Simon, P. A., Wold, C. M., and Asarian-Anderson, J. (2004). Prevalence and predictors of food insecurity among low-income households in los angeles county. *Public Health Nutrition*, 7(6):791–794.

Imron, M. A. and Prasetyo, B. (2020). Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn. *Journal of Soft Computing Exploration*, 1(1):56–62. Accessed: 30 Aug 2024.

Kolisetty, V. and Rajput, D. (2019). A review on the significance of machine learning for data analysis in big data. *Jordanian Journal of Computers and Information Technology*, 06:1.

Pérez-Escamilla, R. and Segall-Corrêa, A. M. (2008). Food insecurity measurement and indicators. *Revista de Nutrição*, 21:15s–26s.

Riyanto, S., Sitanggang, I., Djatna, T., and Atikah, T. (2023). Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, 14.

Sharif, Z. and Ang, M. (2001). Assessment of food insecurity among low income households in kuala lumpur using the radimer/cornell food insecurity instrument - a validation study. *Malaysian journal of nutrition*, 7:15–32.

Silva, M., Raposo, I., Silva, L., Assunção, J., Rolim, T., Souza, A., and Franco, F. (2020). *VULNERABILIDADE SOCIAL, FOME E POBREZA NAS REGIÕES NORTE E NORDESTE DO BRASIL*, pages 1083–1105.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846.

Yuanyuan, S., Yongming, W., Lili, G., Zhongsong, M., and Shan, J. (2017). The comparison of optimizing svm by ga and grid search. In *2017 13th IEEE International Conference on Electronic Measurement Instruments (ICEMI)*, pages 354–360.