**Projeto de Pesquisa e Planejamento de Atividades**

|  |  |  |
| --- | --- | --- |
| **Aluno**: Osvaldo Luiz dos Santos Pereira | | **Data início curso**: 26/10/2021 |
| **Orientador**: Thiago Gentil Ramires | | **Defesa em:**  07/2023 |
| **Curso**: MBA Data Science e Analytics | **Modalidade**: Distância | Turma: 212 |

1. **Título do projeto**

*Aplicações pragmáticas de algoritmos de clusterização em instituições financeiras e bancos de varejo.*

Pragmatic applications of clustering algorithms to finance and retail banking.

1. **Introdução**

Algoritmos de *clusterização* como o *K-Means* (MacQuenn, 1967) foram desenvolvidos desde a década de 1960 para serem utilizados em diversos tipos de aplicações, como por exemplo processamento de sinais, através de métodos de quantização vetorial. Desde então com o crescente desenvolvimento de tecnologias em termos de processamento computacional, as aplicações destes tipos de algoritmos se tornaram cada vez mais pragmáticas, tendo inclusive aplicações em diversos nichos da indústria e empresas (Tang et al., 2022). Neste trabalho o objeto de estudo da segmentação de agências bancárias do setor de bancos de varejo em grupos com características comparáveis com o objetivo de atribuir metas equivalentes à estas agências de forma que a apuração destas metas possua comparação justa, ou seja, supondo duas agências em Estados diferentes, mas em contextos similares, e.g níveis de transações, quantidade de clientes, renda média dos clientes, tamanho, contexto geográfico e social, possam se encontrar em clusters semelhantes.

1. **Objetivo**

Este trabalho tem como objetivo principal apresentar aplicações pragmáticas de algoritmos de *clusterização* em instituições financeiras e bancos de varejo, e desenvolvendo metodologias de análises de dados em projetos nesse nicho empresarial. Será demonstrado que métodos de *clusterização* podem ser utilizados como meio de redução dimensional de dados para tomada de decisão em bancos. O case utilizado será a segmentação de agências bancárias em grupos homogêneos e com variáveis comparáveis dentro destes grupos com o objetivo de serem atribuídas metas de performance para cada um dos clusters encontrados. Ao lidar com um número muito grande de agências, e.g cinco mil ou mais agências espalhadas em diversas regiões do Brasil, é impraticável atribuir cinco mil metas customizadas de performance, e apurar estas metas mensalmente dentre todas as agências, com a criação dos clusters, estes grupos de agências apresentarão comportamentos e características de negócio muito semelhantes, então ao invés de atribuir cinco mil metas, este grupo de agências serão reduzidos a apenas alguns clusters.

1. **Material e Métodos**

Algoritmos como o *K-Means* são utilizados em processamento de imagens, computação gráfica, através de quantização vetorial, mas sua aplicação como métodos de *clusterização* consiste em particionar dados em subconjuntos, que no caso do *K-Means* dá o nome ao método, pois devemos determinar a quantidade k de subconjuntos (clusters) a priori. São classificados como algoritmos de aprendizado não-supervisionado, e podem ser utilizados até mesmo em classificação de textos. Exemplos de aplicações científicas de algoritmos de clusterização é a pesquisas de genoma e *clusterização* de genes (Domeniconi et al., 2007)

Para este trabalho serão utilizados dados sintéticos criados através de métodos estatísticos. Estes dados serão baseados em dados reais, seguindo as distribuições observadas em projetos reais. O intuito de utilizar dados sintéticos é de seguir as Leis Gerais de Proteção de Dados e cláusulas legais de não divulgação (NDA). Outro ponto importante é que a demanda por novas metodologias de análise e criação de algoritmos e modelos de aprendizado de máquina está crescendo numa velocidade muito grande, e o acesso à dados cresce na mesma proporção o que causa o aumento do custo ao acesso à estes dados, portanto é comum que artigos metodológicos e até mesmo criação de novas tecnologias utilizem dados sintéticos nos treinamentos de modelos e desenvolvimento de novas metodologias (Raghunathan, 2021), (Mahmoud, 2021), (Hradec et al, 2022).

Os algoritmos de *clusterização* aplicados serão todos desenvolvidos em linguagem Python. Os dados sintéticos serão baseados em dados reais de agências bancárias de um banco de varejo. Cada uma das observações do *dataset* representará uma das agências, e as colunas serão variáveis que representam características, e comportamentos de cada uma destas agências, como por exemplo tamanho físico da agência, quantidade de clientes de diversos segmentos, quantidade de transações em caixa humano, quantidade de transações em caixa eletrônico, tipos de negócios realizados tanto em volume quanto em valores.

1. **Resultados Esperados**

Após a *clusterização* das agências, os grupos nos quais se encontram deverão possuir dispersão mínima (desvio padrão pela média) de cada uma das variáveis de *clusterização*, de forma que estes grupos sejam homogêneos em termos das características operacionais das agências e sejam comparáveis, a ponto de que estratégias e metas possam ser aplicadas de forma igual em todas as agências do grupo. As agências em cada cluster apresentarão tipos de negócios, em volume e valores similares e com baixa dispersão. Cada um dos *clusters* terá características bem distintas.

1. **Cronograma de Atividades**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Atividades planejadas** | **Mês** | | | | | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Revisão bibliográfica | **X** | X | X |  |  |  |  |  |  |  |
| Criação dos dados sintéticos | **X** | **X** | **X** |  |  |  |  |  |  |  |
| Pesquisa por métodos de clusterização | **X** | **X** | **X** |  |  |  |  |  |  |  |
| Análise dos dados clusterizados |  |  | **X** | **X** | **X** |  |  |  |  |  |
| Testes de performance de diversos algoritmos de clusterização |  |  |  | **X** | **X** | **X** |  |  |  |  |
| Desenvolvimento da metodologia de análise de clusters. |  |  |  |  |  | **X** | **X** |  |  |  |
| Relatório pré-TCC |  |  |  |  |  |  | **X** | **X** |  |  |
| Entrega do TCC |  |  |  |  |  |  |  | **X** | **X** |  |
| Entrega da apresentação de defesa |  |  |  |  |  |  |  |  |  | **X** |

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

1. **Referências Bibliográficas**

MacQueen, J. B. (1967). [*Some Methods for classification and Analysis of Multivariate Observations*](http://projecteuclid.org/euclid.bsmsp/1200512992). Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297. [MR](https://en.wikipedia.org/wiki/MR_(identifier)) [0214227](https://www.ams.org/mathscinet-getitem?mr=0214227). [Zbl](https://en.wikipedia.org/wiki/Zbl_(identifier)" \o "Zbl (identifier)) [0214.46201](https://zbmath.org/?format=complete&q=an:0214.46201). Retrieved 2009-04-07.

Geyang Tang, Rujian Tian, Bingdi Wu. An Overview of Clustering Methods in The Financial World. Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022) Atlantis Press, 26 March 2022. https://doi.org/10.2991/aebmr.k.220307.084

C. Domeniconi , D. Gunopulos , S. Ma , D. Papadopoulos, B. Yan. Locally adaptive metrics for clustering high dimensional data. Data Mining and Knowledge Discovery, Volume 14, Issue 1, February 2007, pp 63–97. <https://doi.org/10.1007/s10618-006-0060-8>

T. E. Raghunathan, Synthetic data, Annual Review of Statistics and Its Application, 8, 129-140, 2021.

K. Dankar, I. Mahmoud. Fake it till you make it: guidelines for effective synthetic data generation, Applied Sciences 11.5 (2021): 2158, 2021.

J. Hradec, M. Craglia, M. Di Leo, S. De Nigris, N. Ostlaender, N. Nicholson, [Multipurpose synthetic population for policy applications](https://publications.jrc.ec.europa.eu/repository/handle/JRC128595), EUR 31116 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-53478-5 (online), doi:10.2760/50072 (online), JRC128595, 2022.