Relatório Técnico-Científico - Introdução à Análise Exploratória de Dados

Antônio G. Nascimento[1], Giovanna O. Araujo[1], Gustavo C. Pereira[1], João M. Neto[1]

[1]Departamento de Sistemas de Informação
Centro Universitário de Excelência (UNEX) – Feira de Santana, BA – Brasil

{242030808, 241031333, 241030517, 241030526}@aluno.unex.edu.br

Abstract. *This technical-scientific report presents a computational approach to Exploratory Data Analysis (EDA), integrated into the Knowledge Discovery in Databases (KDD) process, utilizing 13 statistical functions implemented in pure Python. Developed without external libraries or technologies, the project aims at the direct manipulation of data structures from a predefined dataset. The document details the methodological path, ranging from theoretical foundation to the discussion of obtained results, highlighting the effectiveness of algorithmic logic in pattern extraction and decision-making support.*

Resumo. *Este relatório técnico-científico apresenta uma abordagem computacional para a Análise Exploratória de Dados (AED), integrada ao processo de Knowledge Discovery in Databases (KDD), utilizando 13 métricas estatísticas implementadas em Python, sem bibliotecas ou tecnologias externas visando a manipulação direta de estruturas de dados a partir de um dataset pré-definido. O documento detalha o percurso metodológico, desde a fundamentação teórica até a discussão dos resultados obtidos, evidenciando a eficácia da lógica algorítmica na extração de padrões e suporte à tomada de decisão.*

1. Introdução

No que tange à análise de dados, a estatística desempenha um papel fundamental na identificação de possíveis padrões, que se aplicam a regras de negócio pré-estabelecidas. Exemplos abrangem: Dados relacionados a venda, como a média de compra dos clientes em determinado estabelecimento comercial ou a faixa etária de jogadores em determinada plataforma de jogo *on-line*.

Neste relatório, veremos a estatística aplicada ao *Knowledge Discovery in Databases* (KDD), e como definido por [Boente et al. 2010], a análise desse tipo de dados é inviável sem o auxílio computacional, dessa forma será implementada uma biblioteca desenvolvida especificamente para este estudo em Python, com funções que se utilizam de cálculos estatísticos para o processamento de dados.

2. Fundamentação Teórica

Sob a ótica metodológica da análise estatística de dados, são aplicadas 13 métricas, sendo elas:

Métrica	Equação	Comportamento

Média Aritmética	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	Calcula a média aritmética de uma coluna.
Mediana	Ímpar: $\frac{n+1}{2}$ Par: $\frac{X_{n/2} + X_{(n/2)+1}}{2}$	Calcula a mediana de uma coluna.
Moda	Mo	Encontra a moda de uma coluna
Variância	$s^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}$	Calcula a variância populacional de uma coluna.
Desvio padrão	$s = \sqrt{\frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}}$	Calcula o desvio padrão populacional de uma coluna.
Covariância	$Cov(X,Y) = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ (5)	Calcula a covariância entre duas colunas.
Conjunto de itens únicos	{X,W,W,Y,Y,J}	Retorna o conjunto de itens únicos em uma coluna.
Frequência absoluta	$\sum_{i-1}^{k} f_i = n$	Calcula a frequência absoluta de cada item em uma coluna.
Frequência relativa	$f_r = \frac{f_i}{n}$	Calcula a frequência relativa de cada item em uma coluna.
Frequência acumulada	$F_i = \sum_{j-1}^{i} f_j$ (8)	Calcula a frequência acumulada (absoluta ou relativa) de uma coluna; A frequência é calculada sobre os itens ordenados.

Probabilidade condicional	$$P(A\|B) = \frac{P(A \cap B)}{P(B)}$$	Calcula a probabilidade condicional P(X_i = value1 | X_{i-1} = value2). Este método trata a coluna como uma sequência e calcula a probabilidade de encontrar `value1` imediatamente após `value2`.
Quartis	$$IQR = Q3 - Q1 \ (9)$$	Calcula os quartis (Q1, Q2 e Q3) de uma coluna.
Histogramas por buckets	$$k = 1 + 3,322 \cdot \log_{10}(n)$$	Gera um histograma baseado em buckets (intervalos).

3. Metodologia

O projeto foi implementado a partir de uma análise inicial do dataset pré-estabelecido em “tests.py”, arquivo destinado aos testes unitários das métricas desenvolvidas. A distribuição das atividades entre os quatro integrantes da equipe seguiu um critério de afinidade funcional, agrupando métricas por similaridade lógica — como as variações de frequência (absoluta, relativa e acumulada) — para otimizar a reutilização de código e a consistência. O fluxo de trabalho foi sustentado por um ambiente de versionamento distribuído, com a criação de branches individuais para cada integrante, garantindo a integridade da integração das funções ao núcleo da biblioteca estatística.

4. Resultados e Discussões

Os resultados obtidos através da biblioteca desenvolvida utilizando dicionários e listas do Python, simulando a organização de um *dataframe* convencional, demonstraram precisão matemática idêntica aos outputs de bibliotecas consolidadas, como Pandas e NumPy, validando a eficácia da lógica implementada em Python puro. Ao processar o conjunto de dados experimental, observou-se que as métricas de tendência central e dispersão responderam com acurácia às variações dos dados. Diferente da implementação com bibliotecas externas, a implementação manual permitiu um controle sobre a estrutura de dados, garantindo que as manipulações não alterassem o estado original do dataset proposto.

5. Considerações Finais

Em conclusão, foram implementadas as métricas citadas anteriormente, com testes (posteriormente corrigidos), e casos de borda para tratamento de erros. A criação de

uma biblioteca própria mostrou-se como um ponto desafiador ao utilizar Python para métricas complexas. Nesta prática, foi notado que a aplicação de estatística analítica em *Knowledge Discovery in Databases* é essencial para a etapa de mineração de dados em si. Com uma maior janela de tempo, haveria a possibilidade de implementar cálculos e tratamentos de erros mais complexos, além de utilizar variados *datasets*.

6. Referências

Boente, A. N. P., Goldschmidt, R. R. e Estrela, V. V. (2010). Uma Metodologia de Suporte ao Processo de Descoberta de Conhecimento em Bases de Dados. In: SEGeT – Simpósio de Excelência em Gestão e Tecnologia.

Referências

WALPOLE, R. E.; MYERS, R. H.; MYERS, S. L.; YE, K. Probability & Statistics for Engineers & Scientists. 9. ed. Boston: Pearson, 2012.

TRIOLA, M. F. Introdução à Estatística. 12. ed. Rio de Janeiro: LTC, 2017.