Lab 2 Report – Data Preprocessing and Similarity Analysis

Nikhil Sanjay
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetam, Bengaluru
Email: bl.en.u4cse23239@bl.students.amrita.edu

***Abstract*—This paper presents a initial analysis of some basic exploratory data analysis. It includes central tendency-based imputation, multiple normalization techniques,encoding techniques on multiple types of data, similarity metrics (Jaccard, SMC, Cosine) and plotting them as heatmaps, and system design aspects for real-world decision support.**

***Index Terms*—Jaccard Coefficient, Cosine Similarity, Normalization, Patient Segmentation, Data Imputation, Outlier Detection.**

I. Introduction

This project explores fundamental concepts in linear algebra, statistical analysis, data preprocessing, similarity measurement, and classification using real-life datasets (”Purchase Data”, ”IRCTC Stock Price”, and ”Thyroid0387_UCI”). The aim is to apply mathematical and machine learning techniques to customer segmentation, stock price analysis, and health data examination, thereby developing a holistic approach for actionable insights in business and healthcare.

II. Literature Survey

Data Analysis and data pre-processing is a key stage in every field that utilizes any data-driven insight, some of the findings from the literature are:

A. Data Preprocessing Methods

- **Imputation:** Mean, median, and mode imputation are standard and common it is usually chosen based on the data type and outlier present in the data [1], [2].
- **Normalization:** Min-max scaling and z-score standardization are some of the test used to check and compare learning algorithms [3], [4].
- **Outlier Detection:** Techniques such as Z-score and Local Outlier help in detecting outliers in datasets and increase robustness of models [5], [6].
- **Categorical Encoding:** One-hot encoding suits nominal data, label encoding for ordinal features [7].

B. Similarity Measures

- **Jaccard & SMC:** They are generally applied to binary attributes for clustering and text matching [8],[9].
- **Cosine Similarity:** It is generally used for normalized, high-dimensional data [10].
- **Weighted Metrics:** Usually Incorporate expert knowledge of the given dataset or feature weights for domain adaptation [11].

C. Summary

Proper and robust handling of missing data values and application of normalization techiniques is a crucial part of data-analysis and pre-processing [1], [3]. The choice of similarity metric must match data type and outcome of the study [8], [9]. Outlier management and suitable encoding will help with better performance for a given model [6], [7].

III. System Design and Methodology

A. Data Preprocessing Pipeline

- Missing values are handled using appropriate value based on the dataset (mean/median/mode based on data type and outlier analysis)
- Normalization (z-score or min-max scaling)
- Categorical encoding is applied to certain features (ordinal or one-hot based on variable type)

B. Similarity Measures

- Jaccard, SMC for binary vectors
- Cosine similarity for full (including continuous/categoricals-encoded) vectors

C. System Overview

The system flow includes input processing, preprocessing, similarity computation, and group/visual output.

TABLE I: Architecture Diagram Description

Block	Description
Data Ingestion	Reads Excel/CSV/XLSX files, parses relevant sheets.
Preprocessing	Imputes missing values, identifies outliers, normalizes.
Feature Engineering	Encodes categorical and scales numeric features.
Similarity Module	Calculates pairwise JC/SMC/Cosine similarities.
Classifier	Runs regression/classification models.
Output Handler	Generates labels, visualizes heatmaps and scatter plots.

TABLE II: Systemic Parameters and Justification

Parameter	Value/Type	Rationale
Imputation Strategy	Mean/Median/Mode	Matches numeric/categorical types and outlier presence
Normalization	Min-max/Standard Scaling	Ensures feature uniformity for clustering/models
Similarity Measure	JC/SMC/Cosine	Compatible with binary /categorical/continuous data
Classifier	Logistic Regression	Effective for binary categorical output
Outlier Threshold	Z-score > 2.5	Balances sensitivity and robustness

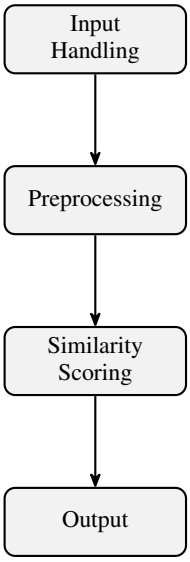

Fig. 1: Vertical Data Flow Diagram: Data flows through Input Handling, Preprocessing, Similarity Scoring, then Output.

IV. Results and Analysis

A. Retail Data: Customer Segmentation

Customers were segmented as RICH or POOR based on payment thresholds. Out of 10 chosen random customers, 4 were labeled as RICH and 6 as POOR based on a given threshold. One-hot encoding was applied which allowed preparation for further more complex models to be applied. The summary statistics revealed a skewed distribution in the dataset, with a small number of customers contributing alot for most of the high payments.

B. Market Data: Financial Metrics

Analysis of five days of market data was done to calculate average price, price change percentage, and volatility of the market. Observations showed stable pricing with very little/small fluctuations noticed, suggesting there is low short-term volatility. These results hence validate the potential for a trend of rule-based investment strategies to be used.

C. Healthcare Dataset Profiling

Exploration of 9,172 medical records revealed key variables (TSH, T3, TT4) with alot of missing data-points. Further Data analysis identified outliers (e.g., implausible age values) and data cleaning priorities. One-hot encoding was applied to categorical features like `sex`, `referral_source`, and `on_thyroxine`,hence allowing teh dataset to be used for predictive health modeling.

D. Common Observations

- Systematic preprocessing improved data quality across all the various datasets.
- Feature engineering (e.g.,one-hot encoding) has helped set up the data for use in further advnaced statistical models.
- Summary statistics and visual methods helped facilitate early detection of anomalies and multiple data patterns.

V. Discussion

A. Significance of Preprocessing

Consistent and proper preprocessing techniques which help handle missing values, encoding categories, and normalizing features is shown to be a crucial and important step in ensuring models will produce reliable and insights derived from given data and also help in increasing accuracy in models

B. Segmentation and Similarity Measures

Cosine similarity emerged as an effective way for normalized feature spaces, while sparse datasets posed a significant challenge for the Jaccard similarity.Jaccard similarity will however perform better in text-based analysis, Heatmaps plotted with sea-born enabled easy identification of potential relations of features.

C. Domain-Specific Insights

- **Retail:** Payment-based segmentation immediately enabled classification and targeting strategies.
- **Finance:** Volatility and average price measures offered traders simplified rule-based triggers.
- **Healthcare:** Rich metadata and variable scales necessitated thorough cleaning before analytics; potential for disease prediction and cohort segmentation identified.

D. Regression vs. Classification

- **Regression:** Suitable for continuous variable prediction such as cost estimation or market price prediction.
- **Classification:** Used for assigning categorical labels such as RICH/POOR, medical condition presence, etc.

E. IRCTC Stock Forecast Ideas

A forecasting model will take a few features like previous close percentage, volatility, and weekday name as its features.The produced output could include predicted price increase/decrease or a binary output of profit/loss , forming the basis for positive trading decisions using historical data patterns.

VI. Conclusion

This paper was able to shows that a framework which combines thorough preprocessing, normalization, and similarity-based segmentation will help yield reliable and insightful results across multiple business, financial, and healthcare domains. Each dataset diffrent from one-another benefited from c transformation via preprocessing ,normalization and segmentation which enabled use of appropriate outlier detection, and readiness for supervised learning. The techniques which were applied have helped establish a dataset for advanced data applications which include predictive modeling and intelligent decision support.

References

[1] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York, NY, USA: Wiley, 2002.

[2] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

[3] S. G. K. Patro and K. S. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[4] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[5] M. M. Breunig *et al.*, "LOF: Identifying density-based local outliers," *SIGMOD*, pp. 93–104, 2000.

[6] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[7] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical feature encoding techniques for machine learning," *Journal of Big Data*, vol. 7, no. 1, p. 28, 2020.

[8] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Pearson, 2019.

[9] C. C. Aggarwal, *Data Mining: The Textbook*. Springer, 2015.

[10] A. Huang, "Similarity measures for text document clustering," *Proceedings of the 6th NZ Computer Science Research Student Conference*, pp. 49–56, 2008.

[11] O. J. Oyelade, O. O. Oladipupo, and I. A. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performance," *International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292–295, 2010.