Linear Discriminant Analysis: Applications and Challenges in Data Analysis

Student: Turcu Ciprian-Stelian

Master: High Performance Computing and Big Data Analytics

E-mail: ciprian.turcu@stud.ubbcluj.ro

Abstract

This study is about the effectiveness of Linear Discriminant Analysis (LDA) in performing dimensionality reduction and classification enhancing on five different datasets, namely Wine Quality, Mall Customers, Heart Disease, Diabetes, and Online Retail. In this paper, we use LDA for enhancing class separability and interpretability of each dataset. As per the analysis, LDA identifies key discriminative features, enabling clear separation among classes. However, LDA performance is based on the assumptions of multivariate normality and homogeneity of variance-covariance. If these assumptions are violated, other methods such as Quadratic Discriminant Analysis (QDA) and Regularized Discriminant Analysis (RDA) may yield better results. The study emphasizes the importance of dimensionality reduction techniques for high-dimensional data and how it could influence the choice of appropriate techniques that are specific to the nature of the data.

Introduction

The technique of Linear Discriminant Analysis (LDA) was picked in this study for its renowned ability in dimensionality reduction and data classification that has a high interpretability. Unlike some of the unsupervised techniques like Principal Component Analysis (PCA), which prioritize variance without regard to class labels, LDA points its focus on finding the maximum separability between predefined categories. Although high accuracy offered by more modern machine learning such as support vector machines (SVM) and neural networks, they can lack transparency on their decisions; LDA offers a mathematical grounded solution that balances a simple and effective decision.

High-dimensional data can often be presented as a challenge in places where the volume of feature space increases exponentially with the number of dimensions, forcing towards sparse data representations and potential overfitting. Dimensionality reduction techniques like LDA, together with other more advanced variants, help alleviate this issue by projecting data onto lower-dimensional spaces that conserve the class separability of the data. With this reduction a great enhancement on computational efficiency can be seen while also improving generalization capability of classifiers.

In this paper, LDA will be applied to five distinct datasets to demonstrate its adaptability and relevance across different domains. The first dataset delves into the relationships of features like alcohol and acidity in order to distinguish wines of varying quality, providing insights into key chemical influences on wine quality. The second dataset has a target of customer clusters generated through K-Means, LDA enabled clear differentiation of customer segments based on their demographics and spending habits. Relationships between features like Spending Score and Annual Income play a big role in understanding customer behaviour patterns. The third dataset containing clinical features like cholesterol and blood pressure were crucial in distinguishing at-risk individuals, uncovering patterns that separate individuals with and without heart disease. The fourth dataset also falls into the health category analysing the major risk factors for diabetes, utilising relationships between glucose levels, BMI, and other health indicators to separate individuals on

their possible diabetes status. The last dataset adds to the information obtained from the second one in terms of LDA efficiency in the domain of customer segmentation. This dataset contained important features like spending habits and product quantities from which we extract regional and behavioural differences among customers.

Theoretical Foundations of Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) represents a major statistical technique heavily used in pattern recognition, machine learning, and data analysis in classifying data points to predefined categories. Originally developed by Sir Roland A. Fisher in 1936, LDA looks to find a linear combination of features that best separates two or more classes of objects or events.

At its core, LDA seeks to project high-dimensionality data onto a lower-dimensional space while keeping the separation between different classes maximized, achieved with modelling the difference between classes while considering variance for each class. One of the key objectives is to find a projection for which the ratio of between-class variance with respect to within-class variance is maximal, therefore ensuring that within the projected space the classes are as distinct as possible. For a dataset that includes multiple classes, LDA calculates the mean vectors for each class as well as the overall mean. It then calculates two matrices: the within-class scatter matrix (S_W), which measures the dispersion of data points within each class, and the between-class scatter matrix (S_B), which measures the scatter between class names. In order to obtain the optimal projection matrix, the solving of the generalized eigenvalue problem for the matrix $S_W^{-1} * S_B$. The eigenvectors that have the largest eigenvalues from the axes of the new feature space, ensure maximum class separability. [1]

Key features and Structure

LDA operates with the several key assumptions:

1. **Multivariate Normality:** Each class is assumed to exert a normal distribution.
2. **Homogeneity of Variance-Covariance:** All classes are assumed to share the same covariance matrix. [2]
3. **Independence:** Each data point in the dataset should be treated as an individual and unrelated observation. Violating the constraint that there should be no direct or systematic relationship between one observation to another one in the dataset, leads to incorrect class boundaries and poor classification performance.

While keeping these assumptions it is easier to determine linear decision boundaries between classes. Linear Discriminant Analysis is structured around the computation of linear discriminant functions, which are linear combinations of input features. These functions allow for the determination of decision surfaces, which partition the decision space into regions corresponding to different classes.

Even if Linear Discriminant Analysis is suitable while the covariance matrices hold, there are situations and scenarios where this assumption is violated, and the classification precision is significantly affected. For such cases more advanced variations of the Linear Discriminant Analysis, can be chosen such as:

1. **Quadratic Discriminant Analysis (QDA):** To overcome the problem of the covariance matrix constraint violation, QDA is used for its class-specific covariance matrices, resulting in quadratic decision boundaries. QDA offers a more flexible technique, but at the cost of

requiring larger sample sizes for reliable estimation due to the increased number of parameters. [3]

2. **Regularized Discriminant Analysis (RDA):** RDA further introduces regularization parameters in order to balance LDA and QDA, providing a middle ground in the case of class covariances differences or limited sized data. [4]
3. **Sparse Linear Discriminant Analysis (Sparse LDA):** Sparse LDA is designed to perform feature selection through regularization methods in order to identify the most discriminative features and, at the same time, reduce overfitting. It is also more interpretable and computationally efficient. In [5] is proposed a sparse LDA algorithm by using thresholding in the estimation of the discriminant vectors to obtain sparsity, outperforming its original version in high-dimensional cases.

Linear Discriminant Analysis comes with great advantages that offer strong usability reasons for data analysis tasks. First key strength is represented by the interpretability of the linear decision boundaries which are straightforward in interpretation, offering a transparent classification method. Furthermore, LDA is computationally less intensive compared to other more complex models, making it suitable in scenarios where the dataset has a large number of entries. At last, projecting the data onto a lower-dimensional space, LDA can reduce the dimensionality of the dataset providing an easier to analyse space.

However, Linear Discriminant Analysis comes with limitations as well. LDA has a high sensitivity to violations of normality and homogeneity of variance-covariance, performance drops as these assumptions are not met. Linear boundaries might not be effective for non-linearly separable data. Lastly as the technique is sensitive to outliers, results may be impacted id the class means and covariances are disproportionate, leading to misclassification.

Real-World Applications of Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) has been shown to be an important tool for dimensionality reduction and data visualization, as well as for pattern recognition, classification, and decision-making with applications in a variety of fields. It has the ability to transform complex data into interpretable components while preserving class separability. We will look into three key sectors where LDA demonstrates its potential:

1. **Healthcare: Enhancing Diagnostics and Preventive Care**

By enhancing patient segmentation, treatment options, and diagnostic accuracy, Linear Discriminant Analysis plays a crucial role in the healthcare industry. LDA helps in early detection and individualized care by analysing clinical data to find patterns that differentiate between illness states.

- **Diabetes Diagnosis:** Healthcare providers are enabled to design targeted solutions and prioritize at-risk groups, ensuring timely care and resource allocation, LDA facilitates the process of classification of patients based on health indicators such as glucose levels and BMI
- **Heart Disease Detection:** Utilising features like blood pressure, cholesterol, and age, LDA separates individuals with hearth disease from healthy patients, supporting precision diagnostics and the development of risk mitigation plans. Moreover, preventive programs can be created based on identifying high-risk populations.

2. **Product Quality Assessment: Refining Standards and Enhancing Consumer Satisfaction**

In industries where the product quality is first, using Linear Discriminant Analysis helps asses and categorize products, enabling continuous improvement and customer satisfaction.

- **Wine Quality Evaluation**: Using chemical properties such as acidity, sugar levels, and alcohol content, we utilize LDA to classify wines into quality categories. In order for winemakers to further refine production processes for maintaining consistency in premium products, identifying discriminative features that influence quality is crucial.
- **Food and Beverage Industry**: In a similar fashion, LDA can analyse chemical and sensory attributes to group products quality or flavour profiles, enhancing the capabilities of brands to tailor offerings to specific market segments or preferences.

3. **Customer Behaviour Analysis: Personalizing Experiences and Driving Engagement**

Customer behaviour is essential for businesses in the design of personalized experiences or better customer engagement. Linear Discriminant Analysis uncovers patterns in customer data aiding the profiling and strategy development.

- **Mall Customer Segmentation:** Entrenched on demographics, spending habits, and income levels, LDA segments shoppers by providing a smaller dimension space of the data that descripts the clusters of high-income, high-spending customers, with potential to premium offerings; lower-income clusters that might utilise targeted budget-friendly promotions. Contributing to the guidance of marketing strategies, loyalty programs, and store layouts.
- **Online Retail Analysis:** Reducing the plane, through LDA, on which features like total spending, purchase frequency, or global location exert their power, enables retailers to identify regional or behavioural differences supporting personalized recommendations, targeted marketing, and optimization of the supply chain.

Processing of Data and Methodologies

In order to sustain this study, we utilized 5 distinct datasets, from 3 topics, in order to explore the effects of Linear Discriminant Analysis, with regards to uncovering patterns, relationships, and insights inside the data. Each dataset required personalized steps for preparation and processing due to its specific structure, features, and objectives. Below we detail the methodologies enforced on each dataset:

1. **Wine Quality Dataset**

The Wine Quality [6] dataset consists of information on various chemical properties of wines, such as acidity, residual sugar, pH, and alcohol content, alongside a target variable pointing to the wine quality on a scale from 3 to 8. The processing of this dataset started by loading the dataset using pandas [7] in order to review its structure and check for missing values or inconsistencies within the entries. The structure allowed for no encoding, due to all of the features being numerical. In order to ensure that features such as alcohol content and residual sugar contribute equally to the analysis, the data was standardized using the StandarScaler from the scikit-

learn [8] library this standardization was a necessary step for the processing of this dataset as features measured on different scales could disproportionately influence the results.

For the selection of the features, all chemical attributes were retained as predictors, while the quality rating served as the target variable. Our primary goal was to utilize Linear Discriminant Analysis to reduce the dimensionality of the dataset while maximizing the separability of wine samples with regards to the quality ratings of the wine. Establishing the number of components for LDA was determined by

$$n_{components} = \min(d, c - 1)$$

where d is the number of features and c is the number of unique wine quality classes. The resulting transformation maps the data onto a lower-dimensional space where the separability between quality categories is maximized.

2. Mall Customer Dataset

The dataset [9] captures demographic and behavioural data, including gender, age, annual income, and spending score. As some categorical features were identified such as gender, the need for encoding the categorical data was revealed, using LabelEncoder [8] this data was transformed into numerical values in order to make it suitable for analysis. As a further data improvement, all numerical features were scaled using the StandardScaler to ensure uniformity and avoid any feature from influencing the results based on its scale.

As this dataset did not contain a clear target feature, the data was segmented into clusters using K-Means clustering. The Clustering of the customers assigned each one to one of five clusters based on their demographic and behavioural patterns. Using these newly created clusters as the target labels for LDA. Transforming the 4 original features into 2 discriminant components eases the visualization and interpretation on the separability of customers segments.

3. Hearth Disease Dataset

Approaching the medical field, we analysed a dataset [10] containing clinical data, observing features such as cholesterol levels, blood pressure, and age, together with a binary target variable indicating the presence or absence of hearth disease. Basic processing procedures were used, such as ensuring completeness of the dataset with no missing values and using the LabelEncoder to encode categorical variables converting them to numerical values.

All numerical features were scaled using StandardScaler in order to ensure that variables such as cholesterol and blood pressure are using the same scale. LDA was utilized on the features of this dataset in order to obtain the target variable which serves as the output class. The high-dimensional space was transformed into a one-dimensional space that maximized the separation between patients with and without hearth disease.

4. Diabetes Dataset

Continuing inside the medical field this dataset [11] focuses on a different major problem in public health. It takes into account features such as glucose levels, BMI, insulin levels, and age, along with a binary target variable indicating whether a patient has diabetes. The preprocessing focused more on scaling the features, rather than encoding or dealing with missing value problems as the dataset utilized only numeric features and none of the data was missing. The StandardScaler was used to ensure that all variables contributed equally to the analysis, particularly since health indicators like glucose and insulin levels are measured on different scales.

The features were used as predictors, while the binary diabetes outcome served as the target variable. LDA was employed to reduce the dataset to a single discriminant component, capturing the maximum separability between diabetic and non-diabetic patients.

5. Online Retail Dataset

The last dataset [12] studied contains transactional data, including product quantities, nuit prices, and customer locations. Preparation of this data included removing rows with missing CutomerID values, in order to ensure the integrity of the analysis. As the data permitted a new feature, TotalPrice was created by multiplying the quantity purchased with the unit price, emphasizing the total value of transactions for each customer. The data was then aggregated at the customer level, summarizing total quantity, total spending, and the country of the first transaction for each customer.

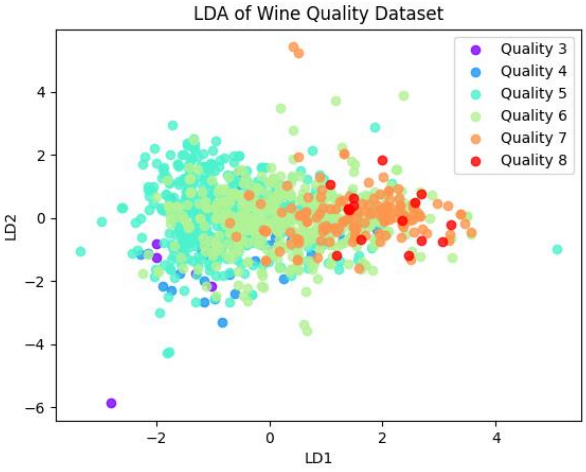
Since the country's attribute is categorical, LabelEncoder was used to encode it and give each country a numerical value. To guarantee consistency, StandardScaler was used to standardize all numerical features. Based on their purchase patterns, customers were then divided into five groupings using K-Means clustering. LDA, which was used to lower the dataset's dimensionality while maintaining the separability of customer segments, used these clusters as the target labels.

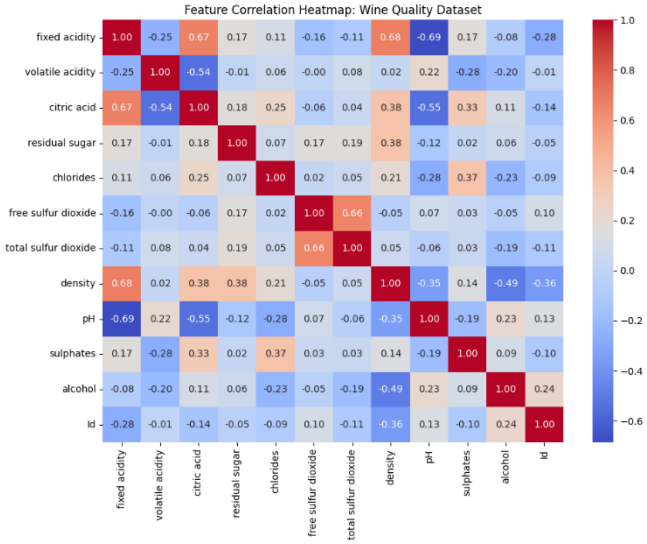
Results

In the first experiment, we applied Linear Discriminant Analysis to the “Wine Quality” dataset. In this context both of the discriminant components are represented by linear combinations of the chemical properties (e.g., acidity, residual sugar, pH, alcohol content). These components capture the variation in the dataset that is most effective at separating wines of different quality levels. In the following we present the obtained results:

- **Discriminant Components:**
  - **LD1:** Emphasizes attributes like alcohol content and acidity, distinguishing higher-quality wines (quality 7 and 8) from lower-quality ones (quality 3 and 4) based on their chemical balance.
  - **LD2:** Captures subtle variations, such as the influence of residual sugar and pH, particularly useful in differentiating medium-quality wines (quality 5 and 6).
- **Segmentation Representation:** The dataset was segmented resembling clusters, grouping wine samples based on their quality ratings. Each cluster emulates the exchange of chemical proprieties and corresponding quality levels:
  - **Quality 3 & 4 (Purple, Blue):** Found at lower LD1 and LD2 values, describe wines with suboptimal chemical profiles, likely marked by imbalanced acidity and lower alcohol content.
  - **Quality 5 (Cyan):** At a centre position in the plot, this group represents standard-quality wines with average chemical, reflecting a balanced yet unremarkable profile.
  - **Quality 6 (Light Green):** Skewed towards higher LD1, this segmentation proposes wines with moderately enhanced chemical profiles, showing improved characteristics such as better acidity balance.

- **Quality 7 & 8 (Orange, Red):** Located at higher LD1 and LD2 values, they promote premium-quality wines. Attributes like higher alcohol content and well-balanced pH distinguish these groups as superior.



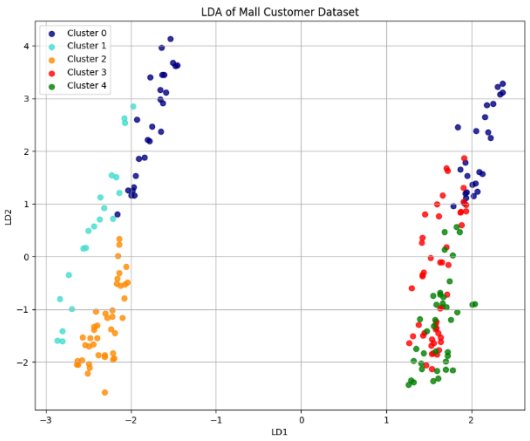


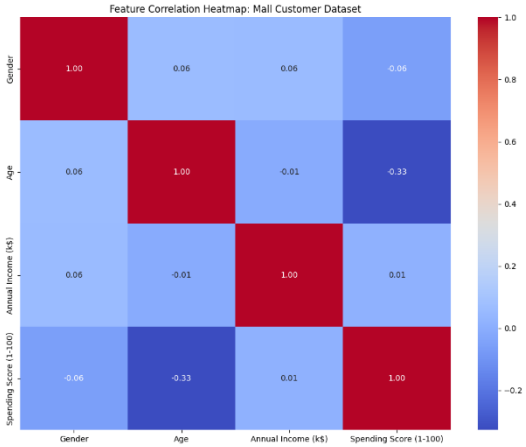
In the heatmap of the feature correlation, we can observe important correlations among chemical features influencing the wine quality. Fixed acidity correlates positively with density (0.68), highlighting intrinsic chemical balance into determining wine structure. There are week correlations as well such as the alcohol content, with most features, confirming an independent role in quality analysis. The pH shows a negative correlation with fixed acidity (-0.69), consistent with expected chemical interactions in acidity regulation.

For the second experiment, LDA was applied to the “Mall Customers” dataset. Linear combinations of demographic and expenditure characteristics (e.g., gender, age, annual income, spending score) are used to represent both discriminant components. The variation in the dataset that best divides up consumer groups is captured by these elements. Here are the results we were able to obtain:

- **Discriminant Components:**
  - **LD1:** Gives insight into spending behaviour relative to the income, differentiating high-income customers with conservative spending from those with higher spending habits.
  - **LD2:** Captures information in the demographic space, such as the influence of age and gender on spending patterns, further aiding in the segmentation of customer profiles.
- **Segmentation Representation:** Data was segmented into clusters, grouping customers based on their demographics and spending scores. Each cluster reflects specific behavioural and demographic traits:
  - **Cluster 0 (Blue):** Oriented toward higher LD1 values, it likely represents high-income customers that prioritize financial prudence through frugal expenditure.
  - **Cluster 1 (Light Blue):** Positioned around lower LD1 values, this group reflects customers of which behaviour shows balanced spending, aligning expenditures with moderate income levels.

- **Cluster 2 (Orange):** Tightly clustered at negative LD2 values, these customers are likely younger with moderate income levels, demonstrating spending patterns influenced by lifestyle choices.
- **Cluster 3 (Red):** Spread along higher LD1 and LD2 values, this cluster represents high-income customers with higher spending scores, indicative of premium spending behaviour and a preference for luxurious lifestyles.
- **Cluster 4 (Green):** Located at lower LD1 and LD2 values, this group likely includes customers with lower income and minimal spending, reflecting limited financial flexibility.



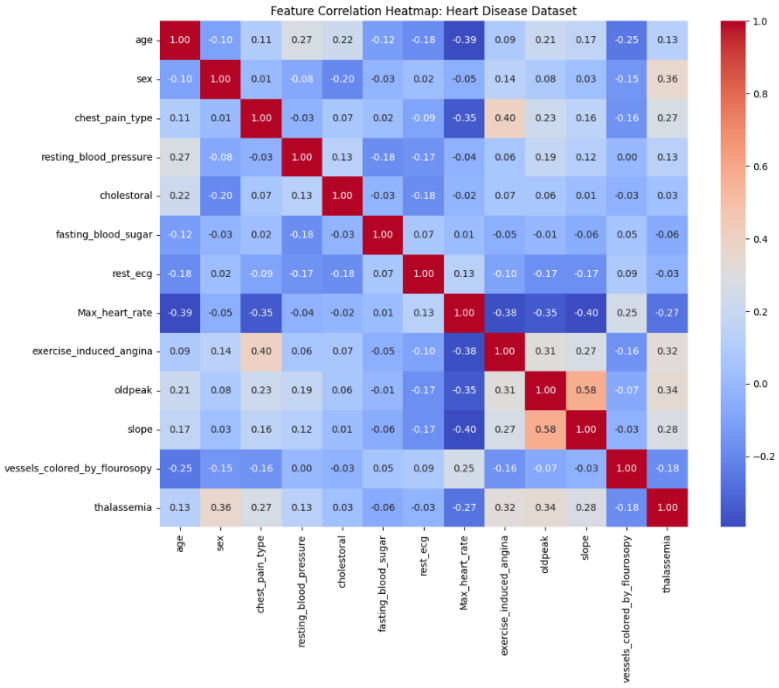


In the Mall Customer dataset heatmap we highlight existing relationships between spending and demographic features. The spending score has a negative correlation with age (-0.33), indicating that younger customers have a tendency to spend more, a trend driven by lifestyle preferences. As for the rest of the features, gender, age, and annual income, they show weak correlation between themselves, boosting independence of demographic traits from spending activity.

For the third experiment, we applied LDA of the "Hearth Disease" dataset. Considering this dataset represents a binary classification problem, only one discriminant component LD1 is derived. This component represents a combination of clinical features (e.g., cholesterol, blood pressure, age) that maximizes the separation between individuals with and without heart disease. Below, we present the obtained results:

- **Discriminant Component:**
  - **LD1:** Offers higher importance to clinical features such as cholesterol and blood pressure, which are critical indicators of heart disease. This component effectively establishes a linear boundary, separating healthy individuals from those at risk.
- **Segmentation Representation:** As this is a linear boundary the data was segmented into 2 classes, based on presence or absence of hearth disease.
  - **Class 0 (Red):** Individuals without heart disease are closely clustered at lower LD1 values. This group is characterized by lower-risk features such as optimal cholesterol and blood pressure levels.
  - **Class 1 (Blue):** Individuals with heart disease are more scattered at higher LD1 values. This class reflects deviations from healthy ranges, including high

cholesterol, irregular heart rates, or elevated blood pressure, indicative of increased risk.

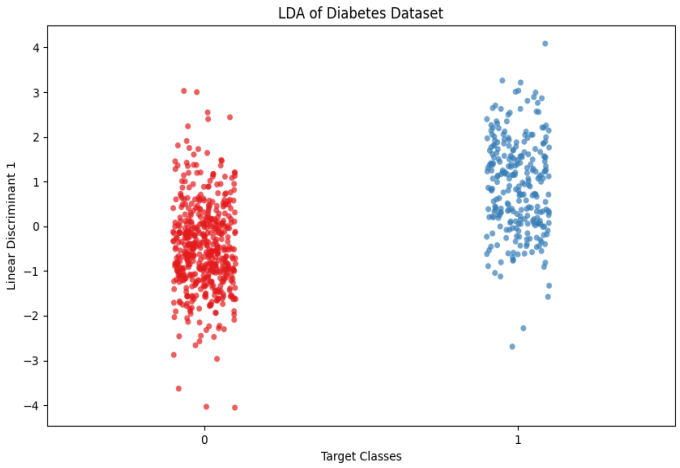


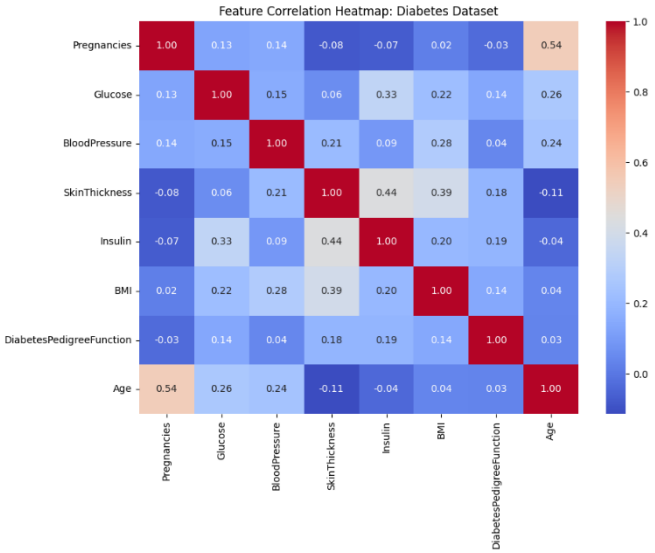
Studying the heatmap of the Hearth Disease dataset we gain critical insights of relationships among the present clinical features that contribute to cardiovascular health. One of the significant relationships is the one between Exercise-induced angina moderately correlating with chest pain type (0.40), reflecting shared responses to physical exertion that are symptomatic of underlying heart conditions. Max hearth rate negatively correlates with age (-0.39), emphasizing the natural drop in cardiovascular capacity as individuals age. As for mostly independent from most of the other features we have cholesterol and fasting blood sugar, having a relatively independent influence on hearth disease risk.

For the fourth experiment, we utilized Linear Discriminant Analysis on the “Diabetes” dataset. Similar to the third this one is also expressed as a binary classification problem, producing only one discriminant component LD1. Based on a combination of health indicators such as glucose, BMI, insulin levels, the component maximizes separation between diabetic and non-diabetic patients. In the following we present the obtained results:

- **Discriminant Component:**
  - **LD1:** Highlights crucial health indicators, such as glucose levels and BMI, which are strong predictors of diabetes. Effectively separating individuals based on the likelihood of having diabetes, emphasizing the most significant diagnostic factors.
- **Segmentation Representation:** The grouping of individuals was segmented into 2 classes, grouping individuals based on their diabetes diagnosis.
  - **Class 0 (Red):** Defines individuals without diabetes, closely located at lower LD1 values. This group is characterized by features like normal glucose levels and BMI, indicating a lower risk of diabetes.
  - **Class 1 (Blue):** Shows people with diabetes, who are scattered distinctly along higher LD1 values. This group includes individuals with elevated glucose levels,

higher BMI, and irregular insulin levels, reflecting a higher risk and presence of diabetes.



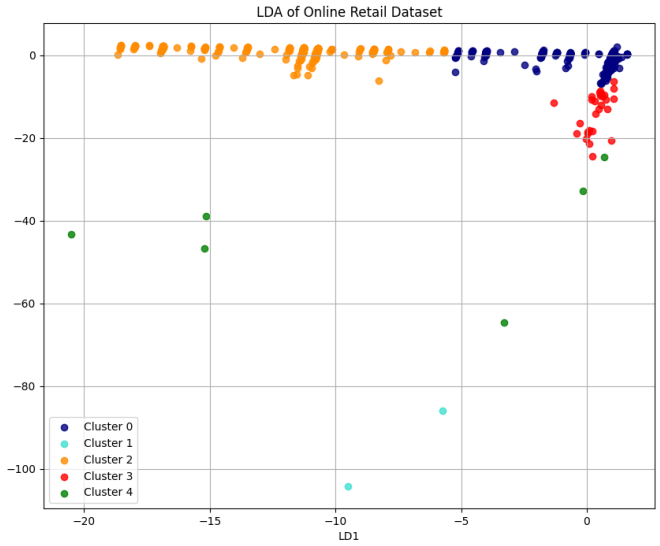


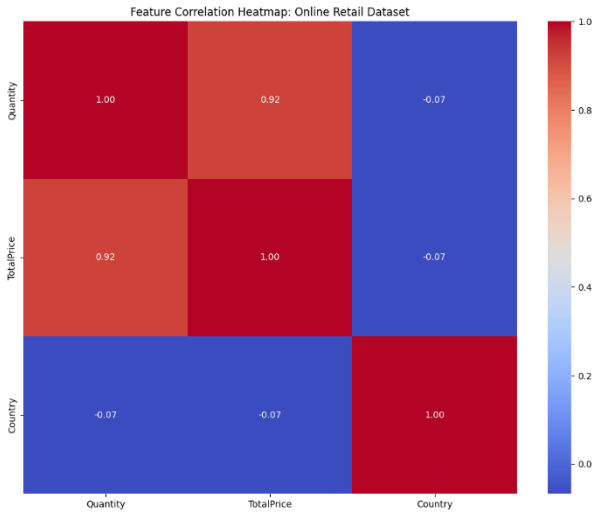
In the Diabetes dataset features like glucose show moderate positive correlations with insulin (0.33) and BMI (0.22), accentuating the metabolic coaction between these factors in diabetes risk. One demographic trend we can extract from the heatmap is that, older individuals have a higher pregnancy count, as the correlation between the features age and pregnancies is at a moderate level (0.54), which may influence diabetes risk indirectly. Skin thickness and blood pressure have weak correlations with other features, indicating their more isolated roles in the dataset.

Lastly in the fifth experiment, we employ LDA on the “Online Retail dataset”. The analysis produced two discriminant components (LD1 and LD2), which are linear combinations of transactional features (e.g., quantity purchased, total spending, customer location) that effectively separate customer segments. Below, we present the obtained results:

- **Discriminant Components:**
  - **LD1:** Emphasizes key transactional features like total spending and purchase frequency, distinguishing high-value customers from low-value ones.
  - **LD2:** Captures geographical variations influenced by the encoded country feature, reflecting differences in purchase behaviour.
- **Segmentation Representation:** The dataset was segmented into distinct clusters, each grouping customers based on their purchasing behaviour:
  - **Cluster 0 (Blue):** Balanced purchases and spending habits are representative for this group of customers, with higher LD1 values. This group indicates towards predictable and steady purchasing behaviour.
  - **Cluster 1 (Light Blue):** Scattered around lower LD1 and LD2 values, this group is characterized by more irregular purchasing patterns or outliers with inconsistent spending.
  - **Cluster 2 (Orange):** With medium to high purchasing behaviour, this group is concentrated near LD1 values that are close to 0. This activity with consistency in total spending, suggests the individuals are regular and engaged shoppers.

- **Cluster 3 (Red):** Spread along lower LD2 values, this cluster suggests customers with high total spending but fewer transactions, potentially representing bulk buyers or luxury shoppers.
- **Cluster 4 (Green):** Scattered towards the farthest regions in LD1 and LD2, this group contains the outliers or customers that do not fit a standard purchasing behaviour, such as low-frequency but high-spending buyers.





As the data in the Online Retail Dataset was aggregated, our heatmap does not have a lot of features to correlate but it contains important information on the correlation of the aggregated data. Relationships between key transactional features. A strong positive correlation (0.92) between quantity purchased and total spending reflects the natural connection where higher purchase volumes lead to greater total expenditures. This correlation underscores the central role of these features in understanding customer behaviour. Interestingly, the encoded country feature shows negligible correlations with both quantity purchased (-0.07) and total spending (-0.07). This suggests that geographical location has little direct impact on purchasing patterns, emphasizing that customer behaviour is more strongly driven by transactional activities than by regional differences.

Overall, the heatmap reveals that while total spending and quantity purchased are tightly coupled, customer location does not significantly influence the primary transactional features. This insight can guide further analysis and strategies focused on customer behaviour, ensuring a stronger emphasis on transactional attributes rather than regional segmentation.

Conclusion

The application of Linear Discriminant Analysis across a range of datasets demonstrates its efficiency in improving class separability and interpretability. The LDA method efficiently performs dimensionality reduction while retaining all the important discriminative information, thus being a very useful technique in data analysis. However, it suffers from assumptions like multivariate normality and equal covariance matrices, which may not hold in every context. For example, when these two conditions do not hold, alternatives like Quadratic Discriminant Analysis (QDA) and Regularized Discriminant Analysis (RDA) relax some of the constraints and introduce more flexibility by allowing different covariance structures for each class, including regularization to handle issues of small sample sizes and outliers. The recent developments in discriminant analysis

have overcome various obstacles like the small sample size problem, noise, outliers, and non-multimodality of class data. These developments led to the realization of robust variants of LDA, hence further improving its performance for a wide range of applications.

The results obtained in this work underscore the strength of LDA in reducing dimensionality while maintaining critical patterns in the data. For example, the reduced dimensionality in the Wine Quality dataset brings out important chemical properties that drive the differentiation of quality. Similarly, dimensionality reduction in the Heart Disease and Diabetes datasets allowed for the determination of such important health indices that would be crucial for the purpose of classification, including cholesterol and glucose levels and BMI. This transformation of complex data into an interpretable format enables clear segmentation using LDA and provides actionable insights in diverse domains.

In summary, though LDA is an important technique for dimensional reduction and classification, the choice of which to use depends on the nature of the data and the assumptions behind each method. Further work can be done by looking into more advanced discriminant methods such as Sparse LDA. These methods could also be integrated into hybrid neural network architectures that would provide both real-time and explainable classification in applications where efficiency and transparency are necessary.

Bibliography

[1] P. P. P. T. T. Xanthopoulos, "Linear Discriminant Analysis," *In: Robust Data Mining. SpringerBriefs in Optimization,* 2013.

[2] "GeeksForGeeks," [Online]. Available: https://www.geeksforgeeks.org/ml-linear-discriminant-analysis. [Accessed 4 11 2024].

[3] Benyamin Ghojogh and Mark Crowley, "Linear and Quadratic Discriminant Analysis: Tutorial," *CoRR,* 2019.

[4] Qiao, Zhihua, Lan Zhou and Jianhua Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data.," *IAENG International Journal of Applied Mathematics 39.1,* 2019.

[5] Jun Shao, Yazhen Wang, Xinwei Deng and Sijian Wang, "April 2011," *Annals of Statistics,* vol. 39, no. 2, 2011.

[6] yasserh, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset/data. [Accessed 10 11 2024].

[7] "Pandas," [Online]. Available: https://pandas.pydata.org/. [Accessed 4 11 2024].

[8] "PyPi," [Online]. Available: https://pypi.org/project/scikit-learn/. [Accessed 4 11 2024].

[9] V. Choudhary, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python/data. [Accessed 10 11 2024].

[10] K. Gangal, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci/data. [Accessed 10 11 2024].

[11] M. Akturk, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data. [Accessed 10 11 2024].

[12] yasserh, "Kaggle," [Online]. Available: https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset/data. [Accessed 10 11 2024].