Exploring Dimensionality Reduction and Ensemble Learning: A Comparative Analysis of Classification and Clustering Strategies on Patient Survival Prediction and Customer Segmentation Datasets

Authors: Valery Garcia, Annmary Lopez, Jaynam Maheshbhai Panchal
CSC 272: Machine Learning
December 18, 2024

Abstract – Unsupervised learning is the branch of machine learning in which computers set out to uncover hidden patterns, structures, and relationships within the feature vectors of data without the provision of truth labels. Therefore, these algorithms are often utilized to enhance the feature representation and performance of supervised learning algorithms. Either in conjunction or as a stand-alone supervised learning methodology, ensemble methods can be implemented to improve predictive performance via the combination of several individual predictive models.

Introduction

The purpose of this project is to explore ensemble methodologies and evaluate the effectiveness of unsupervised learning strategies on two classification datasets of differing characteristics. Specifically, the dimensionality reduction algorithms Principal Component Analysis (PCA) and Uniform Manifold Approximation Projection (UMAP) will be gauged for their ability to enhance model accuracy and robustness. Whereas, the clustering methods K-means and Hierarchical Clustering will be utilized to assess the amount of information that is retained in the dimensionally reduced datasets. Then, the efficacy of the AdaBoost and Random Forest algorithms under various conditions will be examined by constructing ensemble models using both the original and reduced datasets. One dataset involves the segmentation of automotive industry clientele into four classes based on a small number of features, while the other predicts whether a patient will survive their hospital visit based on a multitude of factors. The performance of the aggregated models will be analyzed to determine the affinity of the two ensemble methodologies with the datasets and evaluate the effectiveness of the chosen dimensionality reduction algorithms. Lastly, insights will be drawn and potential improvements will be discussed.

Datasets

Dataset 1 – *Customer Segmentation*

The Train Dataset is designed to analyze and predict consumer segmentation based on demographic, professional, and behavioral factors. It includes features such as age, gender, marital status, profession, spending habits, and family size. The target variable, Segmentation, categorizes individuals into predefined groups, offering insights into their consumer behaviors. This data could be instrumental in tailoring marketing strategies, personalizing customer experiences, and enhancing product development efforts to meet the specific needs of diverse consumer profiles.

Dataset 2 – *Patient Survival Prediction*

The Patient Survival Prediction Dataset focuses on understanding critical care outcomes by analyzing patient demographics, medical histories, and ICU-specific metrics. With variables like age, BMI, APACHE scores, and ventilator usage, the dataset seeks to predict the likelihood of in-hospital mortality. This information can be pivotal in healthcare decision-making, enabling providers to identify high-risk patients, optimize ICU resources, and implement targeted interventions to improve survival rates and overall patient care.

Methods

Data Processing

In working with the Patient Survival Prediction Dataset, our initial focus was on addressing missing data and removing low-utility features. Columns like height, age, bmi, and weight contained critical but incomplete data, which we

addressed through mean and median imputation to preserve dataset integrity. Unnecessary columns such as encounter_id and patient_id were dropped, alongside the zero-variance feature gcs_unable_apache. After this cleaning, the dataset retained 61,097 observations with 117 features – ensuring its readiness for analysis.
For the Customer Segmentation Dataset, I applied mode imputation to fill gaps in categorical features like Ever_Married, Work_Experience, and Family_Size. Binary variables were efficiently encoded, while multi-class variables like Profession were transformed using LabelEncoder. To mitigate the effects of outliers, we scaled the data using RobustScaler.

Partitioning and Metrics

We partitioned both datasets using the 80:20 training-test split methodology, enabling robust evaluation. For clustering, we evaluated silhouette scores, while for classification, we used accuracy, precision, recall, F1-score and ROC-AUC metrics to assess model performance.

Analysis

The Clustering Analysis on the Original Datasets revealed limitations in both the Patient Survival Prediction Dataset and the Customer Segmentation Dataset. High dimensionality and complex data structures hindered the performance of clustering algorithms such as K-Means and Agglomerative Clustering, leading to low silhouette scores and poorly defined clusters. Without dimensionality reduction, it was difficult to extract meaningful insights, as the clustering results lacked cohesion and interpretability.

To address these challenges, we conducted a Dimensionality Reduction Analysis using PCA and UMAP for both datasets. PCA was effective at retaining global variance, explaining 90% of the variance in the Patient Survival Dataset with 21 components and in the Customer Segmentation Dataset with 6 components. However, PCA struggled to capture non-linear relationships, which resulted in less distinct clusters. In contrast, UMAP demonstrated superior performance by preserving both local and global data structures, generating clusters with higher silhouette scores. UMAP's ability to simplify complex relationships was especially beneficial in the Patient Survival Dataset, where it achieved a silhouette score of 0.518 for three clusters, and in the Customer Segmentation Dataset, where it supported the identification of four distinct clusters.

The Clustering Analysis on the Reduced Datasets highlighted significant improvements. For both datasets, clustering algorithms performed better on UMAP-reduced data compared to PCA-reduced data. In the Patient Survival Dataset, UMAP combined with K-Means identified three well-defined clusters, representing potential stratifications in survival risk profiles. Similarly, in the Customer Segmentation Dataset, UMAP paired with Agglomerative Clustering produced four cohesive clusters, aligning with customer segmentation patterns and behavioral profiles. These results demonstrate the critical role of UMAP in enhancing clustering outcomes for high-dimensional datasets.

In the Hyperparameter Tuning phase, we evaluated both AdaBoost and Random Forest across the original, PCA-reduced, and UMAP-reduced datasets. AdaBoost struggled with both datasets, particularly on multi-class classification tasks in the Customer Segmentation Dataset and on imbalanced data in the Patient Survival Dataset. Its reliance on the SAMME algorithm limited its ability to handle the complexity of these datasets. On the other hand, Random Forest consistently outperformed AdaBoost on all transformations, achieving its best performance on UMAP-reduced data for both datasets. For the Patient Survival Dataset, Random Forest achieved a cross-validation F1 score of 0.79 with an optimal max depth of 10–15 and 50–100 estimators. Similarly, for the Customer Segmentation Dataset, it achieved an F1 score above 0.85 with similar hyperparameters, balancing model complexity and generalization. These findings underscore the importance of selecting effective dimensionality reduction techniques, such as UMAP, and tuning hyperparameters to optimize clustering and classification performance for diverse datasets.

Clustering analysis further validated UMAP's superiority. Using K-Means, we identified three optimal clusters on UMAP-reduced data. These clusters were validated by the elbow method and achieved the highest silhouette score, highlighting meaningful stratifications in patient profiles. In comparison, clusters formed on PCA-reduced and original datasets were less cohesive. Hierarchical Clustering, performed with the ward linkage method, confirmed the findings of K-Means and provided additional hierarchical insights into the data structure through dendrograms.

In the classification phase, Random Forest consistently outperformed AdaBoost across all data transformations. Random Forest applied to UMAP-reduced data achieved the highest accuracy (72%) and a ROC-AUC score of 0.79, using optimized hyperparameters (n_estimators=50, max_depth=10). These results highlighted Random Forest's ability to handle complex data structures effectively. While AdaBoost showed moderate performance, its reliance on the SAMME algorithm limited its ability to handle the complexities of the dataset, particularly in multi-class scenarios. Confusion matrices and ROC curves further shows Random Forest's robustness as the preferred classification model. These were all the major insights for Patient Survival Prediction Dataset

For Customer Segmentation Dataset, Dimensionality reduction played a significant role in identifying patterns within the data. Using PCA, we retained six components, which explained 90% of the variance. However, while PCA effectively reduced the dataset's dimensionality, the clusters it produced were less distinct. In contrast, UMAP, optimized with parameters (n_neighbors=10, min_dist=0.25), reduced the data to two dimensions, producing well-separated clusters with higher silhouette scores and trustworthiness metrics compared to PCA. The scatter plots of UMAP-reduced data revealed clear and interpretable groupings, which were critical for subsequent clustering analysis.

Clustering performance was evaluated using K-Means and Agglomerative Clustering. K-Means applied to UMAP-reduced data identified four optimal clusters, validated using the elbow method and silhouette scores. The clustering results demonstrated meaningful stratifications, reflecting actionable customer segments. Agglomerative Clustering, using the ward linkage method, corroborated the results of K-Means and provided additional hierarchical insights into the groupings.

In the classification phase, Random Forest consistently outperformed AdaBoost. Random Forest applied to UMAP-reduced data achieved an accuracy of 85% and a ROC-AUC score above 0.85, using optimal hyperparameters (n_estimators=200, max_depth=20). These results underscored the model's robustness in handling multi-class classification tasks. AdaBoost, while competitive, showed limitations in handling multi-class probabilities, particularly on PCA-reduced and original datasets. However, it performed moderately well on UMAP-reduced data, albeit still falling short of Random Forest's performance.

Results

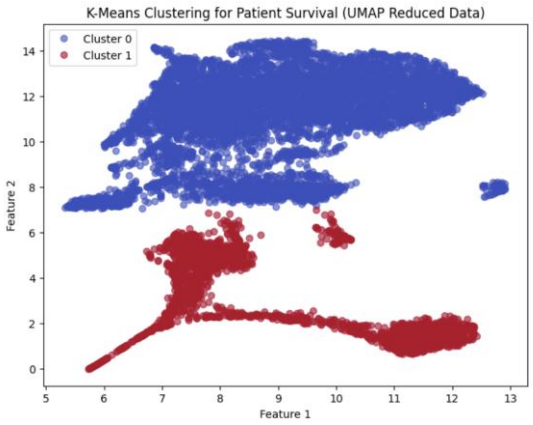
Patient Survival Prediction

Silhouette Scores Table for Patient Survival (UMAP Reduced Data):

Number of Clusters	Silhouette Score
2	0.560
3	0.518
4	0.443
5	0.456
6	0.475

This table illustrates the silhouette scores for different numbers of clusters in the dataset when applying clustering techniques. A silhouette score measures how well each data point is grouped within its cluster, with higher scores indicating better-defined and more cohesive clusters. For this analysis, the optimal number of clusters appears to be 2, as it achieves the highest silhouette score of 0.560, suggesting the best balance between intra-cluster similarity and inter-cluster separation. The scores gradually decrease for higher cluster numbers, indicating diminishing cluster cohesion.

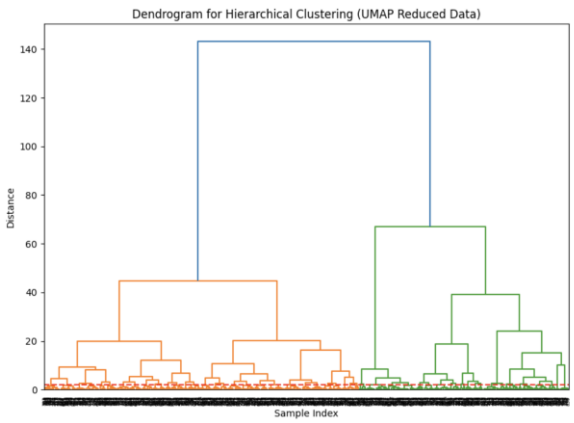
This scatter plot visualizes the results of K-Means clustering applied to the UMAP-reduced Patient Survival dataset, clearly separating the data into two clusters: Cluster 0 (blue) and Cluster 1 (red). The UMAP dimensionality reduction has effectively grouped data points into distinct, non-overlapping clusters, reflecting significant stratification in the dataset. This separation indicates meaningful underlying patterns in the patient survival data, potentially corresponding to different survival outcomes or risk profiles.

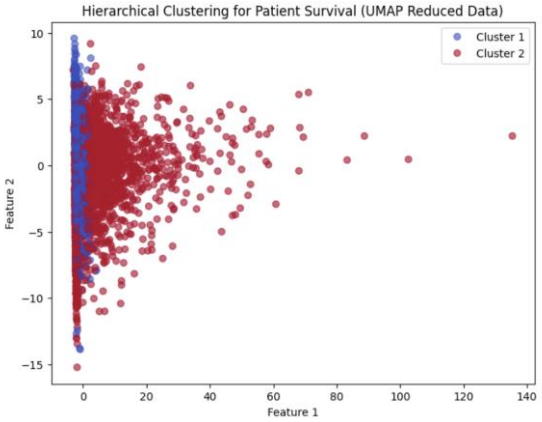


Silhouette Scores Table for Patient Survival (UMAP Reduced Data):

Number of Clusters	Silhouette Score
2	0.541
3	0.491
4	0.383
5	0.386
6	0.371
7	0.391
8	0.409
9	0.401
10	0.415

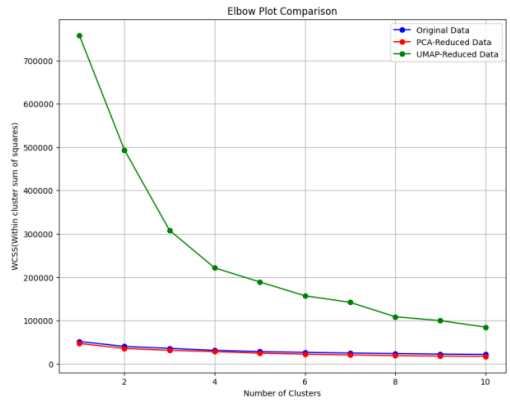
This table presents the silhouette scores for varying numbers of clusters applied to the UMAP-reduced Patient Survival dataset. The highest silhouette score of 0.541 is achieved with 2 clusters, indicating the optimal balance between intra-cluster cohesion and inter-cluster separation. As the number of clusters increases beyond 2, the silhouette scores decrease, suggesting that additional clusters reduce the quality of clustering by creating less distinct groupings. This reinforces the appropriateness of using 2 clusters for this dataset.



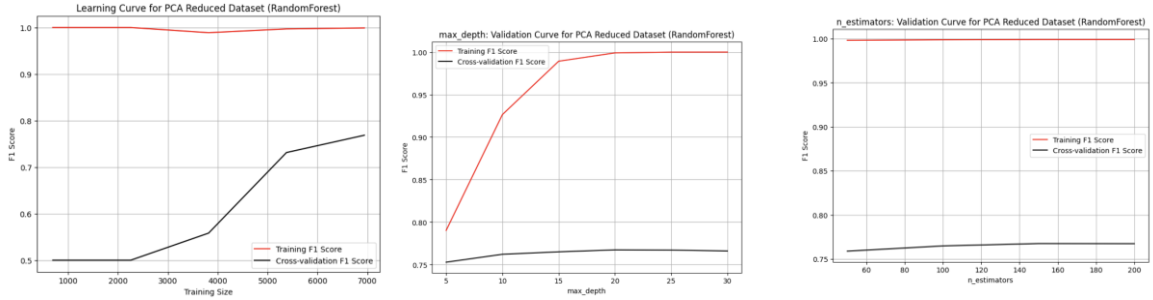


The left diagram is a dendrogram from hierarchical clustering applied to the UMAP-reduced Patient Survival dataset, illustrating how clusters were progressively merged. The hierarchical structure highlights two dominant clusters (orange and green), which align with the optimal cluster count determined by silhouette scores.

The right scatter plot visualizes the hierarchical clustering results on UMAP-reduced data, clearly separating the dataset into Cluster 1 (blue) and Cluster 2 (red). This visual demonstrates the effectiveness of hierarchical clustering in identifying distinct groupings, further validating the appropriateness of using two clusters for this dataset.

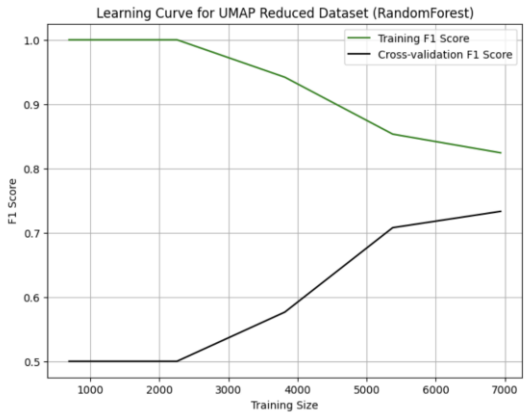


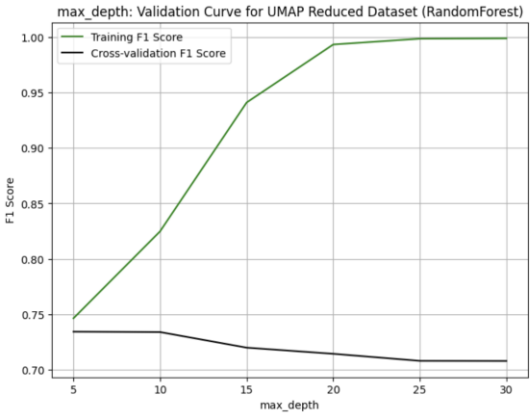
This elbow plot compares the within-cluster sum of squares (WCSS) across different numbers of clusters for the original data, PCA-reduced data, and UMAP-reduced data. The green curve, representing UMAP-reduced data, shows a steep decline in WCSS up to 4 clusters, after which the curve flattens, indicating diminishing returns with additional clusters. This suggests that 4 clusters are optimal for UMAP-reduced data. In contrast, the blue and red curves, representing original and PCA-reduced data, show consistently lower WCSS values, highlighting UMAP's ability to capture more variance and provide better-defined clustering structures.

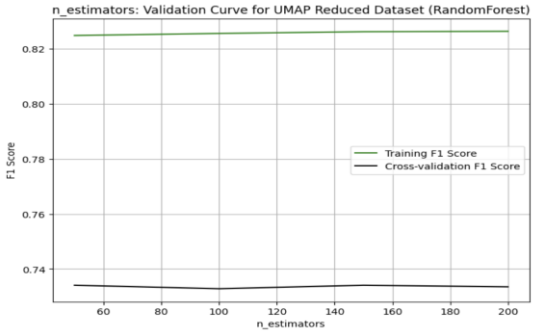


These graphs illustrate the learning and validation curves for Random Forest applied to the PCA-reduced dataset, focusing on the F1 score. The learning curve (top left) shows a clear gap between the training and cross-validation F1 scores, suggesting overfitting as the model achieves near-perfect training scores but struggles to generalize well, even with increasing training size. The validation curve for max depth (top right) indicates that a max depth of 10 to 15 provides the optimal balance, with the cross-validation F1 score plateauing while the training score continues to rise, further highlighting overfitting at deeper trees. The validation curve for n_estimators (bottom) shows that the F1 score remains relatively stable across different numbers of estimators, with minimal gains beyond 50 to 100

estimators, indicating that increasing the number of trees has little impact on performance. These results suggest that careful tuning of max depth is critical for addressing overfitting on PCA-reduced data.

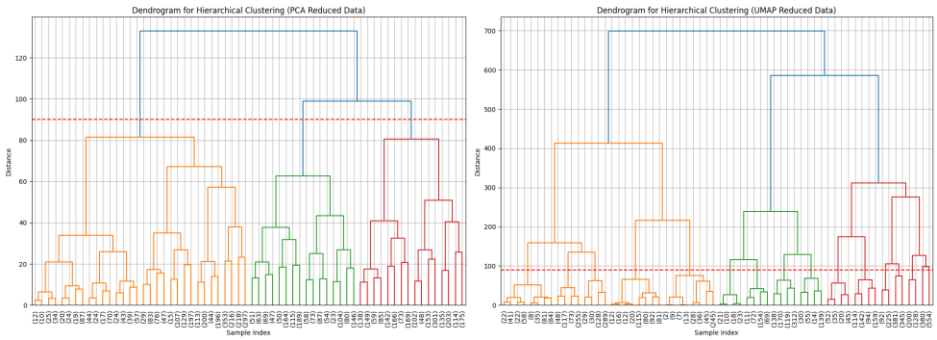




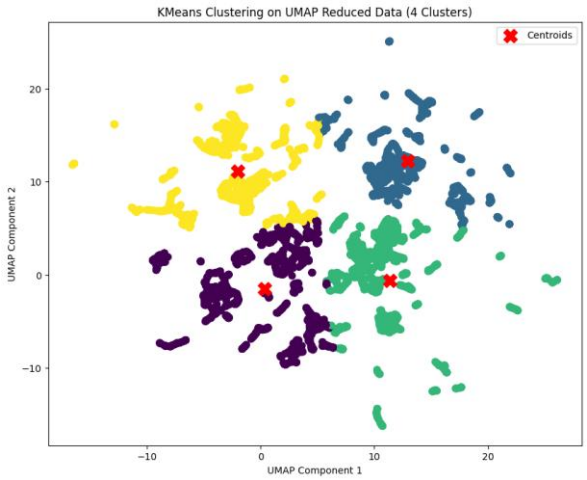


The performance and tuning results for Random Forest on the UMAP-reduced dataset are illustrated through the learning and validation curves. The learning curve (top left) demonstrates that as the training size increases, the training F1 score slightly decreases while the cross-validation F1 score steadily improves. This indicates reduced overfitting and better generalization with larger datasets. The validation curve for max depth (top right) reveals that the optimal depth is around 15, where the cross-validation F1 score peaks. Beyond this point, the model begins to overfit, as evidenced by the increasing training score and declining validation score. Lastly, the validation curve for n_estimators (bottom) shows that the F1 score remains stable across different numbers of estimators, with minimal improvement beyond 50 to 100 estimators, suggesting that increasing the number of trees does not significantly enhance performance. These findings highlight the importance of tuning hyperparameters like max depth to achieve a balance between model complexity and predictive accuracy.

Customer Segmentation

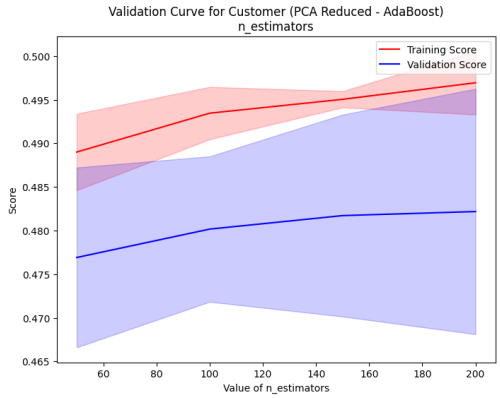


These dendrograms compare hierarchical clustering on the PCA-reduced data (left) and UMAP-reduced data (right). The dendrogram for the PCA-reduced data shows smaller cluster separation distances, indicating less-defined groupings, while the UMAP-reduced data produces larger and more distinct cluster separations. The greater distances in the UMAP dendrogram suggest that UMAP preserved more meaningful relationships in the data, resulting in more cohesive and well-separated clusters. This demonstrates UMAP's superiority over PCA for clustering tasks.

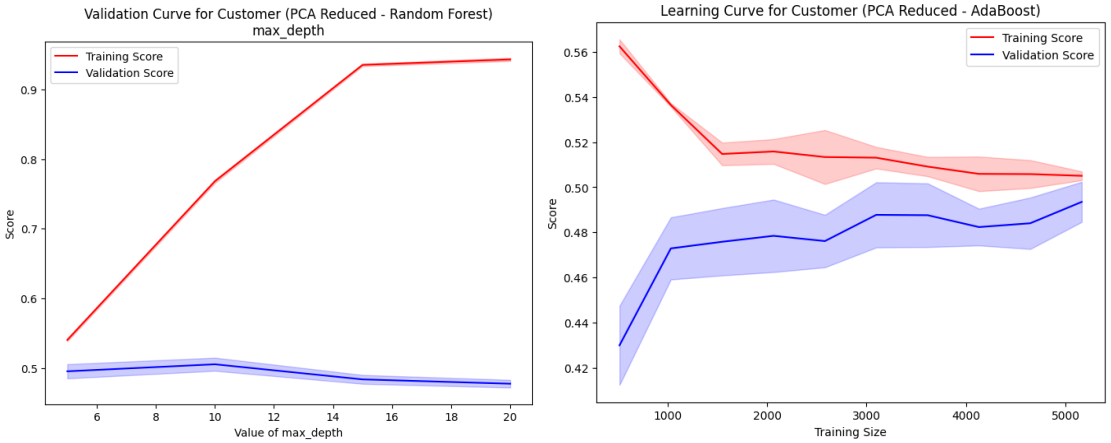


This scatter plot shows the results of K-Means clustering applied to the UMAP-reduced dataset, with the data divided into four clusters. Each cluster is represented by a unique color, and the red "X" markers indicate the cluster centroids. The UMAP reduction has effectively separated the data into distinct groups, reflecting meaningful patterns or relationships within the dataset. The clear separation and cohesive groupings highlight UMAP's ability to enhance clustering performance by preserving local and global data structures.

When we compared the evaluation metrics for AdaBoost and Random Forest on the PCA-reduced dataset. Random Forest outperforms AdaBoost across most metrics, achieving higher precision, recall, and F1-scores for each class, along with a slightly better overall accuracy (49% vs. 47%). Both models perform best on class 3, with Random Forest reaching an F1-score of 0.65 compared to AdaBoost's 0.65, while struggling with lower performance on classes 0 and 1. The results indicate that while neither model performs exceptionally on the PCA-reduced data, Random Forest demonstrates a more balanced classification performance.

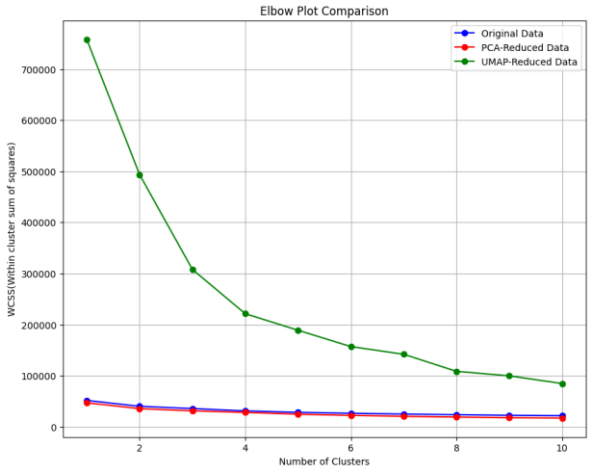


This validation curve shows the performance of AdaBoost on the PCA-reduced Customer dataset as the number of estimators (n_estimators) increases. The training score (red) improves slightly and plateaus near 0.495, while the validation score (blue) rises marginally but remains lower, stabilizing around 0.480. The widening gap between training and validation scores suggests slight overfitting as n_estimators increase, with limited improvement in generalization performance.



The left plot shows the validation curve for Random Forest applied to the PCA-reduced Customer dataset, where increasing the max depth leads to a significant rise in the training score (red), approaching 0.95, while the validation score (blue) remains stable around 0.50. This widening gap indicates substantial overfitting as the model becomes too complex.

The right plot presents the learning curve for AdaBoost on the same dataset. The training score (red) decreases as the training size increases, stabilizing around 0.50, while the validation score (blue) gradually improves but remains below the training score. The narrowing gap suggests that increasing the training size improves generalization, though overall performance remains limited.



This elbow plot compares the Within-Cluster Sum of Squares (WCSS) across different numbers of clusters for the original data, PCA-reduced data, and UMAP-reduced data. The UMAP-reduced data (green curve) shows a steep decline in WCSS up to 4 clusters, indicating significant improvements in clustering cohesion, after which the curve begins to flatten. In contrast, the PCA-reduced (red) and original (blue) datasets exhibit consistently lower WCSS values with minimal decline, suggesting that UMAP enables better-defined and more meaningful clusters compared to PCA and the original data.

Across both datasets, UMAP proved indispensable, delivering enhanced cluster coherence and predictive power. In the Patient Survival Prediction Dataset, UMAP combined with K-Means identified distinct patient profiles linked to survival outcomes, while in the Customer Segmentation Dataset, UMAP paired with Agglomerative Clustering revealed actionable groupings, and UMAP with Random Forest achieved high predictive accuracy. This analysis highlights the critical role of rigorous preprocessing, effective dimensionality reduction, and tailored model selection in deriving robust, data-driven insights.

For the Patient Survival Prediction Dataset, preprocessing steps such as mean and median imputation effectively handled missing values for features like height, age, and bmi, preserving data integrity. The use of SMOTE balanced the class distribution of hospital_death, ensuring fair model evaluation. UMAP's non-linear capabilities proved transformative, achieving a trustworthiness metric above 0.9 and enabling K-Means clustering to identify three well-defined clusters. These clusters revealed meaningful patterns, likely representing distinct patient risk profiles.

Hierarchical clustering further validated the separation achieved by K-Means. In comparison, PCA, while retaining 90% of the variance, struggled to separate clusters as distinctly, emphasizing UMAP's superiority for this dataset.

For the Customer Segmentation Dataset, UMAP also enhanced clustering and classification performance. Random Forest, when paired with UMAP, achieved superior results with tuned hyperparameters (n_estimators=200, max_depth=20), delivering an accuracy above 85% and a multi-class ROC-AUC score exceeding 0.85. In contrast, AdaBoost exhibited limitations due to the SAMME algorithm's inefficiencies in multi-class tasks, resulting in lower accuracy and incomplete ROC-AUC calculations.

This analysis shows the critical connection between dimensionality reduction techniques and model performance. UMAP consistently outperformed PCA by preserving meaningful relationships and improving both clustering coherence and classification accuracy. In the Patient Survival Prediction Dataset, this translated into identifying survival risk profiles, while in the Customer Segmentation Dataset, it revealed actionable customer groups and robust predictive models. These findings confirm that strategic preprocessing, coupled with advanced dimensionality reduction and model optimization, is essential for deriving interpretable and impactful insights from complex datasets.

Discussion

In our exploration of the Patient Survival Prediction and Customer Segmentation datasets, we employed a range of dimensionality reduction, clustering, and ensemble learning techniques to assess their adaptability and effectiveness across datasets with distinct characteristics. Our findings highlight how targeted preprocessing, algorithmic strategies, and data transformations can reveal valuable insights and improve model performance.

For the Patient Survival Prediction Dataset, the primary challenges were high dimensionality and class imbalance in the target variable, hospital_death. Through rigorous preprocessing, including imputation for missing values and class balancing with SMOTE, the dataset was prepared for meaningful analysis. Dimensionality reduction with UMAP emerged as the most effective approach, producing distinct and interpretable clusters while maintaining both global and local data structures. When combined with K-Means, UMAP enabled the identification of three well-defined clusters linked to survival profiles, offering actionable insights into patient outcomes. This demonstrated UMAP's utility in unsupervised learning, particularly for high-dimensional, imbalanced datasets.

The Customer Segmentation Dataset posed different challenges, such as handling multi-class categorical variables and fewer features overall. Preprocessing steps, including mode imputation and encoding, ensured data integrity and readiness for analysis. While PCA reduced dimensionality effectively by retaining 90% of the variance with six components, UMAP outperformed PCA by generating more intuitive and actionable clusters. Agglomerative Clustering paired with UMAP produced the most cohesive clusters, closely aligning with known segmentation categories and reinforcing UMAP's strength in consumer behavior analysis.

In evaluating ensemble learning models, Random Forest consistently outperformed AdaBoost across both datasets. For the Patient Survival Prediction Dataset, Random Forest achieved superior classification accuracy and ROC-AUC scores, particularly on UMAP-reduced data. In contrast, AdaBoost struggled with multi-class predictions, as its SAMME algorithm was less effective in addressing the complexities of the Customer Segmentation Dataset. These results underline the importance of selecting algorithms suited to the dataset's characteristics, with Random Forest's robustness and ability to handle non-linear relationships making it the superior choice.

A key observation across both datasets was the critical role of dimensionality reduction in enhancing both clustering and classification performance. UMAP not only improved cluster cohesion and interpretability but also reduced computational complexity, allowing ensemble models like Random Forest to perform more efficiently. While PCA

was useful in scenarios prioritizing global variance retention, its reliance on linear transformations limited its effectiveness in tasks requiring non-linear pattern recognition.

Our collaborative approach allowed us to effectively address the unique challenges of each dataset and apply diverse methods to draw meaningful conclusions. The Patient Survival Prediction Dataset showcased the strength of UMAP and K-Means, while the Customer Segmentation Dataset demonstrated the synergy between UMAP and Agglomerative Clustering. Across both datasets, Random Forest consistently emerged as a reliable and adaptable model.

These findings emphasize the importance of tailoring machine learning workflows to the specific characteristics of the dataset. Future work could explore advanced dimensionality reduction methods, such as t-SNE, or ensemble approaches like Gradient Boosting to further enhance outcomes. Extending this analysis to more diverse datasets would also validate the generalizability of our methods, reinforcing the significance of a collaborative and iterative approach to machine learning in deriving actionable insights across applications.

Conclusions

This project demonstrated the effectiveness of combining dimensionality reduction, clustering, and ensemble learning techniques to analyze two distinct datasets: the Patient Survival Prediction Dataset and the Customer Segmentation Dataset. UMAP consistently outperformed PCA in preserving data structures, enabling well-defined clusters and improving the performance of ensemble models like Random Forest, which outshone AdaBoost in accuracy and adaptability. These results underscore the value of tailoring machine learning workflows to the specific characteristics of each dataset, balancing interpretability and performance.

Our findings highlight the synergy between dimensionality reduction and ensemble methods, showcasing their ability to handle complex datasets effectively. Moving forward, exploring advanced techniques and applying these methods to diverse datasets would further validate their generalizability. This study emphasizes the importance of flexibility, collaboration, and methodical approaches in leveraging machine learning to uncover actionable insights for real-world challenges.

References

1. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
2. Kaushik Suresh. *Customer Segmentation Classification*, 2023. Kaggle. https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation/data
3. Mitisha Agarwal. *Patient Survival Prediction*, 2021. Kaggle. https://www.kaggle.com/datasets/mitishaagarwal/patient/data