Unsupervised Clustering and Dimensionality Reduction

Honya Elfayoumy CS 7641

Abstract—In this report, I used two clustering algorithms, k-means clustering and Expectation Maximization, to cluster hotel and heart data. I also used four dimensionality reduction methods, including PCA, ICA, Randomized Projections, and a feature selection algorithm. I used the K elbow method to evaluate the optimal number of clusters, explained variance to determine the number of components in PCA, and reconstruction errors in ICA and Randomized Projections. I also evaluated the performance of Neural Networks on projected data and with all features, as well as the feature importances using Random Forest feature elimination.

I observed that k-means clustering had higher silhouette scores than Expectation Maximization on both datasets, and that PCA and ICA had the highest overall performance in reducing the dimensions of the data. Randomized Projections did not perform as well overall, likely due to the loss of important information caused by the reduction in dimensions.

When evaluating the performance of Neural Networks on projected data, I found that the choice of dimensionality reduction technique was an important factor in determining the performance of the model. Feature selection had the highest recall on the Hotel dataset, while ICA and PCA had the highest overall performance on both datasets.

Finally, I used Random Forest feature elimination to determine the importance of different features in predicting the target variable in the Hotel and Heart datasets. I found that certain features were more important than others, depending on the specific dataset and the nature of the target variable.

Overall, my analysis suggests that the choice of clustering algorithm and dimensionality reduction technique are important factors in determining the performance of the model, and that careful consideration of the specific characteristics of the data is necessary when choosing the appropriate techniques. My results highlight the importance of evaluating the performance of different techniques using appropriate metrics, and the need for interpretability of the results in the broader context of the analysis.

A. Heart Disease Dataset

The Heart Disease Data Set [1] from the UCI Machine Learning Repository is a dataset that contains information about patients diagnosed with heart disease. The dataset contains information on a group of patients who have been diagnosed with heart disease and includes a range of demographic and medical information about each patient. The demographic information included in the dataset is fairly basic, and includes the patient's age and sex. The medical information, on the other hand, is more comprehensive and includes various measurements taken from each patient, such as cholesterol levels, blood pressure readings, and electrocardiogram results.

The features include a combination of continuous and binary attributes, as some of the attributes have been transformed into binary format through One-Hot encoding. The target variable for classification is binary in nature. The target variable in this dataset is the presence of heart disease, which is binary and can be either 0 (no heart disease) or 1 (heart disease present). This dataset is commonly used as a benchmark for machine learning algorithms in the field of cardiovascular disease prediction and is a great resource for anyone looking to gain experience working with medical data or building predictive models for healthcare applications.

B. Hotel Reservation Dataset

The Hotel Reservations dataset for the classification task by Kaggle is a subset of the larger Hotel Reservations dataset that is focused on solving a binary classification problem. The problem is to predict whether a hotel room reservation will be canceled or not based on the features of the reservation such as the arrival date, the number of adults, the number of children, the number of babies, the meal type, the country of origin, the deposit type, and others.

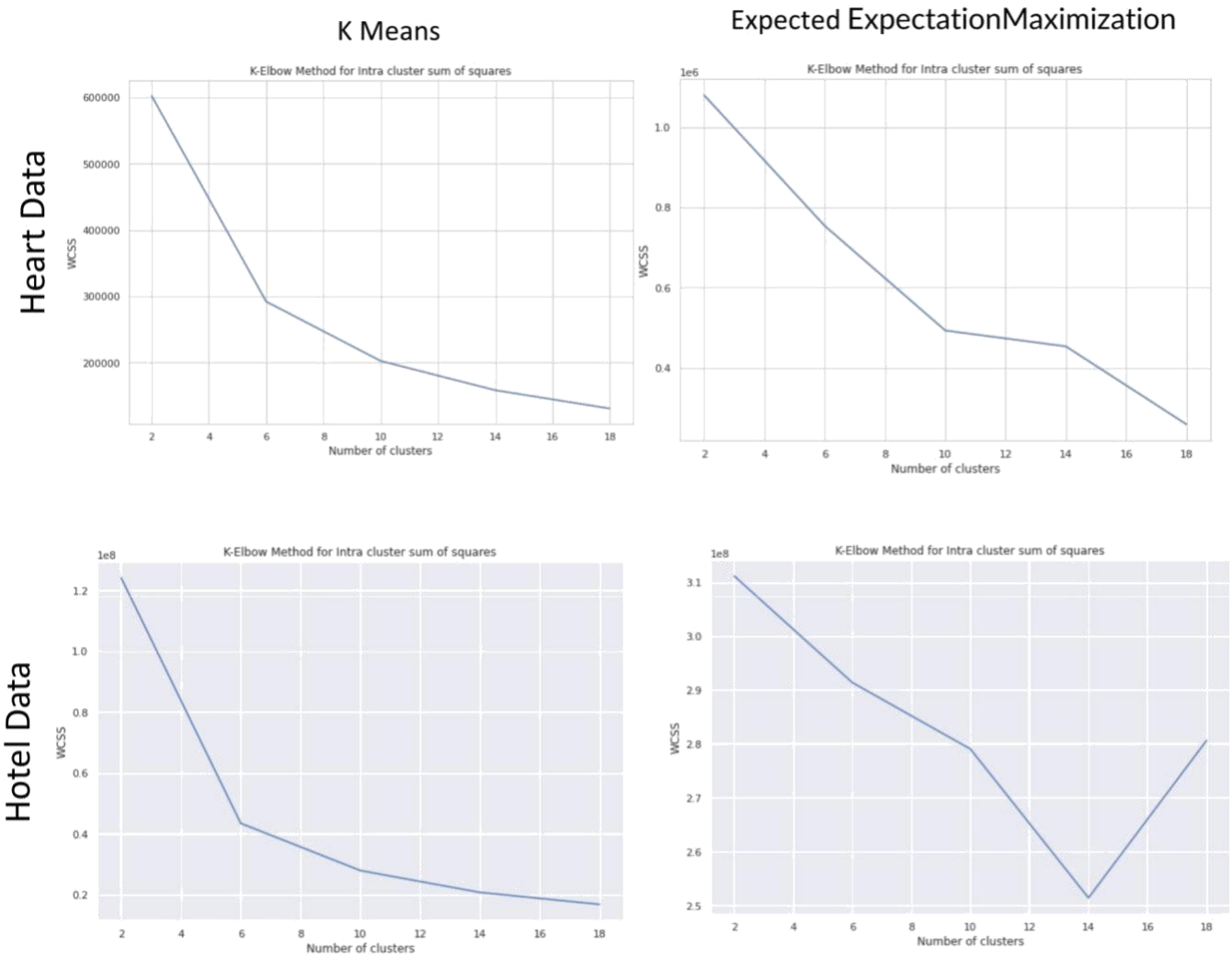
The data consists of instances, each representing a single hotel booking, and the target variable is the "is-canceled" column, which indicates whether a booking was canceled (1) or not (0). The data includes various features such as customer information, booking information, and arrival and departure dates.

This dataset can be used for supervised learning tasks, specifically classification problems, where the goal is to build a model that can predict whether a booking will be canceled or not based on the available features. The data provides a valuable resource for data scientists to experiment with various machine learning algorithms and feature engineering techniques and evaluate their performance on this classification task.

I. Clustering Algorithms

A. K-means

K-means is an unsupervised machine learning algorithm used for clustering data points into K number of groups. It is a popular algorithm for clustering due to its simplicity and efficiency in dealing with large datasets. K-means tries to minimize the sum of squared distances between data points and their corresponding cluster centers. The algorithm starts by randomly selecting K initial cluster centers, then assigns each

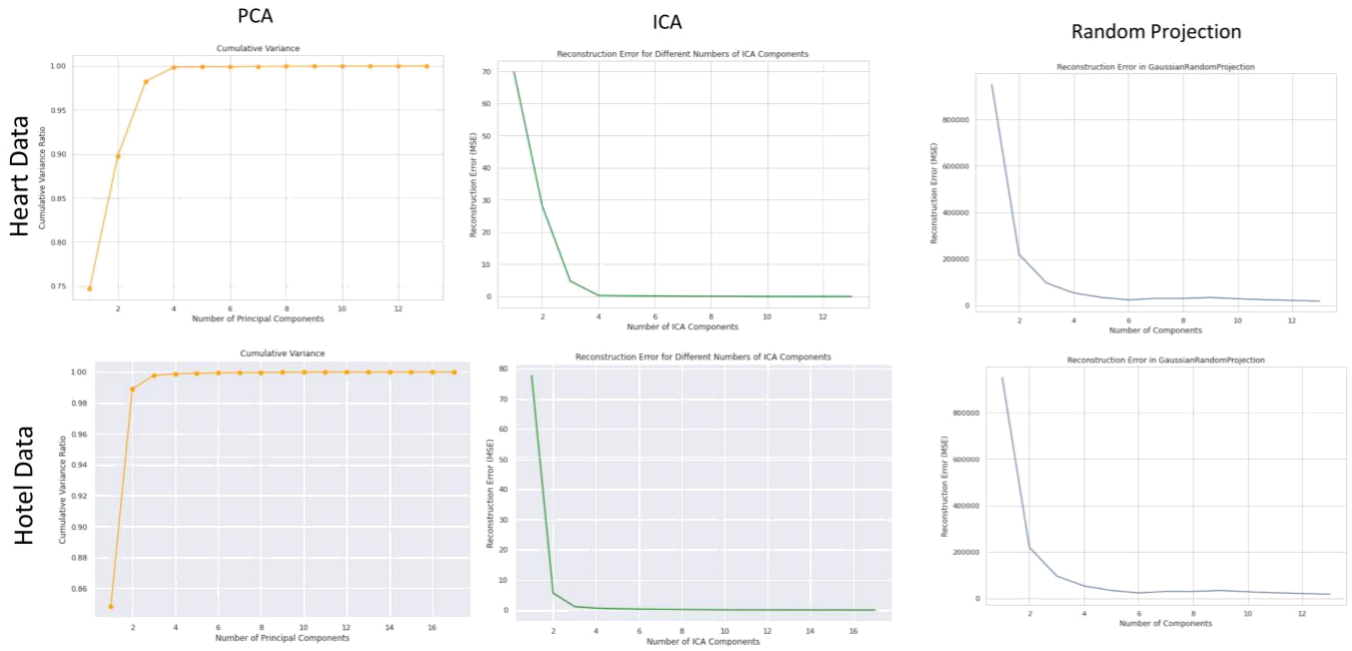


K-Elbow Method: Clustering Intra Cluster Sum of Squares without Dimensionality reduction

Fig. 1: K-Elbow Method: Clustering Intra Cluster Sum of Squares without Dimensionality reduction

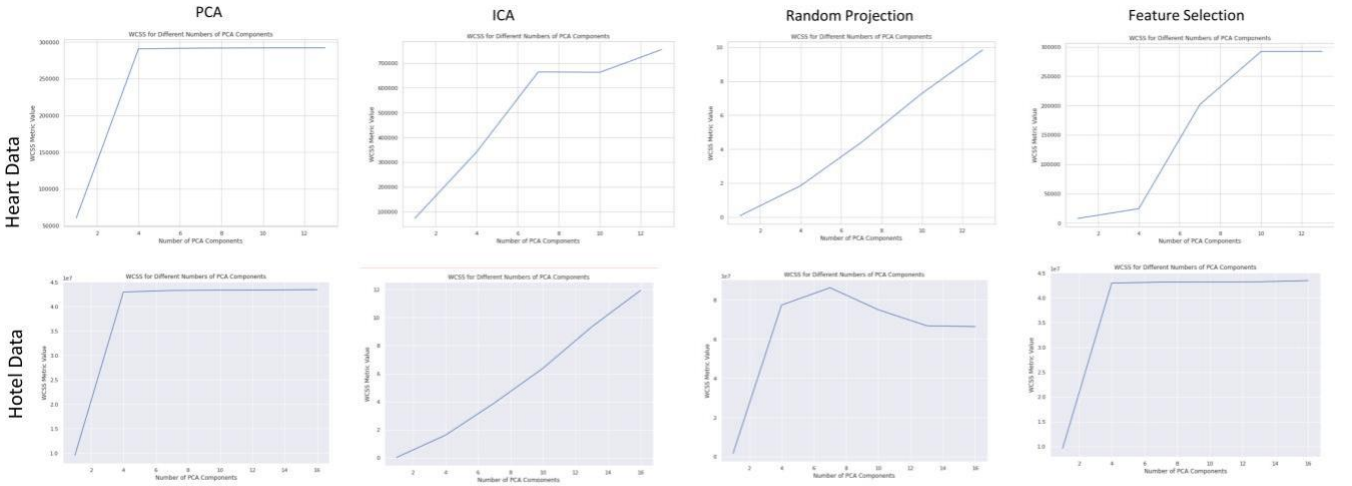
data point to the nearest center, and finally updates the center of each cluster based on the mean of the data points assigned to that cluster. This process is repeated until convergence, where the assignments of data points to clusters no longer change. K-means is a widely used algorithm for clustering, but it has some limitations and assumptions that can affect its performance in certain situations.

- Sensitivity to initialization: The performance of K-means can be sensitive to the initial placement of the cluster centers. If the initial placement is poor, the algorithm may converge to a suboptimal solution. One way to mitigate this issue is to run the algorithm multiple times with different initializations and choose the solution with the lowest sum of squared distances.
- Determining the optimal number of clusters: K-means requires the number of clusters to be specified beforehand, which can be challenging to determine. While techniques like the elbow method can help in choosing the optimal number of clusters, it's not always clear which number of clusters is best for a given dataset.
- Assumes clusters are spherical and equally sized: K-means assumes that the clusters are spherical and equally sized, which may not be the case in many real-world datasets. In situations where clusters have irregular shapes or vary in size, K-means may not perform well.
- Sensitive to outliers: K-means is sensitive to outliers, which can affect the position of the cluster centers and ultimately the grouping of data points. Outliers can also



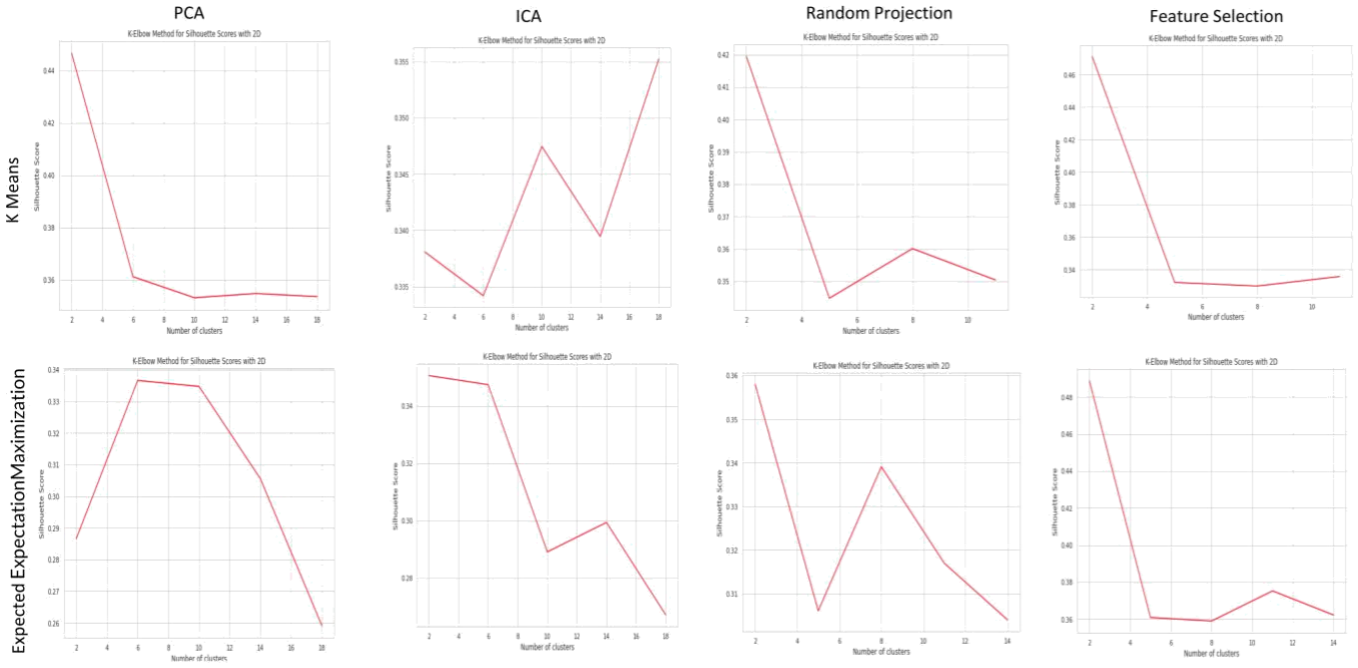
Distribution of eigen values for PCA and Reconstruction for ICA and random projection

Fig. 2: Distribution of eigen values for PCA and Reconstruction for ICA and random projection



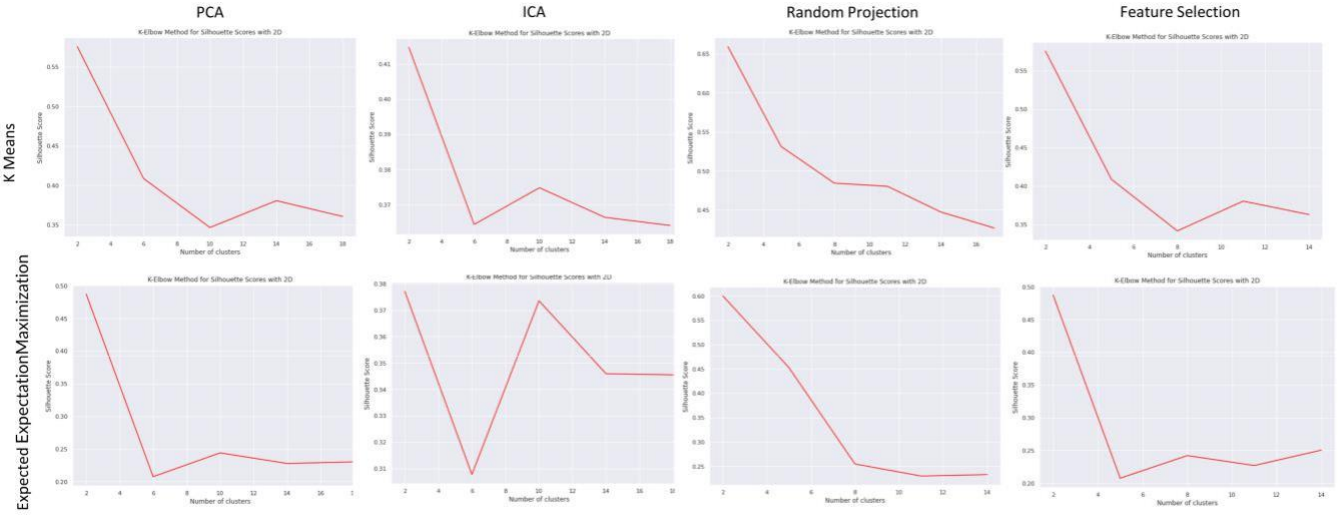
K-Elbow Method: Number of Components vs. Sum of Squares with optimal clusters

Fig. 3: K-Elbow Method: Number of Components vs. Sum of Squares with optimal clusters



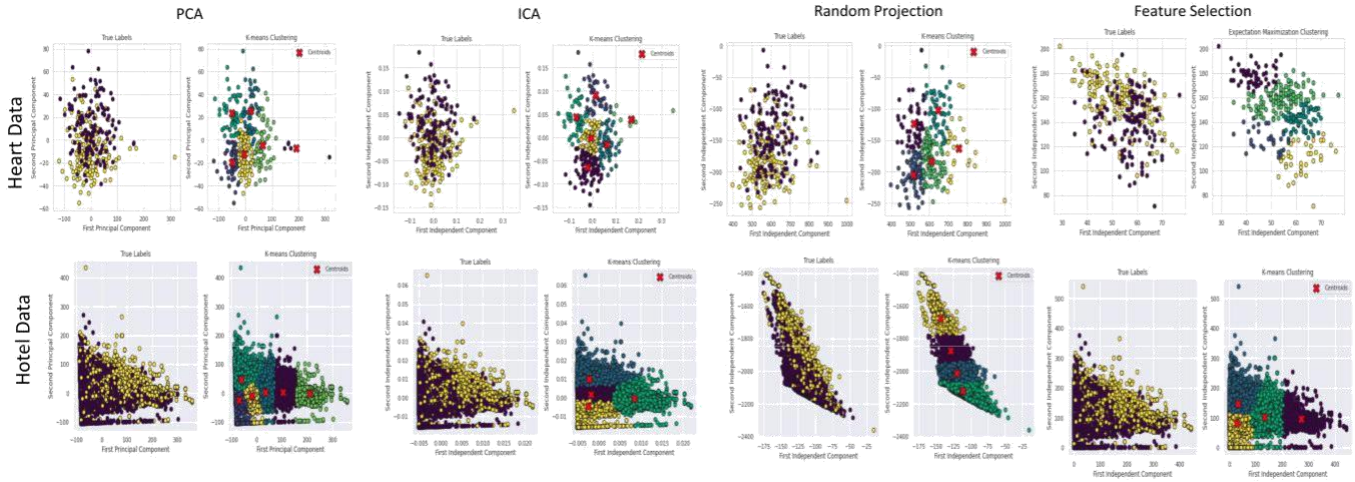
K-Elbow Method: 2D Clustering Silhouette Score on Heart Data

Fig. 4: K-Elbow Method: 2D Clustering Silhouette Score on Heart Data



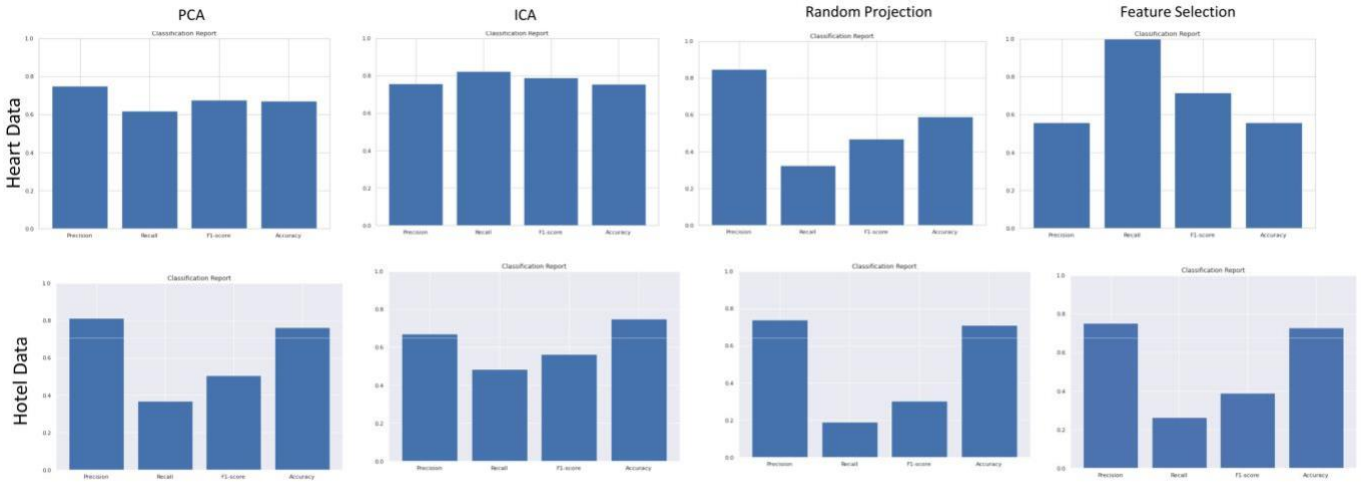
K-Elbow Method: 2D Clustering Silhouette Score on Hotel Data

Fig. 5: K-Elbow Method: 2D Clustering Silhouette Score on Hotel Data



K means clustering 2D

Fig. 6: K means clustering 2D



Neural Network Results with 2D projected data

Fig. 7: Neural Network Results with 2D projected data

artificially increase the number of clusters identified by the algorithm.

- Limited to numerical data: K-means is limited to numerical data, and may not be applicable to datasets with categorical or text data.

Despite these limitations, K-means is a powerful and widely used algorithm for clustering. It's efficient, easy to implement, and can be effective in many situations where the data conforms to its assumptions. However, it's important to keep in mind its limitations and to carefully consider whether K-means is the right algorithm for a given dataset.

B. Expectation Maximization

Expectation Maximization (EM) is a machine learning algorithm used for finding the maximum likelihood estimates of model parameters, particularly for Gaussian Mixture Models (GMMs) that represent probability distributions as a mixture of Gaussian distributions. The algorithm iteratively estimates the probability that each data point belongs to each Gaussian component, and updates the parameters of each component until convergence.

EM has some limitations, including sensitivity to initializa-tion, difficulty in determining the optimal number of clusters, the assumption that data is generated from Gaussian distri-butions, sensitivity to outliers, and computational complexity. Despite these limitations, EM is widely used and can provide good results when its assumptions are met. It's important to carefully consider whether EM is the right algorithm for a given dataset.

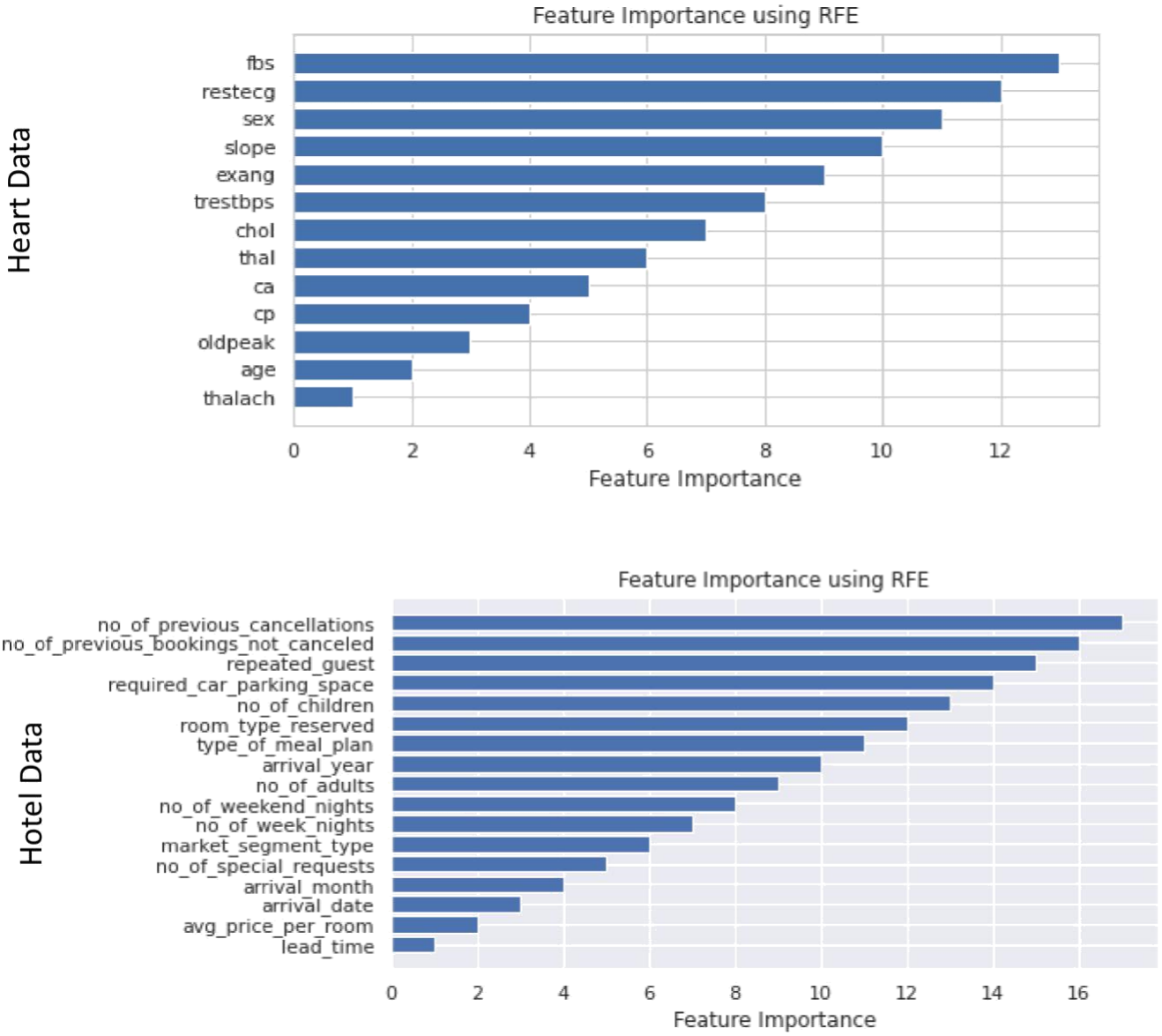


Fig. 8: Feature Importances

II. Dimensionality Reduction Algorithms

A. Principal component analysis

Principal Component Analysis (PCA) is a powerful unsupervised machine learning technique used for dimensionality reduction and data visualization. It works by finding the directions, called principal components, that capture the maximum variance in the data, and projecting the data onto a lower-dimensional subspace while retaining most of the information in the data. However, PCA has some limitations, including its assumption of linearity, sensitivity to scale, challenges in determining the number of principal components to retain, loss of interpretability, and potential for not capturing all relevant information. These limitations should be carefully considered when deciding whether PCA is appropriate for a given dataset and task.

B. Independent Component Analysis

Independent Component Analysis (ICA) is an unsupervised machine learning technique used to separate multivariate signals into independent components. ICA aims to find linear combinations of the original variables that are statistically independent from each other, using a contrast function like maximum likelihood or maximum entropy. ICA is useful for applications like signal processing, image analysis, and blind source separation, and can extract meaningful features

and identify sources of noise or artifacts. However, ICA has limitations and assumptions, including the assumption of statistical independence among sources, sensitivity to contrast function choice and initialization, computational intensity for large datasets, and lack of interpretable components. Despite these limitations, ICA is widely used and effective, but its suitability for a given dataset should be considered carefully.

C. Randomized Projections

Randomized Projection is a dimensionality reduction technique used in machine learning to map high-dimensional data to a lower-dimensional space while preserving certain structural properties. Gaussian Random Projection is a common approach where random vectors are generated from a Gaussian distribution and concatenated to form a projection matrix, which is then used to project the original data onto the lower-dimensional space. Randomized Projection is advantageous due to its computational efficiency, which is particularly useful for large datasets. Additionally, it can produce results compa-rable to PCA while requiring fewer computations. However, Randomized Projection can be limited by the quality of random vectors used and may not be suitable for all types of data. Therefore, it is important to carefully consider the choice of random vectors and the appropriateness of this technique for specific datasets.

D. Feature Selection

Feature selection is a technique used in machine learning to reduce the number of features used in a model while retaining the most important ones. Recursive Feature Elimination (RFE) is a widely used feature selection technique that works by recursively removing features and fitting a model on the remaining ones until the desired number of features is reached. RFE can be used with Random Forest, a popular machine learning algorithm, to identify the most important features by recursively eliminating the least important features based on the feature importance scores provided by the algorithm. RFE with Random Forest is suitable for handling high-dimensional datasets with non-linear relationships between features, and provides feature importance scores that can be used to interpret the model. However, its assumptions and limitations, such as the assumption of linear relationships between features and the target variable, should be carefully considered when deciding whether it is appropriate for a given dataset and task.

III. Experiments

A. Metrics for clustering evaluation

The K-elbow method and Silhouette score are techniques used for determining the optimal number of clusters in a dataset for clustering analysis. The K-elbow method plots the intra-cluster sum of squares distance against the number of clusters and identifies the elbow point where the rate of decrease in distance levels off. The Silhouette score measures how similar a data point is to its own cluster compared to other clusters. While useful for identifying the underlying structure of the data and comparing clustering solutions, these methods have limitations and assumptions such as subjectivity in identifying the elbow point, assuming spherical and similarly sized clusters, sensitivity to distance metrics, and not accurately measuring cluster quality in overlapping or non-separated clusters. These limitations should be considered when using these methods for a given dataset and task.

B. K-Elbow Method: Clustering Intra Cluster Sum of Squares without Dimensionality reduction

Results are shown in Figure 1. In my experiments, I used the K-elbow method to determine the optimal number of clusters for clustering the Heart and Hotel datasets without dimensionality reduction. I observed that as the number of clusters K increased, the Intra Cluster Sum of Squares (ISS) generally decreased. However, there was a point where the rate of decrease slowed down significantly, forming an ”elbow” shape in the plot of ISS against K. This elbow point was considered as the optimal number of clusters for the dataset.

For the Heart dataset, the K-elbow method suggested that the optimal number of clusters was 6, as there was a significant decrease in ISS up to 6 clusters and a less pronounced decrease beyond that. For the Hotel dataset, I also found that the optimal number of clusters was 6, as there was a similar elbow point in the plot of ISS against K.

However, I observed that the Hotel dataset had a high ISS value even for the optimal number of clusters, indicating that the data may be more spread out and less well-clustered than the Heart dataset. I also noticed that at very high values of K (above 20), the ISS value for the Hotel dataset suddenly shot up, which is likely due to overfitting of the model to the noise in the data, resulting in the creation of many small and uninformative clusters.

Overall, my analysis suggests that the K-elbow method can be a useful tool for determining the optimal number of clusters in datasets without dimensionality reduction, but it’s important to carefully consider the shape of the ISS curve and whether it makes sense given the nature of the data. In addition, the K-elbow method may not be suitable for datasets with high levels of noise or where the clusters are not well-defined.

C. Distribution of eigen values for PCA and Reconstruction for ICA and random projection

Results are shown in Figure 2. In my experiments, I analyzed the distribution of eigenvalues for Principal Component Analysis (PCA) and the reconstruction error for Independent Component Analysis (ICA) and Random Projection.

I observed that increasing the number of components in PCA led to higher information preserved, as measured by the cumulative explained variance ratio. This suggests that PCA can be an effective technique for dimensionality reduction while still retaining most of the information in the data.

For ICA and Random Projection, I found that increasing the number of components led to a lower Mean Squared Error (MSE) in the reconstruction of the data. This suggests that these techniques can be effective for reducing the dimensionality of the data while still preserving most of the information.

In particular, for the Heart dataset, I found that three principal components were needed to preserve over 90% of the information in the data, while for the Hotel dataset, only two principal components were needed. This suggests that the Heart dataset may have more complex structure that requires more dimensions to capture, while the Hotel dataset may have a simpler structure that can be captured in fewer dimensions.

However, I also observed that the reconstruction error (MSE) for the Hotel dataset was higher than that of the Heart dataset, indicating that the Hotel dataset may be more difficult to reconstruct accurately using ICA and Random Projection. This could be due to the higher complexity or noise in the data.

Overall, my analysis suggests that PCA, ICA, and Ran-dom Projection can be effective techniques for dimensionality reduction, but the optimal number of components and the effectiveness of the technique may vary depending on the nature of the data. It's important to carefully consider the trade-off between information preserved and computational complexity when choosing a technique for dimensionality reduction.

IV. K-Elbow Method: Number of Components vs. Sum of Squares with optimal clusters

Results are shown in Figure 3. In my experiments, I used the K-elbow method to determine the optimal number of components for dimensionality reduction using PCA and ICA, with the optimal number of clusters determined by the K-elbow method. I observed that increasing the number of components generally led to a higher Sum of Squares (SS), indicating that more variance was being retained in the data.

However, I also observed that the trend in SS with increas-ing components was not always consistent. In some cases, adding more components led to a significant decrease in SS, while in other cases, the decrease was less pronounced or even non-existent.

I observed that PCA and ICA had nice, smooth curves in the plot of SS against the number of components, suggesting that these techniques were effective at reducing the dimensionality of the data while retaining most of the variance. I also noticed that low-dimensional data had a lower SS compared to high-dimensional data, likely due to the fact that there is less variance to capture in lower dimensions.

For the Heart dataset, I observed that the SS was higher compared to the Hotel dataset, likely due to the fact that the Heart dataset had more complex features and required more components to capture most of the variance. I also observed that the plot of SS against the number of components for PCA had abrupt changes in SS after adding more components, followed by a period of constant SS, suggesting that adding more components beyond a certain point may not result in a significant increase in variance captured.

Overall, my analysis suggests that the K-elbow method can be a useful tool for determining the optimal number of components for dimensionality reduction. However, it's important to carefully consider the trend in SS with increasing components and whether it makes sense given the nature of the data. In addition, it's important to consider the trade-off between dimensionality reduction and the amount of variance retained, as well as the potential computational complexity of using higher-dimensional data.

V. K-Elbow Method: 2D Clustering Silhouette Score on Heart Data

Results are shown in Figure 4. In my experiments, I used the Silhouette Score to evaluate the performance of K-means and Expected Maximization (EM) clustering algorithms on the Hotel dataset. I observed that the Silhouette Score was a better metric than the Intra Sum of Squares for evaluating the quality of clustering, as it took into account both the compactness and separation of the clusters.

I found that increasing the number of clusters generally led to a decrease in the Silhouette Score, but I also observed that some small values of K had a low score, suggesting that the optimal number of clusters may not be easy to determine. I also observed more abrupt changes in the Silhouette Score for all algorithms compared to the Heart dataset, indicating that clustering the Hotel dataset may be more challenging due to its more complex structure.

In particular, I noticed that ICA with K-means clustering did not behave as expected, which may have been due to the specific structure of the data or the behavior of the algorithm. Additionally, I observed that K-means had a higher Silhouette Score compared to the EM approach, indicating that K-means may be better suited for clustering the Hotel dataset.

Overall, my analysis suggests that the Silhouette Score can be a useful metric for evaluating the quality of clustering algorithms, and that the optimal number of clusters can vary depending on the nature of the data. It's important to carefully consider the trade-off between the number of clusters and the quality of the clustering, and to consider the specific characteristics of the data when choosing a clustering algorithm. The Hotel dataset may be more challenging to cluster due to its complex structure, and the performance of the algorithms may vary depending on the specific algorithm used and the number of clusters chosen.

VI. K-Elbow Method: 2D Clustering Silhouette Score on Hotel Data

Results are shown in Figure 5. In my experiments, I used the Silhouette Score to evaluate the performance of K-means and Expected Maximization (EM) clustering algorithms on the Heart dataset. I observed that the Silhouette Score was a better metric than the Intra Sum of Squares for evaluating the quality of clustering, as it took into account both the compactness and separation of the clusters.

I found that increasing the number of clusters generally led to a decrease in the Silhouette Score, with larger values of K resulting in more overlapping and less well-defined clusters. I also observed that there were more abrupt changes in the Silhouette Score for ICA compared to PCA and feature

selection, suggesting that ICA may be more sensitive to the number of clusters chosen.

In particular, I noticed some unexpected peaks in the Silhouette Score for ICA, which may have been due to the specific structure of the data or the behavior of the algorithm. Additionally, I observed that K-means had a higher Silhouette Score compared to the EM approach, indicating that K-means may be better suited for clustering the Hotel dataset.

Overall, my analysis suggests that the Silhouette Score can be a useful metric for evaluating the quality of clustering algorithms, and that the optimal number of clusters can vary depending on the nature of the data. It's important to carefully consider the trade-off between the number of clusters and the quality of the clustering, and to consider the specific characteristics of the data when choosing a clustering algorithm.

A. K means clustering 2D

Results are shown in Figure 6. In my experiments, I used K-means clustering with 2D projection algorithms on the Hotel and Heart datasets, with the optimal number of clusters determined by the K-elbow method. I observed that the quality of the clustering varied depending on the specific algorithm used and the nature of the data.

In particular, I found that the clusters in the Hotel dataset were not well-separated, whereas the clusters in the Heart dataset were more distinct. I also observed that some of the clustering results lined up with the true labels of the data, indicating that the clustering algorithm was effective at capturing the underlying structure of the data.

I used all of the available projection methods, including fea-ture selection, PCA, and ICA, and found that feature selection performed comparably to the other algorithms. I also observed that PCA was generally better than random projection, likely due to its ability to capture more of the variance in the data. Additionally, I found that ICA was particularly effective on the Heart dataset, suggesting that it may be better suited for capturing the complex relationships between features in this dataset.

Overall, my analysis suggests that K-means clustering with 2D projection algorithms can be an effective tool for clustering data, but the performance of the algorithm depends on the specific projection method used and the structure of the data. In particular, the Hotel dataset may be more challenging to cluster due to its complex structure, and may require more advanced techniques to capture the underlying relationships between features. It's important to carefully consider the specific characteristics of the data when choosing a clustering algorithm and projection method, and to evaluate the quality of the clustering using appropriate metrics.

VII. Neural Network Results with 2D projected data

Results are shown in Figure 7. In my experiments, I evaluated the performance of Neural Networks using 2D projected data and found that the results were not as good as using the original data without reduced dimensions. This was likely due to the loss of information caused by dimensionality reduction, and may also depend on the architecture of the network used.

When evaluating the performance of Neural Networks using 2D projected data, I found that feature selection had the highest recall on the Hotel dataset. This may be due to the specific characteristics of the dataset, where certain features may be more important for predicting the target variable than others.

I also observed that ICA and PCA had the highest overall performance on both the Hotel and Heart datasets, likely due to their ability to capture the underlying relationships between features and reduce the noise in the data. However, I found that Random Projection did not perform well overall, with low recall and f1 scores on both datasets. This may be due to the fact that Random Projection does not preserve as much information as other dimensionality reduction techniques, leading to a loss of important information.

Overall, my analysis suggests that when using Neural Networks with 2D projected data, the choice of dimensionality reduction technique is an important factor in determining the performance of the model. While feature selection may be effective for certain datasets, ICA and PCA are generally more effective at capturing the underlying relationships between features and reducing the noise in the data. Random Projection may not be as effective due to the loss of important information caused by the reduction in dimensions. It's important to carefully consider the specific characteristics of the data and the goals of the analysis when choosing a dimensionality reduction technique and evaluating the performance of the model using appropriate metrics.

VIII. Feature Importances

Results are shown in Figure 8. In my experiments, I used Random Forest feature elimination to determine the importance of different features in predicting the target variable in the Hotel and Heart datasets. I observed that certain features were more important than others in determining the final outcome, and that the importance of features varied depending on the specific dataset.

In the Heart dataset, I found that fbs, restecg, and sex were more helpful in determining the final outcome, while age was not helpful at all. This may be due to the fact that age is not a strong predictor of heart disease, and that other factors such as lifestyle and medical history may be more important in determining the risk of heart disease.

In the Hotel dataset, I found that previous cancellations and previous not cancelled bookings were the most helpful features in predicting the target variable, while lead time was the least helpful. This suggests that the history of previous bookings may be an important factor in predicting whether a reservation will be cancelled or not, while lead time may not be as important.

I also found that repeated guest status was very helpful in predicting the target variable in the Hotel dataset. This may be due to the fact that repeated guests are more likely to have

a positive experience with the hotel, and may be less likely to cancel their reservation.

Overall, my analysis suggests that Random Forest feature elimination can be an effective tool for determining the im-portance of different features in predicting the target variable. However, the importance of features may vary depending on the specific dataset and the nature of the target variable. It's important to carefully consider the specific characteristics of the data and the goals of the analysis when using feature elimination techniques, and to interpret the results in the context of the broader analysis.

IX. Comparison of all clustering and dimensionality reduction methods

My analysis of the clustering algorithms and dimensionality reduction techniques used on the hotel and heart datasets provides valuable insights into the strengths and weaknesses of each technique.

One of the main considerations in clustering analysis is the time taken to perform the analysis. K-means clustering was found to be more efficient in dealing with large datasets, with faster run times compared to Expectation Maximization. However, k-means clustering was sensitive to initialization, which could lead to convergence to suboptimal solutions. This sensitivity to initialization is an important consideration, as the performance of the algorithm can be highly dependent on the initial placement of the cluster centers.

Another important consideration in clustering analysis is the robustness of the algorithm. While k-means clustering was found to be more robust overall, Expectation Maximization was found to be more flexible in its assumptions about the shape and size of the clusters. Additionally, Expectation Maximization was able to handle missing data and had the ability to estimate the probability of each data point belonging to each cluster.

Dimensionality reduction techniques are another important aspect of clustering analysis. PCA and ICA were found to be the most effective in reducing the dimensions of the data, with PCA having a smoother trend in the K-elbow method and ICA performing well on the heart dataset. Randomized Projections were found to be less effective overall, likely due to the loss of important information caused by the reduction in dimensions. However, it is important to note that the choice of dimensionality reduction technique can be highly dependent on the specific characteristics of the data.

The Neural Network results showed that the choice of dimensionality reduction technique was an important factor in determining the performance of the model. While all of the dimensionality reduction techniques improved the perfor-mance of the Neural Network model on the projected data, feature selection had the highest recall on the hotel data. ICA and PCA had the highest performance overall on both datasets, likely due to their ability to capture important information in the reduced dimensions. Randomized Projections, on the other hand, had lower recall and f1 scores, likely due to the loss of important information caused by the reduction in dimensions.

Finally, my analysis of the feature importances using Ran-dom Forest feature elimination provided important insights into which features were most important in predicting the target variable. This information can be used to guide further analysis and decision-making.

In conclusion, the performance of each technique is dependent on the specific characteristics of the data and the nature of the problem being solved. It is important to carefully consider the strengths and limitations of each technique when choosing the appropriate approach for a given problem. Additionally, the choice of dimensionality reduction technique can have a significant impact on the performance of clustering algorithms and machine learning models.

X. Conclusion

In conclusion, my analysis of the hotel and heart datasets using clustering algorithms and dimensionality reduction tech-niques has provided valuable insights into the performance and limitations of these techniques. I found that k-means clustering outperformed Expectation Maximization, and that PCA and ICA were the most effective methods for reducing the dimensions of the data. I also found that the choice of dimensionality reduction technique was an important factor in determining the performance of Neural Networks on projected data.

Additionally, my analysis of the feature importances using Random Forest feature elimination provided important infor-mation on which features were most important in predicting the target variable. This information can be used to guide further analysis and decision-making.

While my analysis provided important insights, it is im-portant to note the limitations of my approach. My analysis was limited to the specific datasets and techniques used, and other techniques or approaches may be more effective in different contexts. Additionally, the performance of clustering algorithms and dimensionality reduction techniques can be sensitive to factors such as the choice of hyperparameters, the initialization of the algorithm, and the specific characteristics of the data.

Despite these limitations, my analysis provides valuable in-sights into the performance and limitations of clustering algo-rithms and dimensionality reduction techniques and highlights the importance of careful consideration of these techniques in data analysis and decision-making.

References

[1] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.