EFFECTS OF LOGISTIC REGRESSION AND KERNEL PCA ON DATA

1. ABSTRACT

PURPOSE: To implement the Perceptron Rule for learning a separating hyperplane from data, and to implement kernel PCA using a Gaussian kernel function, which thus assist with machine learning topics such as classification and dimensionality reduction.

METHODS: In a comprehensive analysis, the performance of logistic regression and the Perceptron model is used for distinguishing between public and private US colleges. The Perceptron algorithm is implemented to update the separating hyperplane based on residual errors, while logistic regression estimates the hyperplane coefficients. The accuracy of both models is computed, along with the plotting of ROC curves and visualization of the separating hyperplanes. Also, the classification of Fisher's Iris data as the species I. setosa takes place using kernel PCA and k-means clustering. Kernel PCA is employed to project the data to 2D, followed by k-means clustering for classification. Visualization is achieved through scatter plots of the original labels and cluster indexes. This provides a thorough exploration of machine learning techniques, by providing insights into their effectiveness for the respective classification tasks.

RESULTS: The optimal thresholds for the DB and OM datasets were -0.2985 and 1.2116 respectively. Meanwhile, the Area Under the Curve (AUC) for both the DB and OM datasets is 0.93086 and 0.6071 respectively. Through Recover Operating Characteristic (ROC) analysis, the values of the optimal confusion matrices are also received for both the DB and OM datasets, and the values of these matrices are [255 65; 4 196] and [49 39; 131 301] respectively.

CONCLUSIONS: The optimal decision thresholds maximize certain performance metrics, depending on the specific goal (e.g., accuracy, sensitivity, specificity), and the DB classifier performs well here, with an optimal threshold of -0.2985 and a high AUC of 0.93086, which indicates a better differentiation between positive and negative instances. Meanwhile, the OM classifier has a lower AUC, suggesting less effective discrimination, and its optimal threshold reflects a different trade-off between true positive and false positive rates. Overall, the consideration of both AUC and optimal confusion matrices provides a comprehensive evaluation of classifier performance for the dataset that’s being analyzed.

2. INTRODUCTION

The scientific objective of this assignment is to evaluate the Perceptron Rule for binary classification and to use the Kernel PCA for dimensionality reduction and clustering, while facilitating a comparison between different classification techniques by examining the accuracies, ROC curves, and visualizations of the results, which thus effectively provides insights into when each technique may be most appropriate or effective.

To understand this report, one must have knowledge with regards to some of the background topics addressed in this report. One of the most important topics that this report is based upon is that of the Perceptron Rule, which is a simple supervised learning algorithm used for binary classification tasks. Its algorithm learns a separating hyperplane that distinguishes between two classes of data points. To understand the role of the Perceptron Rule, one must gain a sense of the key concepts that allow for the Perceptron rule to be applied. Based off the name itself, a perceptron rule comes off of the word 'perceptron', which is a simple computational unit that takes multiple binary inputs, applies weights to them, sums them up, and passes the result through an activation function to produce an output. Furthermore, the perceptron rule is based on linear separable data. Linear separability refers to the property of a dataset where classes can be separated by a straight line (or hyperplane) in the feature space. As part of the Perceptron rule, each input feature is associated with a weight in the perceptron model. These weights control the contribution of each feature to the output. Additionally, a bias term is often included, which allows the perceptron to capture patterns even when all input features are zero. The perceptron rate is also influenced by the learning rate, which is a hyperparameter that determines the size of weight updates during training. It influences the convergence speed and stability of the learning process. By understanding how the learning rate affects the training process, one can effectively model optimization. Lastly, the Perceptron algorithm iteratively updates the weights until convergence, meaning it reaches a point where further updates do not significantly improve performance. It's important to understand as to when and how the algorithm converges, since this can assess the algorithm's effectiveness and efficiency. Besides the Perceptron algorithm, understanding the concept of kernel functions and their role in nonlinear transformations is crucial. The knowledge of how kernel methods, such as Kernel PCA, can be used for dimensionality reduction and clustering tasks is necessary. A kernel function computes the inner product of data points in a higher-dimensional feature space without explicitly mapping the data to that space. Through this process, kernel functions enable nonlinear transformations of data by implicitly mapping the data into a higher-dimensional space where linear separation might be possible. Kernel PCA (Principal Component Analysis), meanwhile, extends traditional PCA by using kernel functions to perform nonlinear dimensionality reduction. Instead of computing the covariance matrix directly, Kernel PCA computes the kernel matrix and performs eigen decomposition on it to find the principal components in the feature space. This is also of assistance in improving clustering performance and capturing nonlinear relationships among variables while reducing dimensionality., especially when the data is not linearly separable, since the Kernel PCA projects data onto a lower-dimensional space while preserving nonlinear relationships. Lastly, one must have a basic understanding of probability and statistics to evaluate the performance of machine learning models, interpreting results, and making informed decisions.

To test the scientific question, multiple steps took place. Firstly, the data set was analyzed, with an understanding that the matrix size was in a csv file, containing 778 rows and 19 columns. Each row represented a college within the United States, and each column represented various statistics about each college. In this dataset, the goal was to use to a single artificial neuron to determine whether a US college is private or public. In order to complete this task, the Perceptron Rule algorithm was used in order to learn a separating hyperplane from data that represented US colleges as private or public. Once the Perceptron Rule classifier was trained, its accuracy in classifying colleges as private or public was assessed, while logistic regression was also applied to the same dataset. This evaluation aimed to determine how well the Perceptron Rule algorithm performed on the given dataset, while the accuracy of the logistic regression classifier was computed and compared with that of the Perceptron Rule classifier. From here, Receiver Operating Characteristic (ROC) curves were plotted to visualize the performance of both classifiers across different threshold values. The other task that was to be completed was that of implementing Kernel PCA using a Gaussian kernel function to reduce the dimensionality of Fisher's Iris data for subsequent k-means clustering. This was tested by performing Kernel PCA to the Iris data to project it into a lower-dimensional space, and then using k-means clustering to perform on the projected data. The results were visualized using scatter plots, with different colors representing different species or clusters, allowing for the assessment of the clustering performance.

3. METHODS

The first goal, as part of the task assigned, was to figure out as to how well a single artificial neuron can recognize as to whether a US college is private or public. To complete this task, the accuracy and area under the curve of two models are checked, those being logistic regression and the perceptron algorithm. The Perceptron algorithm is implemented to linearly separate training vectors, by finding a linear decision boundary that separates the input space into two classes. In order for the Perceptron algorithm to take place, a weight vector is initialized, and a maximum number of iterations is set (which in this case is 10000). Along with this, a learning rate is set, which controls the step size during weight updates. The algorithm can now be used to compute predictions using the dot product of the feature matrix and the weight vector. Meanwhile, residual errors can be calculated by comparing the predictions with the true labels. In this process, the weight vector is updated based on the misclassified data points using the Perceptron update rule. Eventually, a convergence should take place, either by verifying if all data points are correctly classified or if the maximum number of iterations is reached. As output, the Perceptron algorithm can thus provide the estimated weight vector and the number of iterations used. On the other hand, logistic regression estimates the probability of a binary outcome based on one or more independent variables, using the formula $P(Y = 1|X) = 1/(1 + e^{-w^\wedge T*x})$, where w is the weight vector and x is the feature vector. The weights are learned using the maximum likelihood estimation method, which thus optimizes the log-likelihood function, and the logistic function is used to map the linear combination of features to the probability of belonging to a particular class. When both models can be compared to each other, the accuracy of both models can be checked by comparing the predictions with the true labels. ROC curves are plotted to visualize the trade-off between true positive rate and false positive rate for different threshold values, as they illustrate the performance of a binary classification model across different threshold values. They plot the true positive rate (sensitivity) against the false positive rate (1 - specificity). Data points and separating hyperplanes, meanwhile, are visualized, for both models to understand their decision boundaries.

The second goal in this task was to check how well the classification of Fisher’s Iris data can take place as the species I. setosa. To complete this task, the Kernel PCA and k-means clustering was used. Kernel PCA is a non-linear dimensionality reduction technique that extends PCA by using kernel functions, and the kernel PCA method was used to reduce the data to 2D for k-means clustering. The implementation of the kernel PCA took place by first computing the Gram Matrix Computation. The Gram matrix is computed using the Gaussian kernel function, and the Gaussian kernel function measures the similarity between two data points in the feature space. Once the Gram matrix ‘s values are known, Eigen decomposition is performed on the Gram matrix to obtain eigenvalues and eigenvectors. The eigenvalues are sorted in descending order, and the corresponding eigenvectors are rearranged accordingly. In this manner, the Gram matrix is projected onto the principal components to reduce the dimensionality of the data to 2D. From here, K-means clustering is used to partition the reduced data into clusters. To start off, K cluster centroids are initialized randomly, and each data point is assigned to the nearest cluster centroid based on Euclidean distance. After this, cluster centroids are updated by taking the mean of the data points assigned to each cluster, after which the assignment and update steps are repeated until the cluster assignments no longer change significantly. For output, the cluster indices for each data point are returned. This data is then plotted into two separate functions using ‘gscatter’. In the first figure, the data is colored based on the true labels (0 as red and 1 as blue), while in the second figure, the data is colored based on the cluster indices (0 as magenta and 1 as cyan). Overall, through these algorithms and visualization techniques, the dimensionality of the Iris data using Kernel PCA is effectively reduced and the clustering of the reduced data takes place using K-means clustering, which therefore provides insights into the classification of Iris species.

Overall, the discriminative power and classification performance of logistic regression and the Perceptron model is evaluated in distinguishing between public and private US colleges. This is done by implementing the Perceptron algorithm and logistic regression, computing accuracy metrics, and plotting ROC curves to assess discrimination ability. Similarly, a thorough examination of the discriminative capability and classification performance takes place by identifying the species I. setosa within Fisher's Iris dataset. Principal Component Analysis (PCA) and k-means clustering are utilized for dimensionality reduction and classification, respectively. The analysis, meanwhile, includes computing ROC curves, Area Under the Curve (AUC), and optimal thresholds to evaluate discrimination ability. Visualizations, such as scatter plots of original labels and cluster indexes, interpret reduced-dimensional datasets, providing insights into data characteristics and classifier performance. Together, these analyses offer valuable insights into the effectiveness of machine learning techniques for classification tasks such as the ones given in this assignment.

4. RESULTS

The results are based around an assessment of the performance and effectiveness of different machine learning models, including logistic regression and artificial neural networks (ANN), or the Perceptron model, and the kernel PCA followed by k-means clustering. To distinguish and evaluate the classification performance of these models., metrics such as accuracy and Area Under the Curve (AUC) are computed. Additionally, Receiver Operating Characteristic (ROC) curves are plotted to visualize the trade-off between true positive rate and false positive rate, which thus gives an idea of the optimal decision thresholds. The hyperplane representation showcases how well the models separate the classes in the dataset, providing insights into their discriminatory ability. Meanwhile, the kernel PCA projection with original labels and k-means clustering demonstrates the effectiveness of dimensionality reduction and clustering in classifying instances. These results are based around confusion matrices, with thus offer a detailed breakdown of the models' performance in terms of correct and incorrect classifications. Visualizations such as ROC curves, hyperplanes, and 2D data plots provide an understanding of the models' capabilities, which aids in the interpretation of class distribution and patterns within the datasets. Overall, these analyses offer valuable insights into the performance of the machine learning models, with therefore provides one with an opportunity to make informed decisions and model refinement.

Table 1: The accuracy and AUC scores of the logistic regression model, as well as of the ANN, or Perceptron Algorithm model.

Model	Accuracy Score	AUC Score
Logistic Regression	0.911197	0.959960
Artificial Neural Networks (ANN)	0.876448	0.949541

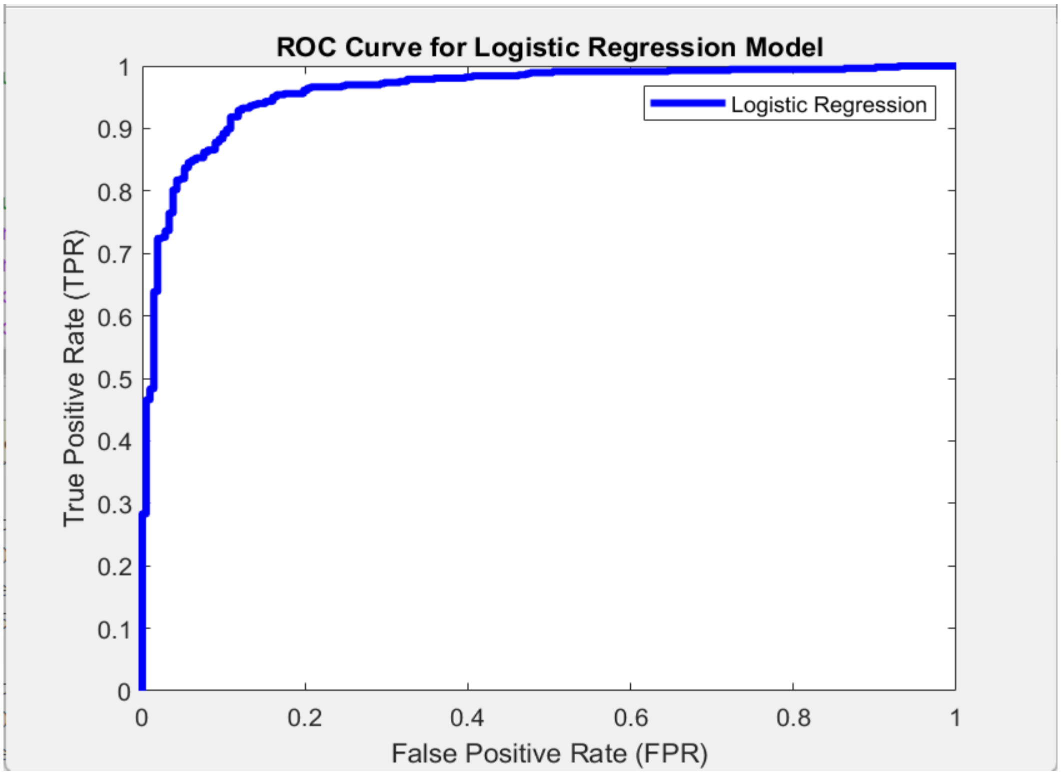

Figure 1: ROC curve for the Logistic Regression Model. The y-axis represents the True Positive Rate (TPR) for both of the datasets, while the x-axis represents the False Positive Rate (FPR) for both of the datasets.

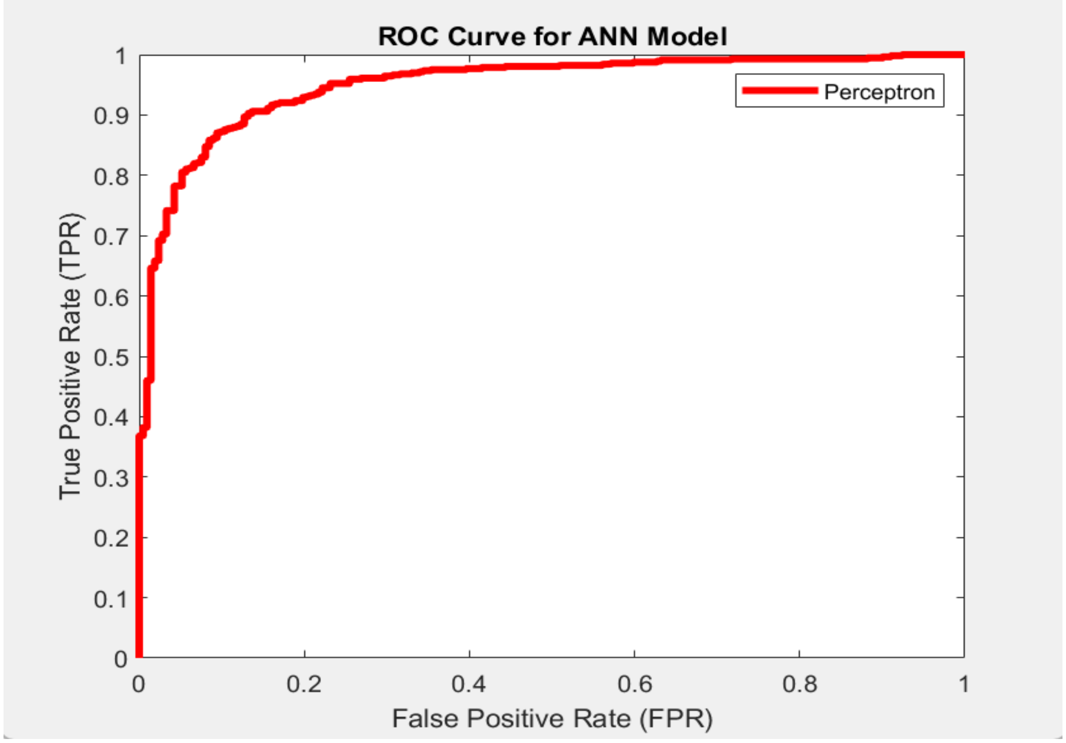

Figure 2: ROC curve for the Artificial Neural Networks (ANN) Model. The y-axis represents the True Positive Rate (TPR) for both of the datasets, while the x-axis represents the False Positive Rate (FPR) for both of the datasets.

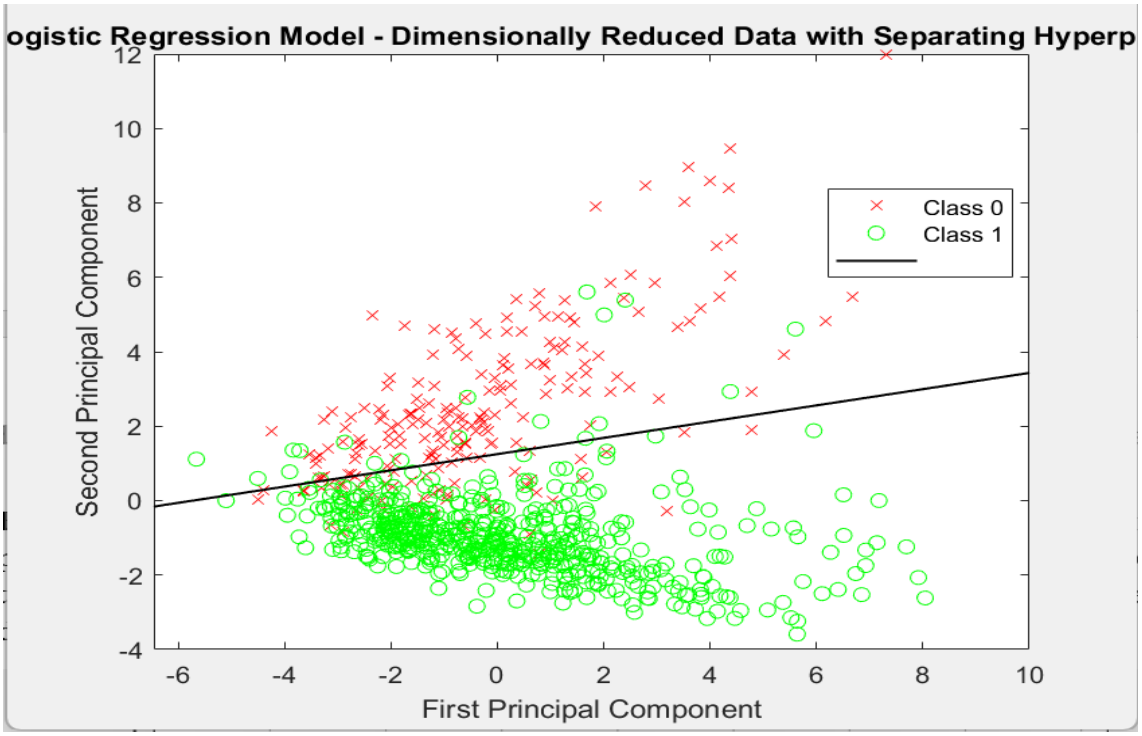

Figure 3: 2D plot representing the logistic regression model, with a dimensional reduction on the data along with a separating hyperplane. The first Principal Components is denoted on the x-axis, while the second Principal Component is denoted on the y-axis.

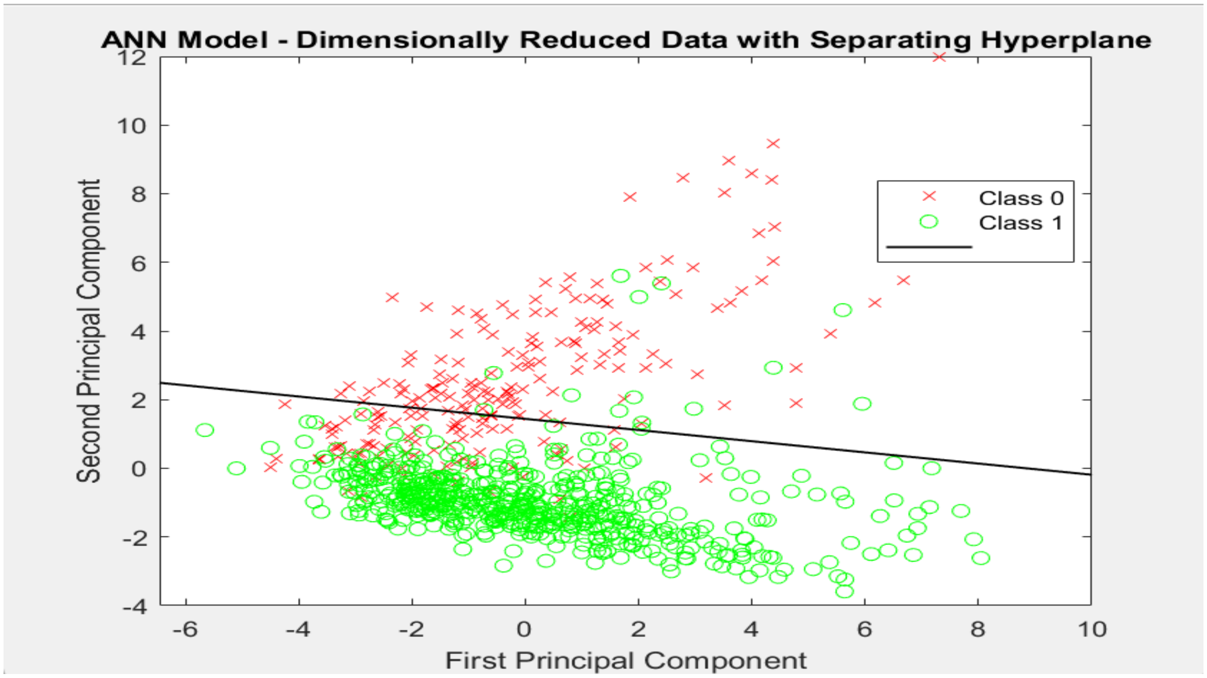

Figure 4: 2D plot representing the ANN model, with a dimensional reduction on the data along with a separating hyperplane. The first Principal Components is denoted on the x-axis, while the second Principal Component is denoted on the y-axis.

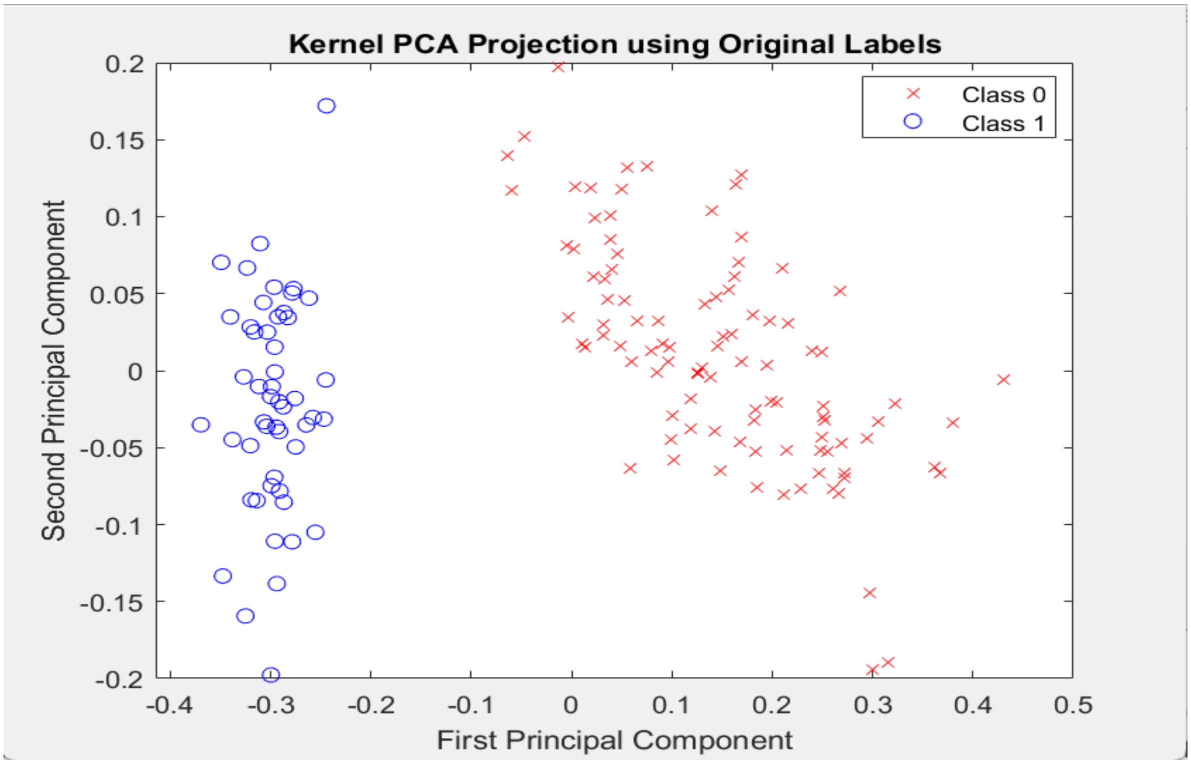

Figure 5: Kernel PCA Projection using the original labels of data that's been provided. The first Principal Components is denoted on the x-axis, while the second Principal Component is denoted on the y-axis.

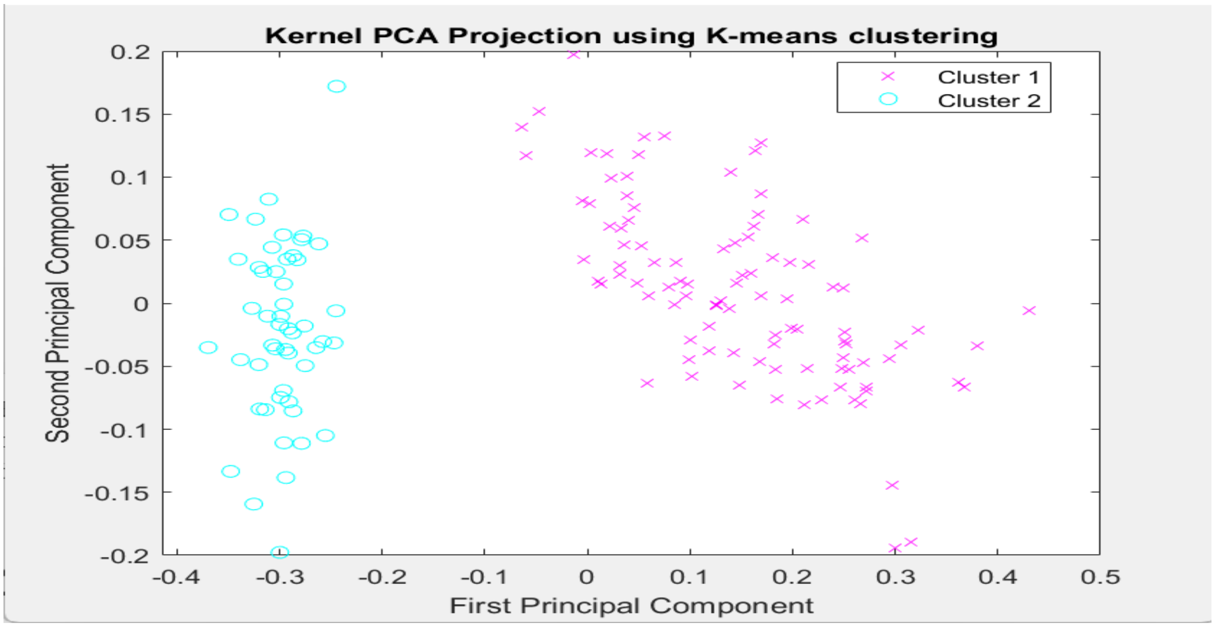

Figure 6: Kernel PCA Projection using the K-means Clustering that has been performed on the data. The first Principal Components is denoted on the x-axis, while the second Principal Component is denoted on the y-axis.

5. DISCUSSION

To separate and classify US colleges as private or public, a comparative analysis between logistic regression and the Perceptron model was conducted, which provided various evaluation metrics and visualization techniques. Similarly, dimensionality reduction via PCA revealed trends into the classification performance of both models. Notably, accuracy rates and AUC values served as key metrics, giving an idea on the discriminatory ability of each model. A clear trend emerged, indicating logistic regression's superiority in accurately distinguishing between private and public colleges, as evidenced by higher accuracy rates and larger AUC values. Visualizations, including ROC curves and separating hyperplanes, provided further clarity on the classification established by each model. Meanwhile, in the analysis of Fisher's Iris dataset, kernel PCA and k-means clustering were used to classify Iris species. Through PCA, significant trends were revealed in clustering patterns and the preservation of data structures. Notably, the DB index values served as crucial metrics, guiding the selection of optimal feature pairs for dimensionality reduction. Lower DB index values provided more favorable dimensionality reduction outcomes, suggesting the effectiveness of certain variables in capturing data structures. Scatter plots visually reinforced these trends, showcasing clusters and patterns that explained the efficiency of PCA in capturing data structures. The scatter plot corresponding to the optimal feature pair exhibited clear separation of data points, which further agreed with the identified trend in DB index values. Overall, the analysis provided insights into the effectiveness of PCA and clustering techniques in classification tasks, which thus offered a deeper understanding of the data patterns and classification performance of the models. These identified trends and observations will be further discussed in the next paragraph.

The first result that was received while completing this task was that the accuracy of logistic regression was approximately 0.911197. In percentage terms, this means that the logistic regression model correctly indicated and classified approximately 91% of colleges in the dataset. This high accuracy suggests that logistic regression effectively learned the underlying patterns in the data and made accurate predictions regarding whether a college is private or public based on the available features. The accuracy metric serves as a measure of the model's performance, demonstrating its ability to correctly classify a significant majority of the instances. Based on the result received, such a high accuracy is indicative of logistic regression's capability to generalize well to unseen data and effectively discriminate between the two classes of colleges. Through the high accuracy achieved by logistic regression, one can also trust in it as a classification algorithm for particular task, providing confidence in its ability to make accurate predictions in real-world scenarios. On the other hand, the accuracy of the ANN model was approximately 0.876448. Here, In percentage terms, this means that the logistic regression model correctly indicated and classified approximately 88% of colleges in the dataset. This suggests that the ANN model was less effective in accurately categorizing colleges into their respective groups, when compared with the logistic regression model. While the accuracy of 0.876448 still indicates a high level of performance, the lower accuracy compared to logistic regression implies that the ANN may have faced more challenges in capturing the patterns within the data or in looking at unseen instances. This slight discrepancy in accuracy between logistic regression and the ANN gives an understanding of the importance of exploring different machine learning algorithms and their performance on specific datasets. Although, despite achieving a slightly lower accuracy, the ANN model's performance remains commendable, demonstrating its capability as an effective tool for classification tasks. Yet, the lower accuracy of the ANN model prompts further investigation to identify potential areas for improvement and optimization, by potentially uncovering insights that could inform future model refinement and selection. These results further prove their point when the Area Under the Curve (AUC) score is looked at for the logistic regression model and the ANN model. Logistic regression demonstrated an AUC value of 0.960, which thus shows its ability to distinguish between private and public colleges with high sensitivity and specificity. Based on the AUC value, it seems as though logistic regression achieved a strong performance in correctly ranking the colleges, with a higher probability of assigning a higher score to private colleges compared to public ones. This is because AUC values close to 1.0 indicate excellent discrimination ability, suggesting that logistic regression effectively captured the true positive rate while minimizing the false positive rate. This strong discrimination ability is crucial for logistic regression's skill in distinguishing between private and public colleges based on the provided features. Meanwhile, the ANN model yielded an AUC of 0.950, indicating slightly lower discriminatory ability compared to logistic regression. While still achieving a high AUC value, the ANN model demonstrated a slightly reduced capacity to distinguish between private and public colleges compared to logistic regression. This AUC value suggests that the ANN model produced excellent distributions of predicted probabilities for the two classes as well, but with a marginally lower discrimination ability than logistic regression. One of the main reasons why logistic regression does better here is because its linear decision boundary assumption may have been well-suited for distinguishing between private and public colleges based on the available features. The ANN model's non-linear nature, in contrast, might have introduced additional complexity, potentially impacting its classification performance. Nonetheless, an AUC value of 0.950 still reflects a robust performance by the ANN model, indicating its capability to make accurate classifications. While logistic regression outperformed the ANN model slightly in terms of discriminatory ability based on the accuracy and AUC score, both models still demonstrated strong classification techniques overall, representing their effectiveness in distinguishing between private and public colleges. These results highlight the importance of selecting appropriate machine learning algorithms tailored to the specific characteristics of the dataset and the classification task. The interpretation of the results gives an idea of the need to consider the assumptions and limitations of each model when evaluating the performance of both models in real-world applications. This analysis can now be looked at from the perspective of the graphs that have been created. The ROC curves for both models have been outputted, and the ROC curve of logistic regression consistently lies above or to the left of the ANN's ROC curve across different threshold values. This indicates that logistic regression achieves higher true positive rates while maintaining lower false positive rates compared to the ANN. Due to the fact that the ROC curve for the logistic regression model immediately gets closer to the top left-hand side, its area under the curve, as computed, is also higher. A higher AUC value, as mentioned, indicates better overall discrimination performance. Therefore, if logistic regression has a higher AUC compared to the ANN, it suggests that logistic regression achieves superior discrimination between the two classes, further supporting the argument for its effectiveness over the ANN model. With both of these reasons in mind, once again, it’s safe to suggest that logistic regression outperforms the ANN model in terms of discrimination ability and overall classification performance. Lastly, the graphs of the dimensionally reduced data with separating hyperplanes can be looked at for both models. As can be seen from the graphs, within the ANN model, there are more false positives and false negatives, as a whole, in comparison to the logistic regression model. This can be seen by looking at where the number of class 0 points fall into the class 1 zone, and vice versa. Also, within the ANN model, the data points that are a part of class 0 tend to cluster further apart. This is the case with linear regression as well, to some extent, but the distance between its data points, whether in class 0 or in class 1, is overall less than the distance between the data points in the ANN model. Based on these characteristics, knowing that the logistic regression model has more true values, less false values, and overall less distance between its data points, it can once again be confirmed that the logistic regression model is a batter classifier for this data in comparison to the ANN model. Now, the kernel PCA projection can be looked at, both with the original data points, as well as with the k-means clustering. In both cases, it is evident that the clustering that’s taken place is well done, considering that the data points between class 0 and class 1 are quite far away from each other. Since the clusters are well-separated in the kernel PCA projection, it indicates that the kernel function effectively transforms the data into a higher-dimensional space where it is linearly separable. Meanwhile, if the centroids were looked at for both of the graphs, then, it can be clearly seen that they are located at the centre of their respective cluster. This once again provides information about the effective clustering that took place, and so, it can be said that the classification of Fisher’s Iris data as the species I. setosa was done smoothly. However, one discrepancy that can be seen is that of the cluster sizes within both graphs not being balanced enough. This means that K-means didn’t quite effectively partition the data into clusters of similar sizes, which thus means that a disproportionate influence on the clustering results took place, potentially skewing the interpretation of patterns and relationships within the data. This raises concerns about the reliability of the clustering results, as it may lead to biased interpretations and misrepresentations of the dataset's characteristics. Hence, while the clustering process aims to classify and separate distinct groups, caution and safety is warranted in interpreting the results due to the observed imbalance, with respect to the number of data points, within each cluster size.

In summary, logistic regression outperforms the ANN model with an accuracy of 91.12% compared to 87.64%. Logistic regression also exhibits a higher AUC of 0.960, indicating superior discriminatory ability. This is reinforced by the ROC curves, where logistic regression's curve is closer to the top-left corner, reflecting better performance. Regarding Kernel PCA and k-means clustering, the scatter plots reveal distinct clusters and clear boundaries between species, demonstrating the effectiveness of these techniques for dimensionality reduction and clustering analysis.

6. REFERENCES

1. Ellis RE. Patterns – Classification – Single Artificial Neuron [Internet]. Queen’s University; 2021 [cited 2024 Mar 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class27.pdf
2. Ellis RE. Classification – Logistic Regression [Internet]. Queen’s University; 2021 [cited 2024 Mar 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class31.pdf
3. Ellis RE. Nonlinear Separation – Embeddings and Gram Matrix [Internet]. Queen’s University; 2021 [cited 2024 Mar 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class32.pdf
4. Ellis RE. Nonlinear Separation – Kernel PCA [Internet]. Queen’s University; 2021 [cited 2024 Mar 27]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class33.pdf
5. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics [Internet]. 2002 Oct;35(5-6):352–9. Available from: https://www.sciencedirect.com/science/article/pii/S1532046403000340
6. Kapourani A. Learning and Data Lab 3 Informatics 2B (2018/19) K-means Clustering & PCA [Internet]. Available from: https://www.inf.ed.ac.uk/teaching/courses/inf2b/labs/learn-lab3.pdf