ECE657A

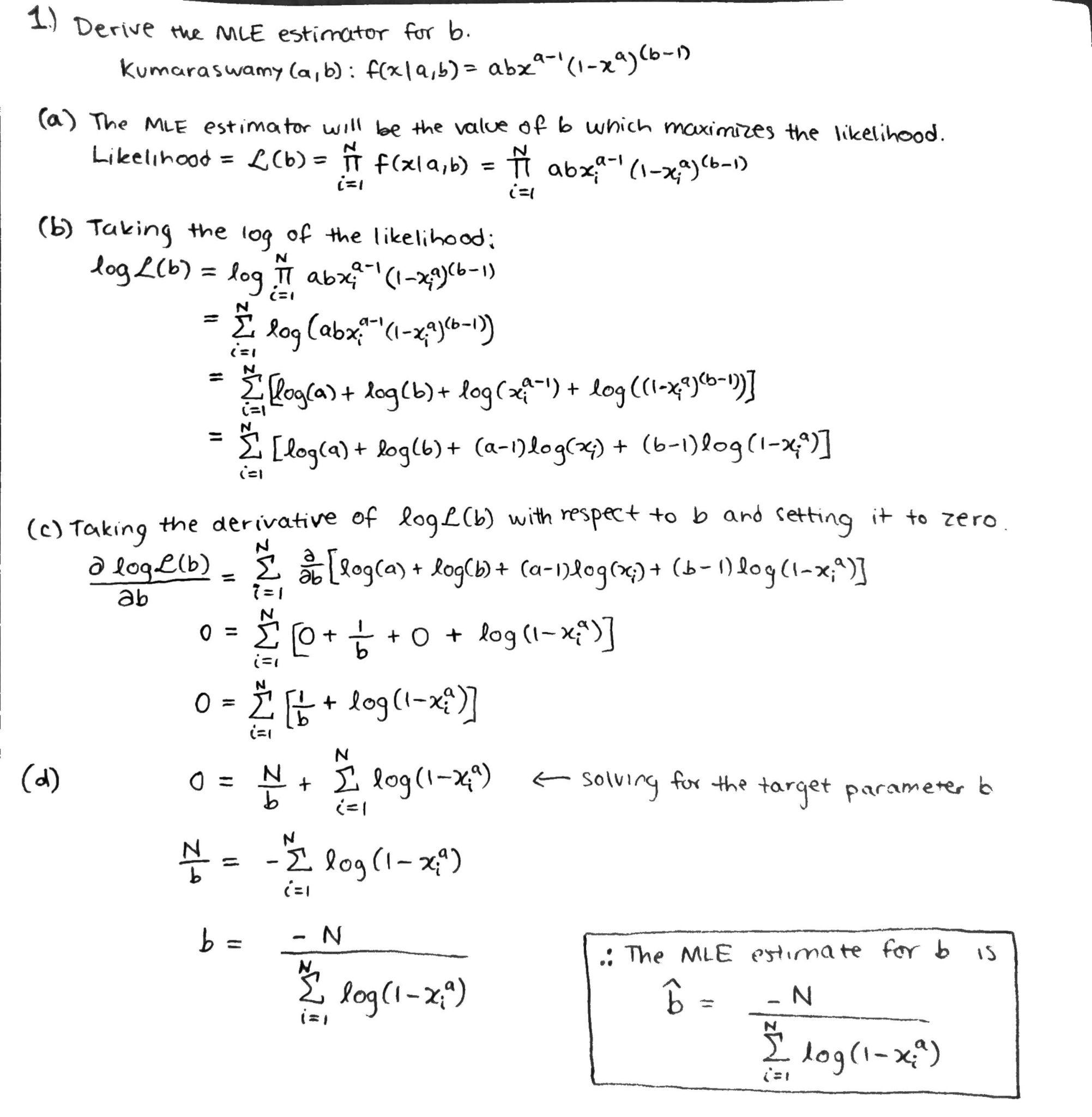
Assignment 2

Munia Humaira and Mishaal Khan

March 2, 2025

#### 1. MLE and MAP Derivation

#### 1.1 Deriving the MLE estimator for



#### 1.2 Deriving the MAP estimator for

#### 

#### 2. Regression

For each dataset, Grid Search Cross Validation with 5 folds was used on the training data to find the best hyperparameter combinations. For the k-NN regression model, the number of neighbours and the weight function were varied. For the random forest regressor model, the number of estimators and their maximum depth were varied. For the gradient tree boosting regressor model, again the number of estimators and their maximum depth were varied. These hyperparameter choices were chosen as they are key factors for these models. Using the best estimator found from grid search, the predictions on the test data were obtained and used to calculate the root mean squared error (RMSE) of the models on the test data.

#### 2.1 Regression with Wine Dataset

The kNN model achieved the lowest RMSE : 0.6104, suggesting that it performs better than the other models in predicting wine quality. The small variance 0.0001 indicates that the model is relatively stable across different folds.

Table 1. Cross validation results for wine dataset

|  | k-NN | Random Forest | Gradient Tree Boosting |
| --- | --- | --- | --- |
| Best hyperparameters from cross validation | No. of neighbours = 14,  Weight function = distance | No. of estimators = 43, maximum depth = 9 | No. of estimators = 43, maximum depth = 6 |
| Mean score across folds for best estimator | -0.408 | -0.446 | -0.422 |
| Variance of scores across folds for best estimator | 0.0001 | 0.0002 | 6.75e-05 |

The Random Forest model performed marginally worse than kNN, as evidenced by its highest RMSE: 0.6603. It may be marginally more sensitive to changes in training data, as seen by the larger variance 0.0002 (in comparison to kNN). While it did not outperform kNN, the Gradient Boosting model did better than Random Forest. It is the most stable model among the three, with the lowest variance of 0.000068 and a comparatively high predictive power i.e., RMSE of 0.6308.

#### 2.2 Regression with Abalone Dataset

For the abalone dataset, the categorical column of “Sex” was one-hot encoded and the numerical features were standardized using sklearn’s StandardScaler. The regression task is to predict the number of rings.

Table 2. Cross validation results for abalone dataset

|  | k-NN | Random Forest | Gradient Tree Boosting |
| --- | --- | --- | --- |
| Best hyperparameters from cross validation | Number of neighbours = 17,  Weight function = distance | Number of estimators = 43, maximum depth = 8 | Number of estimators = 43, maximum depth = 6 |
| Mean score across folds for best estimator | -4.909 | -4.812 | -4.839 |
| Variance of scores across folds for best estimator | 0.180 | 0.093 | 0.062 |

Based on RMSE values, all models had very similar performance, with the random forest model performing the best having a RMSE of 2.057, the gradient tree boosting model having a RMSE of 2.094, and the k-NN performing the worst with a RMSE of 2.104, although the difference between the RMSE values among the models is marginal. The RMSE indicates the error that the model obtains when predicting the number of rings of the test data. A high cross validation mean score and low variance are desired as this indicates a model can generalize well and is not sensitive to the fold’s split [1]. In this case, gradient tree boosting has the lowest variance and highest mean score, while the k-NN has the highest variance and the lowest mean score.

#### 2.3 Regression with Forest Fires Dataset

First, the dataset was checked for missing values, and there were no missing values. The categorical columns of “month” and “day” were one-hot encoded, and the numerical features were scaled with sklearn’s StandardScaler. In the paper “A Data Mining Approach to Predict Forest Fires Using Meteorological Data”, the target variable ‘area’ was transformed by applying the function ln(x+1) [2][3]. This helps to reduce skew from most fires having a small burned area [3].This is the approach taken here.

Table 3 shows the comparison of the forest fires dataset without transforming the area column and with transforming the area column.

Table 3. Regression results for forest fire dataset without and with transformation of output

| Model | Without transformation RMSE (rounded to 3 decimal places) | With transformation RMSE (rounded to 3 decimal places) |
| --- | --- | --- |
| k-NN | 109.337 | 1.502 |
| Random Forest | 109.707 | 1.500 |
| Gradient Tree Boosting | 108.994 | 1.473 |

As seen in Table 3, the RMSE values of the test set are much lower after transforming the area column. Therefore, the remainder of the experiments with the forest fires dataset are performed using the dataset with the transformed area column.

Table 4. Cross validation results for forest fires dataset

|  | k-NN | Random Forest | Gradient Tree Boosting |
| --- | --- | --- | --- |
| Best hyperparameters from cross validation | Number of neighbours = 18,  Weight function = uniform | Number of estimators = 13, maximum depth = 2 | Number of estimators = 3, maximum depth = 2 |
| Mean score across folds for best estimator | -1.973 | -1.943 | -1.916 |
| Variance of scores across folds for best estimator | 0.218 | 0.202 | 0.195 |

Based on RMSE values, the gradient tree boosting model performed the best with a RMSE of 1.473. It also has the highest cross validation mean score and lowest variance among the models. For the random forest model, the RMSE was 1.500. The k-NN model performed the worst, yielding a RMSE of 1.502.

#### 3. Representation Learning

For the representation learning experiments with PCA and LDA, the order of steps from [4] was followed, in which the datasets were split into training and test sets first, then standardized, then the training data was used to generate the scree plot and determine the number of components to choose, and lastly the test set was transformed from the PCA fit on the training data. The other feature extraction method chosen to analyze for each dataset was LDA.

#### 3.1 Representation Learning with Wine Dataset

In Figure 1, the t-SNE image is plotted to visualize the high dimensional data into 2D space. The first two components are shown. However, the separation between different classes cannot be distinguished properly.

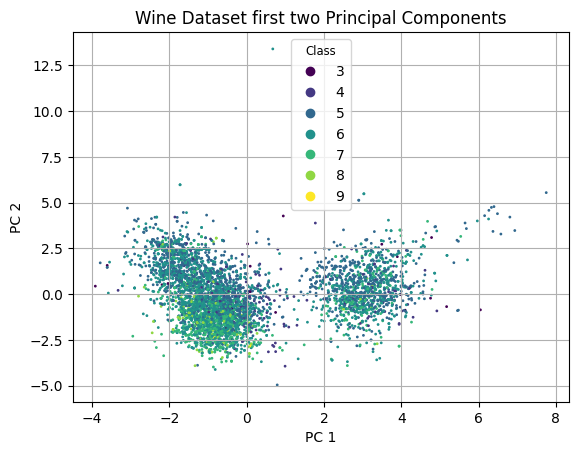
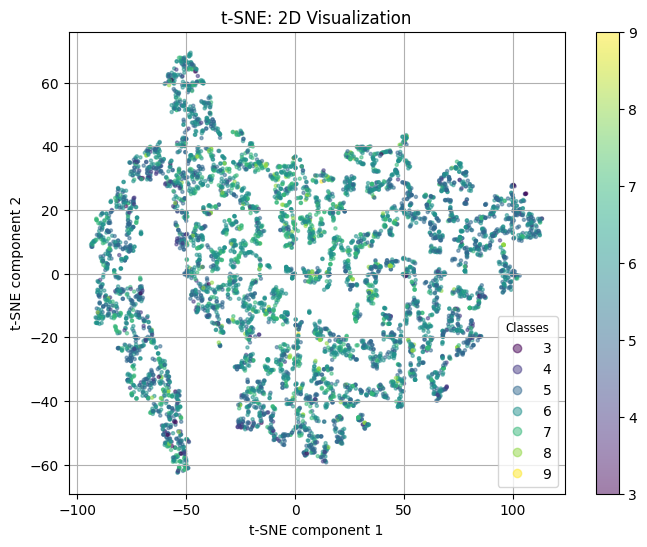


Figure 1. t-SNE 2D visualization of the wine dataset Figure 2. PCA of Wine Dataset

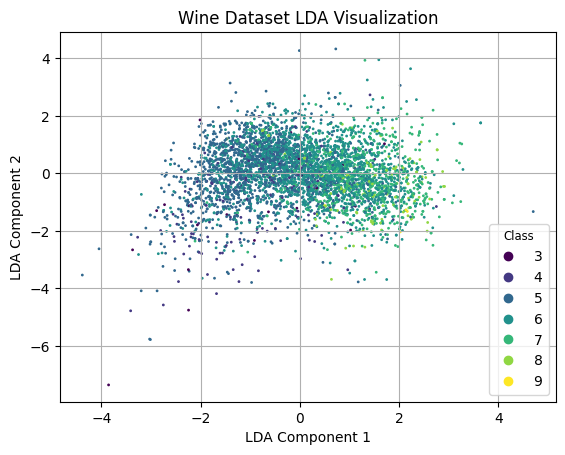
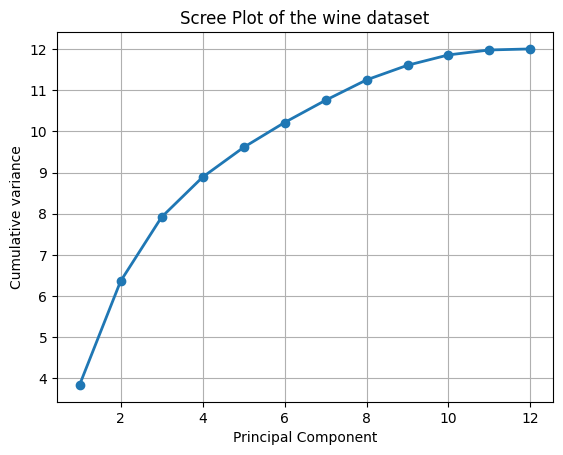


Figure 3. Wine Data Visualization with LDA Figure 4. Scree Plot of the wine dataset

In Figure 2, the first two principal components have been shown. Again, the separation between classes is not distinguishable. It is not apparent what creates the appearance of two separate groups, as there is a mixture of classes in both groups.The total variance explained by the components decreases as expected, with the first component having the highest variance explained of 3.8405 or 32% of the total variance.

The plot of the first two Linear Discriminant Analysis (LDA) components on the training set is displayed in Figure 3. Similar classes points seem to be close together, although not clearly distinguishable. The first component explains 86% of the variance. The scree plot is displayed in Figure 4. The number of reduced components is 6 because the cumulative variance becomes saturated beyond that. The decreased dimensionality representation from PCA and LDA was used to rerun the regression findings for the wine dataset.

#### 3.2 Representation Learning with Abalone Dataset

The abalone dataset after preprocessing has 10 features. Figure 5 shows the t-SNE visualization of the entire abalone dataset, showing the first two components. In this case, the separation between classes is not very clear, at least by observing the first two components. Data points of class 3 are visually close to each other.

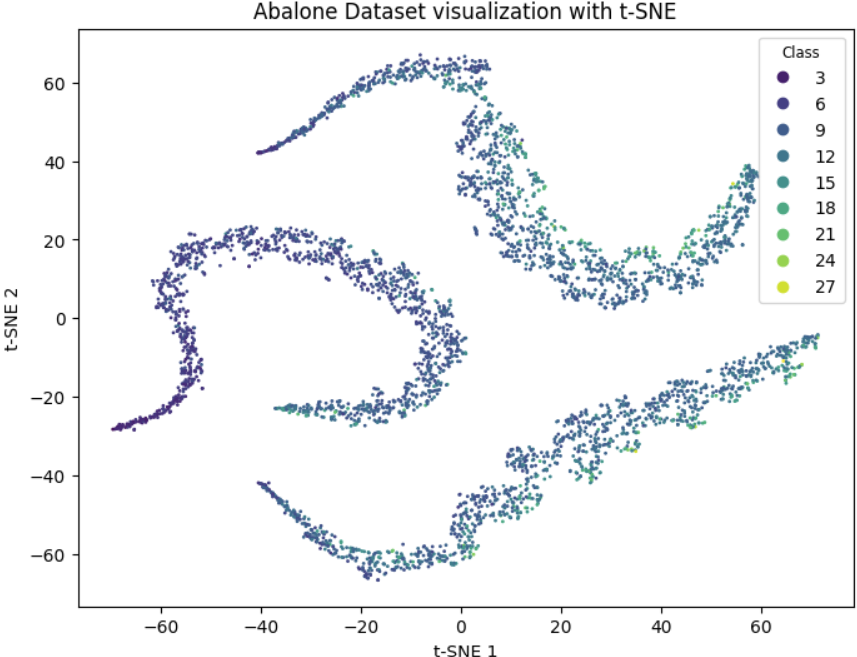
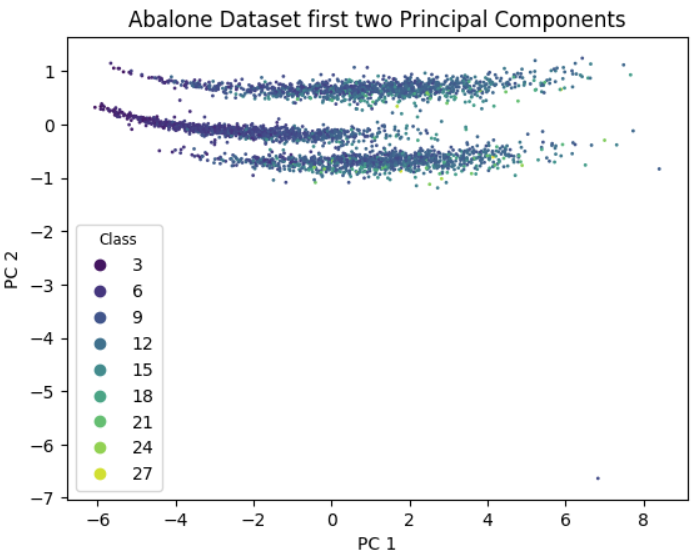


Figure 5. Visualization of abalone dataset with t-SNE Figure 6. Visualization of abalone dataset with PCA

Figure 6 shows the plot of the first two components from PCA on the training set. PCA is an unsupervised linear feature extraction method. Again, there is no clear pattern between the classes. The total variance explained by the components decreases as expected, with the first component having the highest variance explained of 6.441 or 84.001% of the total variance.

Figure 7 shows the plot of the first two components from Linear Discriminant Analysis (LDA) on the training set. There is no clear distinction between the classes but datapoints with a smaller number of rings seem to be located around a similar area and likewise for those with a larger number of rings.The first two components explain most of the variance, with the percentage of variance explained by the first component being 70.787% and that by the second component being 22.017%.

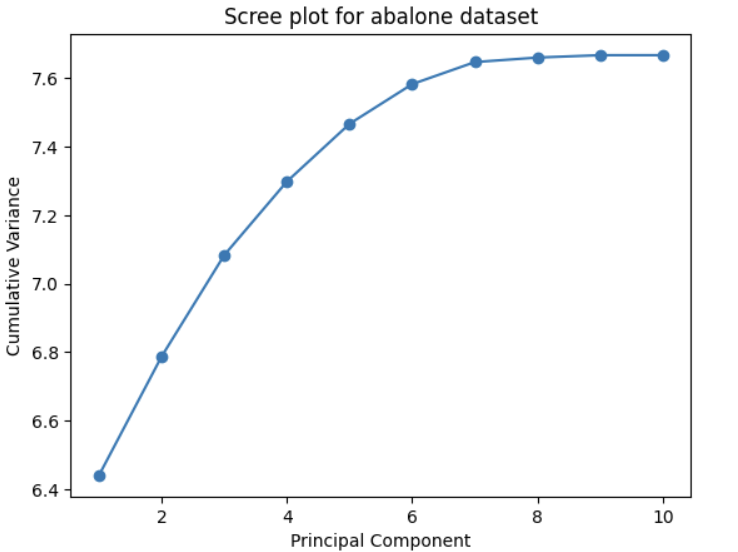
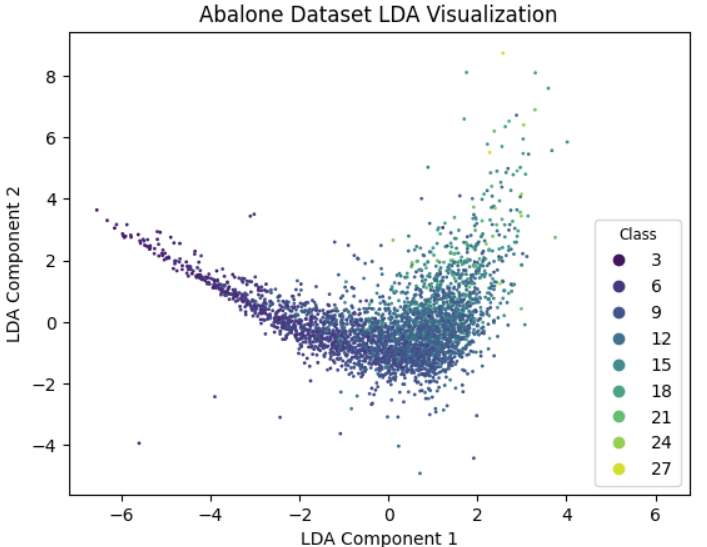


Figure 7. Visualization of abalone dataset with LDA Figure 8. Scree plot of abalone dataset

Figure 8 shows the scree plot of the cumulative variance versus the principal components from PCA. The ‘knee’ point in the curve was taken to be at the 6th component, as beyond that there is not much increase in the cumulative variance, which implies that the first 6 components will be enough.

The regression results were rerun for the abalone using the reduced dimensionality representation from PCA and LDA, using 6 components. The RMSE results are shown in Table 5. Using the dataset reduced through PCA did not result in an improved RMSE, however the dataset reduced through LDA resulted in slight reductions to RMSE values.

#### 3.3 Representation Learning with Forest Fires Dataset

The t-SNE, PCA, LDA and Scree plot visualizations are shown in the FIgure 9, Figure 10, Figure 11 and Figure 12, respectively. For the forest fire dataset, in order for LDA to work, the class (area) needed to be binned, so the plots for PCA and LDA are shown for the binned classes. None of the visualizations show a clear distinction between classes. The first component of the PCA explains 13% of the variance and the first component of the LDA explains 63% of the variance.

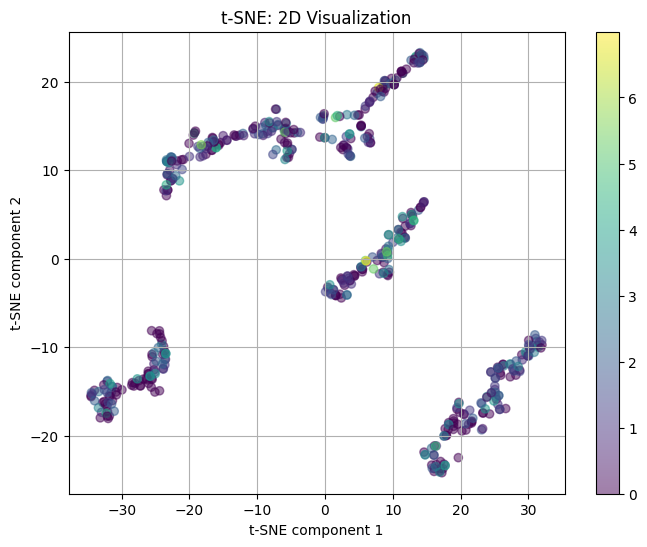
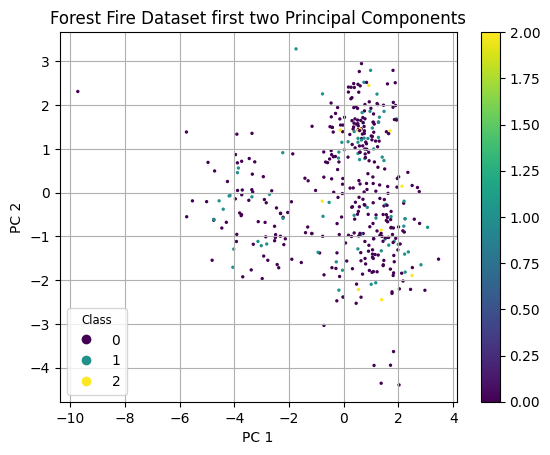


Figure 9. t-SNE visualization of forest fire dataset Figure 10. PCA Visualization of forest fire

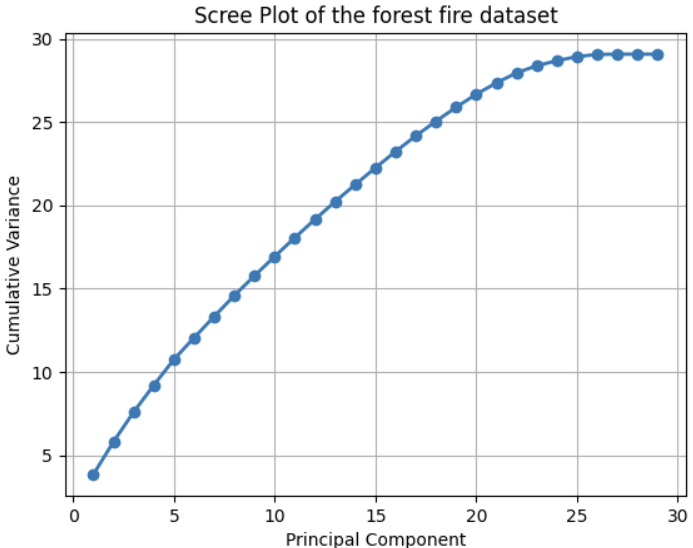
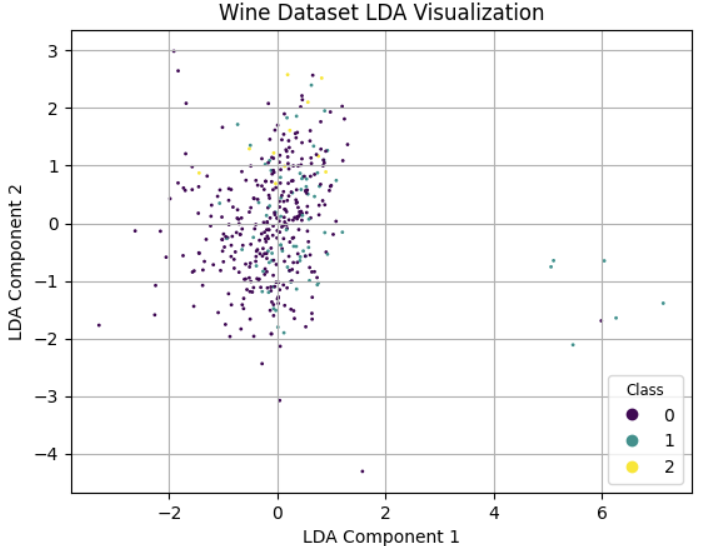


Figure 11. LDA visualization of forest fire Figure 12. Scree plot of forest fire dataset

With the three models, the regression has been performed in the same manner with a reduced component number of 2. The forest fire dataset's LDA and PCA show little variation. Nevertheless, the regression of the reduced dataset resulted in a drop in the RMSE value, though the above plots do not show clear patterns for the components plotted.

The RMSE results for each of the nine datasets are displayed in Table 5. For the abalone dataset, applying LDA resulted in improved RMSE results across the three models. For the forest fires dataset, applying PCA or applying LDA improves RMSE, while for wine no such trend was observed, in fact it had the opposite effect. This may indicate that for the wine dataset, feature extraction is not a good strategy, whereas for the forest fire dataset it is.

Table 5. Summary of performance for each dataset and model

| Dataset | RMSE Result for model choice (rounded to 3 decimal places) | | |
| --- | --- | --- | --- |
| k-NN | Random Forest | Gradient Tree Boosting |
| wine | 0.610 | 0.660 | 0.631 |
| wine-pca | 0.620 | 0.674 | 0.648 |
| wine-lda | 0.615 | 0.663 | 0.657 |
| abalone | 2.104 | 2.057 | 2.094 |
| abalone-pca | 2.153 | 2.067 | 2.083 |
| abalone-lda | 2.046 | 1.997 | 1.984 |
| forest-fires | 1.502 | 1.500 | 1.473 |
| forest-fires-pca | 0.468 | 0.462 | 0.464 |
| forest-fires-lda | 0.467 | 0.456 | 0.458 |

**References**

[1] V. Chugani, “A Comprehensive Guide to K-Fold Cross Validation,” DataCamp, https://www.datacamp.com/tutorial/k-fold-cross-validation (accessed Mar. 2, 2025).

[2] P. Cortez and A. Morais. "Forest Fires," UCI Machine Learning Repository, 2007. [Online]. Available: <https://doi.org/10.24432/C5D88D>.

[3] P. Cortez and A. Morais, “A Data Mining Approach to Predict Forest Fires using Meteorological Data,” *2007*, Available: https://core.ac.uk/download/pdf/55609027.pdf (accessed Mar. 2, 2025).

[4] sarayu gouda, “PCA(Principal Component Analysis) In Python,” Medium, https://medium.com/@sarayupgouda/pca-principal-component-analysis-in-python-f9836c25acb9 (accessed Mar. 2, 2025).