Question 6 : Final Results

Table I : Summary Accuracy Scores for Abalone Dataset

Model	**Setting**	**Dataset**	**Accuracy(%)**
KNN	n_neighbors = 84 (from prev assgn)	abalone-raw	25.1196172248804
KNN	n_neighbors = 84 (from prev assgn)	abalone-pca	25.1196172248804
KNN	***n_neighbors = 84 (from prev assgn)***	***abalone-lda***	***25.717703349282296***
Naive Bayes	Gaussian	abalone-raw	22.0746067673266
Naive Bayes	Gaussian	abalone-pca	22.16995673724322
Naive Bayes	***Gaussian***	***abalone-lda***	***23.318454001088732***
Naive Bayes	***Multinomial***	***abalone-raw***	***16.6625791479242***
Naive Bayes	Multinomial	abalone-pca	16.495086382259405
Naive Bayes	Multinomial	***abalone-lda***	16.495086382259405
Naive Bayes	Complement	abalone-raw	14.7475861673782
Naive Bayes	Complement	abalone-pca	19.29536143024955
Naive Bayes	***Complement***	***abalone-lda***	***23.006790247256685***
Decision Tree	{'max_depth': 4}	abalone-raw	26.26310059307223
Decision Tree	***{'max_depth': 4}***	***abalone-pca***	***26.33515743632352***
Decision Tree	{'max_depth': 3}	abalone-lda	25.8082972
Random Forest	***{'max_depth': 8, 'n_estimators': 183}***	***abalone-raw***	***27.364725095264014***
Random Forest	{'max_depth': 6, 'n_estimators': 23}	abalone-pca	27.029653611437414
Random Forest	{'max_depth': 6, 'n_estimators': 143}	abalone-lda	27.100793628054898
Gradient Tree Boost	***{'max_depth': 3, 'n_estimators': 80}***	***abalone-raw***	***25.35400968398132***
Gradient Tree Boost	{'max_depth': 3, 'n_estimators': 160}	abalone-pca	23.89364811047761
Gradient Tree Boost	{{'max_depth': 3, 'n_estimators': 80}	abalone-lda	24.22851903847807

Table II : Summary Accuracy Scores for WINE Dataset

Model	Setting	Dataset	Accuracy(%)
Gradient Tree Boosting	{'max_depth': 6, 'n_estimators': 150}	wine-pca	46.9617457
Gradient Tree Boosting	***{'max_depth': 6, 'n_estimators': 100}***	***wine-lda***	***49.6547285***
Gradient Tree Boosting	{'max_depth': 6, 'n_estimators': 20}	wine-raw	47.5932374
KNN	n_neighbors = 84 (from prev assgn)	wine-raw	68.3076923
KNN	n_neighbors = 84 (from prev assgn)	wine-pca	68.2307692
KNN	**n_neighbors = 84 (from prev assgn)**	**wine-lda**	**68.5384615**
Naive Bayes	Gaussian	wine-raw	30.1210399
Naive Bayes	Gaussian	wine-pca	44.9918399
Naive Bayes	***Gaussian***	***wine-lda***	***52.9787174***
Naive Bayes	Multinomial	wine-raw	41.512465209924784
Naive Bayes	Multinomial	wine-pca	43.65092674838634
Naive Bayes	***Multinomial***	***wine-lda***	43.65092674838634
Naive Bayes	Complement	wine-raw	36.955421330017174
Naive Bayes	***Complement***	***wine-pca***	***42.75918754071179***
Naive Bayes	Complement	wine-lda	42.44994374370818
Decision Tree	{'max_depth': 6}	wine-pca	45.576171
Decision Tree	***{'max_depth': 1}***	***wine-lda***	***53.1946468***
Decision Tree	{'max_depth': 4}	wine-raw	49.9779831
Random Forest	{'max_depth': 10, 'n_estimators': 183}	wine-pca	48.9016166
Random Forest	***{'max_depth': 4, 'n_estimators': 103}***	***wine-lda***	***54.4722805***
Random Forest	{'max_depth': 8, 'n_estimators': 163}	wine-raw	49.5327175

Best Pipeline:

1. **Abalone:** Among all, Random Forest on abalone without performing dimensionality reduction performed the best. LDA outperformed PCA in all models except for Decision Trees
2. **Wine:** Among all, applying LDA transformation and then classifying using KNN has performed the best. Even between the same classifier, the LDA transformed datasets have performed consistently better than raw or PCA transformed.

Effect of Dimensionality Reduction (on Abalone dataset):

a. **KNN :** A negligible improvement in performance (+0.8%) was observed for LDA
b. **Naive Bayes:** For Gaussian NB, both LDA(~2%) and PCA(~0.1%) performed better by a small margin after dimensionality reduction. For Multinomial NB, all the 6 datasets were observed to perform poorly with only a slight variation in the accuracy scores. For Complement NB, considerable variation in the accuracy scores were seen - Raw(14.74%) < PCA (19.29%)< **LDA (23%)**
c. **Decision Trees:** A slight degradation in performance (-1%) was observed for LDA while the increase was negligible for PCA.
d. **Random Forest:** A slight degradation in performance was observed for both LDA and PCA as compared to the raw dataset
e. **Gradient Tree Boosting:**. A slight degradation in performance was observed for both LDA and PCA as compared to the raw dataset

Effect of Dimensionality Reduction (on Wine dataset):

a. **KNN :** There was negligible difference in performance (± 0.2%)
b. **Naive Bayes:** For Gaussian, significant difference in performance was observed - Raw (30.12%) < PCA (44.99) < **LDA (52.97%).** For Multinomial, a slight improvement was observed for both LDA and PCA. For Complement, both PCA and LDA performed much better than the raw dataset with LDA outperforming PCA
c. **Decision Trees:** Different in performance was observed, but it wasn't as significant as Naive Bayes - PCA (45.57%) < Raw (49.97%) < **LDA (53.19%)**
d. **Random Forest:** Different in performance was observed, as much as Decision Trees. PCA (48.90%) < Raw (49.53%) < **LDA (54.47%)**
e. **Gradient Tree Boosting:** Difference in performance was observed, but it wasn't as much significant (± 1.5%)

As a general trend, for the wine dataset, LDA transformation has performed the best among all the pipelines in most models except in Complement Naive Bayes where PCA performed better.

Additional interesting observations -

- In both the abalone and wine dataset, the tree-based models have performed better at lower values of depth (<10).