DS402 Assignment 6

Name - Gajraj Singh Chouhan
Roll No - B19130

Terminology

- RFC → Random Forest Classifier
- SVC → Support Vector Classifier

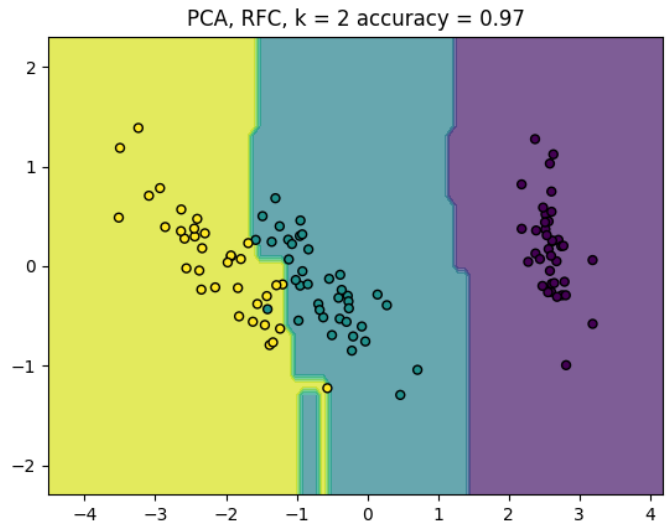
Class Labelling in each dataset

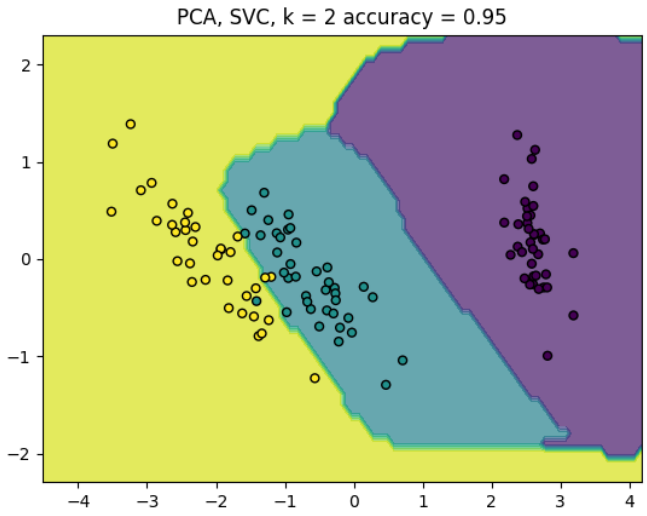
1. Iris →
    - Class Labels - setosa, versicolor, virginica
    - Dimensions - 4
    - Feature Names -
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm

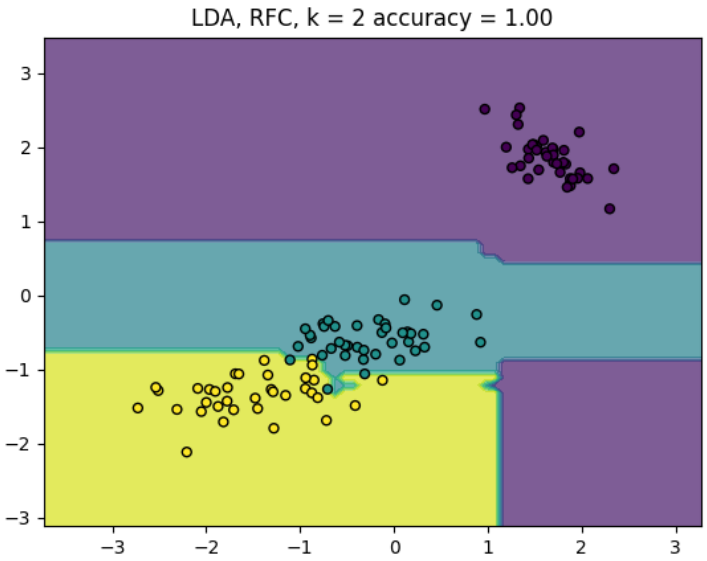
2. Wine →
    - Class Labels - class 0, class 1, class 2
    - Dimensions - 13
    - Feature Names -
        - Alcohol, Malic acid, Ash, Alcalinity of ash, etc.

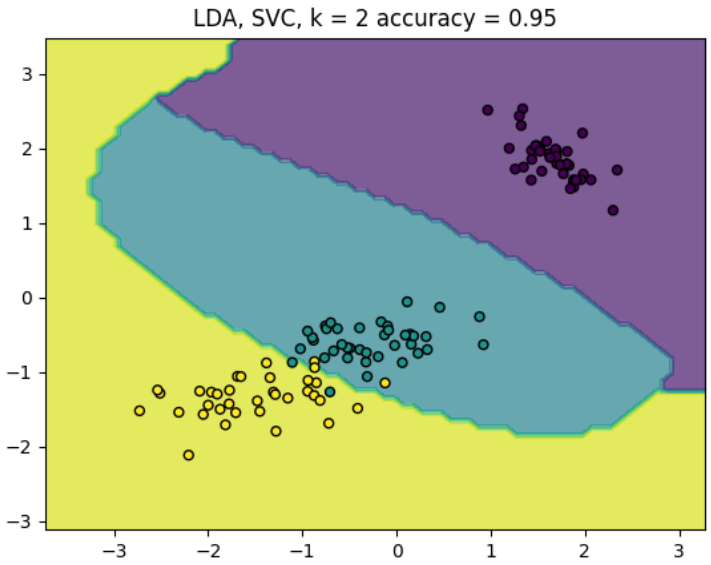
3. Breast Cancer →
    - Class labels - malignant, benign
    - Dimension - 30
    - Feature Names -
        - Radius, texture, perimeter, area etc...

Classification region of Iris dataset using k=2 for both SVM and Random Forest.









			Metrics	total_accuracy	accuracy		precision		recall		f1	
			Classes	total_accuracy	0	1	0	1	0	1	0	1
dataset	transformer	classifier	k									
load_breast_cancer	LDA	SVC	5	0.9231	0.8269	0.9780	0.9556	0.9082	0.8269	0.9780	0.8866	0.9418
		RFC	5	0.9371	0.9231	0.9451	0.9057	0.9556	0.9231	0.9451	0.9143	0.9503
	PCA	SVC	5	0.9441	0.8846	0.9780	0.9583	0.9368	0.8846	0.9780	0.9200	0.9570
		RFC	5	0.9301	0.8846	0.9560	0.9200	0.9355	0.8846	0.9560	0.9020	0.9457
	LDA	SVC	10	0.9301	0.8462	0.9780	0.9565	0.9175	0.8462	0.9780	0.8980	0.9468
		RFC	10	0.9371	0.8846	0.9670	0.9388	0.9362	0.8846	0.9670	0.9109	0.9514
	PCA	SVC	10	0.9580	0.9423	0.9670	0.9423	0.9670	0.9423	0.9670	0.9423	0.9670
		RFC	10	0.9231	0.8846	0.9451	0.9020	0.9348	0.8846	0.9451	0.8932	0.9399
	LDA	SVC	15	0.9231	0.8269	0.9780	0.9556	0.9082	0.8269	0.9780	0.8866	0.9418
		RFC	15	0.9441	0.9038	0.9670	0.9400	0.9462	0.9038	0.9670	0.9216	0.9565
	PCA	SVC	15	0.9580	0.9615	0.9560	0.9259	0.9775	0.9615	0.9560	0.9434	0.9667
		RFC	15	0.9161	0.8846	0.9341	0.8846	0.9341	0.8846	0.9341	0.8846	0.9341
	LDA	SVC	20	0.9371	0.8654	0.9780	0.9574	0.9271	0.8654	0.9780	0.9091	0.9519
		RFC	20	0.9441	0.9231	0.9560	0.9231	0.9560	0.9231	0.9560	0.9231	0.9560
	PCA	SVC	20	0.9231	0.9038	0.9341	0.8868	0.9444	0.9038	0.9341	0.8952	0.9392
		RFC	20	0.9301	0.8846	0.9560	0.9200	0.9355	0.8846	0.9560	0.9020	0.9457
	LDA	SVC	32	0.9301	0.8654	0.9670	0.9375	0.9263	0.8654	0.9670	0.9000	0.9462
		RFC	32	0.9441	0.9038	0.9670	0.9400	0.9462	0.9038	0.9670	0.9216	0.9565
	PCA	SVC	32	0.9231	0.9231	0.9231	0.8727	0.9545	0.9231	0.9231	0.8972	0.9385
		RFC	32	0.9231	0.8654	0.9560	0.9184	0.9255	0.8654	0.9560	0.8911	0.9405

			Metrics	total_accuracy	accuracy			precision			recall			f1		
			Classes	total_accuracy	0	1	2	0	1	2	0	1	2	0	1	2
load_wine	LDA	SVC	4	0.8667	0.7143	0.9500	0.9091	0.9091	0.8261	0.9091	0.7143	0.9500	0.9091	0.8000	0.8837	0.9091
		RFC	4	0.8667	0.8571	0.8000	1.0000	0.8571	0.8889	0.8462	0.8571	0.8000	1.0000	0.8571	0.8421	0.9167
	PCA	SVC	4	0.9778	1.0000	0.9500	1.0000	0.9333	1.0000	1.0000	1.0000	0.9500	1.0000	0.9655	0.9744	1.0000
		RFC	4	0.9333	0.8571	0.9500	1.0000	0.9231	0.9048	1.0000	0.8571	0.9500	1.0000	0.8889	0.9268	1.0000
	LDA	SVC	6	0.9556	0.9286	0.9500	1.0000	1.0000	0.9500	0.9167	0.9286	0.9500	1.0000	0.9630	0.9500	0.9565
		RFC	6	0.8667	0.8571	0.8500	0.9091	0.8571	0.8947	0.8333	0.8571	0.8500	0.9091	0.8571	0.8718	0.8696
	PCA	SVC	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		RFC	6	0.9778	1.0000	0.9500	1.0000	0.9333	1.0000	1.0000	1.0000	0.9500	1.0000	0.9655	0.9744	1.0000
	LDA	SVC	10	0.9778	1.0000	0.9500	1.0000	1.0000	1.0000	0.9167	1.0000	0.9500	1.0000	1.0000	0.9744	0.9565
		RFC	10	0.9778	1.0000	0.9500	1.0000	1.0000	1.0000	0.9167	1.0000	0.9500	1.0000	1.0000	0.9744	0.9565
	PCA	SVC	10	0.9778	0.9286	1.0000	1.0000	1.0000	0.9524	1.0000	0.9286	1.0000	1.0000	0.9630	0.9756	1.0000
		RFC	10	0.9778	1.0000	0.9500	1.0000	0.9333	1.0000	1.0000	1.0000	0.9500	1.0000	0.9655	0.9744	1.0000
	LDA	SVC	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		RFC	13	0.9778	0.9286	1.0000	1.0000	1.0000	0.9524	1.0000	0.9286	1.0000	1.0000	0.9630	0.9756	1.0000
	PCA	SVC	13	0.9778	0.9286	1.0000	1.0000	1.0000	0.9524	1.0000	0.9286	1.0000	1.0000	0.9630	0.9756	1.0000
		RFC	13	0.9778	1.0000	0.9500	1.0000	0.9333	1.0000	1.0000	1.0000	0.9500	1.0000	0.9655	0.9744	1.0000

			Metrics	total_accuracy	accuracy			precision			recall			f1		
			Classes	total_accuracy	0	1	2	0	1	2	0	1	2	0	1	2
dataset	transformer	classifier	k													
load_iris	LDA	SVC	2	0.9474	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333
		RFC	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	PCA	SVC	2	0.9474	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333	1.0000	0.9091	0.9333
		RFC	2	0.9737	1.0000	0.9091	1.0000	1.0000	1.0000	0.9375	1.0000	0.9091	1.0000	1.0000	0.9524	0.9677
	LDA	SVC	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		RFC	4	0.9737	1.0000	1.0000	0.9333	1.0000	0.9167	1.0000	1.0000	1.0000	0.9333	1.0000	0.9565	0.9655
	PCA	SVC	4	0.8947	1.0000	0.9091	0.8000	1.0000	0.7692	0.9231	1.0000	0.9091	0.8000	1.0000	0.8333	0.8571
		RFC	4	0.9737	1.0000	1.0000	0.9333	1.0000	0.9167	1.0000	1.0000	1.0000	0.9333	1.0000	0.9565	0.9655

From test.xlsx

Difference between PCA and LDA

PCA and LDA are two popular dimensionality reduction methods commonly used on data with too many input features. LDA is supervised whereas PCA is unsupervised – PCA ignores class labels. PCA as a technique that finds the directions of maximal variance. In contrast to PCA, LDA attempts to find a feature subspace that maximizes class separability. LDA makes assumptions about normally distributed classes and equal class covariances.

Comparison of classification results using PCA and LDA for each dataset.

The accuracy generally remains high with sometimes gradually increasing with dimensions (k) for all datasets even reaching 100%.

Breast Cancer == around 92-94 %

Wine == highs of 98%

Iris == average of ~96%