This homework focused on using different classification models on cross-validated data to pick the best model for classifying the olive trees area and regain based on the composition of 8 fatty acids. Methods which are used consisted of: Random Forest(RF), Quadratic Linear Discriminant Analysis(QDA), Linear Discriminant Analysis(LDA), K-Nearest Neighbor(KNN), Support Vector Machines(SVM), Naïve Bayes(NB), Decision Tree(DT).

Data preprocessing

There is no missing data within this dataset so there is no need to clean it beforehand. As some of the classifiers are not able to process categorical variables, only numerical attributes are used in defining the classification models. Data has been split into train and test data (70/30) randomly before passing to parameters tuning section to make sure that we keep an unseen data as the holdout test data for prediction purposes.

Parameter tuning

To assess the performance of our models on the training data and pick the most appropriate method, K-fold cross validation has been used. Number of folds have been considered as 10 and "traincontrol" function has been used. For each classification model the fold with highest accuracy would be used to fit the model. For K-NN method, a grid has been defined to assess the performance for different Ks in the range of (1,25).

Classification Models

Area

Accuracy for all the methods in classification task has been summarized in Table 1. All of the classification models except random forest defined by using train function in "CARET" library so we could make a fair comparison about which one should be used. Random forest model has been defined by number of trees equal to 33 and two variables have been used at each split. Although most of the classification models perform very well in classifying the area based on predictors, random forest model has been used to predict the area on the test. Random forest is not impacted by overfitting like other methods and therefore it would be wise to use it on unseen data.

Method	LDA	QDA	KNN	NB	SVM	DT	RF
Accuracy	99.01%	100%	97.03%	96.03%	99.75%	99.5%	100%

Table 1:Classification Accuracy of training data to predict areas

Variable importance plot of random forest model has been shown in Figure 1. According to this Figure, eicosenoic and oleic are the two most important variables in building the random forest models. Out of bag error (OOB) is around 0.25%. The confusion matrix of RF model over the training data has been shown in Table 2. By considering OOB error, it is found that RF reaches accuracy around 99.9975 ≅ 100%.

	North	Sardinia	South	error
North	106	0	0	0
Sardinia	0	70	0	0
South	1	0	0	0.004

Table 2: Confusion matrix of random forest model on Train data

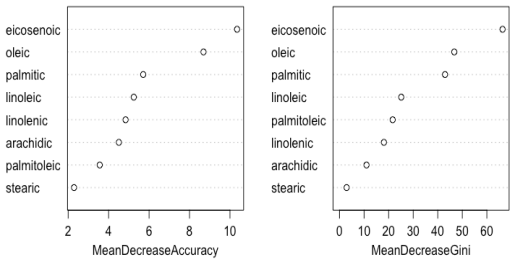

Figure 1:Variable Importance Plot

Random Forest ran on the test data with 167 observations. Comparison of the predicted versus actual lables of test data has been shown in Table 3 which is the confusion matrix. The accuracy of RF method in predicting the area is 100% as shown in Table 3.

	North	Sardinia	South	error
North	45	0	0	0
Sardinia	0	28	0	0
South	1	0	94	0

Table 3: RF model confusion matrix on test data

Region

Accuracy for all the methods in classification task has been summarized in Table 4. All of the classification models except random forest defined by using train function in "CARET" library so we could make a fair comparison about which one should be used. Random forest model has been defined by number of trees equal to 33 and two variables have been used at each split. Random forest model outperforms other methods so it has been used to predict the region on the test data.

Method	LDA	QDA	KNN	NB	SVM	DT	RF
Accuracy	93.6%	96.65%	90.85%	94.31%	95.79%	62.52%	100%

Table 4: Classification Accuracy of training data to predict regions

Variable importance plot of random forest model has been shown in Figure 2. According to this Figure, linoleic, oleic and palmitoleic are the three most important variables in building the random forest models. Out of bag error (OOB) is around 5.43%. The confusion matrix of RF model over the training data has been shown in Table 4. By considering the class error and OOB error, it is found that RF reaches accuracy around 99.96≅ 100%.

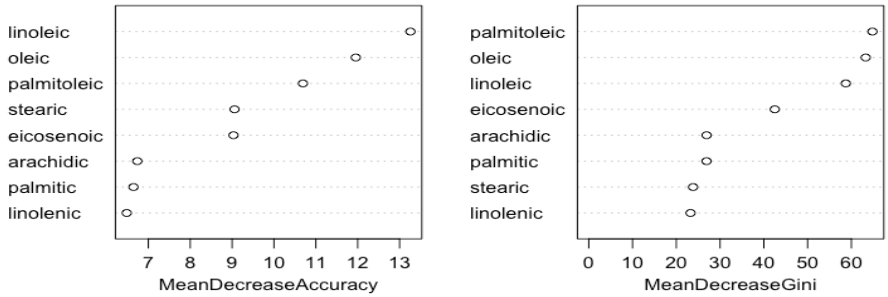

Figure 2:Variable Importance plot

	ApuliaN	ApuliaS	Calabria	LiguriaE	LiguriaW	SardiniaC	SardiniaIn	Sicily	Umbria	error
ApuliaN	17	0	1	0	0	0	0	0	0	0.0555
ApuliaS	0	142	1	0	0	0	0	2	0	0.0206
Calabria	0	1	38	0	0	0	0	1	0	0.05
LiguriaE	0	0	0	34	1	0	0	0	0	0.2857
LiguriaW	0	0	0	1	34	0	0	0	0	0.2857
SardiniaC	0	0	0	0	0	23	1	0	0	0.416
SardiniaIn	0	0	0	0	0	1	45	0	0	0.0217
Sicily	1	7	4	0	0	0	0	14	0	0.4615
Umbria	0	0	0	0	0	0	0	0	36	0

Table 4:Confusion matrix of RF over train data

Random Forest ran on the test data with 167 observations. Comparison of the predicted versus actual labels of test data has been shown in Table 5 which is the confusion matrix. The accuracy of RF method in predicting the region is 95.2% which is still performing good although the number of classes increased.

Predicted\Actual	ApuliaN	ApuliaS	Calabria	LiguriaE	LiguriaW	SardiniaC	SardiniaIn	Sicily	Umbria
ApuliaN	6	0	0	0	0	0	0	1	0
ApuliaS	0	61	1	0	0	0	0	0	0
Calabria	1	0	15	0	0	0	0	3	0
LiguriaE	0	0	0	13	0	0	0	0	0
LiguriaW	0	0	0	1	15	0	0	0	0
SardiniaC	0	0	0	0	0	9	0	0	0
SardiniaIn	0	0	0	0	0	0	19	0	0
Sicily	0	0	0	0	0	0	0	6	0
Umbria	0	0	0	1	0	0	0	0	15

Table 5: region prediction confusion matrix

In the second part of the problem, it has been asked to derive a confusion matrix for Area prediction by using region prediction confusion matrix which has been shown in Table 5. There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria). By considering this information, confusion matrix shown in Table 6 will be derived for area prediction using the region information. For each area, the correct prediction and predictions for other region in the same area are summed up to build the confusion matrix in Table 6.
As the confusion matrix is roughly the same, we could conclude that we got the same amount of results after mapping more complex dataset to simpler one. However, we saw that one misclassification of North to South has been missed so we could conclude that it is possible for us to loose some information but in this problem it was not a big issue.

	North	Sardinia	South
North	45	0	0
Sardinia	0	28	0
South	0	0	94

Table 6: Area confusion matrix derived from region classification